The genome sequence of the psychrophilic archaeon, *Methanococcoides burtonii*: the role of genome evolution in cold-adaptation

Michelle A Allen[1]*, Federico M Lauro[1]*, Timothy J Williams[1], Dominic Burg[1], Khawar S Siddiqui[1], Davide De Francisci[1], Kevin WY Chong[1], Oliver Pilak[1], Hwee H Chew[1], Matthew Z. De Maere[1], Lily Ting[1], Marilyn Katrib[1], Charmaine Ng[1], Kevin R Sowers[3], Michael Y Galperin[4], Iain J Anderson[5], Natalia Ivanova[5], Eileen Dalin[5], Michele Martinez[5], Alla Lapidus[5], Loren Hauser[6], Miriam Land[6], Torsten Thomas[1,2] , and Ricardo Cavicchioli[1§]

[1] School of Biotechnology and Biomolecular Sciences, The University of New South Wales, Sydney, New South Wales 2052, Australia

[2] Centre for Marine Bio-Innovation, The University of New South Wales, Sydney, New South Wales 2052, Australia

[3] Center of Marine Biotechnology, 701 E. Pratt Street, Baltimore, Maryland 21202, USA

[4] NCBI, NLM, National Institutes of Health 8600 Rockville Pike, MSC 3830 Bethesda, Maryland 20894, USA

[5] Department of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA

[6] Oak Ridge National Laboratory, One Bethel Valley Road, Oak Ridge, TN 37831

[§]Corresponding author. Rick Cavicchioli, School of Biotechnology and Biomolecular Sciences, The University of New South Wales, Sydney, New South Wales 2052, Australia
Email: R.Cavicchioli@unsw.edu.au
Phone: (61 2) 9385 3516
Fax: (61 2) 9385 2742

[*]MAA and FML have contributed equally.

Running title: Cold-adapted genome of *Methanococcoides burtonii*

Abstract

Psychrophilic archaea are abundant and perform critical roles throughout the Earth's expansive cold biosphere. Here we report the first complete genome sequence for a psychrophilic methanogenic archaeon, *Methanococcoides burtonii*. The genome sequence was manually annotated including the use of a five tiered Evidence Rating system that ranked annotations from Evidence Rating (ER) 1 (gene product experimentally characterized from the parent organism) to ER5 (hypothetical gene product) to provide a rapid means of assessing the certainty of gene function predictions. The genome is characterized by a higher level of aberrant sequence composition (51%) than any other archaeon. In comparison to hyper/thermophilic archaea which are subject to selection of synonymous codon usage, *M. burtonii* has evolved cold adaptation through a genomic capacity to accommodate highly skewed amino acid content, while retaining codon usage in common with its mesophilic *Methanosarcina* cousins. Polysaccharide biosynthesis genes comprise at least 3.3% of protein coding genes in the genome, and Cell wall/membrane/envelope biogenesis COG genes are over-represented. Likewise, signal transduction (COG category T) genes are over-represented and *M. burtonii* has a high "IQ" (a measure of adaptive potential) compared to many methanogens. Numerous genes in these two over-represented COG categories appear to have been acquired from ε- and δ-proteobacteria, as do specific genes involved in central metabolism such as a novel B form of aconitase. Transposases also distinguish *M. burtonii* from other archaea, and their genomic characteristics indicate they play an important role in evolving the *M. burtonii* genome. Our study reveals a capacity for this model psychrophile to evolve through genome plasticity (including nucleotide skew, horizontal gene transfer and transposase activity) that enables adaptation to the cold, and to the biological and physical changes that have occurred over the last several thousand years as it adapted from a marine, to an Antarctic lake environment.

Keywords

archaea / cold adaptation / genome plasticity / *Methanococcoides burtonii* / psychrophile

Subject Category

Evolutionary Genetics

Introduction

Most (~75%) of the Earth's biosphere is cold (≤5°C), consisting of polar and alpine regions, the deep ocean, terrestrial and ocean sub-surface, caves, the upper atmosphere, seasonally and artificially cold environments. Diverse cold-adapted (psychrophilic) microorganisms have evolved a capacity to proliferate in all these habitats. However, despite the enormous contribution cold environments make to the Earth's biosphere, relatively little is known about the genetic make-up and molecular mechanisms of adaptation of the resident microorganisms and how they drive the critical biogeochemical processes that help to maintain the planet in a habitable state (Cavicchioli 2006; Murray & Grzymski 2007).

The majority of microbiological studies of cold environments have focused on psychrophilic bacteria, and genomic analyses have provided valuable insight into unique characteristics of sea-ice, sediment and planktonic species (Rabus et al. 2004; Methe et al. 2004; Medigue et al. 2005; Riley et al. 2008). In addition to bacteria, it is now clearly recognised that the archaea are numerically, phylogenetically and functionally important members of cold aquatic and terrestrial ecosystems (Cavicchioli 2006; Murray & Grzymski 2007). Their role in global biogeochemical cycles is diverse, including critical roles in the nitrification of soil and ocean waters mediated by ammonia oxidation (Leininger et al. 2006; Konneke et al. 2005; Cavicchioli et al. 2007), and the cycling of simple carbon compounds via methanogenesis and reverse methanogenesis (anaerobic oxidation of methane) (Cavicchioli 2006; Kruger et al. 2005).

Cold anaerobic environments (*e.g.* ocean and lake sediments, permafrost) harbor methanogens that have the capacity to contribute significantly to global carbon emissions through the production of methane as a green house gas. While presently marine environments only contribute approximately 2% of the world's methane flux, this appears to be kept in check by the microbial communities that anaerobically oxidize methane (Moran et al. 2008). The methane cycle is still poorly understood, and gaining an understanding of how these cellular processes occur in the cold will require fundamental knowledge of the molecular mechanisms of cold adaptation of psychrophilic methane-producing and methane-oxidizing archaea (Cavicchioli 2006). Psychrophilic archaea have generally proven to be difficult to isolate. However, methanogens have been isolated from Antarctic lakes, marine sediment in Alaska and the Baltic Sea, a freshwater lake in Switzerland, and Arctic permafrost (Franzmann et al. 1992; Franzmann et al. 1997; Simankova et al. 2001; Chong et al. 2002; von Klein et al. 2002; Singh et al. 2005; Morozova & Wagner 2007; Kendall et al. 2007; Kotsyurbenko et al. 2007).

Methanococcoides burtonii was the first formally characterized archaeal psychrophile, and was isolated from cold (1-2ºC) methane saturated, anaerobic bottom waters (25 m depth) of Ace

Lake, Antarctica (Franzmann et al. 1992). *M. burtonii* is a methylotrophic methanogen utilizing methylamines and methanol, but not H_2:CO_2 or acetate for growth. It is a eurypsychrophile with a relatively broad growth temperature range (-2.5º to 29ºC), in contrast to *Methanogenium frigidum* which is a stenopsychrophile (0º to 18ºC) also isolated from Ace Lake (Franzmann et al. 1992; Franzmann et al. 1997; Reid et al. 2006). Unlike *M. frigidum* which has proven to be very difficult to grow, *M. burtonii* is amenable to laboratory cultivation and has become a model psychrophilic archaeon for dissecting molecular mechanisms of cold adaptation (Cavicchioli 2006). Studies on *M. burtonii* have examined enzyme structure/function, gene regulation, tRNA modification, membrane lipid composition and proteomics (reviewed in Cavicchioli 2006), and intracellular solutes (Costa et al. 2006; Thomas et al. 2001). The draft genome sequences of *M. burtonii* and *M. frigidum* have also enabled comparative genomic analyses assessing the compositional and structural basis of thermal adaptation of proteins for archaea spanning growth temperatures ranging from 0 to 110ºC (Saunders et al. 2003).

The genome sequence of *M. burtonii* was recently completed paving the way for determining for the first time, the genomic basis for growth and survival of a psychrophilic archaeon. Due to the value of the genome sequence and importance of *M. burtonii* to the scientific community, the genome sequence was exhaustively manually annotated to generatate a high quality genome sequence that was then used to evaluate the evolution and biology of *M. burtonii*.

Materials and Methods

Genome sequencing, assembly, automated annotation

DNA was isolated from *M. burtonii* DSM 6242 grown at 23°C [42]. The genome of *M. burtonii* was sequenced at the Joint Genome Institute (JGI) using a combination of 3 kb, 8 kb and 40 kb DNA libraries. All general aspects of library construction and sequencing can be found at the JGI's website (http://www.jgi.doe.gov/). The Phred/Phrap/Consed software package (www.phrap.com) was used to assemble all three libraries and to assess quality [43-45]. Possible mis-assemblies were corrected, and gaps between contigs were closed by editing in Consed, custom primer walks, or PCR amplification. The error rate of the completed genome sequence of *M.burtonii* is less than 1 in 50,000. Putative coding regions were identified using Critica [46], Generation and Glimmer [47], with automated annotation of proteins according to search results from the databases TIGRFam, PRIAM, Pfam, Smart, COGs, Swiss-Prot/TrEMBL and KEGG, as described for other JGI genomes [eg. 48]. Additional curation involved the calculation of homolog, paralog and ortholog gene relationships, and generation of genome statistics. A round of manual curation was performed on the predicted genes of M. burtonii by IMG staff in which 346 changes were made. This corresponds to 13.9% of

predicted genes. The start sites of 136 genes were modified, 62 genes were deleted, and 61 new genes were added. Also 87 genes were identified as pseudogenes. This includes new pseudogenes that were added as well as previously existing genes that were converted to pseudogenes. IMG's curation staff also perform horizontal annotation involving manual investigation of proteins and biochemical pathways, with "IMG term" annotations propagated across genomes according to strict homology criteria (http://imgweb.jgi-psf.org/w/doc/img_er_ann.pdf). In the *M. burtonii* genome 635 proteins were assigned IMG terms by this process.

Manual annotation, Evidence Rating system and comparison genomes

The protein sequence for each gene was searched against the Swiss-Prot and Protein Data Bank (PDB) databases using BLAST [24] to identify the most closely related, experimentally characterized homolog available in the literature. The curated part of Swiss-Prot database was selected because of the extent and quality of annotations associated with each protein sequence [25], and the PDB provides an archive of experimentally-determined three-dimensional protein structures. BLAST matches were examined sequentially, searching for a match with direct experimental verification of the function. Matching proteins were not considered if they had no function listed, a function determined 'by similarity', or no reference cited. Experimental evidence that was considered acceptable to define function included papers detailing the expression and characterization of the protein, protein crystallography studies, and mutation and complementation studies that defined function. Papers that only documented the nucleotide sequence, including genome sequences, were not considered to provide sufficient evidence of function. Nor were unpublished protein crystal structures or the (published or unpublished) crystal structures of hypothetical proteins for which function was not known. In most cases papers were read in sufficient depth to establish if conclusions about function were justified. Choosing published experimental evidence as a prerequisite for defining function probably excluded some valid unpublished data, however it was not possible to assess this data and it was excluded. Careful attention was also given to the recent literature in order to identify valid experimental data (*e.g.* several methanogenesis-related proteins: Mbur_0808, Mbur_0811) that had not yet been updated in protein databases.

After the best experimentally characterized homolog had been determined, InterPro and Pfam domains were checked for their presence in the *M. burtonii* homolog to confirm that all identifiable functional domains were conserved. The graphical representation on the Swiss-Prot BLAST output was used for assessing the arrangement and extent of domain matches throughout the length of the query and matching proteins, thereby providing a rapid way to assess both the global (whole protein) and local (domain and motif) similarity between the *M. burtonii* and the functionally characterized proteins.

An Evidence Rating (ER) system was developed to enable the confidence of functional assignments to be clearly displayed for each gene (Supplementary Information: manual annotation). ER1 indicates that the protein from *M. burtonii* had been experimentally characterized (a self match); ER2, the most closely related functionally-characterized homolog is not from *M. burtonii* but the BLAST alignments share ≥ 35% sequence identity along the entire length of the protein; ER3, the most closely related functionally-characterized homolog shares <35% sequence identity along the length of the protein, but all required motifs/domains for function are present and complete; ER4, an experimentally characterized full-length homolog is not available but conserved protein motifs or domains can be identified; ER5 (hypothetical protein), no functionally characterized homolog can be found, and no characterized protein domains above the Pfam and InterProScan cut-off thresholds can be identified. Comparative genomics (see methods below) was performed using local or external databases (*e.g.* NCBI, IMG) with between 37 and 48 closed archaeal genome sequences. The draft sequence of *M. frigidum* was included in some comparisons (specified in the text).

COG scrambler and ALL scrambler

Two in-house programs, COG_scrambler.pl and ALL_scrambler.pl, were created for identifying significant differences in gene composition between samples. The COG_scrambler program compares all the COGs present in two datasets and calculates which COG categories are statistically over- or under-represented in a dataset by difference of medians analysis. The ALL_scrambler program performs an analogous process but reports which individual COGs are statistically over- or under-represented. All programs were run with a confidence level setting of 0.99 (99% confidence that the difference is not due to chance) and 10,000 resampling replicates. The sample size was 6,000 for the comparison of *M. burtonii* vs methanosarcinal genomes, and 10,000 for all other comparisons. The approach to determine the statistical significance was developed by FML and is similar to that described by Rodriquez-Brito et al (Rodriguez-Brito et al. 2006).

Phylogenetic analyses

To visualize the phylogenetic relationships between proteins in the genome, an in-house perl script "phylo_profiler.pl" was created. This program takes an input file (all *M. burtonii* proteins) and uses BLAST to create a matrix of BLAST scores against selected completed genome sequences. The best BLAST bitscore for each *M. burtonii* protein was then normalised both within and across genomes as previously described (Enault et al. 2003). Pertinent data such as the COG ID, COG category, annotation and ER value were incorporated in the matrix to facilitate various searches and analyses. The matrix produced by phylo_profiler.pl was viewed using BioLayout Express 3D [Freeman et al. 2007) with typical settings as follows: Graph Size vs Corr. Threshold = 95 (Nodes = 630, Edges = 1981,

R^2=0.7987), Start temperature = 250, No. of iterations = 60, K-value modifier = 1.0, Iterations for Burst = 10, Minimum component size = 3. The protein sequences of transposases occurring multiple times (and those identified by proteomics to be expressed in the cell) were aligned using clustalw (Thompson et al. 1994) and used to build phylogenetic trees by neighbour-joining (Saitou et al. 1987) using the Poisson-corrected distance implemented in MEGA4 (Tamura et al. 2007). The stability of the relationships was assessed by 1,000 bootstrap replicates.

Identification of predicted HGT events

The program Alien Hunter (Vernikos & Parkhill 2006) was used to identify genomic islands with altered nucleotide signatures. These islands were further divided in groups depending on their IVOM (Interpolated variable order motif) scores and the ORFs contained within these regions were parsed using custom PERL scripts. The ORFs overlapping an island border were eliminated from further analyses. Each ORF was then phylogenetically assigned according to the best 2 BLAST hits with a minimum bitscore of 50 against a customized version of the NCBI non-redundant (nr) database from which all *M. burtonii* proteins had been eliminated. The phylogenetic distribution was computed using MEGAN (Huson et al. 2007). Codon usage was analyzed with CodonW (Peden 2000). The genes on the primary axis of inertia in the correspondence analysis with values above the mean-value plus one standard deviation were considered efficiently expressed; those below the mean minus one standard deviation were considered not-efficiently expressed. Amino acid percentage composition and principal component analysis were performed using customized PERL scripts and R (Ihaka & Gentleman 1996) as described by Saunders et al. (2003). An artificial genome was generated by removing the coding sequences of efficiently expressed ORFs from the FASTA sequence of the genome using the in-house PERL script “reduce_genome.pl”.

Identification of hypothetical proteins unique to *M. burtonii*

The IMG Phylogenetic Profiler tool was used to search for genes in the *M. burtonii* genome which did not contain a homolog in any of the 814 standard reference genomes (40 completed archaeal genomes, 5 draft archaeal genomes, 19 completed eucaryal genomes, 21 draft eucaryal genomes, 486 completed bacterial genomes and 243 draft bacterial genomes). The cut-off settings for a homolog was minimum amino acid identity of 30%, and maximum e-value of 1×10^{-5} using the present/absent algorithim. The list of proteins obtained was then curated further by removing any proteins that were not annotated as a hypothetical protein (ER5). BLAST searches against the NCBI nr database and the CAMERA Global Ocean Survey combined assembly ORF peptides database were performed (2nd May 2008), and any proteins with matches of e-value smaller than 1×10^{-5} were also discarded in order to obtain the final list of 117 proteins that are unique to *M. burtonii*. Conserved

gene context analysis (Saunders et al. 2005) was also used to infer possible functions for 28 proteins that were known from proteomics analysis to be expressed in the cell (Goodchild et al. 2004a; Goodchild et al. 2004b; Goodchild et al. 2005; Saunders et al. 2005; T. Williams & D. Burg, unpublished results). Some of the unique expressed hypothetical proteins were found to be located in pairs (*e.g.* Mbur_2063 and Mbur_2064) or larger clusters with unique hypothetical proteins that (to date) have not been detected in expression studies (*e.g.* Mbur_0640 to Mbur_0643 and Mbur_0275 to Mbur_0278).

Results and Discussion

M. burtonii genome overview

The completed genome sequence of *M. burtonii* DSM 6242 contains 2575032 bp in a single circular chromosome (Figure 1). Manual annotation (see Materials and Methods) resulted in changes to the functional annotation 1079 of the 2494 genes that had been identified by the auto-annotation process, and functions were assigned to 364 genes that previously had no functional prediction. A complete description of the manual annotation approach is provided in Supplementary Information: manual annotation.

To identify characteristics of the *M. burtonii* genome that distinguishes it from other archaea, the three completed Methanosarcinales genomes, all 14 methanogen genomes, or 39 archaeal genomes were compared to the *M. burtonii* genome using COG_scrambler and ALL_scrambler (Table 1). In all three sets of comparisons, the Signal transduction [T] and Replication, recombination and repair [L] COG categories were significantly over-represented, and the General Function prediction only [R] COG category was under-represented in the *M. burtonii* genome. Individual COGs that were over-represented included signal transduction histidine kinases, RecA-superfamily ATPases and CheY-like response regulators [T], and numerous transposases [L]. In addition, Cell wall, membrane, envelope biogenesis [M] was over-represented in all but the Methanosarcinales genomes.

Gene categories associated with cold adaptation

To identify functional gene categories characteristic of cold adaptation in archaea, genomes from archaeal thermophiles and hyperthermophiles (4 methanogens, or 25 archaea in total) were compared to *M. burtonii* using COG_scrambler and ALL_scrambler (Table 1). The psychrophilic genome of *M. burtonii* was over-represented with Defense mechanisms [V] and Motility [N] and under-represented with Translation [J] and Nucleotide metabolism [F] COG categories compared to both the methanogen and total archaeal genome sets.

In the Defense mechanisms category, the individual COGs that were over-represented in *M. burtonii* were mainly from Type-I restriction modification (RM) systems (COG0286, COG0610, COG0732 and COG4096), and ABC transporters (see below). The RM COGs were present at four locations on the *M. burtonii* genome (Figure S1). One set of restriction endonuclease, methyltransferase and specificity genes was arranged in a typical operon-like RM cluster (Mbur_1841 to Mbur_1843). Another cluster was spaced over 10 genes (Mbur_1213 to Mbur_1222) with numerous genes for hypothetical proteins (ER5) interspersed. Two clusters (Mbur_0506 to Mbur_0509 and Mbur_0480 to Mbur_0484) contain a divergent AAA domain protein (ER4), in addition to the specificity COG0732, methyltransferase COG0286, and a variant helicase as the restriction endonuclease (Mbur_0509 and Mbur_0484). Divergent AAA domain proteins have been implicated in a wide variety of roles, including functioning as transcriptional regulators. The relatively large number of diverse RM systems may be indicative of the exposure of *M. burtonii* to high levels of foreign DNA (see **Cold adaptation is associated with specific signatures of genome evolution**).

The other over-represented COGs from the Defense mechanisms category were COG0577 and COG1136. Proteins belonging to these two COGs were located in six short clusters in the genome, in association with proteins from COG1361 and COG4591 (Figure 2a). COGs 0577 and 4591 both contain proteins that annotated as DUF214-containing protein (ER4). These proteins contain four transmembrane domains each, and one member of the DUF214 family has been characterized as an ABC-transporter permease involved in lipoprotein export (Narita et al. 2003). Interestingly, although COG1361 is designated as S-layer domain proteins, the hypothetical proteins found in COG 1361 form a cluster distinct from the true S-layer proteins, and have similarity to the ABC-transport substrate-binding proteins observed elsewhere in the *M. burtonii* genome (Figure 2b). These proteins each possess a signal peptide motif and one transmembrane domain, similar to ABC-transporter substrate-binding proteins. These clusters of genes may therefore represent novel ABC-transporters.

The over-representation of the putative novel ABC-transporters contrasts with the lack of identifiable ABC transporters for peptides in *M. burtonii*. This is one of the major differences between the genomes of *M. burtonii* and the *Methanosarcina* spp., with the latter containing between 6 and 19 substrate binding proteins for peptide ABC transporters. The lack of peptide transporters is consistent with the inability of *M. burtonii* to use peptides for growth. *M. burtonii* also lacks drug exporters from subfamilies 2 and 3 of the major facilitator superfamily (TC 2.A.1) and proteins from the ACR3 family (2.A.59) and the ArsB family (TC 2.A.45) which are found in *Methanosarcina* spp. The susceptibility that *M. burtonii* may face towards antimicrobial compounds and toxic metalloids may be alleviated by the novel ABC transporters, some of which may also play a role in active transport leading to the formation of extracellular polymeric substances (EPS) that is characteristic of growth of *M. burtonii* in the cold (Reid et al. 2006; see **A large genomic**

committment to polysaccharide biosynthesis). It is also noteworthy that, unlike the *Methanosarcina* species, *M. burtonii* has a coenzyme F_{420}-dependent sulfite reductase (Mbur_0619) similar to the one characterized in *Methanocaldococcus jannaschii* (Johnson et al. 2005) that allows it to tolerate sulfite in the environment. In the waters where *M. burtonii* was isolated, hydrogen sulfide concentrations reach very high levels (8 mM) (Rankin et al. 1999), indicative of high sulfite reductase activity (also see **Linking the evolution of the *M. burtonii* genome to its ecological niche**).

Cold adaptation is associated with specific signatures of genome evolution

All members of the *Methanosarcina* genus (for which *M. burtonii* is closely related) have been shown to possess extremely dynamic genomes (Maeder et al. 2006) where genome rearrangements, acquisition of novel metabolic capabilities (Fournier & Gogarten 2008) and capacity to express altered or foreign genes (Li et al. 2007) contributes to their capacity to colonize a wide variety of ecological niches. To examine the importance of horizontal gene transfer (HGT) to the *M. burtonii* genome, gene-independent analysis of potential HGT was carried out using IVOM scores calculated from Alien Hunter (Vernikos & Parkhill 2006). An extremely high proportion (51%) of the *M. burtonii* genome (higher than for any other completed archaeal genome) was identified as having aberrant sequence composition compared to the average of the genome (Figure 3a). However, across the 48 archaeal genomes, the number of basepairs implicated in HGT did not correlate with growth temperature (Figure 3a). For example, the mesophile *Methanococcus maripaludis* C5 had very little predicted HGT, whereas the hyperthermophile *Methanopyrus kandleri* had almost 30% predicted HGT.

Parametric methods used for the detection of HGT events (such as Alien Hunter) are prone to false positive predictions (Azad & Lawrence 2007). To assess the reason for the high level of atypical nucleotide composition and better assess HGT, we performed a range of further analyses. Putative HGT islands were grouped together based on their IVOM scores and their phylogeny determined. For *M. burtonii* 1042 ORFs were found in 125 genomic islands with altered IVOM scores, 1097 were found outside of these regions and 164 were on the border and discarded from further analysis. The islands were 1313583 nucleotides in length and represented 51.01% of the genome, with an average island length of 10509 nucleotides. For the mesophilic *Methanosarcina* genomes, the islands represented 40.89, 40.79 and 35.36%, for *M. mazei*, *M. acetivorans* and *M. barkeri*, respectively (Table S1). Genes with the highest IVOM scores included a high proportion of unclassified genes and genes from unknown organisms (Figure 3b), and may have arisen through HGT events. The weakly atypical islands contained a high number of genes involved in translation, ribosomal structure and biogenesis (COG category J) that were phylogenetically congruent with the *Methanosarcina* genomes.

There are several lines of evidence that suggest that the ORFs within the weakly atypical islands have not arisen by HGT: 1) Genes encoding ribosomal proteins have high barriers to HGT (Sorek et al. 2007); 2) The codon usage of the ORFs within the islands is very similar to that of the rest of the genome (Figure 3c); 3) As evidenced by the proportion of genes most closely related to the Methanosarcinales genomes (Figure 3b), the phylogenetic distribution of the ORFs is indicative of vertical rather than horizontal transmission. In addition, the presence of a high number of ribosomal proteins, which often display atypical nucleotide composition (Tsirigos & Rigoutsos 2005), suggests alternative reasons for the altered IVOM scores. Initially we hypothesized that the high incidence of predicted HGT genomic islands in *M. burtonii* was caused by variations in codon usage. However, codon usage of the ORFs within the islands was similar to that of the whole genome (Figure 3c) and to the Methanosarcinales genomes (Figure 3d). Moreover, while the usage of specific synonymous codons has been correlated with the ability to grow at high temperatures (Lynn et al. 2002), the codon usage of the psychrophilic archaea, *M. burtonii* and *M. frigidum*, was indistinguishable from that of the mesophilic *Methanosarcina* genomes (Figure 3d).

We subsequently reasoned that the alteration in IVOM scores might be caused by a high incidence of efficiently expressed genes whose protein products have been strongly selected for specific amino acid usage. Principal component analysis of amino acid composition of *M. burtonii* and the three *Methanosarcina* genomes shows a strong correlation (Spearman correlation $P<0.01$) between the PC1 loadings of efficiently expressed genes and the PC2 loadings previously observed (Saunders et al. 2003) for whole genomes that was associated with temperature adaptation (Figure S3). This correlation was not observed when the same analysis was performed with genes that are not efficiently expressed. We therefore conclude that: 1) Efficiently expressed genes in *M. burtonii* tend to have a stronger “psychrophilic” component than those that are not efficiently expressed; 2) *M. burtonii* has unique amino acid usage due to its psychrophilic lifestyle (*e.g.* in comparison to the mesophilic *Methanosarcina* spp).

As a further test, we generated an artificial *M. burtonii* genome where the coding regions of efficiently expressed ORFs were removed, reducing the genome size by 322822 nucleotides (12.54%) and the number of genomic islands from 125 to 105. Of these only 63 were above the original IVOM score and encompass 782651 nucleotides (34.75%) of the artificial genome. That is, the removal of the islands with efficiently expressed ORFs reduces the incidence of regions with atypical IVOM scores by twice as much as the number of nucleotides removed.

Our data highlights the need to carefully assess inferences about HGT when parametric methods are used. For *M. burtonii*, the exceptional amount of genome plasticity (*i.e.* 51.01% altered IVOM scores) is primarily due to a strong selection for specific amino acids associated with cold adaptation present in efficiently expressed genes. This same coding bias is not seen in other genes.

This may indicate that it is a recent evolutionary event and selection is yet to be manifested in other genes, or, that low temperature selection only has a significant effect on genes that need to be efficiently expressed in the cell (*e.g.* high abundance ribosomal proteins and other proteins critical to growth in the environment) (also see **Linking the evolution of the *M. burtonii* genome to its ecological niche**). In comparison to hyper/thermophilic archaea which are subject to selection of synonymous codon usage (Lynn et al. 2002), *M. burtonii* has evolved cold adaptation through a genomic capacity to accommodate highly skewed amino acid content; an ability it has obtained while retaining codon usage in common with related mesophilic *Methanosarcina* spp.

Another feature of the codon usage plot (Figure 3d) is that the pattern of codon usage for archaea with growth temperatures ≤60ºC is similar, and distinct from archaea with higher growth temperatures. A similar distinction at 60ºC was previously noted for the relationship between G+C content of tRNA and growth temperature of archaea (Saunders et al. 2003). In the case of tRNA in *M. burtonii*, high levels of dihydrouridine incorporation provides a mechanism for enhancing flexibility beyond that achievable through G+C content (Noon et al. 2003; Saunders et al. 2003). By analogy for *M. burtonii* proteins, amino acid usage provides a mechanism for enhancing protein psychrophilicity beyond any requirements for specific codon usage.

To further assess possible HGT events, we analyzed the 500 ORFs that could not be reliably assigned to the domain *Archaea* (combined "other" and "NA" categories from Figure 3b). Compared to the whole genome these ORFs were over-represented in COG categories M (Cell wall, membrane, envelope biogenesis) and T (Signal transduction mechanisms) and under-represented in S (Function unknown), F (Nucleotide transport and metabolism) and J (Translation ribosomal structure and biogenesis) (Table 1). We note that the majority (44/70; 63%) of these ORFs were contained in islands with non-significant IVOM scores, illustrating that IVOM scores *per se* do not provide an unambiguous means of identifying HGT events in *M. burtonii*.

The under-representation of F and J is consistent with their central role in cell growth and survival and the inherent barriers for HGT of these genes. However, it was striking that category M was associated with HGT as this COG category is statistically over-represented in the *M. burtonii* genome (Table 1) and the cell wall and lipid membrane have specifically been linked to cold adaptation (see **A large genomic committment to polysaccharide biosynthesis**, and **Lipid biosynthesis**). The 27 category M ORFs are mainly annotated as different types of glycosyltransferases with varying evidence ratings (ER2-ER4), and a number of them appear to have been inherited from the ε-Proteobacteria. In category T, 39 ORFs had atypical phylogenetic assignments. These ORFs mainly belong to COG0642 (Signal transduction histidine kinase) and have ER3 or ER4 confidence ratings. Three that could be reliably assigned to a phylogenetic group (Mbur_1264, Mbur_2201, Mbur_2108) clustered within the δ-Proteobacteria. This suggests that

similar to category M, the over-representation of this Signal transduction category in *M. burtonii* (see ***M. burtonii* genome overview**) is likely to have involved HGT (also see **Signal transduction and adaptive potential**). In addition to these genes in categories M and T, HGT appears to have impacted on specific genes in central metabolism. Phylogenetic analysis of fructose-bisphosphate aldolase (Mbur_1969), serine O-acetyltransferase (Mbur_0414) and aconitase (Mbur_0316) shows that they are most closely related to counterparts from δ-Proteobacteria, and interestingly, *M. burtonii* is the first sequenced archaeon to have the B form of aconitase. Aspartate semialdehyde dehydrogenase (Mbur_0379) does not cluster with archaeal/eucaryal sequences, but does branch within bacterial enzymes.

