

## ORIGINAL ARTICLE

# Genomic and phenotypic differentiation among *Methanosarcina mazei* populations from Columbia River sediment

Nicholas D Youngblut<sup>1,3</sup>, Joseph S Wirth<sup>1,4</sup>, James R Henriksen<sup>1,5</sup>, Maria Smith<sup>2</sup>, Holly Simon<sup>2</sup>, William W Metcalf<sup>1</sup> and Rachel J Whitaker<sup>1</sup>

<sup>1</sup>Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL, USA and <sup>2</sup>Division of Environmental and Biomolecular Systems, Institute of Environmental Health, Oregon Health and Science University, Portland, OR, USA

**Methanogenic archaea are genotypically and phenotypically diverse organisms that are integral to carbon cycling in anaerobic environments. Owing to their genetic tractability and ability to be readily cultivated, *Methanosarcina* spp. have become a powerful model system for understanding methanogen biology at the cellular systems level. However, relatively little is known of how genotypic and phenotypic variation is partitioned in *Methanosarcina* populations inhabiting natural environments and the possible ecological and evolutionary implications of such variation. Here, we have identified how genomic and phenotypic diversity is partitioned within and between *Methanosarcina mazei* populations obtained from two different sediment environments in the Columbia River Estuary (Oregon, USA). Population genomic analysis of 56 *M. mazei* isolates averaging <1% nucleotide divergence revealed two distinct clades, which we refer to as ‘mazei-T’ and ‘mazei-WC’. Genomic analyses showed that these clades differed in gene content and fixation of allelic variants, which point to potential differences in primary metabolism and also interactions with foreign genetic elements. This hypothesis of niche partitioning was supported by laboratory growth experiments that revealed significant differences in trimethylamine utilization. These findings improve our understanding of the ecologically relevant scales of genomic variation in natural systems and demonstrate interactions between genetic and ecological diversity in these easily cultivable and genetically tractable model methanogens.**

The ISME Journal (2015) 9, 2191–2205; doi:10.1038/ismej.2015.31; published online 10 March 2015

## Introduction

Methanogenic archaea comprise a phylogenetically and ecologically diverse group of microorganisms distributed across a wide range of environments (Ferry, 1993). These organisms have a major role in the global carbon cycle, producing an estimated 1 billion tons of methane per year (Thauer, 1998), with *ca* 4% of primary productivity being channeled through methane (Field *et al.*, 1998). Although physiological studies have shown that methanogens vary considerably in substrate utilization, salinity tolerance and other ecological characteristics

(Hedderich and Whitman, 2006), current models of biogenic methane flux do not take methanogen diversity into account (Wania *et al.*, 2010; Riley *et al.*, 2011; Ito and Inatomi, 2012).

Large-scale comparative genomics of closely related strains has become a powerful approach for generating testable hypotheses of the ecological and evolutionary importance of genetic variation (Coleman and Chisholm, 2010; Cadillo-Quiroz *et al.*, 2012; Shapiro *et al.*, 2012; Cornejo *et al.*, 2013). This population genomics approach can only be applied to organisms that are easily cultivable in the laboratory, because linkage between loci is necessary but difficult to obtain using cultivation-independent metagenomic analyses of complex communities. In addition, genetically tractable strains are required for validation of putatively adaptive loci identified through comparative genomics (Ellison *et al.*, 2011).

Currently, population genomic approaches have only been applied to a select few model bacteria and archaea (Shapiro and Polz, 2014), with the exclusion of methanogens, which is likely due to the notorious difficulty in culturing and manipulating these strict anaerobes (Mah and Smith, 1981).

Correspondence: R J Whitaker, Department of Microbiology, University of Illinois at Urbana-Champaign, 601 South Goodwin Avenue, Urbana, IL 61801, USA.  
E-mail: rwhitaker@life.illinois.edu

<sup>3</sup>Current address: Department of Crop and Soil Sciences, Cornell University, 306 Tower Road, Ithaca, NY, USA.

<sup>4</sup>Current address: Department of Microbiology, University of Georgia, 120 Cedar Street, Athens, GA 30602, USA.

<sup>5</sup>Current address: AgBiome, PO Box 14069, Research Triangle Park, NC, USA.

Received 8 October 2014; revised 8 January 2015; accepted 28 January 2015; published online 10 March 2015

For many reasons, the *Methanosarcina* genus provides a tractable experimental system for applying population-level comparative genomics to methanogens. First, members of the genus can be readily and specifically cultivated. Second, the genomes of three *Methanosarcina* type strains (*M. acetivorans* C2A (Sowers *et al.*, 1984), *M. barkeri* Fursaro (Kandler and Hippe, 1977) and *M. mazei* Gö1 (Eggen *et al.*, 1992)) have been extensively studied, and 27 additional *Methanosarcinales* genome assemblies have recently been completed. These data provide a powerful resource for comparative genomics. Third, the genetic system of *Methanosarcina* is highly developed (Kohler and Metcalf, 2012), allowing for explicit testing of hypotheses generated by comparative assessments of natural variation (as applied in Ellison *et al.*, 2011).

Although much is not known of the mechanisms that create and maintain *Methanosarcina* diversity in any given environment, the distinct metabolism of the clade likely has a role. Typical methanogens utilize a limited number of substrates, most commonly reducing CO<sub>2</sub> with H<sub>2</sub> or formate as the electron donor (Ferry, 1993). In contrast, *Methanosarcina* species can utilize a variety of methanogenic pathways, allowing for growth on substrates such as acetate, methanol, trimethylamine (TMA) and H<sub>2</sub>:CO<sub>2</sub> (Kendall and Boone, 2006; Liu and Whitman, 2007). Interestingly, this metabolic diversity varies across the genus, and at least one member, *M. acetivorans* C2A, has seemingly undergone a recent loss of function that prevents growth on H<sub>2</sub>:CO<sub>2</sub> (Kendall and Boone, 2006; Guss *et al.*, 2009). We hypothesize that the potential for plasticity in resource utilization may offer the *Methanosarcina* a basis for divergent evolution through niche partitioning.

Other non-exclusive processes have likely contributed to the creation and maintenance of *Methanosarcina* diversity. For instance, comparative genomics of *M. barkeri* Fursaro, *M. acetivorans* C16 and *M. mazei* Gö1 revealed that large fluctuations in genome size (up to ~1.7 Mb) were likely caused by gene insertions, localized inversions and transpositions (Maeder *et al.*, 2006). In addition, gene gain from bacterial taxa is common in at least some *Methanosarcina* spp. and may often be adaptive (Deppenmeier *et al.*, 2002; Fournier and Gogarten, 2007). Host mobile element dynamics may also have a key role, given that *Methanosarcina* genomes contain a large number of putative mobile element genes and all contain multiple clustered regularly interspaced short palindromic repeats (CRISPRs) (Maeder *et al.*, 2006; Nickel *et al.*, 2013).

In this study, we used a comparative genomics approach to elucidate how genomic and phenotypic variation was partitioned within and between two populations of *Methanosarcina mazei*, with the goal discerning the ecological and evolutionary implications of such variation. The experimental system consisted of three sites along the Columbia River Estuary that varied in salinity and other geochemical

parameters (see Supplementary Materials and Methods). We show evidence that niche partitioning, though adaptive genomic evolution, had a role in creating and maintaining *M. mazei* diversity.

## Materials and methods

### Sample collection

On 22 August 2011 during low tide, sediment samples were collected near the shore at three locations in Youngs Bay and Baker Bay in the Columbia River Estuary as described in the Supplementary Materials and Methods.

### Strain isolation

Direct plating with agar overlays under strictly anaerobic conditions was used for initial *Methanosarcina* strain cultivation. Three sediment dilutions (10<sup>0</sup>, 10<sup>-1</sup> and 10<sup>-2</sup>) were plated on bicarbonate-buffered marine media (Metcalf *et al.*, 1996) or freshwater PIPES-buffered media (described in Supplementary Materials and Methods) containing either 40 mM acetate, 60 mM methanol or 50 mM TMA.

### Genomic sequencing and assembly

Multiplexed libraries were prepared using the Nextera XT DNA Sample Prep Kit (Illumina, San Diego, CA, USA) and sequenced using a HiSeq2000 sequencer (Illumina). Paired-end reads were quality filtered with the FASTX Toolkit v0.0.13, randomly subsampled to one million read pairs per sample (~40–50X coverage) and then assembled as described in the Supplementary Materials and Methods. All isolate genomes were assembled *de novo* in order to avoid biases in synteny and assembly of novel gene content that often result from referenced-based assembly. Putative genomic islands were identified with IslandViewer (Langille and Brinkman, 2009). CRISPRs were identified with CRISPRFinder (Grissa *et al.*, 2007) and classified in accordance with Vestergaard and colleagues (Vestergaard *et al.*, 2014). *cas* genes were manually annotated by searching NCBI nr database via BLASTp and searching the protein family databases CDD (Marchler-Bauer *et al.*, 2005), Pfam (Finn *et al.*, 2008) and COG (Tatusov *et al.*, 2000) with HHsearch (Söding, 2005).

### Whole-genome alignment

Whole-genome alignments (WGAs) were created with mugsy v1.2.3 (Angiuoli and Salzberg, 2011). RAXML v7.2.6 was used to infer a phylogenetic tree from all core (found in all taxa) local collinear blocks (LCBs) in the WGA. The *M. mazei* reference strain genomes were aligned with progressive Mauve v2.3.1 (Darling *et al.*, 2010). Double-cut-and-join distances were calculated with Mauve v2.3.1 (Darling *et al.*, 2004).

