The complete genome sequence of *Moorella thermoacetica* (f. *Clostridium thermoaceticum*)

Elizabeth Pierce,[1] Gary Xie,[2,3] Ravi D. Barabote,[2,3]
Elizabeth Saunders,[2,3] Cliff S. Han,[2,3]
John C. Detter,[2,3] Paul Richardson,[2,4]
Thomas S. Brettin,[2,3] Amaresh Das,[5]
Lars G. Ljungdahl[5] and Stephen W. Ragsdale[1]*

[1]*Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA.*

[2]*Los Alamos National Laboratory, Bioscience Division, Los Alamos, NM, USA.*

[3]*Department of Energy Joint Genome Institute, Walnut Creek, CA, USA.*

[4]*Lawrence Berkeley National Laboratory, Berkeley, CA, USA.*

[5]*Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA.*

Summary

This paper describes the genome sequence of *Moorella thermoacetica* (f. *Clostridium thermoaceticum*), which is the model acetogenic bacterium that has been widely used for elucidating the Wood–Ljungdahl pathway of CO and CO_2 fixation. This pathway, which is also known as the reductive acetyl-CoA pathway, allows acetogenic (often called homoacetogenic) bacteria to convert glucose stoichiometrically into 3 mol of acetate and to grow autotrophically using H_2 and CO as electron donors and CO_2 as an electron acceptor. Methanogenic archaea use this pathway in reverse to grow by converting acetate into methane and CO_2. Acetogenic bacteria also couple the Wood–Ljungdahl pathway to a variety of other pathways to allow the metabolism of a wide variety of carbon sources and electron donors (sugars, carboxylic acids, alcohols and aromatic compounds) and electron acceptors (CO_2, nitrate, nitrite, thiosulfate, dimethylsulfoxide and aromatic carboxyl groups). The genome consists of a single circular 2 628 784 bp chromosome encoding 2615 open reading frames (ORFs), which includes 2523 predicted protein-encoding genes. Of these, 1834 genes (70.13%) have been assigned tentative functions, 665 (25.43%) matched genes of unknown function, and the remaining 24 (0.92%) had no database match. A total of 2384 (91.17%) of the ORFs in the *M. thermoacetica* genome can be grouped in ortholog clusters. This first genome sequence of an acetogenic bacterium provides important information related to how acetogens engage their extreme metabolic diversity by switching among different carbon substrates and electron donors/acceptors and how they conserve energy by anaerobic respiration. Our genome analysis indicates that the key genetic trait for homoacetogenesis is the core *acs* gene cluster of the Wood–Ljungdahl pathway.

Introduction

Here we report the first complete genome sequence of an acetogenic bacterium. Acetogens are obligately anaerobic bacteria that use a mechanism of anaerobic CO and CO_2 fixation, called the reductive acetyl-CoA or the Wood–Ljungdahl pathway as their main mechanism for energy conservation and for synthesis of acetyl-CoA and cell carbon from CO_2 (Drake *et al.*, 1994). Other terms to describe an organism using this type of metabolism are 'homoacetogen' and 'CO_2-reducing acetogen'. We determined the genome sequence of *Moorella thermoacetica*, originally named *Clostridium thermoaceticum*, which has been the model acetogen since its isolation in 1942 (Fontaine *et al.*, 1942) for elucidating the Wood–Ljungdahl pathway. As discussed below, this pathway is found in a broad range of phylogenetic classes, and is used in both the oxidative and reductive directions. Besides being used for energy production and autotrophic carbon assimilation in acetogens (Ljungdahl, 1994; Ragsdale, 1997; Drake *et al.*, 2002), the Wood–Ljungdahl pathway is used for acetate catabolism by sulfate-reducing bacteria (Schauder *et al.*, 1988; Spormann and Thauer, 1988) and aceticlastic methanogens (Ferry, 1999) and for CO_2 fixation by hydrogenotrophic methanogens (Stupperich *et al.*, 1983; Ladapo and Whitman, 1990).

As early as 1932, organisms were discovered that could convert H_2 and CO_2 into acetic acid (Fischer *et al.*, 1932). In 1936, Wieringa reported the first acetogenic bacterium, *Clostridium aceticum* (Wieringa, 1936; 1940), which was reisolated in 1981 (Braun *et al.*, 1981). *Moorella thermoacetica* (Collins *et al.*, 1994), a

clostridium in the *Thermoanaerobacteriaceae* family, attracted wide interest when it was isolated because of its unusual ability to convert glucose almost stoichiometrically to 3 mol of acetic acid (Fontaine *et al.*, 1942). It took many years to uncover the biochemical reactions responsible for this homoacetogenic pathway. The novelty of this pathway was recognized with the discovery that growth of *M. thermoacetica* on glucose involves incorporation of two molecules of CO_2 directly into acetic acid (Barker and Kamen, 1945; Wood, 1952). The isolation of *Acetobacterium woodii* in 1977 and the description of methods to isolate organisms that can grow on H_2 and CO_2 (Balch *et al.*, 1977) led to the isolation of many acetogens that can grow autotrophically, and in 1983 *M. thermoacetica* was shown to grow on H_2/CO_2 (Kerby and Zeikus, 1983). The identification of a hydrogenase in 1982 provided a biochemical basis for autotrophic growth of acetogens on hydrogen (Drake, 1982). Subsequently, acetogens have been shown to use a wide variety of carbon sources and electron donors and acceptors. *Moorella thermoacetica* is a versatile heterotroph, growing on sugars, two-carbon compounds (glyoxylate, glycolate and oxalate), lactate, pyruvate, short-chain fatty acids and methoxylated aromatic compounds. Besides CO_2, electron acceptors include nitrate, nitrite, thiosulfate and dimethylsulfoxide.

Acetogenic bacteria play an important role in the global carbon cycle, producing over 10^{12} kg (10 billion US tons) of acetic acid on a huge scale annually (Wood and Ljungdahl, 1991; Drake *et al.*, 1994), which dwarfs the total output of acetate (an important chemical feedstock) by the world's chemical industry (~10^{10} kg year^{-1}). While most acetogens like *M. thermoacetica* are in the phylum *Firmicutes*, acetogens include *Spirochaetes* (Leadbetter *et al.*, 1999; Salmassi and Leadbetter, 2003), *Deltaproteobacteria* like *Desulfotignum phosphitoxidans* (Schink *et al.*, 2002), and *Acidobacteria* like *Holophaga foetida* (Liesack *et al.*, 1994). Important in the biology of the soil, lakes and oceans, acetogens have been isolated from diverse environments, including the GI tracts of animals and termites (Breznak, 1994; Mackie and Bryant, 1994), rice paddy soils (Rosencrantz *et al.*, 1999), hypersaline waters (Ollivier *et al.*, 1994), surface soils (Peters and Conrad, 1995; Kusel *et al.*, 1999) and deep subsurface sediments (Liu and Suflita, 1993). Like methanogens, acetogens act as a H_2 sink, depleting H_2 that is generated in anaerobic environments during the natural biodegradation of organic compounds. As the build-up of H_2 inhibits biodegradation by creating an unfavourable thermodynamic equilibrium, acetogens enhance biodegradative capacity by coupling the oxidation of hydrogen gas to the reduction of CO_2 to acetate. Acetogens also have been found in a methanogenic mixed population from an army ammunition manufacturing plant waste water treatment facility (Adrian and Arnett, 2004) and a dechlorinating community that has been enriched for bioremediation (Macbeth *et al.*, 2004).

For organisms that house acetogens in their digestive systems, like humans, termites and ruminants (Greening and Leedle, 1989; Tholen and Brune, 1999; Chassard and Bernalier-Donadille, 2006), the acetate generated by microbial metabolism is a beneficial nutrient for the host and for other microbes within the community. In these ecosystems, acetogens can compete directly with hydrogenotrophic methanogenic archaea, or interact syntrophically with acetotrophic methanogens that use H_2 and CO_2 to produce methane (Chassard and Bernalier-Donadille, 2006). Organisms containing the *acs* gene cluster are also enriched in the marine and terrestrial vertebrate and invertebrate gut biomes (Gill *et al.*, 2006; Warnecke *et al.*, 2007). In the termite gut, acetogens are the dominant hydrogen sinks (Pester and Brune, 2007), and it has been proposed that acetate is the major energy source for the termite (Odelson and Breznak, 1983). However, methanogens are the dominant hydrogenotrophs in many environments as methanogens have a lower threshold for H_2 than acetogens (Le Van *et al.*, 1998) and as the energy yield from the conversion of CO_2 and H_2 to methane is greater than that for conversion to acetate (Schink, 1997). Under such conditions, acetogens must often resort to other metabolic pathways for growth and, thus, have a highly diverse metabolic menu that comprises the biodegradation products of most natural polymers like cellulose and lignin, including sugars, alcohols, organic acids and aldehydes, aromatic compounds and inorganic gases like CO, H_2 and CO_2. They also can use a variety of electron acceptors, e.g. CO_2, nitrate, dimethylsulfoxide, fumarate and protons. The *M. thermoacetica* genome sequence described here provides important information related to how acetogens engage this extreme metabolic diversity and how they switch with such facility among different carbon substrates and electron donors and acceptors.

Results

General genome features

The major features of the *M. thermoacetica* genome are listed in Table 1 and shown in Fig. S1. The genome consists of a single circular 2 628 784 bp chromosome that has a GC content of 56% (Fig. S1). Base pair one of the chromosome was assigned within the putative origin of replication. The genome contains 51 tRNAs, three rRNAs (5S, 16S and 23S) and 38 other RNAs predicted by the Rfam. Analysis of 16S rRNA sequences confirmed that *M. thermoacetica* belongs to the phylum *Clostridia* (Fig. 1) and to the class *Thermoanaerobacteriaceae*. Figure 1 also highlights the known homoacetogenic microbes, revealing the wide distribution of phylogenetic classes in

Table 1. General features of the *M. thermoacetica* genome.

Genome size (base pairs)	2628784
G+C content	0.56
Open reading frames	2615
Predicted protein encoding sequences	2523
Predicted sequences encoding RNA genes	92
Coding density	97.90%
rRNA	3
tRNA	51
Genes with function prediction (percentage)	1834 (71.17%)
Number of genes in COG (percentage)	2008 (76.79%)
Number of genes in Pfam (percentage)	2001 (76.52%)
Genes encoding signal peptides (percentage)	159 (6.08%)
Genes encoding transmembrane proteins (percentage)	570 (21.8%)
Genes with orthologues (percentage)	2326 (90.26%)
Genes with paralogues (percentage)	1361 (52.81%)

which this trait appears, clearly indicating that acetogenesis is a metabolic, not a phylogenetic trait. Besides the bacteria shown in Fig. 1, which are related to *M. thermoacetica*, the Wood–Ljungdahl pathway is also found in archaea. On the other hand, homoacetogenesis is a rare occurrence within a particular phylum. For example, two relatives of *M. thermoacetica* in the *Thermoanaerobacteriaceae* family whose genomes have been sequenced, *Thermoanaerobacter tengcongensis* (Bao *et al.*, 2002) and *Thermoanaerobacter ethanolicus* (draft sequence available at JGI), can also grow anaerobically on several mono- and disaccharides and starch (Wiegel and Ljungdahl, 1981); however, among these three strains, only *M. thermoacetica* is homoacetogenic, and only *M. thermoacetica* has acetyl coenzyme A synthase (ACS), a key enzyme in the Wood–Ljungdahl pathway. Fermentation of glucose by *T. tengcongensis* or *T. ethanolicus* yields ethanol and/or acetate, 1.5–2 mol of CO_2 and small amounts of H_2. Thus, the genomic results clearly indicate that the trait of acetogenesis was distributed by horizontal transfer of the core genes of the Wood–Ljungdahl pathway.

The *M. thermoacetica* genome contains 2523 predicted protein-encoding genes, of which 1834 (70.13%) have been assigned tentative functions. Six hundred sixty-five predicted proteins (25.43%) matched proteins of unknown function, and the remaining 24 (0.92%) had no database match. A total of 429 paralogous protein families were identified in *M. thermoacetica* strain ATCC 39073, containing 1307 predicted proteins (49.98% of the total). A total of 2384 (91.17%) of the open reading frames (ORFs) in the *M. thermoacetica* genome can be grouped in orthologue clusters. Fifty-eight pseudogenes were identified in the *M. thermoacetica* genome by the IMG annotation. A complete list of pseudogenes and gene remnants is provided in Table S1. Analysis of these genes in *M. thermoacetica* indicates that 41 out of 58 are redundant genes, with paralogues in the *M. thermoacetica*

genome. The 17 that do not have paralogues encode either putative or hypothetical proteins. These results suggest that elimination of the functions of these redundant genes may be tenuously ascribed to selective pressures.

The Wood–Ljungdahl pathway genes

Figure 2 shows the genes encoding the key enzymes in the Wood–Ljungdahl pathway of CO_2 fixation in which two molecules of CO_2 are reduced to a methyl and a carbonyl group, and are joined with coenzyme A (CoA) to make acetyl-CoA. Acetyl-CoA is then incorporated into cell carbon or converted to acetyl phosphate, whose phosphoryl group is transferred to ADP to generate ATP and acetate, the main growth product of acetogenic bacteria (for reviews see Ljungdahl, 1986; Drake and Daniel, 1997; Ragsdale, 1997). This pathway has been described to consist of an Eastern (blue colouring scheme) and a Western (red) branch (Fig. 2; Ragsdale, 1997).

Genes encoding the Eastern branch of the Wood–Ljungdahl pathway

The genes encoding enzymes in the Eastern branch of the pathway that are involved in the conversion of CO_2 to methyl-H_4folate (Fig. 2) are common to most organisms, though they are expressed at much higher levels in acetogens. As shown in Fig. 3, many of the genes encoding functions of the Eastern branch of the pathway are dispersed on the genome at locations distant from the core genes of the Wood–Ljungdahl pathway, which are involved in converting methyl-H_4folate, CoA and CO_2 to acetyl-CoA. The functional assignments of all these genes (except phosphotransacetylase, below) are backed by biochemical evidence based on protein sequence analysis of well-characterized enzymes.

Conversion of CO_2 to formate is catalysed by the well-characterized selenocysteine-containing formate dehydrogenase (Li *et al.*, 1966; Andreesen and Ljungdahl, 1973; Yamamoto *et al.*, 1983). Moth_2312 and Moth_2313, which are actually one gene, match the sequence deposited by Li, Ljungdahl and Gollin in GenBank (gi 1658401) as the formate dehydrogenase α subunit. The two ORFs were separated in the original genome annotation because they contain the codon for selenocysteine, a stop codon in most contexts. Moth_2314 matches the sequence of the β subunit. In acetogens, formate dehydrogenase catalyses an especially interesting reaction as the reduction of CO_2 to formate ($E_o = -420$) is coupled to the oxidation of NADPH to $NADP^+$ ($E_o = -340$ mV). The *M. thermoacetica* formate dehydrogenase was the first enzyme shown to contain tungsten (Ljungdahl and Andreesen, 1975). Another
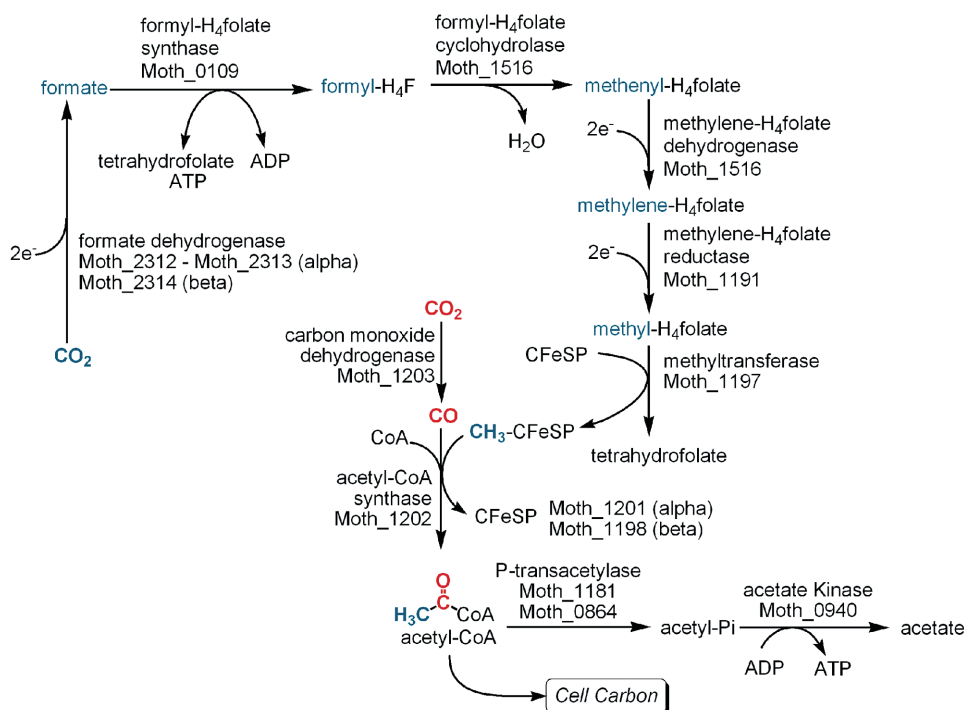
Fig. 1. 16S rRNA-based phylogenetic tree of acetogens and their close neighbours. The tree was derived from 16S rDNA sequence data, and was constructed using the MEGA program neighbour-joining method. Highlighted in yellow are acetogens that have the Wood–Ljungdahl pathway gene clusters listed in Fig. 4.

selenocysteine-containing formate dehydrogenase (Moth_2193), which appears to be part of a formate hydrogen lyase system (Moth_2184–2193), is described below.

