Title

Life in hot carbon monoxide: The complete genome sequence of carboxydothermus hydrogenoformans Z-2901

Journal

PLoS Genetics, 1(5)

ISSN

1553-7390

Authors

Wu, M
Ren, Q
Scott Durkin, A
et al.

Publication Date

2005-12-01

Peer reviewed

Life in Hot Carbon Monoxide: The Complete Genome Sequence of *Carboxydothermus hydrogenoformans* Z-2901

Martin Wu[1], Qinghu Ren[1], A. Scott Durkin[1], Sean C. Daugherty[1], Lauren M. Brinkac[1], Robert J. Dodson[1], Ramana Madupu[1], Steven A. Sullivan[1], James F. Kolonay[1], William C. Nelson[1], Luke J. Tallon[1], Kristine M. Jones[1], Luke E. Ulrich[2], Juan M. Gonzalez[3], Igor B. Zhulin[2], Frank T. Robb[3], Jonathan A. Eisen[1,4,*]

1 The Institute for Genomic Research, Rockville, Maryland, United States of America, **2** Center for Bioinformatics and Computational Biology, School of Biology, Georgia Institute of Technology, Atlanta, Georgia, United States of America, **3** Center of Marine Biotechnology, University of Maryland Biotechnology Institute, Baltimore, Maryland, United States of America, **4** Johns Hopkins University, Baltimore, Maryland, United States of America

We report here the sequencing and analysis of the genome of the thermophilic bacterium *Carboxydothermus hydrogenoformans* Z-2901. This species is a model for studies of hydrogenogens, which are diverse bacteria and archaea that grow anaerobically utilizing carbon monoxide (CO) as their sole carbon source and water as an electron acceptor, producing carbon dioxide and hydrogen as waste products. Organisms that make use of CO do so through carbon monoxide dehydrogenase complexes. Remarkably, analysis of the genome of *C. hydrogenoformans* reveals the presence of at least five highly differentiated anaerobic carbon monoxide dehydrogenase complexes, which may in part explain how this species is able to grow so much more rapidly on CO than many other species. Analysis of the genome also has provided many general insights into the metabolism of this organism which should make it easier to use it as a source of biologically produced hydrogen gas. One surprising finding is the presence of many genes previously found only in sporulating species in the Firmicutes Phylum. Although this species is also a Firmicutes, it was not known to sporulate previously. Here we show that it does sporulate and because it is missing many of the genes involved in sporulation in other species, this organism may serve as a "minimal" model for sporulation studies. In addition, using phylogenetic profile analysis, we have identified many uncharacterized gene families found in all known sporulating Firmicutes, but not in any non-sporulating bacteria, including a sigma factor not known to be involved in sporulation previously.

Citation: Wu M, Ren Q, Durkin AS, Daugherty SC, Brinkac LM, et al. (2005) Life in hot carbon monoxide: The complete genome sequence of *Carboxydothermus hydrogenoformans* Z-2901. PLoS Genet 1(5): e65.

Introduction

Carbon monoxide (CO) is best known as a potent human poison, binding very strongly and almost irreversibly to the iron core of hemoglobin. Despite its deleterious effects on many species, it is also the basis for many food chains, especially in hydrothermal environments such as the deep sea, hot springs, and volcanoes. In these environments, CO is a common potential carbon source, as it is produced both by partial oxidation of organic matter as well as by multiple microbial strains (e.g., methanogens). It is most readily available in areas in which oxygen concentrations are low, since oxidation of CO will convert it to CO_2. In hydrothermal environments, CO use as a primary carbon source is dominated by the hydrogenogens, which are anaerobic, thermophilic bacteria or archaea that carry out CO oxidation using water as an electron acceptor [1]. This leads to the production of CO_2 and H_2. The H_2 is frequently lost to the environment and the CO_2 is used in carbon fixation pathways for the production of biomass. Hydrogenogens have attracted significant biotechnological interest because of the possibility they could be used in the biological production of hydrogen gas.

Hydrogenogens are found in diverse volcanic environments [2–7]. The phylogenetic types differ somewhat depending on the environments and include representatives of bacteria and archaea. *Carboxydothermus hydrogenoformans* is a hydrogenogen that was isolated from a hot spring in Kunashir Island, Russia [2]. It is a member of the Firmicutes Phylum (also known as low GC Gram-positives) and grows optimally at 78 °C. This species has been considered an unusual hydrogenogen, in part because unlike most of the other hydrogenogens, it was believed to be strictly dependent on CO for growth. The other species were found to grow poorly unless CO was supplemented with organic substrates. Thus it was selected for genome sequencing as a potential model obligate CO autotroph.

Surprisingly, initial analysis of the unpublished genome

Received August 8, 2005; Accepted October 19, 2005; Published November 25, 2005
DOI: 10.1371/journal.pgen.0010065

Abbreviations: CDS, protein coding sequence; CO, carbon monoxide; CODH, carbon monoxide dehydrogenase

Editor: Paul M Richardson, Joint Genome Institute, United States of America

* To whom correspondence should be addressed. E-mail: jeisen@tigr.org

A previous version of this article appeared as an Early Online Release on October 19, 2005 (DOI: 10.1371/journal.pgen.0010065.eor).

Synopsis

Carboxydothermus hydrogenoformans, a bacterium isolated from a Russian hotspring, is studied for three major reasons: it grows at very high temperature, it lives almost entirely on a diet of carbon monoxide (CO), and it converts water to hydrogen gas as part of its metabolism. Understanding this organism's unique biology gets a boost from the decoding of its genome, reported in this issue of *PLoS Genetics*. For example, genome analysis reveals that it encodes five different forms of the protein machine carbon monoxide dehydrogenase (CODH). Most species have no CODH and even species that utilize CO usually have only one or two. The five CODH in *C. hydrogenoformans* likely allow it to both use CO for diverse cellular processes and out-compete for it when it is limiting. The genome sequence also led the researchers to experimentally document new aspects of this species' biology including the ability to form spores. The researchers then used comparative genomic analysis to identify conserved genes found in all spore-forming species, including *Bacillus anthracis,* and not in any other species. Finally, the genome sequence and analysis reported here will aid in those trying to develop this and other species into systems to biologically produce hydrogen gas from water.

sequence data led to the discovery that this species is not an obligate CO autotroph [8]. We report here a detailed analysis of the genome sequence of *C. hydrogenoformans* strain Z-2901, the type strain of the species, hereafter referred to simply as *C. hydrogenoformans*.

Results/Discussion

Genome Structure

The *C. hydrogenoformans* genome is a single circular chromosome of 2,401,892 base pairs (bp) with a G+C content of 42.0% (Figure 1, Table 1). Annotation of the genome reveals 2,646 putative protein coding genes (CDSs), of which 1,512 can be assigned a putative function. The chromosome displays two clear GC skew transitions that likely correspond to the DNA replication origin and terminus (Figure 1). Overall, 3.0 % of the genome is made up of repetitive DNA sequences. Included in this repetitive DNA are two large-clustered, regularly interspaced short palindromic repeats (CRISPR, 3.9 and 5.6 kilobases, respectively). Each cluster contains 59 and 84 partially palindromic repeats of 30 bps, respectively (GTTTCAATCCCAGA[A/T]TGGTTCGATTAAAAC). Most repeats within each cluster are identical but they differ for one nucleotide in the middle between clusters. Repeats at ends of the smaller cluster degenerate to some extent. These types of repeats are widespread in diverse groups of bacteria and archaea [9]. The first one-third of the repeat sequence is generally conserved. Although the precise functions of these repeats are unknown, some evidence suggests they are involved in chromosome partitioning [10,11]. In addition, experiments in the thermophilic archaea *Sulfolobus solfataricus* have identified a genus-specific protein binding specifically to the repeats present in that species' genome [11].

One 35-kilobase lambda-like prophage containing 50 CDSs was identified in the genome. It is flanked on one side by a tRNA suggesting this may have served as a site of insertion. Phylogenetic analysis showed this phage is most closely related to phages found in other Firmicutes, particularly the SPP1 phage infecting *Bacillus subtilis*.

As with other members of the Phylum Firmicutes, the directions of leading strand DNA replication and transcription are highly correlated, with 87% of genes located on the leading strand. This gene distribution bias is also highly correlated with the presence of a Firmicutes-specific DNA polymerase PolC in the genome [12]. In *B. subtilis,* PolC synthesizes the leading strand, and another distinct DNA polymerase, DnaE, replicates the lagging strand [13]. In other non-Firmicutes bacteria, DnaE replicates both strands. The asymmetric replication forks of Firmicutes were proposed to contribute to the asymmetry of their gene distributions [12]. One copy of PolC and two copies of DnaE have been identified in *C. hydrogenoformans* genome. At least some of the gene distribution bias can be caused by selection to avoid collision of the RNA and DNA polymerases as well [14,15]. Despite this apparent selection, the lack of significantly conserved gene order across Firmicutes indicates that genome rearrangements still occur at a reasonably high rate.

