Environmental genomics reveals a single species ecosystem deep within the Earth

Dylan Chivian[1,2]*, Eoin L. Brodie[2,3], Eric J. Alm[2,4], David E. Culley[5], Paramvir S. Dehal[1,2], Todd Z. DeSantis[2,3], Thomas M. Gihring[6], Alla Lapidus[7], Li-Hung Lin[8], Stephen R. Lowry[7], Duane P. Moser[9], Paul Richardson[7], Gordon Southam[10], Greg Wanger[10], Lisa M. Pratt[11,12], Gary L. Andersen[2,3], Terry C. Hazen[2,3,12], Fred J. Brockman[13], Adam P. Arkin[1,2,14], Tullis C. Onstott[12,15]

[1]*Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA*
[2]*Virtual Institute for Microbial Stress and Survival, Berkeley, CA*
[3]*Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA*
[4]*Departments of Biological and Civil & Environmental Engineering, MIT, Cambridge, MA*
[5]*Energy & Efficiency Technology Division, Pacific Northwest National Laboratory, Richland, WA*
[6]*Department of Oceanography, Florida State University, Tallahassee, FL*
[7]*Genomic Technology Program, DOE Joint Genomics Institute, Berkeley, CA*
[8]*Department of Geosciences, National Taiwan University, Taipei, Taiwan*
[9]*Division of Earth and Ecosystem Sciences, Desert Research Institute, Las Vegas, NV*
[10]*Department of Earth Sciences, University of Western Ontario, London, ON, Canada*
[11]*Department of Geological Sciences, Indiana University, Bloomington, IN*
[12]*IPTAI NASA Astrobiology Institute, Bloomington, IN*
[13]*Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA*
[14]*Department of Bioengineering, University of California, Berkeley, CA*
[15]*Department of Geosciences, Princeton University, Princeton, NJ*

*To whom correspondence should be addressed:
Dr. Dylan Chivian
Lawrence Berkeley National Laboratory
1 Cyclotron Road, MS Calvin
Berkeley, CA 94720 USA
E-mail: DCChivian@lbl.gov

ONE SENTENCE SUMMARY

DNA from 2.8 km deep in the Earth's crust reveals the genetic complement necessary for a single species ecosystem.

ABSTRACT

DNA from low biodiversity fracture water collected at 2.8 km depth in a South African gold mine was sequenced and assembled into a single, complete genome. This bacterium, *Candidatus Desulforudis audaxviator*, comprises > 99.9% of the microorganisms inhabiting the fluid phase of this particular fracture. Its genome indicates a motile, sporulating, sulfate reducing, chemoautotrophic thermophile that can fix its own nitrogen and carbon using machinery shared with archaea. *Candidatus Desulforudis audaxviator* is capable of an independent lifestyle well suited to long-term isolation from the photosphere deep within Earth's crust, and offers the first example of a natural ecosystem that appears to have its biological component entirely encoded within a single genome.

A more complete picture of life on Earth, and even life *in* the Earth, has recently become possible by extracting and sequencing DNA from an environmental sample, a process called “environmental genomics” or "metagenomics" (*1-8*). This approach allows us to identify members of microbial communities and to characterize the abilities of the dominant members even when isolation of those organisms has proven intractable. However, with a few exceptions (*5, 7*), assembling complete or even near-complete genomes for a substantial portion of the member species is usually hampered by the complexity of natural microbial communities.

In addition to elevated temperatures and a lack of O_2, conditions within Earth’s crust at depths > 1 km are fundamentally different from those of the surface and deep ocean environments. Severe nutrient limitation is believed to result in cell doubling times ranging from 100 to 1,000 years (*9-11*) and as a result subsurface microorganisms might be expected to reduce their reproductive burden and exhibit the streamlined genomes of specialists or spend most of their time in a state of semi-senescence waiting for the return of favorable conditions. Such microorganisms are of particular interest as they permit insight into a mode of life independent of the photosphere.

One bacterium belonging to the *Firmicutes* phylum (Fig. 1a), which we herein name “*Candidatus Desulforudis audaxviator*”, is prominent in small subunit (SSU or 16S) rRNA gene clone libraries (*11-14*) from almost all fracture fluids sampled to date from depths greater than 1.5 km across the Witwatersrand Basin (covering 150 x 300 km near Johannesburg, South Africa). This bacterium was shown in a previous geochemical and 16S rRNA gene study (*11*) to dominate the indigenous microorganisms found in a fracture zone at 2.8 km below land surface at level 104 of the Mponeng mine (MP104).

Although Lin, *et al*. (*11*) discovered that this fracture zone contained the least diverse natural free-living microbial community reported at that time, exceeding the ~80% dominance by the methanogenic archaeon IUA5/6 of a comparatively shallow subsurface community in Idaho (*15*), we were nonetheless surprised when the current environmental genomics study revealed only one species was actually present within the fracture fluid. Furthermore, we found that the single genome that assembled appeared to possess all of the metabolic capabilities necessary for an independent lifestyle. This gene complement was consistent with the previous geochemical and thermodynamic analyses at the ambient ~60°C temperature and pH of 9.3, which indicated formate and H_2 as possessing the greatest potential among candidate electron donors, with sulfate (SO_4^{2-}) reduction as the dominant electron accepting process (*11*).

DNA was extracted from ~5,600 L of filtered fracture water using a protocol that has been demonstrated to be effective on a broad range of bacterial and archaeal species, including recalcitrant organisms (supporting online material, "SOM"). A single, complete, 2.35 megabase pair (Mbp) genome was assembled using a combination of shotgun Sanger sequencing and 454 pyrosequencing (SOM). Similar to other studies that obtained near-complete consensus genomes from environmental samples (*5, 16*), heterogeneity in the population of the dominant species as measured single nucleotide polymorphisms ("SNP") was quite low, showing only 32 positions with a SNP observed more than once (Table S7), suggesting strong selective pressure.

The DNA recovered from the filter, assuming the capture of cells and extraction of DNA from those cells was indeed comprehensive, revealed that this genome represented the only species present in the fluid phase of the fracture. Of the ~0.1% of

95 microbial reads not belonging to *D. audaxviator* (Fig. 1c,d, Tables S5 and S6), about ½
96 represented clear contamination (Table S6), the removal of which resulted in only 22 of
97 29,179 Sanger reads (0.075%) and 59 of 500,008 pyrosequencing reads (0.012%) that
98 could be from other microorganisms. However, even with the great care taken in
99 collecting an uncontaminated sample, it remains possible that some or all of the trace
100 reads are from organisms not indigenous to the fracture. An upper-bound estimate of the
101 contribution of any microorganism other than *D. audaxviator* to the community (Table
102 S6) offered at most only 5 Sanger reads (0.017%) corresponding to γ-Proteobacteria, and
103 at most 9 pyrosequencing reads (0.0018%) corresponding to α-Proteobacteria. Even
104 taking the higher of these proportions suggested that it is unlikely that *D. audaxviator*,
105 and indeed the functioning of the ecosystem, is metabolically dependent upon organisms
106 that would be outnumbered by about 5,000 to 1 (or about 50,000 to 1 from the
107 pyrosequencing data). However, we could not rule out the presence of organisms that
108 might adhere to the surfaces of the fracture or that were smaller than the 0.2 μm filter
109 pore size. It may be that uncaptured microorganisms and bacteriophage, in addition to
110 potential trace species, do play a role in the MP104 ecosystem, perhaps as reservoirs of
111 genetic variation (*17*).

112 We analyzed the genome of *D. audaxviator* using MicrobesOnline
113 (http://www.microbesonline.org) (*18*). If *D. audaxviator* is indeed the solitary resident of
114 this habitat, then its genome should contain the complete genetic complement for
115 maintaining the biological component of the ecosystem prohibiting extreme reduction of
116 its genome. The genome (Table 1), at 2.35 Mbp, was smaller than the 3 Mbp of its
117 nearest sequenced relative *Pelotomaculum thermopropionicum*. It contained 2157

predicted protein coding genes, more than found in streamlined free-living microorganisms, which typically have fewer than 2000 genes (*19*). We found all of the processes necessary for life encoded within the genome, including energy metabolism, carbon fixation, and nitrogen fixation.

Consistent with the thermodynamic evaluation (*11*) that SO_4^{2-} offers the most energetically favorable electron acceptor, the genome possesses the capacity for dissimilatory sulfate reduction (DSR) (Figs. 2, 3, and Table S13) with a gene repertoire like that of other SO_4^{2-} reducing microorganisms (*20*). These genes are present in a set of operons (labeled SR1-SR11 in Fig. 2) and include an extra copy of an archael-type sulfate adenylyltransferase (Sat) (Figure S5) and a H^+-translocating pyrophosphatase, both of which appear to be a consequence of horizontal gene transfer (HGT). High potential electrons enter primarily *via* the activity of a variety of hydrogenases upon H_2 (Table S24).

Carbon assimilation may be from a variety of sources depending on local conditions. The genome contains sugar and amino acid transporters (Fig. 3 and Table S20), suggesting that, at locations where biodensity is high, heterotrophic sources could be used, including recycling of dead cells. At MP104, where biodensity is low, carbon is assimilated from inorganic sources. *D. audaxviator* appeared not to be using the reverse TCA cycle (Table S23), but did have all the machinery of the acetyl-CoA synthesis (Wood-Ljungdahl) pathway (*21, 22*), which utilizes carbon monoxide dehydrogenase (CODH) for the assimilation of inorganic carbon (Figs. 2, 3, S7, and Table S14). Entry of CO_2 substrate into the cell may be accomplished by its anionic species through a putative carbonate ABC transporter or a putative bicarbonate/Na^+ symporter (Fig. 3 and

Table S20). Formate and CO may serve as alternate, more direct, carbon sources in other fractures when sufficiently abundant (Table S2).

The ambient concentration of ammonia in the fracture water ($[NH_3]+[NH_4^+]$ = ~100 μM) (*11*) appears sufficient for *D. audaxviator* (which has an ammonium transporter as well as glutamine synthetase), to obtain its nitrogen from ammonia without resorting to an energetically costly nitrogenase conversion of N_2 to ammonia. Nonetheless, a nitrogenase is present in the genome (Fig. 2 and Table S15) that is more similar to archaeal types, including high temperature variants (*23*), than the nitrogenase of *Desulfotomaculum reducens* (Figs. S4, S8). It may be that *D. audaxviator* is not always presented with sufficient amounts of ammonia, so the versatility provided by the horizontally acquired nitrogenase may have contributed significantly to the success of *D. audaxviator* in colonizing such habitats.

Desulforudis audaxviator shares other genes with archaea that may confer benefits in extreme environments. In addition to the unusual nitrogenase and sulfate adenylyltransferase, acquisitions by ancestors of *D. audaxviator* include (Table S10) a second CODH system (CODH1 in Fig. 2 and Fig. S7), cobalamin biosynthesis protein CobN, and genes for the formation of gas vesicles. It also has two clustered regularly interspaced short palindromic repeat ("CRISPR") regions (Table S12), that are used for viral defense (*24*), occur in the genome with adjacent CRISPR-associated genes ("CAS"), some of which are horizontally shared between *D. audaxviator* and archaea.

D. audaxviator's ability to colonize independently is also assisted by its possession of all of the amino acid synthesis pathways (Table S21). Other factors that may confer fitness in this environment are the ability to form endospores (Table S16) and

the potential for it to grow in deeper, hotter conditions (Table S9). *D. audaxviator* appears capable of sensing nutrients (Table S19) in its environment, and possesses flagella (Table S18) that permit motility along chemical gradients, such as those that occur at the mineral surfaces of the fracture (*25*). One ability that *D. audaxviator* is lacking is a complete system for oxygen resistance (Table S25), suggesting the long-term isolation from O_2.

The MP104 fracture contains the simplest natural environmental microbial community yet described, and has yielded a single, complete genome of an uncultured microorganism using environmental genomics. *Desulforudis audaxviator's* ability to reduce SO_4^{2-} grants access to the most energetically favorable electron acceptor in the fracture zones of the Witwatersrand basin (*26*). Additionally, inherited characteristics of *D. audaxviator*, such as motility, sporulation, and carbon fixation, have been complemented by horizontally acquired systems frequently found in archaea. These abilities have enabled *D. audaxviator* to colonize the deep subsurface, a process that, unlike surface habitats which permit more immediate access, has required fitness throughout the history of the colonization. This "bold traveler" (*audax viator*) has revealed a mode of life isolated from the photosphere, capturing all of the roles necessary for an independent lifestyle and showing that it is possible to encode the entire biological component of a simple ecosystem within a single genome.

REFERENCES AND NOTES

1. A. M. Deutschbauer, D. Chivian, A. P. Arkin, *Curr Opin Biotechnol* **17**, 229 (2006).
2. O. Beja *et al.*, *Environ Microbiol* **2**, 516 (2000).
3. M. R. Rondon *et al.*, *Appl Environ Microbiol* **66**, 2541 (2000).
4. J. C. Venter, *Science* **304**, 66 (2004).
5. G. W. Tyson *et al.*, *Nature* **428**, 37 (2004).
6. S. G. Tringe, *Science* **308**, 554 (2005).
7. M. Strous *et al.*, *Nature* **440**, 790 (2006).
8. D. B. Rusch *et al.*, *PLoS Biol* **5**, e77 (2007).
9. T. J. Phelps, E. M. Murphy, S. M. Pfiffer, D. C. White, *Microbial Ecology* **28**, 335 (1994).
10. B. B. Jørgensen, S. D'Hondt, *Science* **314**, 932 (2006).
11. L. H. Lin *et al.*, *Science* **314**, 479 (2006).
12. D. P. Moser *et al.*, *Appl Environ Microbiol* **71**, 8773 (2005).
13. D. P. Moser *et al.*, *Geomicrobiology Journal* **20**, 517 (2003).
14. T. M. Gihring *et al.*, *Geomicrobiology Journal* **23**, 415 (2006).
15. F. H. Chapelle *et al.*, *Nature* **415**, 312 (2002).
16. V. Zverlov *et al.*, *J Bacteriol* **187**, 2203 (2005).
17. M. L. Sogin *et al.*, *Proc Natl Acad Sci U S A* **103**, 12115 (2006).
18. E. J. Alm *et al.*, *Genome Res* **15**, 1015 (2005).
19. S. J. Giovannoni *et al.*, *Science* **309**, 1242 (2005).
20. M. Mussmann *et al.*, *J Bacteriol* **187**, 7126 (2005).

21. H. L. Drake, S. L. Daniel, *Research in Microbiology* **155**, 869 (2005).

22. M. Wu *et al.*, *PLoS Genet* **1**, e65 (2005).

23. M. P. Mehta, J. A. Baross, *Science* **314**, 1783 (2006).

24. R. Barrangou *et al.*, *Science* **315**, 1709 (2007).

25. G. Wanger, T. C. Onstott, G. Southam, *Geomicrobiology Journal* **23**, 443 (2006).

26. T. C. Onstott *et al.*, *Geomicrobiology Journal* **23**, 369 (2006).

27. L. Lefticariu, L. M. Pratt, E. M. Ripley, *Geochimica. Cosmochim. Acta* **70**, 4889 (2006).

28. We thank Jill Banfield and Gene Tyson for helpful discussion. We thank Jim Bruckner and Brett Baker for assistance with microscopy and Falk Warnecke for advice on 16S FISH. We also thank Thomas Kieft, Grant Zane, and the MicrobesOnline team (Morgan Price, Keith Keller, and Katherine Huang) for advice. We are indebted to Dave Kershaw and colleagues at the Mponeng mine and AngloGold Ashanti Limited, RSA. This work was part of the Virtual Institute for Microbial Stress and Survival (http://vimss.lbl.gov), supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Program:GTL through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. This work was also supported by the NASA Astrobiology Institute through award NNA04CC03A to the IPTAI Team co-directed by LMP and TCO. APA received support from the HHMI. The genome sequence and 16S library sequences reported in this study have been deposited in GenBank under the accession numbers CP000860 and EU730965 - EU731008 respectively.

SUPPORTING ONLINE MATERIAL

www.sciencemag.org/*XXXXXXXXXXX [URL PENDING]*
Materials and Methods
Figs S1 to S8
Tables S1 to S26
References

TABLES

Table 1. General Features of the *Desulforudis audaxviator* genome.

Feature	Value
Genome size (bp)	2,349,476
G+C content (%)	60.9
Predicted protein coding genes (CDS/ORF)	2157
Genes without homology to other organisms (ORFans)	210
Pseudogenes derived from a protein coding gene	83
Average CDS/ORF length (bp)	910
Longest CDS/ORF length (bp)	5601
Percent of genome protein coding (%)	86.8
Ribosomal RNA operons (16S-23s-5S)	2
Transfer RNAs (all amino acids represented, including SeC)	45
Other non-protein coding RNAs	7
CRISPR regions	2
Mobile element (transposons/integrases) gene groups	30
Mobile element genes	83
Other phage-associated genes	18

"bp": base pairs of DNA

FIGURE LEGENDS

Figure 1. Phylogeny and population structure.

(a) Phylogenetic placement of *D. audaxviator* based on protein sequences of universal protein families (Table S3). High bootstrap value supported nodes are indicated with circles. (b) Classifications of SSU rRNA gene clones from PCR amplification of filter extract (Fig. S3). (c) Proportions of Sanger sequencing reads from shotgun clone library of filter extract. Reads classified as *D. audaxviator* by match to assembled genome or by match to sequenced organisms (Table S6). (d) Proportions of 454 pyrosequencing reads directly from filter extract. Reads classified as *D. audaxviator* by match to assembled genome or by match to sequenced organisms (Table S6).

Figure 2. Genome of *D. audaxviator*, with key genes highlighted.

Innermost ring: GC skew (average of (G-C)/(G+C) over 10000 bases, plotted every 1000 bases). Transition at the top (near dnaA) is origin of replication. **Second ring:** G+C content (average of (G+C) over 10000 bases, plotted every 1000 bases), with greater than average value (61%) in blue and below average in red. Below average G+C regions that result from CRISPR sequences are indicated in grey. **Third and fourth rings:** predicted protein coding genes on each strand. Genes with homologs only found within closest clade species (including ORFan genes) are in cyan, genes that are found only within closest clade species and within archaea (resulting from horizontal transfer) in magenta, and all other genes in black. **Outer boxes:** Genes of interest are shown around the ring as operons for sulfate reduction ("SR"), carbon fixation via acetyl-CoA synthesis pathway ("CF"), and nitrogen fixation ("NF"). Horizontally acquired genes

shared with archaea specific to *D. audaxviator* and its nearest relatives are colored according to the key.

Figure 3. Model of the single species ecosystem at MP104.

D. audaxviator's machinery is shown in a cartoon representation, including pathways for sulfate reduction, nitrogen fixation, and carbon fixation. Signal transduction proteins are reported with the number found in parentheses, and have the abbreviations "MCP": methyl-accepting chemotaxis proteins, "HPK": histidine protein kinases, "RR": response regulators. Transporters include approximate substrates. Also shown are the environmental sources of energy and material for the ecosystem, as detailed in Lin, *et al*. (*11*), shown experimentally by Lefticariu, *et al.* (*27*), and described in the SOM.