The presence of multiple genes in *M. burtonii* originating from ε- and δ-Proteobacteria suggests a close, possibly syntrophic relationship, has existed between *M. burtonii* and members of these proteobacterial groups. In this respect, it is noteworthy that several abundant phylotypes similar to known sulfate reducing bacteria of the δ-Proteobacteria, and less abundant signatures of ε-proteobacteria have been detected in sediment and anaerobic water samples of Ace Lake (Bowman et al. 2000) (see **Linking the evolution of the *M. burtonii* genome to its ecological niche**). Our analysis shows that HGT with ε- and δ-Proteobacteria has played a specific role for acquiring genomic capabilities important for environmental adaptation.

Transposons are involved in genome evolution

Phylogenetic profiling was used to look for patterns of gene evolution. Each gene in the *M. burtonii* genome was compared using BLAST against a set of completed genomes to create a profile of the gene's phylogenetic relationships. Two comparisons were made; all completed archaeal genomes, and 431 completed bacterial and archaeal genomes. Genes were clustered according to their phylogenetic occurrence and visualized using BioLayout (Figure S4).

Five of the most tightly-clustered phylogenetic groups were comprised of transposases, with each group having specific phylogenetic occurrence (Table 2; Figure 4). The DDE superfamily of transposases (group 5) appeared to be restricted to members of the Methanosarcinales. Two of the 117 unique hypothetical proteins (Mbur_0556 and Mbur_0576; full list in Table S2) are located in a region adjacent to three cassettes each encoding genes for a DNA-directed DNA polymerase B domain protein and at least two conserved hypothetical proteins. The duplication evident in this region of the genome may also have arisen through transposition, as three transposons are also located nearby. The three cassettes are shared only between *M. burtonii* and *M. mazei*. Several other hypothetical genes appear to have been duplicated, possibly by transposition. Mbur_0169 and

Mbur_0074 share 93% sequence identity and both are located adjacent to a Radical SAM family protein and an alpha/beta hydrolase domain protein, indicative of a duplicated cassette.

Due to approaches used for phylogenetic profiling more distantly related phylogenetic groupings of transposons were detected (Table 2) compared to using BLAST against NCBI nr database to identify best matches (Figure 4). For example, group 2 included matches to *Thermoplasma* species, and groups 3 and 4 to Sulfolobales (Table 2). However, sequences from other methanogens always produced the best BLAST matches (Figure 4), indicating that *M. burtonii* transposases did not appear to have been transferred across distant phylogenetic boundaries. Additional signatures of genes shared between *M. burtonii* and other methanogens include an entire *M. hungatei* CRISPR-region, and a cluster of RNA-directed DNA polymerase genes (*e.g.* possibly viral genes) common to the Methanosarcinales and *M. hungatei*.

Seven transposases have been found to be expressed in the cell during growth (Figure 4, triangles), indicating that transposases are active and not just markers of past genomic changes (Goodchild et al. 2004b; Burg D & Williams T, unpublished results). The three largest groups (1, 2 and 3) were all represented in the expressed list. A further indicator of the activity of transposons is the presence of seven ORFs interrupted by the coding region of a transposase, with five of the ORFs known to be expressed (detected by proteomics) (Table 3). Fragments of five transposases are also present in the genome and are likely to be remnants of transposition events that are gradually being purged from the genome; Mbur_0774 and Mbur_0265 are fragments of a group 2 transposase, Mbur_0771 is a fragment of a group 1 transposase, Mbur_2251 is related to either Mbur_800 or Mbur_2016, and Mbur_0442 is a fragment of a group 6 transposase. The coding region of two of the five transposase fragments have themselves been interrupted by transposable elements; Mbur_0771 possibly by Mbur_0774 (also truncated), and Mbur 2251 by Mbur_2252.

Transposases [L] were one of the main COG categories distinguishing the psychrophilic *M. burtonii* from other archaea (see ***M. burtonii* genome overview**). In association with the above genomic and proteomic data, it appears that transposases play an important role in evolving the genome of *M. burtonii*. Transposase inactivation has previously been linked to cold sensitivity in the deep-sea, cold-adapted bacterium *Photobacterium profundum* SS9 (Lauro et al. 2008), and transposases have been found to be one of the most over-represented COG categories in metagenomic data of cold, deep (4000 m) ocean water samples (DeLong et al. 2006).

Functional capacity is associated with distinct phylogenetic clusters of genes

The largest phylogenetic cluster (106 proteins) was common to all archaeal species and included ribosomal proteins, chaperonins, translation machinery and 89 hypothetical proteins (Table 2). Other clusters with phylogenetic distribution across many archaeal species include groups of signal

transduction proteins, chemotaxis proteins, serine protease inhibitors, O-phosphoseryl-tRNA (Cys) synthetase-linked proteins, and DNA-gyrase-linked proteins. The latter two groups are particularly interesting as they link a well-characterized protein function (ER2) with a protein of unknown function in close phylogenetic association, even though the genes are physically separated on the genome. From this, it can be inferred that the DUF39-domain containing protein Mbur_0311 may be involved in production of Cys-tRNA^{Cys} via phosphoserine, and based on the child-parent relationship of the DUF1119 and peptidase A22 InterPro domains the DUF1119 protein Mbur_1910 may have peptidase function involved in assisting the action of DNA gyrase. In six cases, poorly characterized genes (ER4 and ER5) clustered with well characterized methanogenesis genes (only present in methanogenic archaea). The phylogenetic profile suggests that all the genes in the clusters are involved in methanogen-specific metabolic processes. A similar approach has been used to characterize hypothetical genes known to be expressed in *M. burtonii* (Saunders et al. 2005).

General metabolism and transport

Central metabolism appears overall to be similar to other methanogens. Complete pathways for glycolysis (via a modified Embden-Meyerhof pathway) and gluconeogenesis are present. While no glycogen phosphorylase could be found, other enzymes of starch metabolism are present suggesting that *M. burtonii* can store carbon as starch or glycogen. Glycogen serves as a polysaccharide storage reserve in many methanogens. The only pentose synthesis pathway present is the reverse ribulose monophosphate pathway, similar to other archaea.

M. burtonii possesses carbon monoxide dehydrogenase/acetyl-coenzyme A synthase (CODH/ACS; see **Methanogenesis**), to produce acetyl-CoA from methyl-tetrahydrosarcinapterin and endogenously generated carbon dioxide. Although in a previous report *M. burtonii* was suggested to also be capable of carbon-fixation via RubisCO as it possesses a type III ribulose-1,5-bisphosphate carboxylase/oxygenase (Mbur_2322) (Goodchild et al. 2004b), no identifiable gene for phosphoribulokinase has been found in the completed genome. Further, *M. burtonii* does not grow autotrophically (Franzmann et al. 1992), and there is no evidence for operation of either the reductive acetyl-CoA pathway or Calvin-Benson-Bassham cycle for carbon fixation. Archaeal (type III) RubisCO was recently demonstrated to be involved in AMP metabolism in the archaeon *Thermococcus kodakaraensis* (Sato et al. 2007). In this pathway, AMP phosphorylase (Mbur_0255) and ribose-1,5-bisphosphate isomerase (Mbur_1938) supply RubisCO substrate (ribulose-1,5-bisphosphate, RuBP) from AMP and phosphate. RuBP can then be converted by RubisCO to two 3-phosphoglycerate (3-PGA) molecules, which can then enter central carbon metabolism. As homologous genes for all three enzymes are present, we suggest the *M. burtonii* RubisCO functions in the AMP phosphorylase pathway similar to *T. kodakaraensis*.

M. burtonii also has ADP-dependent (AMP-forming) sugar kinases that are used by many archaea in glycolysis. Given that 3-PGA can be used for ATP generation, these archaea may use the above pathway under anaerobic conditions to utilize AMP when energy levels are low and/or intracellular AMP levels are high (Sato et al. 2007). Alternatively, AMP can be produced from 5-phosphoribosyl-1-pyrophosphate (PRPP) by adenine phosphoribosyltransferase (Mbur_1435), which also recycles the adenine generated by AMP phosphorylase.

Pyruvate synthase (Mbur_2155 - 2158) and pyruvate carboxylase (Mbur_2425 - 2426) provide a pathway for synthesis of oxaloacetate from acetyl-CoA. Synthesis of 2-oxoglutarate from oxaloacetate and acetyl-CoA is likely to take place via a truncated oxidative TCA cycle. The genes involved in the reductive steps of the TCA cycle from oxaloacetate to 2-oxoglutarate are almost completely missing, with only a heterodimeric fumarase (Mbur_0250 – 0251) present. In the oxidative direction, citrate synthase was initially thought to be missing on the basis of the automated genome annotation. However, a citrate synthase with alternate stereo-specificity (known as Re-citrate synthase) has been characterized in a *Clostridium* species, and an ortholog in *M. burtonii* was identified during the manual annotation process (Mbur_1075) (Li et al. 2007). Aconitase (Mbur_0316) is present (see **Cold adaptation is associated with specific signatures of genome evolution**), and Mbur_1073 (isocitrate/isopropylmalate dehydrogenase) is a likely candidate to catalyze the remaining step from isocitrate to 2-oxoglutarate.

M. burtonii has biosynthetic pathways for pyrrolysine in addition to the 20 standard amino acids. Akin to *Methanosarcina* spp., *M. burtonii* has two pathways for cysteine synthesis; the tRNA-dependent pathway (Sauerwald et al. 2005) and the O-acetylserine pathway. Cysteinyl-tRNA synthetase is absent, so the tRNA-dependent pathway is probably the only way to make cysteine for incorporation into proteins. Pyrrolysyl-tRNA synthetase (Mbur_2086) is located adjacent to the putative gene cassette for pyrrolysine synthesis (Mbur_2083 to 2085) (Longstaff et al. 2007); thus pyrrolysine is synthesized before being attached to its tRNA. Evidence for the incorporation of pyrrolysine in methyltransferase enzymes has been provided from proteomic studies (Goodchild et al. 2004a).

The way in which *M. burtonii* fixes nitrogen is unclear from the genome sequence. One *nifH* and one *nifD* homolog are present, but they cluster phylogenetically with group IV nitrogenase-related genes which are not expected to be involved in nitrogen fixation (Raymond et al. 2004). To test the ability of *M. burtonii* to fix nitrogen, cells were grown in defined media with N_2 gas as the sole source of nitrogen (D Burg, HH Chew & M Allen, unpublished results). Growth was observed after several generations and repeated passaging, although rates of growth and growth yield were lower than in media supplemented with an organic or inorganic nitrogen source. Preliminary analysis of *nifH* mRNA levels indicate that gene expression occurs even when cells are grown in complex

media (HH Chew, D Burg & M Allen, unpublished results). These data indicate that *M. burtonii* fixes nitrogen via an as yet, undefined pathway.

Nitrate reductase and nitrate transporters can not be identified in the genome sequence. The only member of the ammonium transporter family in *M. burtonii* (Mbur_0933) is situated adjacent to the gene for the regulatory PII protein (GlnB) (Mbur_0934). Mbur_0933 has 64% sequence identity to the characterized transporter Amt-3 from *A. fuldgidus*. However, it is truncated at both the N-terminus and the C-terminus by approximately 88 amino acids so its functional status is uncertain. Given that ammonia concentrations increase with depth in Ace Lake (Butler et al. 1988; Rankin et al. 1999), it is possible that the ammonia requirements of *M. burtonii* can be met via inefficient facilitated diffusion of ammonium ion, and/or by the passive diffusion of ammonia. After uptake, *M. burtonii* may assimilate ammonia using, (1) the two-step glutamine synthetase (Mbur_1975) and glutamate synthase (Mbur_0092) pathway (GS-GOGAT), and/or (2) glutamate dehydrogenase (GDH) (Mbur_1973). 2-oxoglutarate serves as the carbon skeleton for both pathways, and can be provided by a partial oxidative TCA cycle (see above). The glutamate synthase of *M. burtonii* appears to be a ferredoxin-dependent species. All three proteins are expressed in *M. burtonii* under laboratory growth conditions (Goodchild et al. 2004a; Goodchild et al. 2004b; Goodchild et al. 2005; Burg D & Williams T, unpublished results).

Methanogenesis

As an obligately methylotrophic methanogen, *M. burtonii* obtains energy from the oxidation of methyl groups to carbon dioxide and reduction to methane (Figure 5). Methyltransferases required to initiate metabolism by demethylation of methanol, monomethylamine, dimethylamine and trimethylamine and subsequent methylation of their respective corrinoid binding polypeptides were detected. Multiple copies of the each substrate specific methyltransferase and corrinoid protein were present. However, the second gene encoding dimethylamine methyltransferase (Mbur_2291) is disrupted by a putative transposon and is likely to be non-functional. Three copies of the methyl CoM methyltransferase for transfer of the methyl group from the methanol specific corrinoid, and two copies of the common methyl CoM methyltransferase for methyl transfer from the methylated amine specific corrinoid proteins, were identified. This is similar to observations in the closely related methansarcinal genomes, which also have multiple copies of initiating methyltransferases and cognate corrinoid proteins (Galagan et al. 2002; Deppenmeier et al. 2002; Maeder et al. 2006). However, in the methanosarcinal genomes only single copies of each CoM methyltransferase are present. Methyltransferases specific for methylthiols, which are found in some methanosarcinal species, were not present in *M. burtonii* (Tallant & Krzycki 1997; Tallant et al. 2001). During methylotrophic growth, methyl CoM is disproportionated 3:1 by oxidation to carbon dioxide and

reduction to methane, respectively. All of the genes encoding for enzymes required for methyl reduction to methane and oxidation of methyl CoM via the reversal of the carbon dioxide reduction pathway are present in the *M. burtonii* genome (Figure 5).

Methanosarcina spp. are capable of growth by all four known methanogenic pathways: hydrogen reduction of carbon dioxide, hydrogen reduction of methylated compounds, fermentation of acetate and dismutaton of methylotrophic compounds. The genome of *M. burtonii* provides the first direct insight into the minimal genes required for obligate methylotrophy by a methanogen. Growth with hydrogen requires three hydrogenases: Ech, which catalyzes ferredoxin reduction by hydrogen for subsequent reduction of carbon dioxide to the formyl level; Frh/Fre, which reduce coenzyme F_{420} for reduction of methenyl and methylene groups; Vho, which catalyzes the reduction of the coenzyme CoM reducing electron carrier methanphenazine (Meuer et al. 2002). Except for one protein with similarity to a Coenzyme F_{420}-reducing hydrogenase beta subunit (Mbur_2261), genes encoding these enzymes were not detected in *M. burtonii*; consistent with the inability of this species to grow with hydrogen (Franzmann et al. 1992). During growth, F_{420} dehydrogenase (Fpo) could reduce methanphenazine from reduced F_{420} generated during the oxidation steps of methyl and methylene groups in the methanogenic pathway. Reduced ferredoxin generated by oxidation of formylmethanofuran to carbon dioxide could be providing reducing potential for biosynthesis. *M. acetivorans,* which grows with acetate or methylotrophic substrates but does not grow with hydrogen, also lacks the Ech hydrogenase and although it has *frh* and *vho* genes they are not expressed (Guss et al. 2005). This latter observation combined with the lack of hydrogenase genes in *M. burtonii* confirms that hydrogenases are not required for the methylotrophic pathway.

Hydrogen reduction of methylated compounds requires the specific methyltransferases and the reductive methanogenic pathway after methanopterin. However, this pathway also requires electrons from oxidation of hydrogenase to reduce methyl groups, which are not present in *M. burtonii.* Aceticlastic methanogenesis requires CODH/ACS to catalyze the methylation of methanopterin via dismutation of acetyl CoA. In contrast to two gene copies in aceticlastic *Methanosarcina* spp., *M. burtonii* genome has only one copy of the CODH/ACS operon and lacks genes encoding acetate kinase and phophotransacetylase for generating acetyl CoA from acetate. This is consistent with other non-aceticlastic methanogens that likely use CODH/ACS for carbon assimilation from carbon dioxide. Overall, the limited catabolic capabilities of *M. burtonii* are consistent with the small genome size relative to *Methanosarcina* spp.; a characteristic shared with the genomes of other methanogens with a single catabolic pathway.

In one respect the ability of *Methanosarcina* spp. to grow with all known methanogenic substrates provides them with the ability to adapt their catabolism to changes in substrate availability. In contrast *M. burtonii* is obligately methylotrophic and restricted to methanol and

methylamines as exogenous carbon sources, which raises the question of how this is an advantage for the survival of this species as a specialist. One phenomenon observed for *Methanosarcina* spp. is increased hydrogen production on acetate and methylotrophic substrates when grown in the presence of a sulfate reducing species (Phelps et al. 1995). This also results in a reduction in methane yield per mole of substrate as more substrate is oxidized and reducing equivalents are diverted to hydrogen. This shift in the pathway would effectively reduce the energy yield for the methanogen and enable sulfate reducing bacteria to utilize energy from the non-competitive methylotrophic substrate via the reverse methanogenic pathway of the methanogen. The lack of hydrogenases in *M. burtonii* allows this specialist to utilize methylotrophic substrates without diversion of electrons to sulfate reducers and prevent the indirect utilization of these substrates by the latter. This may have provided a growth advantage to obligate methylotrophs such as *M. burtonii* in Ace Lake when the lake was (in the past) a sulfate rich environment (see **Linking the evolution of the *M. burtonii* genome to its ecological niche**).

Signal transduction and adaptive potential

The signal transduction system of *M. burtonii* is similar to that of its *Methanosarcina* spp. relatives, despite *M. burtonii* having a much smaller genome size. *M. burtonii* is motile by means of a single flagellum (Franzmann et al. 1992) and encodes a chemotaxis system that includes a single methyl-accepting chemotaxis protein (Mbur_0356), its methylase and demethylase (Mbur_0360, Mbur_0399), chemotaxis histidine kinase CheA (Mbur_0361) and chemotaxis response regulator CheY (Mbur_0359). In addition to CheA, *M. burtonii* encodes 29 other two-component histidine kinases, of which 10 contain a (predicted) extracytoplasmic sensor domain and are probably involved in environmental sensing (Galperin 2006). The remaining histidine kinases are intracellular; most of them contain PAS domains and can be involved in sensing of the cellular level of oxygen, CO, NO and other molecules. This could be particularly important for *M. burtonii* because this strict anaerobe might have to cope with the higher solubility of oxygen at low temperatures. Four of the *M. burtonii* histidine kinases have an N-terminal receiver (REC) domain similar to the proteins described recently in the haloarchaea (Galperin 2006), and are likely to participate in complex phosphorelay signal transduction cascades.

Of the 14 response regulators encoded in the *M. burtonii* genome, eight consist of a single stand-alone REC domain and two more have a REC-PAS domain architecture. All these response regulators are likely involved in protein-protein interactions. Similar to other archaea, *M. burtonii* encodes no response regulators of the OmpR, NarL, NtrC, PrrA, or LytR families. However, it encodes a response regulator (Mbur_0695) with a DNA-binding output domain of the GlpR type, which forms a typical operon-like structure with an environmental sensor histidine kinase Mbur_0694. This

regulator is more abundant in cells grown at low temperature, and may form a temperature responsive regulatory system with the histidine kinase (Goodchild et al. 2004a). Two more response regulators (Mbur_0878, Mbur_2185) contain previously uncharacterized output domains, one unique for *M. burtonii* and the other found only in *Methanosarcina* spp. In addition, *M. burtonii* encodes an adenylate cyclase (Mbur_1935) and five predicted Ser/Thr protein kinases, one of the ABC1/AarF family, two of the RIO family, one unusual protein kinase and one Kae1-associated serine threonine kinase. Similar to other archaea, the role of cAMP, presumably synthesized by the adenylate cyclase, remains unknown, as are the cellular targets of the (predicted) protein kinases.

The large number of two-component regulatory systems in *M. burtonii* with limited similarity to known functional proteins may reflect a requirement for complex internal regulation (Ashby 2006), and *M. burtonii* has a high "IQ" (a measure of the adaptive potential of an organism) compared to many methanogens (Galperin 2005). The majority of signal transduction genes (35 out of 45 proteins) have a functional evidence rating of ER4 (indicating a lack of experimentally characterized full length homologs). In addition, 5 ORFs listed as pseudogenes contain histidine kinase domains, and at least 2 of these have been interrupted by transposons and appear to be non-functional. Given the general lack of experimentation on two component regulatory systems in archaea and the specific importance of signal transduction systems to *M. burtonii* (see **Cold adaptation is associated with specific signatures of genome evolution**), there is a compelling reason to experimentally characterize the sensing and response mechanisms of these phosphorelay networks.

Lipid biosynthesis

An important pathway associated with cold adaptation in *M. burtonii* is isoprenoid lipid biosynthesis (Nichols et al. 2004). Synthesis of unsaturated lipids increases during growth at low temperature and is thought to maintain the fluidity, and hence functionality of the membrane in the cold. The DGGGP synthase that was not present in the draft genome, was identified as Mbur_1679 in the closed genome (Figure 6). Phosphomevalonate kinase was also not previously identified, and to date the gene has only been identified in the Sulfolobales (Boucher 2007). Recently *M. jannaschii* was hypothesized to use a modified mevalonate pathway, with the two steps required for conversion of mevalonate phosphate to isopentenyl diphosphate catalyzed by a novel (although not yet experimentally verified) mevalonate phosphate decarboxylase, followed by isopentenyl phosphate kinase (Grochowski et al. 2006). The corresponding enzymes in *M. burtonii* are Mbur_2394 and Mbur_2396, respectively. This mevalonate pathway differs from the steps previously proposed for *M. burtonii*, which requires diphosphomevalonate kinase and diphosphomevalonate decarboxylase (Nichols et al. 2004). In the *M. burtonii* genome, Mbur_2394 and Mbur_2396 lie within a putative

gene cluster associated with lipid biosynthesis, including mevalonate kinase (Mbur_2395), isopentenyl- diphosphate delta-isomerase (Mbur_2397), and a bifunctional enzyme comprising farnesyl pyrophosphate synthetase and geranyl*trans*transferase (Mbur_2399). Other important members of the lipid biosythesis pathway include an experimentally characterized geranylgeranyl reductase responsible for conversion of DGGGP to archaetidic acid (Mbur_1077) (Murakami et al. 2007). This enzyme is thought to play an important role in facilitating the regulation of unsaturation levels by performing selective saturation (similar to plants) (Nichols et al. 2004). Other genes characterized on the basis of COG groupings involved in lipid formation include acetyl-CoA synthetases (3 proteins), acyl-CoA synthetase, pyruvate decarboxylase (alpha and beta subunits located adjacent to each other in the genome), phosphatidyltransferases and cytidyltransferases.

A large genomic committment to polysaccharide biosynthesis

Saccharide or polysaccharide moieties are important modifiers of the isoprenoid lipid membrane (Jahn et al. 2004) and S-layer (Karcher et al. 1993) of archaea, and can be secreted as EPS (Parolis et al. 1996; Paramonov et al. 1998). In *M. burtonii*, genes involved in polysaccharide biosynthesis comprise at least 3.3% of protein coding genes in the genome. In contrast to the 81 *M. burtonii* genes, the mesophilic *M. acetivorans* only contains approximately 30 polysaccharide biosynthesis genes representing 0.6% of its genome (Galagan et al. 2002). Four operon-like clusters of polysaccharide biosynthesis genes (containing 39, 16, 11 and 10 genes) and 5 single glycosyl-transferase genes are distributed around the *M. burtonii* genome (Figure 1). In addition, 5 unique hypothetical proteins are encoded in close proximity to the polysaccharide biosynthesis gene clusters. Five homologs (Mbur_0724, Mbur_0725, Mbur_0726, Mbur_1581, Mbur_2023) of a glycosyl transferase (COG 0438), and four homologs (Mbur_0727, Mbur_1593, Mbur_2028 and Mbur_2225) of a membrane protein involved in the export of O-antigen and teichoic acid (COG2244) are present in four main regions of the genome where they are adjacent to a number of other genes involved in polysaccharide biosynthesis, such as sugar- and N-acetylglucosamine-epimerases and sugar dehydratases (Figure 1). The proteins represented by COG 0438 and COG2244 are homologs to *P. profundum* SS9 proteins, PBPRA2678 and PBPRA2684, respectively. Mutation of the *P. profundum* SS9 genes produces a cold-sensitive phenotype (Lauro et al. 2008; Ferguson et al. personal communication). EPS, including polysaccharides appear to play a role in cell aggregation and biofilm formation of archaea at low temperatures (Reid et al. 2006). Higher levels of EPS are produced by *M. burtonii* growing at low (compared to high) temperatures (Reid et al. 2006), and approximately half of the *M. burtonii* polysaccharide biosynthesis genes are known to be expressed, having been shown to be abundant proteins in proteomic analyses (Goodchild et al. 2004a; Goodchild et al. 2004b; Goodchild et al. 2005; Saunders et al. 2005; Burg D & Williams T, unpublished results). Collectively

these data strongly suggest that polysaccharide biosynthesis plays an important role in the cold adaptation of *M. burtonii*.

Linking the evolution of the *M. burtonii* genome to its ecological niche

Ace Lake is located on Long Peninsula in the Vestfold Hills, East Antarctica (Rankin et al. 1999). During the early Holocene (13000-9400 years ago), Ace Lake was an aerobic freshwater system, becoming a seasonally isolated marine basin (9400-9000 ya) and subsequently open marine basin (9400-5700 ya) (Rankin et al. 1999; Coolen et al. 2004; Cromer et al. 2005). Approximately 5100 ya Ace Lake became a permanently isolated saline lake, and developed meromixis, with an active methane cycle in existence for last ~3000 y. The microbiota in the lake is clearly marine derived. However, isotopic data indicate that all the water now present in the lake is of meteoric origin and that the lake has been mixed for considerable periods prior to its present stable meromixis. Nutrient input presently into the lake is very limited and it is a cold (average ~0°C), oligotrophic system, although inorganic carbon levels are sufficient to lead to carbon dioxide efflux (Rankin et al. 1999). The concentrations of most trace metals are higher than the ocean and are not limiting. Gradients of nutrients occur throughout the lake and concentrations can vary significantly, *e.g.* manganese concentration varies from 78 to 1460 nM within 5 m in the anaerobic zone and this concentration is orders of magnitudes higher than the 5.1 nM of the ocean. The anoxic waters (12-25 m depth) support stable increasing gradients of salt (up to 4.3%), methane (saturated below 20 m, ~5mM) and H_2S (up to 8mM), and decreasing gradients of sulfate (essentially depleted below 19 m) (Rankin et al. 1999).

Close relatives of *M. burtonii* have been identified in several cold ocean water locations in south and north polar marine waters and the deep-sea, including *Methanococcoides alaskense* which has 99.8% 16S rRNA identity (Li et al. 1999; Purdy et al. 2003; Singh et al. 2005; Cavicchioli 2006). Cold, deep sea environments in the Atlantic Ocean have been found rich in extracellular DNA, and may promote opportunities for DNA exchange (Dell'Anno and Danovaro, 2005). The adaptation of *M. burtonii* to the cold will have occurred over millennia in the cold marine environment. However, the large environmental changes that have taken place in Ace Lake since the Holocene are likely to have provided strong selection pressure for ecotypes with genomic variation better suited to the new environment. In this regard it would be valuable to sequence the genome of *M. alaskense* in order to assess this. The specific capacity of *M. burtonii* to evolve through genome plasticity (including nucleotide skew, HGT and transposase activity) appears to have placed it in a strong position to not only adapt to the cold (*e.g.* polysaccharide synthesis, lipid composition, amino acid composition), but to the particular biotic and abiotic conditions that have changed in the lake throughout its recent several thousand year history (*e.g.* central metabolism, novel ABC transporters, coenzyme F_{420}-dependent sulfite reductase).

Acknowledgements

The Australian contingent was supported by funding from the Australian Research Council. The work of IJA, NI, ED, MM, AL, LH and ML was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, Los Alamos National Laboratory under contract No. DE-AC02-06NA25396, and Los Alamos National Laboratory under contract No. DE-AC05-00OR22725. The work of KRS was supported by funding from the US Department of Energy's Office of Science, Biological and Environmental Research Program grant No. DE-FG02-07ER64502 and the National Science Foundation, Division of Cellular and Bioscience grant No. MCB0110762. MYG is supported by the NIH Intramural Research Program at the National Library of Medicine.

References

Ashby MK (2006) Distribution, structre and diversity of "bacterial" genes encoding two-component proteins in the Euryarchaeota. *Archaea* **2:** 11-30

Azad RK, Lawrence JG (2007) Detecting laterally transferred genes: use of entropic clustering methods and genome position. *Nucleic Acids Res.* **35:** 4629-4639

Boucher Y (2007) Lipids: biosynthesis, function and evolution. In *Archaea: Molecular and Cellular Biology* (ed. R. Cavicchioli), pp. 341-353. ASM Press, Washington D.C.