The genes encoding 10-formyl-H$_4$folate synthetase (Moth_0109) and the bifunctional 5,10-methenyl-H$_4$folate cyclohydrolase/5,10-methylene-H$_4$folate dehydrogenase (Moth_1516) are dispersed on the genome and distant from the *acs* gene cluster by at least 300 genes (Fig. 3). In the reaction catalysed by 10-formyl-H$_4$folate synthetase, formate undergoes an ATP-dependent condensation with

H$_4$folate, forming 10-formyl-H$_4$folate. This enzyme has been purified, characterized and sequenced from a number of sources including *M. thermoacetica* (Himes and Wilder, 1968; O'Brien *et al.*, 1975; McGuire and Rabinowitz, 1978; Lovell *et al.*, 1988; 1990). In higher organisms, this enzyme exists as one of the activities of a trifunctional C$_1$-H$_4$folate synthase (also containing cyclohydrolase and dehydrogenase; Staben and Rabinowitz, 1986; Shannon and Rabinowitz, 1988), whereas the bacterial formyl-H$_4$folate synthetases are monofunctional and share similar properties (Lovell *et al.*, 1990). While the

Fig. 2. The Wood–Ljungdahl pathway of CO_2 fixation. Specific *M. thermoacetica* genes predicted or known to encode each step in the pathway are identified by locus tags, which can be used to find gene information at http://img.jgi.doe.gov. The C-1 groups are colour coded with the C-1 unit metabolized through the Eastern or methyl branch of the pathway shown in blue and the C-1 group in the Western or carbonyl branch shown in red.

cyclohydrolase and dehydrogenase are part of a bifunctional protein in *M. thermoacetica*, they are monofunctional proteins in other acetogens, e.g. *Clostridium formicoaceticum* and *Acetobacterium woodii* (Moore *et al.*, 1974; Clark and Ljungdahl, 1982; Ragsdale and Ljungdahl, 1984).

The enzyme that catalyses the last step in the Eastern branch is encoded by the 5,10-methylene-H_4folate reductase gene (Moth_1191). In *M. thermoacetica*, this is an oxygen-sensitive enzyme containing an iron–sulfur cluster, zinc and FAD, which utilizes reduced ferredoxin as an electron donor (Clark and Ljungdahl, 1984; Park *et al.*, 1991). Thauer and colleagues (1977) suggested that the 5,10-methylene-H_4folate reductase may link to an electron transport chain that begins with an electron donor (e.g. hydrogenase, CODH or NADH dehydrogenase)
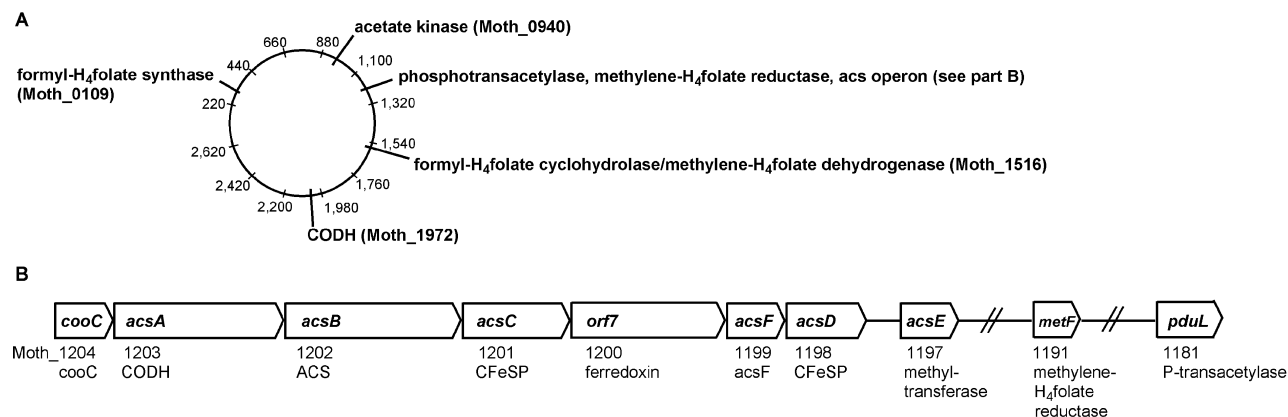

Fig. 3. Arrangement of Wood–Ljungdahl pathway genes in the *acs* gene cluster and in the chromosome.
A. Arrangement of Wood–Ljungdahl pathway genes on the circular chromosome of *M. thermoacetica*. The numbering shows kilobase pairs from the origin of replication.
B. The *acs* gene cluster that contains core Wood–Ljungdahl pathway genes discussed in the text.

through membrane-bound electron carriers, finally to methylene-H$_4$folate. In this scenario, electron transport could result in the generation of a transmembrane proton potential coupled to ATP synthesis. In *A. woodii*, one of the steps involving methylene-H$_4$folate reductase, the methyltransferase or the CFeSP (see Fig. 2) was shown to generate a sodium motive force across the membrane (Heise *et al.*, 1989). However, as the reductase appears to be located in the cytoplasm in *M. thermoacetica* (Clark and Ljungdahl, 1984) and in other organisms, it appeared that this is not the energy-conserving step in the Wood–Ljungdahl pathway (Muller, 2003). On the other hand, the location of the 5,10-methylene-H$_4$folate reductase gene very near genes encoding a hydrogenase (Moth_1193) and heterodisulfide reductase (Moth_1194–1196), which are directly downstream from the *acs* gene cluster (Moth_1197–1204), suggests a possible mechanism of energy conservation involving the reductase. Thus, perhaps methylene-H$_4$folate reductase links to the membrane-associated hydrogenase and/or heterodisulfide reductase to generate the proton motive force.

Genes encoding the Western branch of the Wood–Ljungdahl pathway

The core genes of the Western branch of the Wood–Ljungdahl pathway were first identified in *M. thermoacetica* by a reverse genetics approach to be located in the *acs* gene cluster (Fig. 3B), which contains genes encoding carbon monoxide dehydrogenase (CODH), acetyl-CoA synthase (ACS), methyltransferase (MeTr) and both subunits of the corrinoid iron–sulfur protein (CFeSP) (Roberts *et al.*, 1989). The *acs* gene cluster also includes a gene encoding an iron–sulfur protein of unknown function (*orf7*), and two genes, *cooC* and *acsF*, which encode proteins with sequences that are 44% and 27% identical to that of *Rhodospirillum rubrum cooC*, which is required for nickel insertion into carbon monoxide dehydrogenase (Jeon *et al.*, 2001). Thus, the ubiquitous genes in the Eastern branch of the Wood–Ljungdahl pathway are scattered around the genome, while what can be considered to be the 'core set of Wood–Ljungdahl genes' in the Western part of the pathway are colocalized. The genome sequence reveals an additional carbon monoxide dehydrogenase (Moth_1972), which is similar to an uncharacterized carbon monoxide dehydrogenase from *Clostridium cellulolyticum* and to CODH IV from *Carboxydothermus hydrogenoformans* (Wu *et al.*, 2005), and is unlinked to the *acs* gene cluster.

Phosphotransacetylase has been characterized from *M. thermoacetica* (Drake *et al.*, 1981), but the gene sequence has not been reported. *Moorella thermoacetica* lacks any gene that is homologous to the phosphotransacetylase genes whose products are used for phospho-

rylation of acetate from acetyl-CoA in related organisms, like *Clostridium acetobutyricum* (Boynton *et al.*, 1996). However, Moth_1181 and Moth_0864 are homologous to the recently characterized *pduL*, which is used for 1,2-propanediol degradation by *Salmonella enterica* (Liu *et al.*, 2007). Therefore, Moth_1181 and Moth_0864 are likely to encode phosphotransacetylases in *M. thermoacetica*. The acetate kinase (Moth_0940) gene is distant from and, thus, is not linked to Moth_1181, Moth_0864 or the *acs* gene cluster; however, Moth_1181 is located only 16 genes downstream of the methyltransferase in the Wood–Ljungdahl pathway (or 10 genes further downstream from the gene encoding methylene-H$_4$folate reductase).

Comparison of Wood–Ljungdahl pathway gene clusters

Acetyl-CoA synthase (COG1614, *acsB*) is found in five *Clostridia*, three *Deltaproteobacteria* and three *Dehalococcoidetes* species among the completely sequenced bacteria, and in the almost complete genome sequence of the uncultured bacterium *Kuenenia stuttgartiensis* (Strous *et al.*, 2006), and seems to be a marker for the presence of the Wood–Ljungdahl pathway in bacteria. COG1614 is also found in archaea and corresponds to the ~53 kDa β-subunit that harbours the active site dinickel–iron–sulfur cluster of acetyl-CoA decarbonylase synthase in methanogens and archaeoglobus (Pfam 03598, *cdhC*) (Maupin-Furlow and Ferry, 1996; Gencic and Grahame, 2003). From these genomes with acetyl-CoA synthase genes, only *Dehalococcoides ethenogenes* (Seshadri *et al.*, 2005) and *Dehalococcoides* sp. BAV1 and CBDB1 do not have the complete Wood–Ljungdahl pathway. Though these genomes encode many enzymes that are part of the pathway, they lack homologues of known methylene-tetrahydrofolate reductase, methyltransferase (*acsE*), CO dehydrogenase (*acsA*), phosphotransacetylase (*pta* or *pduL*) and acetate kinase sequences. We have been unable to locate any gene in the *Dehalococcoides* genome that is homologous to any of the CODH (*acsA*) genes from *M. thermoacetica* or *C. hydrogenoformans*. Given the small genome size and metabolic specialization of *D. ethenogenes*, one can speculate that the complete pathway was present in an ancestor, and has partly been lost as the organism's metabolism became limited to its current capabilities. *Alkaliphilus metalliredigenes* (Ye *et al.*, 2004), *C. hydrogenoformans* (Wu *et al.*, 2005), two *Clostridium difficile* strains, two *Desulfitobacterium hafniense* strains (Nonaka *et al.*, 2006), *Syntrophobacter fumaroxidans* (Harmsen *et al.*, 1998), *Deltaproteobacterium* MLMS-1, *K. stuttgartiensis* (Strous *et al.*, 2006) and *Candidatus Desulfococcus oleovorans* Hxd3, formerly *Deltaproteobacterium* Hxd3 (So *et al.*, 2003), appear to have complete Wood–Ljungdahl pathways, although
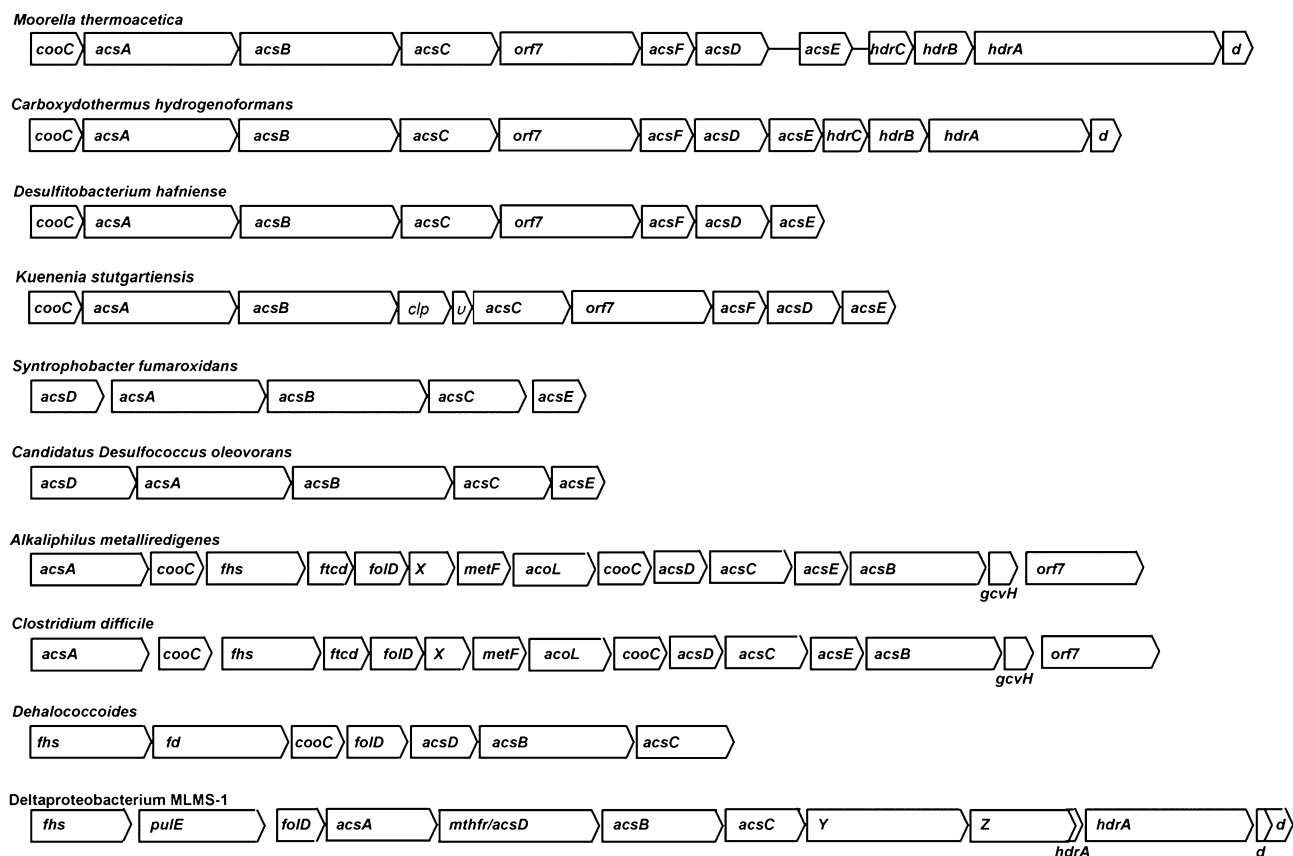
Fig. 4. Wood–Ljungdahl pathway gene clusters described in the text. Abbreviations are as follows: *cooC*: CODH chaperone; *acsA*: CODH; *acs*B: ACS; *acsC*: corrinoid iron–sulfur protein large subunit; *orf7*: unknown function; *acs*F: ACS chaperone similar to *cooC*; *acsD*: corrinoid iron–sulfur protein small subunit; *acsE*: methyltransferase; *hdrC*: heterodisulfide reductase subunit C; *hdrB*: heterodisulfide reductase subunit B; *hdrA*: heterodislfide reductase subunit A; *d*: similar to archaeal F420 reducing hydrogenase, δ-subunit (COG1908); *clp*: similar to *clp* protease; *u*: unknown function; *fhs*: formyl-H₄folate synthase; *ftcd*: formimido-H₄folate cyclodeaminase; *X*: zinc-finger protein of unknown function; *metF*: methylene-H₄folate reductase; *acoL*: dihydrolipoamide dehydrogenase; *gcvH*: glycine cleavage system H protein; *fd*: ferredoxin; *pulE*: Type II secretory pathway, ATPase PulE; *folD*: formyl-H₄folate cyclohydrolase/methylene-H₄folate dehydrogenase; *mthfr/acsD*: methylene-H₄folate reductase/corrinoid iron–sulfur protein small subunit fusion; *Y*: NADPH-dependent oxidoreductase (COG0493); *Z*: molybdopterin oxidoreductase (COG3383).

none of these organisms produces acetate as their main growth product.