Phylogeny and Taxonomy

Analysis of the complete genome of *C. hydrogenoformans* suggests that the taxonomy of this species, as well as some other organisms, needs to be revised. More specifically, phylogenetic analysis based on concatenation of a few dozen markers (Figure 2) reveals a variety of conflicts between the organismal phylogeny and the classification of some of the Firmicutes. For example, *C. hydrogenoformans* is currently considered to be a member of the Family Peptococcaceae in the Order Clostridiales [16]. Thus it should form a clade with the *Clostridium* spp. to the exclusion of other taxa for which genomes are available (e.g., *Thermoanaerobacter tengcongensis,* which is considered to be a member of Thermoanaerobacteriales). The tree, however, indicates that this is not the case and that *T. tengcongensis* and the *Clostridia* spp. are more closely related to each other than either is to *C. hydrogenoformans*. Thus we believe *C. hydrogenoformans* should be placed in a separate Order from Clostridiales.

Perhaps more surprisingly, the concatenated genome tree shows *C. hydrogenoformans* grouping with *Symbiobacterium thermophilum*. *S. thermophilum* is a strictly symbiotic thermophile isolated from compost and is currently classified in the Actinobacteria (also known as high GC Gram-positives) based on analysis of its 16s rRNA sequence [17]. The grouping with Firmicutes is supported by the overall level of similarity of its proteome to other species [18]. We therefore believe the rRNA-based classification is incorrect and that *S. thermophilum* should be transferred to the Firmicutes. Such inaccuracies of the rRNA trees are relatively uncommon and may in this case be due to the mixing of thermophilic and non-thermophilic species into one group. This can cause artifacts when using rRNA genes for phylogenetic reconstruction since the G+C content of rDNA is strongly correlated to optimal growth temperature.

CO Dehydrogenases and Life in CO

Anaerobic species that make use of CO do so using nickel-iron CO dehydrogenase (CODH) complexes [19,20]. These enzymes all appear to catalyze the anaerobic interconversion of CO and CO_2. However, they vary greatly in the cellular role of this conversion and in the exact structure of the complex [19]. Analysis of the genome reveals the presence of five genes encoding homologs of CooS, the catalytic subunit of

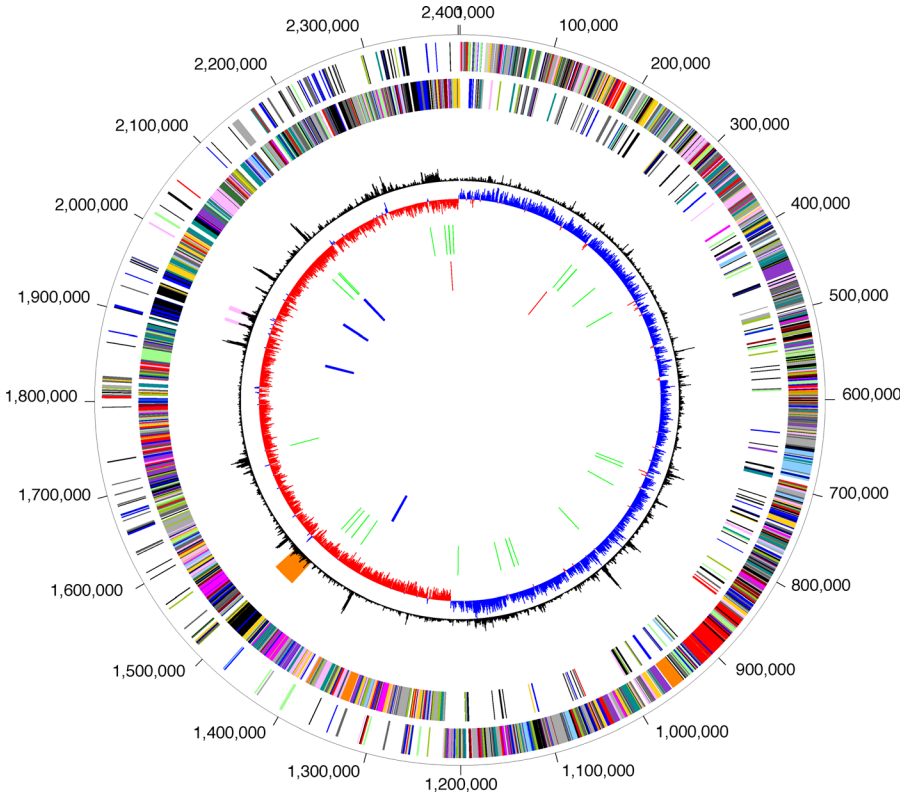

Figure 1. Genomic Organization of *C. hydrogenoformans*

From the outside inward the circles show: (1, 2) predicted protein-coding regions on the plus and minus strands (colors were assigned according to the color code of functional classes; (3) prophage (orange) and CRISPR (pink) regions; (4) χ^2-square score of tri-nucleotide composition; (5) GC skew (blue indicates a positive value and red a negative value); (6) tRNAs (green); (7) rRNAs (blue) and structural RNAs (red).

DOI: 10.1371/journal.pgen.0010065.g001

anaerobic CODHs. These five CooS encoding genes are scattered around the genome, and analysis of genome context, gene phylogeny, and experimental studies in this and other CO-utilizing species suggests they are subunits of five distinct CODH complexes, which we refer to as CODH I-V (Figure 3). The CooS homologs are named accordingly.

Table 1. General Features of the *C. hydrogenoformans* Genome

Feature	Value
Genome size, bp	2,401,892
% G+C	42.0
Predicted protein coding genes (CDSs)	2646
Average CDS length	827
Percent of genome that is coding	91.1
CDSs with assigned function	1512 (57.1%)
Conserved hypothetical CDS[a]	354 (13.4%)
Unknown function CDS[b]	331 (12.5%)
Hypothetical CDS[c]	449 (17.0%)
Transfer RNA	50
Ribosomal RNA	12
Structural RNAs	2
CRISPR regions	2
Prophage	1

[a]Match to genes in other species, but no function known.
[b]Some biochemical function prediction, but cellular role not predictable.
[c]No match to genes in other species.
DOI: 10.1371/journal.pgen.0010065.t001

Specific details about each complex and proposed physiological roles are given in the following paragraphs.

Energy Conservation (CODH-I)

A catalytic subunit (CooS-I, CHY1824) and an electron transfer protein (CooF, CHY1825) of CODH are encoded immediately downstream of a hydrogenase gene cluster (*cooMKLXUH,* CHY1832–27) that is closely related to the one found in *Rhodospirillum rubrum* [21]. These eight proteins form a tight membrane-bound enzyme complex that converts CO to CO_2 and H_2 in vitro [1,22]. In *R. rubrum,* this CODH/hydrogenase complex was proposed to be the site of CO-driven proton respiration where energy is conserved in the form of a proton gradient generated across the cell membrane [21]. Based on the high similarities in protein sequences and their gene organization, this set of genes were suggested to play a similar role in energy conservation in *C. hydrogenoformans* [1]. Consistent with this, this *cooS* gene is in the same subfamily as that from *R. rubrum* (Figure 4).

Carbon Fixation (CODH-III)

Anaerobic bacteria and archaea, such as methanogens and acetogens, can fix CO or CO_2 using the acetyl-CoA pathway (also termed the Wood-Ljungdahl pathway), where two molecules of CO_2, through a few steps, are condensed into one acetyl-CoA, a key building block for cellular biosynthesis and an important source of ATP [23]. The key enzyme of the final step (a CODH/acetyl-CoA synthase complex) has been purified from *C. hydrogenoformans* (strain DSM 6008) cultured

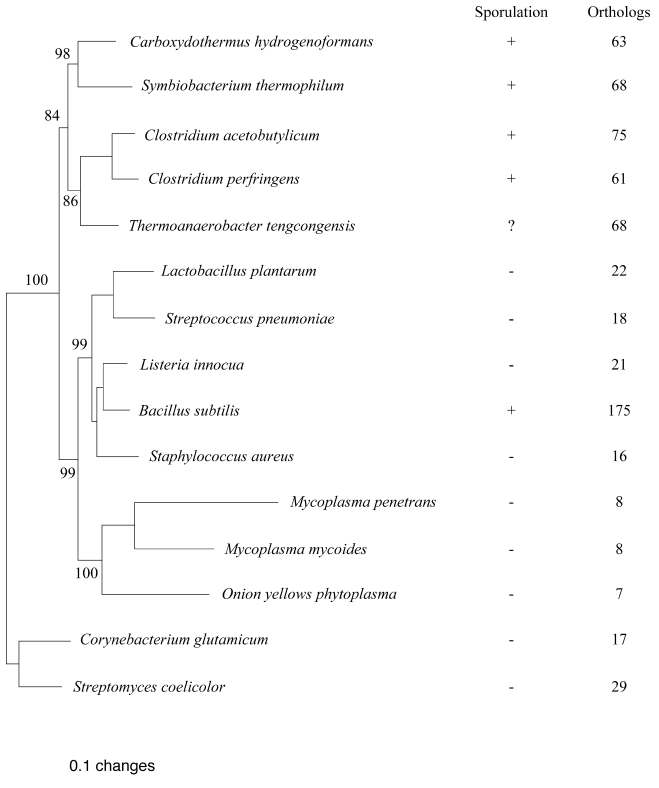

Figure 2. Genome Tree of Representatives of Firmicutes

A maximum likelihood tree was built from concatenated protein sequences of 31 universal housekeeping genes and rooted by two outgroup Actinobacteria (high GC Gram-positives) species: *Corynebacterium glutamicum* and *Streptomyces coelicolor*. Bootstrap support values (out of 100 runs) for branches of interest are shown beside them. Each species' ability to sporulate and its number of putative orthologs of the 175 known *B. subtilis* sporulation genes are also shown.