Bowman JP, Rea SM, McCammon SA, McMeekin TA (2000) Diversity and community structure within anoxic sediments from marine salinity meromictic lakes and a coastal meromictic marine basin, Vestfold Hills, Eastern Antarctica. *Environ. Microbiol.* **2:** 227–237

Butler ECV, Burton HR, Smith JD (1988) Iodine distribution in an Antarctic meromictic saline lake. *Hydrobiologia* **165:** 97-101

Cavicchioli R (2006) Cold adapted Archaea. *Nat. Rev. Microbiol*. **4:** 331-343

Cavicchioli R, DeMaere M, Thomas T (2007) Metagenomic studies reveal the critical and wide-ranging ecological importance of uncultivated archaea: the role of ammonia oxidizers. *BioEssays* **29:** 11–14

Chain PSG, Denef VJ, Konstantinidis KT, Vergez LM, Agullo L, Reyes VL *et al.* (2006) *Burkholderia xenovorans* LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. *Proc. Natl. Acad. Sci. USA* **103:** 15280-15287

Chong SC, Liu Y, Cummins M, Valentine DL, Boone DR (2002) *Methanogenium marinum* sp. nov., a H_2-using methanogen from Skan Bay, Alaska, and kinetics of H_2 utilization. *Antonie van Leeuwenhoek* **81:** 263–270

Coolen MJL, Hopmans EC, Rijpstra WIC, Muyzer G, Schouten S, Volkman JK *et al*. (2004) Evolution of the methane cycle in Ace Lake (Antarctica) during the Holocene: response of methanogens and methanotrophs to environmental changes. *Org. Geochem.* **35:** 1151-1167

Costa J, Empadinhas N, Gonçalves L, Lamosa P, Santos H, da Costa MS (2006) Characterization of the biosynthetic pathway of glucosylglycerate in the archaeon *Methanococcoides burtonii*. *J. Bacteriol.* **188:** 1022-1030

Cromer L, Gibson JAE, Swadling KTM, David RA (2005) Faunal microfossils: Indicators of Holocene ecological change in a saline Antarctic lake. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **221:** 83–97

Dell'Anno A, Danovaro, R (2005) Extracellular DNA plays a key role in deep-sea ecosystem functioning. *Science* **309:** 2179

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311:** 496-503

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

Enault F, Suhre K, Abergel C, Poirot O, Claverie JM (2003) Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinformatics* **19:** i105-i107

Fournier GP, Gogarten JP (2008) Evolution of acetoclastic methanogenesis in *Methanosarcina* via horizontal gene transfer from cellulolytic *Clostridia. J. Bacteriol.* **190:** 1124-1127

Franzmann PD, Liu Y, Balkwill DL, Aldrich HC, Conway de Macario E, Boone DR (1997) *Methanogenium frigidum* sp. nov., a psychrophilic, H_2-using methanogen from Ace Lake, Antarctica. *Int. J. Syst. Bacteriol.* **47:** 1068–1072

Franzmann PD, Springer N, Ludwig W, Conway de Macario E, Rohde M (1992) A methanogenic archaeon from Ace Lake, Antarctica: *Methanococcoides burtonii* sp. nov. *Syst. Appl. Microbiol.* **15:** 573-581

Freeman TC, Goldovsky L, Brosch M, van Dongen S, Mazière P, Grocock RJ *et al.* (2007) Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput. Biol.* **3:** 2032-42

Galagan JE, Nusbaum C, Roy A, Endrizzi MG, Macdonald P, FitzHugh W *et al.* (2002) The Genome of *M. acetivorans* Reveals Extensive Metabolic and Physiological Diversity. Genome Res 12: 532-542

Galperin MY (2005) A census of membrane-bound and intracellular signal transduction proteins in bacteria: Bacterial IQ, extroverts and introverts. *BMC Microbiol.* **5:** 35

Galperin MY (2006) Structural classification of bacterial response regulators: diversity of output domains and domain combinations. *J. Bacteriol.* **188:** 4169-4182

Goodchild A, Raftery M, Saunders NFW, Guilhaus M, Cavicchioli R (2004a) Biology of the cold adapted archaeon, *Methanococcoides burtonii* determined by proteomics using liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **3:** 1164-1176

Goodchild A, Raftery M, Saunders NFW, Guilhaus M, Cavicchioli R (2005) Cold adaptation of the Antarctic archaeon, *Methanococcoides burtonii* assessed by proteomics using ICAT. *J. Proteome Res.* **4:** 473-480

Goodchild A, Saunders NFW, Ertan H, Raftery M, Guilhaus M, Curmi PMG (2004b) A proteomic determination of cold adaptation in the Antarctic archaeon, *Methanococcoides burtonii*. Mol. Microbiol. **53:** 309-321

Grochowski LL, Xu H, White RH (2006) *Methanocaldococcus jannaschii* uses a modified mevalonate pathway for biosynthesis of isopentenyl diphosphate. *J. Bacteriol.* **188:** 3192-3198

Guss AM, Mukhopadhyay B, Zhang JK, Metcalf WW (2005) Genetic analysis of mch mutants in two *Methanosarcina* species demonstrates multiple roles for the methanopterin-dependent C-1 oxidation/reduction pathway and differences in H(2) metabolism between closely related species. *Mol. Microbiol.* **55:** 1671-1680

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res.* **17:** 377-386

Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. *J. Comp. Graph. Stat.* **5:** 299-314

Ivanova NN, Mavromatis K, Chen I-MA, Markowitz VM, Kyrpide NC (2007) Standard operating procedure for the annotations of genomes and metagenomes submitted to the integrated microbial genomes expert review (IMG-ER) system. Available: http://imgweb.jgi-psf.org/w/doc/img_er_ann.pdf

Jahn U, Summons RE, Sturt H, Grosjean E, Huber H (2004) Composition of the lipids of *Nanoarchaeum equitans* and their origin from its host *Ignicoccus* sp. strain KIN4/I. *Arch. Microbiol.* **182:** 404-413

Johnson EF, Mukhopadhyay B (2005) A new type of sulfite reductase, a novel coenzyme F420-dependent enzyme, from the methanarchaeon *Methanocaldococcus jannaschii*. *J. Biol. Chem.* **280:** 38776-38786

Karcher U, Schroder H, Haslinger E, Allmaier G, Schreiner R, Wieland F *et al.* (1993) Primary structure of the heterosaccharide of the surface glycoprotein of *Methanothermus fervidus*. *J. Biol. Chem.* **268:** 26821-26826

Kendall MM, Wardlaw GD, Tang CF, Bonin AS, Liu Y, Valentine DA (2007) Diversity of Archaea in marine sediments from Skan Bay, Alaska, including cultivated methanogens, and description of *Methanogenium boonei* sp. nov. *Appl. Environ. Microbiol.* **73:** 407-414

Klotz MG, Arp DJ, Chain PSG, El-Sheikh AF, Hauser LJ, Hommes NG *et al.* (2006) Complete genome sequence of the marine, chemolithoautotrophic, ammonia-oxidizing bacterium *Nitrosococcus oceani* ATCC 19707. *Appl. Environ. Microbiol.* **72:** 6299-6315

Könneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437:** 543–546

Kotsyurbenko OR, Friedrich MW, Simankova MV, Nozhevnikova AN, Golyshin PN, Timmis KN *et al*. (2007) Shift from acetoclastic to H2-dependent methanogenesis in a west Siberian peat bog at low pH values and isolation of an acidophilic *Methanobacterium strain*. *Appl. Environ. Microbiol.* **73:** 2344-2348

Kruger M, Treude T, Wolters H, Nauhaus K, Boetius A (2005) Microbial methane turnover in different marine habitats. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **227:** 6–17

Lauro FM, Tran K, Vezzi A, Vitulo N, Valle G, Bartlett DG (2008) Large-scale transposon mutagenesis of *Photobacterium profundum* SS9 reveals new genetic loci important for growth at low temperature and high pressure. *J. Bacteriol.* **190:** 1699-1709

Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW *et al*. (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* **442:** 806-809

Li F, Hagemeier CH, Seedorf H, Gottschalk G, Thauer RK (2007) Re-citrate synthase from *Clostridium kluyveri* is phylogenetically related to homocitrate synthase and isopropylmalate synthase rather than to Si-citrate synthase. *J. Bacteriol.* **189:** 4299-4304

Li L, Kato C, Horikoshi K (1999) Microbial diversity in sediments collected from the deepest cold-seep area, the Japan Trench. *Mar. Biotechnol.* **1:** 391–400

Li L, Li Q, Rohlin L, Kim U, Salmon K, Rejtar T *et al.* (2007) Quantitative proteomic and microarray analysis of the Archaeon *Methanosarcina acetivorans* grown with acetate versus methanol. *J. Proteome Res.* **6:** 759-771

Longstaff DG, Larue RC, Faust JE, Mahapatra A, Zhang L, Green-Church KB *et al*. (2007) A natural genetic code expansion cassette enables transmissible biosynthesis and genetic encoding of pyrrolysine. *Proc. Natl. Acad. Sci. USA* **104:** 1021-1026

Lynn DJ, Singer GAC, Hickey DA (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.* **30:** 4272-4277

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

Medigue C, Krin E, Pascal G, Barbe V, Bernsel A, Bertin PN *et al.* (2005) Coping with cold: The genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125. *Genome Res.* **15:** 1325-1335

Methé BA, Nelson KE, Deming JW, Momen B, Melamud E, Zhang X *et al*. (2005) The psychrophilic lifestyle as revealed by the genome sequence of *Colwellia psychrerythraea* 34H through genomic and proteomic analyses. *Proc. Natl. Acad. Sci. USA* **102:** 10913-10918

Meuer J, Kuettner HC, Zhang JK, Hedderich R, Metcalf WW (2002) Genetic analysis of the archaeon *Methanosarcina barkeri* Fusaro reveals a central role for Ech hydrogenase and ferridoxin in methanogenesis and carbon fixation. *Proc. Natl. Acad. Sci. USA* **99:** 5632-5637

Moran JJ, House CH, Vrentas JM, Freeman KH (2008) Methyl sulfide production by a novel carbon monoxide metabolism in *Methanosarcina acetivorans*. *Appl. Environ. Microbiol.* **74:** 540-542

Morozova D, Wagner D (2007) Stress response of methanogenic archaea from Siberian permafrost compared with methanogens from nonpermafrost habitats. *FEMS Microbiol. Ecol.* **61:** 16–25

Murakami M, Shibuya K, Nakayama T, Nishino T, Yoshimura T, Hemmi H (2007) Geranylgeranyl reductase involved in the biosynthesis of archaeal membrane lipids in the hyperthermophilic archaeon *Archaeoglobus fulgidus*. *FEBS J.* **274:** 805-814

Murray AE, Grzymski JJ (2007) Diversity and genomics of Antarctic marine microorganisms. *Phil. Trans. R. Soc. B* **362:** 2259–2271

Narita S-I, Kanamaru K, Matsuyama S-I, Tokuda H (2003) A mutation in the membrane subunit of an ABC transporter LolCDE complex causing outer membrane localization of lipoproteins against their inner membrane-specific signals. *Mol. Microbiol.* **49:** 167-177

Nichols DS, Miller MR, Davies NW, Goodchild A, Raftery M, Cavicchioli R (2004) Cold adaptation in the Antarctic archaeon *Methanococcoides burtonii* involves membrane lipid unsaturation. *J. Bacteriol.* **186:** 8508-8515

Noon KR, Guymon R, Crain PF, McCloskey JA, Thomm M, Cavicchioli R (2003) Influence of temperature on tRNA modification in archaea: *Methanococcoides burtonii* (optimum growth temperature [Topt], 23 degrees C) and *Stetteria hydrogenophila* (Topt, 95 degrees C). *J. Bacteriol.* **185:** 5483-5490

Paramonov NA, Parolis LAS, Parolis H, Boan IF, Anton J, Rodriguez-Valera F (1998) The structure of the exocellular polysaccharide produced by the archaeon *Haloferax gibbonsii* (ATCC 33959). *Carbohyd. Res.* **309:** 89-94.

Parolis H, Parolis LAS, Boan IF, Rodriguez-Valera F, Widmalm G, Manca MC *et al*. (1996) The structure of the exopolysaccharide produced by the halophilic archaeon *Haloferax mediterranei* strain R4 (ATCC 33500). *Carbohyd. Res.* **295:** 147-156.

Peden J (2000) CodonW. Available: http://codonw.sourceforge.net/.

Phelps TJ, Conrad R, Zeikus JG (1985) Sulfate-dependent interspecies H2 transfer between *Methanosarcina barkeri* and *Desulfovibrio vulgaris* during coculture metabolism of acetate or methanol. *Appl. Environ. Microbiol.* **50:** 589-594

Purdy KJ, Nedwell DB, Embley TM (2003) Analysis of the sulfate-reducing bacterial and methanogenic archaeal populations in contrasting Antarctic sediments. *Appl. Environ. Microbiol.* **69:** 3181–3191

Rabus R, Ruepp A, Frickey T, Rattei T, Fartmann B, Stark M *et al*. (2004) The genome of *Desulfotalea psychrophila*, a sulfate-reducing bacterium from permanently cold Arctic sediments. *Environ. Microbiol.* **6:** 887–902

Rankin LM, Gibson JAE, Franzmann PD, Burton HR (1999) The chemical stratification and microbial communities of Ace Lake, Antarctica: a review of the characteristics of a marine-derived meromictic lake. *Polarforschung* **66:** 33–52

Raymond J, Siefert JL, Staples CR, Blankenship RE (2004) The natural history of nitrogen fixation. *Mol. Biol. Evol.* **21:** 541-554

Reid IN, Sparks WB, Lubnow S, McGrath M, Livio M, Valenti J *et al.* (2006) Terrestrial models for extraterrestrial life: methanogens and halophiles at Martian temperatures. *Int. J. Astrobiol.* **5:** 89-97

Riley M, Staley JT, Danchin A, Wang TZ, Brettin TS, Hauser LJ *et al*. (2008) Genomics of an extreme psychrophile, *Psychromonas ingrahamii*. *BMC Genomics* **9:** 210

Rodriguez-Brito B, Rohwer F, Edwards RA (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics 7:* 162

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406-425

Sato T, Atomi H, Imanaka T (2007) Archaeal type III RuBisCOs function in a pathway for AMP metabolism. *Science* **315:** 1003-1006

Sauerwald A, Zhu W, Major TA, Roy H, Palioura S, Jahn D *et al.* (2005) RNA-dependent cysteine biosynthesis in archaea. *Science* **307:** 1969-1972

Saunders NFW, Thomas T, Curmi PMG, Mattick JS, Kuczek E, Slade R *et al.* (2003) Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea, *Methanogenium frigidum* and *Methanococcoides burtonii*. *Genome Res.* **13:** 1580-1588

Saunders NFW, Goodchild A, Raftery M, Guilhaus M, Curmi PMG, Cavicchioli R (2005) Predicted roles for hypothetical proteins in the low-temperature expressed proteome of the Antarctic archaeon *Methanoccoides burtonii*. *J. Proteome Res.* **4:** 464-472

Simankova MV, Parshina SN, Tourova TP, Kolganova TV, Zehnder AJ, Nozhevnikova AN (2001) *Methanosarcina lacustris* sp. nov., a new psychrotolerant methanogenic archaeon from anoxic lake sediments. *Syst. Appl. Microbiol.* **24:** 362–367

Singh N, Kendall MM, Liu Y, Boone DR (2005) Isolation and characterization of methylotrophic methanogens from anoxic marine sediments in Skan Bay, Alaska: description of *Methanococcoides alakenese* sp. nov., and emended description of *Methanosarcina baltica*. *Int. J. Syst. Evol. Microbiol.* **55:** 2531–2538

Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318:** 1449-1452

Tallant TC, Krzycki JA (1997) Methylthiol:Coenzyme M methyltransferase from *Methanosarcina barkeri*, an enzyme of methanogenesis from dimethylsulfide and methylmercaptopropionate. *J. Bacteriol.* **179:** 6902-6911

Tallant TC, Paul L, Krzycki JA (2001) The MtsA subunit of the methylthiol:coenzyme M methyltransferase of *Methanosarcina barkeri* catalyses both half-reactions of corrinoid-dependent dimethylsulfide: coenzyme M methyl transfer. *J. Biol. Chem.* **276:** 4485-4493

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24:** 1596-1599

Thomas T, Kumar N, Cavicchioli R (2001) Effects of ribosomes and intracellular solutes on activities and stabilities of elongation factor 2 proteins from psychrotolerant and thermophilic methanogens. *J. Bacteriol.* **183:** 1974-1982

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673-4680

Tsirigos A, Rigoutsos I (2005) A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res.* **33:** 922-933

Vernikos GS, Parkhill J (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* **22:** 2196-2203

von Klein D, Arab H, Völker H, Thomm M (2002) *Methanosarcina baltica* sp. nov., a novel methanogen isolated from the Gotland Deep of the Baltic Sea. *Extremophiles* **6:** 103–110

Figure legends

Figure 1. Circular representation of the *Methanococcoides burtonii* genome. The circles show (outermost to innermost): 1, DNA coordinates (black); 2, genes on forward strand, colour coded by COG categories; 3, genes on reverse strand, colour coded by COG categories; 4, RNA genes, including tRNAs (green), rRNAs (red) and other RNAs (black); 5, genes involved in putative polysaccharaide/capsule biosynthesis operons (red); 6, GC content; 7, GC skew.

Figure 2. Putative ABC-Transporters involved in cell defense. (A) COG0577, ABC-type antimicrobial peptide transport system, permease component, all annotated DUF214 protein (ER4); COG4591, ABC-type transport system, involved in lipoprotein release, permease component, all annotated DUF214 protein (ER4); COG1136, ABC-type antimicrobial peptide transport system, ATPase component, all annotatated ABC transporter ATPase (ER2); COG1361, S-layer domain, all annotated hypothetical protein (ER5). (B) Unrooted phylogenetic tree of substrate-binding proteins from ABC transporters constructed using the neighbour-joining method from a multiple alignment of amino acid sequences of COG1361 hypothetical proteins (aqua), S-layer proteins (pink) and ABC-transporter substrate-binding proteins (black) from *M. burtonii*. The sequence alignments are in Figure S2.

Figure 3. Predictions of genome plasticity in *M. burtonii.* (A) Percentage of genome predicted to have undergone horizontal gene transfer by the Alien Hunter program. Bars are colour coded according to the maxiumum growth temperature of the organism: purple, <30ºC; blue, 40-49ºC; green, 50-59ºC; yellow, 60-79ºC; orange, 80-99ºC; red, 100ºC and higher. (B) COG category classification (top pie charts), normalized expression (center bar graphs) and phylogenetic assignments (lower pie charts) of the ORFs contained in islands with different IVOM scores. Norm (normal), IVOM ≤ 17.223; low, 17.223 < IVOM < 26; medium, 26 ≤ IVOM < 40; high, 40 ≤ IVOM. (C) Correspondance analysis of codon usage of ORFs in regions with different IVOM scores. Norm (normal), IVOM ≤ 17.223; low, 17.223 < IVOM < 26; medium, 26 ≤ IVOM < 40; high, 40 ≤ IVOM. (D) 106446 ORFs from archaeal genomes represented in Figure 3a, and ORFs from the draft sequence of *M. frigidum* were subjected to correspondance analysis of codon usage (Lynn et al. 2002). Growth temperatures: purple, <30°C; blue, 40-49°C; green, 50-59°C; yellow, 60-79°C; orange, 80-99°C; red, 100°C and higher.

Figure 4. Phylogeny of transposases in *M. burtonii*. 63 transposases grouped in 6 major clusters: Group 1 is found in environmental genomic fragments from deep-sea sediments (GZfos19A5) and has multiple hits to *M. acetivorans* (3) and *M. mazei* (5). Group 2 is found in environmental genomic fragments from deep-sea sediments (GZfos27B6, GZfos9C4, GZfos26D6) and to *M. acetivorans* (1).

Group 3 has multiple hits to *M. acetivorans* (11), *M. mazei* (1) and *M. barkeri* (5) and is probably an IS1-like element. Group 4 has hits to *M. mazei* (2) and is likely to belong to the mutator type. Group 5 contains the DDE superfamily IS4 transposases and has hits to *M. acetivorans* (1), *M. barkeri* (2) and *M. mazei* (1). Group 6 was not detected by phylogenomic profiling and was found in *M. acetivorans* (2) and in genomic fragments from deep-sea sediments (GZfos34A6). Triangles indicate transposases identified during proteomic analysis to be expressed in the cell. Best matches to *M. burtonii* transposases were found by performing BLAST against a customized version of the NCBI nr database from which all *M. burtonii* proteins had been eliminated. The sequence alignments are in Figure S5.

Figure 5. Methanogenesis and biomass production in *M. burtonii.* Colour coding indicates proteins involved in methanogenesis from methanol (red), monomethylamine (MMA; green), dimethylamine (DMA; blue) and trimethylamine (TMA; purple), respectively. Abbreviations are as follows: CH_3OH, methanol; $(CH_3)_3N$, TMA; $(CH_3)_2NH$, DMA; CH_3NH_2, MMA; CH_3-CP, methyl-corrinoid protein; CoM, coenzyme M; CoB, coenzyme B; MePh, oxidized methanophenazine; $MePhH_2$, reduced methanophenazine; F_{420}, coenzyme F_{420}; $F_{420}H2$, reduced coenzyme F_{420}; CH_3-H_4SPT, methyl-tetrahydrosarcinapterin; CH_2=H_4SPT, methylene-tetrahydrosarcinapterin; CH≡H_4SPT, methenyl-tetrahydrosarcinapterin; CHO-H_4SPT, formyl-tetrahydrosarcinapterin; CHO-MFR, formyl-methanofuran; Fd, ferredoxin; MtaB and MtaC, methanol methyltransferase and corrinoid protein ; MttB and MttC, TMA methyltransferase and corrinoid protein; MtbB and MtbC, DMA methyltransferase and corrinoid protein; MtmB and MtmC, MMA methyltransferase and corrinoid protein; Mcr, methyl-CoM reductase; MtaA, methanol:CoM methylase; MtbA, methylamine:CoM methylase; Mcr, methyl-CoM reductase; Hdr, heterodisulfide reductase; Fpo, $F_{420}H_2$ dehydrogenase; Mtr, methyl-H_4SPT:CoM methyltransferase; Mer, methylene-H_4SPT reductase; Mtd, methylene-H_4SPT dehydrogenase; Mch, methenyl-H_4SPT cyclohydrolase; Ftr, formyl-methanofuran: H_4SPT formyltransferase; Fmd or Fwd, formyl-methanofuran dehydrogenase; CODH/ACS, carbon monoxide dehydrogenase/acetyl-CoA synthase.

Figure 6. Lipid biosynthesis in *M. burtonii.* Figure based on analysis of the closed genome sequence (including new gene numbers), incorporating information from previous studies (Nichols et al. 2004; Grochowski et al. 2006; Boucher et al. 2007). Abbreviations are as follows: CoA, coenzyme A; HMG-CoA, 3-hydroxy-3-methylglutaryl-CoA; P, phosphate; IPP, isopentenyl diphosphate; DMAPP, dimethylallyl diphosphate; GPP, geranyl diphosphate; FPP, farnesyl diphosphate; GGPP, geranylgeranyl diphosphate; GGGP, geranylgeranylglyceryl phosphate; DGGGP, digeranylgeranylglyceryl phosphate; CDP, cytidine diphosphoglycerol; G-1-P, glycerol-1-phosphate; DHAP, dihydroxyacetone phosphate.

Supplementary Information

Supplementary Information: manual annotation

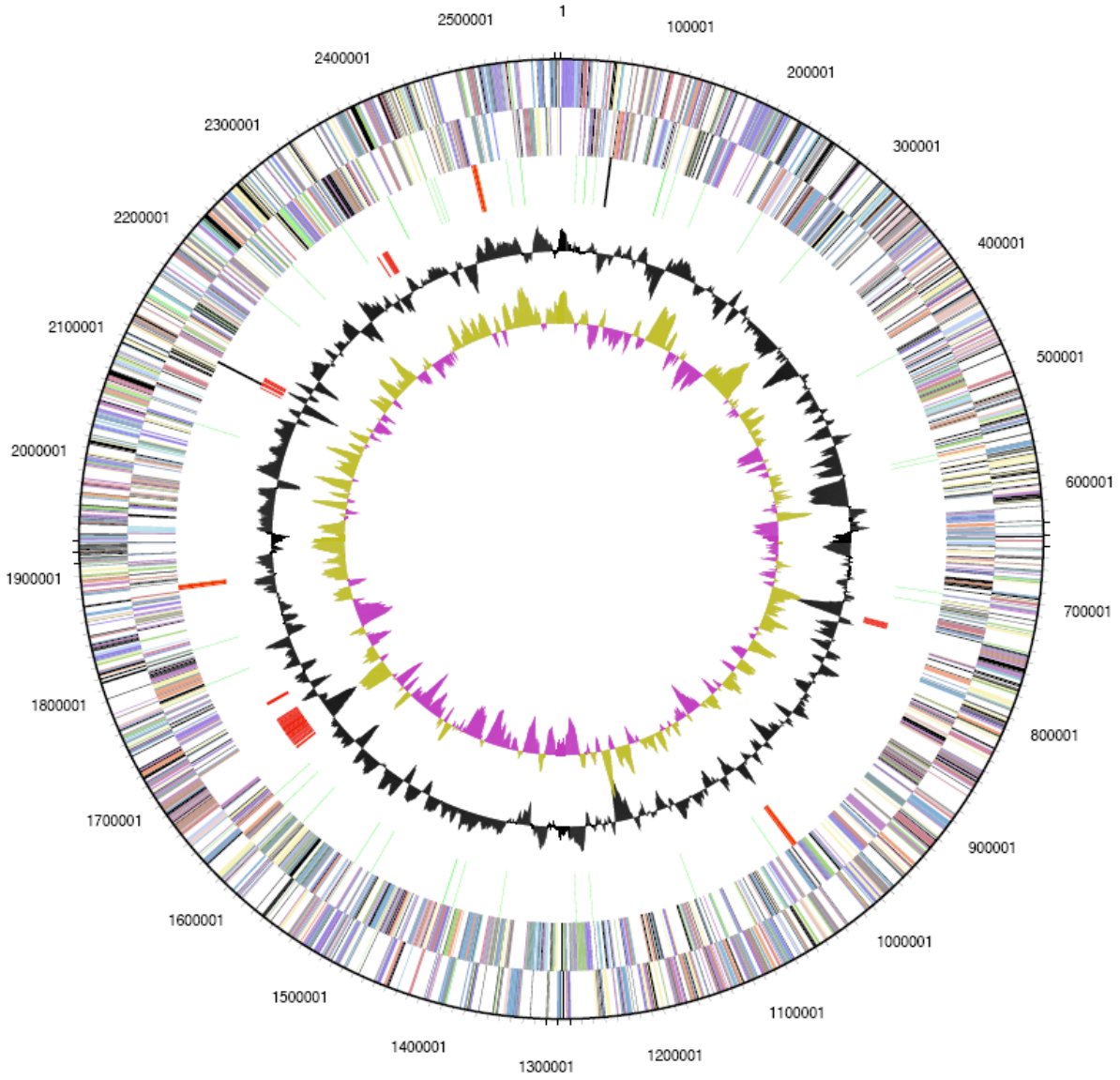

1
100001
200001
300001
400001
500001
600001
700001
800001
900001
1000001
1100001
1200001
1300001
1400001
1500001
1600001
1700001
1800001
1900001
2000001
2100001
2200001
2300001
2400001
2500001

Figure 1.

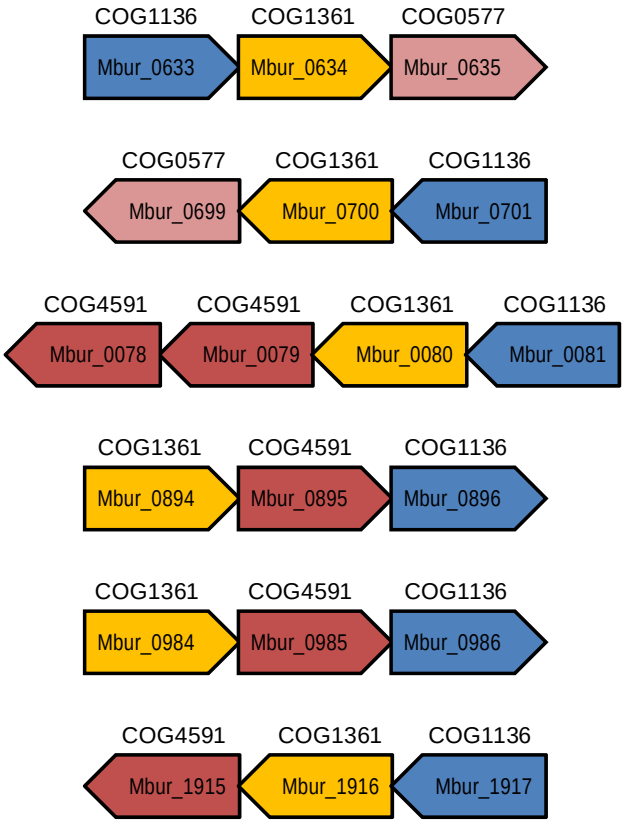

COG1136
COG1361
COG0577
Mbur_0633
Mbur_0634
Mbur_0635
COG0577
COG1361
COG1136
Mbur_0699
Mbur_0700
Mbur_0701
COG4591
COG4591
COG1361
COG1136
Mbur_0078
Mbur_0079
Mbur_0080
Mbur_0081
COG1361
COG4591
COG1136
Mbur_0894
Mbur_0895
Mbur_0896
COG1361
COG4591
COG1136
Mbur_0984
Mbur_0985
Mbur_0986
COG4591
COG1361
COG1136
Mbur_1915
Mbur_1916
Mbur_1917

Figure 2a.

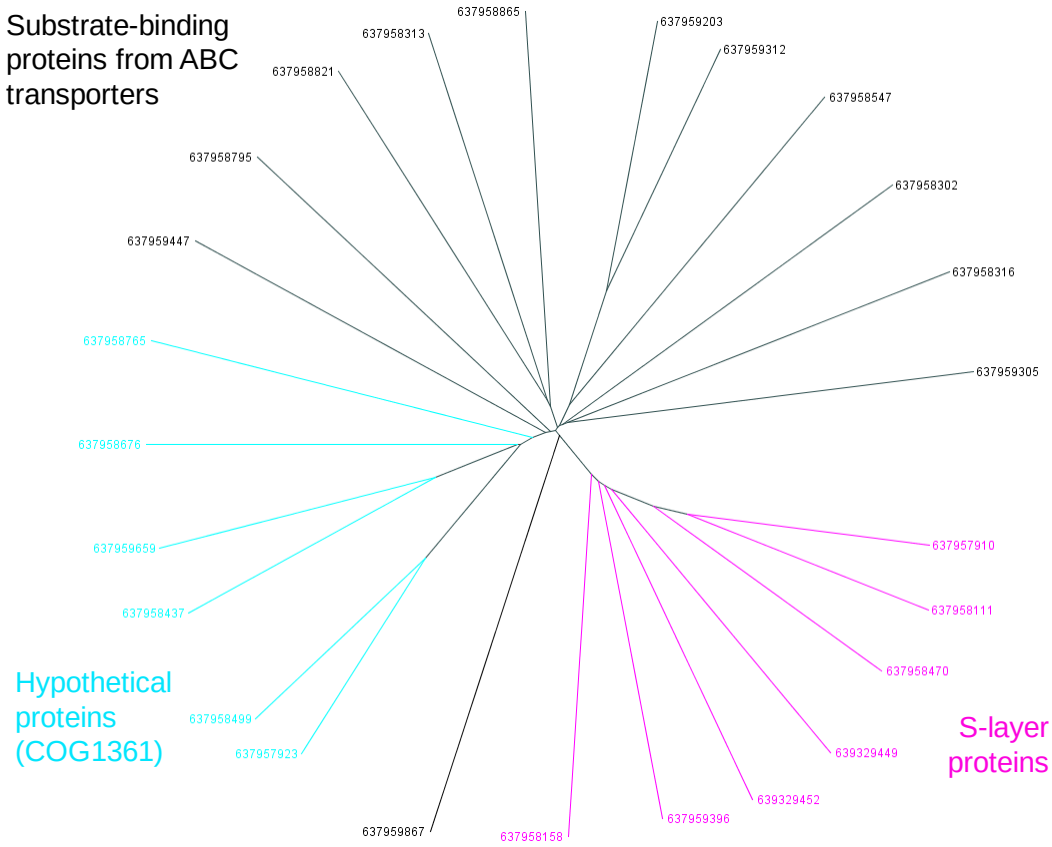

Substrate-binding proteins from ABC transporters
637958865
637958313
637959203
637959312
637958821
637958547
637958795
637958302
637959447
637958316
637958765
637959305
637958676
637959659
637957910
637958437
637958111
637958470
Hypothetical proteins (COG1361)
637958499
637957923
639329449
S-layer proteins
639329452
637959396
637959867
637958158

Figure 2b.