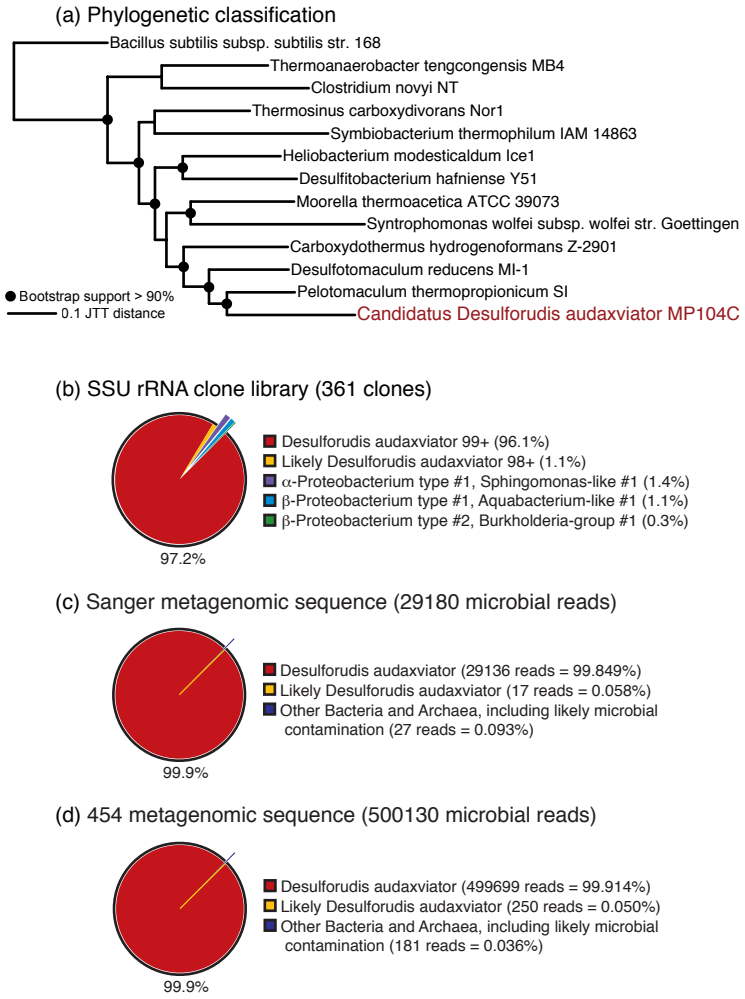

(a) Phylogenetic classification
Bacillus subtilis subsp. subtilis str. 168
Thermoanaerobacter tengcongensis MB4
Clostridium novyi NT
Thermosinus carboxydivorans Nor1
Symbiobacterium thermophilum IAM 14863
Heliobacterium modesticaldum Ice1
Desulfitobacterium hafniense Y51
Moorella thermoacetica ATCC 39073
Syntrophomonas wolfei subsp. wolfei str. Goettingen
Carboxydothermus hydrogenoformans Z-2901
Desulfotomaculum reducens MI-1
Pelotomaculum thermopropionicum SI
Candidatus Desulforudis audaxviator MP104C
Bootstrap support > 90%
0.1 JTT distance
(b) SSU rRNA clone library (361 clones)
Desulforudis audaxviator 99+ (96.1%)
Likely Desulforudis audaxviator 98+ (1.1%)
α-Proteobacterium type #1, Sphingomonas-like #1 (1.4%)
β-Proteobacterium type #1, Aquabacterium-like #1 (1.1%)
β-Proteobacterium type #2, Burkholderia-group #1 (0.3%)
97.2%
(c) Sanger metagenomic sequence (29180 microbial reads)
Desulforudis audaxviator (29136 reads = 99.849%)
Likely Desulforudis audaxviator (17 reads = 0.058%)
Other Bacteria and Archaea, including likely microbial contamination (27 reads = 0.093%)
99.9%
(d) 454 metagenomic sequence (500130 microbial reads)
Desulforudis audaxviator (499699 reads = 99.914%)
Likely Desulforudis audaxviator (250 reads = 0.050%)
Other Bacteria and Archaea, including likely microbial contamination (181 reads = 0.036%)
99.9%

Figure 1

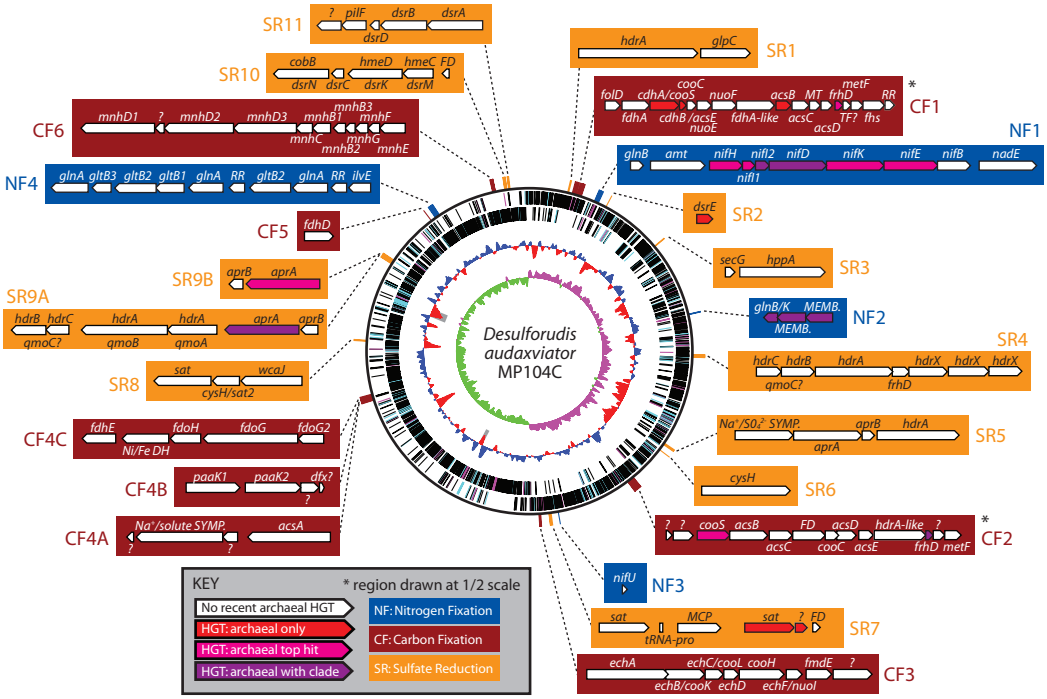

Desulforudis audaxviator MP104C
SR11
SR10
CF6
NF4
CF5
SR9B
SR9A
SR8
CF4C
CF4B
CF4A
SR1
CF1
NF1
SR2
SR3
NF2
SR4
SR5
SR6
CF2
NF3
SR7
CF3
KEY
* region drawn at 1/2 scale
No recent archaeal HGT
HGT: archaeal only
HGT: archaeal top hit
HGT: archaeal with clade
NF: Nitrogen Fixation
CF: Carbon Fixation
SR: Sulfate Reduction

Figure 2

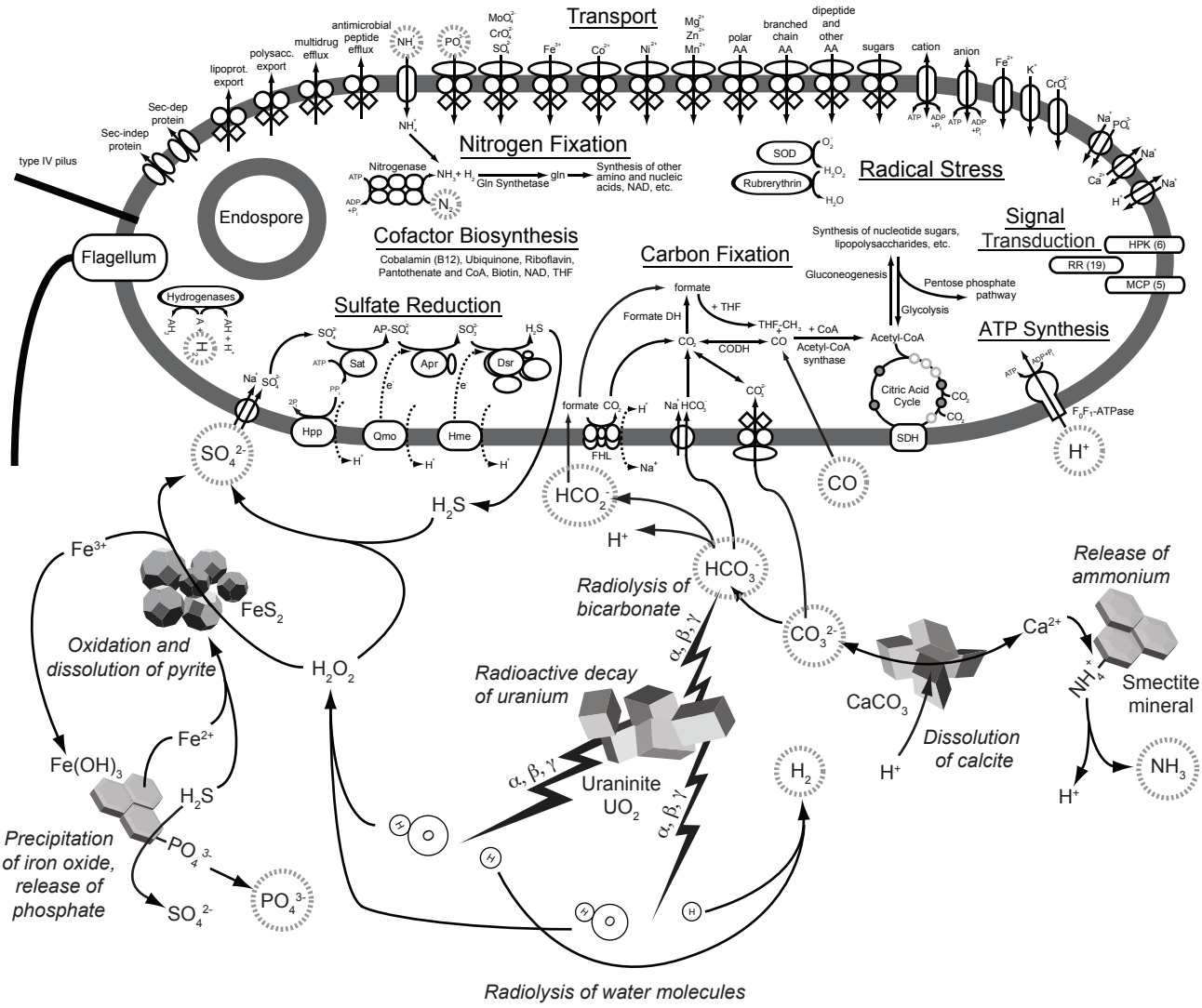

Transport
antimicrobial peptide efflux
multidrug efflux
polysacc. export
lipoprot. export
Sec-dep protein
Sec-indep protein
type IV pilus
Endospore
Flagellum
Nitrogen Fixation
Nitrogenase
Gln Synthetase
gln
Synthesis of other amino and nucleic acids, NAD, etc.
Radical Stress
SOD
Rubrerythrin
Cofactor Biosynthesis
Cobalamin (B12), Ubiquinone, Riboflavin, Pantothenate and CoA, Biotin, NAD, THF
Signal Transduction
HPK (6)
RR (19)
MCP (5)
Carbon Fixation
Synthesis of nucleotide sugars, lipopolysaccharides, etc.
Gluconeogenesis
Pentose phosphate pathway
Glycolysis
Formate DH
+ THF
CODH
Acetyl-CoA synthase
Acetyl-CoA
Citric Acid Cycle
SDH
Sulfate Reduction
Sat
Apr
Dsr
Hpp
Qmo
Hme
Hydrogenases
FHL
ATP Synthesis
F0F1-ATPase
Fe3+
FeS2
Oxidation and dissolution of pyrite
H2O2
Fe2+
Fe(OH)3
H2S
Precipitation of iron oxide, release of phosphate
SO4 2-
PO4 3-
H2S
HCO3 -
H+
HCO3 -
Radiolysis of bicarbonate
α, β, γ
Radioactive decay of uranium
Uraninite UO2
α, β, γ
α, β, γ
H2
CO3 2-
CaCO3
Ca2+
H+
Dissolution of calcite
Release of ammonium
NH4 +
Smectite mineral
NH3
H+
Radiolysis of water molecules

Figure 3

Supporting Online Material for

Environmental genomics reveals a single species ecosystem deep within the Earth

Dylan Chivian[1,2*], Eoin L. Brodie[2,3], Eric J. Alm[2,4], David E. Culley[5],
Paramvir S. Dehal[1,2], Todd Z. DeSantis[2,3], Thomas M. Gihring[6], Alla Lapidus[7],
Li-Hung Lin[8], Stephen R. Lowry[7], Duane P. Moser[9], Paul Richardson[7],
Gordon Southam[10], Greg Wanger[10], Lisa M. Pratt[11,12], Gary L. Andersen[2,3],
Terry C. Hazen[2,3,12], Fred J. Brockman[13], Adam P. Arkin[1,2,14], Tullis C. Onstott[12,15]

[1]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA. [2]Virtual Institute for Microbial Stress and Survival, Berkeley, CA. [3]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA. [4]Departments of Biological and Civil & Environmental Engineering, MIT, Cambridge, MA. [5]Energy & Efficiency Technology Division, Pacific Northwest National Laboratory, Richland, WA. [6]Department of Oceanography, Florida State University, Tallahassee, FL. [7]Genomic Technology Program, DOE Joint Genomics Institute, Berkeley, CA. [8]Department of Geosciences, National Taiwan University, Taipei, Taiwan. [9]Division of Earth and Ecosystem Sciences, Desert Research Institute, Las Vegas, NV. [10]Department of Earth Sciences, University of Western Ontario, London, ON, Canada. [11]Department of Geological Sciences, Indiana University, Bloomington, IN. [12]IPTAI NASA Astrobiology Institute, Bloomington, IN. [13]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA. [14]Department of Bioengineering, University of California, Berkeley, CA. [15]Department of Geosciences, Princeton University, Princeton, NJ.

*To whom correspondence should be addressed:
Dr. Dylan Chivian
Lawrence Berkeley National Laboratory
1 Cyclotron Road, MS Calvin
Berkeley, CA 94720 USA
E-mail: DCChivian@lbl.gov

TABLE OF CONTENTS **PAGE**

I. TAXONOMIC INFORMATION

Inspiration for the name *Candidatus Desulforudis audaxviator.*

"*In Sneffels Joculis craterem quem delibat Umbra Scartaris Julii intra calendas descende, audax viator, et terrestre centrum attinges.*" ("Descend, bold traveler, into the crater of the jokul of Sneffels, which the shadow of Scartaris touches before the kalends of July, and you will attain the center of the earth.")

-- Hidden message deciphered from an Icelandic saga that prompts Professor Lidenbrock to undertake his journey in Jules Verne's "Journey to the Center of the Earth".

Based on its rod-like morphology, its apparent use of the dissimilatory sulfate reduction pathway for energy production, and because of the journey this "*audax viator*" (bold traveler) undertook to live in the extreme depths of the Earth, we have named this organism "*Candidatus Desulforudis audaxviator*". Additionally, as a consensus sequence from a fracture accessed from the 104th level of the Mponeng mine, we have given the genome the strain designation "MP104C".

Taxonomic record for *Candidatus* classification.

Candidatus Desulforudis audaxviator MP104C has been given the NCBI taxonomy ID 477974 and placed in the lineage "cellular organisms; Bacteria; Firmicutes; Clostridia; Clostridiales; Peptococcaceae; Candidatus Desulforudis; Candidatus Desulforudis audaxviator; Candidatus Desulforudis audaxviator MP104C". In accordance with the guidelines of Murray and Stackbrandt (*1*) for the *Candidatus* designation, we offer the following codified taxonomic record for *Candidatus Desulforudis audaxviator* MP104C.

"*Candidatus Desulforudis audaxviator* MP104C" [(*Firmicutes*) NC; G+; R; NAS (GenBank CP000860), oligonucleotide sequence complementary to unique region of 16S rRNA 5'-GCGGGATTTCACCTGCGACTTCTCA-3'; FL (deep subsurface crustal fracture); Anaer., sulfate reducing; T]. Chivian et al., Science [*PUBLICATION INFORMATION TO BE DETERMINED*], 2008.

II. BACKGROUND

Isolation of deep subsurface organisms in South Africa.

South African mines have provided access to microorganism-bearing fluids that emanate from fractures at depths ranging from 0.7 km to 5 km (*2, 3*). Phylogenetic classification of the indigenous microbial species using small subunit (SSU or 16S) rRNA gene analyses of DNA from environmental samples has revealed new genera, families, orders, and in some cases, new candidate phyla of *Archaea* and *Bacteria* (*4, 5*). Of the approximately 280 bacterial and 44 archaeal operational taxomic units (OTUs) identified to date in the South African mines, only 12 mesophilic and thermophilic anaerobic bacteria and one autotrophic methanogen have so far been isolated (*6-10*). Of the bacterial isolates only one belongs to the *Firmicutes* phylum. *Desulforudis audaxviator* has not yet been isolated, which may be due to its extreme sensitivity to O_2 (Table S25).

Desulforudis audaxviator has been prevalent in the 16S rRNA gene clone libraries of thermophilic, sulfidic, moderately saline, alkaline boreholes at Beatrix, Evander, Driefontein, Kloof, and Mponeng Au mines and is the only organism this widely distributed in the Witwatersrand Basin at depths greater than 1.5 km. *D. audaxviator* is found in the deepest and hottest fracture waters to date. The highest temperature determined was based on the hydrogen isotope equilibrium temperature between H_2O and dissolved H_2. During the course of dewatering fracture zones, these temperature estimates and the measured temperatures will change as different depths of the fracture zone contribute water to the borehole. In the case of MP104 the temperature decreased from 62°C to 52°C which, when combined with local heat flow and thermal conductivity data (*11*), suggest that this fracture network extends from 4.2 km to 2.8 km below land surface (kmbls), the latter depth being that of level 104. The fracture water represents a mixture of ~3 million year old paleometeoric water with 0.8-2.5 billion year old, saline, reduced-gas-rich hydrothermal fluid (*3*). H_2 and SO_4^{2-} concentrations tended to be greater in these deeper fractures. Experimental data and theoretical analyses indicate that radiolysis of water directly supplies the H_2 (*12*) and indirectly supplies the SO_4^{2-} by producing H_2O_2 that in turn oxidizes the abundant pyrite in the Witwatersrand quartzite (*13*). Retention of rubrerythrin (Table S25) in the genome of *D. audaxviator* is consistent with recurring exposure to the products of radiolysis.

History of the South African crust.

Unlike surface habitats that permit comparatively instantaneous access, species found in the deep subsurface require fitness throughout the history of their colonization, which in the Witwatersrand basin includes temperatures greater than 60°C, nutrient flux

on the order of 10^{-9} moles $cell^{-1}$ yr^{-1} and pH values ranging from 8.5 to 9.5. The Witwatersrand basin formed between 2.9 to 2.5 Ga and at 2.0 Ga, during the formation of the Vredefort impact structure, it may have had 7 to 10 km more sediment on top than the present day and experienced a peak metamorphic temperature of ~250-300^{o}C. The basin was quiescent until 1.4 Ga dyke swarms from the Pilanesberg alkaline complex to the north of the basin compartmentalized the hydrological structure of the aquifers within the Witwatersrand basin. The 7 to 10 km of overburden was gone by the Permo-Carboniferous glacial period at 280 Ma, because the present day surface outcrops of the nearby Vredefort impact structure reveal signs of glacial scouring. During the Karoo volcanic episode at 200 Ma, however, an additional 2 km of volcanic and sedimentary overburden may have been deposited on top of the Witwatersrand basin.

Fission track apatite thermochronological analyses have revealed that the temperature was 120^{o}C at a depth of 3.7 km in Driefontein mine at 75 Ma and cooled to the present day temperatures at a rate of 1.4^{o}C Myr^{-1} (*11*) as this overburden was removed by uplift and erosion prior to 40 Ma. The South African crust has therefore been moving up and down, heating up and cooling off for billions of years. The fractures tend to seal with burial and open with uplift as lithostatic pressure decreases. Therefore, the period of time between 100 and 40 Ma is probably the most recent time when fluid flow occured into the deeper portions of the crust (*11*). This may date the time of *D. audaxviator's* latest journey into the earth.

Environmental sources of energy and material.

Energy and material for the ecosystem (as shown in Figure 3) comes from the radiolytic production of H_2 and reactive H_2O_2, which in turn reacts with H_2S to produce SO_4^{2-} or with pyrite (FeS_2) to produce SO_4^{2-} and $Fe(OH)_3$ as detailed by Lin, *et al*. (*3*), and shown experimentally by Lefticariu, *et al.* (*13*). The H^+ produced by the cell and released by oxidation reactions dissolves calcite ($CaCO_3$) releasing Ca^{2+} and bicarbonate (HCO_3^-). The Ca^{2+} in turn may exchange with NH_4^+ in chlorite mineral. The HCO_3^- can either be taken up by the putative Na^+/HCO_3^- symporter or it may be radiolytically reduced to formate (HCO_2^-). All three forms of inorganic carbon may be utilized by the Acetyl-CoA carbon fixation pathway, as well as CO. The H_2S produced by the SO_4^2 -reduction pathway can diffuse out of the cell and, in addition to reacting with H_2O_2 to replish SO_4^{2-}, can react with the $Fe(OH)_3$ to regenerate SO_4^{2-} and release PO_4^{3-}. The Fe^{2+} released by this last reaction can combine with H_2S to precipitate FeS or FeS_2.

III. METHODS

Collection of DNA.

Fracture fluid was collected over 3 days (9/27/02-9/30/02) from a borehole located at level 104 (2.8 km below land surface, 1.2 km below sea level) of Mponeng gold mine (26°26'S; 27°26'E), owned and operated by AngloGold, PTY. A Cole Parmer, 0.2 μm effective pore size, double open end, high efficiency, pleated PTFE filter cartridge (http://www.coleparmer.com – EW-06479-52), 8 cm in diameter and 25 cm long was installed on a flowing borehole 15 days after initial intersection of the fracture using an autoclaved expansion packer placed downstream from a large steel ball valve installed by mine contractors. The density of planktonic cells in the fracture fluid, as determined by flow cytometry, was ~3.3×10^4 cells mL^{-1} and ~5.6×10^6 mL of water passed through the filter, yielding a capture of ~1.8×10^{11} cells. The filter consisted of a pleated filter that wrapped around a hard plastic core, but was not actually attached to it, and held in place by a hard plastic outer case with radial slits and hard plastic end caps. Prior to removal, the cartridge was drained of fluid in the mine, removed from its stainless steel canister and carefully wrapped in multiple thicknesses of sterile plastic, placed in a cooler with dry ice and transported to the surface. The cartridge was stored for a couple weeks at -20°C in the field laboratory then transported to Princeton University on dry ice and stored at -80°C until being shipped to Pacific Northwest National Laboratory on dry ice for DNA extraction.