As acetogenesis is a physiological trait that is not phylogenetically conserved (see Fig. 1), we reasoned that co-occurrence of acetyl-CoA synthase with a set of genes only in genomes of acetogenic bacteria could provide clues about which genes are required to confer the ability to grow as an acetogen. Thus, the *M. thermoacetica* genome was compared with those of all *Firmicutes*, *Chlorflexi* and *Deltaproteobacteria* available at IMG, using a BLAST analysis of each *M. thermoacetica* gene against these other genomes, to find *M. thermoacetica* genes with homologues in the other bacterial genomes that also have acetyl-CoA synthase, but without homologues in any of the sequences without acetyl-CoA synthase. The only genes that show this pattern of co-occurrence and co-omission are ACS and both subunits of the corrinoid iron–sulfur protein (CFeSP). The

CFeSP must have a highly conserved function as part of the Wood–Ljungdahl pathway, as homologues of the CFeSP genes are not found in any genome that does not also have acetyl-CoA synthase and other Wood–Ljungdahl pathway genes.

The Wood–Ljungdahl pathway gene clusters in these 12 bacteria can be divided into five groups within which the order of genes is absolutely conserved (Fig. 4). These groupings mostly follow the currently accepted taxonomy found at NCBI. The *M. thermoacetica*, *C. hydrogenoformans*, *D. hafniense* and *K. stuttgartensis* (group A) genes are arranged in the same way, and, like *M. thermoacetica*, *C. hydrogenoformans* has genes encoding heterodisulfide reductase directly following *acsE*. The *K. stuttgartensis* gene cluster is identical in arrangement to the *D. hafniense* gene cluster, except for the insertion of a gene with sequence similar to clp protease sequences and one small gene of unknown function
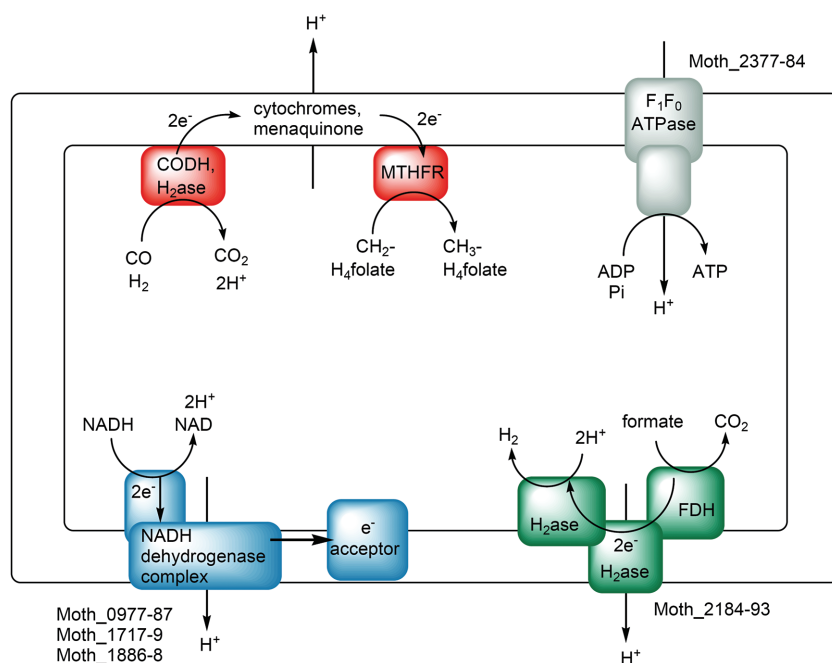
Fig. 5. Possible energy conserving pathways. Shown are pathways for growth by the Wood–Ljungdahl pathway (orange), using formate hydrogenlyase (green), or with alternate electron donors and acceptors, using NADH dehydrogenase I (blue). Under the right growth condition for each system, the system would generate a transmembrane proton gradient that could be used to drive ATP synthesis by the F_1F_0ATPase (grey).

after *acsB*. The *S. fumaroxidans* and *Candidatus D. oleovorans* (group B) gene clusters are alike, and minimal, with only the *acsABCDE* genes. *Candidatus D. oleovorans* has four proteins in the same COG (3640) as the *M. thermoacetica* accessory proteins, and *S. fumaroxidans* has two proteins from this COG, but none of the genes encoding these putative accessory proteins are near the *acs* gene cluster in either species. *A. metalliredigenes* and *C. difficile* (group C) have much larger Wood–Ljungdahl pathway gene clusters, which, besides the genes present in the *M. thermoacetica acs* gene cluster, include genes encoding formyl-tetrahydrofolate synthetase (*fts*), the bifunctional formyl-tetrahydrofolate cyclohydrolase/methylene-tetrahydrofolate dehydrogenase (*folD*), and methylene-tetrahydrofolate reductase, which function in the Eastern branch of the pathway (Fig. 2). The three *Dehalococcoides* species (group D) have gene clusters that include *acsBCD*, *fts* and *folD*; however, they surprisingly lack *acsA*. The *Deltaproteobacterium* MLMS-1 gene cluster (group E) is similar to the *Dehalococcoides* Wood–Ljungdahl gene clusters, but includes *acsA*, genes which belong in COG0493 and COG3383 (see the anaerobic respiration and electron acceptors discussions), one heterodisulfide reductase subunit A gene, and two genes encoding homologues of the archaeal hydrogenase δ-subunit that is found in the HDR gene clusters in *M. thermoacetica*. Based on this analysis, it appears that the core genes of the Wood–Ljungdahl pathway are located in the *acs* gene cluster, which includes *acsA-acsE*, and acquisition of this gene cluster might be sufficient for gaining the ability to grow as an acetogen.

Genes involved in anaerobic respiration

In the Wood–Ljungdahl pathway, one ATP is consumed in the formation of formyl-tetrahydrofolate from tetrahydrofolate and formate in the formyl-tetrahydrofolate synthase reaction (Sun *et al.*, 1969), and one is generated by phosphoryl transfer to ADP in the acetate kinase reaction (Fig. 2) (Schaupp and Ljungdahl, 1974). Therefore, there is no net gain of ATP by substrate level phosphorylation during autotrophic growth. Thus, it is clear that ATP is generated by anaerobic respiration. During growth using the Wood–Ljungdahl pathway, energy should be conserved by transfer of electrons along a membrane-bound electron transfer chain, coincident with generation of a transmembrane proton gradient (Fig. 5). *Moorella thermoacetica* has an F_1F_0ATP synthase, which could use this proton gradient to drive ATP synthesis (Das and Ljungdahl, 1997). The genes encoding the eight subunits of the ATP synthase are Moth_2377–84. *Moorella thermoacetica* has integral membrane-bound electron carriers including *b* cytochromes (see Table 2) and a menaquinone (Gottwald *et al.*, 1975; Hugenholtz *et al.*, 1987; Das *et al.*, 1989) that could be part of this system. There is evidence in the *M. thermoacetica* genome for genes that encode systems that could be associated with energy generation by transmembrane proton gradient formation (Table S2), including NADH dehydrogenase, hydrogenase or formate hydrogen lyase and heterodisulfide reductase.

There is considerable homology of both subunit compositions and sequences among NADH dehydrogenases and some energy conserving NiFe hydrogenases

Table 2. Cytochromes present in the *M. thermoacetica* genome.

Locus tag	Annotation	COG
Moth_0452	Formate dehydrogenase, γ-subunit (cytochrome *b*)[a]	COG2864
Moth_1605	Cytochrome *c* family protein	No cog
Moth_2095	Cytochrome *bd* ubiquinol oxidase, subunit I	COG1271
Moth_2096	Cytochrome *d* ubiquinol oxidase, subunit II	COG1294

a. This formate dehydrogenase is different than the selenocystine-containing formate dehydrogenase identified experimentally by Li, Ljungdahl and Gollin in GenBank (gi 1658401)

(Deppenmeier, 2004; Hedderich, 2004). Examples of the former are mitochondrial complex I and *Escherichia coli* NADH dehydrogenase I. The latter include *E. coli* hydrogenases 3 and 4, which are parts of formate hydrogen lyase complexes (Andrews *et al.*, 1997), the CO-inducible hydrogenase from *Rhodospirillum rubrum* (Fox *et al.*, 1996), a CO dehydrogenase-associated hydrogenase from *C. hydrogenoformans* (Soboh *et al.*, 2002), and the Ech hydrogenase of *Methanosarcina barkeri* (Meuer *et al.*, 1999). Two sets of *M. thermoacetica* genes, Moth_0977–87 and Moth_2184–92, encode proteins homologous to these complexes. A BLAST search of *C. hydrogenoformans*, *E. coli*, *M. barkeri* and *R. rubrum* genomes with genes from Moth_0977–87 (Table 3) shows that these genes are most closely related to NADH dehydrogenase I in bacteria, and to the F_{420}-dependent

hydrogenase of *M. barkeri*. In some bacteria, as shown in Table 3 for *R. rubrum* and *E. coli*, the NADH dehydrogenase operon contains genes encoding subunits E, F and G of the NADH dehydrogenase complex. In *M. thermoacetica*, homologues to these subunits are encoded separately, on two different gene clusters, Moth_ 1717–9 and Moth_ 1886–8. Moth_ 1717 and Moth_ 1888 align only with the first 240 and 280 amino acids of NuoG sequences from *E. coli* and *Thermus thermophilus*. In *T. thermophilus*, this N-terminal portion of the protein has the residues that ligate the three iron–sulfur clusters that are involved in electron transfer through the NADH dehydrogenase complex (Sazanov and Hinchliffe, 2006).

Proteins encoded by Moth_2184–92 (Table 4) are most closely related to *E. coli* hydrogenases 4 and 3 among experimentally characterized proteins. In *C. hydrogeno-*

Table 3. Comparison of Moth_0977–87 with genes encoding *C. hydrogenoformans*, *R. rubrum* and *E. coli* NADH dehydrogenase I and *M. barkeri* F-420-dependent hydrogenase.

M. thermoacetica	*C. hydrogenoformans*	*R. rubrum*	*E. coli*	*M. barkeri*	COG
Moth_0977	CHY_1415	Rru_A1555	b2288[a]	Mbar_A3412	0838
Moth_0978	CHY_1416	Rru_A1556	b2287	Mbar_A3411	0388
Moth_0979	CHY_1417	Rru_A1557	NH	Mbar_A3410	0852
Moth_0980	CHY_1418	Rru_A1558	b2286	Mbar_A3409	0649
NG	NG	Rru_A1559-61[a]	b2285–3[a]	NG	1905, 1894, 1034
Moth_0981	CHY_1419	Rru_A1562	b2282	Mbar_A3408	1005
Moth_0982	CHY_1420	Rru_A1563	b2281	Mbar_A3407	1143
Moth_0983	CHY_1421	Rru_A1564	b2280[a]	Mbar_A3406	0839
NG	NG	NG	NG	Mbar_A3405	0713
Moth_0984	CHY_1422	Rru_A1565	b2279	Mbar_A3404	1009
Moth_0985	CHY_1423	Rru_A1566	b2278	Mbar_A3403	1009
Moth_0986	CHY_1424	Rru_A1567	b2277	Mbar_A3402	1008
Moth_0987	CHY_1425	Rru_A1568	b2276	Mbar_A3401	1007
	Subunit designation				
Moth_0977	A				
Moth_0978	B				
Moth_0979	C				
Moth_0980	D				
NG	EFG				
Moth_0981	H				
Moth_0982	I				
Moth_0983	J				
NG					
Moth_0984	K				
Moth_0985	L				
Moth_0986	M				
Moth_0987	N				

a. These genes do not have reciprocal best hits with genes in the *M. thermoacetica* gene cluster (Moth_0977 to Moth_0987).
The BLAST cut-off is *E*-value $< 1E^{-2}$.
NG, no homologous gene is present in this gene cluster; NH, no gene in this organism is a BLAST hit with the corresponding *M. thermoacetica* gene.

Table 4. Comparison of Moth_2184–92 with hydrogenases from *C. hydrogenoformans* and *E. coli.*

M. thermoacetica	C. hydrogenoformans	E. coli hydrogenase 3	E. coli hydrogenase 4			
Moth_2184	CHY_1830	cooL	b2719	hycG	b2489[a]	hyfI
Moth_2185	CHY_1829	cooX	b2720[a]	hycF	b2488	hyfH
Moth_2186	CHY_1827	cooH	b2721	hycE	b2487[a]	hyfG
Moth_2187	CHY_1832[a]	cooM	NH		b2486	hyfD
Moth_2188	CHY_1832[a]	cooM	b2723[a]	hycC	b2485	hyfF
Moth_2189	NH		NH		b2484	hyfE
Moth_2190	CHY_1831	cooK	b2722[a]	hycD	b2483	hyfC
Moth_2191	CHY_1832[a]	cooM	b2723[a]	hycC	b2482	hyfB
Moth_2192	CHY_1825[a]	cooF	b2724[a]	hycB	b2481	hyfA

	COG					
Moth_2184	3260					
Moth_2185	1143					
Moth_2186	3261					
Moth_2187	1009					
Moth_2188	0651					
Moth_2189	4237					
Moth_2190	0650					
Moth_2191	0651					
Moth_2192	1142					

a. These genes do not have reciprocal best hits with genes in the *M. thermoacetica* gene cluster (Moth_2184 to Moth_2192).
NH, no BLAST hit with *E*-value < $1E^{-2}$ in this gene cluster.

formans, Moth_2184–88 and Moth_2190–92 are most closely related to CO-associated hydrogenase, although there is lower sequence homology between these *M. thermoacetica* genes and CO-associated hydrogenase than between Moth_2184–92 and *E. coli* hydrogenases 3 and 4. Moth_2184–90 are more similar to subunits of hydrogenase 4 than subunits of hydrogenase 3, and Moth_2187–89 encode homologues of *E. coli* hydrogenase four genes *hyfDEF*, which do not have homologues in hydrogenase 3. Hydrogenase 3 does not generate a proton gradient, and probably is used by *E. coli* to metabolize formate produced during fermentative growth. From their analysis of the *hyf* gene cluster, Andrews and colleagues (1997) concluded that hydrogenase 4 is likely to have proton-translocating ability (although this has not been verified experimentally), because, like the NADH dehydrogenase I complex and the *R. rubrum* CO-inducible hydrogenase, and unlike hydrogenase 3 from *E. coli*, *hyfBEF* are homologous to ND2, ND4 and ND5 components of the quinone-reducing and proton translocating machinery in NADH dehydrogenase I. *Moorella thermoacetica* genes Moth_2176–83, which are immediately downstream from this second gene cluster, are homologues of the *hypA-F*, *hycI* and *hycH* genes of *E. coli*, which make up an operon that is required for maturation of the Ni-Fe hydrogenase 3 (Lutz *et al.*, 1991; Blokesch *et al.*, 2004). Immediately upstream of the hydrogenase gene cluster is a gene encoding a selenocysteine-containing formate dehydrogenase (Moth_2193) that is separate from the formate dehydrogenase that feeds formate into the Wood–Ljungdahl pathway. It is likely that Moth_2184–93 encodes a formate

hydrogenlyase system, which by homology to hydrogenase 4, could generate a proton gradient in the presence of formate.