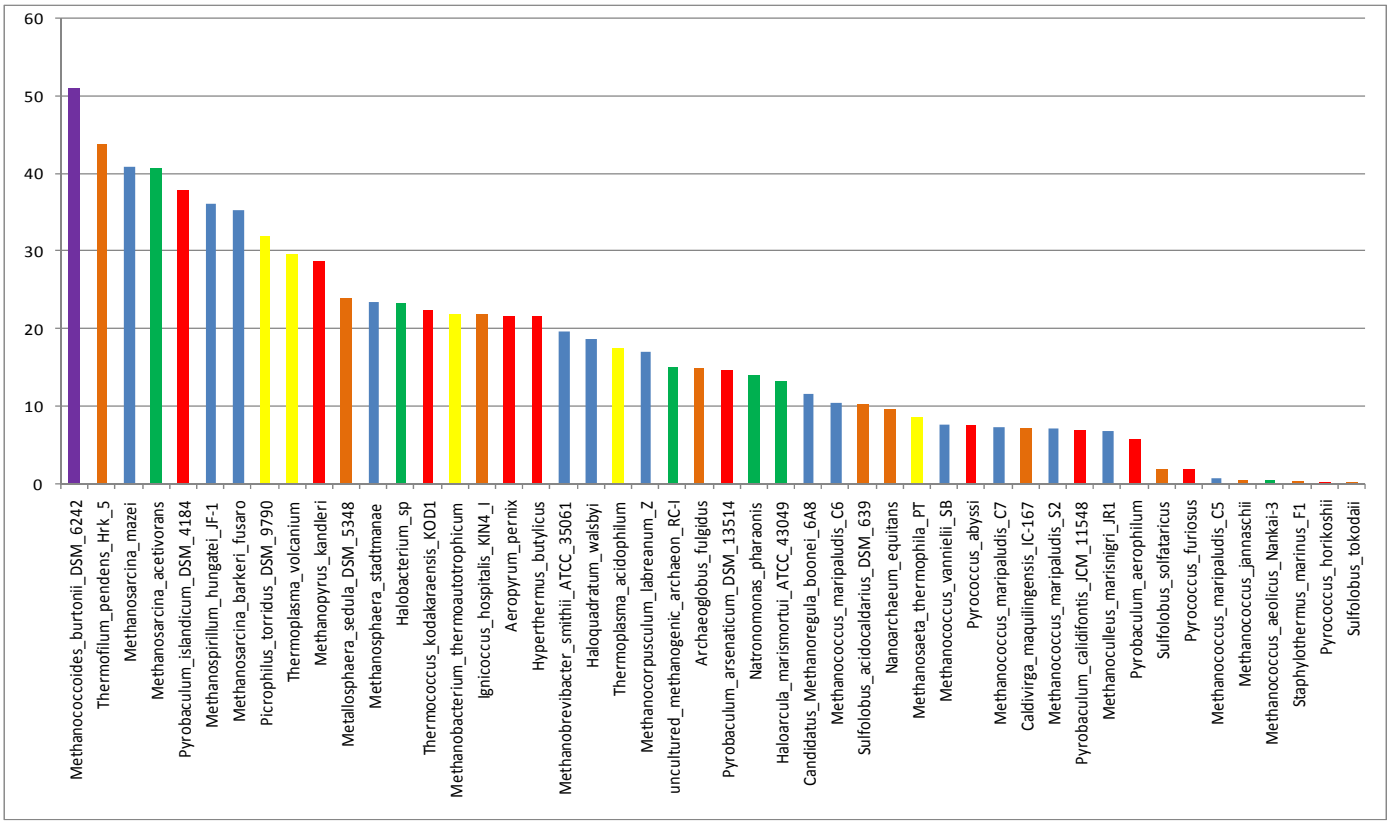

60
50
40
30
20
10
0
Methanococcoides_burtonii_DSM_6242
Thermofilum_pendens_Hrk_5
Methanosarcina_mazei
Methanosarcina_acetivorans
Pyrobaculum_islandicum_DSM_4184
Methanospirillum_hungatei_JF-1
Methanosarcina_barkeri_fusaro
Picrophilus_torridus_DSM_9790
Thermoplasma_volcanium
Methanopyrus_kandleri
Metallosphaera_sedula_DSM_5348
Methanosphaera_stadtmanae
Halobacterium_sp
Thermococcus_kodakaraensis_KOD1
Methanobacterium_thermoautotrophicum
Ignicoccus_hospitalis_KIN4_I
Aeropyrum_pernix
Hyperthermus_butylicus
Methanobrevibacter_smithii_ATCC_35061
Haloquadratum_walsbyi
Thermoplasma_acidophilum
Methanocorpusculum_labreanum_Z
uncultured_methanogenic_archaeon_RC-I
Archaeoglobus_fulgidus
Pyrobaculum_arsenaticum_DSM_13514
Natronomonas_pharaonis
Haloarcula_marismortui_ATCC_43049
Candidatus_Methanoregula_boonei_6A8
Methanococcus_maripaludis_C6
Sulfolobus_acidocaldarius_DSM_639
Nanoarchaeum_equitans
Methanosaeta_thermophila_PT
Methanococcus_vannielii_SB
Pyrococcus_abyssi
Methanococcus_maripaludis_C7
Caldivirga_maquilingensis_IC-167
Methanococcus_maripaludis_S2
Pyrobaculum_calidifontis_JCM_11548
Methanoculleus_marisnigri_JR1
Pyrobaculum_aerophilum
Sulfolobus_solfataricus
Pyrococcus_furiosus
Methanococcus_maripaludis_C5
Methanococcus_jannaschii
Methanococcus_aeolicus_Nankai-3
Staphylothermus_marinus_F1
Pyrococcus_horikoshii
Sulfolobus_tokodaii

Figure 3a.

Figure 3b.

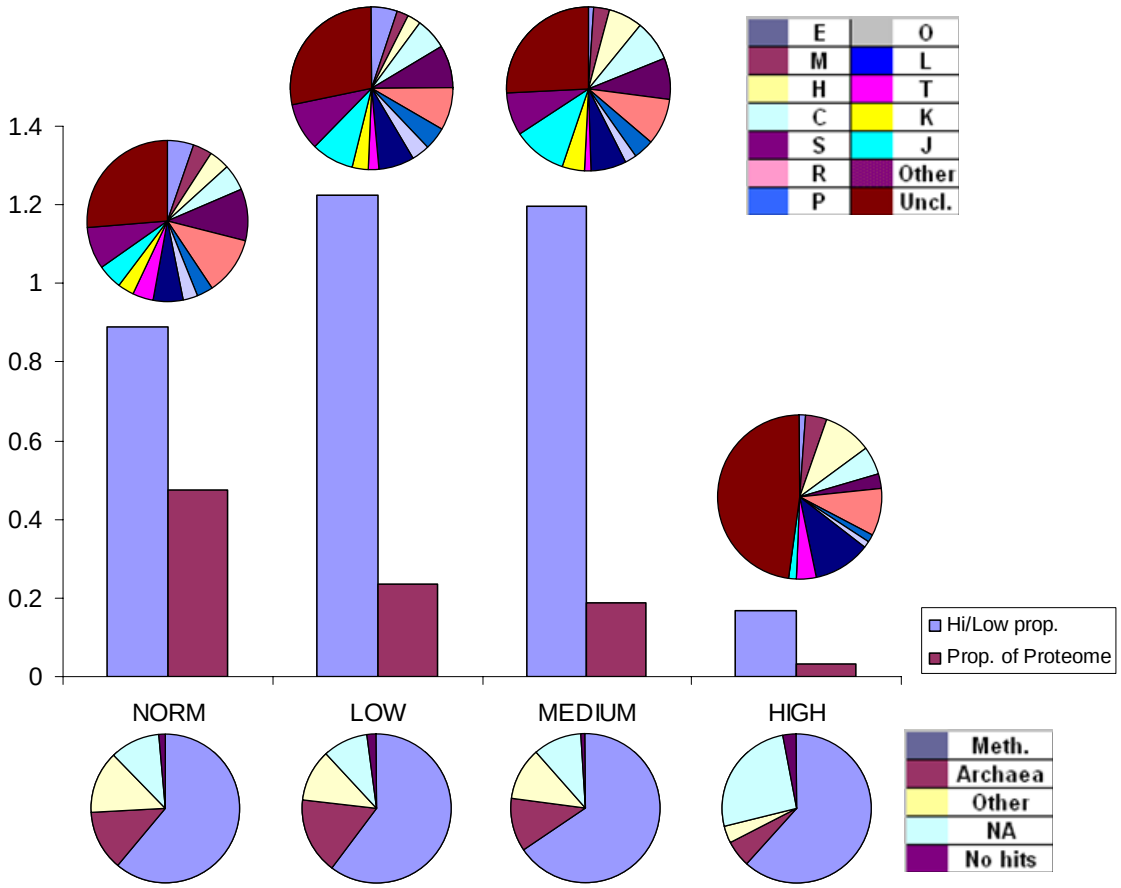

1.4
1.2
1
0.8
0.6
0.4
0.2
0
NORM
LOW
MEDIUM
HIGH
E
O
M
L
H
T
C
K
S
J
R
Other
P
Uncl.
Hi/Low prop.
Prop. of Proteome
Meth.
Archaea
Other
NA
No hits

Figure 3c.

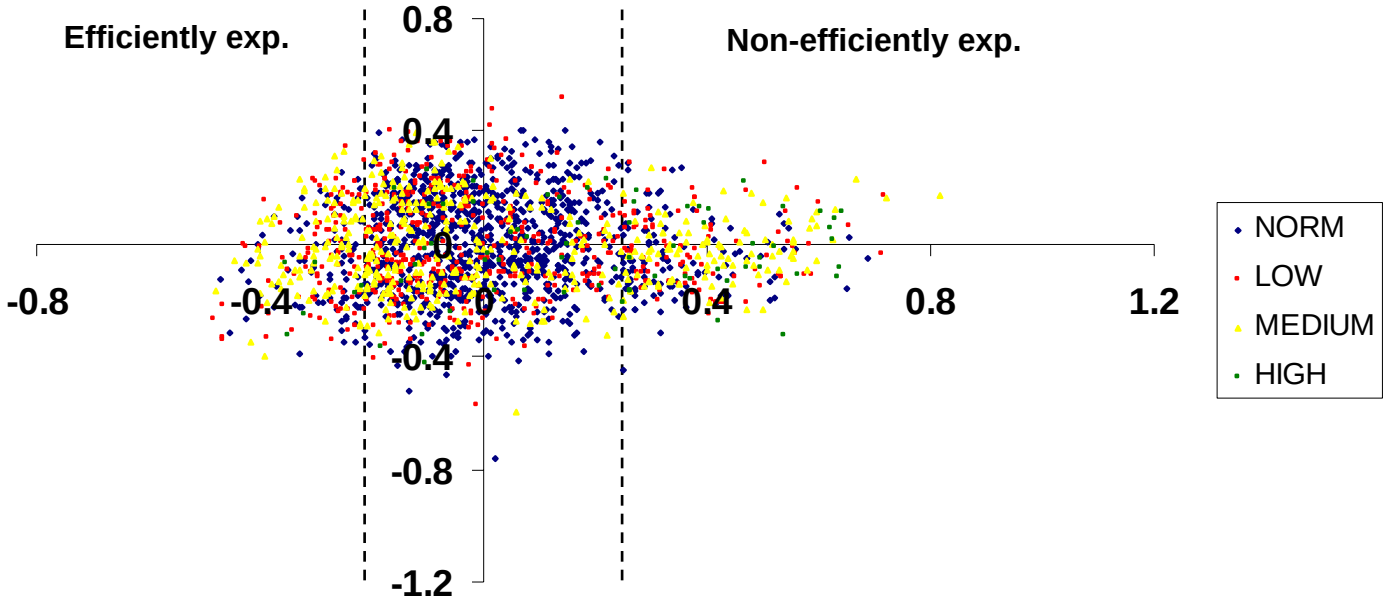

Efficiently exp.
Non-efficiently exp.
0.8
0.4
0
-0.4
-0.8
-1.2
-0.8
-0.4
0
0.4
0.8
1.2
NORM
LOW
MEDIUM
HIGH

Figure 3d.

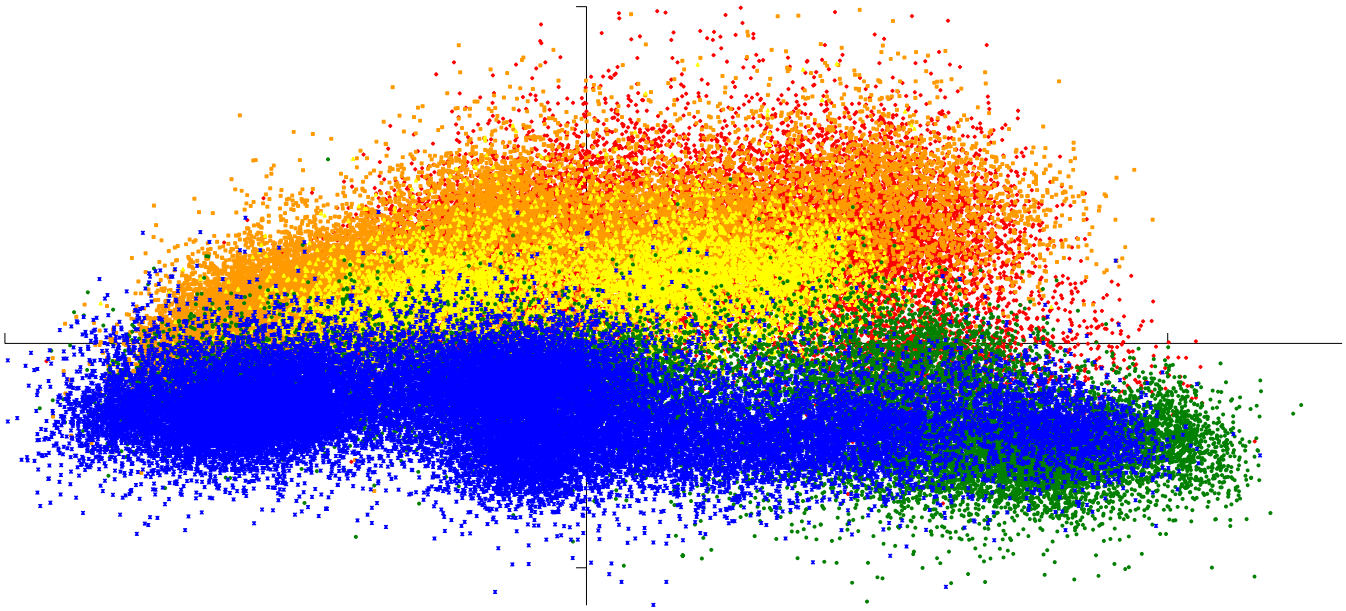
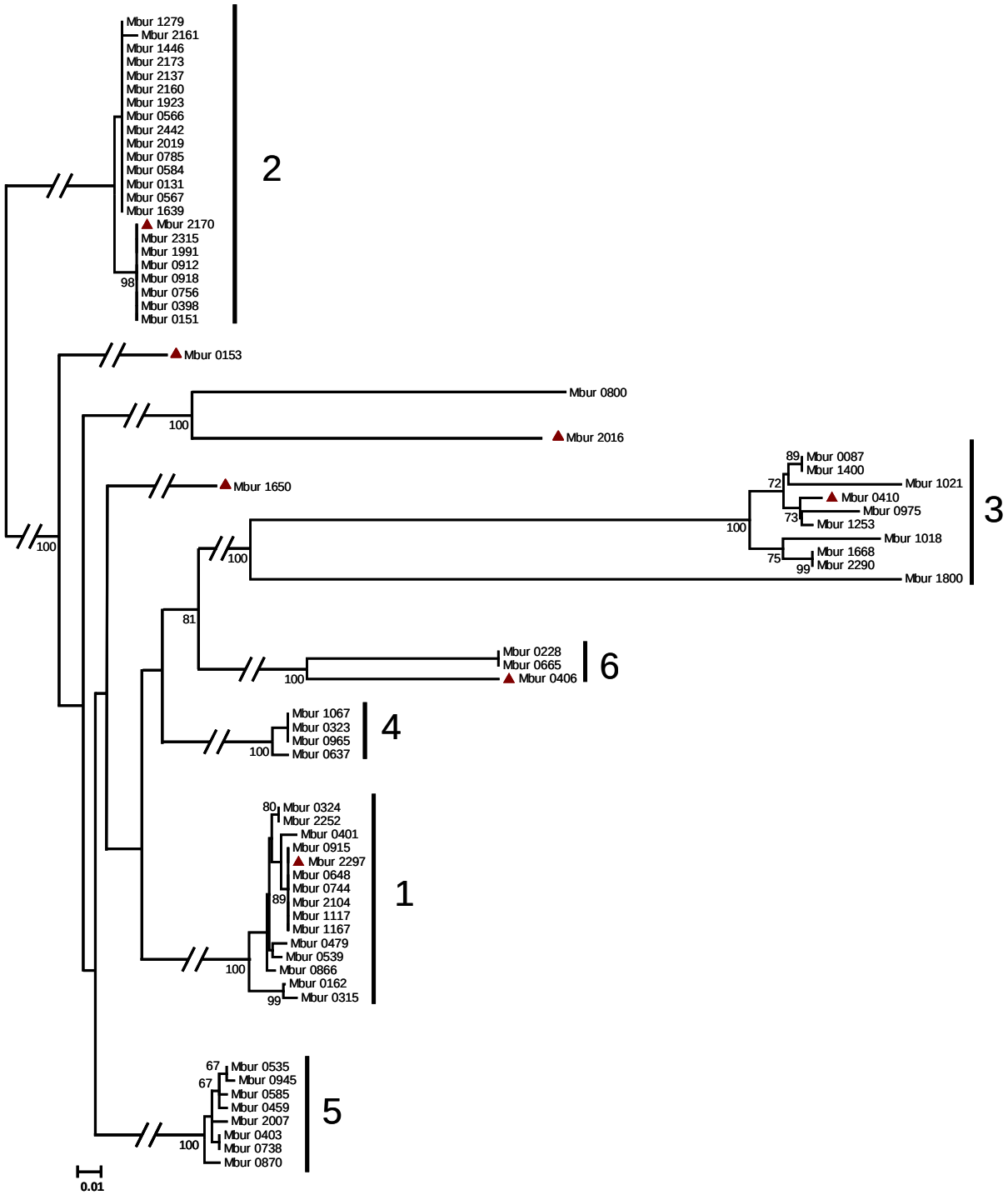


Figure 4.


Mbur 1279
Mbur 2161
Mbur 1446
Mbur 2173
Mbur 2137
Mbur 2160
Mbur 1923
Mbur 0566
Mbur 2442
Mbur 2019
Mbur 0785
Mbur 0584
Mbur 0131
Mbur 0567
Mbur 1639
Mbur 2170
Mbur 2315
Mbur 1991
Mbur 0912
Mbur 0918
Mbur 0756
Mbur 0398
Mbur 0151
98
2
Mbur 0153
Mbur 0800
100
Mbur 2016
Mbur 1650
100
89
Mbur 0087
Mbur 1400
72
Mbur 1021
Mbur 0410
73
Mbur 0975
Mbur 1253
100
Mbur 1018
75
Mbur 1668
99
Mbur 2290
Mbur 1800
3
100
81
Mbur 0228
Mbur 0665
Mbur 0406
100
6
Mbur 1067
Mbur 0323
Mbur 0965
Mbur 0637
100
4
80
Mbur 0324
Mbur 2252
Mbur 0401
Mbur 0915
Mbur 2297
Mbur 0648
Mbur 0744
89
Mbur 2104
Mbur 1117
Mbur 1167
Mbur 0479
Mbur 0539
100
Mbur 0866
Mbur 0162
99
Mbur 0315
1
67
Mbur 0535
67
Mbur 0945
Mbur 0585
Mbur 0459
Mbur 2007
100
Mbur 0403
Mbur 0738
Mbur 0870
5
0.01

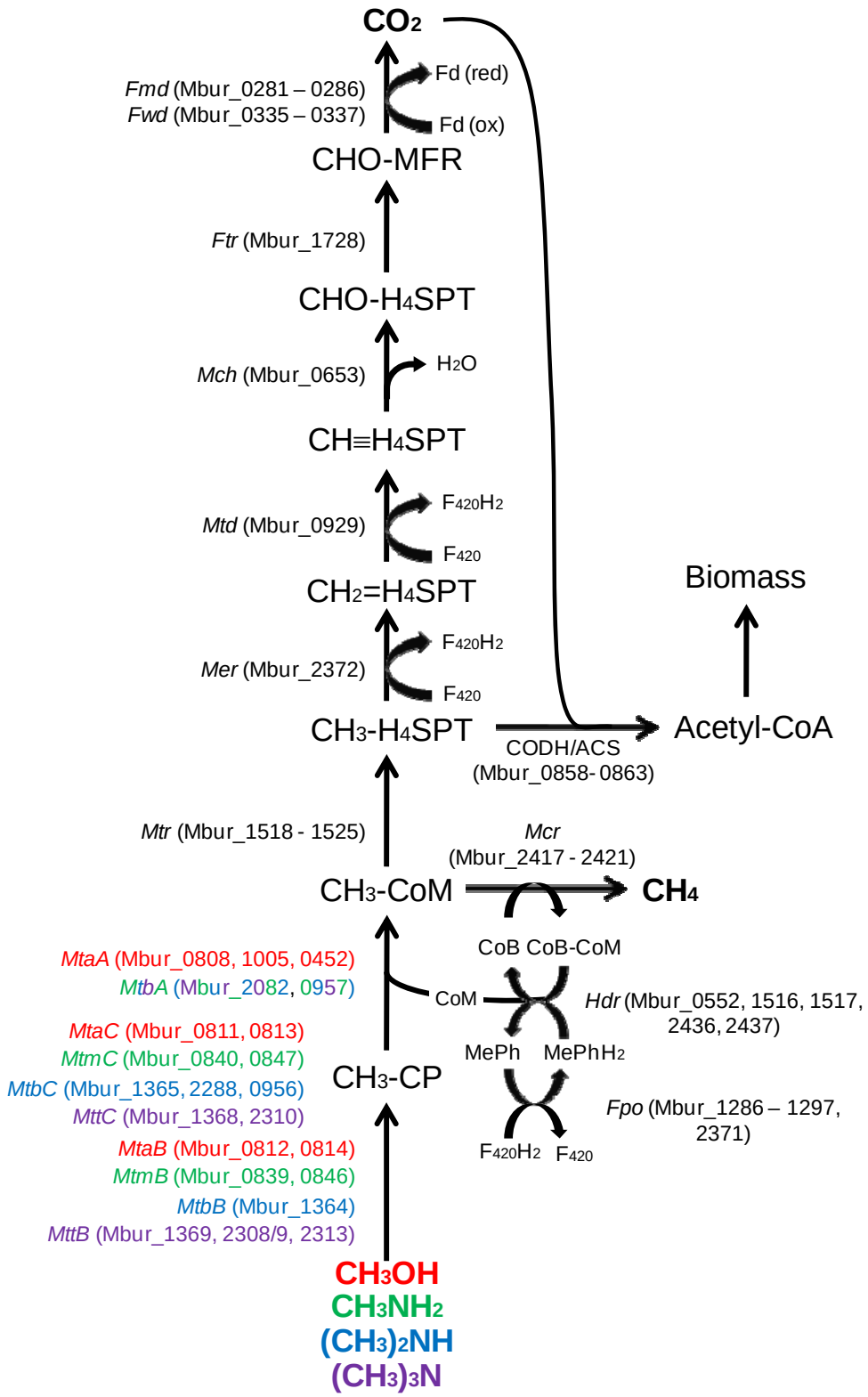

CO2
Fmd (Mbur_0281 – 0286)
Fwd (Mbur_0335 – 0337)
Fd (red)
Fd (ox)
CHO-MFR
Ftr (Mbur_1728)
CHO-H4SPT
Mch (Mbur_0653)
H2O
CH≡H4SPT
Mtd (Mbur_0929)
F420H2
F420
CH2=H4SPT
Mer (Mbur_2372)
F420H2
F420
CH3-H4SPT
CODH/ACS
(Mbur_0858- 0863)
Acetyl-CoA
Biomass
Mtr (Mbur_1518 - 1525)
Mcr
(Mbur_2417 - 2421)
CH3-CoM
CH4
CoB CoB-CoM
MtaA (Mbur_0808, 1005, 0452)
MtbA (Mbur_2082, 0957)
CoM
Hdr (Mbur_0552, 1516, 1517, 2436, 2437)
MtaC (Mbur_0811, 0813)
MtmC (Mbur_0840, 0847)
MtbC (Mbur_1365, 2288, 0956)
MttC (Mbur_1368, 2310)
CH3-CP
MePh
MePhH2
Fpo (Mbur_1286 – 1297, 2371)
F420H2
F420
MtaB (Mbur_0812, 0814)
MtmB (Mbur_0839, 0846)
MtbB (Mbur_1364)
MttB (Mbur_1369, 2308/9, 2313)
CH3OH
CH3NH2
(CH3)2NH
(CH3)3N

Figure 5.

Figure 6.