High molecular weight community DNA was extracted using a rigorous protocol developed for hard-to-lyse Gram-positive bacteria and archaea. The outer plastic case was cut off and the pleated filter removed from the core while it was still frozen, and the pleated filter returned to the freezer. The pleated filter was comprised of 5 layers, the inside (upstream side) stiff net-like layer, a relatively thick pre-filter layer, two filter layers and another net-like layer on the outside. Separating the filter layers from the structural layers of the cartridge filter before carrying out the extraction was required to successfully extract DNA. The first and second filter layers were extracted separately and pooled at the end of the extraction process. For each extraction, the top two filter layers from 150 or 200 cm^2 of the filter were cut into ~1 cm^2 pieces with sterile scissors and placed in 50 mL disposable tubes held in liquid nitrogen. Ten mLs of Bactozyme solution, (cat. no. BZ 160, Molecular Research Center, Inc., Cincinnati, OH 45212) was added to each tube. The filter pieces were wetted by vacuum infiltration and incubated at 50°C for 30 minutes. One mL of a 10% (w/v) SDS solution was added to each tube and 6 rapid freeze/thaw cycles with liquid N_2 and a 50°C water bath were performed. Two hundred μL of Proteinase K (10 mg/mL) was added to each tube and incubated at 50°C for 2 hours. Forty mLs of DNAzol (*14*) (cat. no. DN 127, Molecular Research Center, Inc., Cincinnati, OH 45212) was added to each tube and incubated at 42°C overnight. The supernatant was separated from the filter pieces and particulates by centrifuging at 10,000 x g for 15 minutes. One mL aliquots of the clear supernatant were transferred into 1.7 mL microcentrifuge tubes and the DNA precipitated by adding 600 μL of 100% ethanol and incubating at 4°C overnight.

The DNA was pelleted by centrifugation at 17,000 x g for 30 min and washed with 1 mL 70% ethanol per tube. The DNA was resuspended with 25 µL of sterile water per tube and pooled into one 1.7 mL microcentrifuge tube. The DNA concentration was spectrophotometrically determined by measuring absorbance at 260 nm using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA), and the integrity of the DNA was verified on a 0.6% TBE agarose gel. In 4 extractions, a total of 82 micrograms of DNA was recovered from 650 square centimeters of filter, of which 46 micrograms were high molecular weight (HMW) DNA. DNA was extracted as follows: 11/16/04 extraction: 17 micrograms HMW DNA (249 ng/cm^2); 11/6/2005 and 11/8/2005 extractions: 17 micrograms HMW DNA (70 ng/cm^2 and 93 ng/cm^2 respectively); 4/19/2006 extraction: 12 micrograms HMW DNA (94 ng/cm^2).

Sequencing and assembly.

Sequencing and assembly was done by the DOE Joint Genome Institute (JGI). The high molecular weight DNA extract was used to construct two genomic libraries (~3 kb pUC18 vector and ~8 kb pMCL200 vector) (http://www.jgi.doe.gov/). Double-ended sequencing reactions were carried out using both ET and BigDye terminator chemistry (Perkin Elmer) and resolved using both MegaBase and ABI PRISM 3730 (Applied Biosystems) capillary DNA sequencer. Sanger sequencing (*15*) yielded 31,218 reads of average nominal length 1036 bp for a total of 32.3 Mb (including 29,198 reads with at least 10 contiguous calls with a Phred score ≥ 25 yielding 19.2 Mb of high quality calls). Vector and quality trimming of shotgun data was performed yielding 29,279 reads for a total of 20.7 Mb (average trimmed read length of 708 bp). During the finishing process paired reads information was used to scaffold contigs. Because of the small amount of DNA available, uncaptured gaps between scaffolds were closed using 454 pyrosequencing (*16*) data (750 bp overlapping pseudoreads that are chopped from Newbler (*16*) contigs were assembled together with the Sanger reads) which yielded 56.2 Mb (518,272 reads with an average length of 109 bp). Gap-spanning 454 stretches were confirmed by Sanger sequencing of PCR products performed on source DNA. The reads were assembled using Phrap version SPS-3.57 (*17, 18*) (http://www.phrap.org/), yielding one complete, closed chromosome of length 2,349,476 bp. The assembled genome contained 27900 shotgun Sanger reads and 267 finishing reads. This is the first case when the combination of Sanger and pyrosequencing was applied to the metagenomic assembly finishing. The genome sequence reported in this study has been deposited in GenBank under accession number CP000860. The metagenomic data is available from the Joint Genome Institute (http://www.jgi.doe.gov/) under project number 4000602.

Genome annotation.

We identified and classified the protein and RNA genes using the MicrobesOnline (*19*) annotation pipeline (http://www.microbesonline.org). Protein-coding genes were identified using CRITICA (*20*) and supplemented with non-overlapping high-scoring hits from Glimmer (*21*), and translated into protein sequences assuming the standard microbial genetic code. Additional RNAs were identified using tRNAscan-SE (*22*) and BLASTn (*23*). For each protein-coding gene, we used a comprehensive set of sequence databases to identify conserved domain structure and to provide addition sources of annotations such as Enzyme Commision (EC) numbers, GO terms (*24*), Pfam (*25*) and TIGRfam (*26*) protein sequence family assignments, and membership in COGs (Clusters of Orthologous Groups of proteins) (*27*). Comparison with orthologous sequences (identified as bidirectional best BLASTp hit covering at least 75%) from multiple microbes enables the prediction of operons and regulons (*28*) and allows for viewing the genomic context of a given gene in multiple organisms simultaneously using a tree-based genome browser (http://www.microbesonline.org/treebrowseHelp.html). We applied the operon/regulon predictions and tree-based genome browser extensively in manually curating the annotations of key genes. Genes were subsequently mapped to calls made by the ORNL pipeline, with gene names of the form "DaudXXXX". The annotated *D. audaxviator* genome is accessible *via* MicrobesOnline (http://www.microbesonline.org).

Collection and preparation of samples for microscopy.

Microscopy sample #1 (date: 09/16/02): collected into a 120 ml serum vial. The serum vial was flushed with N_2 gas and autoclaved prior to the field trip. The vial was transported back to the field lab in South Africa within 3 hours and stored in a 4^{o}C refrigerator. Samples were then transported back to USA on blue ice packs, and stored in a 4^{o}C refrigerator. Nothing else was added to the serum vial.

Microscopy sample #2 (date: 11/09/02): collected in sterile 140 mL serum vials, precapped with blue butyl stoppers (Bellco) and preflushed with filtered, industrial grade Argon. Unconcentrated samples were introduced into the vials via 20 Ga syringe needles hooked directly to the flowing Masterflex norprene hose (sterile) off the octopus sampler. Additional concentrated samples were taken off the same flowing sample lines using mediakap filters (0.2 micron). About 2 L was pushed through each of the mediakap filters follwed by backflushing ~60 mL of sample water into waiting small serum vials. All samples were stored 4^{o}C refrigerators at the field lab in South Africa, then at PNNL, then at DRI.

DAPI staining: 1ml of sample #2 was stained w/ 100µl DAPI (3µg/ml) for 10 minutes in the dark. Stained samples were filtered (Poretics, polycarbonate, black, 0.22µm pore, 25mm; Osmonics, Inc) and viewed using 100x-oil emersion lens and epifluorescent microscopy with appropriate filters.

Scanning electron microscopy (SEM): both sample #1 and sample #2 were filtered though 0.4µm Isopore membrane filters (millipore) then processed through an ethanol dehydration series (25, 50, 75, and 100% v/v ethanol) with each treatment lasting 30 min. The samples were then critically point dried in a SamDri® Critical Point Drier (Tesumis Inc.) to preserve the structure of the cells. The filter papers were mounted on aluminum stubs with carbon adhesive tabs, coated with palladium-gold alloy to reduce charging artifacts and imaged at 5 kV using a LEO 1540XB Field Emission SEM.

CARD-FISH protocol: Catalyzed-reporter deposition fluorescence *in situ* hybridization (CARD-FISH) was performed. A 25 bp probe for *Candidatus Desulforudis audaxviator* was designed using the software package ARB (*29*) according to recommendations by Hugenholtz, *et al*. (*30*). The probe was checked for homology to all sequences available in the Greengenes database (*31*) as of March 2008. The probe was synthesized and 5' labeled with Horse-Radish Peroxidase (Invitrogen, CA).

Probe name	**Probe sequence**	**Bases**	**Modification**
DLO1_HRP	GCG GGA TTT CAC CTG CGA CTT CTC A	25	5' Horse-Radish Peroxidase

CARD-FISH was performed essentially as described by Sekar *et al.* (*32*). Samples were fixed by addition of 0.2 µm filtered 96% ethanol to a final concentration of 50% (v/v). Fixed samples were filtered through 0.2µm black polycarbonate filters that were cut into sections using a sterile scalpel. Filters sections were air dried, dipped into 0.2% (w/v) low-melting-point agarose and placed on glass slides and air dried at 35°C for 10 min. Filter sections were then dehydrated in 96% ethanol for 1 min and air dried. For cell permeabilization, agarose embedded filter sections were incubate in lysozyme (10 mg/ml) at 37°C for 60 min and achromopeptidase (60 U/ml) at 37°C for 30 min. Sections were then incubated in 0.01 M HCl for 10 min at RT to inactivate endogenous peroxidases (to avoid false positive signals due to non-specific tyramide deposition) before washing with mobio grade water (0.2 µm filtered, autoclaved, DEPC treated) and 0.2 µm filtered 96% ethanol. Filter sections were placed on glass slides and 400 µl of hybridization buffer (containing 20% formamide and 0.5 ng probe DLO1_HRP μl^{-1}). Slides were incubated in sealed Petri dishes overnight at 35°C. Filter sections were washed in prewarmed (37°C) washing buffer. Filter sections were then incubated in 1 x PBS amended with 0.05% of Triton X-100 followed by incubation in substrate mix (1 parts of CY3-labeled tyramide and 100 parts of amplification buffer [1 x PBS, 0.0015% H_2O_2, 0.1% blocking reagent (PBS + 1% BSA]) at 37°C for 10 min in the dark. Filter sections were then washed in 1x PBS amended with 0.05% Triton X-100 and then with mobio grade water followed by 96% ethanol. Filter sections were then mounted with VECTASHIELD HardSet Mounting Medium with DAPI (Vector Laboratories, CA). Epifluorescence images were

taken using filters for DAPI and CY3 spectra using a Leica DMRX microscope.

16S rRNA gene amplification for PhyloChip and clone library analysis.

The 16S rRNA gene was amplified from gDNA extracts using modified (degeneracies removed) universal primers 27F (5' AGAGTTTGATCCTGGCTCAG) and 1492R (5' GGTTACCTTGTTACGACTT) for bacteria and 4Fa (5' TCCGGTTGATCCTGCCRG 3') combined with 1492R for archaea. Each PCR reaction mix contained: 1X Ex Taq buffer, 0.8mM dNTP mixture, 0.02U/μL Ex Taq polymerase (TaKaRa Bio Inc, Japan), 0.4mg/mL bovine serum albumin (BSA), and 300nM each primer and 36ng gDNA. PCR conditions were as follows: 1 cycle of 3 min at 95°C, followed by 25 cycles (35 for Archaea) of 30 sec at 95°C, 30 sec at annealing temperature (gradient of 8 temperatures between 48-58°C), and 1 min at 72°C, with a final extension for 7 min at 72°C. PCR products from the eight different annealing temperatures were combined, concentrated by precipitation and resuspended in DEPC treated water. Lack of a visible band following gel electrophoresis suggested archaea were absent or in low numbers.

16S rRNA amplicon analysis by clone library sequencing.

Bacterial 16S rRNA amplicon pools amplified as for PhyloChip analysis were ligated to pCR4-TOPO vectors (Invitrogen, CA), using an insert to vector ratio of 3:1 to maximize diversity of amplicons recovered. Ligated plasmids were transformed into *E. coli* TOP10 chemically competent cells according to the manufacturer's recommended protocol (Invitrogen, CA). Three hundred eighty four clones were randomly selected by a robotic picker and inserts were sequenced bi-directionally using M13 vector specific primers. Sequences were primer and vector screened using cross_match, quality scored using Phred and assembled into contigs using Phrap (*17, 18*). Sequences were trimmed to retain only bases Phred ≥q20 and high quality contigs were tested for chimeras (one of which was removed from further analysis) using Bellerophon version 3 (http://greengenes.lbl.gov/cgi-bin/nph-bel3_interface.cgi).

16S rRNA amplicon analysis by PhyloChip hybridization.

PhyloChip analysis was essentially as described previously (*33-35*). Results are given in Table S5. For bacteria, 780 ng of 16S rRNA gene amplicons were spiked with internal controls consisting of synthetic 16S rRNA gene fragments and non-16S rRNA gene

fragments. Despite the lack of visible PCR amplicons from archaeal reactions an aliquot from those combined reactions was also included in the amplicon mix to be analyzed by PhyloChip. This mix was fragmented, to a size range of 50-200 bp in length using DNAse I (0.02 U/µg DNA, Invitrogen, CA, USA) in One-Phor All buffer (Amersham, NJ, USA) according to Affymetrix's standard protocol, with incubation at 25˚C for 10 min, followed by enzyme denaturation at 98˚C for 10 min. Biotin labeling was performed using an Affymetrix Gene Labeling Reagent and terminal deoxynucleotidyl transferase (Promega, WI, USA) according to Affymetrix technical expression manual (http://www.affymetrix.com/support/technical/manual/expression_manual.affx). The labeled DNA was then denatured (99˚C for 5 min) and hybridized to the 'PhyloChip' DNA microarray in 100 mM MES (morpholineethanesulfonic acid) buffer, pH 6.6, containing 1 M NaCl, 20 mM EDTA, 0.01% Tween 20, 100 µg of herring sperm DNA/ml, 500 µg of bovine serum albumin (BSA)/ml, and 0.5 nM control biotin-oligonucleotide B3. Arrays were hybridized at 48˚C overnight (> 16 hr) at 60 rpm and washed and stained according to the Affymetrix technical expression manual.

Arrays were scanned using a GeneArray Scanner (Affymetrix, CA, USA). The scan was recorded as a pixel image and analyzed using standard Affymetrix software (Microarray Analysis Suite, version 5.1) that reduces the data to an individual signal value for each probe. Background probes were identified as those producing intensities in the lowest 2% of all intensities. The average intensity of the background probes was subtracted from the fluorescence intensity of all probes. The noise value (N) was considered the variation in pixel intensity signals observed by the scanner as it read the array surface. The standard deviation of the pixel intensities within each of the identified background cells was divided by the square root of the number of pixels comprising that cell. The average of the resulting quotients was then used for N in the calculations described below.

Probe pairs scored as positive were those that met two criteria: (i) the intensity of fluorescence from the perfectly matched probe (PM) was greater than 1.3 times the intensity from the mismatched control (MM), and (ii) the difference in intensity, PM minus MM, was at least 130 times greater than the squared noise value (>130 N^2). The positive fraction (PosFrac) was calculated for each probe set as the number of positive probe pairs divided by the total number of probe pairs in a probe set. An OTU was considered present in the sample when over 90% of its assigned probe pairs are positive (PosFrac > 0.90). Hybridization intensity (referred to as intensity) was calculated in arbitrary units (a.u.) for each probe set as the trimmed average (maximum and minimum values removed before averaging) of the PM minus MM intensity differences across the probe pairs in a given probe set.

Sequence analysis of 16S rRNA gene libraries and comparison with PhyloChip data.

Sequences were aligned to the Greengenes 7,682-character format using the NAST web-server (http://greengenes.lbl.gov/NAST) (*31, 36*). Similarity to public database records was calculated with DNADIST (*37*) using the DNAML-F84 option assuming a transition:transversion ratio of 2.0 and an A, C, G, T 16S rRNA gene base frequency of 0.2537, 0.2317, 0.3167, 0.1979, respectively. This was calculated empirically from all records of the Greengenes 16S rRNA gene multiple sequence alignment over 1,250 nucleotides in length. The Lane mask (*38*) was used to restrict similarity observations to 1,287 conserved columns (lanes) of aligned characters. Three cloned sequences from this study were rejected from further analysis when <1,000 characters could be compared to a lane-masked reference sequence. Sequences were assigned to a taxonomic node using a sliding scale of similarity thresholds (*39*). Phylum, class, order, family, sub-family, or OTU placement was accepted when a clone surpassed similarity thresholds of 80%, 85%, 90%, 92%, 94%, or 97%, respectively. For example, when similarity to nearest database sequence was <94%, the clone was considered to represent a novel sub-family and a novel class was denoted when similarity was <85%. Diversity estimates (Shannon-Weaver index (*40*) and the non-parametric richness estimator Chao1 (*41*)) were calculated using the software DOTUR (*42*) with the clone distance matrix as input and a furthest-neighbor clustering algorithm. Dominance in clone libraries was calculated as 1-Shannon evenness index (1-E) where evenness (E) is represented as follows: E = H/lnS, where H = Shannon-Weaver diversity index and S is the total richness in a sample. Results are given in Table S5.

Reducing the impact of the dominant species on assessment of 16S rRNA gene sequence diversity.

PhyloChip microarray data indicated that other bacterial species besides *Candidatus Desulforudis audaxviator* were present in the gDNA extracts. However, the initial SGNY clone library analysis showed little evidence for this (Fig. S3). We hypothesized that the extreme dominance of *Candidatus Desulforudis audaxviator* in this system made detection of less abundant species by clone library or shotgun metagenomics problematic without a significant sequencing effort. To overcome this obstacle we succeeded (Fig. S3) in reducing the dominance of the *D. audaxviator* template in the PCR reaction by selective restriction enzyme digestion. Using the data obtained from the PhyloChip and previous studies of this fracture water system (*3*) we identified the other possible templates in the gDNA extract and selected a restriction enzyme (*Sal*I) that would digest the *D. audaxviator* 16S rRNA gene making it unavailable for amplification, while minimizing digestion of other less abundant 16S rRNA gene templates (an online tool, 'Seq and Destroy' was written for this purpose and can be accessed at http://greengenes.lbl.gov/cgi-bin/nph-seq_and_destroy.cgi). gDNA was pre-digested with 20U *Sal*I and 36ng of digested DNA was added to PCR reactions which were carried out as for the intact gDNA 16S rRNA gene libraries. Aliquots from the pooled products of these PCR reactions were ligated, transformed and sequenced as described above.

Sequences were also vector screened, quality checked, assembled, trimmed and chimera screened as described for the intact gDNA. The SGNY and SGNX library results are given in Figure S3, in particular the phylogenetic tree of Figure S3d. Comparison with the phylogenetic microarray results is given in Figure S5b. The clone library sequences have been submitted to GenBank with accession numbers EU730965 - EU731008.

IV. FIGURES AND TABLES

Table S1. Abbreviations used in tables.

Column headings are as follows:
Gene: the locus id.
Name: the gene name.
Description: functional assignment of the gene, usually taken from a protein family, or sometimes from a homologous gene in another organism if membership in a protein family is not confident for the *D. audaxviator* gene (likely as a result of the undersampling of the protein family). The following protein sequence families are used: "COG": clusters of orthologous groups (*27*), "PFAM" or "PF": protein families (*25*), "TIGRFAM", "TIGR", or "TF": TIGR protein families (*26*), "SM": SMART protein families (*43*), and "SSF": SUPERFAMILY protein families (*44*).
Len: the length of the gene, in amino acids for protein-coding genes, and in base pairs for non-protein-coding genes (including pseudogenes)
CH id: the amino acid identity of the closest homolog in another species, or "N/A" if no homolog is found.
CH species: the species name of the closest homolog in another species (usually abbreviated according to the table below), or "ORFan" if no homolog is found. At the time of most of these analyses, we did not have the complete genome sequence for *Pelotomaculum thermopropionicum* SI nor *Desulfotomaculum reducens* MI-1. We also did not have any genomic sequence for the other relatives *Syntrophomonas wolfei* subsp. wolfei str. Goettingen (with the exception of the analysis of the signal transduction genes Table S19), *Heliobacterium modesticaldum* Ice1, *Thermosinus carboxydivorans* Nor1, and *Clostridium novyi* NT.
Notes: notes pertinent to the gene. Some of the abbreviations used include: "ds": downstream, "us": upstream, "hh": hitchhiking (meaning present in operon primarily providing different functionality), "annot.": source organism from which annotation was taken.