### Core and variable gene analysis

Genes were called and annotated using the Rapid Annotation using Subsystem Technology server (Aziz *et al.*, 2008). The ITEP toolkit (Benedict *et al.*, 2014) was used to group genes from all isolates and type strains into putative homologs through Markov Chain Clustering (via the MCL program) of BLASTp maximum bitscore ratios (0.4 cutoff, 2.0 inflation parameter) (Enright *et al.*, 2002). Amino-acid sequences of gene clusters were aligned with mafft v7.037b (Katoh and Standley, 2013), reverse translated with PAL2NAL v14 (Suyama *et al.*, 2006) and used to infer maximum likelihood phylogenies with RAxML v7.2.6 (Stamatakis, 2006). Quantification of dN/dS, mean sequence identity and  $F_{ST}$  values for core genes was performed with SNAP, Mothur v1.24.0 and Arlecore v3.5.1.3, respectively (Korber, 2000; Schloss *et al.*, 2009; Excoffier and Lischer, 2010). Both Mowgli (Nguyen *et al.*, 2013) and the Quartet Decomposition Server (Mao *et al.*, 2012) were used to assess inter-clade recombination of core genes.

### Data submission

The *mcrA* 454 pyrosequencing reads have been deposited in the sequence read archive with the accession numbers SAMN02709813–SAMN02709815. All 63 genome assemblies have been deposited at DDBJ/EMBL/GenBank under the accessions numbers JJOR00000000–JJRB00000000, and all raw shotgun Illumina reads for each culture were deposited in the sequence read archive with the accession number SRP040701.

### Statistics and plotting

Statistical evaluations were performed in R (R Development Core Team, 2010). The circular genome plots were created with Circos (Krzywinski *et al.*, 2009), all other plots were produced with R using the ggplot2 package (Wickham, 2009). All custom Perl scripts used in this study are available at [https://github.com/nyoungb2/pop\\_genome](https://github.com/nyoungb2/pop_genome).

### Methane production assays

Methane production, as proxy for culture growth, was monitored using a Hewlett Packard 5890 Series II gas chromatograph (Hewlett Packard, Wilmington, DE, USA) with a flame ionization detector and a column of stainless steel filled with 80/120 Carbo-pack B/3% SP-1500 (Supelco, Bellefonte, PA, USA) heated to 225 °C.

Further methodological details are provided in the Supplementary Materials and Methods.

## Results

### A spatially heterogeneous distribution of methanogens

Samples were collected from three sites in the Columbia River Estuary, which will be referred to as

‘Youngs Bay—Back’ (YBB) near the freshwater inlet of Youngs River, ‘Youngs Bay—Mouth’ (YBM) near the mouth of Youngs Bay, and ‘Baker Bay’ (BB) on the northwest side of BB. YBB and YBM are separated by ~4 km and are ~20 km and ~23 km from BB, respectively. The geochemistry of the Youngs Bay sediment samples was relatively similar as compared with the BB sample. The BB sample had a higher pH (7.5 vs 6.6 and 6.3), and contained substantially less  $\text{NH}_4^+$  and Fe. In addition, these samples establish a salinity gradient ranging from highest in BB to lowest in YBB (see Supplementary Materials and Methods). In all, 454 pyrosequencing of methyl co-enzyme M reductase (*mcrA*) amplicons revealed large differences in methanogen community composition and structure (Supplementary Figure 1). A single taxon in the *Methanoregula* genus dominated both sites in Youngs Bay, while the BB site harbored a relatively large fraction of *Methanosarcina* taxa (49% of all reads in the sample) that were evenly distributed across the genus (Supplementary Figure 1).

Of 128 *Methanosarcina* isolates obtained from the three sites (Supplementary Figure 2), 40% belonged to the *M. mazei* clade. Most of the *M. mazei* isolates originated from the Youngs Bay samples (YBM: 26, YBB: 26, BB: 4). Although the clade was relatively rare in the environment (Supplementary Figure 1), the large number of closely related *M. mazei* isolates provided a powerful data set that could be effectively used to resolve the recent evolution of genomic and phenotypic variation within this clade. Thus, we focused on this cohort of *M. mazei* isolates for our comparative genomics analyses.

### Inter-clade differentiation by substrate

Illumina paired-end sequencing and *de novo* genome assembly of all 56 *M. mazei* isolates produced draft assemblies of similar quality (Table 1), with on average, an N50 of 42.5 kbp  $\pm$  1 (s.d.), 3937 coding sequences  $\pm$  57 (s.d.) and a genome length of 4.07 Mb  $\pm$  0.05 (s.d.). Genome assembly fragmentation was likely caused by multiple factors including transposases, mobile elements, other paralogous genes and repeat elements (Supplementary Table 3). A WGA of these assemblies with seven closed *M. mazei* genomes produced 1331 ‘core’ (found in all taxa) LCBs  $\geq$  500 bp in length, representing 71% of the mean genome length (2.89 of 4.07 Mb). The high number of LCB breakpoints suggests that the genome assemblies often varied in assembly breakpoint positions. In addition, this highly fragmented WGA reflects a loss of synteny within the *M. mazei* clade, with large rearrangements and inversions in the *M. mazei* genomes (Supplementary Figure 3). A dendrogram of double-cut-and-join distance values (a measure of synteny) showed that the loss of synteny increased with genetic distance, with *M. mazei* TMA representing one extreme (Supplementary Figure 3).

A phylogeny inferred from all core LCBs revealed that the isolates from both samples were distributed

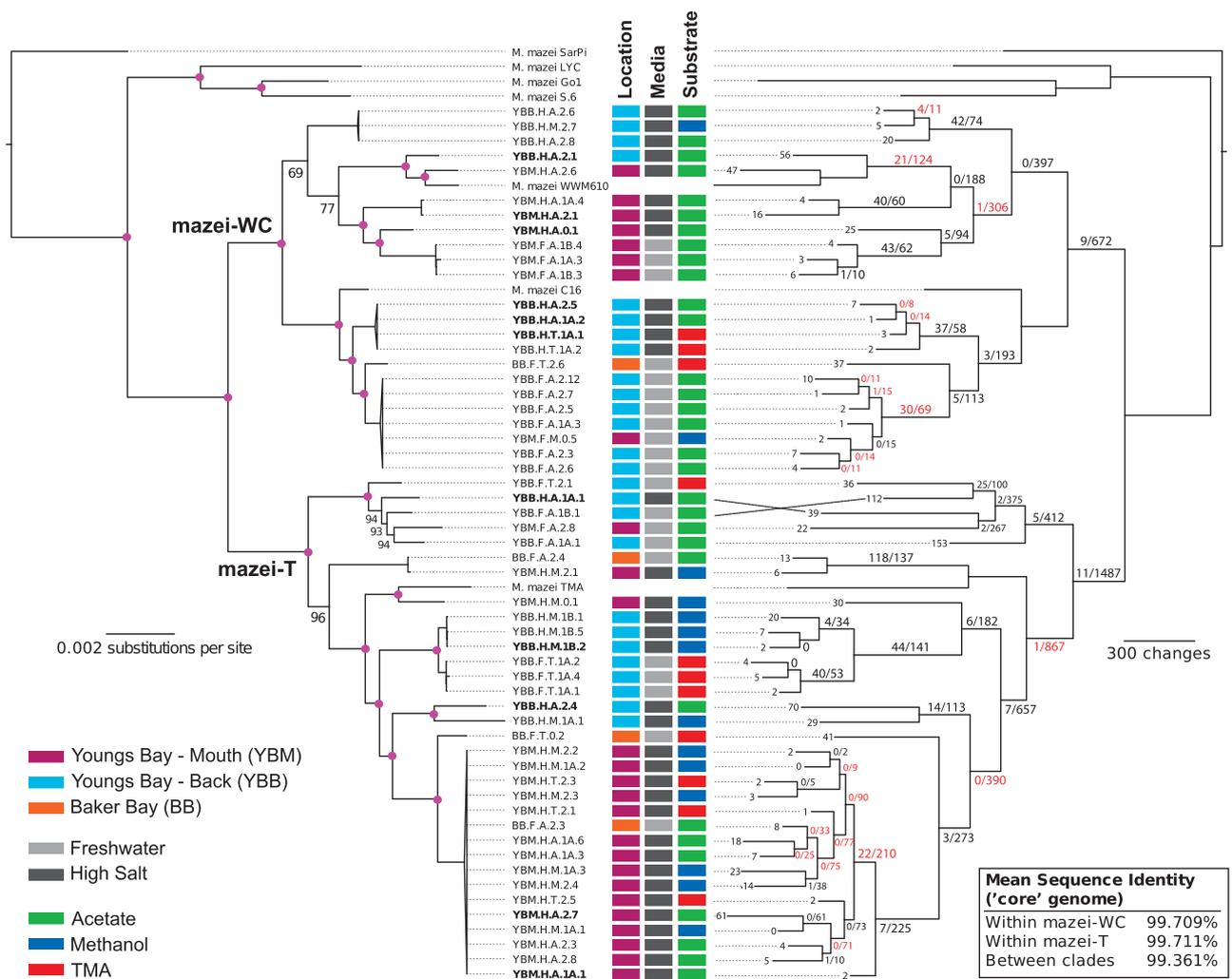
between two major, well-resolved clades, with ~99.4% average nucleotide identity between the clades and ~99.7% average nucleotide identity within each clade (Figure 1). The first clade harbored the reference strains *M. mazei* WWM610 and *M.*

*mazei* C16 ('mazei-WC' clade), whereas the second clade contained *M. mazei* TMA ('mazei-T' clade). Within both of these clades, many isolates shared nearly clonal core genomes, with only tens to hundreds of single-nucleotide polymorphisms

**Table 1** Genome assembly summary of the 56 *M. mazei* isolates

	Number of scaffolds	Number of contigs	N50 (kb)	Maximum scaffold length (kb)	Total length (Mb)	Number of CDS	Coverage
Minimum	123	151	20.3	67	3.97	3817	42.3
Mean	187	215	42.5	170	4.07	3937	44.2
Median	167	196	43.3	155	4.08	3940	44.1
Maximum	374	446	77.1	433	4.23	4068	45.4
S.D.	56	58	1.0	78	0.05	57	0.7

Abbreviation: CDS, coding sequence. See Supplementary Table 1 for individual assembly summaries. 'Coverage' is defined as the mean number of reads mapped to each nucleotide position in the assembly.