DOI: 10.1371/journal.pgen.0010065.g002

under limited CO supply and shown to be functional in vitro [24]. Genes encoding this complex and other proteins predicted to be in this pathway are clustered in the genome (CHY1221–7). This cluster is very similar to the *acs* operon from the acetogen *Moorella thermoacetica* which encodes the acetyl-CoA pathway machinery [25]. The phylogenetic tree also shows that CooS-III is in the same subfamily as the corresponding gene in the *M. thermoacetica acs* operon (Figure 4), suggesting they have the same biological functions. In addition, all the genes in the acetyl-CoA pathway have been identified in the *C. hydrogenoformans* genome and activities of some of those gene products have been detected (Figure 5), prompting us to propose that this organism carries out autotrophic fixation of CO through this pathway. This is consistent with the observation that key enzymes for the other known CO_2 fixation pathways, such as the Calvin cycle, the reverse tricarboxylic acid cycle, and 3-hydroxypropionate cycle are apparently not encoded in the genome.

Oxidative Stress Response (CODH-IV)

C. hydrogenoformans, though an anaerobe, has to deal with oxidative challenges present in the environment from time to time. Unlike aerobes, many anaerobes are proposed to use an alternative oxidative stress protection mechanism that depends on proteins such as rubrerythrin [26,27]. With few exceptions, rubrerythrin-like proteins have been found in complete genomes of all anaerobic and microaerophilic microbes but are absent in aerobic microbes [28]. Rubrerythrin is thought to play a role in the detoxification of reactive oxygen species by reducing the intermediate hydrogen peroxide, although the exact details remain elusive [28,29]. *C. hydrogenoformans* encodes three rubrerythrin homologs. One of them forms an operon with genes encoding CooS-IV, a CooF homolog, and a NAD/FAD-dependent oxidoreductase (CHY0735–8, Figure 3), suggesting that their functions are related. Here we speculate that this operon encodes a multisubunit complex where electrons stripped from CO by the CODH are passed to rubrerythrin to reduce hydrogen peroxide to water, with CooF and the NAD/FAD-dependent oxidoreductase acting as the intermediate electron carriers. Therefore, CODH-IV may play an important role in oxidative stress response by providing the ultimate source of reductants.

Others

Two other homologs of CooS are encoded in the genome. The gene encoding CooS-II (CHY0085) was originally cloned with the neighboring *cooF* (CHY0086) [30] and the complex was purified as functional homodimers [1]. This complex (CODH-II) is membrane-associated and an in vitro study showed it might have an anabolic function of generating NADPH [1]. Its structure has been solved [31]. The role of CooS-V (CHY0034) is more intriguing as it is the most deeply branched of the CooSs (Figure 4) and is not flanked by any genes with obvious roles in CO-related processes.

Aerobic bacteria metabolize CO using drastically different CODHs that are unrelated to the anaerobic ones. The CODHs from aerobes are dimers of heterotrimers composed of a molybdoprotein (CoxL), a flavoprotein (CoxM), and an iron-sulfur protein (CoxS) and belong to a large family of molybdenum hydroxylases including aldehyde oxidoreductases and xanthine dehydrogenases [32]. These enzymes characteristically demonstrate high affinity for CO, and the oxidation is typically coupled to CO_2 fixation via the reductive pentose phosphate cycle.

C. hydrogenoformans has one gene cluster (CHY0690–2) homologous to the *coxMSL* cluster in *Oligotropha carboxidovorans,* the most well-studied aerobic CODHs. However, our phylogenetic analysis showed that the *C. hydrogenoformans* homolog of CoxL does not group within the CODH subfamily. Therefore, we conclude that it is unlikely that this gene cluster in *C. hydrogenoformans* encodes a CODH, although that needs to be tested. Of the available published and unpublished genomes, only *R. rubrum* appears to have both an anaerobic CODH and a close relative of the aerobic *O. carboxidovorans* CODH. Accordingly, *R. rubrum,* a photosynthetic bacterium, can grow in the dark both aerobically and anaerobically using CO as an energy source.

Structures of both the Mo- and Ni-containing enzymes have been published recently. The crystal structure of CooS-II from *C. hydrogenoformans* is a dimeric enzyme with dual Ni-containing reaction centers each connected to the enzyme surface by 70-Å hydrophobic channels through which CO transits [31]. This channeling, also confirmed experimentally [33,34], explains the mechanism of CO use as a central metabolic intermediate despite its low solubility and generally low concentration in geothermal environments.

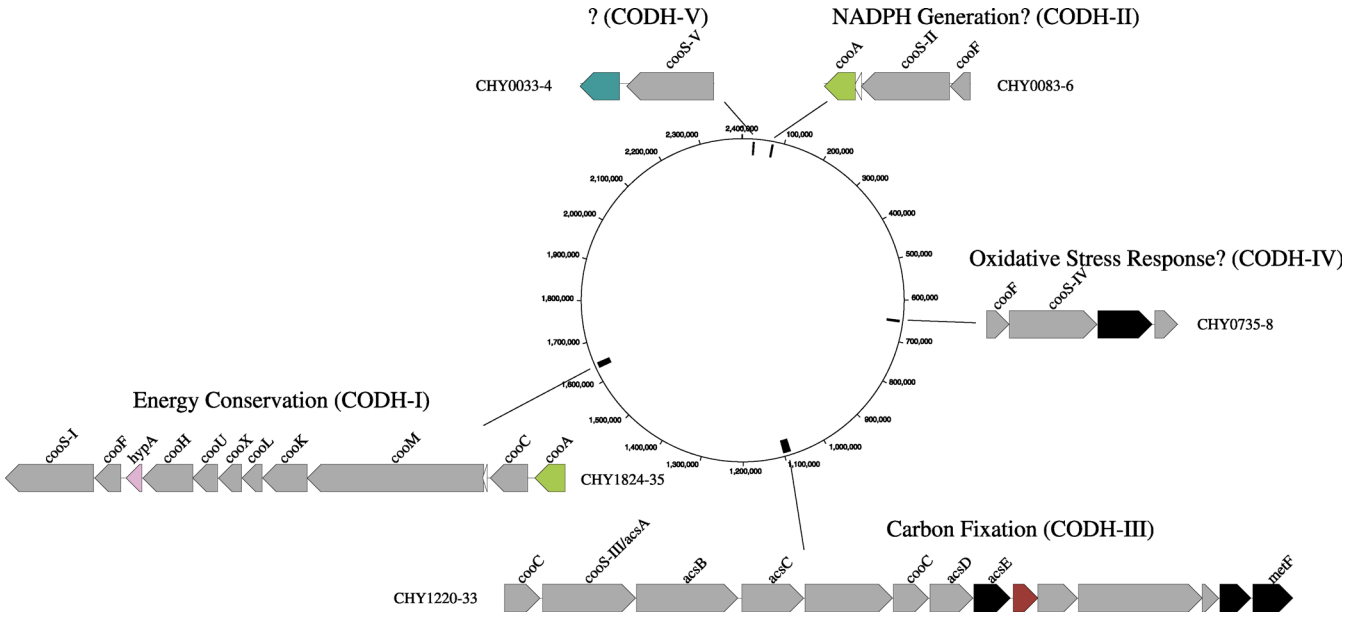

Figure 3. Genome Locations of Genes Predicted to Encode Five CODH Complexes

The genome locations of the genes encoding the five CooS homologs (labelled CooS I-V) are shown. Also shown are neighboring genes that are predicted to encode the five distinct CODH complexes (CODH I-V) with each CooS homolog. Possible cellular roles for four of the five CODH complexes are indicated.

DOI: 10.1371/journal.pgen.0010065.g003

Sporulation

The *C. hydrogenoformans* genome encodes a large number of homologs of genes involved in sporulation in other Firmicutes, spanning all stages of sporulation (Table 2). Among those are the master switch gene *spo0A* and all sporulation-specific sigma factors, σ^H, σ^E, σ^F, σ^G, and σ^K. However, sporulation has not been previously reported for this species. With this in mind, we set out to re-examine the morphology of *C. hydrogenoformans* cells and found endospore-like structures when cultures were stressed (Figure 6).