Additionally, species names have been abbreviated as follows:

Archaea	
A. pernix	*Aeropyrum pernix* K1
A fulgidus	*Archaeoglobus fulgidus* DSM 4304
Halo. NRC-1	*Halobacterium sp. NRC-1*
M. maripaludis	*Methanococcus maripaludis*
M. jannaschii	*Methanocaldococcus jannaschii* DSM 2661
M. kandleri	*Methanopyrus kandleri* AV19
M. acetivorans	*Methanosarcina acetivorans* C2A
M. barkeri	*Methanosarcina barkeri* str. fusaro
M. mazei	*Methanosarcina mazei* Goe1
M. hungatei	*Methanospirillum hungatei* JF-1
M. stadtmanae	*Methanosphaera stadtmanae* DSM 3091

Archaea	
M. thermautotrophicus	*Methanothermobacter thermautotrophicus* Δ*H*
N. pharaonis	*Natronomonas pharaonis* DSM 2160
P. aerophilum	*Pyrobaculum aerophilum* str. IM2
P. abyssi	*Pyrococcus abyssi GE5*
P. furiosus	*Pyrococcus furiosus* DSM 3638
S. solfataricus	*Sulfolobus solfataricus* P2
S. tokodaii	*Sulfolobus tokodaii* str. 7
T. kodakaraensis	*Thermococcus kodakaraensis* KOD1
T. acidophilum	*Thermoplasma acidophilum* DSM 1728
T. volcanium	*Thermoplasma volcanium* GSS1

Bacteria	
A. tumefaciens	*Agrobacterium tumefaciens* str. C58 (Cereon)
A. variabilis	*Anabaena variabilis* ATCC 29413
H. marismortui	*Haloarcula marismortui* ATCC 43049
A. dehalogenans	*Anaeromyxobacter dehalogenans* 2CP-C
A. aeolicus	*Aquifex aeolicus* VF5
A. vinelandii	*Azotobacter vinelandii* AvOP
B. anthracis Sterne	*Bacillus anthracis* str. Sterne
B. cereus	*Bacillus cereus* ZK
B. clausii	*Bacillus clausii* KSM-K16
B. halodurans	*Bacillus halodurans* C-125
B. licheniformis	*Bacillus licheniformis* DSM 13
B. subtilis	*Bacillus subtilis* subsp. subtilis
B. thuringiensis	*Bacillus thuringiensis* serovar konkukian str. 97-27
B. japonicum	*Bradyrhizobium japonicum* USDA 110
B. pseudomallei	*Burkholderia pseudomallei* K96243
B. xenovorans	*Burkholderia xenovorans* LB400

Bacteria	
L. sakei	*Lactobacillus sakei* subsp. sakei 23K
Leptospira interrogans	*Leptospira interrogans* L1-130
M. magneticum	*Magnetospirillum magneticum* AMB-1
M. succiniciproducens	*Mannheimia succiniciproducens* MBEL55E
M. aqueolei	*Marinobacter aqueolei*
M. thermoacetica	*Moorella thermoacetica* ATCC 39073 (Previously named *Clostridium thermoaceticum*)
M. avium	*Mycobacterium avium* K10
M. bovis	*Mycobacterium bovis* AF2122/97
N. winogradskyi	*Nitrobacter winogradskyi* Nb-255
N. oceani	*Nitrosococcus oceani* ATCC 19707
N. farcinica	*Nocardia farcinica* IFM 10152
N. punctiforme	*Nostoc punctiforme* PCC 73102
Nos. sp. PCC 7120	*Nostoc* sp. PCC 7120
O. iheyensis	*Oceanobacillus iheyensis* HTE831
P. carbinolicus	*Pelobacter carbinolicus* str. DSM 2380
P. luteolum	*Pelodictyon luteolum* DSM 273

C. hydrogenoformans	*Carboxydothermus hydrogenoformans* Z-2901	P. thermopropionicum	*Pelotomaculum thermopropionicum* SI
C. muridarum	*Chlamydia muridarum* Nigg	Pir. sp. 1	*Pirellula* sp. 1
C. chlorochromatii	*Chlorobium chlorochromatii* CaD3	P. gingivalis	*Porphyromonas gingivalis* W83
C. tepidum	*Chlorobium tepidum* TLS	P. haloplanktis	*Pseudoalteromonas haloplanktis* TAC125
C. acetobutylicum	*Clostridium acetobutylicum* ATCC 824	R. eutropha	*Ralstonia eutropha* JMP134
C. perfringens	*Clostridium perfringens*	R. etli	*Rhizobium etli* CFN 42
C. tetani	*Clostridium tetani* E88	R. palustris	*Rhodopseudomonas palustris* HaA2
C. glutamicum	*Corynebacterium glutamicum* ATCC 13032	R. rubrum	*Rhodospirillum rubrum* ATCC 11170
D. aromatica	*Dechloromonas aromatica* RCB	R. albus	*Ruminococcus albus*
D. ethenogenes	*Dehalococcoides ethenogenes* 195	S. ruber	*Salinibacter ruber* DSM 13855
Dehalo. sp. CBDB1	*Dehalococcoides* sp. CBDB1	S. baltica	*Shewanella baltica* OS155
D. geothermalis	*Deinococcus geothermalis* DSM 11300	S. frigidimarina	*Shewanella frigidimarina* NCIMB 400
D. hafniense DCB2	*Desulfitobacterium hafniense* DCB-2	Silic. sp. TM1040	*Silicibacter* sp. TM1040
D. hafniense Y51	*Desulfitobacterium hafniense* Y51	S. avermitilis	*Streptomyces avermitilis* MA-4680
D. audaxviator	*Desulforudis audaxviator* MP104C	S. thermophilum	*Symbiobacterium thermophilum* IAM 14863
D. psychrophila	*Desulfotalea psychrophila* LSv54	Syn. sp. JA-2	*Synechococcus* sp. JA-2-3B'a(2-13)
D. reducens	*Desulfotomaculum reducens* MI-1	Syn. sp. JA-3	*Synechococcus* sp. JA-3-3Ab
D. desufuricans G20	*Desulfovibrio desulfuricans* G20	Syn. sp. PCC 6803	*Synechocystis* sp. PCC 6803
D. vulgaris DP4	*Desulfovibrio vulgaris* DP4	S. wolfei	*Syntrophomonas wolfei* subsp. wolfei str. Goettingen
DvH	*Desulfovibrio vulgaris* Hildenborough	T. tengcongensis	*Thermoanaerobacter tengcongensis* MB4T
D. vulgaris HB	*Desulfovibrio vulgaris* Hildenborough	T. elongatus	*Thermosynechococcus elongatus* BP-1
D-monas spp.	*Desulfuromonas* spp.	T. maritima	*Thermotoga maritima* MSB8
E. faecalis	*Enterococcus faecalis* V583	T. thermophilus HB8	*Thermus thermophilus* HB8
E. coli K12	*Escherichia coli* K12	T. thermophilus HB27	*Thermus thermophilus* HB27
F. tularensis	*Francisella tularensis* subsp. Tularensis SCHU S4	T. denitrificans	*Thiobacillus denitrificans* ATCC 25259
G. kaustophilus	*Geobacillus kaustophilus* HTA426	T. denticola	*Treponema denticola* ATCC 35405
G. metallireducens	*Geobacter metallireducens* GS-15	V. splendidus	*Vibrio splendidus* 12B01
G. sulfurreducens	*Geobacter sulfurreducens* PCA	V. vulnificus	*Vibrio vulnificus* CMCP6
Jann. sp. CCS1	*Jannaschia* sp. CCS1	W. succinogenes	*Wolinella succinogenes* DSM 1740
K. pneumoniae	*Klebsiella pneumoniae*	X. campestris	*Xanthomonas campestris* pv. campestris str. 8004

Table S2. Range of geochemical parameters for *D. audaxviator* bearing fracture water samples.

Table S2 summarizes the range of geochemical parameters recorded at four boreholes where *D. audaxviator* is found (conditions specific to MP104 may be found in (*3*)). SO_4^{2-} is the dominant electron acceptor followed by inorganic carbon. The most abundant electron donor is CH_4 followed by H_2, C_2H_6, C_3H_8, acetate, CO, n-C_4, formate, iso-C_4, and propanoate. Concentrations are in Molar units. CO concentrations are highest at EV818 Hole 6 (Evander mine, 1.8-2 kmbls in quartzite) and DR 546 Hole 1 (Dreifontein mine, 3.3 kmbls in metavolanic rock).

	Average	Minimum	Maximum
Depth (kmbls.)	2.800	1.300	>3.35
pH	8.6	7.3	9.3
pe	-1.94	-3.10	-1.13
T^{o}C	48.2	39.0	61.7
TOC	$2.5X10^{-4}$	$1.1X10^{-4}$	$4.3X10^{-4}$
DOC	$2.3X10^{-4}$	$1.4X10^{-4}$	$4.0X10^{-4}$
acetate	$3.3X10^{-5}$	$7.3X10^{-6}$	$8.8X10^{-5}$
formate	$2.5X10^{-6}$	$1.9X10^{-7}$	$7.6X10^{-6}$
propanoate	$4.0X10^{-7}$	$1.2X10^{-7}$	$2.7X10^{-6}$
CH_4	$1.0X10^{-2}$	$1.5X10^{-3}$	$1.7X10^{-2}$
C_2H_6	$6.1X10^{-4}$	$8.0X10^{-6}$	$1.3X10^{-3}$
C_3H_8	$7.2X10^{-5}$	$7.8X10^{-7}$	$1.9X10^{-4}$
iso-C_4	$5.5X10^{-7}$	$1.0X10^{-8}$	$1.4X10^{-6}$
n-C_4	$1.1X10^{-5}$	$1.0X10^{-7}$	$3.0X10^{-5}$
CO	$1.6X10^{-5}$	$6.6X10^{-8}$	$1.2X10^{-4}$
H_2	$9.6X10^{-4}$	$2.9X10^{-7}$	$3.7X10^{-3}$
DIC	$1.3X10^{-4}$	$7.3X10^{-6}$	$3.7X10^{-4}$
HS^-	$8.4x10^{-4}$	$3.1x10^{-6}$	$1.5x10^{-3}$
SO_4^{2-}	$5.8x10^{-4}$	$7.7x10^{-6}$	$3.1x10^{-3}$

$S_2O_3^{2-}$	8.2×10^{-7}	1.8×10^{-8}	6.0×10^{-6}
O_2	$<3.1 \times 10^{-6}$	$<3.1 \times 10^{-6}$	$<3.1 \times 10^{-6}$
NO_2^-	5.1×10^{-7}	2.2×10^{-7}	4.6×10^{-6}
NO_3^-	1.4×10^{-6}	6.5×10^{-8}	1.4×10^{-5}
N_2	2.8×10^{-3}	1.6×10^{-4}	5.7×10^{-3}
NH_3	6.7×10^{-5}	1.6×10^{-9}	4.1×10^{-4}
PO_4^{2-}	1.0×10^{-7}	3.6×10^{-8}	1.8×10^{-7}
F	1.0×10^{-4}	3.7×10^{-6}	1.6×10^{-4}
Cl	5.2×10^{-2}	2.6×10^{-2}	1.7×10^{-1}
Br	1.3×10^{-4}	6.1×10^{-5}	4.2×10^{-4}
Li	6.0×10^{-5}	7.2×10^{-7}	2.4×10^{-4}
Na	3.4×10^{-2}	1.0×10^{-2}	1.0×10^{-1}
Mg	1.9×10^{-5}	2.1×10^{-7}	2.3×10^{-4}
K	1.9×10^{-4}	7.9×10^{-5}	5.6×10^{-4}
Ca	7.5×10^{-3}	1.6×10^{-3}	3.9×10^{-2}
Sr	7.9×10^{-5}	3.2×10^{-5}	2.4×10^{-4}
Ba	5.8×10^{-6}	1.7×10^{-6}	2.5×10^{-5}
Al	1.8×10^{-6}	7.4×10^{-9}	2.5×10^{-5}
Si	2.1×10^{-4}	3.9×10^{-5}	5.8×10^{-4}
Mn	1.7×10^{-5}	1.8×10^{-9}	4.3×10^{-4}
Fe	8.1×10^{-7}	1.8×10^{-8}	5.2×10^{-6}
Cr	2.1×10^{-7}	1.9×10^{-8}	6.7×10^{-7}
Co	6.6×10^{-8}	1.7×10^{-8}	1.1×10^{-6}
Ni	8.2×10^{-8}	1.7×10^{-8}	3.4×10^{-7}
Cu	7.8×10^{-8}	1.6×10^{-8}	3.1×10^{-7}
Zn	2.0×10^{-7}	1.5×10^{-8}	1.0×10^{-6}
As	6.9×10^{-7}	2.1×10^{-8}	6.9×10^{-6}

W	$1.7x10^{-6}$	$5.4x10^{-9}$	$4.7x10^{-6}$
U	$1.7x10^{-7}$	$4.2x10^{-8}$	$4.0x10^{-7}$

Table S3. Proteins used to build phylogenetic tree of Fig. 1.

The universally distributed COGs that do not have ambiguous alignments (*45*) that were used to build the phylogenetic tree in Fig. 1. The tree was from a concatenated multiple sequence alignment built using MUSCLE (*46*), and determined by maximum likelihood by PHYML (*47*) with 100 replicates for bootstrapping (sampling with replacement), using the JTT amino acid substitution model (*48*). Genomes in which COGs were found in multiple copies and therefore excluded from the analysis for those species are indicated in the Notes.

Gene	Name	Description	Len	Notes
Daud2218	ychF	COG0012 Predicted GTPase, probable translation factor	327	extra Desulfotomaculum reducens
Daud1378	pheS	COG0016 Phenylalanine-tRNA synthethase alpha subunit	340	
Daud0220	rpsL	COG0048 Ribosomal protein S12	125	
Daud0221	rpsG	COG0049 Ribosomal protein S7	157	
Daud0606	rpsB	COG0052 Ribosomal protein S2	245	
Daud0211	rplK	COG0080 Ribosomal protein L11	143	
Daud0212	rplA	COG0081 Ribosomal protein L1	232	
Daud0225	rplC	COG0087 Ribosomal protein L3	214	
Daud0230	rplV	COG0091 Ribosomal protein L22	114	
Daud0231	rpsC	COG0092 Ribosomal protein S3	219	
Daud0235	rplN	COG0093 Ribosomal protein L14	123	
Daud0237	rplE	COG0094 Ribosomal protein L5	181	
Daud0239	rpsH	COG0096 Ribosomal protein S8	132	
Daud0240	rplF	COG0097 Ribosomal protein L6	182	
Daud0242	rpsE	COG0098 Ribosomal protein S5	169	
Daud0250	rpsM	COG0099 Ribosomal protein S13	124	

Daud0251	rpsK	COG0100 Ribosomal protein S11	129	
Daud0330	rplM	COG0102 Ribosomal protein L13	146	
Daud0331	rpsI	COG0103 Ribosomal protein S9	131	
Daud0013	serS	COG0172 Seryl-tRNA synthetase	426	
Daud0933	rpsO	COG0184 Ribosomal protein S15	90	
Daud0234	rpsQ	COG0186 Ribosomal protein S17	86	
Daud0232	rplP	COG0197 Ribosomal protein L16	144	
Daud0244	rplO	COG0200 Ribosomal protein L15	147	
Daud0245	prlA	COG0201 Preprotein translocase subunit SecY	425	
Daud0253	rpoA	COG0202 DNA-directed RNA polymerase, alpha subunit	316	
Daud0241	rplR	COG0256 Ribosomal protein L18	124	
Daud1865	leuS	COG0495 Leucyl-tRNA synthetase	827	
Daud0252	rpsD	COG0522 Ribosomal protein S4 and related proteins	209	extra Clostridium novyi , Symbiobacterium thermophilum

Table S4(a,b). Counts of closest homologs in sequenced organisms.

Supporting the phylogenetic assignment of *D. audaxviator* in Fig. 1, Table S4(a) reports the number of homologs from each microorganisms that provide the closest homolog (as determined by possessing the highest BLASTp bit score) to a protein-coding gene in *D. audaxviator*. To ascertain whether there was bias caused by undercounting homologous genes that are very close to the top hit, we also report Table S4(b), which gives the number of homologs that are high-scoring (within 25 bits of the highest scoring homolog) from each microorganism. In both views, the *Desulfotomaculum*-clade member *Pelotomaculum thermopropionicum* (*49*), a syntrophic propionate oxidizer, has the greatest number of genes that are closest to those found in *D. audaxviator*. The genome of *P. thermopropionicum* was unfortunately incomplete when this analysis was performed, so it is likely even more closely related to *D. audaxviator* than this partial comparison suggests. *Desulfotomaculum reducens* (*50*) is second (unfortunately, it was also incomplete at the time of this analysis), followed by *Moorella thermoacetica* (*51*) (previously named *Clostridium thermoaceticum*), and *Carboxydothermus hydrogenoformans* (*52*). At the time of this analysis, we unfortunately also did not have genomic sequence for the other relatives *Syntrophomonas wolfei* subsp. wolfei str. Goettingen, *Heliobacterium modesticaldum* Ice1, *Thermosinus carboxydivorans* Nor1, and *Clostridium novyi* NT.

Table S4(a).

Organism	Count
Pelotomaculum thermopropionicum SI	667
Desulfotomaculum reducens MI-1	397
Moorella thermoacetica ATCC 39073	214
Carboxydothermus hydrogenoformans Z-2901	135
Thermoanaerobacter tengcongensis MB4T	59
Symbiobacterium thermophilum IAM 14863	34
Geobacillus kaustophilus HTA426	22
Desulfitobacterium hafniense DCB-2	22
Desulfitobacterium hafniense Y51	21
Desulfuromonas spp.	20
Methanosarcina acetivorans C2A	17
Dehalococcoides ethenogenes 195	16
Lactobacillus sakei subsp. sakei 23K	15
Pelobacter carbinolicus str. DSM 2380	15
Archaeoglobus fulgidus DSM 4304	13
Thermus thermophilus HB8	12
Methanothermobacter thermautotrophicus DH	12
Geobacter sulfurreducens PCA	12
Methanosarcina barkeri str. fusaro	12
Geobacter metallireducens GS-15	10
Dehalococcoides sp. CBDB1	10

Table S4(b).

Organism	Count
Pelotomaculum thermopropionicum SI	968
Desulfotomaculum reducens MI-1	914
Moorella thermoacetica ATCC 39073	636

Carboxydothermus hydrogenoformans Z-2901	464
Desulfitobacterium hafniense Y51	240
Thermoanaerobacter tengcongensis MB4T	237
Desulfitobacterium hafniense DCB-2	220
Symbiobacterium thermophilum IAM 14863	219
Geobacillus kaustophilus HTA426	156
Geobacter metallireducens GS-15	132
Geobacter sulfurreducens PCA	131
Bacillus halodurans C-125	127
Desulfuromonas spp.	113
Bacillus thuringiensis serovar konkukian 97-27	101
Bacillus licheniformis DSM 13	97
Pelobacter carbinolicus str. DSM 2380	97
Bacillus cereus ZK	96
Bacillus subtilis	93
Bacillus anthracis str. Ames	92
Bacillus clausii KSM-K16	89
Clostridium acetobutylicum ATCC 824	83
Desulfovibrio desulfuricans G20	78
Dehalococcoides ethenogenes 195	77
Methanosarcina acetivorans C2A	76
Oceanobacillus iheyensis HTE831	76
Desulfovibrio vulgaris Hildenborough	70

Figure S1. Relationship to sequenced organisms and environmental clones.

The 16S gene of this phylotype is almost identical to a 16S clone from a biofilm found in a Danish heating system (1506 of 1510 positions are the same, with 2 of the 4 differing positions an unidentified nucleotide in the Danish clone sequence). The 16S gene is also very similar to a 16S clone (~96% identity) found in the Gulf of Cadiz (where runoff from mining occurs and is the location of

undersea mud volcanoes), a 16S clone (~96% identity) from an aquifer in New Mexico and 16S clones (~96% identity) from the vents of the Juan de Fuca ridge in the North Eastern Pacific Ocean (Fig. S1 of SOM). Among isolated organisms, its 16S most resembles those from the *Desulfotomaculum* clade, many of which were derived from subsurface environments, but the identity to the closest 16S of a *Desulfotomaculum*, *D. kuznetsovii*, is only ~90% (well below the generally-accepted genus cutoff of 97%), and so it does not belong in this genus.

Phylogenetic tree based on 16S rDNA sequences from both sequenced organisms and environmental clones (some of which were truncated). Sequences aligned with MUSCLE (*46*). Tree determined by maximum likelihood with PHYML (*47*) using HKY substitution model (*53*). High bootstrap value supported nodes are indicated by circles. Note that the topology is slightly different for the placement of some relatives (e.g. *Symbiobacterium* and *Thermoanaerobacter*), due to the decreased amount of information (fewer positions) available from the 16S sequence compared with the protein tree as well as the lack of intermediate species with which to build the tree. Environmental clones correspond to the accession numbers AF517773 (OGL-7B Rainbow hydrothermal vent mid-Atlantic ridge), AY225657 (AT-s30 Rainbow hydrothermal vent mid-Atlantic ridge), AF325224 (P-3 Piceance basin Colorado), DQ208688 (RL50JIII Japanese mine), AY753399 (SK18A Heating system biofilm Skanderborg Denmark), AY181047 (1026B3 Juan de Fuca ridge NE Pacific), AY181044 (1026B117 Juan de Fuca ridge NE Pacific), AY753389 (SK21 Heating system biofilm Skanderborg Denmark), AY122603 (OSS-33 New Mexico aquifer), DQ004670 (CAMV300B902 Gulf of Cadiz Spain/Portugal).