**Figure 1** The tree on the left is a maximum likelihood inference (GTR- $\Gamma$  model; 100 bootstrap replicates) from WGA for all *M. mazei* isolates ( $n = 56$ ) and type strains. Purple nodes have a bootstrap support of 100. Isolation metadata is mapped onto the tree (see the legend). Bold leaf labels highlight isolates used for methane production assays (see Figure 6). The tree on the right is a maximum parsimony inference of gene presence-absence (100 bootstrap replicates). Node labels indicate (genes specific to and found in all descendants)/(genes specific to and found in any descendants). Red node labels highlight nodes with a bootstrap support of <70. 'BB', 'YBM', 'YBB' identify cultures from Baker Bay, Youngs Bay (at the mouth) and Youngs Bay (near the back), respectively. 'F' and 'H' identify cultures isolated on freshwater or marine media, respectively. 'A', 'M' and 'T' identify isolates grown on acetate, methanol or TMA, respectively.

(SNPs) differentiating these strains. Nearly clonal isolates were predominantly, but not entirely, isolated from the same location (for example, isolate YBM.F.M.0.5 in Figure 1). The prevalence of the clades did not vary significantly by location ( $\chi^2$ ,  $P$ -value < 0.161) or media ( $\chi^2$ ,  $P$ -value < 0.168), but did by substrate ( $\chi^2$ ,  $P$ -value < 0.005). Strains in mazei-WC were primarily isolated using acetate, whereas similar numbers of mazei-T strains were isolated on each substrate (Figure 1).

We excluded the BB isolates from further analyses because of their under-representation in the *M. mazei* clade, which could bias quantitative analyses.

#### Gene flow between clades

Of the putative orthologous genes present among the 52 *M. mazei* isolates from Youngs Bay, 2144 were found in a single copy in all genomes ('core') and 4112 varied in copy number and/or were absent in some isolates ('variable'). Of the core genes, 67 had poor alignments likely due to artificial gene truncations caused by unassembled genome regions (see Supplementary Materials and Methods) and were excluded from subsequent analyses.

To explore the impact of horizontal gene transfer on the evolutionary trajectory of mazei-WC and mazei-T, we inferred inter-clade (mazei-WC vs mazei-T) transfer of core genes. Tree reconciliation with Mowgli (Nguyen *et al.*, 2013) on all well-supported gene trees ( $n = 29$ ) supported a scenario of transfer for all but one gene, indicating continued gene flow between clades (Supplementary Table 4). Manual assessment of these gene trees suggested a true signal of gene transfer in all but four trees, where two had low bootstrap support ( $\leq 70$ ) for the nodes inferred as transferred, while two others appeared to contain falsely clustered paralogs. Only 1 of these 28 genes was located within a putative genomic island.

In support of these findings, 661 of 1341 (49.3%) suitable core gene trees were found to contain  $\geq 1$  well-supported quartet (bootstrap value of  $> 70$ ) that was incongruent with the mazei-T and mazei-WC bifurcation. Manual inspection of a subset of these gene trees indicated that this signal was likely a result of both recombination and incomplete lineage sorting. Of the 661 gene trees with incongruent quartets,  $< 3\%$  were located in a putative genomic island.

#### Partitioned variation across the core genome

Diversifying selection acting within a population will increase genetic variation between sub-populations, while decreasing variation within (that is, fixation) (Hartl and Clark, 1998). To assess the level of genetic variation and fixation between mazei-T and mazei-WC, we calculated the mean sequence identity and  $F_{ST}$  values for all core genes and mapped them onto the reference genome *M. mazei* C16 (Figure 2). Levels of genetic divergence and fixation varied across the genome, with certain

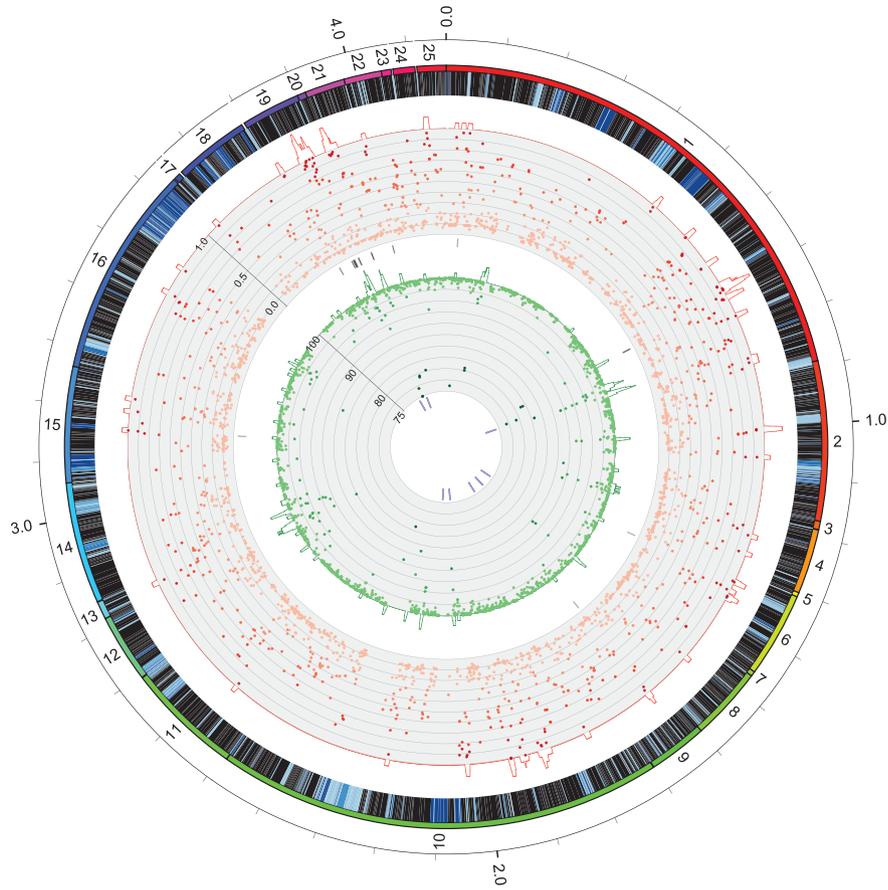
narrow regions containing relatively high numbers of (nearly) fixed genes ( $F_{ST} > 0.9$ ) and/or highly divergent genes ( $< 97\%$  sequence identity). Throughout the genome, a total of 15 genes were both (nearly) fixed and highly divergent (Table 2). Putative annotations of these genes consisted of three molybdenum-containing formylmethanofuran dehydrogenase subunits (FmdA, FmdB and FmdE), three molybdate transporter proteins (ModA, ModB and ModE), a uridylate kinase, a type II secretion protein, a universal stress protein and multiple hypothetical proteins (Table 2). Six of the 15 genes were located within 26 kbp of each other. Furthermore, these six genes were present in an  $\sim 80$  kbp region ( $\sim 3.85$ – $3.93$  Mb on *M. mazei* C16) that contained the highest density of fixation and genetic divergence, with 16 (nearly) fixed genes and 55 divergent genes (Figure 2). These genes also colocalized in an  $\sim 80$  kbp region on *M. mazei* WWM610 and *M. mazei* TMA.

To test for evidence of diversifying selection between clades in these and other genomic regions, we calculated relative rates on nonsynonymous to synonymous substitutions (dN/dS) for each core gene between mazei-WC and mazei-T. The median dN/dS value between the two *M. mazei* clades was  $0.121 \pm 0.374$  (s.d.). Eight core genes had a dN/dS value of  $> 1$ , although none of these were also (nearly) fixed and divergent. Most of these genes had hypothetical annotations, although a putative phosphoglycerate mutase and an endonuclease III were among them (Supplementary Table 5). We also calculated dN/dS between all *M. mazei* isolates and the *de novo* assembled genomes of seven isolates most closely related to *M. lacustris* strains (Supplementary Figure 1—'\*\*\*'; Supplementary Table 2) to assess how dN/dS scaled with evolutionary distance (Kryazhimskiy and Plotkin, 2008). Incorporating the *M. lacustris*-like strains decreased detection of positive selection, with no genes showing a dN/dS value of  $> 1$ .

To assess the potential of geographic population structure within each clade, we subdivided the isolates by clade and location (YBM or YBB) and calculated  $F_{ST}$  values for all core genes. No signal of intra-clade geographic population structure was observed, with only 0 and 2 genes (nearly) fixed by location within mazei-WC and mazei-T clades, respectively.

#### Clade-specific gene content

Adaptive gene gain or loss specific to either mazei-T or mazei-WC would contribute to the genetic and ecological segregation between the clades and manifest as gene presence in all (or most) members in one clade but not the other. Alternatively, rapid local adaptation or niche convergence could manifest as gene presence and absence patterns following geographic location or phenotype instead of the evolutionary history of the core genome. We assessed these scenarios by examining the distribution of variable genes across the 52 *M. mazei* isolates. A maximum parsimony tree based on the presence and



**Figure 2** Genes and genomic regions of each *M. mazei* isolate mapped onto the genome of *M. mazei* C16. The outer ring with numbered labels designates LCBs shared among *M. mazei* C16, *M. mazei* WWM610, *M. mazei* TMA (as in Figure 3). The second-most inner ring designates core and variable regions among all *M. mazei* isolates, with black indicating core regions, dark shades of blue indicating presence in most genomes and light shades indicating presence in few genomes. The red and green line plot shows the density of core genes with an  $F_{ST}$  value  $> 0.9$  in a 10 kb window (maximum density of 5) and a mean sequence identity of  $< 97\%$  (maximum density of 6), respectively. The scatterplots of red or green dots displays the  $F_{ST}$  value for each core gene (value range of  $-0.1$  to  $1$ ) and mean sequence identity values (value range of  $75\%$  to  $100\%$ ), respectively. Black bars between these two plots highlight genes that have an  $F_{ST}$  value  $> 0.9$  and a mean sequence identity of  $< 97\%$ . Purple bars highlight genes with a dN/dS value of  $> 1$ .

absence of particular genes within each genome was congruent with the core gene phylogeny (Figure 1). Except for the highly similar strain clusters, the majority of branch length was distributed among terminal branches, implying a high level of strain-specific gene content. We found 60 and 192 genes specific to the mazei-WC and mazei-T clades, respectively (Table 3). Furthermore, 672 and 1487 variable genes were exclusively found in mazei-WC and mazei-T, although not necessarily present in (nearly) all members of the clade (Figure 1). Of the clade-specific genes, 75% (45 genes) and 72% (139 genes) had ambiguous ‘hypothetical’ annotations in the mazei-WC and mazei-T clades, respectively. None or very few ( $\leq 2$ ) genes were specific to location, media or substrate that were not also specific to a clade, suggesting that the distribution of these genes has followed the evolution of the core genome.