In methanogenic archaea, hydrogenases distinct from the Ech hydrogenase discussed above are also involved in energy generation as parts of membrane-bound electron transport systems, along with heterodisulfide reductase. In these systems, electrons are transferred from hydrogen to the disulfide formed between coenzyme M and coenzyme B during the last step in methanogenesis. Both two-subunit haem *b*-containing and three-subunit (non-haem) heterodisulfide reductases have been characterized (Setzke *et al.*, 1994; Deppenmeier, 2004). Although it is not an archaeon, *M. thermoacetica* has three heterodisulfide reductase gene clusters that are homologous to the three-subunit heterodisulfide reductase from *Methanothermobacter marburgensis* (Table 5), one of which (Moth_1196) is directly after the CODH/ACS gene cluster. Moth_0441 (a heterodisulfide reductase A subunit) includes about 200 amino acids (from 400 to 600) that do not align with HdrA domains from *Methanosarcina* or most bacterial hits. These amino acids align with lower *E*-values with sequences from *Syntrophobacter fumaroxidans*, *Methanosarcina mazei*, an uncultured archaeon, *Desulfotomaculum reducens*, two *Geobacter* species and *Candidatus D. oleovorans*. Moth_1194 aligns with HdrA, except between amino acids 150 and 550. This region aligns with COG0493, which is made of the NADPH-dependent glutamate synthase β-chain, and related oxidoreductases. Moth_0811 and Moth_0812 make up one selenocysteine-containing gene, which aligns with other HdrA sequences along its full length.

Table 5. Heterodisulfide reductase gene clusters in the *M. thermoacetica* genome.

Locus tag	Annotation	COG
Moth_0439	Heterodisulfide reductase subunit C-like (sequence is shorter)	COG1150
Moth_0440	Heterodisulfide reductase subunit B	COG2048
Moth_0441	Heterodisulfide reductase subunit A	COG1148
Moth_0442[a]	Similar to archaeal F420 reducing hydrogenase, δ-subunit	COG1908
Moth_0809	Heterodisulfide reductase subunit C	COG1150
Moth_0810	Heterodisulfide reductase subunit B	COG2048
Moth_0811	Heterodisulfide reductase selenocysteine containing subunit A	COG1148
Moth_0812	Heterodisulfide reductase selenocysteine containing subunit A	COG1148
Moth_0813[a]	Selenocysteine containing methyl viologen-reducing hydrogenase, δ-subnit	COG1908
Moth_1193[a]	Similar to archaeal F420 reducing hydrogenase, δ-subunit	COG1908
Moth_1194	NADPH-dependent glutamate synthase β-chain and related oxidoreductases (COG0493), Heterodisulfide reductase subunit A (COG1148)	COG1148 COG0493
Moth_1195	Heterodisulfide reductase subunit B	COG2048
Moth_1196	Heterodisulfide reductase subunit C	COG1150
Moth_1452	Unknown function	COG0247
Moth_1453	Heterodisulfide reductase A	COG1148
Moth_1454[a]	FAD-dependent oxidoreductase/NAD(P)H-dependent reductase	COG1251

a. These *M. thermoacetica* genes are annotated by similarity to COGs 1908 and 1251, not by alignment with specific protein sequences.

Genes involved in autotrophic growth

One-carbon compounds that *M. thermoacetica* can use for growth include H_2/CO_2, CO, formate, methanol and methyl groups from many methoxylated aromatic compounds. The enzymes that catalyse transfer of the methyl group of methoxylated aromatic compounds to tetrahydrofolate are encoded by *mtvABC* (Moth_0385–7 and Moth_1316–8) and have been purified and identified biochemically by enzymatic function and protein sequence analysis (Naidu and Ragsdale, 2001). This pathway is shown in Fig. 6, along with that of a methanol methyltransferase system (MtaABC, encoded by (Moth_1208–9 and Moth_2346) was recently characterized from *M. thermoacetica* (Das *et al.* 2007). *Moorella thermoacetica* has two nearly identical *mtvA* genes, each of which is next to a gene that we assign as *mtvB*, and a predicted cobalamin-binding *mtvC* gene. The putative *mtvB* genes are 80% identical to each other, and the *mtvC* genes differ by one amino acid. Both *mtvA* and *mtvC* genes are similar to cobalamin-dependent methionine synthase (5-methyltetrahydrofolate:homocysteine S-methyltransferase, encoded by *metH* in *E. coli*).

The two *mtvB* genes are most similar to uroporphyrinogen decarboxylase sequences, among experimentally characterized proteins, although the function of MtvB is clearly associated with methyl transfer. In addition to these *mtv* genes, *M. thermoacetica* encodes several homologues of MtvA, MtvB and MtvC (Table 6). It is possible that some of these homologues are used for growth on methyl donors other than vanillate and methanol. Many homologues of the *M. thermoacetica* MtvA and C proteins are also found in *D. hafniense*, which can also demethylate methoxylated aromatic compounds (Neumann *et al.*, 2004). These are arranged in clusters with *A* and *C* genes

and sequences aligning either with *M. thermoacetica* MtvB or sequences of similar function from *Acetobacterium dehalogenans*. One of the *M. thermoacetica* MtvB homologues (Moth_1314) is fused to a domain that has not been functionally identified, and does not align with any COG identified at NCBI. Sequences that align with this whole fusion are found in *D. hafniense*, *Archaeoglobus fulgidus* and *A. metalliredigenes*.

Genes involved in heterotrophic growth

Moorella thermoacetica can grow heterotrophically on fructose, glucose, xylose, ethanol, *n*-propanol, *n*-butanol (Beaty and Ljungdahl, 1991), oxalate and glyoxylate (Daniel and Drake, 1993), glycolate (Seifritz *et al.*, 1999), pyruvate (Barker and Kamen, 1945) and lactate (Gößner *et al.*, 1999). *Moorella thermoacetica* encodes a complete Embden–Meyerhof–Parnas pathway and all enzymes that are part of the oxidative branch of the pentose phosphate pathway (Fig. S2). Glucose and xylose are converted stoichiometrically to acetate by *M. thermoacetica* (Fontaine *et al.*, 1942; Barker and Kamen, 1945). It seems likely that fructose is metabolized by the same pathway as glucose (Fig. S3), as growth yields of the two substrates are very similar (Andreesen *et al.*, 1973).

Ethanol, propanol and butanol are oxidized to acetate, propionate and butyrate by *M. thermoacetica* (Beaty and Ljungdahl, 1991). Eight genes from four COGs containing alcohol dehydrogenases are present in the genome (Table S3). Proteins encoded by three genes from COG2414, tungsten-containing aldehyde:ferredoxin oxidoreductases, might catalyse the second step in growth on these alcohols.

The pathways for growth of *M. thermoacetica* on the two-carbon substrates oxalate, glyoxylate and glycolate
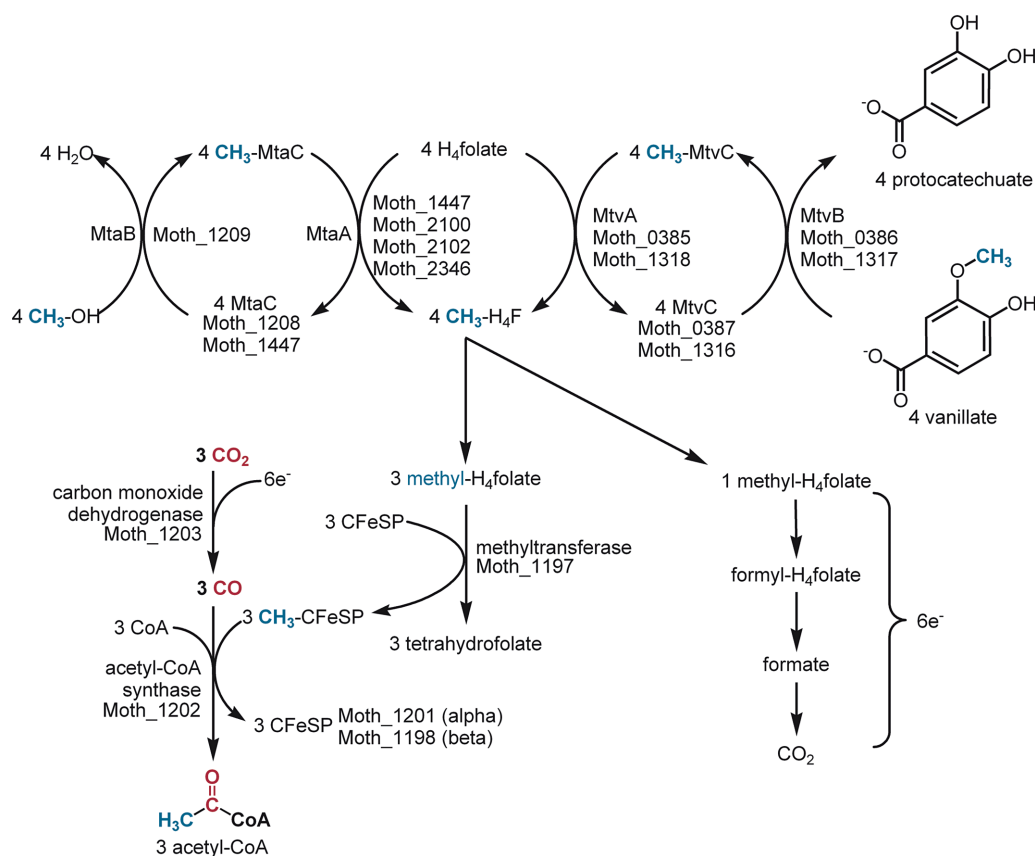
Fig. 6. Methyl groups from methanol and vanillate linked to the Wood–Ljungdahl pathway. Demethylation of vanillate or methanol provides preformed methyl groups to the Wood–Ljungdahl pathway. Three moles of acetyl-CoA are formed from 4 mol of methanol or vanillate, because oxidation of one methyl group provides six electrons that reduce 3 mol of CO_2 to 3 mol of CO, which can be used in the synthesis of three acetyl-CoA molecules. The C-1 units are labelled as in Fig. 2.

Table 6. MtvA, MtvB and MtvC homologues in the *M. thermoacetica* genome.

Locus tag	Annotation	COG
Moth_0385	MtvA	COG1410
Moth_0386	MtvB	COG0407
Moth_0387	MtvC	COG5012
Moth_1197	AcsE (methyl-THF corrinoid FeS protein methyltransferase)	COG1410
Moth_1206	MtvC homologue	COG5012
Moth_1207	MtvA homologue	COG1410
Moth_1208	MtaC	COG5012
Moth_1209	MtaB	No COG
Moth_1213	MtvC homologue	COG1410
Moth_1311	MtvC homologue	COG5012
Moth_1314	MtvB plus an N-terminal domain	COG0407
Moth_1316	MtvC	COG5012
Moth_1317	MtvB	COG0407
Moth_1318	MtvA	COG1410
Moth_1447	MtaA/MtaC	COG0407, COG 5012
Moth_2100	MtaA	COG0407
Moth_2102	MtaA	COG0407
Moth_2114	MtvC homologue/histidine kinase fusion protein	COG5012, COG2972
Moth_2115	MtvA homologue	COG1410
Moth_2116	MtvC homologue	COG5012
Moth_2346	MtaA	COG0407

are not known. The stoichiometries of acetate production suggest that they are oxidized by two, four and six electrons respectively (Seifritz *et al.*, 1999). Given differences in growth on these three substrates, it seems likely that *M. thermoacetica* uses different pathways for their oxidation. Only partial pathways known to be involved in glycolate metabolism in other organisms are present in *M. thermoacetica*. In *E. coli*, the glycolate metabolism genes are located in an operon, which includes *glcA* (encodes a glycolate permease), *glcB* (encodes malate synthase), *glcG* (function is unknown), *glcF, E* and *D* (encode the three subunits of glycolate oxidase) and *glcC* (encodes a transcriptional activator) (Pellicer *et al.*, 1996), and the first enzymatic step is oxidation of glycolate to glyoxylate via an FMN-dependent oxidase (GlcDEF). The *M. thermoacetica* genes are arranged in two clusters, one containing *glcDEF* homologues (Moth_0037–39), and the other containing homologues to *glcACF*, an unknown protein, another *glcF*, and a sequence that aligns with *glcD* and *E* (Moth_2303–8). The glycolate permease gene (*glcA*) in *E. coli* and the proposed *M. thermoacetica glcA* are very similar to L-lactate permeases (Núñez *et al.*, 2001). No homologue of malate synthase was found in *M. thermoacetica*, but *M. thermoacetica* does have a gene, Moth_1954, with 51% sequence identity to a possible glyoxylate reductase that has been purified from *Thermococcus litoralis* (Ohshima *et al.*, 2001). This enzyme catalyses formation of malate from glyoxylate and acetyl-CoA and is homologous to D-lactate dehydrogenase.

Moorella thermoacetica is able to grow on lactate and pyruvate. Lactate is probably converted to pyruvate by Moth_1826, which is similar to both *Clostridium acetobutylicum* and *Thermotoga maritima* L-lactate dehydrogenases. Pyruvate is metabolized using pyruvate:ferredoxin oxidoreductase (PFOR), which catalyses the oxidative cleavage of pyruvate and attachment of the two carbon fragment to CoA, forming acetyl-CoA, CO_2 and reduced ferredoxin (Fig. S4). PFOR is also essential in sugar metabolism, as it couples the Embden–Meyerhoff–Parnas pathway to the Wood–Ljungdahl pathway. In sugar metabolism, four additional electrons are available from the glyceraldehyde-phosphate dehydrogenase reaction, so 1 mol of glucose (i.e. 2 mol of pyruvate) is required to supply the eight electrons for reducing 2 mol of CO_2 to acetate. The PFOR that has been purified from *M. thermoacetica* is a homodimer, with a subunit mass of 120 kDa (Furdui and Ragsdale, 2002). The gene encoding this enzyme, Moth_0064, is a fusion of three COGs, COG0674, COG1014, COG1013, which are annotated as α-, γ- and β-subunits. Five other sets of genes belonging to the same COGs as PFOR (Table S4) could encode authentic PFORs, but they could also encode other proteins in the oxoglutamate oxidoreduc-

tase family, like 2-ketoisovalerate oxidoreductase, indolepyruvate oxidoreductase and 2-ketoglutarate oxidoreductase. Many of the residues that are known to be important in the well-characterized PFOR from *Desulfovibrio africanus* are conserved in the PFOR-like gene clusters of *M. thermoacetica*. These include residues that bind iron–sulfur clusters, thiamine pyrophosphte and pyruvate. Conservation of these residues suggests that these *M. thermoacetica* gene clusters likely encode enzymes that perform similar functions to PFOR. Variation in the residue that forms a hydrophobic interaction with the methyl group of pyruvate in the *D. africanus* enzyme could be due to different substrate specificity of the protein products of these different gene clusters. For example, the replacement of this residue with a glycine residue could accommodate a substrate larger than pyruvate.

Genes involved in utilization of electron acceptors other than CO_2

Moorella thermoacetica can use electron acceptors other than CO_2 during heterotrophic growth. When nitrate and CO_2 are both supplied, nitrate is used preferentially (Seifritz *et al.*, 1993). In medium containing a carbon source for heterotrophic growth, addition of nitrate stimulates growth, and reducing equivalents from H_2 and one-carbon substrates are used to make ammonium, while acetate production is decreased. Work by Fröstl *et al.*, 1996; Arendsen *et al.*, 1999 suggests that carbon incorporation through the Wood–Ljungdahl pathway is repressed by nitrate. Nitrite (Seifritz *et al.*, 2003), dimethylsulfoxide and thiosulfate (Beaty and Ljungdahl, 1991) are also electron sinks for *M. thermoacetica*. Some of these electron acceptors could be reduced by oxidoreductases similar to DMSO and TMAO reductases found in *E. coli* and many other organisms. These proteins use molybdopterin cofactors, are often membrane associated, and in some cases are transported to the periplasm by twin-arginine transport (TAT) systems.

Moorella thermoacetica genes similar to the *E. coli* DMSO and TMAO reductases have been identified. The *E. coli* DMSO reductase has three subunits. DmsA is the molybdopterin-containing catalytic subunit (COG0243), DmsB is an iron–sulfur electron transfer subunit (COG0437), and DmsC is a membrane anchor subunit (COG3302). TMAO reductase has a cytochrome subunit instead of DmsC and B-like subunits. *Moorella thermoacetica* has eight genes belonging to COGs 0243 and 3383, which are made up of molybdopterin-containing proteins, shown in Table 7. Five of these genes are adjacent to COG0437 proteins. One homologue of the *E. coli* DmsC subunit is also present in *M. thermoacetica*. No oxidoreductase subunit homologous to the pentaheme

Table 7. Gene clusters containing predicted molybdopterin subunits in COGs 0243 and 3383.