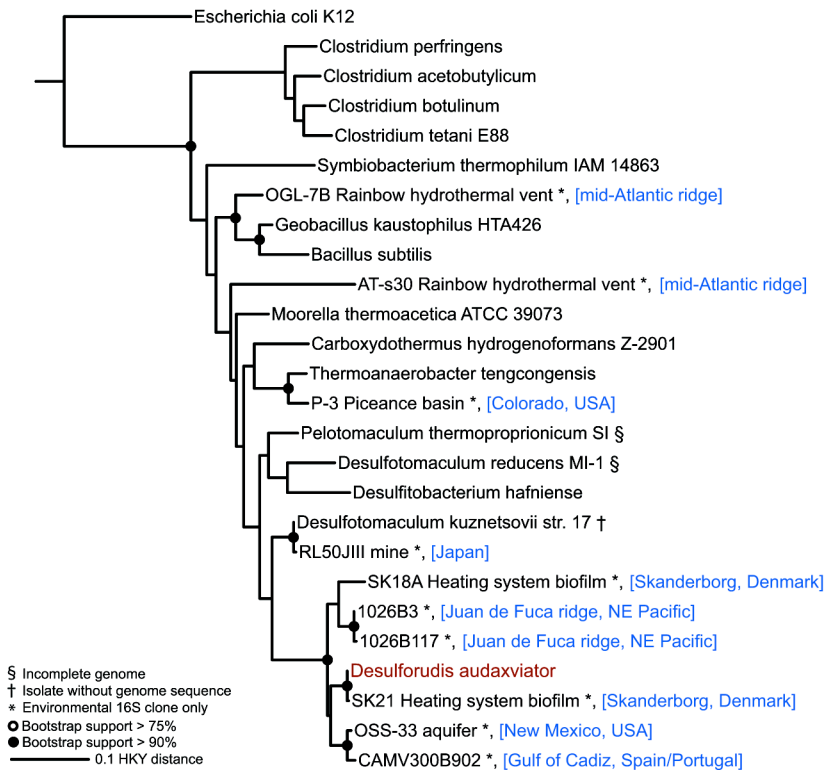

Escherichia coli K12
Clostridium perfringens
Clostridium acetobutylicum
Clostridium botulinum
Clostridium tetani E88
Symbiobacterium thermophilum IAM 14863
OGL-7B Rainbow hydrothermal vent *, [mid-Atlantic ridge]
Geobacillus kaustophilus HTA426
Bacillus subtilis
AT-s30 Rainbow hydrothermal vent *, [mid-Atlantic ridge]
Moorella thermoacetica ATCC 39073
Carboxydothermus hydrogenoformans Z-2901
Thermoanaerobacter tengcongensis
P-3 Piceance basin *, [Colorado, USA]
Pelotomaculum thermoproprionicum SI §
Desulfotomaculum reducens MI-1 §
Desulfitobacterium hafniense
Desulfotomaculum kuznetsovii str. 17 †
RL50JIII mine *, [Japan]
SK18A Heating system biofilm *, [Skanderborg, Denmark]
1026B3 *, [Juan de Fuca ridge, NE Pacific]
1026B117 *, [Juan de Fuca ridge, NE Pacific]
Desulforudis audaxviator
SK21 Heating system biofilm *, [Skanderborg, Denmark]
OSS-33 aquifer *, [New Mexico, USA]
CAMV300B902 *, [Gulf of Cadiz, Spain/Portugal]
§ Incomplete genome
† Isolate without genome sequence
* Environmental 16S clone only
Bootstrap support > 75%
Bootstrap support > 90%
0.1 HKY distance

Figure S2. Microscopy.

Figure S2(a). DAPI stain fluorescence microscopy.

MP104 microscopy sample #2 was stained with DAPI and imaged (see Methods). Across many images and based on flow cytometry on the sample, only one "rod-like" vegetative morphotype was observed, consistent with the single genome assembled from the DNA. Bright points in image below may represent spores, which could have formed after collection of the sample as it was not fixed at any point during the ~6 years from collection to imaging. Due to the 4°C storage temperature compared with the ~60 °C conditions at MP104, it is unlikely that the reverse, germination, has occured. Image contrast enhanced uniformly increase clarity. Image courtesy James Bruckner, DRI.

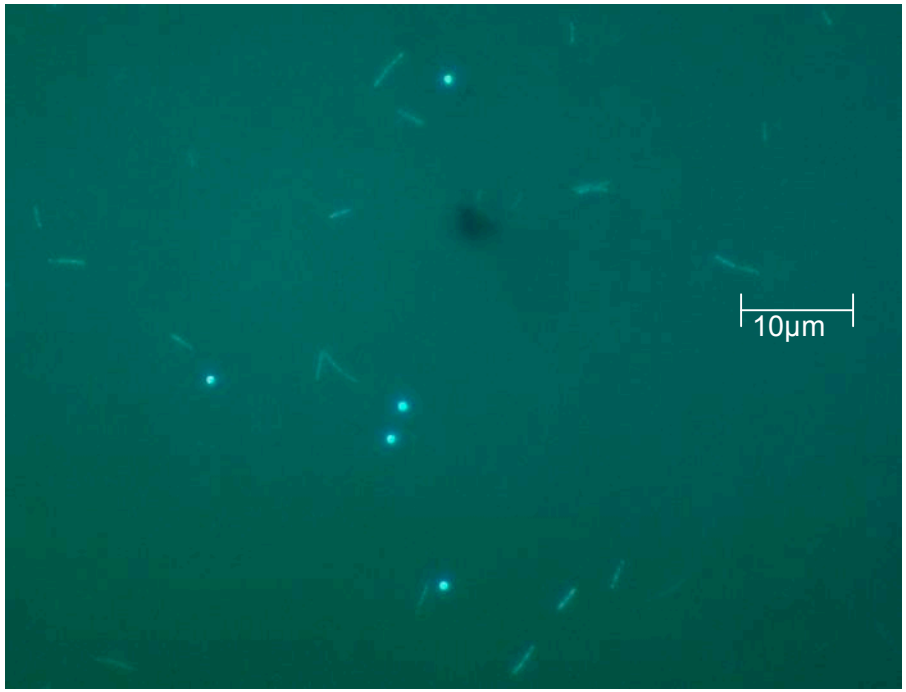
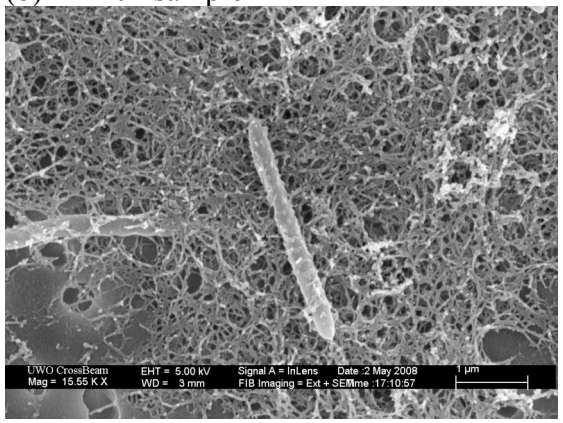


Figure S2(b,c). SEM of MP104 sample #1 and DR9.

Scanning electron micrographs were taken of (b) MP104 microscopy sample #1 (see Methods) and (c) a sample from a different mine, 648 meters down in borehole D8A, where *D. audaxviator* is also the predominant organism (*2*). Only one morphotype was observed at both locations.

(b) MP104 sample #1



(c) D8A

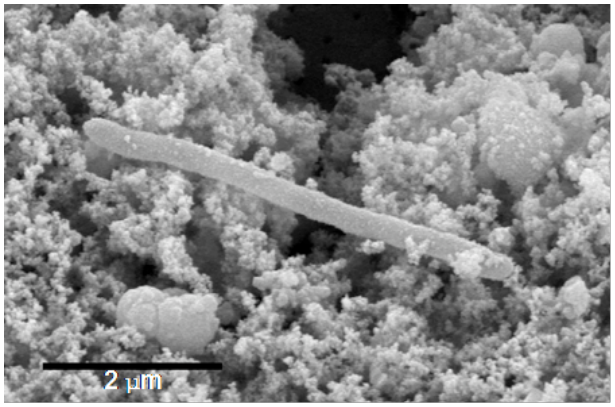
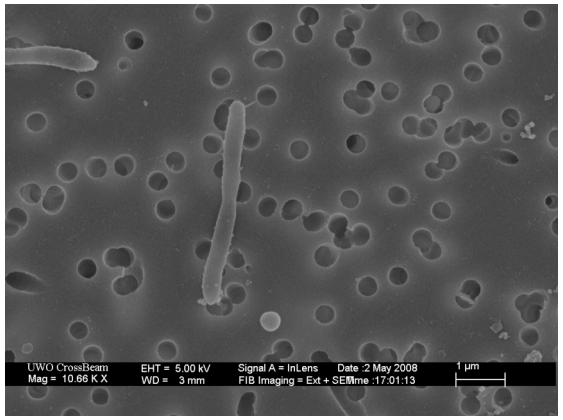


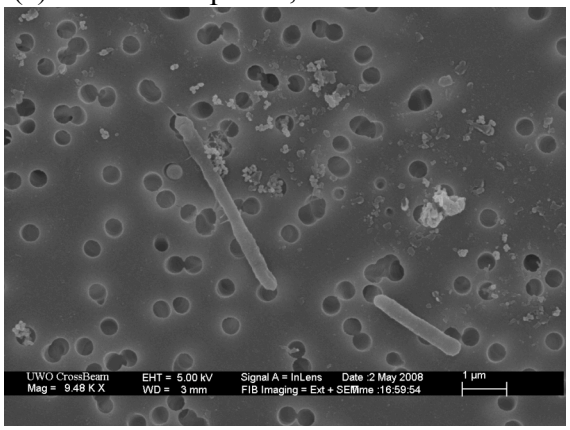
Figure S2(d,e,f). SEM of Microscopy sample #2.

MP104 microscopy sample #2 was also morphologically consistent in SEM, showing only rod-like cells and small spherical objects that may be spores in all views, three of which are shown in (d-f). Sample #2, like sample #1, was also left unfixed so any spores could have formed after sample collection, but given the 4^{o}C storage temperature it is unlikely that the reverse, germination, has occured. The dark objects with bright edges in the images are filter pores and the tiny globular objects are probably various minerals.

(d) MP104 sample #2, field 1



(e) MP104 sample #2, field 2



(f) MP104 sample #2, field 3

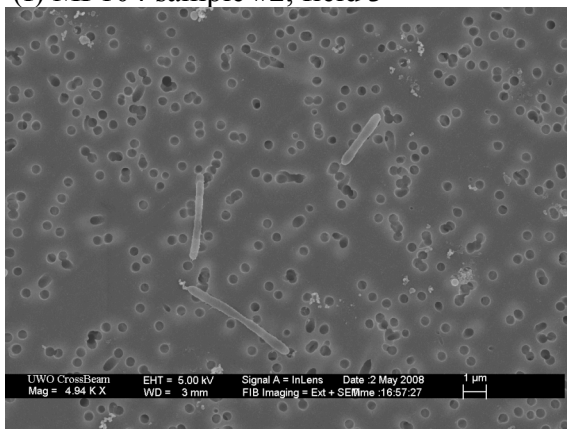


Figure S2(g). Catalyzed-reporter Deposition Fluorescent *In Situ* Hybridization (CARD-FISH) microscopy of sample #2. Epifluorescence images showing a rod-shaped cell viewed at the emission maxima for (g.a) the nucleic acid stain DAPI (461 nm) and (g.b) Cy3 (565 nm). The white scale bars represent 5 µm. We were able to photograph only a single DAPI stained cell in over 50 fields of view across 5 filter sections. This cell was also Cy3 stained, which demonstrates the presence of the horse-radish peroxidase labeled probe specific to Candidatus *Desulforudis audaxviator* 16S rRNA. We suspect that the low number of stained cells may have been due to the relatively harsh permeabilization procedure used. Additionally, the use of ethanol as a fixative results in significant precipitation from these fracture water concentrates which obscures the fields of view making focusing extremely difficult.

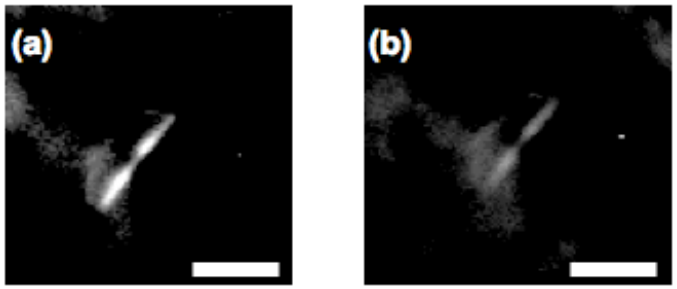


Figure S3(a,b,c,d). 16S rRNA gene PCR amplification of gDNA.

Figure S3(a). Proportions of 16S clones in intact (SGNY) and pre-digested (SGNX) libraries.

Distribution of clone sequences in 16S rRNA gene libraries generated from intact gDNA isolated from Mponeng fracture water (SGNY library) and selectively pre-digested Mponeng fracture water gDNA (SGNX library).

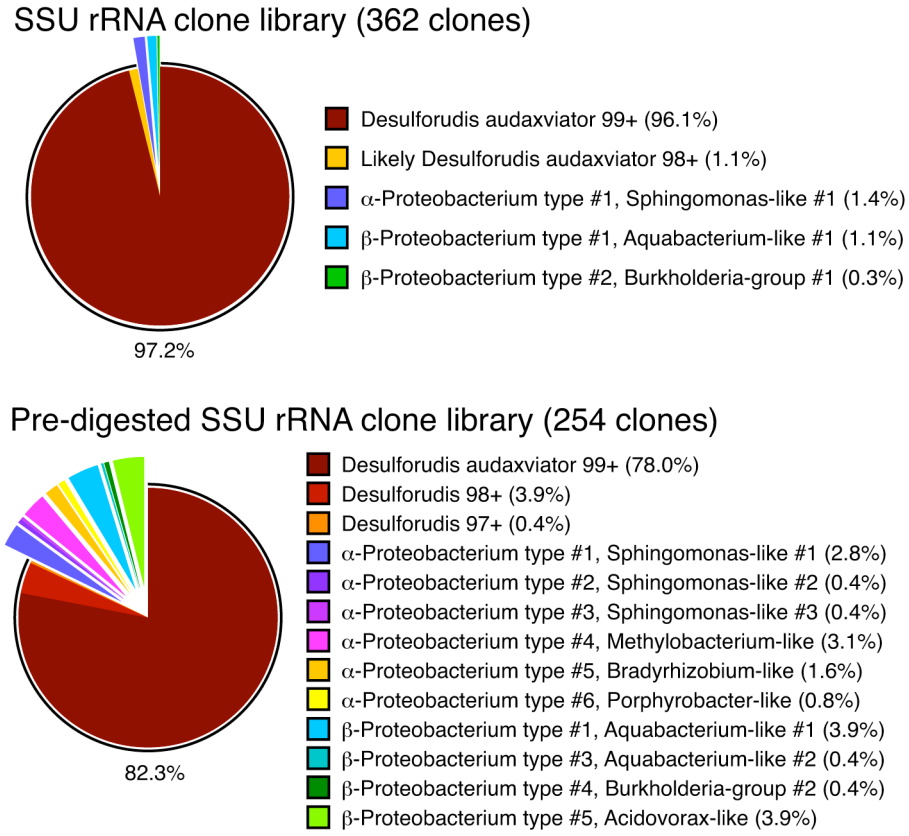


Figure S3(b). Diversity of Acid Mine Drainage (AMD) community (16S clones).

Classification of 16S clones from Tyson, *et al*. (*54*). Metagenomic sequencing of AMD samples more closely followed population structure indicated by 16S clone analysis than we found for South African gold mine MP104 sample.

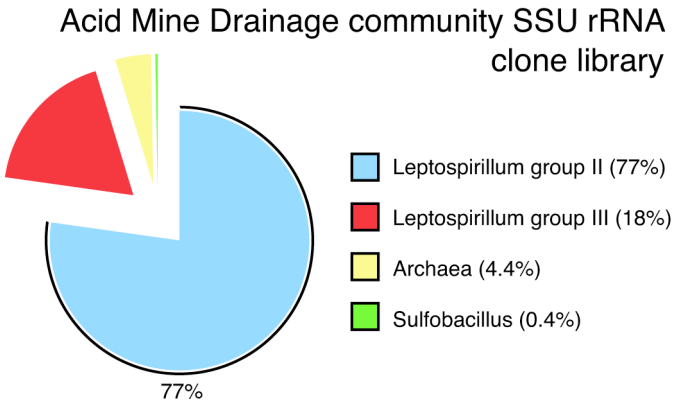


Figure S3(c). Phylogenetic assignment of 16S clones.

Categorization of the sequences from regular SGNY and predigested SGNX library, clustered at 99%, is shown below, with one representative for each cluster of clones. Phylogenetic tree based on 16S rRNA gene sequences from both sequenced organisms and environmental clones. Sequences aligned with MUSCLE (*46*). Tree determined by maximum likelihood with PHYML (*47*) using HKY substitution model (*53*). High bootstrap value supported nodes are indicated by circles.

Clones from MP104 filter extract have names of the form "Clone MP104-MF-*", with clones from the intact SGNY library in green and clones from the pre-digested SGNX library in blue. The number of members in each cluster larger than one is shown in orange parentheses. For reference, also shown in purple are clones from the South African subsurface, including those found at the MP104 site by Lin, *et al.* (*3*). Clones from the fracture water ("Clone MP104-<date>-*") are in purple. Clones from the service water ("Clone MP104-SW-*") are in red. Reference 16S rRNA gene sequences for classification are black. Hypervariable regions of the 16S gene were suppressed using a Lane mask (*38*) from *Clostridium botulinum* type F.

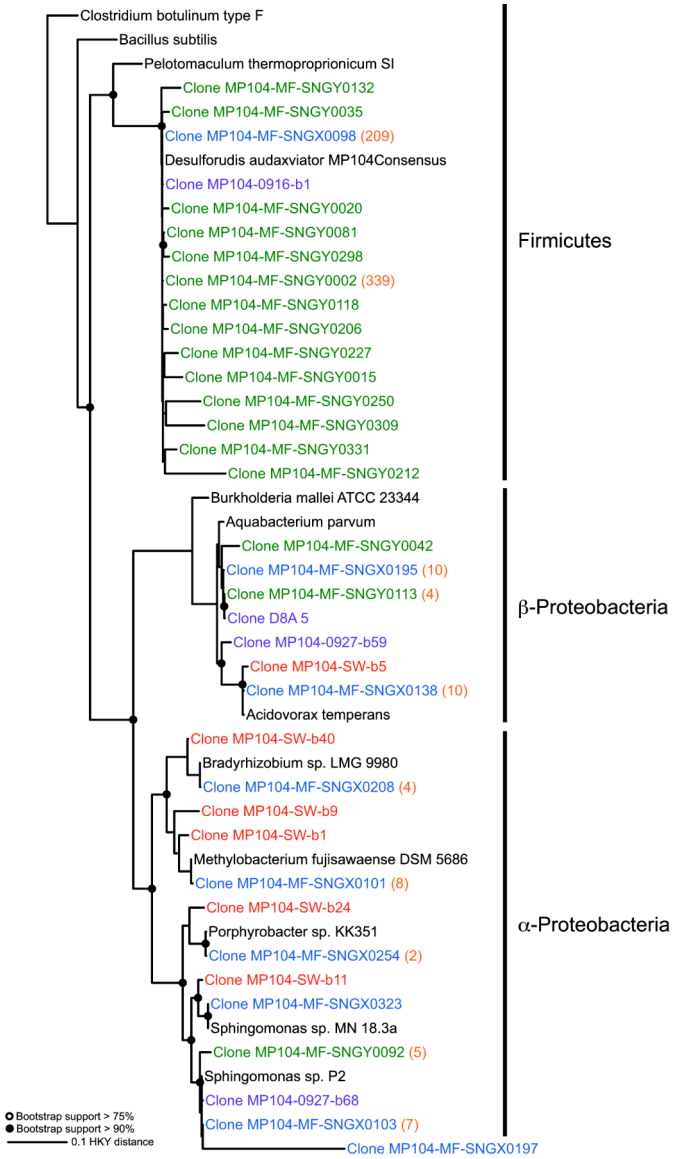
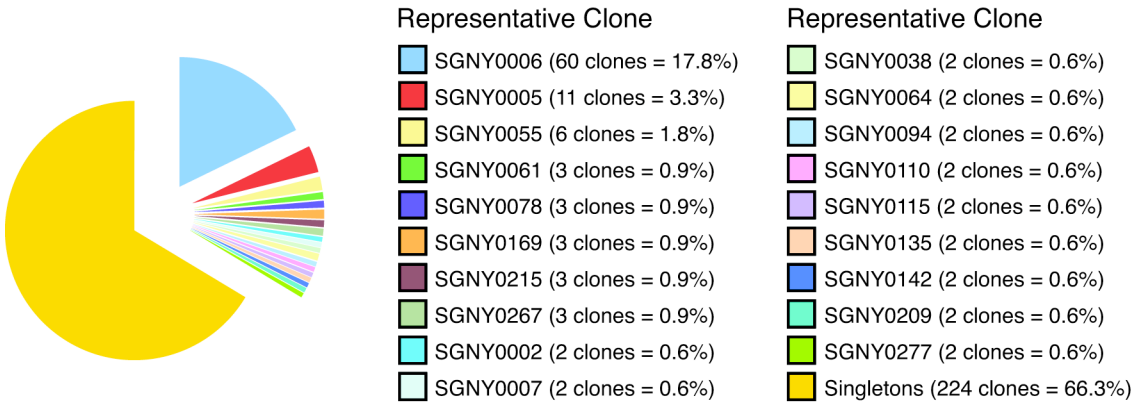

Clostridium botulinum type F
Bacillus subtilis
Pelotomaculum thermoproprionicum SI
Clone MP104-MF-SNGY0132
Clone MP104-MF-SNGY0035
Clone MP104-MF-SNGX0098 (209)
Desulforudis audaxviator MP104Consensus
Clone MP104-0916-b1
Clone MP104-MF-SNGY0020
Clone MP104-MF-SNGY0081
Clone MP104-MF-SNGY0298
Clone MP104-MF-SNGY0002 (339)
Clone MP104-MF-SNGY0118
Clone MP104-MF-SNGY0206
Clone MP104-MF-SNGY0227
Clone MP104-MF-SNGY0015
Clone MP104-MF-SNGY0250
Clone MP104-MF-SNGY0309
Clone MP104-MF-SNGY0331
Clone MP104-MF-SNGY0212
Firmicutes
Burkholderia mallei ATCC 23344
Aquabacterium parvum
Clone MP104-MF-SNGY0042
Clone MP104-MF-SNGX0195 (10)
Clone MP104-MF-SNGY0113 (4)
Clone D8A 5
Clone MP104-0927-b59
Clone MP104-SW-b5
Clone MP104-MF-SNGX0138 (10)
Acidovorax temperans
β-Proteobacteria
Clone MP104-SW-b40
Bradyrhizobium sp. LMG 9980
Clone MP104-MF-SNGX0208 (4)
Clone MP104-SW-b9
Clone MP104-SW-b1
Methylobacterium fujisawaense DSM 5686
Clone MP104-MF-SNGX0101 (8)
Clone MP104-SW-b24
Porphyrobacter sp. KK351
Clone MP104-MF-SNGX0254 (2)
Clone MP104-SW-b11
Clone MP104-MF-SNGX0323
Sphingomonas sp. MN 18.3a
Clone MP104-MF-SNGY0092 (5)
Sphingomonas sp. P2
Clone MP104-0927-b68
Clone MP104-MF-SNGX0103 (7)
Clone MP104-MF-SNGX0197
α-Proteobacteria
Bootstrap support > 75%
Bootstrap support > 90%
0.1 HKY distance

Figure S3(d). Diversity in *Desulforudis* 16S rRNA gene clones.