Mapping genes specific to mazei-WC or mazei-T onto the reference genomes revealed that many of these genes co-located within variable regions throughout the genome (Figure 3). For genes specific to mazei-WC, we found one instance of co-location

between multiple clade-specific genes and a putative genomic island. Eight such regions were observed for mazei-T; most of which were flanked by transposases, integrases and tRNAs, suggesting they were introduced by mobile genetic elements.

Clade-specific genes may have been acquired largely from closely related taxa or from more divergent taxa as have often been observed (Ochman *et al.*, 2000; Cohan and Koeppl, 2008). We used BLASTp (E-value  $< 1e-30$ ) to determine the best match of each clade-specific gene in NCBI’s non-redundant protein database. Bacterial homologs comprised 75% and 67% of the best hits for genes specific to mazei-WC and mazei-T, respectively. Among these, *Proteobacteria* and *Firmicutes* comprised a total of  $\sim 40\%$  of the best hits in each clade, with a majority of hits to members of the *Deltaproteobacteria*, *Clostridia* and *Bacilli*.

*Clade-specific defense systems*  
CRISPR-associated (Cas) and restriction modification (RM) genes were among the clade-specific genes that

**Table 2** The annotations of all (nearly) fixed and divergent core genes ( $F_{ST} > 0.9$ ; sequence identity  $< 97\%$ )

$F_{st}$	Sequence identity	$dN/dS$	Annotation	M. mazei C16		
				Start (bp)	End (bp)	Locus tag
0.96	76.2	0.99	Hypothetical protein	37 190	36 978	MSMAC_0031
0.95	81.3	0.14	Molybdate transport system permease protein (ModB)	716 776	716 111	MSMAC_0591
0.94	81.8	0.17	Molybdate ABC transporter (ModA)	717 652	716 959	MSMAC_0592
1.00	96.6	0.10	Hypothetical protein	1 332 362	1 332 541	MSMAC_1099
0.92	96.5	0.74	Hypothetical protein	1 626 669	1 626 457	MSMAC_1330
1.00	96.2	0.23	Hypothetical protein	3 159 383	3 159 144	MSMAC_2568
0.97	95.8	0.07	Uridylate kinase	3 810 286	3 809 585	MSMAC_3079
0.93	94.1	0.41	Hypothetical protein	<b>3 854 997</b>	<b>3 854 029</b>	<b>MSMAC_3114</b>
0.96	97.0	0.04	Formylmethanofuran dehydrogenase (molybdenum) subunit B (FmdB)	<b>3 860 164</b>	<b>3 858 864</b>	<b>MSMAC_3118</b>
0.99	96.9	0.05	Formylmethanofuran dehydrogenase (molybdenum) subunit A (FmdA)	<b>3 863 192</b>	<b>3 861 438</b>	<b>MSMAC_3121</b>
1.00	96.1	0.14	Formylmethanofuran dehydrogenase (molybdenum) subunit E (FmdE)	<b>3 864 821</b>	<b>3 864 252</b>	<b>MSMAC_3123</b>
0.94	76.9	0.10	S-layer domain	<b>3 878 347</b>	<b>3 877 286</b>	<b>MSMAC_3136</b>
1.00	74.6	0.67	Molybdenum transporter ATP-binding protein (ModE)	<b>3 879 523</b>	<b>3 879 362</b>	<b>MSMAC_3137</b>
0.98	96.8	0.04	Type II secretion system protein	3 923 368	3 921 692	MSMAC_3172
0.92	80.3	0.20	Universal stress protein	3 994 295	3 993 388	MSMAC_3233

The bold values highlight genes co-located in a ~25 kbp region.

could be identified by their putative annotations (Table 3). The clade specificity of *cas* genes appeared to be a result of differences in CRISPR subsystems between the clades (Figure 4). Specifically, subtype III-B was found in the majority of mazei-WC members, but only in two members of mazei-T, whereas subtypes I-D and III-A were solely present in members of the mazei-T clade. In addition, a novel variant classified here as subtype VIII-3 was found in all members of the mazei-T clade, except for the most basal members. Instead, these strains possessed subtype III-A and or III-B systems.

We assessed whether the spacer content in the repeat spacer arrays of these strains corresponded to the difference in CRISPR subsystems between clades. Of the 2240 unique spacer sequences, 3.5% and 4.2% were specific to the mazei-WC and mazei-T clades, respectively (Supplementary Table 6). No spacers were specific to YBB isolates. We found 4.1% to be specific to YBM isolates; however, these spacers were also specific to the mazei-T clade and all fell into the nearly clonal clade of 15 YBM isolates. No spacers were confined to isolates from specific media or growth substrates. In contrast, 17% of spacers were specific to a subtype. These data show that partial segregation of CRISPR subtypes between clades largely dictated spacer content. Within CRISPR subtypes, the leader end generally showed substantially more spacer content variation than the trailer end (Supplementary Figure 4), indicating a history of sequential additions from the leader end.

In accordance with segregated CRISPR diversity, each clade contained a specific suite of RM systems. RM genes were highly prevalent among the isolates, with 88 gene clusters annotated as single copy RM genes. The type I RM system genes specific to mazei-

WC were adjacent when mapped to each reference genome and located adjacently to a putative genomic island. The type I subunit R (restriction) gene specific to mazei-T was adjoining to S (specificity) and M (modification) subunits, and the type III subunit R gene was adjacent to an M subunit.

#### A putative clade-specific metabolic gene cassette

A set of mazei-WC specific genes was annotated as CoB-CoM heterodisulfide reductase subunits A and C (HdrAC), formate dehydrogenase subunits A and B (FdhAB), and a methylviologen-reducing hydrogenase (MvhD). These genes along with *hdrB* were immediately adjacent and always found in a conserved order: *fdhAB, mvhD, hdrABC* (~3.17 Mb in *M. mazei* C16). Moreover, the genes either overlapped or were separated by  $\leq 170$  bp (Figure 5a), suggesting a transcriptional unit. This putative operon was found in all members of mazei-WC, including *M. mazei* C16, but excluding *M. mazei* WWM610 and the two closest isolates, suggesting a recent loss in this subclade. This gene set was also found in a sparsely distributed set of *Methanosarcina* strains, namely *M. horonobensis* HB1 (Shimizu et al., 2011), *M. sp.* Naples 100 and *M. calensis* Cali. This sparse distribution among the *Methanosarcina* spp. indicates that either this putative operon was in the last common ancestor of the genus and lost in many separate events or was gained in separate events in multiple clades throughout the genus. Phylogenies of each of these six genes supported the former hypothesis, with each being either congruent with a concatenated core gene phylogeny or held low bootstrap support for nodes displaying incongruences.

**Table 3** Genes only found in all members of the mazei-T or mazei-WC clade after accounting for artificial absences (that is, present in at least half of a nearly clonal group of isolates)

Clade specificity	Number of genes	Annotation	Present in all members?
mazei-T	144	Hypothetical protein	Y (7)
	7	CRISPR-associated protein	
	4	Glycosyl transferase	
	3	DNA helicase	Y (1)
	3	vrlJ, vrlQ and vrlP	
	2	Chaperone protein dnaK	
	2	DNA sulfur modification protein (dndB and dndD)	
	2	duf324 domain-containing protein	
	2	Sensory transduction histidine kinase	
	1	3'-Phosphoadenosine 5'-phosphosulfate sulfurtransferase dndC	
	1	ATPase involved in DNA repair, sbcC	
	1	Cell surface protein	
	1	Dipeptide transport system permease protein dppB (tc 3.a.1.5.2)	
	1	Dolichol-phosphate mannosyltransferase (ec 2.4.1.83)	
	1	Endonuclease III (ec 4.2.99.18)	
	1	grpE protein hsp-70 cofactor	
	1	Helicase (snf2/rad54 family)	
	1	Huntingtin interacting protein e-like protein	Y
	1	Hypothetical protein bvu 3741	
	1	Lead, cadmium, zinc and mercury transporting ATPase (ec 3.6.3.3) (ec 3.6.3.5)	
	1	Methyltransferase (ec 2.1.1.-)	
	1	Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein oppA (tc 3.a.1.5.1)	
	1	Peptidase c14, caspase catalytic subunit p20	
	1	Putative ATP-binding protein	
	1	Putative serine/threonine protein kinase	
	1	Signal peptidase I (ec 3.4.21.89)	
	1	TPR repeat	
	1	Transposase	Y
	1	Type I restriction modification system, restriction subunit R (ec 3.1.21.3)	
	1	Type III restriction enzyme, res subunit	
	1	ubiE/coq5 methyltransferase	
	1	ycfA-like	Y
	mazei-WC	45	Hypothetical protein
3		Type I restriction modification system (subunits M, R and S)	Y (2)
2		CoB-CoM heterodisulfide reductase (hdrA and hdrC)	
2		Formate dehydrogenase (fdhA and fdhB)	
1		DNA-cytosine methyltransferase (ec 2.1.1.37)	
1		Endonuclease	
1		Internalin	
1		Methyl-viologen-reducing hydrogenase (mvhD)	
1		Retron-type RNA-directed DNA polymerase (ec 2.7.7.49)	
1		Tetratricopeptide TPR 2	
1		t/g-specific DNA glycosylase( ec:3.2.2.- )	
1		Transposase	Y