We then used phylogenetic profile analysis to look for other possible sporulation genes in the genome. Phylogenetic profiling works by grouping genes according to their distribution patterns in different species [35]. Proteins that function in the same pathways or structural complexes frequently have correlated distribution patterns. Phylogenetic profile analysis identified an additional set of 37 potential sporulation-related genes (Figure 7). Those genes are generally *Bacillales*- and *Clostridiales*-specific, consistent with the fact that endospores have so far only been found in these and other closely related Firmicutes. Most of the novel genes are conserved hypothetical proteins, whereas a few are putative membrane proteins. In support, a few of those novel sporulation genes have been shown to be involved in *Bacillus subtilis* sporulation by experimental studies [36,37]. The rest of the genes are thus excellent candidates for encoding known sporulation functions that have not been assigned to genes or previously unknown sporulation activities. Strikingly, within this group of genes, in addition to other known sporulation-specific sigma factors (σ^E, σ^F, σ^G, and σ^K), we identified a sigma factor (CHY1519) that was not known to be associated with sporulation previously. σ^I, its putative ortholog in *B. subtilis,* has shown some association with heat shock [38]. It remains to be determined experimentally whether this sigma factor is involved in sporulation, and if so, the regulatory network it controls.

A search of known sporulation-related genes in *B. subtilis* against *C. hydrogenoformans* revealed that many of them are missing in the genome. Of the 175 *B. subtilis* sporulation-related genes we compiled from the genome annotation and literature [39,40], half have no detectable homologs in *C. hydrogenoformans* using BLASTP with an E-value cutoff of 1e-5. Putative orthologs defined by mutual-best-hit methodology are present for only one third of those genes in *C. hydrogenoformans.* Among those missing genes are *spo0B* and *spo0F,* which encode the key components of the complex phosphorelay pathway in *B. subtilis* that channels various signals such as DNA damage, the ATP level, and cell density to the master switch protein Spo0A and therefore governs the cell's decision to enter sporulation. *C. hydrogenoformans* hence uses either a simplified version of this pathway or an alternative signal transduction pathway to sense the environmental or physiological stimuli. A large number of genes involved in the protective outer layer (cortex, coat, and exosporium) formation, spore germination, and small acid-soluble spore protein synthesis, among a few genes in various stages of spore development, are also missing. A similar, but slightly different, set of genes are missing in the other spore-forming *Clostridia* species as well [41]. Absence of those genes is more pronounced in non-spore-forming Firmicutes such as *Listeria* spp., *Staphylococcus* spp., and *Streptococcus* spp., as they lack all the sporulation-specific genes. When overlaid onto the phylogeny of Firmicutes (Figure 2), this observation can be explained by either multiple independent gene-loss events along branches leading to non-*Bacillus* species or by independent gene-gain events along branches leading to *Bacillus* and *Clostridia,* or by both. Whatever the history is of the sporulation evolution, the core set of sporulation genes

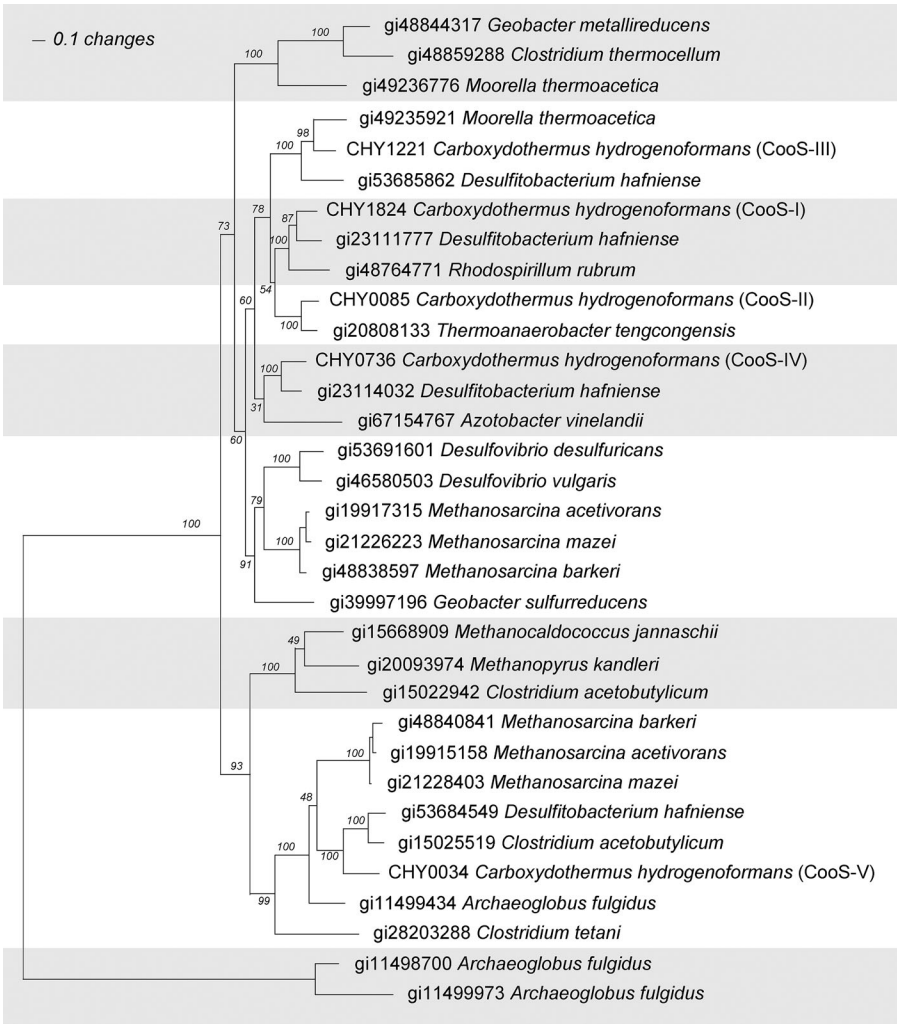

Figure 4. Phylogenetic Tree of CooS Homologs

The figure shows a maximum-likelihood tree of CooS homologs. The tree indicates the five CooS homologs in *C. hydrogenoformans* are not the result of recent duplications but instead are from distinct subfamilies. The other CooS homologs included in the tree were obtained from the NCBI nr database and include some from incomplete genome sequences generated by United States Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/).

DOI: 10.1371/journal.pgen.0010065.g004

shared by *Bacillus* and *Clostridia* might be close to a "minimal" sporulation set, as so far only these two groups have been found to be capable of producing endospores. Alternatively, some spore specific functions may be carried out by non orthologous genes in different species, which would prevent us from identifying them by this type of analysis.

Strictly Dependent on CO?

Until very recently, *C. hydrogenoformans* was thought to be an autotroph strictly depending on CO for growth. An overview of the genome reveals features related to its autotrophic lifestyle. For example, it has lost the entire sugar phosphotransferase system and encodes no complete pathway for sugar compound degradation. However, many aspects of the gene repertoire are suggestive of heterotrophic capabilities. For example, among the transporters encoded in the genome are ones predicted to import diverse carbon compounds including formate, glycerol, lactate, C4-dicarboxylate (malate, fumarate, or succinate; the binding receptor for this has three paralogs in the genome), 2-keto-3-deoxygluconate, 2-oxoglu-

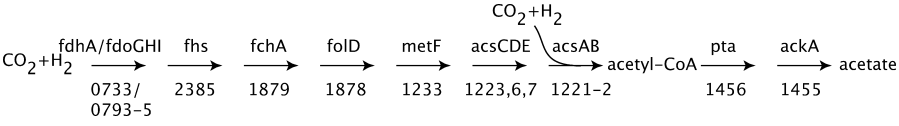

Figure 5. Predicted Complete Acetyl-CoA Pathway of Carbon Fixation in *C. hydrogenoformans*

Genes predicted to encode each step in the acetyl-CoA pathway of carbon fixation were identified in the genome. The locus numbers are indicated on the figure.

DOI: 10.1371/journal.pgen.0010065.g005

Table 2. Orthologs of Known *Bacillus subtilis* Sporulation Genes in *C. hydrogenoformans*