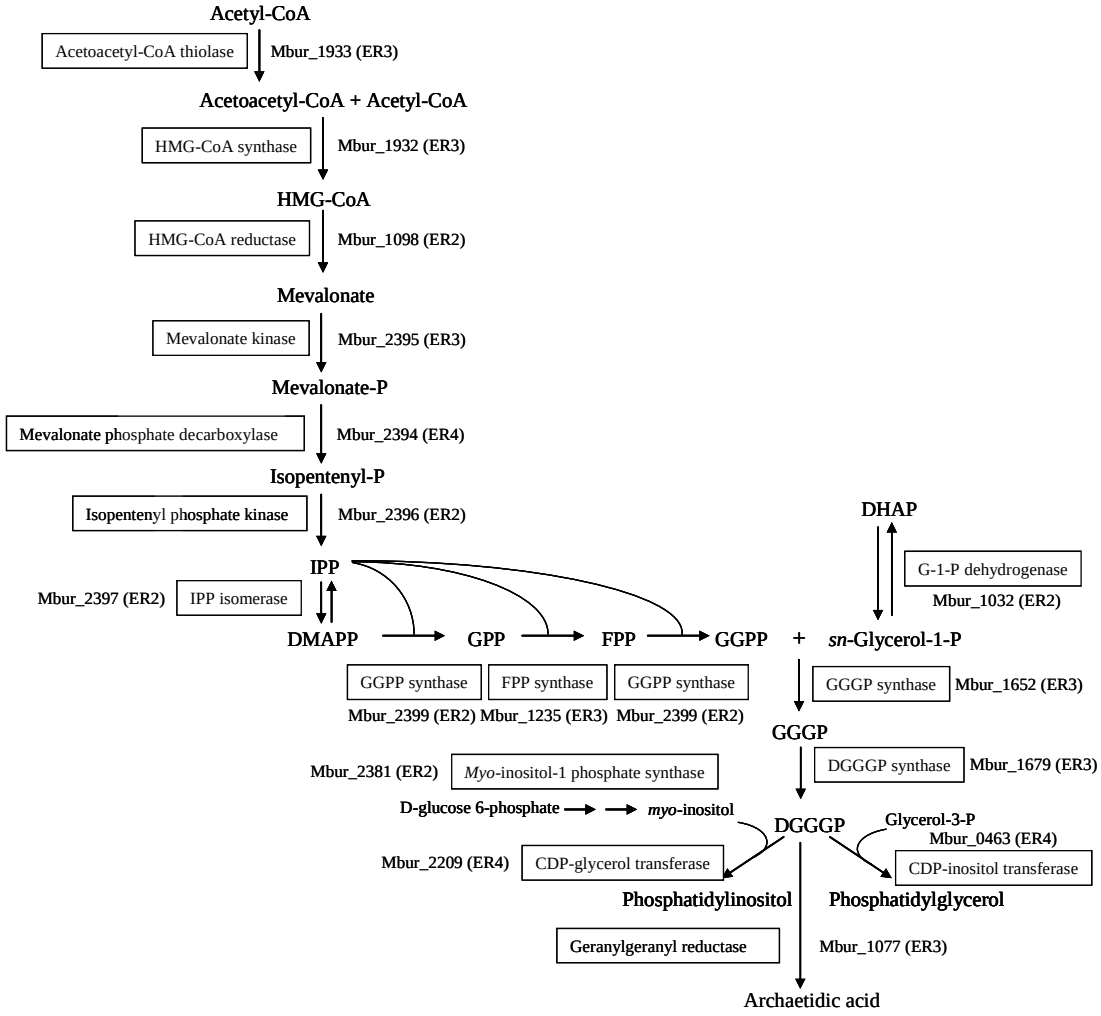

Acetyl-CoA
Acetoacetyl-CoA thiolase
Mbur_1933 (ER3)
Acetoacetyl-CoA + Acetyl-CoA
HMG-CoA synthase
Mbur_1932 (ER3)
HMG-CoA
HMG-CoA reductase
Mbur_1098 (ER2)
Mevalonate
Mevalonate kinase
Mbur_2395 (ER3)
Mevalonate-P
Mevalonate phosphate decarboxylase
Mbur_2394 (ER4)
Isopentenyl-P
Isopentenyl phosphate kinase
Mbur_2396 (ER2)
IPP
Mbur_2397 (ER2)
IPP isomerase
DMAPP
GPP
FPP
GGPP
+
sn-Glycerol-1-P
DHAP
G-1-P dehydrogenase
Mbur_1032 (ER2)
GGPP synthase
FPP synthase
GGPP synthase
Mbur_2399 (ER2)
Mbur_1235 (ER3)
Mbur_2399 (ER2)
GGGP synthase
Mbur_1652 (ER3)
GGGP
Mbur_2381 (ER2)
Myo-inositol-1 phosphate synthase
DGGGP synthase
Mbur_1679 (ER3)
D-glucose 6-phosphate
myo-inositol
DGGGP
Glycerol-3-P
Mbur_0463 (ER4)
Mbur_2209 (ER4)
CDP-glycerol transferase
CDP-inositol transferase
Phosphatidylinositol
Phosphatidylglycerol
Geranylgeranyl reductase
Mbur_1077 (ER3)
Archaetidic acid

Table 1. COG Categories with a Statistically Significant Difference in Abundance

	COG category	*M. burtonii* atypical ORFs[a] vs. whole genome	*Methanosarcina* spp. (3)	Comparison with Methanogens		Archaea	
				All (14)	Thermophilic (4)	All (39)	Thermophilic (25)
Over-represented in *M. burtonii*	Replication [L]		X	X	X	X	X
	Signal transduction [T]	X	X	X	X	X	X
	Translation [J]		X				
	Coenzyme metabolism [H]		X			X	X
	Cell wall [M]	X		X	X	X	X
	Intracellular trafficking [U]			X	X	X	X
	Motility [N]				X	X	X
	Function unknown [S]						X
	Defense mechanisms [V]				X		X
Under-represented in *M. burtonii*	General function only [R]		X	X	X	X	X
	Inorganic ion metabolism [P]		X	X			
	Defense mechanisms [V]		X				
	Function unknown [S]	X		X	X		
	Energy production [C]			X	X		X
	Lipid metabolism [I]					X	X
	Amino acid metabolism [E]					X	X
	Carbohydrate metabolism [G]					X	X
	Translation [J]	X			X		X
	Nucleotide metabolism [F]	X			X		X
	Secondary metabolites [Q]						X

X, denotes that the abundance of the COG was statistically significant at the 99% confidence level. Grey shaded boxes highlight differences related to temperature (all methanogens vs hyper/thermophilic methanogens, and all archaea vs hyper/thermophilic archaea)

[a] 500 proteins of *M. burtonii* that were not phylogenetic congruent with the domain *Archaea* (combined "other" and "NA" categories from Figure 4b)

Table 2. Clusters Formed from Phlogenetic Profiling using Genes from *M. burtonii*

Cluster ID	Proteins represented in clusters	No.	Phylogenetic distribution	Notes
Transposons group 1	Transposase (ER4)	15	*Halobacterium sp., M. acetivorans, M. mazei, M. hungatei*	
Transposons group 2	Transposase (ER3) with 1 (ER4)	24	*M. acetivorans, T. volcanium, T. acidophilum*	
Transposons group 3	Transposase (ER4)	10	*M. acetivorans, M. barkeri, M. mazei, S. tokodaii, S. solfataricus*	
Transposons group4	Transposase, mutator type (ER3)	4	*M. mazei, M. barkeri, S. solfataricus*	
Transposons group 5	IS4 Transposase, DDE superfamily (ER4)	8	*M. barkeri, M. mazei, M. acetivorans*	
Methanosarcina-like cluster	3 hypothetical proteins (ER5), 3 DUF1608 (ER4), 2 DUF1699 (ER4), metalloenzyme superfamily (ER4), Sodium/solute symporter (ER3)	10	*M. thermophila, M. acetivorans, M. barkeri, M. mazei*	DUF1608 (IPR06457) – linked to an InterProDomain associated with archaeal S-layers (IPR06454). DUF1699 – archaeal proteins with very conserved sequences.
Methanogen cluster	Methanogenesis pathway genes (21), also acetylglutamate kinase, ornithine acetyl transferase, nitrogenase and associated proteins, many ER4 DUF proteins, one hypothetical protein (ER5)	46	All methanogens, *A. fulgidus*	
M. mazei-specific	8 hypothetical proteins, 3 proteins with DNA-directed DNA-polymerase B domains (ER4), sodium/glutamate symporter (ER3), bacterial Ig-like domain protein (ER4)	13	*M. mazei*	6 of the genes (3 hypothetical, 3 DNA-polymerase B domain) are in a region of novel hypothetical genes and have possibly duplicated/moved by transposase activity.
Methanosarcina-specific	29 hypothetical proteins (ER5), pyrrolysyl-tRNA synthetase (ER2), tRNA-dihydrouridine synthase (ER2), fructose-1,6-bisphosphatase (ER2), restriction endonuclease (ER4), N6 DNA methyylase	33	*M. acetivorans, M. barkeri, M. mazei*	

	(ER2), nicotinate phosphoribosyl transferase (ER2), 8 DUF or other domain proteins (ER4)			
Largest cluster	Several ribosomal proteins, transposases, 3 thermosome subunit (chaperonin), GTP phosphohydrolase, translation elongation factor, RNase L inhibitor, a few DUF or other domain ER4 proteins, but largely all hypotheticals (89)	106	Fairly even across all archaea, moderate peak at *Nanoarchaeum equitans*	
Signal transduction	Signal transduction histidine kinases (14 were ER4, 4 were ER3)	18	*M. aeolicus, M. maripaludis C5, M. maripaludis C7, M. maripaludis S2, M. vannielii, M. marisnigri, M. thermophila, M. acetivorans, M. mazei, M. barkeri, M. hugatei, N. pharaonis, M. thermoautotrophicum, A. fulgidus, H. marismortui, Halobacterium NRC-1, H. walsbyi*	
Chemotaxis	Chemotaxis proteins CheW, CheY, CheB, CheA, CheD and CheR	6	*A. fulgidus, H. marismortui, Halobacterium NRC-1, M. maripaludis C5 C7 and S2, M. vannielii, M. marisnigri, M. acetivorans, M. mazei, M. hungatei, N. pharaonis, P. abyssi, P. horikoshii, T. kodakaraensis*	Form an operon Mbur_0357 to Mbur_0364
Methanogenesis	Mono-, di- and tri-methylamine corrinoid proteins	6	*M. smithii, M. aeolicus, M. maripaludis C5 C7 and S2, M. acetivorans, M. barkeri, M. mazei, M. stadtmanae, S. marinus, T. pendens*	*S. marinus* and *T. pendens* in this grouping of methanogenesis related proteins.
Methanogenesis 2	2 methanol corrinoid proteins and two conserved methanogen proteins with ferredoxin domains (one, Mbur_0809, found very close to the methanol corrinoid	4	*M. smithii, M. acetivorans, M. barkeri, M. mazei, M. stadtmanae*	

	proteins Mbur_0812 and Mbur_0814)			
Methanogenesis 3	3 methanol-specific methyltransferase mtaA and one methylamine-specific methyltransferase mtbA	4	*M. smithii,M. maripaludis C5 C7 and S2, M. acetivorans, M. barkeri, M. mazei, M. stadtmanae*	One *mtbA* gene not clustering with these (Mbur_0957) may have a different phylogenetic occurrence.
CRISPR region	7 CRISPR-associated proteins (ER4) and one chaperone (ER2)	8	*M. hungatei*	Appears to have come directly from *M. hungatei*
Hypotheticals	Hypothetical protein (ER5)	9	*M. acetivorans*	
Viruses	RNA-directed DNA polymerase (reverse transcriptase) (ER4)	4	*M. acetivorans, M. barkeri, M. mazei, M. hungatei*	Possibly replicating RNA from prophages.
Serpins	Serine protease inhibitor (ER3)	3	*M. marisnigri, M. acetivorans, M. mazei, P. aerophilum, P. arsenaticum, P. calidifontis, P. islandicum, T. kodakaraensis*	
ER2, ER3 and ER4 all clustered together	O-phosphoseryl-tRNA (Cys) synthetase (ER2), Sep-tRNA:Cys-tRNA synthase (ER3) and CBS-domain and DUF39-domain containing protein (ER4)	3	*A. fulgidus, M. thermoautotrophicum, M. aeolicus, M. jannaschii, M. maripaludis C5 C7 and S2, M. vannielii, M. labreanum, M. marisnigri, M. kandleri, M. thermophila, M. acetivorans, M. barkeri, M. mazei, M. hungatei* (all methanogens except *M. smithii* and *M. stadtmanae*)	Linkage of a ER4 protein with ER2 and ER3 proteins may infer a common function. These three proteins are linked by phyloprofile in each *Methanosarcina* genome.
ER2 and ER4 proteins	DNA gyrase subunits A and B (ER2) and a protein of unknown function DUF1119 (ER4)	3	*A. fulgidus, H. marismortui, Halobacterium NRC-1, H. walsbyi, M. labreanum, M. marisnigri, M. thermophila, M. acetivorans, M. barkeri, M. mazei, M. hungatei, N. pharaonis, P. torridus, T. acidophilum, T. volcanium*	DUF protein is possibly associated with DNA-gyrase activity. The proteins are spaced out across the genome for *M. burtonii,* Methanosarcinales and for *M. hungatei*. All three proteins are expressed in *M. burtonii*. The DUF1119 domain is a child of IPR006639::Peptidase A22, presenilin signal peptide, suggesting it may have a peptidase function.

Table 3. Genes Interrupted by Transposases

Gene number, function and ER	**Transposon**
Mbur_0914[a]; histidine kinase/PAS domain; ER4	Mbur_0915 (Group 1)
Mbur_1066; NADPH-dependent FMN reductase; ER4	Mbur_1067 (Group 4)
Mbur_1638[a]; hypothetical; ER5	Mbur_1639 (Group 2)
Mbur_1664[a]; hypothetical; ER5	Mbur_1668 (Group 3)
Mbur_2136[a]; histidine kinase/PAS domain; ER4	Mbur_2137 (Group 2)
Mbur_2172; hypothetical; ER5	Mbur_2173 (Group 2)
Mbur_2291[a]; dimethylamine methyltransferase	Mbur_2290 (Group 3)
[a] detected by mass spectrometry in proteomics study to be expressed in the cell	

SUPPLEMENTARY INFORMATION

Table S1. Islands with Atypical Nucleotide Composition (IVOM scores) Detected by Alien Hunter[1]

Organism	IVOM islands	IVOM ORFs	Non-IVOM ORFs	Border ORFs	Total island bp	Average island bp	% of genome
M. barkeri	198	1476	1931	199	1723283	8703	35.36
M. acetivorans	202	1553	2758	229	2346310	11615	40.79
M. mazei	155	1214	1979	177	1675167	10808	40.89
M. burtonii	125	1042	1097	164	1313583	10509	51.01

[1] Vernikos GS, Parkhill J (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. Bioinformatics 22: 2196-2203.

Table S2. Unique Hypothetical Proteins (ER 5) in the *M. burtonii* Genome

Locus Tag	Expressed	Gene context[a]	Length (aa)
Mbur_0103	yes	Near other unique hypothetical proteins	316
Mbur_0105			262
Mbur_0106			120
Mbur_0129			171
Mbur_0140	yes	Downstream of the sodium/hydrogen antiporter gene cluster	81
Mbur_0173			294
Mbur_0236			86
Mbur_0237			227
Mbur_0241			108
Mbur_0275			103
Mbur_0276			110
Mbur_0277			124
Mbur_0278			230
Mbur_0287			95
Mbur_0293			130
Mbur_0343	yes	Near Flagellin/ flagella gene cluster	334
Mbur_0366			192
Mbur_0383			129
Mbur_0385			137
Mbur_0402			97
Mbur_0405			92
Mbur_0440	yes	Near amylase and hydrolase enzyme cluster, facing the opposite direction to Mbur_0441	117
Mbur_0445			246
Mbur_0492			183
Mbur_0494			89
Mbur_0495			91
Mbur_0542			79
Mbur_0551			96
Mbur_0554			107
Mbur_0556	yes	Near 3 DNA directed DNA polymerase B subunits	165
Mbur_0559			234
Mbur_0574			167
Mbur_0576	yes	Near 3 DNA directed DNA polymerase B subunits	429
Mbur_0640			100
Mbur_0641	yes	Adjacent to other unique hypothetical proteins	85
Mbur_0642			79
Mbur_0643			128
Mbur_0668			308
Mbur_0670			90
Mbur_0672			194
Mbur_0691			102

Mbur_0713	yes	Nothing discernable	178
Mbur_0735			176
Mbur_0764			96
Mbur_0789			145
Mbur_0794			87
Mbur_0823			375
Mbur_0928			74
Mbur_0930			579
Mbur_0931			152
Mbur_0988			201
Mbur_0995			138
Mbur_0996			160
Mbur_1071			78
Mbur_1072			157
Mbur_1097			91
Mbur_1099			189
Mbur_1108			177
Mbur_1122			625
Mbur_1146			160
Mbur_1161	yes	Near RNA polymerase cluster	175
Mbur_1189			196
Mbur_1197			87
Mbur_1210			146
Mbur_1223			127
Mbur_1248			79
Mbur_1251			107
Mbur_1298	yes	Adjacent to $F_{420}H_2$ dehydrogenase cluster	168
Mbur_1386	yes	In operon with electron transport complex Rnf	134
Mbur_1402			77
Mbur_1404			104
Mbur_1405			74
Mbur_1432	yes	Nothing discernable	134
Mbur_1484	yes	Downstream from the condensin operon	241
Mbur_1493	yes	Neighbouring, and encoded divergently, to the condensin operon	121
Mbur_1500	yes	Part of the condensin operon	258
Mbur_1507			89
Mbur_1530			185
Mbur_1544			176
Mbur_1636			75
Mbur_1637	yes	Near cell surface and filamentation proteins	103
Mbur_1661			116
Mbur_1708			84
Mbur_1720			130
Mbur_1721			128
Mbur_1726	yes	Nothing discernable	526
Mbur_1727			95

Mbur_1747			208
Mbur_1756	yes	Near a string of 5 CBS-domain proteins and a RecA ATPase	161
Mbur_1768			104
Mbur_1780			79
Mbur_1797			140
Mbur_1798			118
Mbur_1819			269
Mbur_1832			134
Mbur_1845			82
Mbur_1855			96
Mbur_1922			66
Mbur_1939			79
Mbur_1981	yes	Nothing discernable	405
Mbur_1996	yes	Nothing discernable	130
Mbur_2057	yes	In an operon with other hypothetical proteins and an ATPase	182
Mbur_2061			147
Mbur_2063	yes	Next to a DsbD-like protein (transfers electrons from periplasmic protein disulfide bond isomerase (DsbC) to a thioredoxin)	391
Mbur_2064	yes	Next to a DsbD-like protein (transfers electrons from periplasmic protein disulfide bond isomerase (DsbC) to a thioredoxin)	210
Mbur_2098	yes	Nothing discernable	126
Mbur_2103			236
Mbur_2187			91
Mbur_2211			81
Mbur_2213	yes	Nothing discernable	184
Mbur_2224	yes	Within a putative polysaccharide synthesis operon	188
Mbur_2226	yes	Within an operon with glycosyl transferases and a polysaccharide biosynthesis protein	662
Mbur_2235			95
Mbur_2295			92
Mbur_2314			117
Mbur_2369	yes	In between clusters for tryptophan biosynthesis and methanogenesis	151
Mbur_2443			93

[a] Conserved gene context analysis (Saunders et al 2005) was examined for genes that were known by proteomics analysis to be expressed in the cell

Saunders NFW, Goodchild A, Raftery M, Guilhaus M, Curmi PMG, et al. (2005) Predicted roles for hypothetical proteins in the low-temperature expressed proteome of the Antarctic archaeon *Methanoccoides burtonii*. J Proteome Res 4: 464-472.

Figure S1. Type-I restriction modification gene clusters in *M. burtonii.* COG0732, restriction endonuclease S subunits; COG0286, Type I restriction modification system methyltransferase (M subunit); COG0610, Type I site-specific restriction-modification system, R (restriction) subunit and related helicases; COG4096, Type I site-specific restriction-modification system, R (restriction) subunit and related helicases; COG1205, distinct helicase family with a unique C-terminal domain including a metal-binding cysteine cluster; COG2865, predicted transcriptional regulator containing an HTH domain and an uncharacterized domain shared with the mammalian protein Schlafen; IL, Interleukin; Fic, filamentation induced by cAMP. Unlabelled proteins are hypothetical proteins (ER5). Diagrammatic representation of gene arrangements (not to scale).

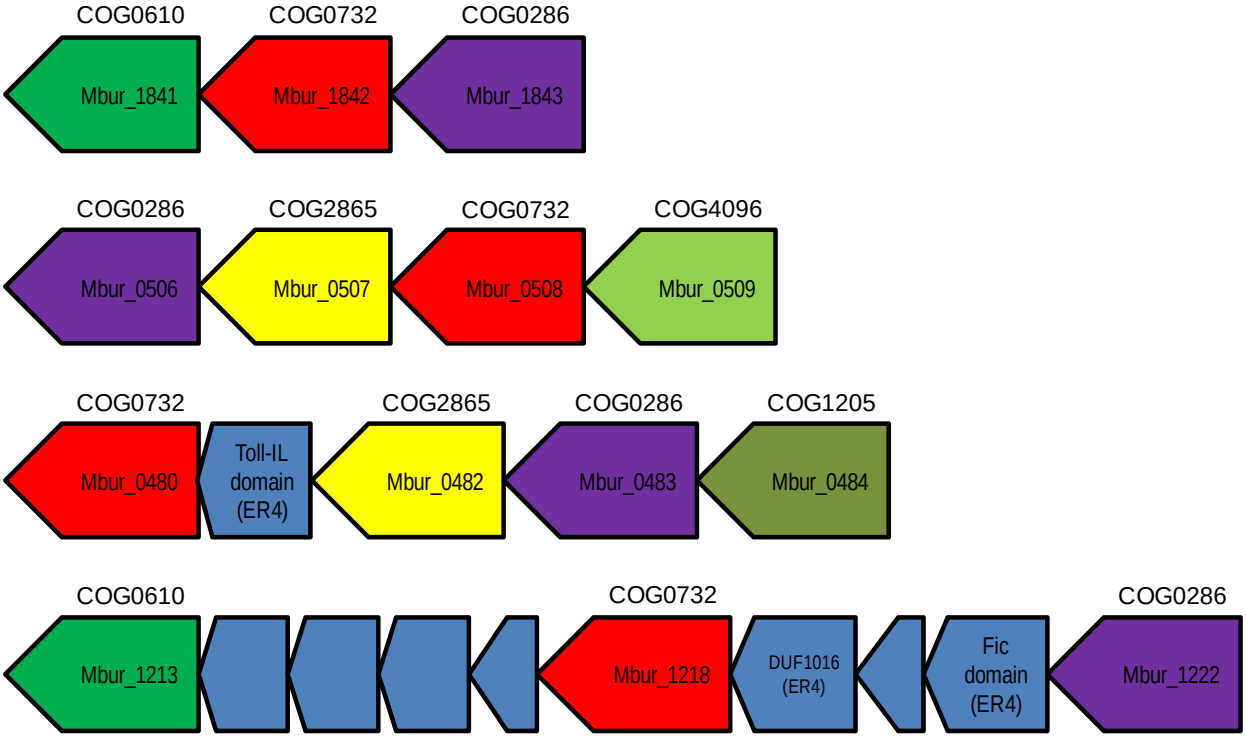
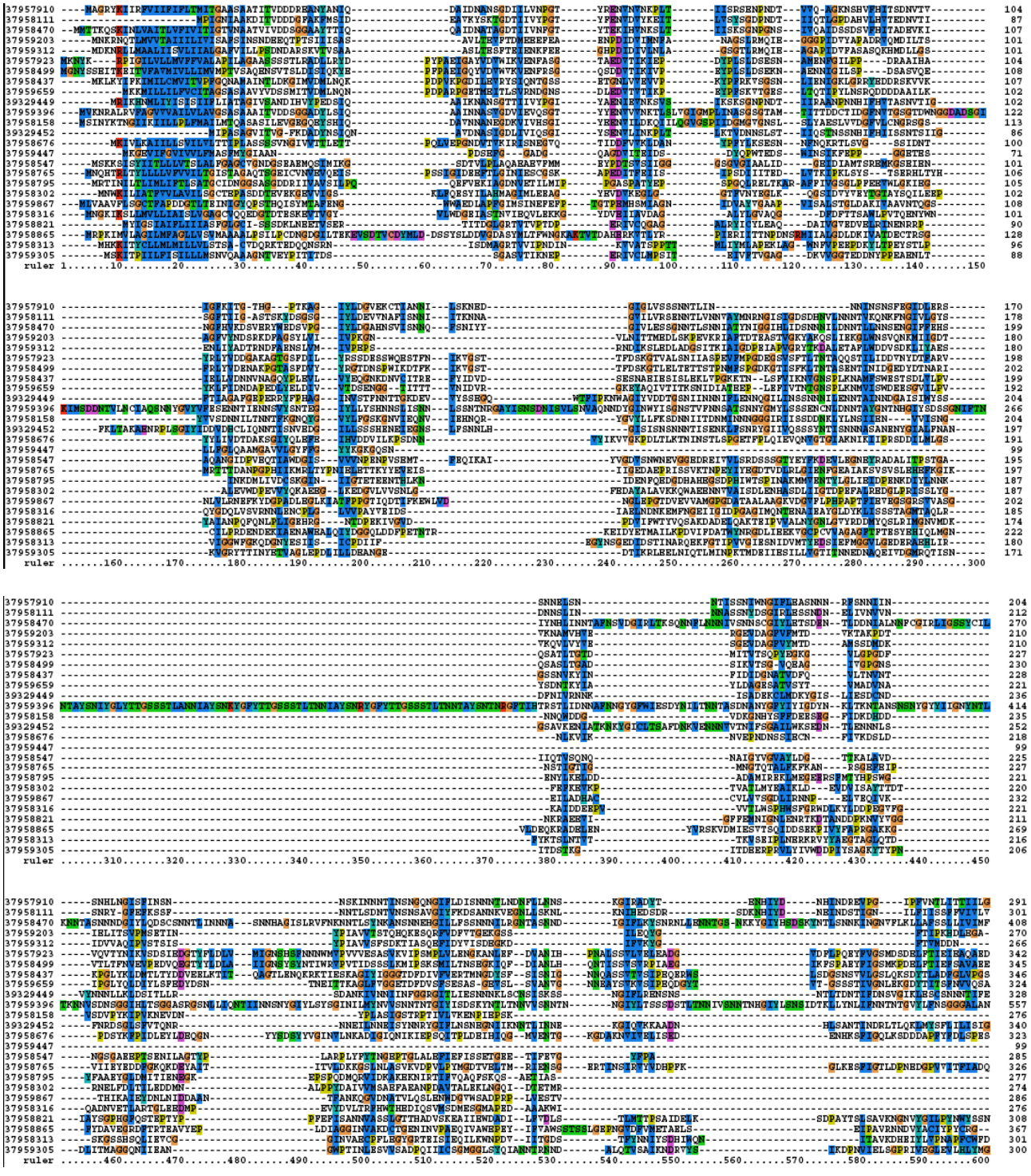


Figure S2. Alignment of ABC transporter sequences. Sequences used for Figure 2. Multiple alignment of the COG1361 hypothetical proteins (637958765, 637958676, 637959659, 637958437, 637958499 and 637957923), ABC-transporter substrate-binding proteins (637959867, 637959447, 637958795, 637958821, 637958313, 637958865, 637959203, 637959312, 637958567, 637958302, 637959305) and S-layer proteins (637958158, 637959396, 639329452, 639329449, 637958470, 637958111 and 637957910) from *M. burtonii* constructed using ClustalX 1.83 with default parameters and colour scheme (Thompson et al 1997). Sequences are denoted by their IMG Gene Object ID number. Note that S-layer protein 637959396 is 1200 aa long and was truncated at aa737 in the alignment. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 24: 4876-4882.



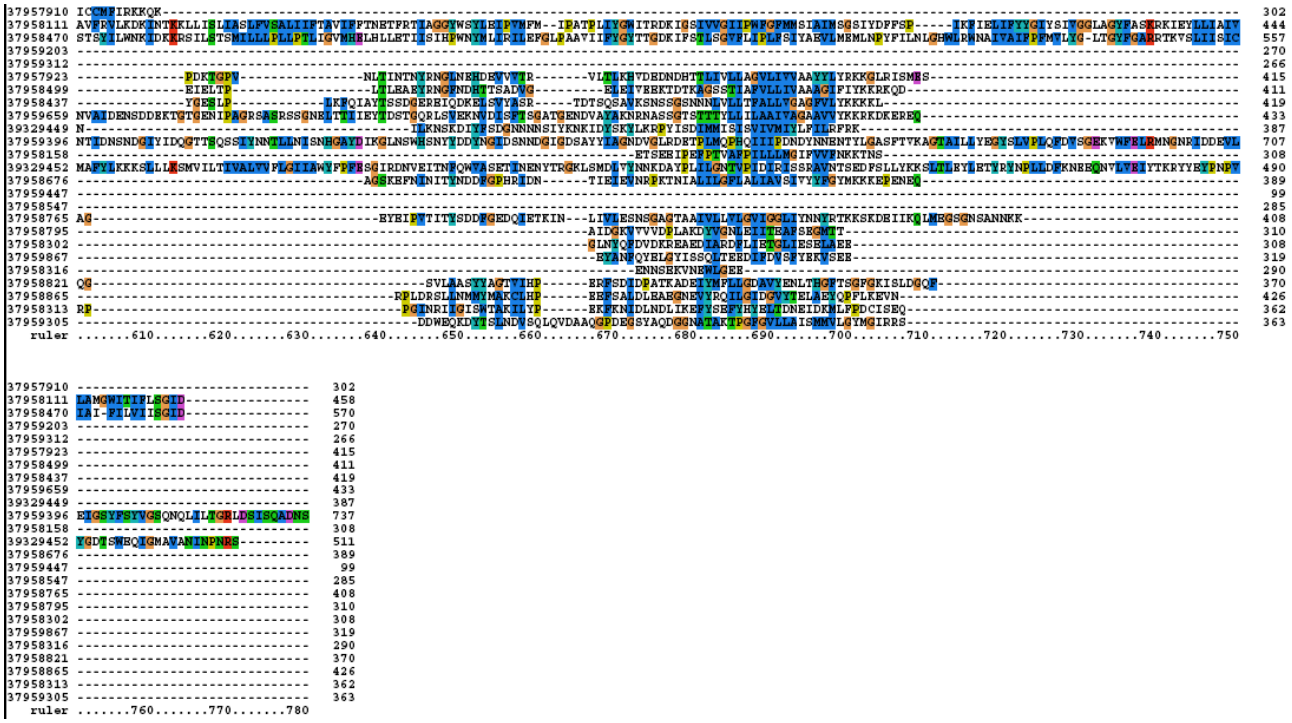


Figure S3. Principle Components Analysis of amino acid composition. (A) Component scores for organisms. (B) Component loadings for amino acids from archaeal proteomes. (C) Component scores for amino acids from non-efficiently expressed proteins of *M. burtonii*. (D) Component scores for amino acids from efficiently expressed proteins of *M. burtonii*. Ape, *Aeropyrum pernix*; Afu, *Archaeoglobus fulgidus*; Cma, *Caldivirga maquilingensis*; Hma, *Haloarcula marismortui*; Halo, *Halobacterium* sp. NRC-1; Hbut, *Hyperthermus butylicus*; Ihos, *Ignicoccus hospitalis*; Msm, *Methanobrevibacter smithii*; Mbur, *Methanococcoides burtonii*; Mae, *Methanococcus aeolicus*; Mjan, *Methanococcus jannaschii*; MmC5, *Methanococcus maripaludis* C5; MmC7, *Methanococcus maripaludis* C7; MmS2, *Methanococcus maripaludis* S2; Mvan, *Methanococcus vannielii*; Mmar, *Methanoculleus marisnigri*; Mfri, *Methanogenium frigidum*; Mkan, *Methanopyrus kandleri*; Mthe, *Methanosaeta thermophila*; Mac, *Methanosarcina acetivorans*; Mbar, *Methanosarcina barkeri*; Mmaz, *Methanosarcina mazei*; Nequ, *Nanoarchaeum equitans*; Npha, *Natromonas pharaonis*; Paer, *Pyrobaculum aerophilum*; Parse, *Pyrobaculum arsenaticum*; Pcali, *Pyrobaculum calidifontis*; Pisla, *Pyrobaculum islandicum*; Paby, *Pyrococcus abyssi*; Pfuri, *Pyrococcus furiosus*; Phori, *Pyrococcus horikoshii*; Smar, *Staphylothermus marinus*; Saci, *Sulfolobus acidocaldarius*; Ssolf, *Sulfolobus solfataricus*; Stoko, *Sulfolobus tokodaii;* Tkoda, *Thermococcus kodakaraensis*; Tpen, *Thermophilum pendens*; Taci, *Thermoplasma acidophilum*; Tvol, *Thermoplasma volcanium*.