Locus tag	Annotation	COG
Moth_0450	Molybdopterin subunit, TAT sequence	COG0243
Moth_0451	Iron sulfur cluster containing subunit	COG0437
Moth_0452	Cytochrome b556 subunit of formate dehydrogenase	COG2864
Moth_1379	TorD	COG3381
Moth_1381	Iron sulfur cluster containing subunit	COG0437
Moth_1382	Molybdopterin containing subunit	COG0243
Moth_1384	DmsC type membrane anchor subunit	COG3302
Moth_1385	Iron sulfur cluster containing subunit	COG0437
Moth_1386	Molybdopterin subunit, TAT sequence	COG0243
Moth_1388	Iron sulfur cluster containing subunit	COG0437
Moth_1389	Molybdopterin subunit, TAT sequence	COG0243
Moth_1390	TorD	COG3381
Moth_1906	Polysulfide reductase, integral membrane protein	COG5557
Moth_1907	Iron sulfur cluster containing subunit	COG0437
Moth_1908	Molybdopterin subunit, TAT sequence	COG0243
Moth_1909	Molybdopterin containing subunit	COG3383
Moth_2193	Se-containing formate hydrogen lyase subunit	COG3383
Moth_2312–13	Se-containing formate dehydrogenase subunit	COG3383
Moth_2314	Predicted NADPH-dependent oxidoreductase	COG0493

cytochrome *c* family TorC subunit of the TMAO reductase in *E. coli* (gi 2506330, COG3005; Gon *et al.*, 2001) was found in *M. thermoacetica*. This subunit connects TMAO reductase in *E. coli* with respiration via the quinone pool. Thus, in *M. thermoacetica*, it is likely that one of the electron transfer proteins discussed above in the *Anaerobic Respiration* section couples TMAO and DMSO-derived electrons to the quinone pool.

Analysis of the N-terminal sequences of the molybdopterin oxidoreductases shows that three of the genes include likely TAT signal sequences. *Moorella thermoacetica* encodes TatA and TatC proteins (Moth_1278 and Moth_1279), but does not have a *tatB* gene. This is also true of *Bacillus subtilis*, which uses a system composed of A and C subunits (Jongbloed *et al.*, 2004). Two genes similar to *torD*, flanking a gene cluster that encodes three molybdopterin-containing oxidoreductases, two of which have TAT sequences, might be chaperones for the TAT system (Hatzixanthis *et al.*, 2005; Genest *et al.*, 2006).

In addition to the molybdopterin oxidoreductases, *M. thermoacetica* has additional genes that could be involved in growth on electron acceptors other than CO_2, e.g. sulfite and nitrate (Table S5). One gene cluster has three genes that have about 80% sequence identity with the dissimilatory sulfite reductase (*dsrAB*) and the gene *dsrD*, of unknown function, from *Desulfotomaculum thermocisternum* (Larsen *et al.*, 1999). Directly upstream from these genes are a gene similar to the membrane component of *E. coli* nitrate reductase, *narI*, a gene similar to *nrfD* which is a transmembrane component of *E. coli* nitrite reductase, two predicted iron–sulfur containing proteins, and a cytochrome *c*. The dissimilatory sulfite reductase does not have any likely transmembrane regions, but the additional genes might serve to connect the reductase

to a membrane-bound respiratory pathway. One other possible sulfite reductase gene cluster, Moth_1629–32, has *dsrAB* genes, and two genes similar to *dsrC* from *A. fulgidus*, whose function is not fully resolved. The genes in this cluster have only about 30% sequence identity to the most closely related genes in GenBank. The four genes in this second cluster have no predicted transmembrane domains.

Expanded gene groups in M. thermoacetica *based on comparative COG analysis*

The *M. thermoacetica* genome was compared with those of *T. ethanolicus* and *T. tengcongensis* to obtain insights into *M. thermoacetica*'s wide metabolic range. We analysed the placement of *M. thermoacetica*, *T. ethanolicus* and *T. tengcongensis* genes in COGs to identify functions with increased representation in *M. thermoacetica*. Table S6 shows some of the COGs with increased representation along with the number of genes belonging to each COG. The largest expansions in *M. thermoacetica*, relative to the other *Thermoanaerobacteriaceae* genome sequences, are in COGs that could encode proteins involved in electron transfer to or from growth substrates, including molybdopterin oxidoreductases in the dimethylsulfoxide reductase family detailed above, and in the xanthine dehydrogenase family (Kisker *et al.*, 1999), with subunits in COGs 1529, 1319 and 2080. The xanthine dehydrogenase subunits might be part of a pathway of purine degradation including amidohydrolase genes (COG0402), which are adjacent to two of the xanthine dehydrogenase genes. The COGs containing heterodisulfide reductase and sulfite reductase discussed above are absent from *T. ethanolicus* and *T. tengcongensis*,

except for one subunit A homologue in *T. ethanolicus*. *Moorella thermoacetica* has 13 genes representing two COGs of proteins that are likely to contain iron–sulfur clusters, and that seem to have diverse functions in the *M. thermoacetica* genome. COG0247 is part of the possible glycolate oxidase proposed above as the homologue of *E. coli glcF*. COG0247 sequences are also adjacent to a heterodisulfide reductase A gene (Moth_1453) that is unlinked to any other heterodisulfide reductase subunits, and to one of the sulfite reductase gene clusters. A gene encoding a COG0437 protein is part of the same sulfite reductase gene cluster, and other COG0437 proteins are predicted to be encoded by genes adjacent to many of the molybdopterin oxidoreductase genes, and are found in two gene clusters with aldehyde-ferredoxin oxidoreductases (COG2414).

Moorella thermoacetica has four genes in COG1251, which contains NAD(P)H-dependent nitrite reductase subunits similar to *E. coli*'s cytoplasmic nirB. However, *M. thermoacetica* lacks the gene encoding the other subunit of the enzyme found in *E. coli*, nirD (COG2146). One COG1251 gene is within an apparent glutamate synthase operon (Moth_1289–94), while another is adjacent to a gene cluster containing COG0437, an aldehyde-ferredoxin oxidoreductase and a gene similar to *E. coli*'s MoaD, which is involved in molybdopterin biosynthesis (Moth_0721–24). The third COG1251 gene is in a gene cluster with COG0243 and the lone heterodisulfide reductase A gene (Moth_1452–54), while the fourth is next to a hypothetical protein sequence of unknown function.

Another large expansion is in COGs related to methyl transfer reactions. Included in these COGs are the methyltransferase of the Wood–Ljungdahl pathway (Fig. 2) and other genes that could feed methyl groups into this pathway, including *mtvA* and *mtvC* homologues, and COG0407 which includes *mtvB* (Table 6). Four COG0407 genes are positioned close to *mtvA* and *mtvC* homologues, and four are adjacent to two putative sugar phosphate permeases that are members of family 1 of the major facilitator superfamily, in COG2271, which is also expanded in *M. thermoacetica*.

Another expanded COG group contains components of ABC-type nitrate/sulfonate/bicarbonate transport systems. These are represented by four gene clusters, each containing an ATPase, a permease component and a periplasmic component. Each of the *Thermoanaerobacter* species encodes one complete transporter from these three COGs.

Genes involved in cofactor synthesis

The only vitamin required for *M. thermoacetica*'s growth is nicotinic acid, a precursor of NAD and NADP (Lundie and Drake, 1984). *Moorella thermoacetica* appears to have genes encoding the phosphoribosyl transferase (Moth_1836), adenylyltransferase (Moth_0564) and glutamine-dependent NAD synthetase (Moth_0553) needed for conversion of nicotinic acid to NAD (Moth_1836, Moth_0564 and Moth_0553).

As only one vitamin is required for growth, genes encoding pathways for synthesis of all other cofactors should be present in the genome. Two metabolic routes for the synthesis of tetrahydrofolate are the *de novo* synthesis pathway (Fig. S5) and a salvage pathway. In *E. coli*, the *de novo* synthesis pathway requires the activities of the proteins encoded by the genes *folE*, *folB*, *folK*, *folP*, *folC* and *folA* to produce the final product, 6-(S)-tetrahydrofolate. *Moorella thermoacetica* has no *folA* gene, but the isoenzyme encoded by *ydgB* (*folM*) is present in three copies. The activities of two other enzymes from this pathway have been measured in *E. coli*, but the genes that encode the proteins are not yet known, so no comparison is possible with the *M. thermoacetica* genome. The functions of the tetrahydrofolate salvage pathway in *M. thermoacetica*, which regenerates tetrahydrofolate from 5- or 10-formyl-tetrahydrofolate could be accomplished by methylene-tetrahydrofolate dehydrogenase/methenyl-tetrahydrofolate cyclohydrolase of the Wood–Ljungdahl pathway (Moth_1516) and Moth_2045, a phosphoribosylglycinamide formyl-transferase.

Moorella thermoacetica has a complete pathway of cobalamin biosynthesis. Cobalamin is a central cofactor in the autotrophic metabolism of *M. thermoacetica*, linking the methyl and carbonyl branches of the Wood–Ljungdahl pathway, with the methylated corrinoid iron–sulfur protein serving as a substrate of the acetyl-CoA synthase reaction. *Moorella thermoacetica* contains over 20 different types of cobalamin or precursors that total 300–700 nmol g^{-1} of cells (Ljungdahl *et al.*, 1966). The levels of corrinoids and their structural diversity in anaerobic bacteria have been reviewed (Stupperich *et al.*, 1990). Both the anaerobic branch of corrin ring synthesis and the common branch of the adenosyl-cobalamin pathway are intact. Although several genes in the aerobic pathway of corrin ring synthesis are also present in the *M. thermoacetica* genome, these genes are not specific to the aerobic route and genes that are unique to the aerobic pathway are absent in *M. thermoacetica*. Almost all of the cobalamin biosynthesis genes are located within a single gene cluster (Moth_1087–1104) in the *M. thermoacetica* genome. ABC transporters, which are encoded by *cbiM*, *N*, *Q* and *O*, mediate high affinity transport of cobalt ions for B$_{12}$ synthesis. Three gene clusters (Moth_1219–1222, Moth_1861–1863 and Moth_2427–2429) have combinations of *cbiM*, *N, Q* and *O* genes.

As expected from the numbers of DMSO reductase and xanthine oxidase family proteins encoded in the

M. thermoacetica genome, this organism has genes that could be used to synthesize molybdenum cofactors (Fig. S6). Most of these were found based on their homology to *E. coli* and *Arabidopsis thaliana* genes (Schwartz, 2005). *Moorella thermoacetica* has homologues of all *E. coli* molybdenum cofactor synthesis genes except for *moaE*, which encodes the large subunit of molybdopterin synthase. Some *M. thermoacetica* homologues of *mogA*, *moaC*, *moaA* and *moeA* are found in one cluster of genes (Moth_2132–2137), along with a gene belonging to pfam03473, which is a family of predicted sulfur-carrying proteins which are sometimes associated with molybdenum cofactor sulfurase domains, as in molybdopterin synthase. Another gene cluster (Moth_1819–1824) has a gene homologous to *mobB* from *E. coli*, the function of which is not known, another moeA homologue, a gene that matches the C-terminal half of *mobA*, and three subunits of a possible ABC-type molybdate transporter, homologous to *modABC* from *E. coli*.

Transporter genes

Two hundred and ten transport genes, constituting approximately one hundred and thirty transport systems, were identified in the *M. thermoacetica* genome (Table S7). These include six α-helical membrane channels for the transport of anions, cations and small solutes. The genome encodes 61 secondary permeases, which transport a variety of substrates that include amino acids, sugars, anions, cations, nucleotides and toxic drugs. Sixteen belong to the Major Facilitator Superfamily (MFS), two are RND-type transporters, six are DMT-type drug and metabolite transporters and three belong to the MOP superfamily. One hundred three proteins encoded in the genome constitute approximately 35 ABC-type transport systems, which are predicted to transport anions, cations, peptides, amino acids, vitamins, heavy metal ions and toxic drugs. The genome also encodes four polypeptides that form a single fructose-type sugar-transporting phosphotransferase system. Three cation-transporting P-type ATPases were also recognized in the genome. Approximately 19 putative transporters with unknown substrates were also identified in the genome.

Motility genes

Moorella thermoacetica is not motile, although stains show peritrichous flagella (Fontaine *et al.*, 1942). Thirty-five *M. thermoacetica* genes encode proteins forming entire components and substructure of flagella, such as motor, switch complex, basal body and filament (Fig. S7). These proteins fall into at least six groups: hook-associated proteins (FlgK and FlgL), basal and hook proteins (FlgE), capping proteins (FliD), biosynthesis secretory proteins (FlhA, FlhB, FliF, FliH and FliI), flagella formation proteins, motor proteins (FliG and FliM) and basal proteins (FlgC and FlgB). However, genes encoding the accessory flagellar assembly proteins, FlgA, FlhC, FlhD and FliT, could not be identified in the *M. thermoacetica* genome. The organization of the genes for the general component of the flagellar proteins export apparatus appears to be similar to that found in *Bacillus subtilis* (Macnab, 2004). A similar gene cluster exists in *Thermoanaerobacter tengcongensis*, which also does not have functional flagella in cultures, but seems to have many of the genes for flagellar assembly and chemotaxis (Xue *et al.*, 2001; Bao *et al.*, 2002). From the genome annotation, *M. thermoacetica* also has an elaborate cascade of chemotaxis genes, including six chemoreceptors, together with a complete set of flagellar genes, most located within three large clusters of about 70 genes (Moth_0740–7, Moth_0764–91, Moth_0797-Moth_0805). The only exception is CheZ, which is not found in the *Moorella* or *Thermoanaerobacter tengcongensis* genomes. CheZ, a protein known to accelerate dephosphorylation of the response regulator CheY, has only been found in a few non-thermophilic Gram-negative bacteria, such as *E. coli* (Blattner *et al.*, 1997), *Pseudomonas aeruginosa* (Stover *et al.* 2000) and *Vibrio cholerae* (Heidelberg *et al.* 2000), and it is not essential for flagellar motility and chemotaxis (Scharf *et al.*, 1998). The presence of genes involved in flagellar structure and movement in both non-motile *M. thermoacetica* and *T. tengcongensis* suggests either that they might be inactive or activated only under certain environmental conditions or that they have lost function only recently before divergence of these two species. In the *T. tengcongensis* proteomics study, two key gene products of the flagella formation, motor proteins (FliG and FliM) and basal proteins (FlgC and FlgB) were not detected, supporting the suspicion that incomplete gene expression may cause the absence of flagella in the cultured cells (Wang *et al.*, 2004).

Discussion

This paper describes the first genomic sequence of an acetogen. The genomic results clearly indicate that the trait of acetogenesis is a metabolic, not a phylogenetic trait. To identify a set of genes that are responsible for the acetogenic lifestyle, the *M. thermoacetica* genome was compared with those of all *Firmicutes*, *Chlorflexi* and *Deltaproteobacteria* sequences available at IMG. Surprisingly, the only *M. thermoacetica* genes with homologues in the other bacterial genomes that also have acetyl-CoA synthase, but lack homologues in any of the genome sequences without acetyl-CoA synthase, are acetyl-CoA synthase itself and the two subunits of the corrinoid iron–sulfur protein (CFeSP). While the genes encoding the

H$_4$folate-dependent enzymes in the Eastern part of the pathway are scattered around the genome, the 'core set of Wood–Ljungdahl genes' (*acsA-acsE*) in the Western part of the pathway are colocalized. Being required for general C-1 metabolism in all organisms, the Eastern branch genes are ubiquitous; thus, we suggest that acquisition of the *acs* gene cluster by horizontal transfer might be sufficient for gaining the ability to utilize the Wood–Ljungdahl pathway. For some organisms, this would confer the ability to grow as a homoacetogen; for others (sulfate reducers and aceticlastic methanogens), the ability to oxidize acetate. Acquisition of the core Wood–Ljungdahl genes would allow a wide variety of organisms, including archaea and a broad range of bacterial phyla (*Firmicutes*, *Chlorflexi* and *Deltaproteobacteria*), to fix CO and CO$_2$ by the reductive acetyl-CoA pathway. The simple strategy of the Wood–Ljungdahl pathway of successively joining two one-carbon compounds to make a two-carbon compound has been envisioned to be the earliest form of metabolism (Martin and Russell, 2007).