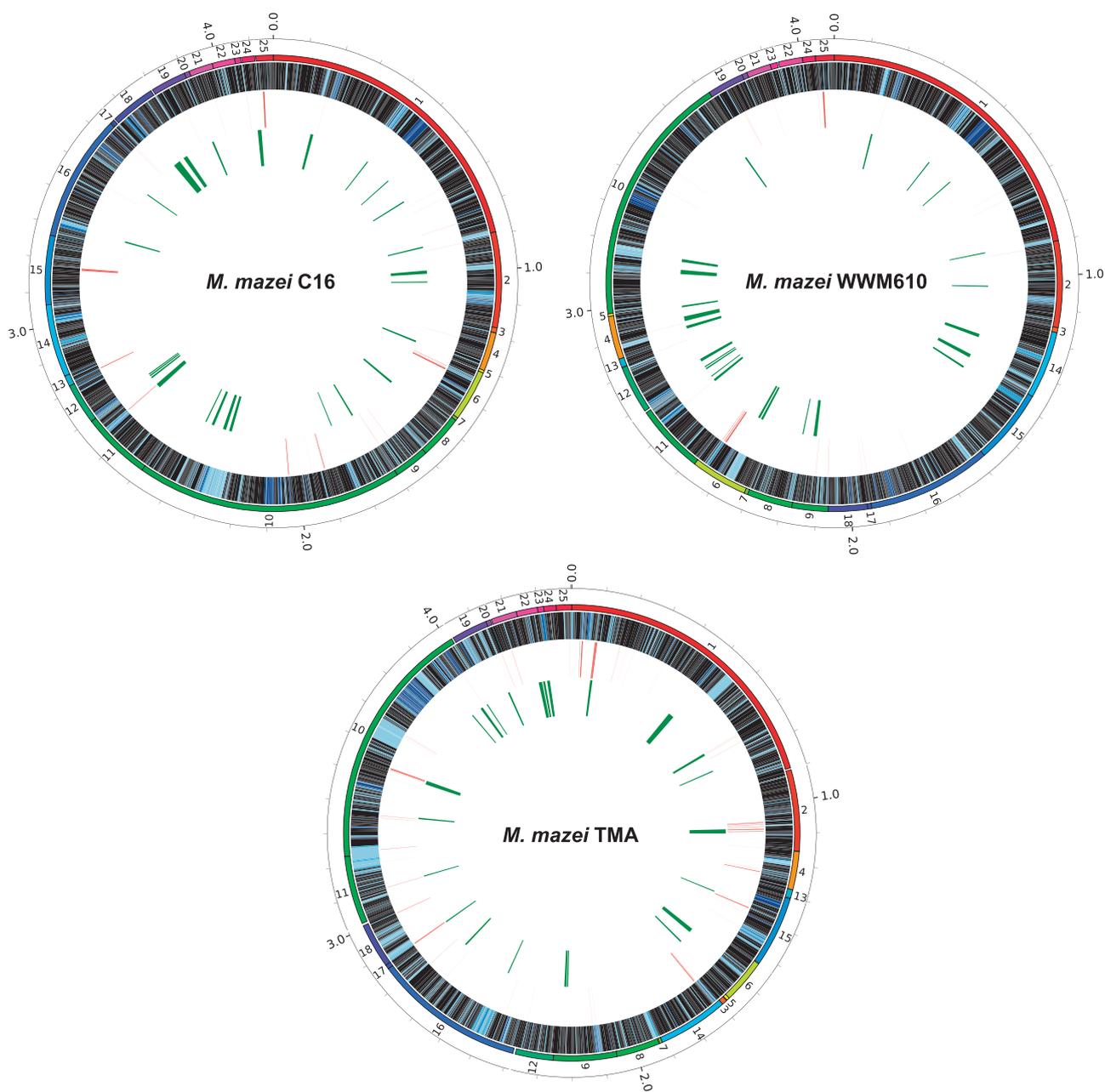
Numbers in parentheses in the fourth column refer to the number of genes present in all members.

All 52 *M. mazei* isolates possessed both of the Hdr paralogs found in other *Methanosarcina* spp. (Figure 5b). Unrooted phylogenies of each Hdr homolog showed that the HdrABC adjacent to FdhAB and MvhD appeared to be a third paralog (HdrA3B3C3) (Figure 5; Supplementary Figure 5). In addition, the genes were distant paralogs that seemed to have diverged before the inception of the *Methanosarcina* genus. Although divergent, the 4Fe-4 S, flavin adenine dinucleotide and zinc-binding domains of genes found in biochemically characterized HdrABC genes appeared to be conserved in HdrA3B3C3.

The MvhD adjacent to HdrA3B3C3 was divergent from the characterized delta subunit of methylviologen-reducing hydrogenase in *M. maripaludis*, with only 40% sequence identity. However, all four

cysteines were conserved in the 2Fe-2 S binding domain. FdhB was homologous to FdhB1 in *M. maripaludis* (BLASTp; E-value =  $2e-101$ ) although divergent with 43% sequence identity. The FdhB homologs in *Methanosarcinales* form a distinct clade from any methanogen clade lacking cytochromes (Supplementary Figure 6). In contrast to FdhB, FdhA does not appear to be homologous to any characterized FdhA and had no homology to any coding sequences outside of the *Methanosarcinales*.

*Differences in substrate utilization rates between clades*  
We tested the possibility of niche differentiation between the mazei-WC and mazei-T clades by directly comparing methane production rates of

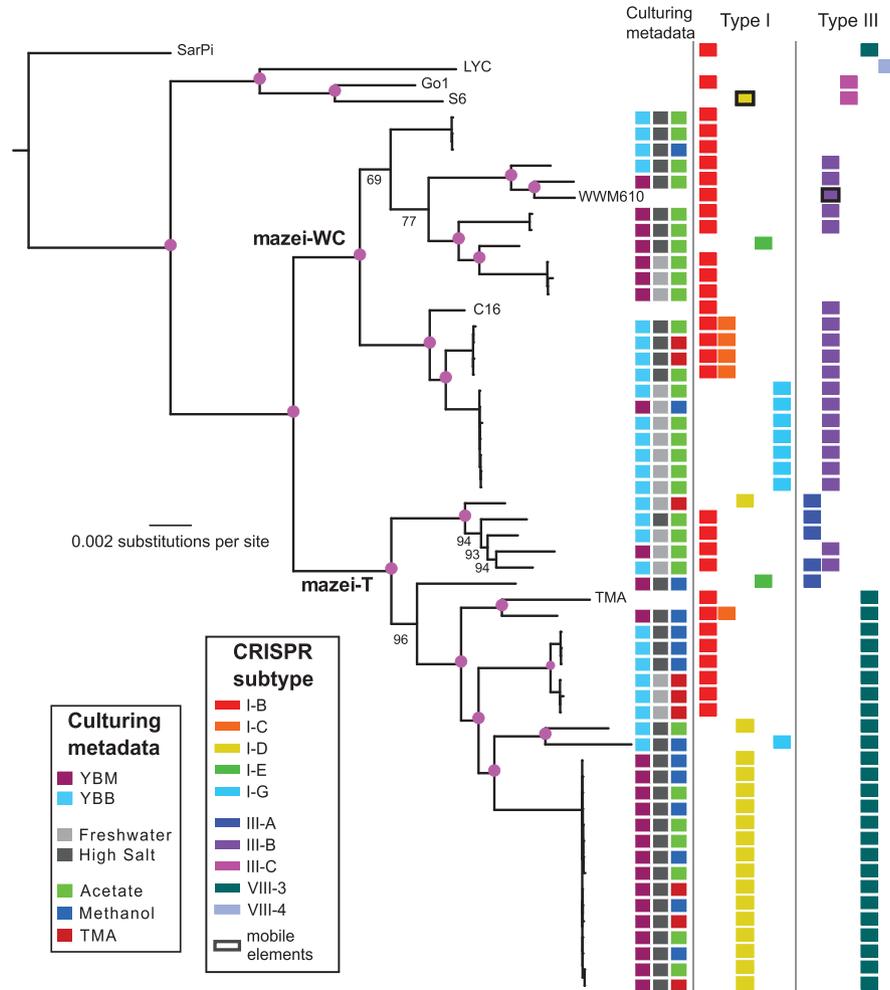


**Figure 3** Mazei-WC and mazei-T clade-specific gene clusters mapped onto closed reference genomes: *M. mazei* C16 and *M. mazei* WWM610 for mazei-WC, and *M. mazei* TMA, for mazei-T. The outer ring with numbered labels designates LCBs shared among the three genomes. The second-most inner ring designates clade-specific core and variable regions. Black regions are found in all isolates and references in the mazei-WC or mazei-T clades ('core'), whereas shades of blue indicate how many genomes in the clade contain that region, with darker shades indicating presence in more genomes. The red bars refer to clade-specific gene locations. The green bars highlight regions identified as potential genomic islands.

isolates from each clade when growing on acetate, methanol or TMA. Optical density could not be used to assess growth because most of the cultures formed macroscopic aggregates even in the marine medium (0.4 M NaCl). In an attempt to control for growth biases introduced during isolation, we performed direct comparisons between isolates from each clade that, when possible, were obtained from the same location and isolated on the same medium and substrate (Figure 6). We found no significant differences in maximum methane production rates between

the mazei-WC and mazei-T clades when isolates were grown on acetate or methanol (Supplementary Table 7). However, the mazei-WC and mazei-T isolates showed significantly different maximum methane production rates on TMA (Supplementary Table 7), with mazei-WC isolates generally showing substantially higher rates than mazei-T isolates.