Clones from the intact SGNY library that matched at 99% or better (without corroboration of polymorphism) to the 2 identical consensus 16S rRNA genes of the assembly were further clustered into groups, with 100% identity within each group. Most, if not all, of the variation in the sequences from the assembly is likely due to PCR amplification error, which typically has about 1.5 errors in 1400 bases with the Taq polymerase in the protocol used.



Additionally, examination of the 16S clones in the SGNY library that were close to the *Desulforudis audaxviator* consensus 16S sequence but below 99% identity (12 distant clones: SGNY 0015, 0020, 0081, 0118, 0132, 0206, 0212, 0227, 0250, 0298, 0309, 0331) revealed that the polymorphisms in those sequences were not reliable. Each clone sequence was obtained by the phrap assembly of a single forward and a single reverse Sanger read, which tended to overlap only for the middle 400 bases. In this region, it was possible to determine which of the putative polymorphisms in the clone sequence were confirmed by both the forward and reverse read. Of the 170 putative polymorphisms found in these 12 clones in the overlapping ~400 base region, only 12 polymorphisms were corroborated by both the forward and reverse read. Based on this indicated untrustworthiness of putative polymorphisms based on a single read and the non-negligible error rate of the polymerase used in the PCR reactions, we elected to consider only those polymorphisms found in more than 1 clone (ignoring the forward and reverse reads since ¾ of the clone length was covered by only one of the reads) as somewhat reliable. Applying the rule of corroboration of the same polymorphism at that position from at least one of the 352 *D.*

audaxviator 16S clones in the SGNY library over the full length of each clone, we find that of the 519 putative polymorphisms in the 12 distant clones, only 178 at 116 positions are confirmed by a duplicate observation in the full set of 352 16S clones (changing the % identity for SGNY 0015: from 97.4% to 99.1%; 0020: 98.5% to 99.6%; 0081: 98.7% to 99.7%; 0118: 98.6% to 99.5%; 0132: 97.2% to 99.1%; 0206: 98.7% to 99.2%; 0212: 93.0% to 98.5%; 0227: 97.9% to 98.7%; 0250: 95.2% to 98.3%; 0298: 98.2% to 99.4%; 0309: 93.1% to 98.1%; 0331: 97.7% to 99.2%). Additionally, of the near set of 340 clones (after removal of the 12 more distant clones) there are 868 putative polymorphisms, with 464 confirmed at 170 positions. Combining the results for both the near and distant clones yields 642 confirmed at 198 positions. However, even if the most conservative means of obtaining the rate of polymorphism is used, 170 positions out of the consensus 16S length of 1692 bases yields a rate of about 10% at a depth of coverage of 340X (this stands in stark contrast to the 0 confirmed SNPs in the 16S region found in the Sanger metagenomic reads). While a substantial fraction of the difference between the 16S clone library SNP rate and the SNP rate inferred from the metagenomic sequencing (32 confirmed SNPs throughout the entire genome at a depth of coverage of about 8X) may be true heterogeneity and can be explained by the greatly increased likelihood of observing a duplicate polymorphism with 340 / 8 = 42.5 times the number of sequences, the discrepancy may instead indicate that 16S PCR libraries are inherently less reliable for obtaining high-fidelity sequence than whole-genome shotgun approaches, with a greater number of systematic PCR or sequencing errors contributing an increased probablilty of two errors occurring that confirm one another.

Table S5(a,b). Phylogenetic microarray analysis.

Table S5(a). Comparison of PhyloChip and 16S rRNA gene clone libraries.
Analysis of prokaryotic species composition in fracture water gDNA extracts used for metagenome analysis.

		Clone library and microarray comparisons			
16S rRNA library SGNY	**Phylogenetic node**	**Percent clones assigned to array groups**	**Array only**	**Array and cloning**	**Cloning only**
Non-digested fracture water DNA	Phylum	100	24	2	0
	Class*	99	45	4	0
	Order	3	68	1	1
	Family	3	95	1	1
	Subfamily	3	108	1	1
	OTU	3	174	1	1
Shannon diversity (at 99%)	0.77				
Number of taxa (at 99%)	25				
Dominance (1-E)	0.76				
Maximum Chao1 richness estimate (at 99%)	102				
16S rRNA library SGNX	**Phylogenetic node**	**Percent clones assigned to array groups**	**Array only**	**Array and cloning**	**Cloning only**
Pre-digested fracture water DNA	Phylum	100	24	2	0
	Class*	99	45	4	0
	Order	18	67	2	1
	Family	18	94	2	2
	Subfamily	18	107	2	3
	OTU	17	174	1	6
Shannon diversity (at 99%)	1.26				
Number of taxa (at 99%)	25				
Dominance (1-E)	0.61				
Maximum Chao1 richness estimate (at 99%)	160				
Percent increase in diversity detected due to selective pre-digestion (calculated from Shannon index and Chao1)	61-64				

* the large increase in clones assigned at class level is due to the divergence between the dominant *Desulforudis audaxviator* sequence and the organisms represented on the PhyloChip

Table S5(b). Microbial identifications by various methods.

Comparison of fracture water microbial counts by both 16S rRNA gene composition (PhyloChip, intact, and pre-digested 16S PCR amplification of gDNA) and by BLASTn and BLASTx matches of metagenomic sequence reads (Sanger and 454 pyrosequencing reads), including likely microbial contamination. Note that BLASTx matches of metagenomic sequence reads are not necessarily highly specific and may offer misleading assignments as they do not properly account for horizontal transfers.

Group	PhyloChip	Intact 16S Clones	Pre-digested 16S Clones	Sanger Reads	454 Reads
Archaea; Crenarchaeota; Thermoprotei	no	0	0	0	1
Archaea; Euryarchaeota; Thermoplasmata	yes	0	0	0	0
Archaea; Euryarchaeota; Methanomicrobia	no	0	0	1	4
Archaea; Euryarchaeota; Thermococci	no	0	0	0	1
Bacteria; Acidobacteria; Acidobacteria	yes	0	0	0	1
Bacteria; Acidobacteria; Acidobacteria-5	yes	0	0	0	0
Bacteria; Acidobacteria; Solibacteres	yes	0	0	0	0
Bacteria; Actinobacteria; Actinobacteria	yes	0	0	1	8
Bacteria; Actinobacteria; BD2-10 group	yes	0	0	0	0
Bacteria; BRC1; Unclassified	yes	0	0	0	0
Bacteria; Bacteroidetes; Bacteroidetes	yes	0	0	0	2
Bacteria; Bacteroidetes; Flavobacteria	yes	0	0	0	0
Bacteria; Bacteroidetes; KSA1	yes	0	0	0	0
Bacteria; Bacteroidetes; Sphingobacteria	yes	0	0	1	0
Bacteria; Bacteroidetes; Unclassified	yes	0	0	0	0
Bacteria; Chloroflexi; Anaerolineae	yes	0	0	0	0
Bacteria; Chloroflexi; Chloroflexi-4	yes	0	0	0	0
Bacteria; Chloroflexi; Dehalococcoidetes	yes	0	0	0	2
Bacteria; Chloroflexi; Thermomicrobia	yes	0	0	0	0
Bacteria; Chloroflexi; Unclassified	yes	0	0	1	0
Bacteria; Coprothermobacteria; Unclassified	yes	0	0	0	0
Bacteria; Cyanobacteria; Cyanobacteria	yes	0	0	4	1
Bacteria; Deinococcus-Thermus; Unclassified	yes	0	0	0	1
Bacteria; Firmicutes; Bacilli	yes	0	0	0	0
Bacteria; Firmicutes; Catabacter	yes	0	0	0	7
Bacteria; Firmicutes; Clostridia	yes	3	0	1	11
Bacteria; Firmicutes; Desulfotomaculum	yes	339	209	29153	499949

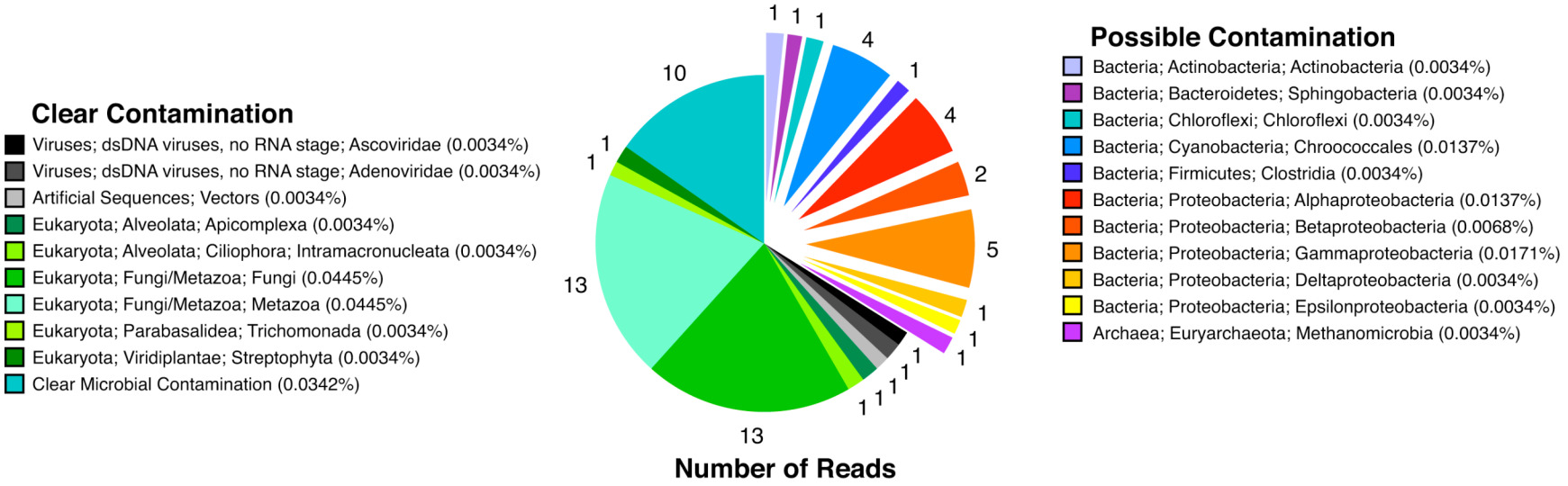
Bacteria; Firmicutes; Mollicutes	no	0	0	0	1
Bacteria; Firmicutes; Symbiobacteria	yes	0	0	0	0
Bacteria; Firmicutes; gut clone group	yes	0	0	0	0
Bacteria; Gemmatimonadetes; Unclassified	yes	0	0	0	0
Bacteria; Lentisphaerae; Unclassified	yes	0	0	0	0
Bacteria; Marine group A; mgA-2	yes	0	0	0	0
Bacteria; NC10; NC10-1	yes	0	0	0	0
Bacteria; NC10; NC10-2	yes	0	0	0	0
Bacteria; Natronoanaerobium; Unclassified	yes	0	0	0	0
Bacteria; Nitrospira; Nitrospira	yes	0	0	0	0
Bacteria; OP9/JS1; OP9	yes	0	0	0	0
Bacteria; OP10; CH21 cluster	yes	0	0	0	0
Bacteria; Planctomycetes; Planctomycetacia	yes	0	0	0	0
Bacteria; Proteobacteria; Alphaproteobacteria	yes	5	23	4	14
Bacteria; Proteobacteria; Betaproteobacteria	yes	5	22	2	5
Bacteria; Proteobacteria; Deltaproteobacteria	yes	0	0	1	11
Bacteria; Proteobacteria; Epsilonproteobacteria	yes	0	0	1	0
Bacteria; Proteobacteria; Gammaproteobacteria	yes	0	0	10	109
Bacteria; Proteobacteria; Unclassified	yes	0	0	0	0
Bacteria; Spirochaetes; Spriochaetes	yes	0	0	0	0
Bacteria; Synergistes; Unclassified	yes	0	0	0	0
Bacteria; Termite group 1; Unclassified	yes	0	0	0	0
Bacteria; Unclassified; Unclassified	yes	0	0	0	0
Bacteria; Verrucomicrobia; Unclassified	yes	0	0	0	1
Bacteria; Verrucomicrobia; Verrucomicrobia	yes	0	0	0	1
Bacteria; WS3; Unclassified	yes	0	0	0	0
Bacteria; Marine group A; mgA-2	yes	0	0	0	0
Total	N/A	352	254	29180	500130
Fraction Desulforudis (%)	N/A	96.307	82.283	99.907	99.964

Table S6(a,b,c,d,e,f). Sanger and 454 reads that don't match *D. audaxviator* assembly.

In order to ascertain the upper-bound on the proportion of organisms other than *D. audaxviator*, we assumed that the stray reads (that were not clear contamination) were not a consequence of error in the sequencing and did not result from contamination. We further assumed that each coarse taxonomic assignment corresponds to a single species (e.g. all reads classified as belonging to γ-Proteobacteria are from the same species). Under such assumptions, we find that the greatest proportion of other species in the sample indicated by Sanger sequencing would be from γ-Proteobacteria (5 reads = 0.0171%), α-Proteobacteria (4 reads = 0.0137%), and Cyanobacteria (4 reads = 0.0137%). The greatest proportion of other species in the sample indicated by 454 pyrosequencing would be from α-Proteobacteria (9 reads = 0.0018%), γ-Proteobacteria (8 reads = 0.0016%), and Clostridia (8 reads = 0.0016%).

(a) Organisms other than *Desulforudis* found in Sanger reads (sets D and E), including likely and possible contamination.

Percentages based on 29218 reads that match *Desulforudis* or other organisms (sets B, C1, C2, D, and E), including contamination.



(b) Categorization of Sanger reads.

We were very careful in our classification of reads, and attempted to be as thorough and rigorous as possible. Some of the Sanger reads suffered from errors in the sequencing and unreliable or unknown base calls, complicating the analysis (e.g. 2020 of the 31218 total Sanger reads did not have a run of at least 10 contiguous base calls with a Phred score ≥ 25). Additionally, classification of reads using protein BLAST is potentially misleading given the inability to determine whether the gene was subject to horizontal transfer. The Sanger reads were taken through a series of assignments as follows: 1. Reads were matched to the *D. audaxviator* assembly using BLASTn with a strong mismatch penalty of -3, a gap initiation penalty of -5 (the default value), a gap extension penalty of -2 (the default value), a strict e-value threshold of 1e-50, and additionally were required to match ≥ 75% of the read length at ≥ 97% nucleotide identity, yielding set B. 2. Under the assumption that usually about 80-90% of microbial genomes are protein-coding and that the signal from protein sequences would be more robust to sequencing error and sensitive for classification based on more remote relatives, the remaining reads were scanned using BLASTx (translating the reads) using default values against a protein sequence collection that combined the non-redundant protein database (NR) obtained from the NCBI on Aug. 10, 2007 with the *D. audaxviator* genome in all 6 frames of translation (NR+Da6). We chose to err on the side of sensitivity for detecting the presence of other organisms and so permitted a loose e-value cutoff of 0.1, followed by visual inspection of the alignments. We classified the read as likely a weaker member of the assembly (set C1) if the top hit was to the *D. audaxviator* translation, had a bit score of at least 50 bits and was 1.2 times or more greater than the bit score of the next-highest non-*D. audaxviator* hit. We classified the read as likely belonging to another organism (set E) if the bit score of the top hit was > 1.2 times or more greater than the top *D. audaxviator* hit. Lastly, we classified the read as difficult to assign but still possibly a member of the assembly (set D) if the bit scores of the top *D. audaxviator* hit and the top non-*D. audaxviator* hits were within a factor of 1.2 of each other. We reassigned reads from set D to set C2 when the top non-*D. audaxviator* hit was a near relative and had an identity equal to or below (or was considerably shorter) the top *D. audaxviator* hit. We also removed reads from sets D and E when the match was exceptionally untrustworthy and likely a consequence of sequencing error, indicated by low-identity matches between low-complexity non-hydrophobic P,G,A,S,T,H-rich sequences (usually collagen-like sequences from animals) by visual inspection of the alignments, and discarded them from further analysis. Remaining sequences that did not hit the assembly nor known protein sequences were likely not microbial in origin, and discarded from further analysis. Attention was also paid to whether legitimate reads were eukaryotic in origin, from common contaminants or likely contaminants from organisms sequenced at the JGI, with such reads discarded from further analysis. Lastly, the remaining reads that did not match the assembly and did not contain protein-coding sequences were checked for nucleotide similarity to sequenced genomes, with a small number matching Eukaryotic and microbial contamination only (results not shown), and were discarded from further analysis. The remaining reads that do not match anything are mostly of poor quality and appear to be

the result of systematic error in the sequencing process, and are of a proportion consistent with that found for sequencing projects of clonal microorganisms (Alex Copeland, personal communication). We discarded the unassignable reads from further analysis.

Set	Set Description	Number of Reads
A	all Sanger reads (including PCR walking reads)	31218
B	reads matching D. audaxviator assembly with BLASTn	27696
~B	reads not matching assembly with BLASTn	3522
C1	reads from set ~B that clearly best match D. audaxviator with BLASTx	1440
C2	reads from set ~B that best match D. audaxviator and near relatives with BLASTx	17
D	reads from set ~B that match D. audaxviator and other organisms with BLASTx	4
E	reads from set ~B that clearly best match other organisms with BLASTx	61
F	reads from set E that are almost certain contamination	43
G	reads from sets D and E that are not almost certain contamination	22

From these data, we calculate that the number of legitimate (not including likely Eukaryotic or microbial contamination) Sanger reads (sets B + C1 + C2 + D + G) is 29179. We may also determine that the proportion strictly belonging to *D. audaxviator* (sets B + C1 + C2) is 29153 / 29179 = 99.9109%, and, erring on the side of favoring other organisms by including set D, the proportion that belong to other organisms (set G) is 22 / 29179 = 0.0754%.