The genomic analysis also revealed a second CODH (Moth_1972), which has not been characterized biochemically. Other genomes contain multiple CODHs, like *C. hydrogenoformans*, which contains five CODHs (Wu *et al.*, 2005). The only genes that could be identified to encode a phosphotransacetylase in *M. thermoacetica* (Moth_1181 and Moth_0864), are homologous to the recently characterized novel enzyme from *Salmonella enterica* called *pduL* (Liu *et al.*, 2007); thus, it is important to test this function by performing biochemical assays on these putative phosphotransacetylases.

Perhaps not surprisingly, given the metabolic diversity of acetogenic microbes, the largest COG expansions in *M. thermoacetica*, relative to the other *Thermoanaerobacteriaceae* genome sequences, involve electron transfer proteins. How electron donors and acceptors interact with respiratory pathways in *M. thermoacetica* also remains an open question. Although the enzymes in the Wood–Ljungdahl pathway have been characterized, it has been unclear how acetogens gain energy, as there is no net ATP synthesis by substrate-level phosphorylation. Homologues of NADH dehydrogenase I (Table 3) and archaeal heterodisulfide reductase (Table 5) are present in the genome. Other genes identified that may play important roles in electron transfer linked proton translocation are Moth_2184–90, with significant homology to the genes encoding the subunits of *E. coli*'s hydrogenase 4 (Table 4), which has been proposed to have proton translocating ability (Andrews *et al.*, 1997). These various sets of genes could encode partial or complete membrane-bound complexes that would function in anaerobic respiration by generating a transmembrane proton gradient for use by the F$_1$F$_0$ATP synthase (Moth_2377–84). Localization of genes encoding homologues of heterodisulfide

reductase (Moth_1194-Moth_1196), a hydrogenase (Moth_1193) and 5,10-methylene-H$_4$folate reductase (Moth_1191) near the *acs* gene cluster are consistent with a suggestion that the reductase and a hydrogenase might link to CODH to generate a proton motive force, driving anaerobic respiration in acetogens (Thauer *et al.*, 1977).

Moorella thermoacetica contains eight genes encoding molybdopterin-containing proteins, that are likely to encode the reductases that allow growth on nitrate (Fröstl *et al.*, 1996), dimethylsulfoxide and thiosulfate (Beaty and Ljungdahl, 1991), as well as reductases for other electron acceptors not yet identified (Table 7). These molybdopterin-dependent proteins would then link to the membrane-bound respiratory systems described in the previous paragraph. The genomic information described here should help identify the specific membrane-associated proteins that are recruited to allow growth on these various electron acceptors. It is likely that these electron transfer components are differentially expressed under different growth conditions and the large number of them might be part of *M. thermoacetica*'s strategy for fine-tuning its growth to specific environmental conditions. The presence of so many different sets of genes that are likely to encode proteins involved in energy generation by electron and proton transport (for example, the presence of three heterodisulfide reductase gene clusters) could be part of this strategy. That cytochrome *b*, which is found in membranes of *M. thermoacetica* cells grown using the Wood–Ljungdahl pathway, is downregulated in cells grown on nitrate (Arendsen *et al.*, 1999) indicates that the respiratory pathway used in connection with the Wood–Ljungdahl pathway is different than the one used during growth on nitrate, at least in some components.

Other expanded gene groups that were identified in *M. thermoacetica* include methyltransferases, which are required for growth on various methyl group donors (methanol, methoxylated aromatics), and ABC-type nitrate/sulfonate/bicarbonate transport systems, which also perhaps reflect the ability of acetogens to grow on a wide variety of substrates. Methyltransferase genes that link to the Wood–Ljungdahl pathway (besides *acsCDE*) include *mtaABC* (Moth_1208-1209) encoding the methanol:methyl-H$_4$folate methyltransferase system, *mtvABC* (Moth_1316–1318 and Moth_0385–0387), which transfers the methyl group of methoxylated aromatic compounds to H$_4$folate, and several other methyltransferase genes whose specific biochemical functions have not been assigned.

We were unable to predict the pathways by which *M. thermoacetica* metabolizes the two-carbon organic acids oxalate, glyoxylate and glycolate (Seifritz *et al.*, 1999). The most reduced compound among these is glycolate, which is oxidized to glyoxylate in *E. coli* (Pellicer *et al.*, 1996) by an FMN- and H$_2$O$_2$-dependent glycolate

oxidase, which is encoded by the *glcDEF* genes that are homologous to genes located in two gene clusters in *M. thermoacetica*. However, it seems unlikely that the anaerobe *M. thermoacetica* uses an oxidase mechanism. Homologues of malate synthase (GlcB), which catalyses formation of malate from glyoxylate and acetyl-CoA, or glyoxylate carboligase (converts 2 mol of glyoxylate to form tartronate semialdehyde), were not found in *M. thermoacetica*, which is consistent with the lack of enzymatic activities for these proteins in extracts of cells grown on glyoxylate (Seifritz *et al.*, 2002). Perhaps malate is formed from glyoxylate and acetyl-CoA by a glyoxylate reductase homologue (Moth_ 1954). Homologues of proteins in the pathway of oxalate metabolism in *Oxalobacter formigenes*, which begins with oxalyl-CoA formation, were also missing from the *M. thermoacetica* genome. A hint for the oxalate metabolic pathway is that a 33 kDa protein is induced when *M. thermoacetica* is grown on oxalate (Daniel and Drake, 1993); however, no sequence information for this protein has yet been described. Postgenomic studies will likely uncover how acetogens use the *glc* gene cluster and/or other genes to metabolize oxalate, glyoxylate and glycolate under anaerobic conditions. Identifying the mechanism of growth on oxalate should be particularly interesting, because during oxalate-dependent growth, *M. thermoacetica* uses CO_2 preferentially over nitrate as a terminal electron acceptor. Oxalate is the only substrate for *M. thermoacetica* for which this preference has been found.

Knowing the complete genome sequence of *M. thermoacetica* will be a valuable tool to help understand the growth of this organism and the acetogenic lifestyle. The genome sequence enables protein expression studies and microarrays, which should address important questions arising from annotation of this genome sequence. For example, growth of *M. thermoacetica* on oxalate and glyoxylate, followed by 2D electrophoresis analysis of proteins could uncover shifts in protein expression that would uncover what pathways are used. Protein and gene expression studies, on cells grown on different electron acceptors, could help identify the pathways of anaerobic respiration in *M. thermoacetica*. Such experiments will uncover how electrons are passed from electron donor to electron acceptor during growth, and how *M. thermoacetica* switches among different donors and acceptors, depending on what substrates are available in its environment. Similar studies may reveal how *M. thermoacetica* grows in different environments and in different ecological settings; for example, how its metabolism and gene expression patterns change when an acetogen is part of a microbial community or in coculture with a methanogen or sulfate reducer that can act as a H_2 sink. In turn, proteomic and microarray studies should help us improve the annotation of the genome, by allowing us to identify specific biochemical functions and metabolic roles for genes of unknown or partially known function.

Experimental procedures

DNA extraction and purification

Moorella thermoacetica cells were grown at 55°C in the presence of 100 mM glucose (18 g l^{-1}) under 100% CO_2 in the medium described by Ljungdahl and Andreesen (Ljungdahl and Andreesen, 1978), containing yeast extract and tryptone. Another sample of DNA was prepared from cells grown at 58°C on 200 mM methanol under 100% CO_2 in Drake's medium (Seifritz *et al.*, 1993). The DNA was then isolated according to the procedure of Marmur (Marmur, 1961) and stored frozen at −20°C in 10 mM Tris/HCl, pH 8.0 and 1 mM EDTA at a concentration of 438 μg ml^{-1}.

Genome sequencing

The genome of *M. thermoacetica* was sequenced at the Joint Genome Institute (JGI) using a combination of 3, 6, 7 and 37 kb (fosmid) DNA libraries. All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov/. Draft assemblies were based on 41 897 total reads. All four libraries provided 18× coverage of the genome. The Phred/Phrap/Consed software package (http://www.phrap.com) was used for sequence assembly and quality assessment (Han and Chain, 2006). After the shotgun stage, reads were assembled with parallel phrap (High Performance Software, LLC). Possible misassemblies were corrected with Dupfinisher (Abraham *et al.*, 2006) or transposon bombing of bridging clones (Epicentre Biotechnologies, Madison, WI). Gaps between contigs were closed by editing in Consed, custom primer walk and PCR amplification (Roche Applied Science, Indianapolis, IN). A total of 3345 additional reactions were necessary to close gaps and to raise the quality of the finished sequence. The completed genome sequence of *M. thermoacetica* contains 45 242 reads, achieving an average of eightfold sequence coverage per base with an error rate less than 1 in 100 000.

Annotation

Automated gene modelling was completed by combining results from Critica, Generation and Glimmer modelling packages, and comparing the translations to GenBank's nonredundant database using the basic local alignment search tool for proteins (BLASTP). Predicted coding sequences were subjected to manual analysis using JGI's gene models quality assessment pipeline. The predicted functions were further analysed by using the Integrated Microbial Genomes (IMG) annotation pipeline (http://img.jgi.doe.gov; Markowitz *et al.*, 2006).

Comparative analysis

Comparative analysis of *M. thermoacetica* ATCC 39073, *Thermoanaerobacter tengcongensis* MB4 and *Thermoa-*

naerobacter ethanolicus 39E was performed by using a set of tools available at IMG. Unique and orthologous *M. thermoacetica* genes were identified by using BLASTP (cut-off scores of $E < 10^{-2}$ and 20% identity and reciprocal hits with cut-offs of $E < 10^{-5}$ and 30% respectively). The transporter database (TCDB; http://www.tcdb.org; Saier *et al.*, 2006) was used to identify all the transporters in the genome of *M. thermoacetica*. Twin arginine transport signal sequences were predicted by the TatP server at http://www. cbs.dtu.dk/services/TatP/ (Bendtsen *et al.*, 2005). Metabolic pathways were constructed with reference to MetaCyc (Caspi *et al.*, 2006) and PubMed. Identification numbers given in the text for individual genes are IMG locus tag numbers, which can be used to search for genes at the Integrated Microbial Genomes website.

Phylogenetic trees

The phylogenetic tree was made using prealigned 16S rRNA sequences derived from the Ribosomal Database site (http:// rdp.cme.msu.edu/index.jsp) (Cole *et al.*, 2007). Manual alignment adjustments were made as needed with the assistance of the BioEdit multiple alignment tool of Hall (http://www.mbio. ncsu.edu/BioEdit/bioedit.html). The refined multiple alignment was used as input for generation of a phylogenetic tree using the program package version 3.1 (Kumar *et al.*, 2004). The Neighbor Joining (Saitou and Nei, 1987), Minimum Evolution (Nei *et al.*, 1998; Takahashi and Nei, 2000), UPGMA (Nei and Kumar, 2000) and Maximum Parsimony (Fitch, 1971) methods with the default settings were employed to obtain phylogenetic trees. Both the distance- and character-based trees yielded similar clusters and arrangements of taxa.

Nucleotide sequence accession number

The sequence data described here have been deposited in GenBank under Accession No. CP000232.

Acknowledgements

References

Abraham, L.O., Li, Y., and Zamble, D.B. (2006) The metal- and DNA-binding activities of *Helicobacter pylori* NikR. *J Inorg Biochem* **100:** 1005–1014.

Adrian, N.R., and Arnett, C.M. (2004) Anaerobic biodegradation of hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX) by *Acetobacterium malicum* strain HAAP-1 Isolated from a methanogenic mixed culture. *Curr Microbiol* **48:** 332–340.

Andreesen, J.R., and Ljungdahl, L.G. (1973) Formate dehydrogenase of *Clostridium thermoaceticum*: incorporation of selenium-75, and the effects of selenite, molybdate, and tungstate on the enzyme. *J Bacteriology* **116:** 867–873.

Andreesen, J.R., Schaupp, A., Neurauter, C., Brown, A., and Ljungdahl, L.G. (1973) Fermentation of glucose, fructose, and xylose by *Clostridium thermoaceticum*: effect of metals on growth yield, enzymes, and the synthesis of acetate from CO_2. *J Bacteriology* **114:** 743–751.

Andrews, S.C., Berks, B.C., McClay, J., Ambler, A., Quail, M.A., Golby, P., and Guest, J.R. (1997) A 12-cistron *Escherichia coli* operon (hyf) encoding a putative proton-translocating formate hydrogenlyase system. *Microbiology* **143:** 3633–3647.

Arendsen, A.F., Soliman, M.Q., and Ragsdale, S.W. (1999) Nitrate-dependent regulation of acetate biosynthesis and nitrate respiration by *Clostridium thermoaceticum*. *J Bacteriol* **181:** 1489–1495.

Balch, W.E., Schoberth, S., Tanner, R.S., and Wolfe, R.S. (1977) *Acetobacterium*, a new genus of hydrogen-oxidizing, carbon-dioxide-reducing, anaerobic bacteria. *Int J Syst Bacteriol* **27:** 355–361.

Bao, Q., Tian, Y., Li, W., Xu, Z., Xuan, Z., Hu, S., *et al.* (2002) A complete sequence of the *Thermoanaerobacter tengcongensis* genome. *Genome Res* **12:** 689–700.

Barker, H.A., and Kamen, M.D. (1945) Carbon dioxide utilization in the synthesis of acetic acid by *Clostridium thermoaceticum*. *Proc Natl Acad Sci* **31:** 219–225.

Beaty, S.P., and Ljungdahl, L.G. (1991) Growth of *Clostridium thermoaceticum* on methanol, ethanol, propanol, and butanol in medium containing either thiosulfate or dimethylsulfoxide. In *Abstracts of the Annual Meeting of the American Society for Microbiology*, abstr. K-131, p. 236. Washington, DC: American Society for Microbiology.

Bendtsen, J., Nielsen, H., Widdick, D., Palmer, T., and Brunak, S. (2005) Prediction of twin-arginine signal peptides. *BMC Bioinformatics* **6:** 167.

Blattner, F.R., Plunkett, G. III, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1462.

Blokesch, M., Paschos, A., Bauer, A., Reissmann, S., Drapal, N., and Böck, A. (2004) Analysis of the transcarbamoylation-dehydration reaction catalyzed by the hydrogenase maturation proteins HypF and HypE. *Eur J Biochem* **271:** 3428–3436.

Boynton, Z.L., Bennett, G.N., and Rudolph, F.B. (1996) Cloning, sequencing, and expression of genes encoding phosphotransacetylase and acetate kinase from *Clostridium acetobutylicum* ATCC 824. *Appl Environ Microbiol* **62:** 2758–2766.

Braun, M., Mayer, F., and Gottschalk, G. (1981) *Clostridium aceticum* (Wieringa), a microorganism producing acetic acid from molecular hydrogen and carbon dioxide. *Arch Microbiol* **128:** 288–293.

Breznak, J.A. (1994) Acetogenesis from carbon dioxide in termite guts. In *Acetogenesis.* Drake, H.L. (ed.). New York, USA: Chapman & Hall, pp. 303–330.

Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* **34:** D511–D516.

Chassard, C., and Bernalier-Donadille, A. (2006) H_2 and acetate transfers during xylan fermentation between a butyrate-producing xylanolytic species and hydro-

genotrophic microorganisms from the human gut. *FEMS Microbiol Lett* **254:** 116–122.

Clark, J.E., and Ljungdahl, L.G. (1982) Purification and properties of 5,10-methenyltetrahydrofolate cyclohydrolase. *J Biol Chem* **257:** 3833–3836.

Clark, J.E., and Ljungdahl, L.G. (1984) Purification and properties of 5,10-methylenetetrahydrofolate reductase, an iron-sulfur flavoprotein from clostridium formicoaceticum. *J Biol Chem* **259:** 10845–10889.

Cole, J.R., Chai, B., Farris, R., Wang, Q., Kulam-Syed-mohideen, A.S., McGarrell, D., *et al.* (2007) The ribosomal database project (RDP-II): introducing *myRDP* space and quality controlled public data. *Nucleic Acids Res* **35:** D169–D172.

Collins, M., Lawson, P., Willems, A., Cordoba, J., Fernandez-Garayzabal, J., Garcia, P., *et al.* (1994) The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. *Int J Syst Bacteriol* **44:** 812–826.