We investigated the genetic basis of this phenotypic difference by assessing the degree of genetic differentiation and selection on genes shown to potentially affect TMA metabolism (Supplementary Table 8). We



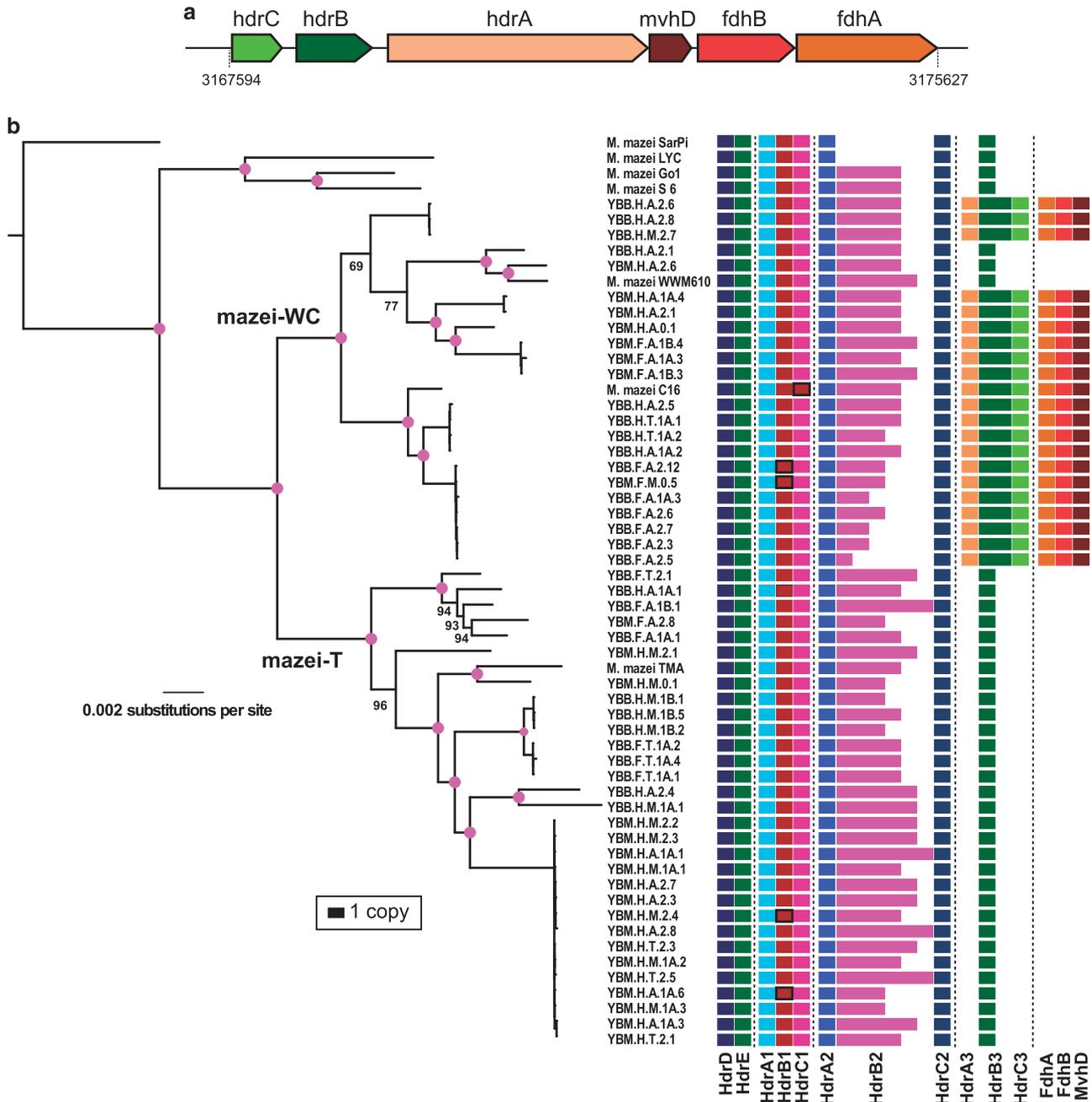
**Figure 4** CRISPR systems present in each isolate are mapped onto a maximum likelihood phylogeny similar to Figure 1 (BB isolates removed). CRISPRs are classified as in Vestergaard and colleagues (Vestergaard *et al.*, 2014). Black box outlines ('mobile elements') identify CRISPR systems in which *cas* gene continuity was disrupted by intervening putative mobile element genes.

analyzed genes involved in the methyl group transfer from methylated C1 compounds to CoM, which canonically involves substrate-specific methyltransferases (MT1) and more generalized methyltransferases (MT2) facilitating methyl group transfer from the substrate to a specific corrinoid protein and then to CoM (Paul *et al.*, 2000). In addition, substrate-specific permeases facilitate substrate uptake and putative regulatory genes control expression of the various metabolic pathways (Krätzer *et al.*, 2009). We note that dimethylamine (DMA) and subsequently monomethylamine are produced from TMA metabolism (Paul *et al.*, 2000; Bose *et al.*, 2006); therefore, genes involved in all three metabolic pathways may affect growth on TMA. Also, most of these focal genes exist as highly similar paralogs that fall into single gene clusters in our data set; thus, we manually divided paralogous clusters to accurately quantify differentiation and selection of homologs.

Although we did not observe differentiation or selection for TMA-specific genes (Supplementary Table 8), the gene encoding the DMA-specific

corrinoid protein MtbC2 was close to fixation ( $F_{ST} = 0.72$ ). Unexpectedly, one paralog for each step in the canonical methanol:CoM pathway (*mtaB1*, *mtaC1* and *mtaA1*) were close to fixation ( $F_{ST} = 0.8$ , 0.73 and 0.93). SNPs in these DMA- and methanol-specific genes were mostly synonymous SNPs ( $dN/dS \leq 0.15$ ). The *mtmC1* gene was the only instance of substantial gene length variation between clades; however, this was likely caused by an artificial gene truncation (gene start < 90 bp from a contig end).

We additionally assessed all 75 genes that Krätzer *et al.* (2009) identified as differentially expressed in *M. mazei* Gö1 when grown on methanol versus TMA (Supplementary Table 8). Besides the genes previously mentioned, three genes were nearly fixed between clades ( $F_{ST} > 0.9$ ): a putative beta-ketoacyl synthase/thiolase and two hypothetical proteins. Again, the (nearly) fixed SNPs were mostly synonymous ( $dN/dS < 0.21$ ). A putative catalase differed in both copy number and gene length between clades because of a gene fission event in mazei-T.



**Figure 5** (a) The position of the *hdrABC-fdhAB-mvhD* gene cassette in *M. mazei* C16. (b) Gene copy number of Hdr, Fdh and Mvh paralogs is mapped onto a maximum likelihood phylogeny similar to Figure 1 (BB isolates removed). Black box outlines highlight genes that were not called by Rapid Annotation using Subsystem Technology (RAST) but were identified *via* tBLASTn.

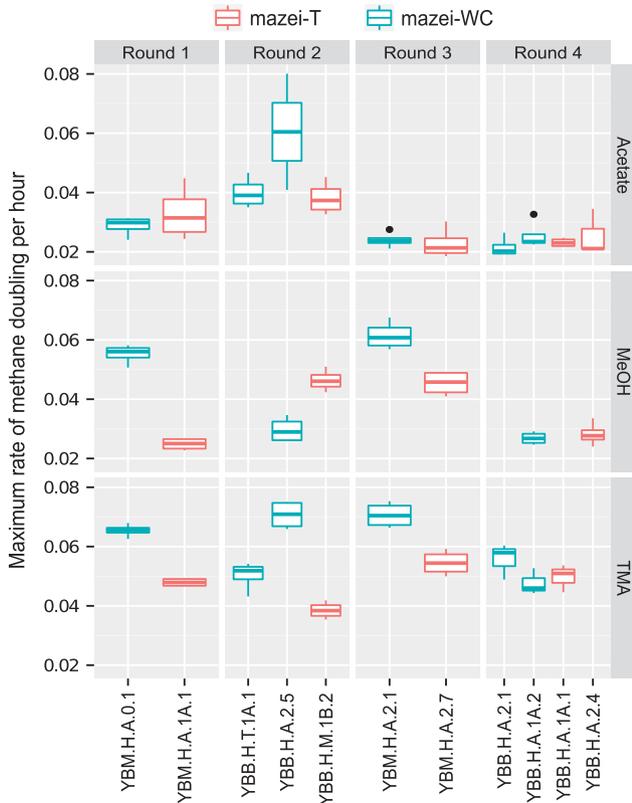
## Discussion

Through comparative population-level genomics and methane production assays, we have elucidated how genotypic and phenotypic diversity is partitioned among methanogen populations inhabiting a natural environment. Furthermore, our findings provide testable hypotheses on the genetic mechanisms that have created and maintained this diversity. Although *M. mazei* taxa comprised nearly half of the *Methanosarcinales* isolates that we obtained, the clade was rare in each sediment sample (Supplementary Figure 1).

Nevertheless, this cohort of isolates provided a powerful data set for robustly elucidating the taxonomic scale on which natural variation occurs in these methanogens and also the potential ecological and evolutionary relevance of such variation.