(c) Classification of Sanger reads for sets C2, D, and E

Set C2:

Likely D. audax.	Num. of Reads	Best D.audax. Ident	Best Ident.	Best Aln. Len	Best E-value	Best Bit Score	Closest Species Other Than D. audaxviator	Partial Classification
*	11	100	92.86	56	1E-23	107	Pelotomaculum thermopropionicum SI	Bacteria; Firmicutes; Clostridia
*	4	100	93.1	29	4E-08	60.8	Thermosinus carboxydivorans Nor1	Bacteria; Firmicutes; Clostridia
*	1	86.57	80	55	3E-18	96.3	Carboxydothermus hydrogenoformans Z-2901	Bacteria; Firmicutes; Clostridia
*	1	86.81	68.13	91	3E-28	129	Desulfitobacterium hafniense Y51	Bacteria; Firmicutes; Clostridia

Set D:

Likely D. audax.	Possible Contam.	Contam. Code	Num. of Reads	Best D.audax. Ident	Best Ident.	Best Aln. Len	Best E-value	Best Bit Score	Closest Species Other Than D. audaxviator	Partial Classification
*		d	1	89.74	74.62	197	1E-79	300	Synechococcus sp. JA-3-3Ab	Bacteria; Cyanobacteria; Chroococcales
*		c	1	80.21	75.64	78	7E-26	120	Methanococcoides burtonii DSM 6242	Archaea; Euryarchaeota; Methanomicrobia
*		c	1	96.67	83.33	30	1E-05	53.9	Syntrophobacter fumaroxidans MPOB	Bacteria; Proteobacteria; Deltaproteobacteria
	*	c	1	50.85	41.1	73	0.034	43.9	Thiomicrospira denitrificans ATCC 33889	Bacteria; Proteobacteria; Epsilonproteobacteria

[a] eukaryotic or viral contamination
[b] human-associated
[c] strain sequenced at JGI
[d] very close relative sequenced at JGI (Genus-level or closer)

Set E:

Likely Contam.	Possible Contam.	Contam. Code	Num. of Reads	Best Ident.	Best Aln. Len	Best E-value	Best Bit Score	Closest Species	Partial Classification
*		a	5	83.78	74	3E-31	122	Gibberella zeae	Eukaryota; Fungi/Metazoa group; Fungi
*		a	4	78.26	69	3E-47	124	Homo sapiens	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	2	53	200	6E-50	201	Candida albicans SC5314	Eukaryota; Fungi/Metazoa group; Fungi
*		a	2	50	30	8E-04	33.1	Aspergillus clavatus NRRL 1	Eukaryota; Fungi/Metazoa group; Fungi
*		a	1	40	90	0.085	41.6	Cloning vector pNOT218	artificial sequences; vectors
*		a	1	36.67	30	0	22.3	Porcine adenovirus A	Viruses; dsDNA viruses, no RNA stage; Adenoviridae
*		a	1	35.71	70	0.024	42.7	Pan troglodytes	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	33.33	84	0.027	43.1	Aspergillus niger	Eukaryota; Fungi/Metazoa group; Fungi
*		a	1	33.33	27	0	25.8	Debaryomyces hansenii CBS767	Eukaryota; Fungi/Metazoa group; Fungi

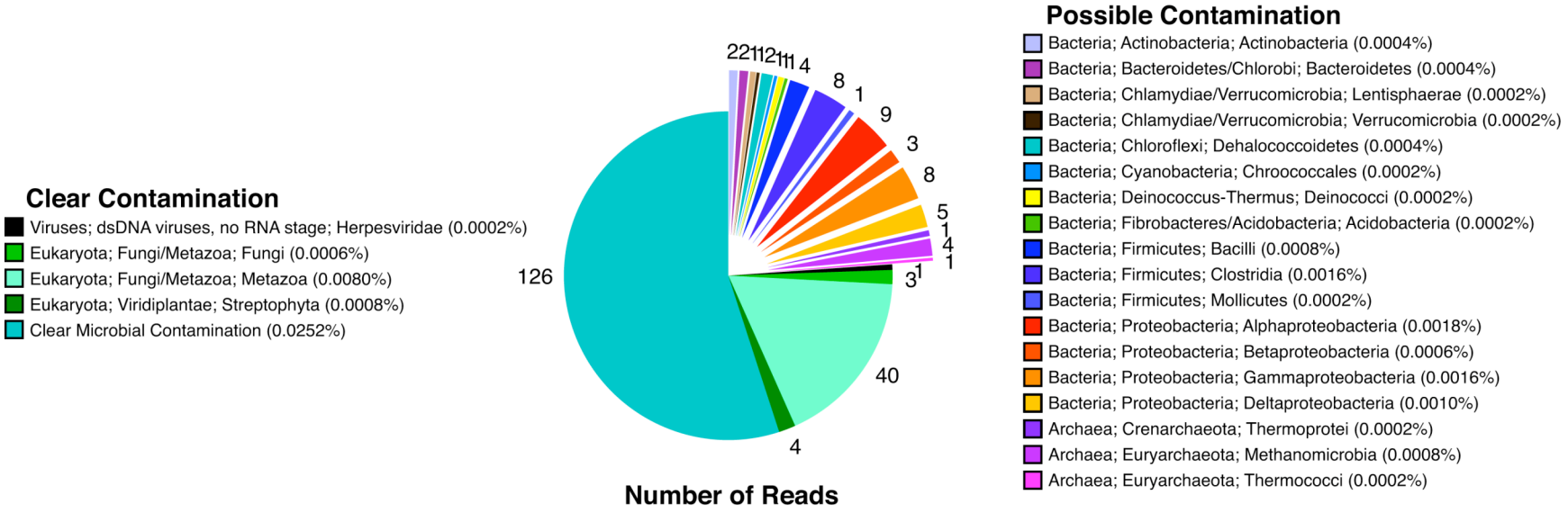
*		a	1	32.81	128	0.01	44.7	Paramecium tetraurelia	Eukaryota; Alveolata; Ciliophora; Intramacronucleata
*		a	1	32.47	77	0.094	42	Medicago truncatula	Eukaryota; Viridiplantae; Streptophyta
*		a	1	32.14	140	2E-05	52.8	Aspergillus terreus NIH2624	Eukaryota; Fungi/Metazoa group; Fungi
*		a	1	31.82	66	0	30.8	Schistosoma japonicum	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	31.43	70	0.077	39.3	Mus musculus	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	30.87	298	2E-20	103	Trichomonas vaginalis G3	Eukaryota; Parabasalidea; Trichomonada
*		a	1	30.77	65	0.012	42	Biomphalaria glabrata	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	30.65	62	0.062	40.8	Rattus norvegicus	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	29.13	103	0.05	42.7	Microcotyle sebastis	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	28.81	59	0	23.9	Drosophila melanogaster	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	28.66	157	1E-04	50.1	Macaca mulatta	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	28.47	144	5E-04	48.9	Gallus gallus	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	28.26	92	0.05	43.5	Aspergillus nidulans FGSC A4	Eukaryota; Fungi/Metazoa group; Fungi
*		a	1	26.19	42	0	23.9	Plasmodium yoelii yoelii str. 17XNL	Eukaryota; Alveolata; Apicomplexa
*		a	1	21.69	249	0.046	42.7	Spodoptera frugiperda ascovirus 1a	Viruses; dsDNA viruses, no RNA stage; Ascoviridae
*		c	6	100	59	1E-26	124	Escherichia coli B	Bacteria; Proteobacteria; Gammaproteobacteria
*		d	1	99.07	215	2E-87	325	Pseudomonas fluorescens bv. A	Bacteria; Proteobacteria; Gammaproteobacteria
	*		2	96.3	54	2E-22	110	Acinetobacter baumannii ATCC 17978	Bacteria; Proteobacteria; Gammaproteobacteria
*		d	3	92.61	176	1E-70	270	Pseudomonas fluorescens Pf-5	Bacteria; Proteobacteria; Gammaproteobacteria
	*	c	2	89.8	49	2E-18	97.4	Cyanothece sp. CCY0110	Bacteria; Cyanobacteria; Chroococcales
	*	d	1	83.5	309	5E-142	507	Bradyrhizobium sp. ORS278	Bacteria; Proteobacteria; Alphaproteobacteria
	*		1	74.07	216	4E-87	325	Janthinobacterium sp. Marseille	Bacteria; Proteobacteria; Betaproteobacteria
	*	c	1	71.43	154	1E-48	196	Rhodopseudomonas palustris	Bacteria; Proteobacteria;

								CGA009	Alphaproteobacteria
	*	c	2	56.47	85	2E-14	83.6	Shewanella sediminis HAW-EB3	Bacteria; Proteobacteria; Gammaproteobacteria
	*	c	1	55.17	29	0.02	41.2	Caulobacter sp. K31	Bacteria; Proteobacteria; Alphaproteobacteria
	*	d	1	52.52	139	2E-34	137	Pseudomonas sp. 14-3	Bacteria; Proteobacteria; Gammaproteobacteria
	*	c	1	46.81	235	1E-29	134	Burkholderia sp. 383	Bacteria; Proteobacteria; Betaproteobacteria
	*	c	1	46.67	15	0	21.2	Cytophaga hutchinsonii ATCC 33406	Bacteria; Bacteroidetes; Sphingobacteria
	*		1	35.89	248	7E-30	134	Stappia aggregata IAM 12614	Bacteria; Proteobacteria; Alphaproteobacteria
	*	c	1	35.56	180	9E-21	104	Alkaliphilus metalliredigens QYMF	Bacteria; Firmicutes; Clostridia
	*	c	1	33.77	77	0.027	44.3	Herpetosiphon aurantiacus ATCC 23779	Bacteria; Chloroflexi; Chloroflexi
	*		1	31.11	90	0.063	41.6	Thermosynechococcus elongatus BP-1	Bacteria; Cyanobacteria; Chroococcales
	*		1	30.4	125	0.019	42.7	Streptomyces avermitilis MA-4680	Bacteria; Actinobacteria; Actinobacteria

[a] eukaryotic or viral contamination
[b] human-associated
[c] strain sequenced at JGI
[d] very close relative sequenced at JGI (Genus-level or closer)

(d) Organisms other than *Desulforudis* found in 454 reads (sets D and E), including likely and possible contamination.

Percentages based on 500178 reads that match *Desulforudis* or other organisms (sets B, C1, C2, D, and E), including contamination.



(e) Categorization of 454 reads.

As with the Sanger reads, we were very careful in our classification of the 454 reads, and again attempted to be as thorough and rigorous as possible. Some of the 454 reads suffered from errors in the sequencing and unreliable base calls, especially some that were unusually much longer than the ~100 bases of the normal reads. Additionally, classification of such short reads using protein BLAST is less trustworthy than the longer Sanger reads and should be viewed mostly at a coarse taxonomic level such as class, and again, such classifications are potentially misleading given the inability to determine whether the gene was subject to horizontal transfer. The 454 reads were taken through the same series of assignments as the Sanger reads, but with a couple variations, as follows: 1. Reads were matched to the *D. audaxviator* assembly using BLASTn with a strong mismatch penalty of -3, a weakened gap

initiation penalty of -1 (not the default value, as miscounted homopolymer runs are frequent and cause short alignments with default gap penalties), a weakened gap extension penalty of -1 (again, not the default value), and additionally were required to match at least 40 bases and ≥ 75% of the read length at ≥ 97% nucleotide identity, yielding set B. 2. Under the assumption that usually about 80-90% of microbial genomes are protein-coding and that the signal from protein sequences would be more robust to sequencing error and sensitive for classification based on more remote relatives, the remaining reads were scanned using BLASTx (translating the reads) using default values against a protein sequence collection that combined the non-redundant protein database (NR) obtained from the NCBI on Aug. 10, 2007 with the *D. audaxviator* genome in all 6 frames of translation (NR+Da6). We chose to err on the side of sensitivity for detecting the presence of other organisms and so permitted a loose e-value cutoff of 0.1, followed by visual inspection of the alignments. We classified the read as likely a weaker member of the assembly (set C1) if the top hit was to the *D. audaxviator* translation, had a bit score of at least 50 bits and was 1.2 times or more greater than the bit score of the next-highest non-*D. audaxviator* hit. We classified the read as likely belonging to another organism (set E) if the bit score of the top hit was > 1.2 times or more greater than the top *D. audaxviator* hit. Lastly, we classified the read as difficult to assign but still possibly a member of the assembly (set D) if the bit scores of the top *D. audaxviator* hit and the top non-*D. audaxviator* hits were within a factor of 1.2 of each other. We reassigned reads from set D to set C2 when the top non-*D. audaxviator* hit was a near relative and had an identity equal to or below (or was considerably shorter) the top *D. audaxviator* hit. We also removed reads from sets D and E when the match was exceptionally untrustworthy and likely a consequence of sequencing error, indicated by low-identity matches between low-complexity non-hydrophobic P,G,A,S,T,H-rich sequences (usually collagen-like sequences from animals) by visual inspection of the alignments, and discarded them from further analysis. Remaining sequences that did not hit the assembly nor known protein sequences were likely not microbial in origin, and discarded from further analysis. Attention was also paid to whether legitimate reads were eukaryotic in origin, from common contaminants or likely contaminants from organisms sequenced at the JGI, with such reads discarded from further analysis. Lastly, the remaining reads that did not match the assembly and did not contain protein-coding sequences were checked for nucleotide similarity to sequenced genomes, with a small number matching Eukaryotic and microbial contamination only (results not shown), and were discarded from further analysis. The remaining reads that do not match anything are mostly of poor quality and appear to be the result of systematic error in the sequencing process, and are of a proportion consistent with that found for sequencing projects of clonal microorganisms (Alex Copeland, personal communication). We discarded the unassignable reads from further analysis.

Set	Set Description	Number of Reads
A	all 454 reads	518272
B	reads matching D. audaxviator assembly with BLASTn	493380
~B	reads not matching assembly with BLASTn	24892
C1	reads from set ~B that clearly best match D. audaxviator with BLASTx	6319
C2	reads from set ~B that match other organisms with BLASTx,but D. audaxviator with equal or higher identity and above 90%, with BLASTx	250
D	reads from set ~B that match other organisms and D. audaxviator with BLASTx	23
E	reads from set ~B that clearly best match other organisms with BLASTx	206
F	reads from set E that are almost certain contamination	170
G	reads from sets D and E that are not almost certain contamination	59

From these data, we calculate that the number of legitimate (not including likely contamination) 454 reads (sets B + C1 + C2 + D + G) is 500008. We may also determine that the proportion strictly belonging to *D. audaxviator* (sets B + C1 + C2) is 499949 / 500008 = 99.9882%, and, erring on the side of favoring other organisms by including set D, the proportion that belong to other organisms (set G) is 59 / 500008 = 0.0118%.

(f) Classification of 454 reads for sets C2, D, and E

Set C2:

Likely D. audax.	Num. of Reads	Best D.audax. Ident	Best Ident.	Best Aln. Len	Best E-value	Best Bit Score	Closest Species Other Than D. audaxviator	Partial Classification
*	46	100	100	15	0.034	36.2	Pelotomaculum thermopropionicum SI	Bacteria; Firmicutes; Clostridia
*	17	100	100	21	8E-07	55.8	Desulfotomaculum reducens MI-1	Bacteria; Firmicutes; Clostridia
*	14	100	100	20	0.005	43.1	Carboxydothermus hydrogenoformans Z-2901	Bacteria; Firmicutes; Clostridia
*	7	100	100	21	0.002	44.7	Symbiobacterium thermophilum IAM 14863	Bacteria; Firmicutes; Bacilli
*	6	100	100	18	7E-05	43.5	Syntrophobacter fumaroxidans MPOB	Bacteria; Proteobacteria; Deltaproteobacteria

*	5	100	100	18	0.045	40	Gloeobacter violaceus PCC 7421	Bacteria; Cyanobacteria; Gloeobacteria
*	4	100	100	26	8E-07	55.8	Clostridium kluyveri DSM 555	Bacteria; Firmicutes; Clostridia
*	4	100	100	23	8E-07	55.8	uncultured bacterium	Bacteria; Firmicutes; Clostridia
*	3	100	100	18	0.058	39.7	Candidatus Kuenenia stuttgartiensis	Bacteria; Planctomycetes; Planctomycetacia
*	2	100	100	27	4E-07	57	Bacillus amyloliquefaciens FZB42	Bacteria; Firmicutes; Bacilli
*	2	100	100	23	9E-06	52.4	Dorea longicatena DSM 13814	Bacteria; Firmicutes; Clostridia
*	2	100	100	22	5E-04	46.6	Alkaliphilus metalliredigens QYMF	Bacteria; Firmicutes; Clostridia
*	2	100	100	21	6E-10	53.1	Cyanothece sp. CCY0110	Bacteria; Cyanobacteria; Chroococcales
*	2	100	100	16	0.003	43.9	Methanospirillum hungatei JF-1	Archaea; Euryarchaeota; Methanomicrobia
*	2	100	100	13	0.016	33.1	Anaeromyxobacter sp. Fw109-5	Bacteria; Proteobacteria; Deltaproteobacteria
*	1	100	100	24	9E-06	52.4	Bacillus clausii KSM-K16	Bacteria; Firmicutes; Bacilli
*	1	100	100	19	0.003	43.9	Thermotoga maritima	Bacteria; Thermotogae; Thermotogae
*	1	100	100	18	0.027	40.8	Clostridium tetani E88	Bacteria; Firmicutes; Clostridia
*	1	100	100	17	0.076	39.3	Verminephrobacter eiseniae EF01-2	Bacteria; Proteobacteria; Betaproteobacteria
*	1	100	100	17	0.059	39.7	Wolbachia endosymbiont strain TRS of Brugia malayi	Bacteria; Proteobacteria; Alphaproteobacteria
*	2	96.67	96.67	30	3E-09	60.1	Geobacter sulfurreducens PCA	Bacteria; Proteobacteria; Deltaproteobacteria
*	2	100	96.3	27	6E-08	59.7	Candidatus Desulfococcus oleovorans Hxd3	Bacteria; Proteobacteria; Deltaproteobacteria
*	1	100	96.15	26	2E-06	54.7	Delftia acidovorans SPH-1	Bacteria; Proteobacteria; Betaproteobacteria
*	13	100	95.65	23	3E-04	47.4	Moorella thermoacetica ATCC 39073	Bacteria; Firmicutes; Clostridia
*	6	100	95.65	23	1E-05	51.6	Syntrophomonas wolfei str. Goettingen	Bacteria; Firmicutes; Clostridia
*	7	100	95.45	22	0.009	42.4	Desulfitobacterium hafniense Y51	Bacteria; Firmicutes; Clostridia
*	1	100	95.45	22	2E-04	47.8	Geobacter lovleyi SZ	Bacteria; Proteobacteria; Deltaproteobacteria
*	3	100	95.24	21	0.012	42	Geobacillus thermodenitrificans NG80-2	Bacteria; Firmicutes; Bacilli
*	1	100	95.24	21	0.003	43.9	alpha proteobacterium HTCC2255	Bacteria; Proteobacteria;

								Alphaproteobacteria
*	1	100	95.24	21	0.002	44.7	Synechococcus sp. BL107	Bacteria; Cyanobacteria; Chroococcales
*	2	100	95	20	0.005	43.1	Thermosipho melanesiensis BI429	Bacteria; Thermotogae; Thermotogae
*	1	95	95	20	0.007	42.7	Syntrophus aciditrophicus SB	Bacteria; Proteobacteria; Deltaproteobacteria
*	1	100	95	20	8E-04	45.8	Stigmatella aurantiaca DW4/3-1	Bacteria; Proteobacteria; Deltaproteobacteria
*	3	94.74	94.74	19	0.046	40	Clostridium cellulolyticum H10	Bacteria; Firmicutes; Clostridia
*	2	94.74	94.74	19	0.074	39.3	Synechococcus sp. WH 5701	Bacteria; Cyanobacteria; Chroococcales
*	2	100	94.74	19	0.001	45.1	Thermotoga petrophila RKU-1	Bacteria; Thermotogae; Thermotogae
*	1	100	94.74	19	0.002	44.3	Desulfotomaculum geothermicum	Bacteria; Firmicutes; Clostridia
*	1	100	94.12	17	0.078	39.3	Sphingomonas elodea	Bacteria; Proteobacteria; Alphaproteobacteria
*	1	100	93.75	16	0.015	41.6	Geobacillus tepidamans	Bacteria; Firmicutes; Bacilli
*	1	95.65	91.67	24	8E-04	45.8	Ruminococcus obeum ATCC 29174	Bacteria; Firmicutes; Clostridia
*	1	100	91.67	24	6E-04	46.2	Pedobacter sp. BAL39	Bacteria; Bacteroidetes/Chlorobi; Bacteroidetes
*	2	95.45	90.91	22	0.002	44.7	Anaeromyxobacter dehalogenans 2CP-C	Bacteria; Proteobacteria; Deltaproteobacteria
*	2	92	90.91	22	2E-04	48.1	Methanosaeta thermophila PT	Archaea; Euryarchaeota; Methanomicrobia
*	1	95	90	20	0.045	40	Methanocaldococcus jannaschii DSM 2661	Archaea; Euryarchaeota; Methanococci
*	1	100	90	20	0.009	42.4	Clostridium paraputrificum	Bacteria; Firmicutes; Clostridia
*	1	100	90	20	0.002	44.7	Roseiflexus sp. RS-1	Bacteria; Chloroflexi; Chloroflexi
*	2	100	89.47	19	0.035	40.4	Geobacter uraniumreducens Rf4	Bacteria; Proteobacteria; Deltaproteobacteria
*	1	100	89.47	19	6E-04	46.2	delta proteobacterium MLMS-1	Bacteria; Proteobacteria; Deltaproteobacteria
*	7	96.3	89.29	28	1E-06	55.1	Thermosinus carboxydivorans Nor1	Bacteria; Firmicutes; Clostridia
*	6	100	88.89	18	0.007	42.7	Heliobacillus mobilis	Bacteria; Firmicutes; Clostridia
*	1	94.44	88.89	18	0.045	40	Rhodopseudomonas palustris BisB18	Bacteria; Proteobacteria; Alphaproteobacteria
*	1	100	88.89	18	0.06	39.7	Clostridium botulinum A str. ATCC 3502	Bacteria; Firmicutes; Clostridia

*	1	100	88	25	3E-07	57.4	Clostridium beijerinckii NCIMB 8052	Bacteria; Firmicutes; Clostridia
*	4	100	87.5	24	0.001	45.4	Thermoanaerobacter ethanolicus X514	Bacteria; Firmicutes; Clostridia
*	1	100	87.5	24	9E-06	52.4	Photobacterium sp. SKA34	Bacteria; Proteobacteria; Gammaproteobacteria
*	1	100	87.5	24	2E-04	47.8	Sulfobacillus acidophilus	Bacteria; Firmicutes; Bacilli
*	1	95.45	86.96	23	0.001	45.4	Bradyrhizobium sp. ORS278	Bacteria; Proteobacteria; Alphaproteobacteria
*	1	100	86.67	30	3E-06	53.9	Rhodopseudomonas palustris CGA009	Bacteria; Proteobacteria; Alphaproteobacteria
*	1	100	86.36	22	4E-04	47	Halothermothrix orenii H 168	Bacteria; Firmicutes; Clostridia
*	1	100	86.36	22	0.004	43.5	Rhodobacter sphaeroides ATCC 17025	Bacteria; Proteobacteria; Alphaproteobacteria
*	1	91.3	85.71	21	0.035	40.4	Geobacillus kaustophilus HTA426	Bacteria; Firmicutes; Bacilli
*	1	90.48	85.71	21	0.026	40.8	Ralstonia pickettii 12J	Bacteria; Proteobacteria; Betaproteobacteria
*	1	100	85.71	21	3E-04	47.4	Streptomyces avermitilis MA-4680	Bacteria; Actinobacteria; Actinobacteria
*	1	96	84	25	8E-05	49.3	Beggiatoa sp. PS	Bacteria; Proteobacteria; Gammaproteobacteria
*	1	100	82.61	23	3E-04	47.4	Planctomyces maris DSM 8797	Bacteria; Planctomycetes; Planctomycetacia
*	1	100	82.61	23	0.016	41.6	Methanococcus maripaludis C7	Archaea; Euryarchaeota; Methanococci
*	1	94.12	82.35	34	2E-07	57.8	Streptomyces atroolivaceus	Bacteria; Actinobacteria; Actinobacteria
*	1	100	81.82	22	0.005	43.1	Haloarcula marismortui ATCC 43049	Archaea; Euryarchaeota; Halobacteria
*	1	100	81.82	22	0.005	43.1	Desulfovibrio desulfuricans G20	Bacteria; Proteobacteria; Deltaproteobacteria
*	4	88.89	81.48	27	1E-05	48.9	Thermoanaerobacter tengcongensis MB4	Bacteria; Firmicutes; Clostridia
*	1	100	81.48	27	4E-05	50.1	Azotobacter vinelandii AvOP	Bacteria; Proteobacteria; Gammaproteobacteria
*	1	100	81.25	32	4E-06	53.5	Pseudomonas mendocina ymp	Bacteria; Proteobacteria; Gammaproteobacteria
*	1	95.24	80	20	0.045	40	Caldicellulosiruptor saccharolyticus DSM 8903	Bacteria; Firmicutes; Clostridia
*	1	100	80	25	4E-04	47	Geobacter metallireducens GS-15	Bacteria; Proteobacteria;