Locus	Gene	Description
CHY1978	*spo0A*	Stage 0 sporulation protein A
CHY0010	*spo0J*	Stage 0 sporulation protein J
CHY0370	*obg*	spo0B-associated GTP-binding protein
CHY0009	*soj*	Sporulation initiation inhibitor protein soj
CHY1960	*spoIIAB*	Anti-sigma F factor
CHY2541	*spoIID*	Stage II sporulation protein D
CHY1517	*spoIID*	Putative stage II sporulation protein D
CHY0212	*spoIIE*	Putative stage II sporulation protein E
CHY2057	*spoIIGA*	Putative sporulation specific protein SpoIIGA
CHY1965	*spoIIM*	Putative stage II sporulation protein M
CHY1923	*spoIIP*	Putative stage II sporulation protein P
CHY0408	*spoIIP*	Putative sporulation protein
CHY2054	*spoIIR*	Stage II sporulation protein R
CHY0206		Putative stage II sporulation protein D
CHY2007	*spoIIIAA*	Putative sporulation protein
CHY2006	*spoIIIAB*	Putative sporulation protein
CHY2005	*spoIIIAC*	Putative sporulation protein
CHY2004	*spoIIIAD*	Putative sporulation protein
CHY2003	*spoIIIAE*	Putative sporulation protein
CHY2001	*spoIIIAG*	Putative sporulation protein
CHY2534	*spoIIID*	Stage III sporulation protein D
CHY1159	*spoIIIE*	DNA translocase FtsK
CHY0004	*spoIIIJ*	Sporulation associated-membrane protein
CHY1916	*spoIVA*	Stage IV sporulation protein A
CHY1979	*spoIVB*	Putative stage IV sporulation protein B
CHY1957	*spoVAC*	Stage V sporulation protein AC
CHY1956	*spoVAD*	Stage V sporulation protein AD
CHY1955	*spoVAE*	Stage V sporulation protein AE
CHY0960	*spoVB*	Stage V sporulation protein B
CHY1152	*spoVFA*	Dipicolinate synthase, A subunit
CHY1153	*spoVFB*	Dipicolinate synthase, B subunit
CHY1391	*spoVK*	Stage V sporulation protein K
CHY1202	*spoVR*	Stage V sporulation protein R
CHY1171	*spoVS*	Stage V sporulation protein S
CHY0202	*spoVT*	Stage V sporulation protein T
CHY2272	*cotJC*	cotJC protein
CHY0786	*cotJC*	cotJC protein
CHY1463	*sspD*	Small acid-soluble spore protein
CHY1464	*sspD*	Small acid-soluble spore protein
CHY1175	*sspF*	Small acid-soluble spore protein
CHY1465		Putative small acid-soluble spore protein
CHY1941	*spmA*	Spore maturation protein A
CHY1940	*spmB*	Spore maturation protein B
CHY0958		Small acid-soluble spore protein
CHY1160		Putative spore cortex-lytic enzyme
CHY1756	*sleB*	Putative spore cortex-lytic enzyme
CHY0336	*gerKA*	Spore germination protein GerKA
CHY1404	*gerKB*	Spore germination protein
CHY0337	*gerKC*	Spore germination protein
CHY0305	*gerM*	Putative germination protein GerM
CHY1950		Putative spore germination protein
CHY0143		RNA polymerase sigma factor
CHY2056	*sigE*	RNA polymerase sigma-E factor
CHY1959	*sigF*	RNA polymerase sigma-F factor
CHY2055	*sigG*	RNA polymerase sigma-G factor
CHY2333	*sigH*	RNA polymerase sigma-H factor
CHY0617	*sigK*	RNA polymerase sigma-K factor
CHY1462	*gpr*	Spore protease
CHY2672		Sigma-K processing regulatory protein BofA
CHY0424		Putative sporulation protein

DOI: 10.1371/journal.pgen.0010065.t002

tarate, and amino acids. In addition there is a diverse array of signal transduction pathways including chemotaxis not commonly found in the genomes of autotrophs (see below). Consistent with these observations, Henstra et al. recently

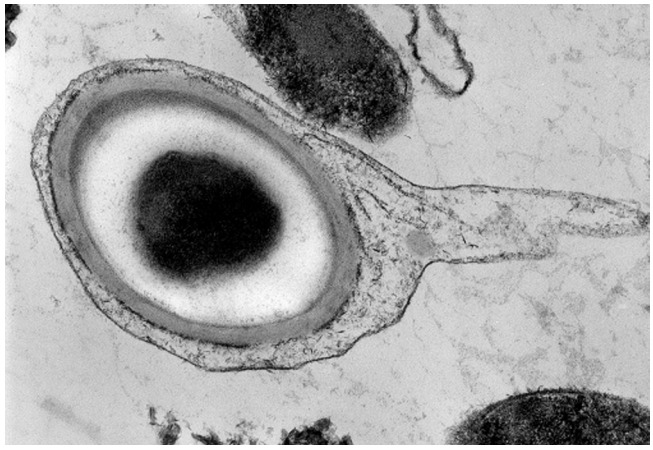

Figure 6. An Electron Micrograph of a *C. hydrogenoformans* Endospore

The finding of homologs of many genes involved in sporulation in other species led us to test whether *C. hydrogenoformans* also could form an endospore. Under stressful growth conditions, endospore-like structures form. We note that even though homologs could not be found in the genome for many genes that in other species are involved in protective outer-layer (cortex, coat, and exosporium) formation, those structures seem to be visible and intact.

DOI: 10.1371/journal.pgen.0010065.g006

showed that formate, lactate, and glycerol could be utilized as carbon source provided 9,10-anthraquinone-2,6-disulfonate was used as the electron acceptor [8]. Similarly, sulfite, thiosulfate, sulfur, nitrate, and fumarate were reduced with lactate as electron donor, although heterotrophic growth was relatively slow compared with cultures growing on pure CO [8]. It is not known what electron acceptors are likely to be coupled to these pathways in the isolation locale of *C. hydrogenoformans,* however it is clear that there is a more versatile complement of energy sources than initially concluded by Svetlichny et al. [2].

In terms of autotrophic lifestyle, although *C. hydrogenoformans* and *S. thermophilum* are close phylogenetically, they have gone separate ways in their lifestyles. *S. thermophilum* is an uncultivable thermophilic bacterium growing as part of a microbial consortium [18], while *C. hydrogenoformans* is a hot-spring autotroph that can survive efficiently on CO as its sole carbon and energy source. Accordingly, their metabolic capabilities are very different and only half of their proteomes are homologous. It is not clear why *S. thermophilum* is dependent on other microbes. Unlike other symbiotic microorganisms, no large-scale genome reductions have occurred in *S. thermophilum* [18]. On the other hand, *C. hydrogenoformans* has evolved to live preferably on CO, possibly by acquiring and/or expanding its complement of CODHs. As a result, it has lost many genes associated with a heterotrophic lifestyle, such as the phosphotransferase transporter system, and may be on the verge of becoming an obligate autotroph. Even though *C. hydrogenoformans* is more closely related to *S. thermophilum* than to *T. tengcongensis,* an anaerobic thermophile isolated also from freshwater hot springs [42], *C. hydrogenoformans* actually shares slightly less genes with *S. thermophilum* than with *T. tengcongensis.*

Signal Transduction

C. hydrogenoformans is poised to respond to diverse environmental cues through a suite of signal transduction pathways

CHY1367 C4-dicarboxylate response regulator
CHY1529 degV family protein
CHY2346 putative DNA-binding protein
CHY1959 sigF RNA polymerase sigma-F factor
CHY2034 conserved hypothetical protein
CHY1391 spoVK stage V sporulation protein K
CHY0786 cotJC cotJC protein
CHY2600 SCP-like extracellular protein
CHY2481 putative phosphoesterase
CHY1978 spo0A stage 0 sporulation protein A
CHY2617 conserved hypothetical protein
CHY1943 conserved hypothetical protein
CHY2007 putative sporulation protein
CHY2055 sigG RNA polymerase sigma-G factor
CHY1913 pheB ACT domain protein pheB
CHY2057 putative sporulation specific protein SpoIIGA
CHY2611 YabG peptidase,U57 family
CHY0171 putative membrane protein
CHY0408 putative sporulation protein
CHY1940 spore maturation protein B
CHY2541 stage II sporulation protein D
CHY0212 putative stage II sporulation protein E
CHY1457 putative membrane protein
CHY1965 putative stage II sporulation protein M
CHY2006 putative sporulation protein
CHY2003 putative sporulation protein
CHY2001 putative sporulation protein
CHY1462 gpr spore protease
CHY1560 conserved hypothetical protein
CHY1589 conserved hypothetical protein
CHY1648 putative membrane protein
CHY1916 stage IV sporulation protein A
CHY1955 stage V sporulation protein AE
CHY1956 stage V sporulation protein AD
CHY1979 putative stage IV sporulation protein B
CHY1957 stage V sporulation protein AC
CHY2054 stage II sporulation protein R
CHY2056 sigE RNA polymerase sigma-E factor
CHY0020 conserved hypothetical protein
CHY0336 gerKA spore germination protein GerKA
CHY0337 spore germination protein
CHY0424 putative sporulation protein
CHY0202 spoVT stage V sporulation protein T
CHY2622 transcriptional regulator, AbrB family
CHY2004 putative sporulation protein
CHY1463 small acid-soluble spore protein
CHY1464 small acid-soluble spore protein
CHY2053 PRC-barrel domain protein
CHY2005 putative sporulation protein
CHY2534 stage III sporulation protein D
CHY0617 sigK RNA polymerase sigma-K factor
CHY1487 rpoZ DNA-directed RNA polymerase, omega subunit
CHY0423 conserved hypothetical protein
CHY0329 putative ATP-dependent protease La
CHY1171 spoVS stage V sporulation protein S
CHY1593 putative lipoprotein
CHY1519 RNA polymerase sigma factor
CHY0207 conserved hypothetical protein
CHY1161 conserved hypothetical protein
CHY0305 putative germination protein GerM
CHY0021 putative membrane protein
CHY1390 conserved hypothetical protein
CHY0038 putative membrane protein
CHY1043 putative glycosyl transferase
CHY2349 ATP:guanido phosphotransferase domain protein
CHY2350 uvrB/uvrC motif domain protein
CHY1843 glpP glycerol uptake operon antiterminator regulatory protein
CHY1960 anti-sigma F factor
CHY0441 CBS domain protein
CHY1452 conserved hypothetical protein
CHY0278 putative membrane protein
CHY0651 transcription regulator, Fur family
CHY1082 conserved hypothetical protein
CHY2676 conserved hypothetical protein
CHY1489 conserved hypothetical protein
CHY1155 aspartate kinase, monofunctional class
CHY2271 N-acetylmuramoyl-L-alanine amidase
CHY0544 vanW domain protein
CHY1941 spore maturation protein A