### *Evolution of the mazei-WC and mazei-T clades*

Phylogenetic inference based on WGA revealed that *M. mazei* isolates were partitioned into two distinct clades (Figure 1), *mazei*-WC and *mazei*-T. Although *M. mazei* isolates in both clades originated almost



**Figure 6** The box plots show the distributions of maximum rate of methane production (concentration doubling per hour) for 3–4 replicates of each culture on each of the three substrates. Maximum methane production rates were calculated from the maximum slope of logistic curves fit onto time series of methane concentrations spanning >600 h. Rounds denote separate growth experiments directly comparing isolates from each clade, while controlling for isolation source, media and substrate. Boxes represent in inter-quartile range (IQR), with the center line representing the median value. Whiskers indicate the lowest and highest values within 1.5 X the IQR from the first and third quartiles, respectively. Black dots represent outliers beyond the whiskers. Absent data points were caused by growth inhibition of >1 replicate culture because of the accidental introduction of oxygen.

equally from each Youngs Bay site, BB isolates were largely absent. This finding can be explained by the high prevalence of other diverse *Methanosarcina* clades in BB, which either outcompeted *M. mazei* or increased the sampling depth required to obtain *M. mazei* isolates from our *Methanosarcina*-targeted culturing approach.

In accordance with the core genome, variable gene content was largely partitioned between clades, suggesting clade-specific gene gain and loss (Figure 1). The very low number of variable genes specific to isolates from a particular sampling location does not support a scenario of adaptive gene acquisition. In contrast, the two clades significantly differ in the number of strains isolated on particular substrates, indicating niche partitioning in regards to substrate utilization. However, our evidence of continued gene flow between mazei-T and mazei-WC suggests that ecological barriers have not lead to complete species barriers.

Divergent selection for adaptive loci in a background of recombination can create heterogeneous genomic landscapes, where genetic differentiation accumulates in some regions, whereas the homogenizing effects of gene flow limit divergence in other regions (Nosil *et al.*, 2009). The high prevalence of gene fixation in only certain regions of the core genome supports this scenario (Figure 2). Consistent with the maintenance of diversity through ecological differentiation, we did observe a clade-specific difference in regards to methane production rates during growth on TMA (Figure 6), which shows that these clades are phenotypically distinct. Further support of this phenotypic distinction was evidenced by the statistically significant inter-clade difference of which growth substrates yielded isolates. Although growth on a particular substrate does not preclude the ability to grow on another, this significant inter-population partitioning by substrate suggests that mazei-WC and mazei-T possess distinct growth characteristics.

By investigating the suite of variable genes and core alleles that are fixed between the two clades, we have identified some possible sources of ecological differentiation between the mazei-WC and mazei-T clades.

#### *The influence of mobile genetic elements*

The majority of variable genes in this data set were strain specific, indicating their rapid movement through these genomes. Numerous genes specific to mazei-T or mazei-WC colocalize to genomic islands, which suggests that they were introduced via mobile elements (Figure 3). Many of these genes appear to have been acquired by inter-domain gene transfer from multiple bacterial phyla, in accordance with previous reports of bacterial gene acquisition in the *Methanosarcinaceae* (Deppenmeier *et al.*, 2002; Fournier and Gogarten, 2007). Gene gain appeared to occur asymmetrically between clades, with the mazei-T clade possessing a much larger number of putative genomic islands and clade-specific genes (Figure 3).

The prevalence, diversity and distribution of defense systems in these strains are consistent with mobile elements being active in these populations. Interestingly, the mazei-T and mazei-WC clades differed in genes associated with both RM systems and CRISPR systems (Table 3). The clade specificity of both innate and adaptive immune systems against foreign genetic elements suggests that the maintenance of diversity within the YBB and YBM environments is at least partially due to differences in susceptibility to viral pathogens or other genetic mobile elements. The correspondence between specific CRISPR spacers and the presence or absence of a particular CRISPR-Cas system supports this conclusion. The identified genomic islands in both clades had hallmarks of integrated viral or plasmid DNA (for example, flanking mobile element proteins, integrases and tRNAs), and thus further supports the

hypothesis that mobile elements have influenced the genomic evolution of mazei-T and mazei-WC.

#### Genomic signatures of a potentially adaptive gene cassette

We identified a gene cassette putatively encoding HdrA3B3C3 with FdhAB and MvhD among the mazei-WC isolates and certain type strains (Figure 5). The *hdrABC* paralog along with *mvhD* and *fdhAB* are highly conserved within the mazei-WC clade and appear to be polycistronic, suggesting that the gene set encodes a protein complex. *Methanococcus maripaludis* has been experimentally shown to grow on formate in a pathway coupling formate oxidation to the reduction of CO<sub>2</sub> to formyl-MFR by additionally coupling the reaction to the reduction of the CoM-CoB heterodisulfide through flavin-based electron bifurcation (Costa *et al.*, 2013). The pathway involves FdhAB for formate oxidation and a methylviologen-reducing hydrogenase acting as the adaptor subunit for electron bifurcation by HdrABC. However, no *Methanosarcina* strain has been shown to grow on formate, including the *Methanosarcina* type strains that we found to contain the *fdhAB-mvhD-hdrABC* cassette. Therefore, we hypothesize that the gene cassette encodes an enzyme that acts to produce formate, possibly for use as a precursor of purine biosynthesis. This metabolic difference between mazei-T and mazei-WC may have had a role in maintaining genomic diversity between these populations.

As additional support for this hypothesized pathway, four subunits of the molybdenum-containing formylmethanofuran dehydrogenase (Fmd) are fixed or nearly fixed between mazei-T and mazei-WC. Fmd along with Fwd are two isozymes that are known to catalyze the reduction of CO<sub>2</sub> to formyl-MFR (Schmitz *et al.*, 1992 a,b). The Fwd operon is also present among all *M. mazei* strains, but each gene in the operon has low  $F_{ST}$  values (<0.37). Polymorphisms in FmdC and FmdF, the other genes in the FmdEFACDB operon, were both also (nearly) fixed, but these genes were absent from a strain in mazei-T, and therefore had not been included in our analysis of core gene fixation. Although polymorphisms in the Fmd operon were (nearly) fixed between clades, both synonymous and nonsynonymous genes were (nearly) fixed, leading to low dN/dS values (<0.3). This signal of weak purifying selection on the Fmd operon may be due to changing selective pressures (that is, initially positive but changed to weak purifying) from when the gene cassette was initially gained in mazei-WC or lost in mazei-T.

Interestingly, multiple genes involved in molybdenum transport along with molybdopterin synthesis were also divergent and nearly fixed between clades. Two of these genes were located ~1 Mb downstream of the Fmd operon in *M. mazei* C16, suggesting that the fixation was not due to linkage. Given that Fmd requires molybdenum to function (Schmitz *et al.*, 1992 a,b; Costa *et al.*, 2010), the introduction or loss of

the mazei-WC clade-specific cassette may have altered cellular requirements for molybdopterin.

#### Genetic signatures of differing substrate utilization rates

The significant difference in TMA utilization rates between mazei-T and mazei-WC conforms to the hypothesis that the two clades have diverged because of niche partitioning (Figure 6; Supplementary Table 7). Although our search for the genetic basis of this difference did identify some putative adaptive genes, we did not find a 'smoking gun' that clearly signaled adaptive differentiation. The near fixation of a DMA-specific methyltransferase gene (*mtaB1*) may have resulted in differing TMA utilization rates by altering utilization rates of the DMA produced by TMA metabolism (Paul *et al.*, 2000). The near fixation of the methanol-specific genes *mtaB1*, *mtaC1* and *mtaA1* does not correspond to the TMA utilization data (Figure 6) and highlights the possibility that fixation of these genes may be caused by linkage to the actual adaptive loci or by neutral evolutionary processes (Barrett and Hoekstra, 2011). This alternative hypothesis is supported by the rarity of nonsynonymous SNPs among all nearly fixed genes identified in this analysis. We note that the observed phenotypic difference may be caused by variation at the regulatory level, especially given that the methylamine:CoM and methanol:CoM pathways are highly regulated (Bose *et al.*, 2006; Bose and Metcalf, 2008; Opulencia *et al.*, 2009).

## Conclusions

The *M. mazei* isolates that we obtained from the Columbia River Estuary differ by <1% at their core genomes, and yet, we have shown that they comprise two clades distinct in genotype and phenotype. Niche partitioning by primary metabolism and interactions with foreign genetic elements likely had a direct effect on the spatial and temporal distribution of these clades, with each clade being governed by different environmental pressures. The highly resolved taxonomic scale at which we have observed ecological differentiation should be taken into account in future studies of methanogen diversity. Our hypotheses on which loci were likely involved in the ecological differentiation between mazei-T and mazei-WC can be specifically tested using genetic and physiological methods already established for *Methanosarcina* (Kohler and Metcalf, 2012), whereas extending our comparative genomics methodology to other environments and methanogen clades will ascertain the generality these hypotheses. The findings of this study help to establish the taxonomic scales at which genomic and phenotypic variation dictate the spatial and temporal distribution of methanogens, which is needed in order to determine how these key factors in the carbon cycle will behave in future environmental scenarios.

## Conflict of Interest

The authors declare no conflict of interest.

## Acknowledgements

We thank Maya Errabolu for helping collect gas chromatography measurements, Gargi Kulkarni for isolating *M. mazei* WWM610 and everyone involved in the Genomes to Life: Biological Systems Research on the Role of Microbial Communities in Carbon Cycling project, especially Matt Benedict, Judy Luke and Sarah Reinhart. We also thank Mary Elizabeth Metcalf, Nicolai Müller, Petra Kohler, Madeline López-Muñoz, Thom Mand, Andrew He Fu and Jeremy Ellermeier for providing assistance with anaerobic cultivation and gas chromatography analysis. This material is based on work supported in part by the Department of Energy under grant no. DE-SC0005348.