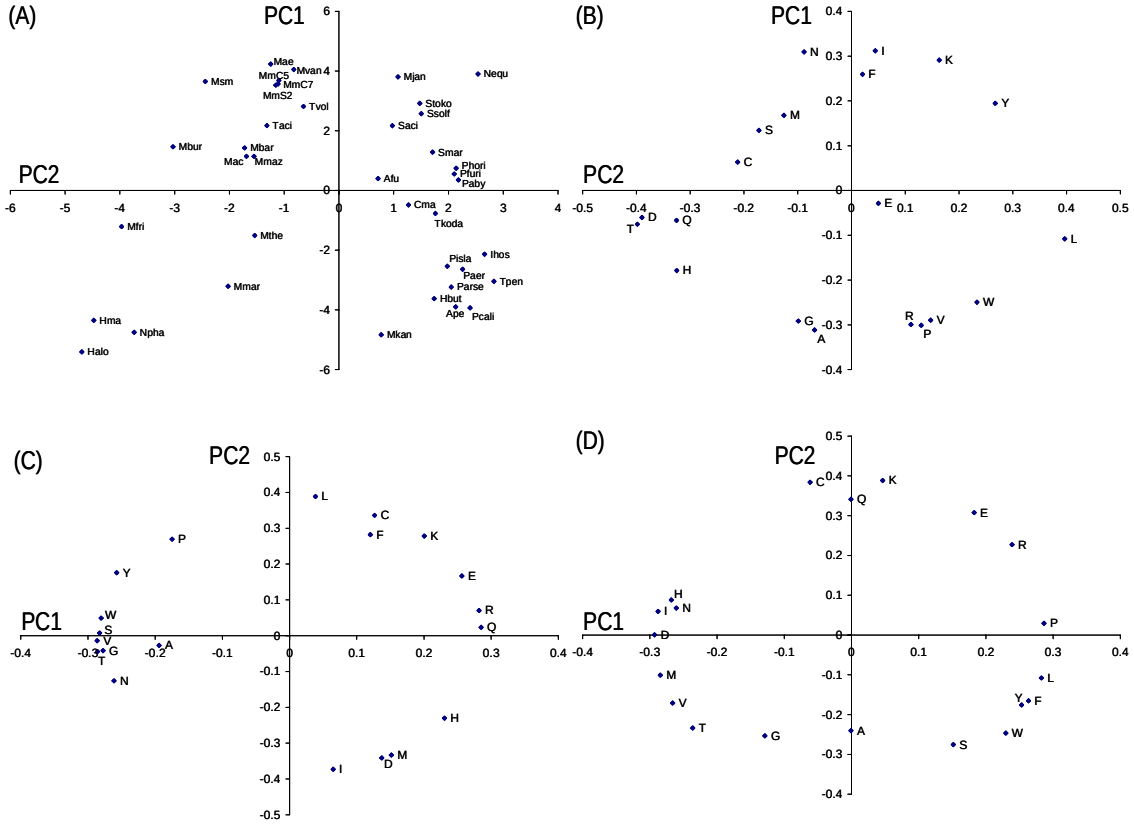


Figure S4. *M. burtonii* gene clusters. (A) Screen shot of Biolayout Express 3D showing three gene clusters (top left panel) obtained from the comparison of the *M. burtonii* genome with all archaeal genomes. The central gene-cluster has been highlighted, and the name, annotation and other details for each gene in the cluster are shown (top right panel) along with the phylogenetic profile (bottom panel). (B) Signal transduction cluster (top left panel) obtained from the comparison of the *M. burtonii* genome with all archaeal genomes.

Figure S4a.

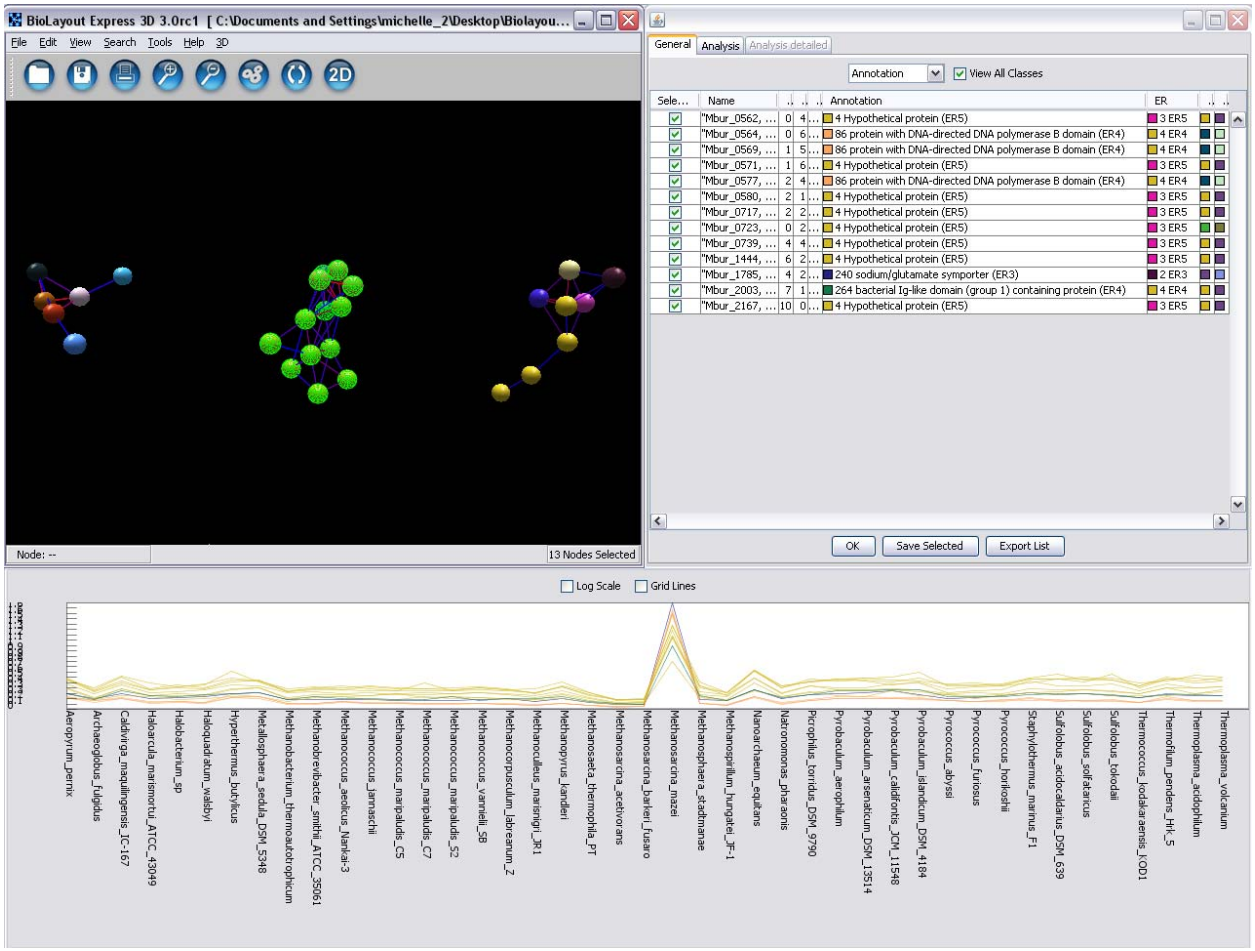


Figure S4b.

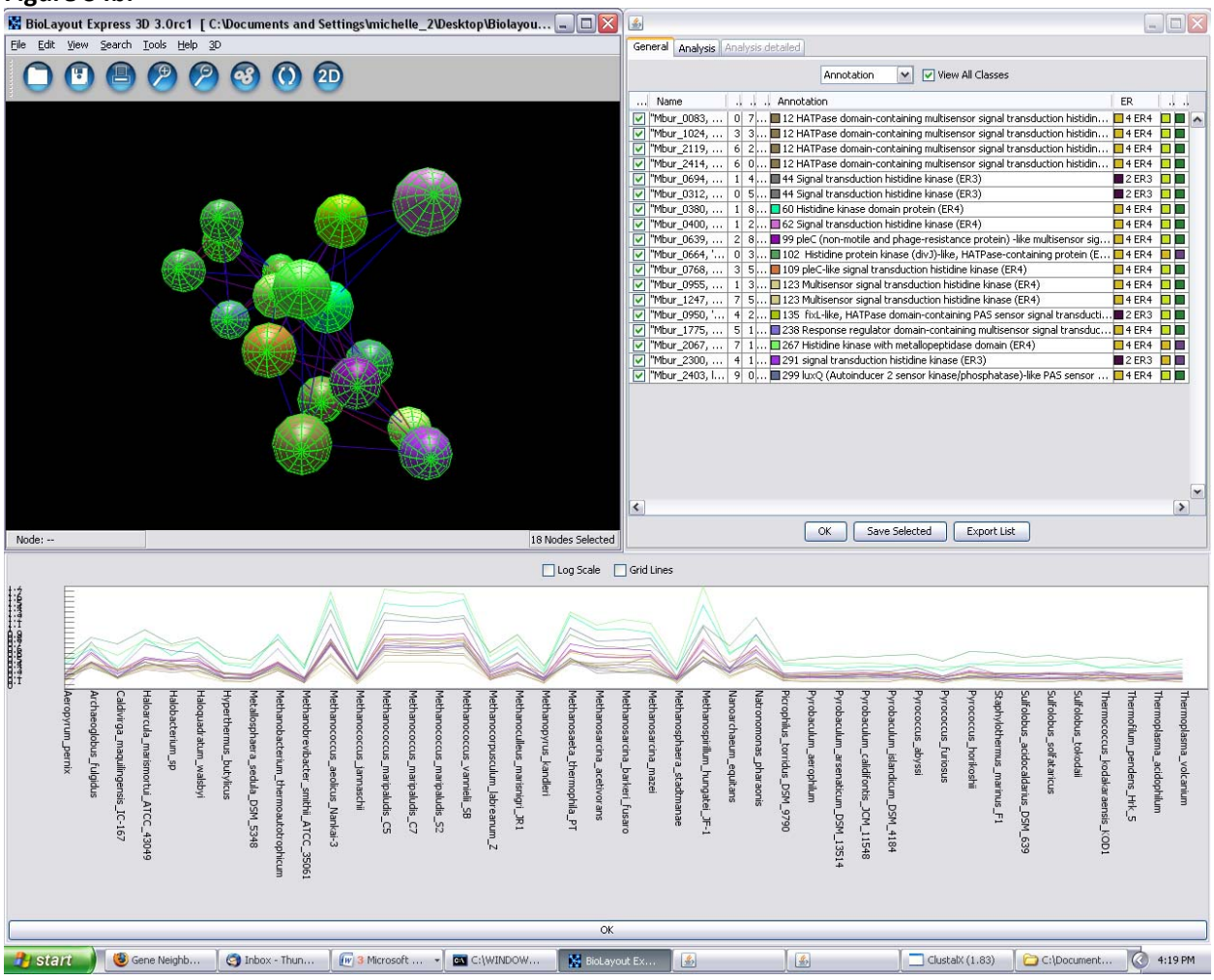

BioLayout Express 3D 3.0rc1 [C:\Documents and Settings\michelle_2\Desktop\Biolayou...
File Edit View Search Tools Help 3D
2D
Node: --
18 Nodes Selected
General
Analysis
Analysis detailed
Annotation
View All Classes
Name
Annotation
ER
OK
Save Selected
Export List
Log Scale
Grid Lines
Aeropyrum_pernix
Archaeoglobus_fulgidus
Caldivirga_maquilingensis_IC-167
Haloarcula_marismortui_ATCC_43049
Halobacterium_sp
Haloquadratum_walsbyi
Hyperthermus_butylicus
Metallosphaera_sedula_DSM_5348
Methanobacterium_thermoautotrophicum
Methanobrevibacter_smithii_ATCC_35061
Methanococcus_aeolicus_Nankai-3
Methanococcus_jannaschii
Methanococcus_maripaludis_C5
Methanococcus_maripaludis_C7
Methanococcus_maripaludis_S2
Methanococcus_vannielii_SB
Methanocorpusculum_labreanum_Z
Methanoculleus_marisnigri_JR1
Methanopyrus_kandleri
Methanosaeta_thermophila_PT
Methanosarcina_acetivorans
Methanosarcina_barkeri_fusaro
Methanosarcina_mazei
Methanosphaera_stadtmanae
Methanospirillum_hungatei_JF-1
Nanoarchaeum_equitans
Natronomonas_pharaonis
Picrophilus_torridus_DSM_9790
Pyrobaculum_aerophilum
Pyrobaculum_arsenaticum_DSM_13514
Pyrobaculum_calidifontis_JCM_11548
Pyrobaculum_islandicum_DSM_4184
Pyrococcus_abyssi
Pyrococcus_furiosus
Pyrococcus_horikoshii
Staphylothermus_marinus_F1
Sulfolobus_acidocaldarius_DSM_639
Sulfolobus_solfataricus
Sulfolobus_tokodaii
Thermococcus_kodakaraensis_KOD1
Thermofilum_pendens_Hrk_5
Thermoplasma_acidophilum
Thermoplasma_volcanium
OK
start
Gene Neighb...
Inbox - Thun...
Microsoft ...
C:\WINDOW...
BioLayout Ex...
ClustalX (1.83)
C:\Document...
4:19 PM

Figure S5. Alignment of transposase sequences. Sequences used for Figure 4. Transposase cluster alignments in interleaved NEXUS format.

```
Cluster 1
Mbur_0866 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQRIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_2252 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQKIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_2297 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQRIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_2104 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQRIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_1167 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQRIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_1117 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQRIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_0915 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQRIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_0744 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQRIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_0648 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQRIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_0401 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQRIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_0324 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQKIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_0539 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQTIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_0479 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQRIIELQAVRITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_0162 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQRIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]
Mbur_0315 MNTKRKEILAVYEQGPEAVVTLVTTLYDIIAEQQTIIELQAARITELEERVKKLEEQLKKNSRNSSKPPSTDVFINEKPK [ 80]

Mbur_0866 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERSLEDIEVKDHEKRQVFDIPPIKLQVTEHRAEIKSC [160]
Mbur_2252 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERSLEDIEVKDHEKRQVFDIPPIKLQVTEHRAEIKSC [160]
Mbur_2297 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERSLEDIEVKDHEKRQVFDIPPIKLQVTEHRAEIKSC [160]
Mbur_2104 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERSLEDIEVKDHEKRQVFDIPPIKLQVTEHRAEIKSC [160]
Mbur_1167 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERSLEDIEVKDHEKRQVFDIPPIKLQVTEHRAEIKSC [160]
Mbur_1117 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERSLEDIEVKDHEKRQVFDIPPIKLQVTEHRAEIKSC [160]
Mbur_0915 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERSLEDIEVKDHEKRQVFDIPPIKLQVTEHRAEIKSC [160]
Mbur_0744 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERSLEDIEVKDHEKRQVFDIPPIKLQVTEHRAEIKSC [160]
Mbur_0648 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERSLEDIEVKDHEKRQVFDIPPIKLQVTEHRAEIKSC [160]
Mbur_0401 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERS---------------------------------- [160]
Mbur_0324 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERS---------------------------------- [160]
Mbur_0539 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERS---------------------------------- [160]
Mbur_0479 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERS---------------------------------- [160]
Mbur_0162 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERS---------------------------------- [160]
Mbur_0315 TKSRRKKSGKKPGGQKDHPGTTLRMVDVPDEVIIHKVHKCSNCERSLEDIEVKDHEKRQVFDIPPIKLQVTEHRAEIKSC [160]

Mbur_0866 PHCGCKNKATFSEKVKQPTQYGLRLASLAVYLHDYQLLPYERSCELRADVCGCEISPATLARAEKTCFEKLEDFEQQIKN [240]
Mbur_2252 PHCGCKNKATFSEKVKQPTQYGLRLASLAVYLHDYQLLPYDRSCELLADVCGCEISPATLARAEKTCFEKLEDFEQQIKN [240]
Mbur_2297 PHCGCKNKATFSEKVKQPTQYGLRLASLAVYLHDYQLLPYERSCELLADVCGCEISPATLARAEKTCFEKLEDFEQQIKN [240]
Mbur_2104 PHCGCKNKATFSEKVKQPTQYGLRLASLAVYLHDYQLLPYERSCELLADVCGCEISPATLARAEKTCFEKLEDFEQQIKN [240]
Mbur_1167 PHCGCKNKATFSEKVKQPTQYGLRLASLAVYLHDYQLLPYERSCELLADVCGCEISPATLARAEKTCFEKLEDFEQQIKN [240]
Mbur_1117 PHCGCKNKATFSEKVKQPTQYGLRLASLAVYLHDYQLLPYERSCELLADVCGCEISPATLARAEKTCFEKLEDFEQQIKN [240]
Mbur_0915 PHCGCKNKATFSEKVKQPTQYGLRLASLAVYLHDYQLLPYERSCELLADVCGCEISPATLARAEKTCFEKLEDFEQQIKN [240]
Mbur_0744 PHCGCKNKATFSEKVKQPTQYGLRLASLAVYLHDYQLLPYERSCELLADVCGCEISPATLARAEKTCFEKLEDFEQQIKN [240]
Mbur_0648 PHCGCKNKATFSEKVKQPTQYGLRLASLAVYLHDYQLLPYERSCELLADVCGCEISPATLARAEKTCFEKLEDFEQQIKN [240]
Mbur_0401 ------------------------------------------------------------------LEDFEQQIKN [240]
Mbur_0324 ------------------------------------------------------------------LEDFEQQIKN [240]
Mbur_0539 ------------------------------------------------------------------LEDFEQQIKN [240]
Mbur_0479 ------------------------------------------------------------------LEDFEQQIKN [240]
Mbur_0162 ------------------------------------------------------------------LEDFEQQIKN [240]
Mbur_0315 PHCGCKNKATFSEKVKQPTQYGLRLASLAVYLHDYQLLPYERSCELRADVCGCEISPATLARAEKTCFEKLEDFEQQIKN [240]

Mbur_0866 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSEAMNVMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_2252 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSDAMNAMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_2297 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSDAMNAMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_2104 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSDAMNAMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_1167 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSDAMNAMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_1117 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSDAMNAMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_0915 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSDAMNAMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_0744 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSDAMNAMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_0648 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSDAMNAMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_0401 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSEAMNAMGILPNFNGTVVHYFWKSYYKYDCDHSICNAH [320]
Mbur_0324 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSDAMNAMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_0539 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYCPHQKRGSDAMNAMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_0479 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYCPHQKRGSEAMNAMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_0162 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSEAMNAMGILPNFNGTVVHDFWKSYYKYDCDHSICNAH [320]
Mbur_0315 FLIESPVINCDETGMRIEGKRQWLHVASTNKMTCYYPHQKRGSEAMNAMGILPNFNGTVVHYFWKSYYKYDCDHSICNAH [320]

Mbur_0866 LLRELTSVSENDDQLWSKAMNILLIDVKKSVDQIREMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_2252 LLRELTSVSENDDQLWSKAMNILLIDVKKSVDQIREMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSK-------- [400]
Mbur_2297 LLRELTSVSENDNQLWSKAMNILLIDVKKSVDQIRGMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_2104 LLRELTSVSENDNQLWSKAMNILLIDVKKSVDQIRGMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_1167 LLRELTSVSENDNQLWSKAMNILLIDVKKSVDQIRGMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_1117 LLRELTSVSENDNQLWSKAMNILLIDVKKSVDQIRGMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_0915 LLRELTSVSENDNQLWSKAMNILLIDVKKSVDQIRGMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_0744 LLRELTSVSENDNQLWSKAMNILLIDVKKSVDQIRGMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_0648 LLRELTSVSENDNQLWSKAMNILLIDVKKSVDQIRGMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_0401 LLRELTSVSENDNQLWSKAMNILLIDVKKSVDQIRKMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_0324 LLRELTSVSENDDQLWSKAMNILLIDVKKSVDQIREMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_0539 LLRELTSVSENDDQLWSKAMNILLIDVKKSVDQIREMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_0479 LLRELTSVSENDDQLWSKAMNILLIDVKKSVDQIREMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_0162 LLRELTSVSENDDQLWSKAMNILLLMSKNQLT-IREMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]
Mbur_0315 LLRELTSVSENDDQLWSKAMNILLLMSKNQLT-IREMSGCMKPERIKEFEDWYGQIIHIGIEENPQLQAKSKKRGRTKQT [400]

Mbur_0866 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSRVRSYISTVKKNQVPVMDAIR [480]
```

```
Mbur_2252 -------------NDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSRVRSYISTVKKNQVPVMDAIR [480]
Mbur_2297 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSRVRSYISTVKKNQIPVMDAIR [480]
Mbur_2104 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSRVRSYISTVKKNQIPVMDAIR [480]
Mbur_1167 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSRVRSYISTVKKNQIPVMDAIR [480]
Mbur_1117 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSRVRSYISTVKKNQIPVMDAIR [480]
Mbur_0915 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSRVRSYISTVKKNQIPVMDAIR [480]
Mbur_0744 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSRVRSYISTVKKNQIPVMDAIR [480]
Mbur_0648 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSRVRSYISTVKKNQIPVMDAIR [480]
Mbur_0401 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSRVRSYISTVKKNQVPVMDAIR [480]
Mbur_0324 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLA---------------------------------------------- [480]
Mbur_0539 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSWVRSYISTVKKNQVPVMDAIR [480]
Mbur_0479 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSRVRSYISTVKKNQVPVMDAIR [480]
Mbur_0162 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFLRVRSYISTVKKNQVPVMDAIR [480]
Mbur_0315 TAKNLLDRFIGYKNDILRFMHDLKVPFENNLAERDVRMMKVQQKISGTFRSMQGALIFSRVRSYISTVKKNQVPVMDAIR [480]

Mbur_0866 NAIAGMPFIPTIV [493]
Mbur_2252 NAIAGMPFIPTIV [493]
Mbur_2297 NAIAGMPFIPTIV [493]
Mbur_2104 NAIAGMPFIPTIV [493]
Mbur_1167 NAIAGMPFIPTIV [493]
Mbur_1117 NAIAGMPFIPTIV [493]
Mbur_0915 NAIAGMPFIPTIV [493]
Mbur_0744 NAIAGMPFIPTIV [493]
Mbur_0648 NAIAGMPFIPTIV [493]
Mbur_0401 NAIAGMPFIPTIV [493]
Mbur_0324 ------------- [493]
Mbur_0539 NAIAGMPFIPTIV [493]
Mbur_0479 NAIAGMPFIPTIV [493]
Mbur_0162 NAIAGMPFIPTIV [493]
Mbur_0315 NAIAGMPFIPTIV [493]
```

Cluster 2

```
Mbur_0398 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_0756 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_0912 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_0918 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_1991 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_0151 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_2170 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_2315 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_2173 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_2442 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_2160 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_2137 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_2019 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_1923 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_1639 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_1446 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_1279 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_0785 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_0584 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_0567 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_0566 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_0131 MSLTNFAFKEEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYINKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]
Mbur_2161 ---------QEYKRLENLGDKLSEIESLIDWKPFRPIIAEMYVNKTEFGGRPNVDEIVMLKMLVLQQWHGLSDPELERQA [ 80]

Mbur_0398 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERIAQSGKEDEIWNEMQRQLDKKGLRIKQGMIQDATFI [160]
Mbur_0756 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERIAQSGKEDEIWNEMQRQLDKKGLRIKQGMIQDATFI [160]
Mbur_0912 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERIAQSGKEDEIWNEMQRQLDKKGLRIKQGMIQDATFI [160]
Mbur_0918 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERIAQSGKEDEIWNEMQRQLDKKGLRIKQGMIQDATFI [160]
Mbur_1991 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERIAQSGKEDEIWNEMQRQLDKKGLRIKQGMIQDATFI [160]
Mbur_0151 TDRISFRKFLGHVKSEIQELQIINLQQFPAKIPDHTTVWAFRERIAQSGKEDEIWNEMQRQLDKKGLRIKQGMIQDATFI [160]
Mbur_2170 TDRISFRKFLGHVKSEIQELQIINLQQFPAKIPDHTTVWAFRERIAQSGKEDEIWNEMQRQLDKKGLRIKQGMIQDATFI [160]
Mbur_2315 TDRISFRKFLGHVKSEIQELQIINLQQFPAKIPDHTTVWAFRERIAQSGKEDEIWNEMQRQLDKKGLRIKQGMIQDATFI [160]
Mbur_2173 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_2442 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_2160 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_2137 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_2019 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_1923 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_1639 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_1446 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_1279 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_0785 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_0584 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_0567 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_0566 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_0131 TDRISFRKFLG----------------FPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]
Mbur_2161 TDRISFRKFLGHVKSEIQELQIINLQQFPAKIPDHTTVWAFRERISQAGKEDEIWNEMQRQLNKKGLKIKQGMIQDATFI [160]

Mbur_0398 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_0756 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_0912 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_0918 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_1991 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_0151 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
```

```
Mbur_2170 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_2315 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_2173 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_2442 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_2160 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_2137 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_2019 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_1923 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_1639 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_1446 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_1279 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_0785 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_0584 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_0567 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_0566 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_0131 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]
Mbur_2161 HADPGHANLDTPRGNEAKTRRCKDGTWTKKASKSHFGYKLHTIEDTEYDLIRRYRTTTASVHDSQVDLSEEGEVVYRDRG [240]

Mbur_0398 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_0756 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_0912 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_0918 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_1991 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_0151 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_2170 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_2315 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_2173 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_2442 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_2160 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_2137 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_2019 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_1923 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_1639 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_1446 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_1279 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_0785 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_0584 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_0567 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_0566 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_0131 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]
Mbur_2161 YFGAISKGYDATMQRGVRGHPIGIRDKMRNKRISRKRAKGERPYAVIKNVFTSGFVRVTTLARVNVKMAITAFSYNLYQL [320]

Mbur_0398 RTIRRKSLG [329]
Mbur_0756 RTIRRKSLG [329]
Mbur_0912 RTIRRKSLG [329]
Mbur_0918 RTIRRKSLG [329]
Mbur_1991 RTIRRKSLG [329]
Mbur_0151 RTIRRKSLG [329]
Mbur_2170 RTIRRKSLG [329]
Mbur_2315 RTIRRKSLG [329]
Mbur_2173 RTIRRKSLG [329]
Mbur_2442 RTIRRKSLG [329]
Mbur_2160 RTIRRKSLG [329]
Mbur_2137 RTIRRKSLG [329]
Mbur_2019 RTIRRKSLG [329]
Mbur_1923 RTIRRKSLG [329]
Mbur_1639 RTIRRKSLG [329]
Mbur_1446 RTIRRKSLG [329]
Mbur_1279 RTIRRKSLG [329]
Mbur_0785 RTIRRKSLG [329]
Mbur_0584 RTIRRKSLG [329]
Mbur_0567 RTIRRKSLG [329]
Mbur_0566 RTIRRKSLG [329]
Mbur_0131 RTIRRKSLG [329]
Mbur_2161 RTIRRKSLG [329]
```