Figure 7. Phylogenetic Profile Analysis of Sporulation in *C. hydrogenoformans*

For each protein encoded by the *C. hydrogenoformans* genome, a profile was created of the presence or absence of orthologs of that protein in the predicted proteomes of all other complete genome sequences. Proteins were then clustered by the similarity of their profiles, thus allowing the grouping of proteins by their distribution patterns across species. Examination of the groupings showed one cluster consisting of mostly homologs of sporulation proteins. This cluster is shown with *C. hydrogenoformans* proteins in rows (and the prediced function and protein ID indicated on the right) and other species in columns with presence of a ortholog indicated in red and absence in black. The tree to the left represents the portion of the cluster diagram for these proteins. Note that most of these proteins are found only in a few species represented in red columns near the center of the diagram. The species corresponding to these columns are indicated. We also note that though most of the proteins in this cluster, for which functions can be predicted, are predicted to be involved in sporulation and some have no predictable functions (highlighted in blue). This indicates that functions of these proteins' homologs have not been characterized in other species. Since these proteins show similar distribution patterns to so many proteins with roles in sporulation, we predict that they represent novel sporulation functions.

DOI: 10.1371/journal.pgen.0010065.g007

and processes. The organism has 83 one-component regulators and 13 two-component systems (including two chemotaxis systems), which are average numbers for such a genome size [43] (Table S1). Many of the genes encoding these two-component systems are next to transporters, possibly being involved in regulation of solute uptake, while others are adjacent to oxidoreductases. *C. hydrogenoformans* also possesses an elaborate cascade of chemotaxis genes, including 11 chemoreceptors, and a complete set of flagellar genes, most located within a large cluster of about 70 genes (CHY0963–1033). Chemotaxis allows microbes to respond to environmental stimuli by swimming toward nutrients or away from toxic chemicals. Generally, a heavy commitment to chemotaxis is not a characteristic of autotrophic microorganisms [44], and it is possible that *C. hydrogenoformans* is responding to gradients of inorganic nutrients, or gases such as CO, O_2, H_2, or CO_2.

Critical for sensing CO, two CooA homologs occur in the *C. hydrogenoformans* genome, both of which are encoded within operons containing *cooS* genes. CooA proteins are heme proteins that act as both sensors for CO as well as transcriptional regulators. They belong to the cyclic adenosine monophosphate receptor protein family and induce CO-related genes upon CO binding [45]. CHY1835, encoding CooA1, is at the beginning of the *R. rubrum*-like *coo* operon. CHY0083, encoding CooA2, is at the end of the operon possibly involved in NADPH generation from CO [1] (Figure 3).

C. hydrogenoformans lacks certain subfamilies of transcription factors that are present in its close *Clostridia* relatives, such as those utilizing the following helix-turn-helix domains: iron-dependent repressor DNA-binding domain, LacI, PadR, and DeoR (Pfam nomenclature). The genome does not encode any proteins of the LuxR family, which are usually abundant in both one-component (e.g., quorum-sensing regulators) and two-component systems.

The largest family of transcriptional regulators in *C. hydrogenoformans* is sigma-54- dependent activators. Eight such regulators comprise one-component systems (CHY0581, CHY0788, CHY1254, CHY1318, CHY1359, CHY1376, CHY1547, and CHY2091) and another one is a response regulator of the two-component system (CHY1855). Seven one-component sigma-54-dependent regulators have at least one PAS domain as a sensory module. PAS domains are known to often contain redox-responsive cofactors, such as FAD, FMN, and heme and serve as intracellular oxygen and redox sensors [46]. Overall, there are 18 PAS domains in *C. hydrogenoformans*. It is a very significant number compared to only two PAS domains in *Moorella thermoacetica* (similar genome size) and nine in *Desulfitobacterium hafniense* (a much larger genome). The most abundant sensory domain of bacterial signal transduction, the LysR substrate-binding domain, which binds small molecule ligands, is present only in six copies in *C. hydrogenoformans* (there are 36 copies in *D. hafniense*), re-enforcing the notion that redox sensing via PAS domains might be the most critical signal transduction event for this organism.

The most intriguing signal transduction protein in *C. hydrogenoformans* is the sigma-54-dependent transcriptional regulator that has an iron hydrogenase-like domain as a sensory module (CHY1547). This domain contains 4Fe-4S clusters and is predicted to use molecular hydrogen for the reduction of a variety of substrates. Its fusion with the sigma-54 activator and the DNA-binding HTH__8 domain in the CHY1547 protein strongly suggests that this is a unique regulator that activates gene expression in *C. hydrogenoformans* in response to hydrogen availability. Interestingly, it is located immediately upstream of a ten-gene cluster encoding a Ni/Fe hydrogenase (CHY1537–46). Iron hydrogenases similar to the one in CHY1547 can be identified in several bacterial genomes including *S. thermophilum, Dehalococcoides ethenogenes,* and some *Clostridia;* however, they are not associated with DNA-binding domains. The only organisms where we found a homologous sigma-54 activator are *M. thermoacetica, Geobacter metallireducens, G. sufurreducens,* and *Desulfuromonas acetoxidans.*

Selenocysteine-Containing Proteins

C. hydrogenoformans possesses all known components of the selenocysteine (Sec) insertion machinery (CHY1803:SelA, CHY1802:SelB, CHY2058:SelD) and the Sec tRNA. A total of 12 selenocysteine-containing proteins (selenoproteins) were identified in *C. hydrogenoformans* genome by the Sec/Cys homology method (Table 3). For each of them, an mRNA stem-loop structure, the signature of the so-called Sec Insertion Sequence (SECIS) required for the Sec insertion, is present immediately downstream of the UGA codon. Although most of the identified selenoproteins are redox proteins, as has been shown for other bacteria and archaea [47], three are novel. Two are transporters (CHY0860, CHY0565), while the third is a methylated-DNA-protein-cysteine methyltransferase (CHY0809), a suicidal DNA repair protein that repairs alkylated guanine by transferring the alkyl group to the cysteine residue at its active site. It is striking that although this protein has been found in virtually every studied organism, only the one in *C. hydrogenoformans* has selenocysteine in place of cysteine at its active site. Therefore, this selenoprotein most likely evolved very recently, probably from a cysteine-containing protein. Similar patterns exist for the two selenocysteine-containing transporters, suggesting invention of new selenoproteins is an ongoing process in *C. hydrogenoformans.*

Translational Frameshifts

Analysis of the genome identified many potential cases of frameshifted genes. They are identified by having significant sequence similarity in two reading frames to a single homolog in another species. Examination of sequence traces suggests they are not sequencing errors. Some of these appear to be programmed frameshifts. Programmed frameshifting is a ubiquitous mechanism cells use to regulate translation or generate alternative protein products [48]. The frameshift in the gene *prfB* (CHY0163), encoding the peptide chain release factor 2, is a well-studied example of programmed frameshift that actually regulates its own translation [48].

However, many of the detected frameshifts appear to be the result of mutations from an ancestral un-frameshifted state. This is best exemplified by examination of the frameshift in the *cooS*-III gene (CHY1221), which as described above is predicted to encode one of the key components of the acetyl-CoA carbon fixation pathway. In cultures of another strain of this species (DSM 6008), a functional full-length (i.e., unframeshifted) version of this protein has been purified [24] and sequence comparisons of the gene from that strain with ours revealed many polymorphisms, including a deletion in our strain that gave rise to this frameshift (unpublished data). Studies of DSM 6008 show that in cultures grown in excess CO, the acetyl-CoA synthase (ACS, CHY1222) existed predominantly as monomer and only trace amount of CODH-III/ACS complex could be detected. On the other hand, when the CO supply was limited, CODH-III/ACS complex became the dominant form. It is plausible that CODH-III is not absolutely required for carbon fixation when the CO supply is high. Thus the frameshift and other mutations in *cooS-III* in Z-2901 may reflect the fact that it has been serially cultured in excess CO in the lab for many years. The putative lab-acquired mutations in Z-2901 are yet another reason to sequence type strains of species that have been directly acquired from culture collections and not submitted to extended laboratory culturing [49].

Conclusion

Living solely on CO is not a simple feat and the fact that *C. hydrogenoformans* does it so well makes it a model organism for this unusual metabolism. Our analysis of the genome sequence, and phylogenomic comparisons with other species, provide insights into this species' specialized metabolism. Perhaps most striking is the presence of genes that apparently encode five distinct carbon monoxide dehydrogenase complexes. Analysis of the genome has also revealed many new perspectives on the biology and evolution of this species, for example, leading us to propose its reclassification, providing further evidence that it is not a strict autotroph and revealing a previously unknown ability to sporulate. The analysis reported here and the availibility of the complete genome sequence should catalyze future studies of this organism and the hydrogenogens as a whole.