## References

- Angiuoli SV, Salzberg SL. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **27**: 334–342.
- Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA et al. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**: 75.
- Barrett RDH, Hoekstra HE. (2011). Molecular spandrels: tests of adaptation at the genetic level. *Nat Rev Genet* **12**: 767–780.
- Benedict MN, Henriksen JR, Metcalf WW, Whitaker RJ, Price ND. (2014). ITEP: an integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics* **15**: 8.
- Bose A, Metcalf WW. (2008). Distinct regulators control the expression of methanol methyltransferase isozymes in *Methanosarcina acetivorans* C2A. *Mol Microbiol* **67**: 649–661.
- Bose A, Pritchett MA, Rother M, Metcalf WW. (2006). Differential regulation of the three methanol methyltransferase isozymes in *Methanosarcina acetivorans* C2A. *J Bacteriol* **188**: 7274–7283.
- Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML et al. (2012). Patterns of gene flow define species of thermophilic archaea. *PLoS Biol* **10**: e1001265.
- Cohan FM, Koeppel AF. (2008). The origins of ecological diversity in prokaryotes. *Curr Biol* **18**: R1024–R1034.
- Coleman ML, Chisholm SW. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc Natl Acad Sci USA* **107**: 18634–18639.
- Cornejo OE, Lefébure T, Bitar PDP, Lang P, Richards VP, Eilertson K et al. (2013). Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*. *Mol Biol Evol* **30**: 881–893.
- Costa KC, Lie TJ, Xia Q, Leigh JA. (2013). VhuD facilitates electron flow from H<sub>2</sub> or formate to heterodisulfide reductase in *Methanococcus maripaludis*. *J Bacteriol* **195**: 5160–5165.
- Costa KC, Wong PM, Wang T, Lie TJ, Dodsworth JA, Swanson I et al. (2010). Protein complexing in a methanogen suggests electron bifurcation and electron delivery from formate to heterodisulfide reductase. *Proc Natl Acad Sci USA* **107**: 11050–11055.
- Darling ACE, Mau B, Blattner FR, Perna NT. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**: 1394–1403.
- Darling AE, Mau B, Perna NT. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**: e11147.
- Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R et al. (2002). The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. *J Mol Microbiol Biotechnol* **4**: 453–461.
- Eggen RIL, Geerling ACM, PWJD Groot, Ludwig W, De Vos WM. (1992). Methanogenic bacterium Gö1: an acetoclastic methanogen that is closely related to *Methanosarcina frisia*. *Syst Appl Microbiol* **15**: 582–586.
- Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL et al. (2011). Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proc Natl Acad Sci USA* **108**: 2831–2836.
- Enright AJ, Van Dongen S, Ouzounis CA. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**: 1575–1584.
- Excoffier L, Lischer HEL. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **10**: 564–567.
- Ferry JG. (1993). *Methanogenesis: Ecology, Physiology, Biochemistry & Genetics*. Chapman & Hall: New York.
- Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**: 237–240.
- Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, Hotz H-R et al. (2008). The Pfam protein families database. *Nucleic Acids Res* **36**: D281–D288.
- Fournier GP, Gogarten JP. (2007). Evolution of acetoclastic methanogenesis in *Methanosarcina* via horizontal gene transfer from cellulolytic clostridia. *J Bacteriol* **190**: 1124–1127.
- Grissa I, Vergnaud G, Pourcel C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**: W52–W57.
- Guss AM, Kulkarni G, Metcalf WW. (2009). Differences in hydrogenase gene expression between *Methanosarcina acetivorans* and *Methanosarcina barkeri*. *J Bacteriol* **191**: 2826–2833.
- Hartl D, Clark A. (1998). *Principles of Population Genetics* 3rd edn Sinauer Associates: Sunderland, MA.
- Hedderich R, Whitman W. (2006). Physiology and biochemistry of the methane-producing Archaea. *Prokaryotes* **2**: 1050–1079.
- Ito A, Inatomi M. (2012). Use of a process-based model for assessing the methane budgets of global terrestrial ecosystems and evaluation of uncertainty. *Biogeosciences* **9**: 759–773.
- Kandler O, Hippe H. (1977). Lack of peptidoglycan in the cell walls of *Methanosarcina barkeri*. *Arch Microbiol* **113**: 57–60.
- Katoh K, Standley DM. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.
- Kendall MM, Boone DR. (2006). The order methanosarcinales. *Prokaryotes* **3**: 244–256.
- Kohler PRA, Metcalf WW. (2012). Genetic manipulation of *Methanosarcina* spp. *Front Microbiol* **3**: 259.

- Korber B. (2000). HIV signature and sequence variation analysis. *Comput Anal HIV Mol Seq* **4**: 55–72.
- Krätzer C, Carini P, Hovey R, Deppenmeier U. (2009). Transcriptional profiling of methyltransferase genes during growth of *Methanosarcina mazei* on trimethylamine. *J Bacteriol* **191**: 5108–5115.
- Kryazhimskiy S, Plotkin JB. (2008). The population genetics of dN/dS. *PLoS Genet* **4**: e1000304.
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D *et al.* (2009). Circos: an information aesthetic for comparative genomics. *Genome Res* **19**: 1639–1645.
- Langille MGI, Brinkman FSL. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* **25**: 664–665.
- Liu Y, Whitman WB. (2007). Metabolic, phylogenetic, and ecological diversity of the methanogenic archaea. *Ann N Y Acad Sci* **1125**: 171–189.
- Maeder DL, Anderson I, Brettin TS, Bruce DC, Gilna P, Han CS *et al.* (2006). The *Methanosarcina barkeri* genome: comparative analysis with *Methanosarcina acetivorans* and *Methanosarcina mazei* reveals extensive rearrangement within methanosarcinal genomes. *J Bacteriol* **188**: 7922–7931.
- Mah RA, Smith MR. (1981). The methanogenic bacteria. In: Starr MP, Stolp H, Trüper HG, Balows A, Schlegel HG (eds) *The Prokaryotes*. Springer: Berlin Heidelberg, pp 948–977.
- Mao F, Williams D, Zhaxybayeva O, Poptsova M, Lapierre P, Gogarten JP *et al.* (2012). Quartet decomposition server: a platform for analyzing phylogenetic trees. *BMC Bioinform* **13**: 123.
- Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M *et al.* (2005). CDD: a conserved domain database for protein classification. *Nucleic Acids Res* **33**: D192–D196.
- Metcalf WW, Zhang JK, Shi X, Wolfe RS. (1996). Molecular, genetic, and biochemical characterization of the *serC* gene of *Methanosarcina barkeri* Fusaro. *J Bacteriol* **178**: 5797–5802.
- Nguyen TH, Ranwez V, Pointet S, Chifolleau A-MA, Doyon J-P, Berry V. (2013). Reconciliation and local gene tree rearrangement can be of mutual profit. *Algorithms Mol Biol* **8**: 12.
- Nickel L, Weidenbach K, Jäger D, Backofen R, Lange SJ, Heidrich N *et al.* (2013). Two CRISPR-Cas systems in *Methanosarcina mazei* strain Gö1 display common processing features despite belonging to different types I and III. *RNA Biol* **10**: 779–791.
- Nosil P, Funk DJ, Ortiz-Barrientos D. (2009). Divergent selection and heterogeneous genomic divergence. *Mol Ecol* **18**: 375–402.
- Ochman H, Lawrence JG, Groisman EA. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299–304.
- Opulencia RB, Bose A, Metcalf WW. (2009). Physiology and posttranscriptional regulation of methanol:coenzyme M methyltransferase isozymes in *Methanosarcina acetivorans* C2A. *J Bacteriol* **191**: 6928–6935.
- Paul L, Ferguson DJ, Krzycki JA. (2000). The trimethylamine methyltransferase gene and multiple dimethylamine methyltransferase genes of *Methanosarcina barkeri* contain in-frame and read-through amber codons. *J Bacteriol* **182**: 2520–2529.
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna: Austria, <http://www.R-project.org>.
- Riley WJ, Subin ZM, Lawrence DM, Swenson SC, Torn MS, Meng L *et al.* (2011). Barriers to predicting changes in global terrestrial methane fluxes: analyses using CLM 4 Me, a methane biogeochemistry model integrated in CESM. *Biogeosci Discuss* **8**: 1733–1807.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.
- Schmitz RA, Albracht SPJ, Thauer RK. (1992a). A molybdenum and a tungsten isoenzyme of formylmethanofuran dehydrogenase in the thermophilic archaeon *Methanobacterium wolfei*. *Eur J Biochem* **209**: 1013–1018.
- Schmitz RA, Albracht SPJ, Thauer RK. (1992b). Properties of the tungsten-substituted molybdenum formylmethanofuran dehydrogenase from *Methanobacterium wolfei*. *FEBS Lett* **309**: 78–81.
- Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G *et al.* (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**: 48–51.
- Shapiro BJ, Polz MF. (2014). Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol* **22**: 235–247.
- Shimizu S, Upadhye R, Ishijima Y, Naganuma T. (2011). *Methanosarcina horonobensis* sp. nov., a methanogenic archaeon isolated from a deep subsurface Miocene formation. *Int J Syst Evol Microbiol* **61**: 2503–2507.
- Söding J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**: 951–960.
- Sowers KR, Baron SF, Ferry JG. (1984). *Methanosarcina acetivorans* sp. nov., an acetotrophic methane-producing bacterium isolated from marine sediments. *Appl Environ Microbiol* **47**: 971–978.
- Stamatakis A. (2006). RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinforma Oxf Engl* **22**: 2688–2690.
- Suyama M, Torrents D, Bork P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**: W609–W612.
- Tatusov RL, Galperin MY, Natale DA, Koonin EV. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**: 33–36.
- Thauer RK. (1998). Biochemistry of methanogenesis: a tribute to Marjory Stephenson:1998 Marjory Stephenson Prize Lecture. *Microbiology* **144**: 2377–2406.
- Vestergaard G, Garrett RA, Shah SA. (2014). CRISPR adaptive immune systems of archaea. *RNA Biol* **11**: 156–167.
- Wania R, Ross I, Prentice IC. (2010). Implementation and evaluation of a new methane model within a dynamic global vegetation model: LPJ-WHyMe v1.3.1. *Geosci Model Dev* **3**: 565–584.
- Wickham H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer: New York.

Supplementary Information accompanies this paper on The ISME Journal website (<http://www.nature.com/ismej>)