Daniel, S.L., and Drake, H.L. (1993) Oxalate- and glyoxylate-dependent growth and acetogenesis by *Clostridium thermoaceticum*. *Appl Environ Microbiol* **59:** 3062–3069.

Das, A., and Ljungdahl, L. (1997) Composition and primary structure of the F1F0 ATP synthase from the obligately anaerobic bacterium *Clostridium thermoaceticum*. *J Bacteriol* **179:** 3746–3755.

Das, A., Hugenholtz, J., Van Halbeek, H., and Ljungdahl, L.G. (1989) Structure and function of a menaquinone involved in electron transport in membranes of *Clostridium thermoautotrophicum* and *Clostridium thermoaceticum*. *J Bacteriol* **171:** 5823–5829.

Das, A., Fu, Z.-Q., Tempel, W., Liu, Z.-J., Chang, J., Chen, L., *et al.* (2007) Characterization of a corrinoid protein involved in the C1 metabolism of strict anaerobic bacterium *Moorella thermoacetica*. *Proteins: Struct. Funct. Bioinf.* **67:** 167–176.

Deppenmeier, U. (2004) The membrane-bound electron transport system of *Methanosarcina* species. *J Bioenerg Biomembr* **36:** 55–64.

Drake, H.L. (1982) Demonstration of hydrogenase in extracts of the homoacetate-fermenting bacterium *Clostridium thermoaceticum*. *J Bacteriol* **150:** 702–709.

Drake, H.L., and Daniel, S.L. (1997) Acetogenic bacteria: what are the *in situ* consequences of their diverse metabolic versatilities? *BioFactors* **6:** 13–24.

Drake, H.L., Daniel, S.L., Matthies, C., and Küsel, K. (1994) Acetogenesis, acetogenic bacteria, and the acetyl-CoA pathway: past and current perspectives. In *Acetogenesis*. Drake, H.L. (ed). New York, USA: Chapman & Hall, pp. 3–60.

Drake, H.L., Hu, S.I., and Wood, H.G. (1981) Purification of five components from *Clostridium thermoaceticum* which catalyze synthesis of acetate from pyruvate and methyltetrahydrofolate. Properties of phosphotransacetylase. *J Biol Chem* **256:** 11137–11144.

Drake, H.L., Kusel, K., and Matthies, C. (2002) Ecological consequences of the phylogenetic and physiological diversities of acetogens. *Antonie Van Leeuwenhoek* **81:** 203–213.

Ferry, J.G. (1999) Enzymology of one-carbon metabolism in methanogenic pathways. *FEMS Microbiol Rev* **23:** 13–38.

Fischer, F., Lieske, R., and Winzer, K. (1932) Biologische gasreaktionen. II. Gber die bildung von essigs ure bei der biologischen umsetzung von kohlenoxyd und kohlens ure mit wasserstoff zu methan. *Biochem Z* **245:** 2–12.

Fitch, W.M. (1971) Towards defining the course of evolution: minimum change for a specific tree topology. *Syst Zool* **20:** 406–416.

Fontaine, F.E., Peterson, W.H., McCoy, E., Johnson, M.J., and Ritter, G.J. (1942) A new type of glucose fermentation by *Clostridium thermoaceticum* n. sp. *J Bacteriol* **43:** 701–715.

Fox, J.D., He, Y., Shelver, D., Roberts, G.P., and Ludden, P.W. (1996) Characterization of the region encoding the CO-induced hydrogenase of *Rhodospirillum rubrum*. *J Bacteriol* **178:** 6200–6208.

Fröstl, J.M., Seifritz, C., and Drake, H.L. (1996) Effect of nitrate on the autotrophic metabolism of the acetogens *Clostridium thermoautotrophicum* and *Clostridium thermoaceticum*. *J Bacteriol* **187:** 4597–4603.

Furdui, C., and Ragsdale, S.W. (2002) The roles of coenzyme a in the pyruvate: ferredoxin oxidoreductase reaction mechanism: rate enhancement of electron transfer from a radical intermediate to an iron-sulfur cluster. *Biochemistry* **41:** 9921–9937.

Gencic, S., and Grahame, D.A. (2003) Nickel in subunit beta of the acetyl-CoA decarbonylase/synthase multienzyme complex in methanogens: catalytic properties and evidence for a binuclear Ni-Ni site. *J Biol Chem* **278:** 6101–6110.

Genest, O., Seduk, F., Ilbert, M., Méjean, V., and Iobbi-Nivol, C. (2006) Signal peptide protection by specific chaperone. *Biochem Biophys Res Comms* **339:** 991–995.

Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312:** 1355–1359.

Gon, S., Giudici-Orticoni, M.-T., Mejean, V., and Iobbi-Nivol, C. (2001) Electron transfer and binding of the c-type cytochrome TorC to the trimethylamine N-oxide reductase in *Escherichia coli*. *J Biol Chem* **276:** 11545–11551.

Gößner, A., Devereux, R., Ohnemüller, N., Acker, G., Stackebrandt, E., and Drake, H.L. (1999) Thermicanus aegyptius gen. nov., sp. nov., isolated from oxic soil, a fermentative microaerophile that grows commensally with the thermophilic acetogen Moorella thermoacetica. *Appl Environ Microbiol* **65:** 5124–5133.

Gottwald, M., Andreesen, J.R., LeGall, J., and Ljungdahl, L.G. (1975) Presence of cytochrome and menaquinone in *Clostridium formicoaceticum* and *Clostridium thermoaceticum*. *J Bacteriol* **122:** 325–328.

Greening, R.C., and Leedle, J.A.Z. (1989) Enrichment and isolation of *Acetitomaculum ruminis*, gen. nov., sp. nov.: acetogenic bacteria from the bovine rumen. *Arch Microbiol* **151:** 399–406.

Han, C.S., and Chain, P. (2006) Finishing repeat regions automatically with Dupfinisher. In *Proceeding of the 2006 International Conference on Bioinformatics and Computational Biology*. Valafar, H. (ed). Las Vegas, USA: CSREA Press, pp. 141–146.

Harmsen, H., Van Kuijk, B., Plugge, C., Akkermans, A., De Vos, W., and Stams, A. (1998) Syntrophobacter fumaroxi-

dans sp. nov., a syntrophic propionate-degrading sulfate-reducing bacterium. *Int J Syst Bacteriol* **48:** 1383–1387.

Hatzixanthis, K., Clarke, T.A., Oubrie, A., Richardson, D.J., Turner, R.J., and Sargent, F. (2005) Signal peptide–chaperone interactions on the twin-arginine protein transport pathway. *Proc Natl Acad Sci* **102:** 8460–8465.

Hedderich, R. (2004) Energy-converting [NiFe] hydrogenases from archaea and extremophiles: ancestors of complex I. *J Bioenerg Biomembr* **36:** 65–75.

Heidelberg, J.F., Elsen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., *et al.* (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae. Nature* **406:** 477–484.

Heise, R., Muller, V., and Gottschalk, G. (1989) Sodium dependence of acetate formation by the acetogenic bacterium *Acetobacterium woodii. J Bacteriol* **171:** 5473–5478.

Himes, R.H., and Wilder, T. (1968) Formyltetrahydrofolate synthetase. Effect of pH and temperature on the reaction. *Arch Biochem Biophys* **124:** 230.

Hugenholtz, J., Ivey, D.M., and Ljungdahl, L.G. (1987) Carbon monoxide-driven electron transport in *Clostridium thermoautotrophicum* membranes. *J Bacteriol* **169:** 5845–5847.

Jeon, W.B., Cheng, J., and Ludden, P.W. (2001) Purification and characterization of membrane-associated CooC protein and its functional role in the insertion of nickel into carbon monoxide dehydrogenase from *Rhodospirillum rubrum. J Biol Chem* **42:** 38602–38609.

Jongbloed, J.D.H., Grieger, U., Antelmann, H., Hecker, M., Nijland, R., Bron, S., and van Dijl, J.M. (2004) Two minimal Tat translocases in *Bacillus. Mol Microbiol* **54:** 1319–1325.

Kerby, R., and Zeikus, J.G. (1983) Growth of *Clostridium thermoaceticum* on H_2/CO_2 or CO as energy source. *Curr Microbiol* **8:** 27–30.

Kisker, C., Schindelin, H., Baas, D., Rétey, J., Merkenstock, R.U., and Kroneck, P.M.H. (1999) A structural comparison of molybdenum cofactor-containing enzymes. *FEMS Microbiol Rev* **22:** 503–521.

Kumar, S., Tamura, K., and Nei, M. (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* **5:** 150–163.

Kusel, K., Wagner, C., and Drake, H.L. (1999) Enumeration and metabolic product profiles of the anaerobic microflora in the mineral soil and litter of a beech forest. *FEMS Microbiol Ecol* **29:** 91–103.

Ladapo, J., and Whitman, W. (1990) Method for isolation of auxotrophs in the methanogenic archaebacteria: role of the acetyl-CoA pathway of autotrophic CO_2 fixation in *Methanococcus maripaludis. PNAS* **87:** 5598–5602.

Larsen, Ø., Lien, T., and Birkeland, N.-K. (1999) Dissimilatory sulfite reductase from *Archaeoglobus profundus* and *Desulfotomaculum thermocisternum*: phylogenetic and structural implications from gene sequences. *Extremophiles* **3:** 63–70.

Leadbetter, J.R., Schmidt, T.M., Graber, J.R., and Breznak, J.A. (1999) Acetogenesis from H_2 plus CO_2 by spirochetes from termite guts. *Science* **283:** 686–689.

Le Van T.D., Robinson, J.A., Ralph, J., Greening, R.C., Smolenski, W.J., *et al.* (1998) Assessment of reductive acetogenesis with indigenous ruminal bacterium popula-

tions and *Acetitomaculum ruminis. Appl Environ Microbiol* **64:** 3429–3436.

Li, L.-F., Ljungdahl, L., and Wood, H.G. (1966) Properties of nicotinamide adenine dinucleotide phosphate-dependent formate dehydrogenase from *Clostridium thermoaceticum. J Bacteriol* **92:** 405–412.

Liesack, W., Bak, F., Kreft, J.U., and Stackebrandt, E. (1994) *Holophaga foetida* gen. nov., sp. nov., a new, homoacetogenic bacterium degrading methoxylated aromatic compounds. *Arch Microbiol* **162:** 85–90.

Liu, S., and Suflita, J.M. (1993) H_2-CO_2-dependent anaerobic O-demethylation activity in subsurface sediments and by an isolated bacterium. *Appl Environ Microbiol* **59:** 1325–1331.

Liu, Y., Leal, N.A., Sampson, E.M., Johnson, C.L.V., Havemann, G.D., and Bobik, T.A. (2007) PduL Is an evolutionarily distinct phosphotransacylase involved in B12-dependent 1,2-propanediol degradation by *Salmonella enterica* serovar Typhimurium LT2. *J Bacteriol* **189:** 1589–1596.

Ljungdahl, L.G. (1986) The autotrophic pathway of acetate synthesis in acetogenic bacteria. *Annu Rev Microbiol* **40:** 415–450.

Ljungdahl, L.G. (1994) The acetyl-CoA pathway and the chemiosmotic generation of ATP during acetogenesis. In *Acetogenesis*. Drake, H.L. (ed.). New York, USA: Chapman & Hall, pp. 63–87.

Ljungdahl, L.G., and Andreesen, J.R. (1975) Tungsten, a component of active formate dehydrogenase from *Clostridium thermoaceticum. FEBS Lett* **54:** 279–282.

Ljungdahl, L.G., and Andreesen, J.R. (1978) Formate dehydrogenase, a selenium-tungsten enzyme from *Clostridium thermoaceticum. Methods Enzymol* **53:** 360–372.

Ljungdahl, L.G., Irion, E., and Wood, H.G. (1966) Role of corrinoids in the total synthesis of acetate from CO_2 by *Clostridium thermoaceticum. Fed Proc* **25:** 1642.

Lovell, C.R., Przybyla, A., and Ljungdahl, L.G. (1988) Cloning and expression in *Escherichia coli* of the *Clostridium thermoaceticum* gene encoding thermostable formyltetrahydrofolate synthetase. *Arch Microbiol* **149:** 280–285.

Lovell, C.R., Przybyla, A., and Ljungdahl, L.G. (1990) Primary structure of the thermostable formyltetrahydrofolate synthetase from *Clostridium thermoaceticum. Biochemistry* **29:** 5687–5694.

Lundie, L.L., Jr and Drake, H.L. (1984) Development of a minimally defined medium for the acetogen *Clostridium thermoaceticum. J Bacteriol* **159:** 700–703.

Lutz, S., Jacobi, A., Schlensog, V., Böhm, R., Sawers, G., and Böck, A. (1991) Molecular characteriyation of an operon (hyp) necessary for the activity of the three hydrogenase isoenzymes in *Escherichia coli. Mol Microbiol* **5:** 123–135.

Macbeth, T.W., Cummings, D.E., Spring, S., Petzke, L.M., and Sorenson, K.S., Jr (2004) Molecular characterization of a dechlorinating community resulting from *in situ* biostimulation in a trichloroethene-contaminated deep, fractured basalt aquifer and comparison to a derivative laboratory culture. *Appl Environ Microbiol* **70:** 7329–7341.

McGuire, J.J., and Rabinowitz, J.C. (1978) Studies on the mechanism of formyltetrahydrofolate synthetase. The *Peptococcus aerogenes* enzyme. *J Biol Chem* **253:** 1079.

Mackie, R.I., and Bryant, M.P. (1994) Acetogenesis and the rumen: syntrophic relationships. In *Acetogenesis*. Drake, H.L. (ed). New York, USA: Chapman & Hall, pp. 331–364.

Macnab, R.M. (2004) Type III flagellar protein export and flagellar assembly. *Biochim Biophys Acta* **1694:** 207–217.

Markowitz, V.M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., *et al.* (2006) The integrated microbial genomes (IMG) system. *Nucleic Acids Res* **34:** D344–D348.

Marmur, J. (1961) A procedure for the isolation of deoxyribonucleic acid from microorganisms. *J Mol Biol* **3:** 208–218.

Martin, W., and Russell, M.J. (2007) On the origin of biochemistry at an alkaline hydrothermal vent. *Philos Trans R Soc Lond B Biol Sci* **362:** 1887–1925.

Maupin-Furlow, J.A., and Ferry, J.G. (1996) Analysis of the CO dehydrogenase/acetyl-coenzyme A synthase operon of *Methanosarcina thermophila. J Bacteriol* **178:** 6849–6856.

Meuer, J., Bartoschek, S., Koch, J., Künkel, A., and Hedderich, R. (1999) Purification and catalyc properties of Ech hydrogenase from *Methanosarcina barkeri. Eur J Biochem* **265:** 235–335.

Moore, M.R., O'Brien, W.E., and Ljungdahl, L.G. (1974) Purification and characterization of nicotinamide adenine dinucleotide-dependent methylenetetrahydrofolate dehydrogenase from *Clostridium formicoaceticum. J Biol Chem* **249:** 5250–5253.

Muller, V. (2003) Energy conservation in acetogenic bacteria. *Appl Environ Microbiol* **69:** 6345–6353.

Naidu, D., and Ragsdale, S.W. (2001) Characterization of a three-component vanillate O-demethylase from *Moorella thermoacetica. J Bacteriol* **183:** 3276–3281.

Nei, M., and Kumar, S. (2000) *Molecular Evolution and Phylogenetics*. New York, USA: Oxford University Press.

Nei, M., Kumar, S., and Takahashi, K. (1998) The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. *Proc Natl Acad Sci USA* **95:** 12390–12397.

Neumann, A., Engelmann, T., Schmitz, R., Greiser, Y., Orthaus, A., and Diekert, G. (2004) Phenyl methyl ethers: novel electron donors for respiratory growth of *Desulfitobacterium hafniense* and *Desulfitobacterium* sp. strain PCE-S. *Arch Microbiol* **181:** 245–249.

Nonaka, H., Keresztes, G., Shinoda, Y., Ikenaga, Y., Abe, M., Naito, K., *et al.* (2006) Complete genome sequence of the dehalorespiring bacterium desulfitobacterium hafniense Y51 and comparison with *Dehalococcoides ethenogenes* 195. *J Bacteriol* **188:** 2262–2274.