Table 3. Selenoproteins Identified in *C. hydrogenoformans* Genome

Locus	Description
CHY2058	Selenide, water dikinase
CHY2392	Glycine reductase, selenoprotein A
CHY2393	Glycine reductase, selenoprotein B
CHY0733	NAD-dependent formate dehydrogenase, alpha subunit
CHY0740	Dehydrogenase
CHY0793	Formate dehydrogenase-O, major subunit
CHY0809	Methylated-DNA-protein-cysteine methyltransferase
CHY0860	Cation-transporting ATPase, E1-E2 family
CHY0930	Heterodisulfide reductase, subunit A
CHY0931	Hydrogenase, methyl-viologen-reducing type, delta subunit
CHY1095	Thioredoxin domain selenoprotein/cytochrome C biogenesis family protein
CHY0565	Mercuric transport protein, putative

DOI: 10.1371/journal.pgen.0010065.t003

Materials and Methods

Medium composition and cultivation. *C. hydrogenoformans* Z-2901were cultivated under strictly anaerobic conditions in a basal carbonate-buffered medium composed as described [2]. However, 1.5 g l^{-1} $NaHCO_3$, 0.2 g l^{-1} $Na_2S \cdot 9\ H_2O$, 0.1 g l^{-1} yeast extract, and 2 μmol l^{-1} $NiCl_2$ were used instead of reported concentrations, and the Na_2S concentration was lowered to 0.04 g l^{-1}. Butyl rubber-stoppered bottles of 120 ml contained 50 ml medium. Bottles were autoclaved for 25' at 121 °C. Gas phases were pressurized to 170 kPa and were composed of 20% CO_2 and either 80% of N_2, H_2, or CO. Sporulation was induced by the addition of 0.01 mM $MnCl_2$ to the medium and by a transient heat shock treatment (100 °C for 5 min).

EM of *C. hydrogenoformans* endospore. Samples were fixed with 5% glutaraldehyde for 2 h and 1% OsO_4 for 4 h at 4 °C and then embedded in Epon-812. The thin sections were stained with uranyl acetate and lead citrate according to the method described by Miroshnichenko et al. [50]. The samples were observed and photographed using a JEOL JEM-1210 electron microscope.

Genome sequencing. Genomic DNA was isolated from exponential-phase cultures of *C. hydrogenoformans* Z-2901. This strain was acquired by Frank Robb from Vitali Svetlitchnyi (Bayreuth University, Germany) in 1995 after being serially grown in culture since its original isolation in 1990. Cloning, sequencing, assembly, and closure were performed as described [51,52]. The complete sequence has been assigned GenBank accession number CP000141 and is available at http://www.tigr.org.

Annotation. The gene prediction and annotation of the genome were done as previously described [51,52]. CDSs were identified by Glimmer [53]. Frameshifts or premature stop codons within CDSs were identified by comparison to other species and confirmed to be "authentic" by either their high quality sequencing reads or re-sequencing. Repetitive DNA sequences were identified using the REPUTER program [54].

Comparative genomics. To identify putative orthologs between two species, both of their proteomes were BLASTP searched against a local protein database of all complete genomes with an E-value cutoff of 1e-5. Species-specific duplications were identified and treated as one single gene (super-ortholog) for later comparison. Pair-wise mutual best-hits were then identified as putative orthologs.

Genome tree construction. Protein sequences of 31 housekeeping genes *(dnaG, frr, infC, nusA, pgk, pyrG, rplA, rplB, rplC, rplD, rplE, rplF, rplK, rplL, rplM, rplN, rplP, rplS, rplT, rpmA, rpoB, rpsB, rpsC, rpsE, rpsI, rpsJ, rpsK, rpsM, rpsS, smpB, tsf)* from genomes of interest were aligned to pre-defined HMM models and ambiguous regions were auto-trimmed according to an embedded mask. Concatenated alignments were then used to build a maximum likelihood tree using phyml [55].

Phylogenetic profile analysis. For each protein in *C. hydrogenoformans*, its presence or absence in every complete genome available at the time of this study was determined by asking whether a putative ortholog was present in that species (see above). Proteins were then grouped by their distribution patterns across species (bits of 1 and 0, 1 for presence and 0 for absence) using the CLUSTER program and the clusters were visualized using the TREEVIEW program (http://rana.lbl.gov/EisenSoftware.htm). Species were weighted by their closeness to each other to partially remove the phylogenetic component of the correlation [56].

Identification of selenoproteins. Each CDS of *C. hydrogenoformans* that ends with stop codon TGA was extended to the next stop codon TAA or TAG. It was then searched with BLASTP against the nraa database. A protein with a TGA codon pairing with a conserved Cys site was identified as a putative selenoprotein. The secondary structure

of the mRNA immediately downstream of the TGA codon was also checked using MFOLD [57] to look for a possible stem-loop structure.

Supporting Information

Table S1. Regulatory Genes in *Clostridia* Species

Found at DOI: 10.1371/journal.pgen.0010065.st001 (22 KB DOC).

Acknowledgments

We would like to thank The Institute for Genomic Research's (TIGR) Bioinformatics Department for supporting the infrastructure associated with genome annotation and analysis; the TIGR IT Department for general IT support; Dan Haft for discussions regarding selenoproteins; Anne Ciecko, Kristi Berry, and Chris Larkin for initial work on genome closure; Patrick Eichenberger, Richard Losick, and Jim Brannigan for discussions about sporulation; Vitali Svetlichnyi for general discussions about the organism; Terry Utterback and Tamara Feldblyum for coordinating the shotgun sequencing; William C. Nierman for making the genomic libraries; and Claire M. Fraser for supporting the selection of this genome as part of the DOE project referenced below. The sequencing, annotation, and analysis of the genome were supported by United States Department of Energy, Office of Biological Energy Research, Co-Operative Agreement DE-FC0295ER61962. Support for other work in association with this publication came from the National Science Foundation (NSF/MCB-02383387 for FTR and JMG) and the National Institutes of Health (GM072285 for LEU and IBZ).

Competing interests. The authors have declared that no competing interests exist.

Author contributions. MW, FTR, and JAE conceived and designed the experiments. JMG and FTR performed the physiology experiments. MW, QR, LEU, IBZ, FTR, and JAE analyzed the data. MW, IBZ, FTR, and JAE wrote the paper. LJT and KMJ closed the genome. ASD, SCD, LMB, RJD, RM, SAS, JFK, and WCN annotated the genome. ■