Cluster 3

```
Mbur_0087 LCMNCPRCKSSNHKKNGKIDGRQRYKCHDCGYSYSVEIKSTASPMSVKRQALQLYLEGLGFRSIGRFLGVSHVSVQKWIK [ 80]
Mbur_1400 LCMNCPRCKSSNHKKNGKIDGRQRYKCHDCGYSYSVEIKSTASPMSVKRQALQLYLEGLGFRSIGRFLGVSHVSVQKWIK [ 80]
Mbur_0975 --MNCPRCKSSNHKKNGKIDGRQRYKCHDCEYNYSVELKSTASPMSVKRQALQLYLEGLGFRSIGRFLRVSHVSVQKWIK [ 80]
Mbur_1253 --MNCPRCKSFNHKKNGKIDGRQRYKCHDCEYNYSVELKSTASPMSVKRQALQLYLEGLGFRSIGRFLGVSHVSVQKWIK [ 80]
Mbur_0410 --MNCPKCKSSNHKKNGKIDGLQRYKCHDCEYNYSVELKSTASPMSVKRQALQLYLEGLGFRSIGRFLGVSHVSVQKWIK [ 80]
Mbur_1021 --MNCSRCKSSNHKKNGKINRFQRYKCHDCGYN-LVEIKSTARPMSVKRQALQLYLEGLGFRSIGRFLGVSHVSVQKWIK [ 80]
Mbur_1668 --MNCPRCKSSSHKKNGKIDGRQRYKCHDCGYSYSVEIKSTASPMSVKRQALQLYLEGLGFRSIGRLLGVSHVSVQKWIK [ 80]
Mbur_2290 LCMNCPRCKSSSHKKNGKIDGRQRYKCHDCGYSYSVEIKSTASPMSVKRQALQLYLEGLGFRSIGRLLGVSHVSVQKWIK [ 80]
Mbur_1018 --MNCPRCKSSNHKKNGKIDGHQRYKCHDCGYSYSVEIKSTASPMSVKRQALQLYLEGLGFRSIGRLLGVSHVSVQKWIK [ 80]
Mbur_1800 --MNCPRCKSSNHKKNGKIDGRQRYKCHDCEYNYSVELKSTASPMSVKRQALQLYLEGLGFRSIGRLLGVSHVSVQKWIK [ 80]

Mbur_0087 KFGRELEDLKSENEISIVELDEMHTYIGNKKNIVGSGLLLIEMEKGSSTALLVAGERKQD-NSGKNRGR---RLEKLITG [160]
Mbur_1400 KFGRELEDLKSENEISIVELDEMHTYIGNKKNIVGSGLLLIEMEKGSSTALLVAGERKQD-NSGKNRGR---RLEKLITG [160]
Mbur_0975 KFGRELENLKSENEILIVELDEMHTYIGNKTNIVGSGLLLIEMEKGSSTALLVAGERKQD-NSGKNRGR---RLEKLITG [160]
Mbur_1253 KFGRELEDLKSENEISIVELDEMHTYIGNKKNIVGSGLLLIEMEKGS-TALLVAGERKQD-NSGKNRGR---RLEKLITG [160]
Mbur_0410 KFGRELEDLKSENEISIVELDEMHTYIGNKKNIVGSGLLLIEMEKGSSTALLVAGERKQD-NSGKNRGR---RLEKLITG [160]
Mbur_1021 KFGRELEDLKSENEISIVELDEMHTYIGNKKNIVGSGLLLIEMEKGSSTALLVAGERKQD-NSGKKRGR---RLEKLITG [160]
Mbur_1668 KFGRELEDIKSENEISIVELDEMHTYIGNKKNIVGSGLLLIEMEKGSSTALLVAGGRKQD-SSGKSRGR---RLEKLITG [160]
Mbur_2290 KFGRELEDIKSENEISIVELDEMHTYIGNKKNIVGSGLLLIEMEKGSSTALLVAGGRKQD-SSGKSRGR---RLEKLITG [160]
```

```
Mbur_1018 KFGRELEDLKSENEILIVELDEMHTYVGNKKNSVGSGLLLIQMEKGSSTALLVAGGRKQDKSSGKSRGR---RLEKLITG [160]
Mbur_1800 KFGRELEDIKSENEISIVELDEMHTYIGNKKILLD---LDCCRWKKVHRLLFWQGEGNRTKALEKVKGEGDWRSDDSLEG [160]

Mbur_0087 GHMQSLFQRKFTLNPKLKHILLKDITAYSGTFWQDEESQSVILRVLKCSTLFFFSTEIKNRCLI [224]
Mbur_1400 GHMQSLFQRKFTLNPKLKHILLKDITAYSGTFWQDEESQSVILRVLKCSTLFFFSTEIKNRCLI [224]
Mbur_0975 GHMQSLFQRKFTLNPKLKHILLKDITAYSGTFWQDEESQSVILRVLKCSTLLFFSTEIKNLCLI [224]
Mbur_1253 GHMQSLFQRKFTLNPKLKHILLKDITAYSGTFWQDEESQSVILRVLKCSTLFFFSTEIKSLCLV [224]
Mbur_0410 GHMQSLFQRKFTLNPKLKHILLKDITAYSGTFWQDEESQSVILRVLKCSTLFFFSTEIKNRCLI [224]
Mbur_1021 GHMQSLFQRKFTLNPKLKHILLKDITAL-GTFWQDEESQNVIPRVLKCSTLFFFST-IK----- [224]
Mbur_1668 GHMQSFFQRKFILNPKLKHILLKDITAYSGTFWQDEESQSVILRVLKCSTLFFFSTEIKSLCLV [224]
Mbur_2290 GHMQSFFQRKFILNPKLKHILLKDITAYSGTFWQDEESQSVILRVLKCSTLFFFSTEIKSLCLV [224]
Mbur_1018 GHMQSFFQSKFILNPKLKHILLKDITAL-GTFQ-GEESQSVILRVLKCSTLFFFSTEIKSLCLI [224]
Mbur_1800 ICRVSSRENSYSIQSNIYCRIQHN--ALSGKIEKKVKVLYESNAEVLCSSFDEVQKRISYVL-- [224]

Cluster 4
Mbur_0965 MELYDLISDYLNDSNVSLKPLITYFLNEVMEQEAIEQSGAGKHERSITRTAHRNGYRDRSLTTRHGELTLKKPQLRDFPF [ 80]
Mbur_1067 MELYDLISDYLNDSNVSLKPLITYFLNEVMEQEAIEQSGAGKHERSITRTAHRNGYRDRSLTTRHGELTLKKPQLRDFPF [ 80]
Mbur_0323 MELYDLISDYLNDSNVSLKPLITYFLNEVMEQEAIEQSGAGKHERSITRTAHRNGYRDRSLTTRHGELTLKKPQLRDFPF [ 80]
Mbur_0637 MELYDLISDYLNDSNVSLKPLITYFLNEVMEQEAIEQSGAGKHERSITRTAHRNGYRDRSLTTRHGELTLKKPQLRDFPF [ 80]

Mbur_0965 TTQVFERYSRTEKAIENAIVESYVQGVSTRKVEKIISQLGVESISRSRVSRIAQDLDKTVHGFMNKPIEHEIKYLYVDAT [160]
Mbur_1067 TTQVFERYSRTEKAIENAIVESYVQGVSTRKVEKIISQLGVESISRSRVSRIAQDLDKTVHGFMNKPIEHEIKYLYVDAT [160]
Mbur_0323 TTQVFERYSRTEKAIENAIVESYVQGVSTRKVEKIISQLGVESISRSRVSRIAQDLDKTVHGFMNKPIEHEIKYLYVDAT [160]
Mbur_0637 TTQVFERYSRTEKAIENAIVESYVQGVSTRKVEKIISQLGVESISRSRVSRIAQDLDKTVHGFMNKPIEHEIKYLYVDAT [160]

Mbur_0965 YLKVRDRVRYVNKAVFIVAGVKNDGYREILGVKIADSEEAMFWEEMFTDLKERGLRGVELVISDGHKGIQRAVERQFLGA [240]
Mbur_1067 YLKVRDRVRYVNKAVFIVAGVKNDGYREILGVKIADSEEAMFWEEMFTDLKERGLRGVELVISDGHKGIQRAVERQFLGA [240]
Mbur_0323 YLKVRDRVRYVNKAVFIVAGVKNDGYREILGVKIADSEEAMFWEEMFTDLKERGLRGVELVISDGHKGIQRAVERQFLGA [240]
Mbur_0637 YLKVRDRVRYVNKAVFIVAGVKNDGYREILGVKIADSEEAMFWEEMFTDLKKRGLRGVELVISDGHKGIQRAVERQFLGA [240]

Mbur_0965 SWQMCIVHLERLILKKLPRKHHKEAMESFKEVQDDQTKLLGLMVEWDRPGFEKAAETIERFQHGLTNYQAFPKEHWKRIK [320]
Mbur_1067 SWQMCIVHLERLILKKLPRKHHKEAMESFKEVQDDQTKLLGLMVEWDRPGFEKAAETIERFQHGLTNYQAFPKEHWKRIK [320]
Mbur_0323 SWQMCIVHLERLILKKLPRKHHKEAMESFKEVQDDQTKLLGLMVEWDRPGFEKAAETIERFQHGLTNYQAFPKEHWKRIK [320]
Mbur_0637 SWQMCMVHLERLILKKLPRKHHKEAMESFKEVQDDQTKLLGLMVEWDRPGFEKAAETIERFQHGLTNYQAFPKEHWKRIK [320]

Mbur_0965 TTNMIERLNKEVKRRSKVVGAFPNDESLMRLVIAVLIDQNENWITGNRYLTMED [374]
Mbur_1067 TTNMIERLNKEVKRRSKVVGAFPNDESLMRLVIAVLIDQNENWITGNRYLTMED [374]
Mbur_0323 TTNMIERLNKEVKRRSKVVGAFPNDESLMRLVIAVLIDQNENWITGNRYLTMED [374]
Mbur_0637 TTNMVEILNKEVKRRSKVVGAFPNDESLMRLVVAVLIDQNENWITGNRYLTMED [374]

Cluster 5
Mbur_0459 MYQSMKYYYDVSDGICTRDSIANYLNTDNASICQFLYFLDIDDIASYVESSYYADKDWHFRYKVSSMIKLIVVKCYRNLS [ 80]
Mbur_0585 ----MKYYYDVSDGICTRDSIANYLNTDNASICQFLYFLDIDDIASYVESSYYADKDWHFRYKVSSMIKLIVVKCYRNLS [ 80]
Mbur_0535 MYQPMKYYYDVSDGICTRDSIANYLNTDNASICQFLYFLDIDDIASYVESSYYADKDWHFRYKVSSMIKLIVVKCYRNLS [ 80]
Mbur_0945 MYQSMKYYYDVSDGICTRDSIANYLNTDNASICQFLYFLDIDDIASYVESSYYTDKDWHFRYKVSSMIKLIVVKCYRNLS [ 80]
Mbur_0403 ------------------------------------------------------------MIKLIVVKCYRNLS [ 80]
Mbur_0738 ------------------------------------------------------------MIKLIVVKCYRNLS [ 80]
Mbur_0870 MYQSMKYYYDVSDGICTRDSIANYLNTDNASICQFLYFLDIDDIASYVESSYYTDKDWHFRYKVSSMIKLIVVKCYRNLS [ 80]
Mbur_2007 MCQSMKYYYDVSDGICTRDSIANYLNTDNASICQFLYFLDIDDIASYVESSYYTDKDWHFRYKVSSMIKLIVVKCYRNLS [ 80]

Mbur_0459 FEKTISTLTKEEAQLLSFEDNNGIMNLPSPATLHHFVIYRLGKTGLDEVMFKIGKNISKNTKIRDAKTDSTPLEASRYDK [160]
Mbur_0585 FEKTISTLTKEEAQLLSFEDNNGIMNLPSPATLHHFVKYRLVKTGLDEVMFKIGKNISKNTKIRDAKTDSTPLEASRYDK [160]
Mbur_0535 FEKTISTLTKEEAQLLSFEDNNGIMNLPSPATLHHFVKYRLGKTGLDEVMFKIGKNISKNTKIRDAKTDSTPLEASRYDK [160]
Mbur_0945 FEKTISTLTKEEAQLLSFEDNNGIMNLPSPATLHHFVKYRLGKTGLDEVMFKIGKNISKNTKIRDAKTDPTPLEASRYDK [160]
Mbur_0403 FEKTISTLTKEEAQLLSFEDNNGIMNLPSPATLHHFVKYRLGKTGLDEVMFKIGKNISKNTKIRDAKTDSTPLEASRYDK [160]
Mbur_0738 FEKTISTLTKEEAQLLSFEDNNGIMNLPSPATLHHFVKYRLGKTGLDEVMFKIGKNISKNTKIRDAKTDSTPLEASRYDK [160]
Mbur_0870 FEKTISTLTKEEAQLLSFEDNNGIMNLPSPATLHHFVKYRLGKTGLDEVMFKIGKNISKNTKIRDAKTDSTPLEASRYDK [160]
Mbur_2007 FEKTISTLTKEEAQLLSFEDNNGIMNLPSPATLHHFVKYRLGKTGLDEVMFKIGKNISKNTKIRDAKTDSTPLEASRYDK [160]

Mbur_0459 YADYNPHYNCKMDKAHITMIGTLPIYMTHTKGASHDSPELKKHIDALVEMGVDIDTYALDGGYDSFRNHADIWYKLNAKP [240]
Mbur_0585 YADYNPHYNCKMDKAHITMIGTLPIYMTHTKGASHDSPELKKHIDALVEMGVDIDTYALDGGYDSFRNHADIWYKLNAKP [240]
Mbur_0535 YADYNPHYNCKMDKAHITMIGTLPIYMTHTKGASHDSPELKKHIDALVEMGVDIDTYALDGGYDSFRNHADIWYKLNAKP [240]
Mbur_0945 YADYNPHYNCKMDKAHITMIGTLPIYMTHTKGASHDSPELKKHIDALVEMGVDIDTYALDGGYDSFRNHADIWYKLNAKP [240]
Mbur_0403 YADYNPHYNCKMDKAHITMIGPLPIYMTHTKGASHDSPELKKHIDALVEMGVDIDTYALDGGYDSFRNHADIWYKLNAKP [240]
Mbur_0738 YADYNPHYNCKMDKAHITMIGPLPIYMTHTKGASHDSPELKKHIDALVEMGVDIDTYALDGGYDSFRNHADIWYKLNAKP [240]
Mbur_0870 YADYNPHYNCKMDKAHITMIGTLPICMTHTKGASHDSPELKKHIDALVEMGVDIDTYALDGGYDSFRNHADIWYKLNAKP [240]
Mbur_2007 YADYNPHYNCKMDKAHITMIGTLPIYMTHTKGASHDSPELKKHIDALVEMGVDIDTYALDGGYDSFRNHADIWYKLNAKP [240]

Mbur_0459 VIAYSSDSKVQYEGMMERIDHWVNKMWKLGGSIHMKYEEKLHFLYENGREKQVGMHLRNKNIKDDGFDEDYSHRGECERV [320]
Mbur_0585 VIAYSSDSKVQYEGMMERIDHWVNKMWKLGGSIHMKYEEKLHFLYENGREKQVGMHLRNKNIKDDGFDEDYSHRGECERV [320]
Mbur_0535 VIAYSSDSKVQYEGMMERIDHWVNKMWKLGGSIHMKYEEKLHFLYENGREKQVGMHLRNKNIKDDGFDEDYSHRGECERV [320]
Mbur_0945 VIAYSSDSKVQYEGMMERIDHWVNKMWKLGGSIHMKYEEKLHFLYENGREKQVGMHLRNKNIKDDGFDEDYSHRGECERV [320]
Mbur_0403 VIAYSSDSKVQYEGMMERIDHWVNKMWKLGGSIHMKYEEKLHFLYENGREKQVGMHLRNKNIKDDGFDEDYSHRGECERV [320]
Mbur_0738 VIAYSSDSKVQYEGMMERIDHWVNKMWKLGGSIHMKYEEKLHFLYENGREKQVGMHLRNKNIKDDGFDEDYSHRGECERV [320]
Mbur_0870 VIAYSSDSKVQYEGMMERIDHWVNKMWKLGGSIHMKYEEKLHFLYENGREKQVGMHLRNKNIKDDGFDEDYSHRGECERV [320]
Mbur_2007 VIAYFSDSKVQYEGMMERIDHWVNKMWKLGGSIHMKYEEKLHFLYENGREKQGGMHLRNKNIKDDGFDEDYSHRGECERV [320]

Mbur_0459 HNHIKWTVKFDIRGMKNGSKKLYSVMNFVAYQLLVATNLQNGVKETNSFANYV [373]
Mbur_0585 HNHIKWTVKFDIRGMKNGSKKLYSVMNFVAYQLLVATNLQNGVKETNSFANYV [373]
Mbur_0535 HNHIKLTVKFDIRGMKNGSKKLYSVMNFVAYQLLVATNLQNGVKETNSFANYV [373]
Mbur_0945 HNHIKLTVKFDIRGMKNGSKKLYSVMNFVAYQLLVATNLQNGVKETNSFANYV [373]
Mbur_0403 HNHIKWTVKFDIRGMKNSSKKLYSVMNFVAYQLLVATNLQNGVKETNSFANYV [373]
Mbur_0738 HNHIKWTVKFDIRGMKNSSKKLYSVMNFVAYQLLVATNLQNGVKETNSFANYV [373]
Mbur_0870 HNHIKWTVKFDIRRMKNSSKKLYSVMNFVAYQLLVATNLQNGVKETNSFTNYV [373]
Mbur_2007 HNHIKWTVKFDIRGMKNSSKKLYSVMNFVAYQLLVATNLQNGVKETNSFANYV [373]
```

Cluster 6

```
Mbur_0228 -------------MEDKELIQITLGLLSPWFVKDIDLNTSKRRMDIYLDFSKGTKFSCPVCNKLSELHDTKKKVWRHLDF [ 80]
Mbur_0665 -------------MEDKELIQITLGLLSPWFVKDIDLNTSKRRMDIYLDFSKGTKFSCPVCNKLSELHDTKKKVWRHLDF [ 80]
Mbur_0406 VGNLIFISSNSISMEDKELLQIALGLSSPWFVKDIDLNTSKKRMDIYLDFTKGTKFPCPICNKLSELHDTKEKVWRHLDF [ 80]

Mbur_0228 FHYETYLHARVPRTKCNEHGVKLVNVPWTRQNTGFTLFFEALIVA----ISKEMTVSAIAEMINIHEDSVWRILTHYVNK [160]
Mbur_0665 FHYETYLHARVPRTKCNEHGVKLVNVPWTRQNTGFTLFFEALIVA----ISKEMTVSAIAEMINIHEDSVWRILTHYVNK [160]
Mbur_0406 FHHETYLHTRVPRTKCNEHGVKLVNVPWTRLNTGFTLFLEALFVACPKKMSKKMTVSAIADMVNGHEDSLWRILSHYVKE [160]

Mbur_0228 AAAKMDLSGLDTIGVDEISVKKGHSYVTLFYDLNKSRIIHIENGKKRSVFKKFREYLSKKIDPDNIKYISMDMYPAFKGG [240]
Mbur_0665 AAAKMDLSGLDTIGVDEISVKKGHSYVTLFYDLNKSRIIHIENGKKRSVFKKFREYLSKKIDPDNIKYISMDMYPAFKGG [240]
Mbur_0406 SMVKTDLSNLDIIGVDEISVKKGHSYVTLFYDLNKDRVIHIENGKRKDGFRKFRDFLSTKTNPDNIKYISMDMSPAFKGG [240]

Mbur_0228 AREYFPNAKIVYD-------KFHIVKMMNDAIDKVRRSEYQSNKDLGKTRFMWLKNPENLSDREIAKIQSIKDLDTKTAK [320]
Mbur_0665 AREYFPNAKIVYD-------KFHIVKMMNDAIDKVRRSEYQSNKDLGKTRFMWLKNPENLSDREIAKIQSIKDLDTKTAK [320]
Mbur_0406 AKEYFPNAKVVFDKYREDVGKFHIVKMMNDAIDKVRRSEYQSNKELGKTRFMWLKNPENLTDREIIKIKSIKDLDTKTAK [320]

Mbur_0228 AYKFKLALQRLWEIKNMDVAREYIEKWHYWGTHSNIKEVITLAKMIKRNSHGILESIKGNISNGVVEGLNNKIKTAFKRS [400]
Mbur_0665 AYKFKLALQRLWEIKNMDVAREYIEKWHYWGTHSNIKEVITLAKMIKRNSHGILESIKGNISNGVVEGLNNKIKTAFKRS [400]
Mbur_0406 AYKFKLALKRFWEIKNIRVAREYLEKWYYWGTHSNIKEIITLAKMIKRNSYGILESIKGDNSNGVVEGLNNKIKTAFKRS [400]

Mbur_0228 YGLKTEVYRNKMIFLMAGKLSLPTRC [426]
Mbur_0665 YGLKTEVYRNKMIFLMAGKLSLPTRC [426]
Mbur_0406 YGLKTEMYRNTMIFLMAGKLKLPTRS [426]
```

Supplemenatry Information: manual annotation
Note: Figures and Tables for this Supplementary Information are denoted by "SI".

Background. Genome sequences for more than 750 cellular forms of life have been completed since the first genome was published in 1995 [1], the vast majority of which have been for microorganisms. The most well characterized microorganisms have the largest number of experimentally defined functions assigned to their individual genes. *Escherichia coli* K-12, first isolated in 1922, has been the subject of more than 100,000 published studies [2] with 15,000 publications cited in the EcoCyc database describing properties linked to the experimentally defined functions for 66% of the 4460 gene [3]. In contrast, most microorganisms have less than a handful of their genes experimentally characterized, with many new species having no experimental validation for genes in their newly sequenced genomes.

As a result of the general lack of experimental evidence for the function of most genes, genome annotation relies heavily on the capacity to infer function using auto-annotation pipelines that search, on various levels, for protein homology. A typical annotation pipeline includes identification of protein coding genes, and assignment of gene function on the basis of BLAST searches and multiple alignments with Hidden Markov models with data from a wide range of sources such as the Swiss-Prot, Pfam, TIGRFam, COG, and InterPro databases ([8-9], http://imgweb.jgi-psf.org/w/doc/img_er_ann.pdf, http://www.jcvi.org/cms/research/projects/annotation-service/overview/). Genomic context and synteny may also be employed [10].

Despite the development in sophistication of auto-annotation pipelines, important limitations persist. These range from the propagation of erroneous annotations to conflicts between the results of different annotation pipelines [11-14] Databases such as the curated part of UniProt offer the benefit of providing evidence for why a particular gene function has been assigned [15]. This is important, as a particularly common source of error is the assignment of function based solely on sequence similarity.

In response to the problems with automated annotations, the majority of commonly used annotation platforms, including IMG-ER, MaGe, Artemis, GenDB, and Manatee, have incorporated a capacity for manual editing [16], and the annotation of a number of genomes has been improved by manual curation. While most eucaryal metazoan genomes have been intensively curated (*e.g. Arabidopsis thaliana* [17] and *Apis mellifera* [18]), most microbial genomes have not. Microbial genome papers that do cite manual annotation typically provided no details of how it was performed [19]. Approximately 30 microbial genome papers do provide some details of the manual annotation process. In some cases the manual curation corrected structural features, such as placement of start codons, (*e.g. Clavibacter michiganensis* [20]) while others appraised the data used to infer function (*e.g.* hits to various databases, performing of protein alignments) in order to identify erroneous functional assignments [14, 21]. Although highly desirable, in only a small number of cases did manual curation extend to explicitly identifying the primary experimental data from the literature which supported the assigned function of each protein (*e.g. E. coli* [3] , *Helicobacter pylori* [22], *Pseudoalteromonas haloplanktis* [23]).

Approach to manual annotation. The protein sequence for each gene was searched against the Swiss-Prot and Protein Data Bank (PDB) databases using BLAST [24] to identify the most closely related, experimentally characterized homolog available in the literature. The curated part of Swiss-Prot database was selected because of the extent and quality of annotations associated with each protein sequence [25], and the PDB provides an archive of experimentally-determined three-dimensional protein structures. BLAST matches were examined sequentially, searching for a match with direct experimental verification of the function. Matching proteins were not considered if they had no function listed, a function determined 'by similarity', or no reference cited. Experimental evidence that was considered acceptable to define function included papers detailing the expression and characterization of the protein, protein crystallography studies, and mutation and complementation studies that defined function. Papers that only documented the nucleotide sequence, including genome sequences, were not considered to provide sufficient evidence of function. Nor were unpublished protein crystal structures or the (published or unpublished) crystal structures of hypothetical proteins for which function was not known. In most cases papers were read in sufficient depth to establish if conclusions about function were justified. Choosing published experimental evidence as a prerequisite for defining function probably excluded some valid

unpublished data, however it was not possible to assess this data and it was excluded. Careful attention was also given to the recent literature in order to identify valid experimental data (*e.g.* several methanogenesis-related proteins: Mbur_0808, Mbur_0811) that had not yet been updated in protein databases.

After the best experimentally characterized homolog had been determined, InterPro and Pfam domains were checked for their presence in the *M. burtonii* homolog to confirm that all identifiable functional domains were conserved. The graphical representation on the Swiss-Prot BLAST output was used for assessing the arrangement and extent of domain matches throughout the length of the query and matching proteins, thereby providing a rapid way to assess both the global (whole protein) and local (domain and motif) similarity between the *M. burtonii* and the functionally characterized proteins.

An Evidence Rating (ER) system was developed to enable the confidence of functional assignments to be clearly displayed for each gene (Figure SI1), providing a ranking from ER1 (experimental characterization of the product from the parent organism) to ER5 (hypothetical). A team of twelve researchers performed the manual annotation. The annotation rate generally varied from 1 gene per 10 min to 1 per hour, with the rate dependent on the individual's prior and gained expertise, prior familiarity with the particular gene or associated biological process (*e.g.* pathway), and the time taken to determine if an experimentally characterized matching gene was present in the literature. The total time spent manually annotating the 2494 genes was estimated as 1900 hours. Individuals were initially allocated genes present in Clusters of Orthologous Genes (COG) categories, with genes not present in COGs annotated by the senior Research Associate. Some genes belonged to multiple COGs, leading to annotations being verified more than once. Evaluating the range of data (*e.g.* literature, BLAST scores, domain matches), rationalizing conflicting data, and deciding on an accurate functional description and ER value was not trivial. Initial training, discussions about disputed annotations and experience gained, all contributed to streamlining and improving efficiency. After the completion of manual annotation for all genes, a senior member of the team reviewed the annotations.

Greatly improving annotation accuracy. The completed genome sequence of *M. burtonii* DSM 6242 contains 2575032 bp in a single circular chromosome with 2494 predicted genes of which 2431 are CDSs. Prior to the manual annotation process 1478 CDSs (61%) were allocated a functional prediction and 953 (39%) were not (Table SI1). During the manual annotation, the wording for 1059 genes was changed. While some were minor changes, such as semantic changes from an accepted name to the standard EC nomenclature, or the addition of a gene symbol designation, many were important changes (see examples in Table SI2; full data is available in Table SI4 and SI5).

The annotations of 725 genes were changed to a more specific annotation on the basis of evidence from the literature, 34 of which described distinctly different gene functions. The majority of these 34 genes appear to have been erroneously named based on the presence of a high-scoring Pfam or InterPro domain. For example, Mbur_1229 was originally designated shikimate/quinate 5-dehydrogenase, a gene involved in the biosynthesis of aromatic amino acids, and was changed after manual inspection to glutamyl-tRNA reductase, representing a gene involved in the first step in tetrapyrrole biosynthesis by the C5 pathway (Table SI2). All top 100 matches to Mbur_1229 in Swiss-Prot were to glutamyl-tRNA reductase, as was the top match from the PDB (E-value $2.3x10^{-67}$, sequence identity 151/386 aa (39%), positive matches 236/386 aa (61%)). The aberrant initial annotation was traced to the presence of a domain corresponding to a region of 144 aa in the center of the protein called shikimate_DH (PF01488) with an E-value of $1.3x10^{-50}$. This domain scored higher than the glutamyl-tRNAGlu reductase domains that were also identified (PF05201 and PF00745, E-values $9.9x10^{-44}$ and $9.1x10^{-22}$, respectively) and thus was adopted by the auto-annotation pipeline. Experimentally characterized glutamyl-tRNAGlu reductase contains all three Pfam domains, providing confidence that the assigned Mbur_1229 is a glutamyl-tRNAGlu reductase. Several other examples are detailed in Table SI2, with full data available in Table SI4.

For two (Mbur_1982, Mbur_0434) of the 34 genes, it is unclear how the original erroneous annotation had occurred. In three cases, the experimental characterization of an *M. burtonii* homolog was published within the last year and the corrected name had not entered the Swiss-Prot database prior to auto-annotation (Mbur_1195 [26], and Mbur_1938 and Mbur_2118 [27]). However, information sufficient

to assign a distinctly different function to 31 of the 34 genes has been readily available in the public domain for a long time.

For the majority of genes that received a more specific annotation (533 out of 725), information was uncovered that led to a refinement of the initial annotation. The main types of improvements were, 1) modifications from a family designation to a specific protein (*e.g.* Mbur_2113), 2) change from a pathway designation to a specific protein within that pathway (*e.g.* Mbur_ 1702), 3) modifications that determined substrate specificity (*e.g.* for ABC transporters such as Mbur_1015, and for others such as corrinoid methyltransferase Mbur_0840), 4) determination of particular protein subunits within multi-subunit protein complexes (*e.g.* Mbur_1178), 5) assignment of known functions to domains for those described with a "domain of unknown function", and 6) the determination of membership of a protein family or conserved domain for 115 genes previously described as hypothetical. In addition, of the 158 genes described as "pseudogenes", 91 possessed similarity to known proteins or domains and were given appropriate annotations, and 67 were reclassified as hypothetical proteins due to the absence of discernable functional information (see below, "Identifying incorrect and omitted genes").

The annotations of 334 genes were changed to a less specific annotation. Examples of this include, 1) genes where the annotation was changed to reflect a broader category (*e.g.* Mbur_0175) including ABC-type transporters where the evidence for substrate specificity was lacking (*e.g.* Mbur_2133), 2) genes of a specific enzyme family that did not fulfill criteria for their specific functional annotation (*e.g.* Mbur_2429), and 3) genes initially named due to the presence of a particular domain or motif where the evidence for the presence of these features was not sufficiently strong (*e.g.* Mbur_0504). In addition, 27 genes originally assigned a function did not have experimental evidence or domain matches to support the annotations and were reclassified as hypothetical proteins, ER5 (see below "The value of the Evidence Rating System")

The benefits of improving the accuracy of the functional annotations are illustrated by comparing the vitamin B12 biosynthesis pathway inferred from auto-annotation *vs* manual annotation (Figure SI2). Auto-annotation provided evidence for five (Mbur_2355, Mbur_2356, Mbur_2358, Mbur_2359 and Mbur_2360) genes involved in the aerobic adenosylcobalamin biosynthesis pathway, and two copies of the genes required for late insertion of cobalt, cobalt chelatase *cobN* (Mbur_0865 and Mbur_2260). Two other genes relevant to vitamin B12 synthesis were also auto-annotated, *cbiX* and *cysGB*, although these are involved early in the anaerobic pathway. Manual annotation revealed that functional assignments for CbiX and CysGB were sound, while the two CobN homologs lacked a key domain required for CobN function and were both reassigned as a magnesium/cobalt chelatase-domain containing protein. Three (Mbur_2356, Mbur_2358 and Mbur_2360) of the five gene products originally assigned to the aerobic pathway exhibited highest levels of identity to experimentally characterized homologs from *Salmonella enterica* serovar Typhimurium, which uses the anaerobic pathway [28], and three more genes (Mbur_1701, Mbur_1702 and Mbur_2357) involved in the anaerobic pathway were identified. Six additional genes (Mbur_0191, Mbur_2087, Mbur_2090, Mbur_2091, Mbur_2093, Mbur_2098) involved in the vitamin B12 pathway were also identified. The manual functional annotations of these individual genes provided the basis for complete pathway reconstruction, which revealed strong evidence for the anaerobic biosynthesis pathway, consistent with the anaerobic growth properties of the organism. The ease with which the anaerobic pathway was identified not only illustrates how the manual annotation expedited the identification of (in hindsight) an expected pathway, but illustrates the confidence that may therefore be assumed when identifying less expected findings.

The value of the Evidence Rating system

The ER values provide a rapid means of assessing the certainty of the functional assignment for each gene. Four proteins from *M. burtonii* were assigned ER1 as they have been experimentally characterized; elongation factor 2 (Mbur_1171) [5], two proteins involved in formation of the compatible solute glycosylglycerate, glucosyl-3-phosphoglycerate synthase (Mbur_0737) [6] and glucosyl-3-phosphoglycerate phosphatase (Mbur_0736) [6], and a protein with a cold shock domain that complements a cold sensitive defect in *E. coli* (Mbur_1438) [7] and appears to form part of the archaeal exosome [29]. Forty-two percent of proteins in the *M. burtonii* genome had strong (ER2) or moderate (ER3) similarity to experimentally characterized proteins from other organisms, and 32% of the proteins had similarity to protein domains or families but no experimentally characterized homolog (ER4) (Table SI1). Of the 1478

genes assigned a function by the auto-annotation pipeline, the certainty of functional assignments was greatly clarified. Three of the genes have established functions (ER1), 570 have high confidence of function (ER2), 400 have a moderate level of confidence, and the functions assigned to 478 are tentative (ER4). Twenty-seven of the 1478 contained no evidence of function and were reclassified at hypothetical (ER5).

The original auto-annotation listed 953 genes (39% of the genome) with no predicted function (hypothetical). This was reduced to 25% (ER5) after manual annotation. Functional assignments with moderate or greater certainty were assigned to 65 of the hypothetical genes (ER1 + ER2 + ER3, line 2 in Table SI1), and tentative functions for a further 299 (ER4, line 2 in Table SI1). The reassignment of hypothetical proteins was in part facilitated by applying protein threading [30, 31] to the 953 hypothetical proteins. A total of 17 proteins had a threading z-score above 8, indicating similarity to known proteins at the superfamily/fold level or better. Manual annotation had already assigned a function and rating of ER4 or higher to 13 of these proteins, and the remaining 4 were reassigned to ER4 on the basis of the threading results (Mbur_0102, Mbur_1050, Mbur_1069 and Mbur_1971).