Núñez, F.M., Pellicer, M.-T., Badía, J., Aguilar, J., and Baldomà, L. (2001) The gene *yghK* linked to the *glc* operon of *Escherichia coli* encodes a permease for glycolate that is structurally and functionally similar to L-lactate permease. *Microbiology* **147:** 1069–1077.

O'Brien, W.E., Brewer, J.M., and Ljungdahl, L.G. (1975) Chemical, physical and enzymatic comparisons of formyltetrahydrofolate synthetases from thermo- and mesophilic clostridia. In *Proceedings of the International Symposium on Enzymes and Proteins from Thermophilic Microorganisms, Structure and Functions*. Zuber, H. (ed). Basel, Switzerland: Birkhauser-Verlag, p. 249.

Odelson, D.A., and Breznak, J.A. (1983) Volatile fatty acid production by the hindgut microbiota of xylophagous termites. *Appl Environ Microbiol* **45:** 1602–1613.

Ohshima, T., Nunoura-Kominato, N., Kudome, T., and Sakuraba, H. (2001) A novel hyperthermophilic archaeal glyoxylate reductase from *Thermococcus litoralis*. Characterization, gene cloning, nucleotide sequence and expression in *Escherichia coli. Eur J Biochem* **268:** 4740–4747.

Ollivier, B., Caumette, P., Garcia, J.-L., and Mah, R.A. (1994) Anaerobic bacteria from hypersaline environments. *Microbiol Rev* **58:** 27–38.

Park, E.Y., Clark, J.E., DerVartanian, D.V., and Ljungdahl, L.G. (1991) 5,10-methylenetetrahydrofolate reductases: iron-sulfur-zinc flavoproteins of two acetogenic clostridia. In *Chemistry and Biochemistry of Flavoenzymes*. Miller, F. (ed). Boca Raton, Fl, USA: CRC Press, pp. 389–400.

Pellicer, M.-T., Badía, J., Aguilar, J., and Baldomà, L. (1996) *glc* locus of *Escherichia coli*: characterization of genes encoding the subunits of glycolate oxidase and the *glc* regulator protein. *J Bacteriol* **178:** 2051–2059.

Pester, M., and Brune, A. (2007) Hydrogen is the central free intermediate during lignocellulose degradation by termite gut symbionts. *ISME J* **1:** 551–565.

Peters, V., and Conrad, R. (1995) Methanogenic and other strictly anaerobic bacteria in desert soil and other oxic soils. *Appl Environ Microbiol* **61:** 1673–1676.

Ragsdale, S.W. (1997) The Eastern and Western branches of the Wood/Ljungdahl pathway: how the East and West were won. *Biofactors* **6:** 3–11.

Ragsdale, S.W., and Ljungdahl, L.G. (1984) Purification and properties of NAD-dependent 5,10-methylenetetrahydrofolate dehydrogenase from *Acetobacterium woodii. J Biol Chem* **259:** 3499–3503.

Roberts, D.L., James-Hagstrom, J.E., Garvin, D.K., Gorst, C.M., Runquist, J.A., Baur, J.R., *et al.* (1989) Cloning and expression of the gene cluster encoding key proteins involved in acetyl-CoA synthesis in *Clostridium thermoaceticum*: CO dehydrogenase, the corrinoid/FeS protein, and methyltransferase. *Proc Natl Acad Sci* **86:** 32–36.

Rosencrantz, D., Rainey, F.A., and Janssen, P.H. (1999) Culturable populations of *Sporomusa* spp. and *Desulfovibrio* spp. in the anoxic bulk soil of flooded rice microcosms. *Appl Environ Microbiol* **65:** 3526–3533.

Saier, M.H. Jr, Tran, C.V., and Barabote, R.D. (2006) TCDB: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res* **34:** D181–D186.

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

Salmassi, T.M., and Leadbetter, J.R. (2003) Analysis of genes of tetrahydrofolate-dependent metabolism from cultivated spirochaetes and the gut community of the termite *Zootermopsis angusticollis. Microbiology* **149:** 2529–2537 (cover illustration).

Sazanov, L.A., and Hinchliffe, P. (2006) Structure of the hydrophilic domain of respiratory complex I from thermus thermophilus. *Science* **311:** 1430–1436.

Scharf, B.E., Fahrner, K.A., and Berg, H.C. (1998) CheZ has no effect on flagellar motors activated by CheY13DK106YW. *J Bacteriol* **180:** 5123–5128.

Schauder, R., Preuß, A., Jetten, M.S., and Fuchs, G. (1988) Oxidative and reductive acetyl CoA/carbon monoxide dehydrogenase pathway in *Desulfobacterium autotrophicum*. *Arch Microbiol* **151:** 84–89.

Schaupp, A., and Ljungdahl, L.G. (1974) Purification and properties of acetate kinase from *Clostridium thermoaceticum*. *Arch Microbiol* **100:** 121–129.

Schink, B. (1997) Energetics of syntrophic cooperation in methanogenic degradation. *Microbiol Mol Biol Rev* **61:** 262–280.

Schink, B., Thiemann, V., Laue, H., and Friedrich, M.W. (2002) Desulfotignum phosphitoxidans sp. nov., a new marine sulfate reducer that oxidizes phosphite to phosphate. *Arch Microbiol* **177:** 381–391.

Schwartz, G. (2005) Molybdenum cofactor biosynthesis and deficiency. *Cell Mol Life Sci* **62:** 2792–2910.

Seifritz, C., Daniel, S.L., Gößner, A., and Drake, H.L. (1993) Nitrate as a preferred electron sink for the acetogen *Clostridium thermoaceticum*. *J Bacteriol* **175:** 8008–8013.

Seifritz, C., Fröstl, J.M., Drake, H.L., and Daniel, S.L. (1999) Glycolate as a metabolic substrate for the acetogen *Moorella thermoacetica*. *FEMS Microbiol Lett* **170:** 399–405.

Seifritz, C., Fröstl, J.M., Drake, H.L., and Daniel, S.L. (2002) Influence of nitrate on oxalate- and glyoxylate-dependent growth and acetogenesis by *Moorella thermoacetica*. *Arch Microbiol* **178:** 457–464.

Seifritz, C., Drake, H.L., and Daniel, S.L. (2003) Nitrite as an energy-conserving electron sink for the acetogenic bacterium *Moorella thermoacetica*. *Curr Microbiol* **46:** 329–333.

Seshadri, R., Adrian, L., Fouts, D.E., Eisen, J.A., Phillippy, A.M., Methe, B.A., *et al.* (2005) Genome sequence of the PCE-dechlorinating bacterium *Dehalococcoides ethenogenes*. *Science* **307:** 105–108.

Setzke, E., Hedderich, R., Heiden, S., and Thauer, R. (1994) H_2: heterodisulfide oxidoreductase complex from *Methanobacterium thermoautotrophicum.* Composition properties. *Eur J Biochem* **220:** 139–148.

Shannon, K.W., and Rabinowitz, J.C. (1988) Isolation and characterization of the *Saccharomyces cerevisiae* MIS1 gene encoding mitochondrial C_1-tetrahydrofolate synthase. *J Biol Chem* **263:** 7717.

So, C.M., Phelps, C.D., and Young, L.Y. (2003) Anaerobic transformation of alkanes to fatty acids by a sulfate-reducing bacterium, strain Hxd3. *Appl Environ Microbiol* **69:** 3892–3900.

Soboh, B., Linder, D., and Hedderich, R. (2002) Purification and catalytic properties of a CO-oxidizing: H_2-evolving enzyme complex from *Carboxydothermus hydrogenoformans*. *Eur J Biochem* **269:** 5712–5721.

Spormann, A.M., and Thauer, R.K. (1988) Anaerobic acetate oxidation to CO_2 by *Desulfotomaculum acetoxidans* – demonstration of enzymes required for the operation of an oxidative acetyl-CoA/carbon monoxide dehydrogenase pathway. *Arch Microbiol* **150:** 374–380.

Staben, C., and Rabinowitz, J.C. (1986) Nucleotide sequence of the *Saccharomyces cerevisiae ADE3* gene encoding C_1-tetrahydrofolate synthase. *J Biol Chem* **261:** 2986.

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., *et al.* (2000) Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* **406:** 959–964.

Strous, M., Pelletier, E., Mangenot, S., Rattei, T., Lehner, A., Taylor, M.W., *et al.* (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440:** 790–794.

Stupperich, E., Hammel, K.E., Fuchs, G., and Thauer, R.K. (1983) Carbon monoxide fixation into the carboxyl group of acetyl coenzyme A during autotrophic growth of *Methanobacterium*. *FEBS Lett* **152:** 21–23.

Stupperich, E., Eisinger, H.-J., and Schurr, S. (1990) Corrinoids in anaerobic bacteria. *FEMS Microbiol. Lett* **87:** 355–360.

Sun, A.Y., Ljungdahl, L., and Wood, H.G. (1969) Total Synthesis of acetate from CO_2 II. Purification and properties of formyltetrahydrofolate synthetase from *Clostridium thermoaceticum*. *J Bacteriol* **98:** 842–844.

Takahashi, K., and Nei, M. (2000) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Mol Biol Evol* **17:** 1251–1258.

Thauer, R.K., Jungermann, K., and Decker, K. (1977) Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol Rev* **41:** 100–180.

Tholen, A., and Brune, A. (1999) Localization and *in situ* activities of homoacetogenic bacteria in the highly compartmentalized hindgut of soil-feeding higher termites (*Cubitermes* spp.). *Appl Environ Microbiol* **65:** 4497–4505.

Wang, J., Xue, Y., Feng, X., Li, X., Wang, H., Li, W., *et al.* (2004) An analysis of the proteomic profile for *Thermoanaerobacter tengcongensis* under optimal culture conditions. *Proteomics* **4:** 136–150.

Warnecke, F., Luginbuhl, P., Ivanova, N., Ghassemian, M., Richardson, T.H., Stege, J.T., *et al.* (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450:** 560–565.

Wiegel, J., and Ljungdahl, L.G. (1981) *Thermoanaerobacter ethanolicus* gen. nov., spec. nov., a new, extreme thermophilic, anaerobic bacterium. *Arch Microbiol* **128:** 343–348.

Wieringa, K.T. (1936) Over het verdwijnhnen van waterstof en koolzuur onder anaerobe voorwaarden. *Antonie Van Leeuwenhoek* **3:** 263–273.

Wieringa, K.T. (1940) The formation of acetic acid from carbon dioxide and hydrogen by anaerobic spore-forming bacteria. *Antonie Van Leeuwenhoek* **6:** 251–262.

Wood, H.G. (1952) Fermentation of 3,4-C^{14}- and 1-C^{14}-labeled glucose by *Clostridium thermoaceticum*. *J Biolog Chem* **199:** 579–583.

Wood, H.G., and Ljungdahl, L.G. (1991) Autotrophic character of the acetogenic bacteria. In *Variation in Autotrophic Life*. Shively, J.M., and Barton, L.L. (eds). New York, USA: Academic Press, pp. 201–250.

Wu, M., Ren, Q., Durkin, A.S., Daugherty, S.C., Brinkac, L.M., Dodson, R.J., *et al.* (2005) Life in hot carbon monoxide: the complete genome sequence of *Carboxydothermus hydrogenoformans* Z-2901. *PLoS Genetics* **1:** e65.

Xue, Y., Xu, Y., Liu, Y., Ma, Y., and Zhou, P. (2001) *Thermoanaerobacter tengcongensis* sp. nov., a novel anaerobic, saccharolytic, thermophilic bacterium isolated

from a hot spring in Tengcong, China. *Int J Syst Evol Microbiol* **51:** 1335–1341.

Yamamoto, I., Saiki, T., Liu, S., and Ljungdahl, L. (1983) Purification and properties of NADP-dependent formate dehydrogenase from *Clostridium thermoaceticum*, a tungsten-selenium-iron protein. *J Biol Chem* **258:** 1826–1832.

Ye, Q., Roh, Y., Carroll, S.L., Blair, B., Zhou, J., Zhang, C.L., and Fields, M.W. (2004) Alkaline anaerobic respiration. Isolation and characterization of a novel alkaliphilic and metal-reducing bacterium. *Appl Environ Microbiol* **70:** 5595–5602.

Supplementary material

The following supplementary material is available for this article online:

Table S1. Pseudogenes found in the *M. thermoacetica* genome.

Table S2. Parameters for BLAST alignments of some *M. thermoacetica* sequences mentioned in the text with GenBank sequences.

Table S3. Genes belonging to alcohol and aldehyde dehydrogenase COGs and pfams.

Table S4. Five *M. thermoacetica* gene clusters made up of COGs found in pyruvate:ferredoxin oxidoreductase.

Table S5. Dissimilatory sulfite reductase gene cluster.

Table S6. COGs with greater representation in the *M. thermoacetica* genome than in genomes of two closely related bacteria.

Table S7. Transporters encoded in the genome of *M. thermoacetica*.

Fig. S1. Circular representations of the genome of *M. thermoacetica* ATCC 39073. The outermost two circles indicate start sites of genes and assigned function (coloured by COG categories). Circle 1 consists of forward-strand gene products. Circle 2 consists of reverse-strand gene products. Colours represent the following functional categories: amino acid biosynthesis, cyan; biosynthesis of cofactors, brown; cell envelope, light gray; cellular processes, light blue; central intermediary metabolism, dark salmon; energy metabolism, green; fatty acid and phospholipid metabolism, orange; other categories, salmon; protein fate, dark gray; purines, pyrimidines, nucleosides, and nucleotides, light green; regulatory functions, light salmon; replication, blue; transcription and translation, magenta; transport and binding proteins, yellow; unassigned, black; unknown function, red; circle 3, RNA genes (tRNAs green, sRNAs red); circle 4, pseudogenes,

circle 5, IS elements; circle 6, G+C content; circle 7, GC skew [(G-C/G+C), khaki indicates values > 1, purple < 1].

Fig. S2. Pathway proposed for metabolism of xylose. Locus tags for genes encoding individual steps are shown.

Fig. S3. Pathway proposed for metabolism of glucose and fructose. Locus tags for genes encoding individual steps are shown.

Fig. S4. Pyruvate from sugar metabolism provides electrons and carbon for acetate synthesis by the Wood–Ljungdahl pathway.

Fig. S5. Tetrahydrofolate synthesis. *Moorella thermoacetica* genes which could be involved in synthesis of tetrahydrofolate from guanosine triphosphate and chorismate. The sequences of genes encoding enzymes which catalyse the transformations of 7,8-dihydroneopterin 3′-triphosphate to dihydroneopterin phosphate and dihydroneopterin phosphate to dihydroneopterin are not known.

Fig. S6. Synthesis of molybdenum cofactor and bis-molybdopterin guanine dinucleotide cofactor from guanine triphosphate. No *M. thermoacetica* genes were found with homology to the large subunit of molybdopterin synthase, MoaE.

Fig. S7. Components of the bacterial flagellar proteins in *M. thermoacetica*. The figure was derived from the KEGG web server (Kanehisa *et al.*, 2006; http://www.genome.jp/kegg/). The gene IDs that correspond to the labelled flagellar proteins are as follows: FlgB (Moth_0768), FlgC (Moth_0769), FlgD (Moth_0777), FlgE (Moth_0779), FlgG (Moth_0764), FlgK (Moth_0746), FlgL (Moth_0747), FlgL (Moth_0760), FlgM (Moth_0744), FlgN (Moth_0745), FlhA (Moth_0790), FlhB (Moth_0789), FliD (Moth_0765), FliE (Moth_0770), FliF (Moth_0771), FliG (Moth_0772), FliH (Moth_0773), FliI (Moth_0774), FliJ (Moth_0775), FliK (Moth_0776), FliM (Moth_0804), FliN (Moth_0784), FliO (Moth_0785), FliP (Moth_0786), FliQ (Moth_0787), FliR (Moth_0788), FliS (Moth_0766). Genes encoding FlgA, FlhC, FlhD, FliT proteins could not be identified in the *M. thermoacetica* genome. However, additional putative flagellar genes with unknown functions are encoded in the genome, such as Moth_0780 (FlbD), Moth_0783 (FliL), Moth_0791 (FlhF), Moth_0797 (FlgE-like), Moth_0798 (FlgE-like).

This material is available as part of the online article from http://www.blackwell-synergy.com

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.