References

1. Svetlitchnyi V, Peschel C, Acker G, Meyer O (2001) Two membrane-associated NiFeS-carbon monoxide dehydrogenases from the anaerobic carbon-monoxide-utilizing eubacterium *Carboxydothermus hydrogenoformans*. J Bacteriol 183: 5134–5144.
2. Svetlichny VA, Sokolova TG, Gerhardt M, Ringpfeil M, Kostrikina NA, et al. (1991) *Carboxydothermus hydrogenoformans* gen. nov., sp. nov., a CO-utilizing thermophilic anaerobic bacterium from hydrothermal environments of Kunashir Island. Syst Appl Microbiol 14: 254–260.
3. Svetlichny V, Sokolova T, Kostrikina N, Lysenko A (1994) A new thermophilic anaerobic carboxydotrophic bacterium *Carboxydothermus restrictus* sp. nov. Mikrobiologiya 63: 294–297.
4. Bonch-Osmolovskaya E, Miroshnichenko M, Slobodkin A, Sokolova T, Karpov G, et al. (1999) Biodiversity of anaerobic lithotrophic prokaryotes in terrestrial hot springs of Kamchatka. Microbiology 68: 343–351.
5. Sokolova TG, Gonzalez JM, Kostrikina NA, Chernyh NA, Tourova TP, et al. (2001) *Carboxydobrachium pacificum* gen. nov., sp. nov., a new anaerobic, thermophilic, CO-utilizing marine bacterium from Okinawa Trough. Int J Syst Evol Microbiol 51: 141–149.
6. Sokolova TG, Gonzalez JM, Kostrikina NA, Chernyh NA, Slepova TV, et al. (2004) *Thermosinus carboxydivorans* gen. nov., sp. nov., a new anaerobic, thermophilic, carbon-monoxide-oxidizing, hydrogenogenic bacterium from a hot pool of Yellowstone National Park. Int J Syst Evol Microbiol 54: 2353–2359.
7. Sokolova TG, Kostrikina NA, Chernyh NA, Tourova TP, Kolganova TV, et al. (2002) *Carboxydocella thermautotrophica* gen. nov., sp. nov., a novel anaerobic, CO-utilizing thermophile from a Kamchatkan hot spring. Int J Syst Evol Microbiol 52: 1961–1967.
8. Henstra AM, Stams AJ (2004) Novel physiological features of *Carboxydothermus hydrogenoformans* and *Thermoterrabacterium ferrireducens*. Appl Environ Microbiol 70: 7236–7240.
9. Mojica FJ, Diez-Villasenor C, Soria E, Juez G (2000) Biological significance of a family of regularly spaced repeats in the genomes of *Archaea, Bacteria,* and mitochondria. Mol Microbiol 36: 244–246.
10. Mojica FJ, Ferrer C, Juez G, Rodriguez-Valera F (1995) Long stretches of short tandem repeats are present in the largest replicons of the *Archaea Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. Mol Microbiol 17: 85–93.
11. Peng X, Brugger K, Shen B, Chen L, She Q, et al. (2003) Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in *Sulfolobus* genomes. J Bacteriol 185: 2410–2417.
12. Rocha E (2002) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol 10: 393–395.
13. Dervyn E, Suski C, Daniel R, Bruand C, Chapuis J, et al. (2001) Two essential DNA polymerases at the bacterial replication fork. Science 294: 1716–1719.
14. Brewer BJ (1988) When polymerases collide: Replication and the transcriptional organization of the *E. coli* chromosome. Cell 53: 679–686.
15. Liu B, Alberts BM (1995) Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex. Science 267: 1131–1137.
16. Garrity GM, Bell JA, Lilburn TG (2004) Taxonomic outline of the prokaryotes. Bergey's manual of systematic bacteriology, Second Edition. New York: Springer-Verlag. 2816 p.
17. Ohno M, Shiratori H, Park MJ, Saitoh Y, Kumon Y, et al. (2000) *Symbiobacterium thermophilum* gen. nov., sp. nov., a symbiotic thermophile that depends on co-culture with a *Bacillus* strain for growth. Int J Syst Evol Microbiol 50 Pt 5: 1829–1832.
18. Ueda K, Yamashita A, Ishikawa J, Shimada M, Watsuji TO, et al. (2004) Genome sequence of *Symbiobacterium thermophilum,* an uncultivable bacterium that depends on microbial commensalism. Nucleic Acids Res 32: 4937–4944.
19. Ferry JG (1995) CO dehydrogenase. Annu Rev Microbiol 49: 305–333.
20. Ragsdale SW, Kumar M (1996) Nickel-containing carbon monoxide dehydrogenase/acetyl-CoA synthase(,). Chem Rev 96: 2515–2540.
21. Fox JD, He Y, Shelver D, Roberts GP, Ludden PW (1996) Characterization of the region encoding the CO-induced hydrogenase of *Rhodospirillum rubrum*. J Bacteriol 178: 6200–6208.
22. Soboh B, Linder D, Hedderich R (2002) Purification and catalytic properties of a CO-oxidizing:H2-evolving enzyme complex from *Carboxydothermus hydrogenoformans*. Eur J Biochem 269: 5712–5721.
23. Ragsdale SW (2004) Life with carbon monoxide. Crit Rev Biochem Mol Biol 39: 165–195.
24. Svetlitchnyi V, Dobbek H, Meyer-Klaucke W, Meins T, Thiele B, et al. (2004) A functional Ni-Ni-[4Fe-4S] cluster in the monomeric acetyl-CoA synthase from *Carboxydothermus hydrogenoformans*. Proc Natl Acad Sci U S A 101: 446–451.
25. Roberts DL, James-Hagstrom JE, Garvin DK, Gorst CM, Runquist JA, 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/Fe-S protein, and methyltransferase. Proc Natl Acad Sci U S A 86: 32–36.
26. Adams MW, Jenney FE Jr., Clay MD, Johnson MK (2002) Superoxide reductase: Fact or fiction? J Biol Inorg Chem 7: 647–652.
27. Lynch MC, Kuramitsu HK (1999) Role of superoxide dismutase activity in the physiology of *Porphyromonas gingivalis*. Infect Immun 67: 3367–3375.
28. Weinberg MV, Jenney FE Jr., Cui X, Adams MW (2004) Rubrerythrin from the hyperthermophilic archaeon *Pyrococcus furiosus* is a rubredoxin-dependent, iron-containing peroxidase. J Bacteriol 186: 7888–7895.
29. Sztukowska M, Bugno M, Potempa J, Travis J, Kurtz DM Jr. (2002) Role of rubrerythrin in the oxidative stress response of *Porphyromonas gingivalis*. Mol Microbiol 44: 479–488.
30. Gonzalez JM, Robb FT (2000) Genetic analysis of *Carboxydothermus hydrogenoformans* carbon monoxide dehydrogenase genes cooF and cooS. FEMS Microbiol Lett 191: 243–247.
31. Dobbek H, Svetlitchnyi V, Gremer L, Huber R, Meyer O (2001) Crystal structure of a carbon monoxide dehydrogenase reveals a [Ni-4Fe-5S] cluster. Science 293: 1281–1285.
32. Schubel U, Kraut M, Morsdorf G, Meyer O (1995) Molecular characterization of the gene cluster coxMSL encoding the molybdenum-containing carbon monoxide dehydrogenase of *Oligotropha carboxidovorans*. J Bacteriol 177: 2197–2203.
33. Maynard EL, Lindahl PA (1999) Evidence of a molecular tunnel connecting the active sites for CO2 reduction and acetyl-coA synthesis in acetyl-coA synthase from *Clostridium thermoaceticum*. J Am Chem Soc 121: 9221–9222.
34. Seravalli J, Ragsdale SW (2000) Channeling of carbon monoxide during anaerobic carbon dioxide fixation. Biochemistry 39: 1274–1277.
35. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285–4288.
36. Eichenberger P, Jensen ST, Conlon EM, van Ooij C, Silvaggi J, et al. (2003) The sigmaE regulon and the identification of additional sporulation genes in *Bacillus subtilis*. J Mol Biol 327: 945–972.
37. Molle V, Fujita M, Jensen ST, Eichenberger P, Gonzalez-Pastor JE, et al. (2003) The Spo0A regulon of *Bacillus subtilis*. Mol Microbiol 50: 1683–1701.
38. Zuber U, Drzewiecki K, Hecker M (2001) Putative sigma factor SigI (YkoZ) of *Bacillus subtilis* is induced by heat shock. J Bacteriol 183: 1472–1475.
39. Stragier P, Losick R (1996) Molecular genetics of sporulation in *Bacillus subtilis*. Annu Rev Genet 30: 297–241.
40. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390: 249–256.
41. Nolling J, Breton G, Omelchenko MV, Makarova KS, Zeng Q, et al. (2001) Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. J Bacteriol 183: 4823–4838.

42. Bao Q, Tian Y, Li W, Xu Z, Xuan Z, et al. (2002) A complete sequence of the *T. tengcongensis* genome. Genome Res 12: 689–700.
43. Ulrich LE, Koonin EV, Zhulin IB (2005) One-component systems dominate signal transduction in prokaryotes. Trends Microbiol 13: 52–56.
44. Faguy DM, Jarrell KF (1999) A twisted tale: The origin and evolution of motility and chemotaxis in prokaryotes. Microbiology 145 (Pt 2): 279–281.
45. He Y, Shelver D, Kerby RL, Roberts GP (1996) Characterization of a CO-responsive transcriptional activator from *Rhodospirillum rubrum*. J Biol Chem 271: 120–123.
46. Taylor BL, Zhulin IB (1999) PAS domains: Internal sensors of oxygen, redox potential, and light. Microbiol Mol Biol Rev 63: 479–506.
47. Kryukov GV, Gladyshev VN (2004) The prokaryotic selenoproteome. EMBO Rep 5: 538–543.
48. Farabaugh PJ (1996) Programmed translational frameshifting. Annu Rev Genet 30: 507–528.
49. Ward N, Eisen J, Fraser C, Stackebrandt E (2001) Sequenced strains must be saved from extinction. Nature 414: 148.
50. Miroshnichenko ML, Bonch-Osmolovskaya EA, Neuer A, Kostrikina NA, Chernych NA, et al. (1989) *Thermococcus stetteri* sp. nov., a new extremely thermophilic marine sulfur metabolizing archaebacterium. Syst Appl Microbiol 12: 257–262.
51. Eisen JA, Nelson KE, Paulsen IT, Heidelberg JF, Wu M, et al. (2002) The complete genome sequence of *Chlorobium tepidum* TLS, a photosynthetic, anaerobic, green-sulfur bacterium. Proc Natl Acad Sci U S A 99: 9509–9514.
52. Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, et al. (2004) Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: A streamlined genome overrun by mobile genetic elements. PLoS Biol 2: e69. DOI: 10.1371/journal.pbio.0020069
53. Salzberg SL, Delcher AL, Kasif S, White O (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res 26: 544–548.
54. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, et al. (2001) REPuter: The manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29: 4633–4642.
55. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
56. Eisen JA, Wu M (2002) Phylogenetic analysis and gene functional predictions: Phylogenomics in action. Theor Popul Biol 61: 481–487.
57. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31: 3406–3415.
58. Onyenwoke RU, Brill JA, Farahi K, Wiegel J (2004) Sporulation genes in members of the low G+C Gram-type-positive phylogenetic branch (Firmicutes). Arch Microbiol 182: 182–192.

Note Added in Proof

It has come to our attention that a complementary comparison of sporulation genes in various Firmicutes was published in 2004 [58]. This study identified homologs of known sporulation genes in Firmicutes by experimental methods and genome analysis. The authors then used these results to study the evolution of sporulation and known sporulation genes.