The ER values revealed major differences in the level of understanding about functional processes represented by different COG categories (Figure SI3). The nucleotide transport and metabolism COG category [F] contains the greatest proportion of well characterized genes (75% ER2), followed by the translation, ribosomal structure and biogenesis [J], amino acid transport and metabolism [E], and coenzyme transport and metabolism [H] categories. Other COG categories where the majority of genes had a characterized homolog (at least 75% of genes in ER2 or ER3) include lipid transport and metabolism [I], inorganic ion transport and metabolism [P], secondary metabolites biosynthesis, transport and catabolism [Q], cell motility [N], and carbohydrate transport and metabolism [G]. Confident functions (ER1, ER2 or ER3) were also assigned to 38% of genes (90 genes) in the general function prediction only [R] category. Achieving this greatly improves the overall extent of functional characterization of genes in the genome.

In addition to category [S] (function unknown), only a small proportion of confident functional predictions (75% ER4) exist in the signal transduction mechanisms [T] category. This suggests that the genes involved in signal transduction have not been well characterized, not only in *M. burtonii*, but also in other archaea. The transcription [K] and posttranslational modification, protein turnover and chaperones [O] categories also had a high proportion of genes with ER4 values (45% and 41%, respectively). Experimental analyses of genes from these COG categories (S, T, K and O) provide the best chance for clarifying gene functions, thereby advancing the understanding of archaeal biology.

Identifying incorrect and omitted genes

Manual annotation uncovered several genome errors including missing tRNA genes and incorrectly identified pseudogenes; defined as genes corresponding to <30% of a COG hit, or with more than one internal stop codon or frameshift (http://imgweb.jgi-psf.org/cgi-bin/archaeal_qa/main.cgi). Auto-annotation listed 49 tRNA genes, with the notable absence of tRNA-trp or tRNA-pyl. Using ARAGORN [32] and tRNAscan-SE [33] a further 3 potential tRNA genes were identified, including the missing tRNA-trp and two additional tRNA-ser genes. The tRNA-pyl could not be identified using these programs, but was subsequently identified with reference to the literature [34]. No selenocysteine tRNA was found, however the archaeal SELB protein homolog (Mbur_1757) was present and may play a role in pyrrolysine rather than selenocysteine incorporation [35].

During the initial curation 158 CDSs were annotated as pseudogenes. Manual annotation provided a functional annotation (ER4 or higher) for 91 of these (Table SI3). Previous proteomics analyses shows that 30 of the 158 CDSs are expressed in the cell ([36-38] and D. Burg and T. Williams, unpublished results; a total of 1023 proteins detected by proteomics are identified in Table SI5). These include genes involved in diverse cellular processes (*e.g.* metabolism, replication, transposases) and that have a range of ER values (ER2, 6; ER3, 7; ER4, 11; ER5, 6). The ER2 genes were archaeal DNA polymerase II small subunit, $F_{420}H_2$ dehydrogenase subunit J and four pyrrolysine-containing methyltransferases. Synthesis of methylamine methyltransferases has been shown to involve the incorporation of the amino acid pyrrolysine at an amber stop codon [39, 40], and proteomics data supporting its incorporation in a trimethylamine methyltransferase from *M. burtonii* has been reported [36]. The lack of recognition of pyrrolysine in auto-annotation pipelines provides an explanation for the incorrect assignment of the methyltransferases as

pseudogenes. However, the reason for the assignment of the majority of genes as pseudogenes was not apparent.

Despite the knowledge that errors within automated annotations may be up to 49% of the genome [11], for many purposes, such as searching for the presence of a gene or for comparative genomics, users of genome data are likely to assume that the bulk of annotations in any given genome sequence are equally valid and may equate to ER2. The fact that this is a bad assumption, is well illustrated by the fact that only 319 protein encoding genes (13% of the genome) initially given a functional assignment retained the same annotation wording and were given ER2 (Table SI1). While 522 genes (21%) retained their hypothetical protein (ER5) status, the remaining 66% of genes were altered with either an improvement in wording accuracy or clarification of the level of confidence associated with the annotation, or both (Table SI1).

The ER system ensures it is readily apparent whether the gene itself has been characterized, if it is a homolog of a characterized protein, or if it shares only a single domain in common with a protein of interest, or is truly hypothetical. This will facilitate the choosing of gene/protein targets (*e.g.* high likelihood of target having function of interest *vs* target clearly a hypothetical protein) for a broad range of experimental (*e.g.* pathway analysis, adaptation studies) and computational studies (*e.g.* gene context analysis to infer the function of hypothetical proteins [30, 41]).

The manual annotation of the *M. burtonii* genome has revealed the extent to which an auto-annotated genome can be improved to provide a substantially better interpretation of the data, and we heartily endorse the provision and ongoing development of user-friendly manual curation websites like IMG-ER which facilitate such efforts. The annotations are as accurate as possible based on data available in databases and the literature as of March 2008. This now serves as a reference date for iterative functional annotations by our research group which will be made accessible via the IMG-ER website, and for others to perform analyses targeting genes of interest or groups of genes with a defined level of understanding about their function (*i.e.* from established function, ER1, to hypothetical, ER5). Our experience with *M. burtonii* suggests that an average sized microbial genome of 3000 genes could be manually annotated by a team of 20 researchers in ~2400 hours; on average, ~120 hours per person. This is achievable by an individual group and collaborators, or by teams of dedicated specialists chosen for their expertise in specific areas that cover the breadth of genome biology.

References

1. R.D. Fleischmann, et al., Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, Science 269 (1995) 496-512.
2. J. Lederberg, *E. coli* K-12, Microbiology Today 31 (2004) 116.
3. P.D. Karp, et al., Multidimensional annotation of the *Escherichia coli* K-12 genome, Nucleic Acids Res. 35 (2007) 7577-7590.
4. R. Cavicchioli, Cold-adapted archaea, Nature Reviews Microbiology 4 (2006) 331-343.
5. T. Thomas, R. Cavicchioli, Effect of temperature on stability and activity of elongation factor 2 proteins from Antarctic and thermophilic methanogens, J. Bacteriol. 182 (2000) 1328-1332.
6. J. Costa, et al., Characterization of the biosynthetic pathway of glucosylglycerate in the archaeon *Methanococcoides burtonii*, J. Bacteriol. 188 (2006) 1022-1030.
7. L. Giaquinto, et al., Structure and function of cold shock proteins in Archaea, J. Bacteriol. 189 (2007) 5738-5748.
8. V.M. Markowitz, et al., The integrated microbial genomes (IMG) system, Nucleic Acids Res. 34 (2006) D344-D348.
9. V.M. Markowitz, et al., The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions, Nucleic Acids Res. 36 (2008) D528-D533.
10. D. Vallenet, et al., MaGe: a microbial genome annotation system supported by synteny results, Nucleic Acids Res. 34 (2006) 53-65.
11. C.E. Jones, A.L. Brown, U. Baumann, Estimating the annotation error rate of curated GO database sequence annotations, BMC Bioinformatics 8 (2007) 170.
12. C. Andorf, D. Dobbs, V. Honavar, Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach, BMC Bioinformatics 8 (2007) 284.

13. W.R. Gilks, B. Audit, D. De Angelis, S. Tsoka, C.A. Ouzounis, Modeling the percolation of annotation errors in a database of protein sequences, Bioinformatics 18 (2002) 1641-1649.
14. M.Y. Galperin, E.V. Koonin, Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption, In Silico Biol. 1 (1998) 55-67.
15. C.H. Wu, H. Huang, L.-S.L. Yeh, W.C. Barker, Protein family classification and functional annotation, Comp. Biol. Chem. 27 (2003) 37-47.
16. K. Bryson, et al., AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system, Nucleic Acids Res. 34 (2006) 3533-3545.
17. B.J. Haas, et al., Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols and the final release, BMC Biology 3 (2005) 7.
18. C.G. Elsik, et al., Community annotation: procedures, protocols and supporting tools, Genome Res. 16 (2006) 1329-1333.
19. S.V. Angiuoli, et al., Toward an online repository of standard operating procedures (SOPs) for (meta)genomic annotation, OMICS 12 (2008) 137-141.
20. K.-H. Gartemann, et al., The genome sequence of the tomato-pathogenic Actinomycete *Clavibacter michiganensis* subsp. michiganensis NCPPB382 reveals a large island involved in pathogenicity, J. Bacteriol. 190 (2008) 2138-2149.
21. I.V. Tetko, et al., MIPS bacterial genomes functional annotation benchmark dataset, Bioinformatics 21 (2005) 2520-2521.
22. I.G. Boneca, et al., A revised annotation and comparative analysis of *Helicobacter pylori* genomes, Nucleic Acids Res. 31 (2003) 1704-1714.
23. C. Medigue, et al., Coping with cold: The genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125, Genome Res. 15 (2005) 1325-1335.
24. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403-410.
25. A. Gattiker, et al., Automated annotation of microbial proteomes in SWISS-PROT, Comp. Biol. Chem. 27 (2003) 49-58.
26. A. Hecker, et al., An archaeal orthologue of the universal protein Kae1 is an iron metalloprotein which exhibits atypical DNA-binding properties and apurinic-endonuclease activity in vitro, Nucleic Acids Res. 35 (2007) 6042-6051.
27. T. Sato, H. Atomi, T. Imanaka, Archaeal type III RuBisCOs function in a pathway for AMP metabolism, Science 315 (2007) 1003-1006.
28. J.R. Roth, J.G. Lawrence, M. Rubenfield, S. Kieffer-Higgins, G.M. Church, Characterization of the cobalamin (Vit B_{12}) biosynthetic genes of *Salmonella typhimurium*, J. Bacteriol. 175 (1993) 3303-3316.
29. E.V. Koonin, Y.I. Wolf, L. Aravind, Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach, Genome Res. 11 (2001) 240-252.
30. N.F.W. Saunders, et al., Predicted roles for hypothetical proteins in the low-temperature expressed proteome of the Antarctic archaeon *Methanoccoides burtonii*, J. Prot. Res. 4 (2005) 464-472.
31. N.F.W. Saunders, et al., Mechanisms of thermal adaptation revealed from the genomes of the Antarctic archaea *Methanogenium frigidum* and *Methanococcoides burtonii*, Genome Res. 13 (2003) 1580-1588.
32. D. Laslett, B. Canback, ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences, Nucleic Acids Res. 32 (2004) 11-16.
33. T. Lowe, S.R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, Nucleic Acids Res. 25 (1997) 955-964.
34. J.A. Krzycki, The direct genetic encoding of pyrrolysine, Curr. Opin. Microbiol. 8 (2005) 706-712.
35. M. Ibba, D. Söll, Aminoacyl-tRNAs: setting the limits of the genetic code, Genes. Dev. 18 (2004) 731-738.
36. A. Goodchild, et al., A proteomic determination of cold adaptation in the Antarctic archaeon, *Methanococcoides burtonii*, Mol. Microbiol. 53 (2004) 309-321.

37. A. Goodchild, M. Raftery, N.F.W. Saunders, M. Guilhaus, R. Cavicchioli, Cold adaptation of the Antarctic archaeon, *Methanococcoides burtonii* assessed by proteomics using ICAT, J. Prot. Res. 4 (2005) 473-480.
38. A. Goodchild, M. Raftery, N.F.W. Saunders, M. Guilhaus, R. Cavicchioli, Biology of the cold adapted archaeon, *Methanococcoides burtonii* determined by proteomics using liquid chromatography-tandem mass spectrometry, J. Prot. Res. 3 (2004) 1164-1176.
39. D.G. Longstaff, et al., A natural genetic code expansion cassette enables transmissible biosynthesis and genetic encoding of pyrrolysine, PNAS USA 104 (2007) 1021-1026.
40. L. Paul, D.J. Ferguson Jr., J.A. Krzycki, The trimethylamine methyltransferase gene and multiple dimethylamine methyltransferase genes of *Methanosarcina barkeri* contain in-frame and read-through amber codons, J. Bacteriol. 182 (2000) 2520-2529.
41. L.L. Grochowski, H. Xu, R.H. White, Identification and characterization of the 2-phospho-L-lactate guanylyltransferase involved in coenzyme F420 biosynthesis, Biochemistry 47 (2008) 3033-3037.
42. P.D. Franzmann, N. Springer, W. Ludwig, E. Conway de Macario, M. Rohde, A methanogenic archaeon from Ace Lake, Antarctica: *Methanococcoides burtonii* sp. nov., Syst. Appl .Microbiol. 15 (1992) 573-581.
43. B. Ewing, P. Green, Base-calling of automated sequencer traces using phred. II. Error probabilities, Genome Res. 8 (1998) 186-194.
44. B, Ewing, L. Hillier, M.C. Wendl, P. Green, Base-calling of automated sequencer traces using phred. I. Accuracy assessment, Genome Res. 8 (1998) 175-185.
45. D. Gordon, C. Abajian, P. Green, Consed: a graphical tool for sequence finishing, Genome Res. 8 (1998) 195-202.
46. J.H. Badger, G.J. Olsen, CRITICA: Coding region identification tool invoking comparative analysis, Mol. Biol. Evol. 16 (1999) 512-524.
47. A.L. Delcher, D. Harmon, S. Kasif, O. White, S.L. Salzberg, Improved microbial gene identification with GLIMMER, Nucleic Acids Res. 27 (1999) 4636-4641.
48. M.G. Klotz, et al., Complete genome sequence of the marine, chemolithoautotrophic, ammonia-oxidizing bacterium *Nitrosococcus oceani* ATCC 19707, Appl. Environ. Microbiol. 72 (2006) 6299-6315.
49. D.A. Rodionov, A.G. Vitreschak, A.A. Mironov, M.S. Gelfand, Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes, J. Biol. Chem. 278 (2003) 41148-41159.
50. C.A. Roessner, A.I. Scott, Fine-tuning our knowledge of the anaerobic route to cobalamin (Vitamin B12), J. Bacteriol. 188 (2006) 7331-7334.

Figure SI1. Evidence Rating System. ER1 indicates that the protein from *M. burtonii* had been experimentally characterized (a self match); ER2, the most closely related functionally-characterized homolog is not from *M. burtonii* but the BLAST alignments share ≥ 35% sequence identity along the entire length of the protein; ER3, the most closely related functionally-characterized homolog shares <35% sequence identity along the length of the protein, but all required motifs/domains for function are present and complete; ER4, an experimentally characterized full-length homolog is not available but conserved protein motifs or domains can be identified; ER5 (hypothetical protein), no functionally characterized homolog can be found, and no characterized protein domains above the Pfam and InterProScan cut-off thresholds can be identified. tRNA and rRNA genes were processed in an analogous way, and assigned ER2.

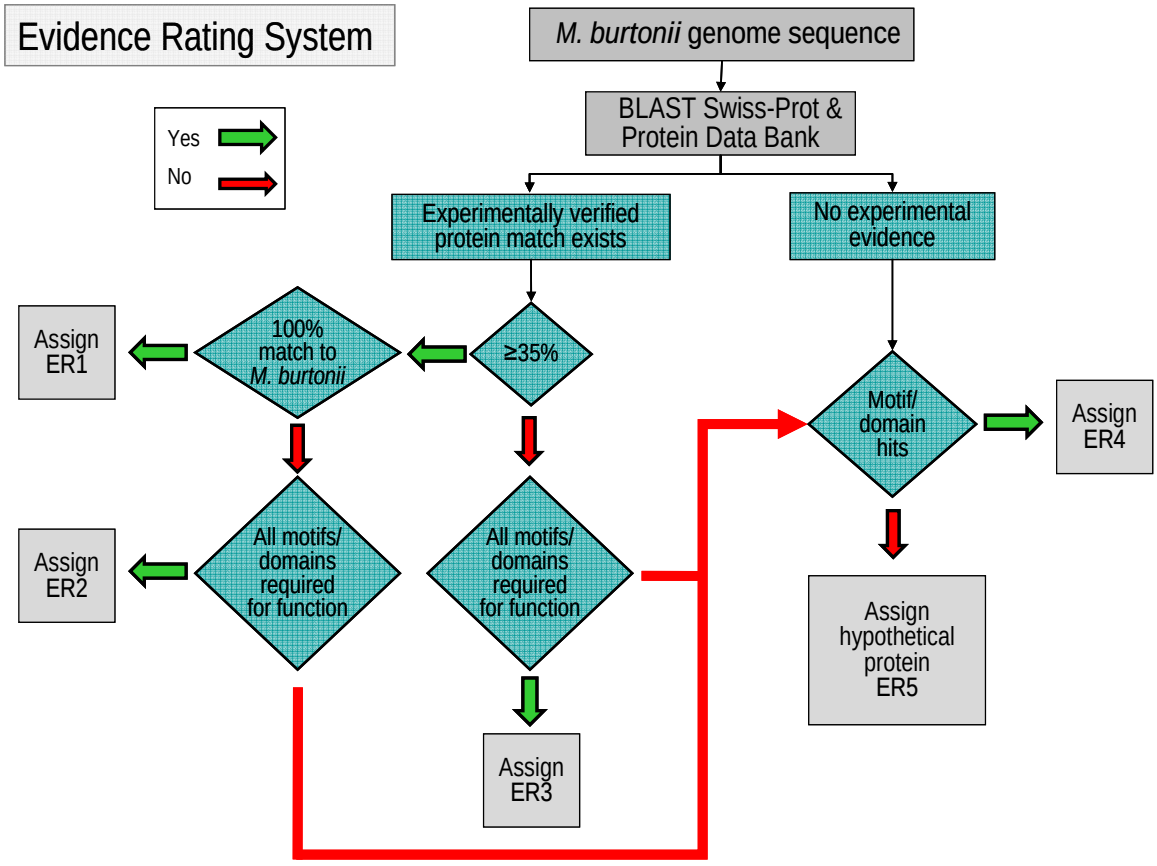


Figure SI2. Vitamin B12 biosynthesis pathway from auto- and manually-annotated genomes. Inferred adenosylcobalamin biosynthesis pathway in *M. burtonii* based on initial annotations (A) and after manual curation (B). Figure modified from [49, 50]. Genes in the aerobic pathway are named according to *Pseudomonas denitrificans*, all others are named according to *Salmonella enterica* serovar Typhimurium except where archaeal alternatives are known. 5,6-dimethylbenzimidazole (DMB), niacinamide mononucleotide (NMN).

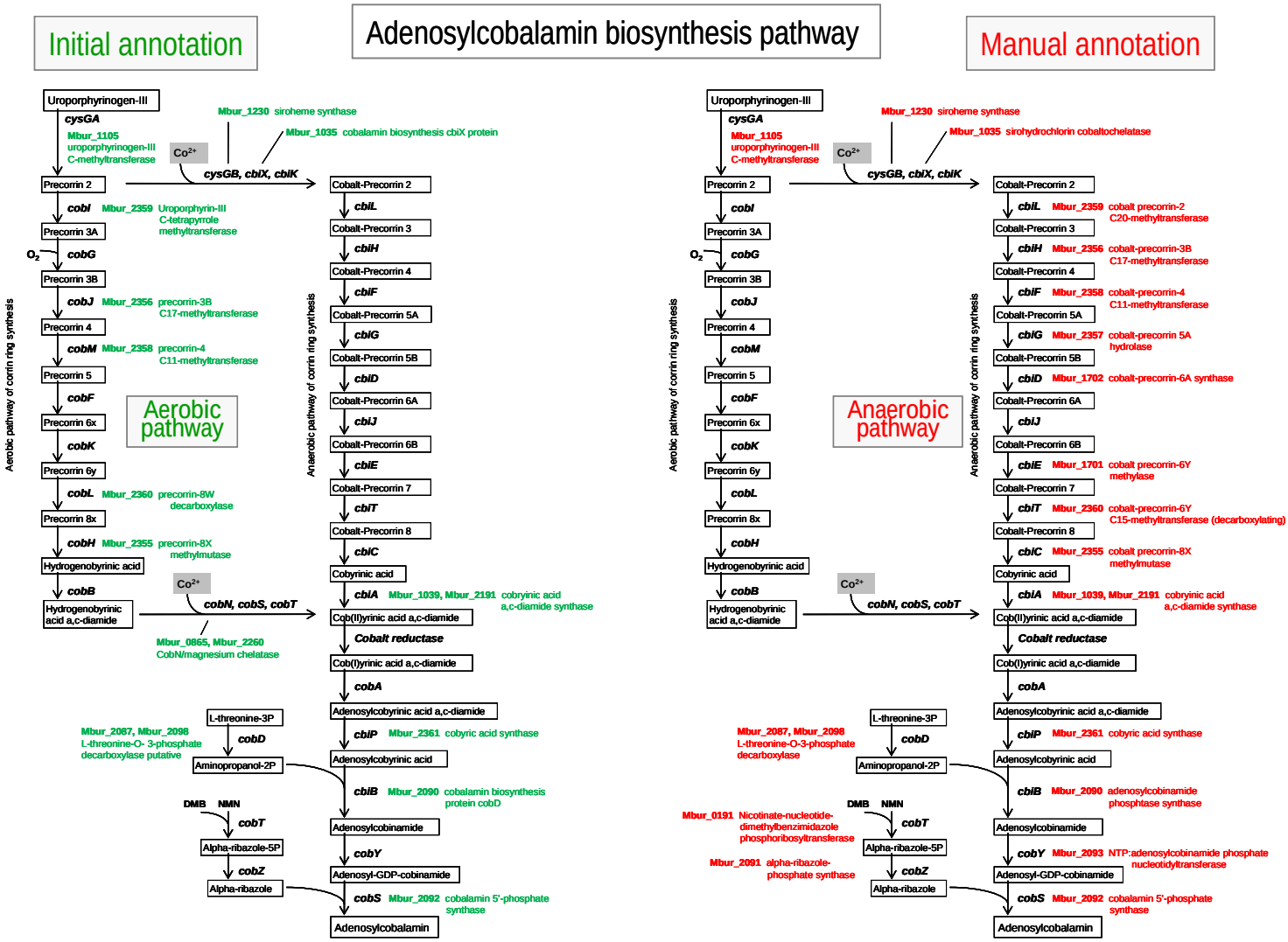


Figure SI3. Distribution of Evidence Rating values across COG categories. COG Categories are: Amino acid transport and metabolism [E]; Carbohydrate transport and metabolism [G]; Cell cycle control, cell division, chromosome partitioning [D]; Cell motility [N]; Cell wall/membrane/envelope biogenesis [M]; Chromatin structure and dynamics [B]; Coenzyme transport and metabolism [H]; Defense mechanisms [V]; Energy production and conversion [C]; Function unknown [S]; General function prediction only [R]; Inorganic ion transport and metabolism [P]; Intracellular trafficking, secretion, and vesicular transport [U]; Lipid transport and metabolism [I]; Nucleotide transport and metabolism [F]; Posttranslational modification, protein turnover and chaperones [O]; RNA processing and modification [A]; Replication, recombination and repair [L]; Secondary metabolites biosynthesis, transport and catabolism [Q]; Signal transduction mechanisms [T]; Transcription [K]; Translation, ribosomal structure and biogenesis [J].

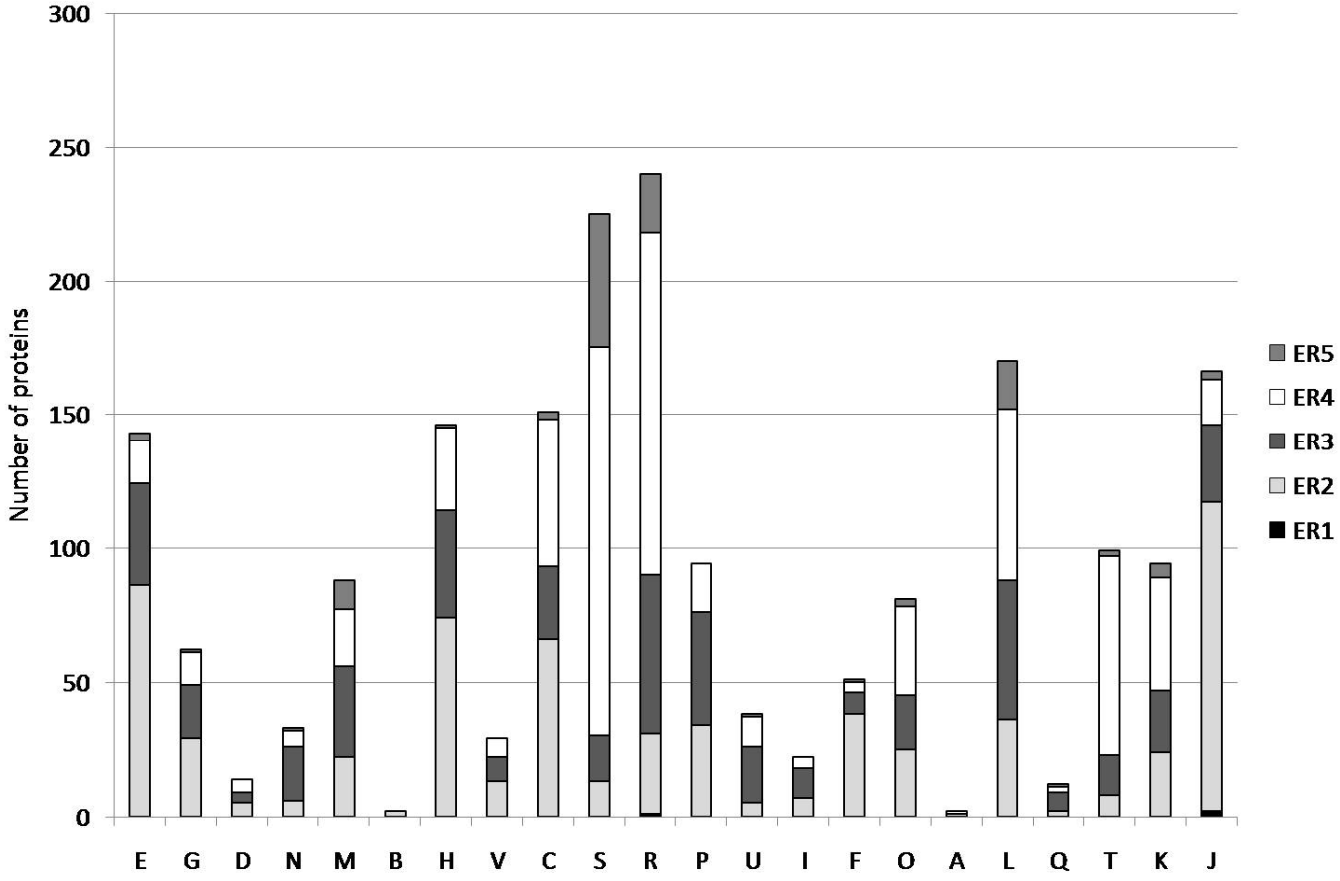


Table SI1. Annotation wording changes and evidence rating distribution of *M. burtonii* proteins.

	Total	Evidence rating distribution				
		ER1	ER2	ER3	ER4	ER5
Proteins with function prediction by automated pipeline	1478	3	570	400	478	27
Proteins without function prediction by automated pipeline	953	1	26	38	299	589
No change in wording	1372	1	319	235	295	522
More specific annotation wording	725	3	262	153	240	67
Less specific annotation wording	334	0	15	50	242	27
All *M. burtonii* proteins	2431	4	596	438	777	616

Table SI2. Examples of changes in annotation wording

	Locus Tag	Former annotation from pipeline	New annotation	Comments
No change	Mbur_2416	gamma-glutamyl phosphate reductase	Glutamate-5-semialdehyde dehydrogenase (proA) (EC 1.2.1.41) (ER2)	semantic change to the Enzyme Commission's accepted protein name
	Mbur_0777	FeoA	Ferrous iron transport protein A (feoA) (ER3)	no change, but more complete and informative name used
	Mbur_0053	hypothetical protein	hypothetical protein (ER5)	no change
More specific	Mbur_0711	Isopropylmalate/citramalate/homocitrate synthase	(R)-citramalate synthase (cimA) (EC 2.3.3.-) (ER2)	The first annotation includes several functions, however the evidence points to only one of these
	Mbur_1229	Shikimate/quinate 5-dehydrogenase	glutamyl-tRNA reductase (EC 1.2.1.70) (ER2)	completely different, specific annotation
	Mbur_2446	methyltransferase	SAM-dependent methyltransferase (ER3)	Cofactor usage specificity was determined
	Mbur_0796	pyridoxal phosphate-dependent enzyme	Sep-tRNA:Cys-tRNA synthase (EC 2.5.1.-) (ER3)	specific pyridoxal-phosphate-binding enzyme was identified
	Mbur_0906	hypothetical protein	THUMP-domain containing protein (ER4)	a domain with putative activity was identified
	Mbur_1929	Phosphoadenosine phosphosulfate reductase	Phosphoadenosine phosphosulfate reductase fused to RNA-binding PUA and 4Fe-4S binding domains (ER4)	clarification of multiple domains not usually found together - suggests reductase activity not likely
	Mbur_1291	(pseudogene)	F420H2 dehydrogenase subunit J (ER2)	Initially no annotation provided at all
Less Specific	Mbur_2133	ABC nitrate/sulfonate/bicarbonate transporter, ATPase subunit	ABC transporter ATPase subunit (ER2)	Evidence of specific substrate for ABC transporter not present, hence broadened to wider, family category
	Mbur_0336	tungsten formylmethanofuran dehydrogenase subunit B	Formylmethanofuran dehydrogenase subunit B (EC 1.2.99.5) (ER3)	evidence of which metal cofactor is used was not available
	Mbur_1805	phosphoesterase, RecJ-like protein	Nucleic acid binding protein (ER3)	evidence exists only for nucleic acid binding function, not phosphoesterase activity
	Mbur_1754	5-methyltetra-hydropteroyltriglutamate--homocysteine methyltransferase	hypothetical protein (ER5)	no functional or domain evidence

Table SI3. Pseudogene annotations

Evidence rating	Number of genes	Number Expressed
ER2	11	6
ER3	21	7
ER4	59	11
ER5	67	6
Total	158	30 (19%)

Table SI4. Proteins with substantially different manual annotations compared to the initial auto-annotation. See excel file with same name.

Table SI5. *M. burtonii* full list of genes and annotations. See excel file with same name. Note that the data is accessible in searchable format via the IMG website. Annotation data have also been submitted to the NCBI GenBank database under accession number XXXXXXXX.