								Deltaproteobacteria
*	1	100	80	20	0.015	41.6	Lactobacillus casei ATCC 334	Bacteria; Firmicutes; Bacilli
*	1	95	78.95	19	0.06	39.7	Pelodictyon luteolum DSM 273	Bacteria; Bacteroidetes/Chlorobi; Chlorobi
*	1	100	78.95	19	0.035	40.4	Clostridium thermocellum ATCC 27405	Bacteria; Firmicutes; Clostridia
*	1	93.1	78.57	28	2E-05	51.6	Bacillus cereus subsp. cytotoxis NVH 391-98	Bacteria; Firmicutes; Bacilli
*	1	93.94	78.12	32	1E-06	55.5	Methanosarcina barkeri str. Fusaro	Archaea; Euryarchaeota; Methanomicrobia
*	1	100	77.42	31	1E-06	55.1	Geobacter bemidjiensis Bem	Bacteria; Proteobacteria; Deltaproteobacteria
*	1	100	77.27	22	0.001	45.1	Clostridium perfringens SM101	Bacteria; Firmicutes; Clostridia
*	1	100	76	25	4E-05	50.1	Burkholderia xenovorans LB400	Bacteria; Proteobacteria; Betaproteobacteria
*	1	100	76	21	0.007	42.7	Sulfurovum sp. NBC37-1	Bacteria; Proteobacteria; Epsilonproteobacteria
*	1	100	74.19	31	3E-06	53.9	Desulfovibrio vulgaris DP4	Bacteria; Proteobacteria; Deltaproteobacteria
*	1	100	73.91	23	0.021	41.2	Clostridium novyi NT	Bacteria; Firmicutes; Clostridia
*	2	96.43	72.73	33	7E-06	52.8	Bacillus sp. B14905	Bacteria; Firmicutes; Bacilli
*	1	90.9	72.73	22	0.045	40	Clostridium sp. OhILAs	Bacteria; Firmicutes; Clostridia
*	1	100	72.73	33	4E-07	57	Chlorobium chlorochromatii CaD3	Bacteria; Bacteroidetes/Chlorobi; Chlorobi
*	1	100	72	25	0.005	43.1	Shewanella woodyi ATCC 51908	Bacteria; Proteobacteria; Gammaproteobacteria
*	1	100	69.23	39	1E-04	48.9	Lactobacillus plantarum WCFS1	Bacteria; Firmicutes; Bacilli
*	1	100	67.74	31	0.06	39.7	Deinococcus radiodurans R1	Bacteria; Deinococcus-Thermus; Deinococci
*	1	95	60.71	28	0.078	39.3	Leptospira interrogans serovar Lai str. 56601	Bacteria; Spirochaetes; Spirochaetes
*	1	100	54.29	35	0.021	41.2	Marinobacter aquaeolei VT8	Bacteria; Proteobacteria; Gammaproteobacteria
*	1	100	53.12	32	0.035	40.4	Haemophilus influenzae R3021	Bacteria; Proteobacteria; Gammaproteobacteria

Set D:

Likely D. audax.	Possible Contam.	Contam. Code	Num. of Reads	Best D. audax. Ident	Best Ident.	Best Aln. Len	Best E-value	Best Bit Score	Closest Species Other Than D. audaxviator	Partial Classification
	*	a	1	82.5	93.1	29	4E-09	63.5	Aspergillus nidulans FGSC A4	Eukaryota; Fungi/Metazoa group; Fungi
	*	a	1	64.29	61.11	36	0.002	44.3	Apis mellifera	Eukaryota; Fungi/Metazoa group; Metazoa
	*	a	1	75	53.33	15	0	22.7	Vitis vinifera	Eukaryota; Viridiplantae; Streptophyta
	*	a	1	77.27	44.74	38	0.059	39.7	Ajellomyces capsulatus NAm1	Eukaryota; Fungi/Metazoa group; Fungi
	*	c	1	71.79	95	20	3E-04	45.1	Victivallis vadensis ATCC BAA-548	Bacteria; Chlamydiae/Verrucomicrobia; Lentisphaerae
	*	d	1	91.67	94.44	36	1E-11	72	Clostridium acetobutylicum ATCC 824	Bacteria; Firmicutes; Clostridia
	*	d	1	78.57	90.48	21	0.02	41.2	Bacillus halodurans C-125	Bacteria; Firmicutes; Bacilli
	*	d	1	87.5	88	25	2E-09	45.1	Thermoanaerobacter tengcongensis MB4	Bacteria; Firmicutes; Clostridia
	*	c	1	77.78	86.67	15	0.009	30.4	Methanocorpusculum labreanum Z	Archaea; Euryarchaeota; Methanomicrobia
	*	d	1	84.38	86.21	29	8E-07	55.8	Haemophilus influenzae 3655	Bacteria; Proteobacteria; Gammaproteobacteria
	*	c	1	72.97	79.41	34	8E-07	55.8	Clostridium cellulolyticum H10	Bacteria; Firmicutes; Clostridia
	*		1	94.12	77.27	22	0.007	42.7	Mycoplasma hyopneumoniae 7448	Bacteria; Firmicutes; Mollicutes
	*	d	1	84.62	75	24	0.003	43.9	Pelobacter propionicus DSM 2379	Bacteria; Proteobacteria; Deltaproteobacteria
	*	c	1	94.44	72	25	0.021	41.2	Syntrophobacter fumaroxidans MPOB	Bacteria; Proteobacteria; Deltaproteobacteria
	*		1	70.73	69.44	36	5E-07	56.6	Myxococcus xanthus DK 1622	Bacteria; Proteobacteria; Deltaproteobacteria

	*		2	74.07	69.23	26	0.059	39.7	Ruminococcus torques ATCC 27756	Bacteria; Firmicutes; Clostridia
	*	d	1	76.92	69.23	26	0.007	42.7	Bacillus clausii KSM-K16	Bacteria; Firmicutes; Bacilli
	*	c	1	77.5	67.5	40	1E-05	52	Acidobacteria bacterium Ellin345	Bacteria; Fibrobacteres/Acidobacteria group; Acidobacteria
	*	c	1	81.25	66.67	33	4E-04	47	Syntrophomonas wolfei str. Goettingen	Bacteria; Firmicutes; Clostridia
	*	d	1	91.67	64.71	34	0.001	45.4	Desulfovibrio vulgaris str. Hildenborough	Bacteria; Proteobacteria; Deltaproteobacteria
	*	d	1	57.89	52.78	36	0.046	40	Dehalococcoides sp. CBDB1	Bacteria; Chloroflexi; Dehalococcoidetes
	*		1	73.08	45	40	0.012	40	Bacteroides vulgatus ATCC 8482	Bacteria; Bacteroidetes/Chlorobi; Bacteroidetes

[a] eukaryotic or viral contamination
[b] human-associated
[c] strain sequenced at JGI
[d] very close relative sequenced at JGI (Genus-level or closer)

Set E:

Likely Contam.	Possible Contam.	Contam Code	Num. of Reads	Best Ident.	Best Aln. Len	Best E-value	Best Bit Score	Closest Species	Partial Classification
*		a	20	100	32	4E-13	76.6	Homo sapiens	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	3	100	21	0.001	45.1	Pan troglodytes	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	2	100	19	5E-08	50.1	Rattus norvegicus	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	81.82	11	0	22.3	Lymnaea stagnalis	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	61.11	18	0.062	30.8	Saimiriine herpesvirus 2	Viruses; dsDNA viruses, no RNA stage; Herpesviridae
*		a	4	59.46	37	0.078	39.3	Mus musculus	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	2	57.5	40	0.046	40	Bos taurus	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	3	54.84	31	0.099	38.9	Oryza sativa (japonica cultivar-group)	Eukaryota; Viridiplantae; Streptophyta

*		a	1	54.05	37	0.035	40.4	Xenopus laevis	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	51.35	37	0.045	40	Strongylocentrotus purpuratus	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	2	51.11	45	0.044	40	Tetraodon nigroviridis	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	47.5	40	0.02	41.2	Canis familiaris	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	45.83	24	0	22.7	Danio rerio	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	42.11	57	0.1	38.9	Macaca mulatta	Eukaryota; Fungi/Metazoa group; Metazoa
*		a	1	35.42	48	0.046	40	Sclerotinia sclerotiorum 1980	Eukaryota; Fungi/Metazoa group; Fungi
*		c	47	100	35	7E-14	79.3	Shewanella baltica OS195	Bacteria; Proteobacteria; Gammaproteobacteria
*		c	33	100	21	7E-05	49.3	Shewanella baltica OS223	Bacteria; Proteobacteria; Gammaproteobacteria
*		c	12	100	34	1E-12	75.1	Shewanella baltica OS185	Bacteria; Proteobacteria; Gammaproteobacteria
*		b	6	100	23	2E-04	47.8	Propionibacterium acnes KPA171202	Bacteria; Actinobacteria; Actinobacteria
*		d	4	100	23	5E-10	50.8	Shewanella oneidensis MR-1	Bacteria; Proteobacteria; Gammaproteobacteria
*		c	4	100	21	0.007	42.7	Desulfovibrio vulgaris str. DP4	Bacteria; Proteobacteria; Deltaproteobacteria
*		b/d	2	100	37	3E-15	84	Staphylococcus epidermidis ATCC 12228	Bacteria; Firmicutes; Bacilli
*		c	2	100	25	2E-07	57.8	Acidovorax sp. JS42	Bacteria; Proteobacteria; Betaproteobacteria
*		d	2	100	23	3E-04	47.4	Desulfovibrio vulgaris str. Hildenborough	Bacteria; Proteobacteria; Deltaproteobacteria
*		d	1	100	36	1E-12	75.1	Streptococcus thermophilus LMG 18311	Bacteria; Firmicutes; Bacilli
*		d	1	100	33	0.044	40	Pseudomonas putida GB-1	Bacteria; Proteobacteria; Gammaproteobacteria
*		c	1	100	32	9E-11	68.9	Shewanella baltica OS155	Bacteria; Proteobacteria; Gammaproteobacteria
*		d	1	100	30	6E-09	62.8	Pseudomonas stutzeri A1501	Bacteria; Proteobacteria; Gammaproteobacteria

*		c	1	100	27	1E-07	58.9	Syntrophomonas wolfei str. Goettingen	Bacteria; Firmicutes; Clostridia
*		c	1	96.67	30	7E-08	59.3	Thermoanaerobacter ethanolicus ATCC 33223	Bacteria; Firmicutes; Clostridia
*		c	2	94.12	17	0.099	35	Shewanella putrefaciens CN-32	Bacteria; Proteobacteria; Gammaproteobacteria
*		c	2	93.33	30	3E-09	63.9	Bradyrhizobium sp. BTAi1	Bacteria; Proteobacteria; Alphaproteobacteria
*		d	3	91.3	23	1E-04	48.9	Bradyrhizobium japonicum USDA 110	Bacteria; Proteobacteria; Alphaproteobacteria
*		d	1	91.3	23	8E-04	45.8	Clostridium botulinum A str. ATCC 3502	Bacteria; Firmicutes; Clostridia
	*		1	90.91	22	4E-05	43.5	Acinetobacter sp. ADP1	Bacteria; Proteobacteria; Gammaproteobacteria
	*		1	87.5	32	2E-07	57.8	Xanthomonas campestris str. ATCC 33913	Bacteria; Proteobacteria; Gammaproteobacteria
	*	d	1	84.38	32	9E-08	58.9	Polynucleobacter sp. QLW-P1DMWA-1	Bacteria; Proteobacteria; Betaproteobacteria
	*	c	2	83.33	30	2E-05	51.2	Nitrobacter winogradskyi Nb-255	Bacteria; Proteobacteria; Alphaproteobacteria
	*	d	1	81.25	16	0.009	34.7	Dehalococcoides sp. CBDB1	Bacteria; Chloroflexi; Dehalococcoidetes
	*	d	2	80.65	31	6E-07	56.2	Pseudomonas entomophila L48	Bacteria; Proteobacteria; Gammaproteobacteria
	*		1	75.76	33	1E-05	52	Symbiobacterium thermophilum IAM 14863	Bacteria; Firmicutes; Bacilli
	*		1	75.68	37	2E-08	61.2	Pyrococcus furiosus	Archaea; Euryarchaeota; Thermococci
	*	c	1	72.73	33	1E-06	55.1	Cenarchaeum symbiosum A	Archaea; Crenarchaeota; Thermoprotei
	*	c	1	71.43	28	0.015	41.6	Candidatus Methanoregula boonei 6A8	Archaea; Euryarchaeota; Methanomicrobia
	*	b	1	70	30	3E-05	50.8	Mycobacterium smegmatis str. MC2 155	Bacteria; Actinobacteria; Actinobacteria
	*	c	1	68.75	32	0.005	43.1	Thermoanaerobacter ethanolicus X514	Bacteria; Firmicutes; Clostridia

	*	c	1	66.67	30	0.099	38.9	Thiobacillus denitrificans ATCC 25259	Bacteria; Proteobacteria; Betaproteobacteria
	*	c	1	64.71	34	7E-06	52.8	Zymomonas mobilis subsp. mobilis ZM4	Bacteria; Proteobacteria; Alphaproteobacteria
	*	d	1	64.29	28	0.002	44.3	Bacillus halodurans C-125	Bacteria; Firmicutes; Bacilli
	*	c	1	63.33	30	0.004	43.5	Geobacter uraniumreducens Rf4	Bacteria; Proteobacteria; Deltaproteobacteria
	*		1	62.5	16	0	24.6	Agrobacterium tumefaciens str. C58	Bacteria; Proteobacteria; Alphaproteobacteria
	*	d	1	62.16	37	1E-04	48.9	Vibrio cholerae 2740-80	Bacteria; Proteobacteria; Gammaproteobacteria
	*	c	1	61.76	34	2E-04	47.8	Clostridium cellulolyticum H10	Bacteria; Firmicutes; Clostridia
	*	d	1	60.87	23	0.02	41.2	Pseudomonas aeruginosa C3719	Bacteria; Proteobacteria; Gammaproteobacteria
	*	c	1	59.38	32	5E-04	46.6	Opitutaceae bacterium TAV2	Bacteria; Chlamydiae/Verrucomicrobia; Verrucomicrobia
	*	c	2	58.33	24	0.06	39.7	Methanococcoides burtonii DSM 6242	Archaea; Euryarchaeota; Methanomicrobia
	*		1	56.67	30	0.009	42.4	Algoriphagus sp. PR1	Bacteria; Bacteroidetes/Chlorobi; Bacteroidetes
	*	d	1	51.11	45	0.078	39.3	Synechococcus sp. PCC 7002	Bacteria; Cyanobacteria; Chroococcales
	*	c	1	50	46	0.077	39.3	Methylobacterium sp. 4-46	Bacteria; Proteobacteria; Alphaproteobacteria
	*	d	1	48.72	39	0.059	39.7	Burkholderia pseudomallei 668	Bacteria; Proteobacteria; Betaproteobacteria
	*		1	47.17	53	0.079	39.3	Thermus thermophilus HB27	Bacteria; Deinococcus-Thermus; Deinococci
	*		1	47.06	34	0.078	39.3	Sagittula stellata E-37	Bacteria; Proteobacteria; Alphaproteobacteria
	*	d	1	44.44	18	0	23.1	Xanthobacter autotrophicus Py2	Bacteria; Proteobacteria; Alphaproteobacteria
	*	c	1	42.5	40	0.045	40	Halorhodospira halophila SL1	Bacteria; Proteobacteria; Gammaproteobacteria
	*		1	42.11	19	0	21.2	Streptomyces aculeolatus	Bacteria; Actinobacteria; Actinobacteria
	*	c	1	37.21	43	0	21.2	Methylobacterium extorquens	Bacteria; Proteobacteria;

								PA1	Alphaproteobacteria
	*	c	1	35.59	59	0.009	42.4	Parvibaculum lavamentivorans DS-1	Bacteria; Proteobacteria; Alphaproteobacteria

[a] eukaryotic or viral contamination
[b] human-associated
[c] strain sequenced at JGI
[d] very close relative sequenced at JGI (Genus-level or closer)

Table S7(a,b). Single base substitutions (SNPs) found in Sanger reads.

Examination of the Sanger reads permitted an estimate of the degree of genetic variation in the *D. audaxviator* population. Two or more reads corroborate the observation of only 32 positions with single nucleotide polymorphisms ("SNP") in the population in the entire 2.35 Mbp genome. Twelve of the SNPs occur within the same ORFan gene (Daud1974). Several other genes possess two SNPs, yielding a total of 11 genes that exhibit a SNP. Comparison with the unusually homogeneous *Leptospirillum* group II population (*54*) with a similar read depth, showed 60-fold less polymorphism and means that the reads came largely from a dominant near-clonal strain. Regrettably, without many orders of magnitude more sequencing it is impossible to access the polymorphism that might be present in any rarer sub-types. This insufficient sequence information for rarer sub-types prohibits estimation of the recombination rate and the effective population size, without which we cannot determine of the number of generations that have occurred since founding. We additionally do not have enough SNPs to ascertain whether genes are under purifying, neutral, or positive selection, with the possible exception of Daud1974 which has 8 synonymous and 3 non-synonymous SNPs, suggesting purifying selection for this gene.

Since a pronounced excess of cells were sampled (~1.8×10^{11} cells) compared with the number of Sanger reads corresponding to the assembly (~2.9×10^{4} reads), we may expect that each read came from a different cell and may therefore be used to ascertain nucleotide polymorphism. We examined the Sanger reads that strongly matched to the assembly (sets B and C1 from Table S6b) for polymorphism with respect to the consensus. We choose to investigate several approaches for identifying the single base substitutions and found a very low number, even with the least stringent parameters, precluding discrimination of sub-populations in the fashion of Tyson, *et al*.(*54*). We chose to not attempt to identify SNPs from the 454 sequence data because such data, to our knowledge, does not yet offer reliable quality scores, and additonally suffers from artifacts primarily as a consequence of homopolymeric stretches of sequence. While the Sanger reads, when including low-quality calls, gave an average depth of coverage of 11.356 X we found using

only the higher-quality calls did not greatly reduce the average depth of coverage (the lowest was 8.033 X) and therefore felt justified in focusing on the more trustworthy data. The five approaches we used to identify the single-base substitutions were (from most stringent to most permissive) 1. a minimum base call quality (Phred score(*17, 18*)) of at least 15 and a minimum of two identical observations of the mutation, 2. allowing only a single observation, but requiring a Phred score of at least 25, but ignoring the first 50 bases of each read (based on an increased rate of non-matching bases even for higher Phred scores at the beginning of the reads), 3. a Phred score of at least 25 and using the entire read, 4. a Phred score of at least 20 and ignoring the first 50 bases, and 5. a Phred score of at least 20 and using the entire read. The number and location of SNPs identified by these five approaches are given in Table S7a, with the rate calculated from the genome length of 2,349,476 bp.

Table S7a. SNP statistics.

Category	Phred≥15 +duplicates	Phred≥25 +skip50	Phred≥25	Phred≥20 +skip50	Phred≥20
Intergenic	7	25	31	87	98
Pseudogenes	0	5	5	10	10
RNA	0	2	2	8	8
Synonymous in protein-coding genes	11	59	68	201	214
Non-synonymous in protein-coding genes	14	91	102	316	339
Total SNPs	32	182	208	622	669
Average rate of SNPs	0.0014%	0.0077%	0.0089%	0.026%	0.028%
Average depth of coverage	9.567 X	8.033 X	8.106 X	8.800 X	8.895 X

Of interest were the numerous synonymous (and repeatedly observed) mutations within the ORFan gene Daud1974 (which may have been recently acquired), as well as the non-synonymous mutation in the H^+ translocating pyrophosphatase (gene Daud0308), which appears to have been horizonatally acquired from archaea. Such ORFan and horizontally transferred genes may either be adapting to the new host, or may be subject to less selective pressure than more established genes. Additionally, many of the SNPs found with the less stringent parameters lay within transposons, possibly a consequence of reduced selective pressure on such regions of the genome, although some fraction of the inferred SNPs within transponsons may instead be attributable to the challenge in perfectly assigning such reads to an assembly that contains multiple identical or near-identical regions.

Table S7b. Reliable SNPs.

The reliable SNP identifications (those with multiple observations) are given. If the mutation occurs within a protein-coding gene, the effect on the amino acid sequence of the mutation is indicated. The base and the position are given with respect to the positive strand in the assembly, whereas the codon change in protein-coding genes is with respect to the coding strand. Intergenic SNPs are indicated with "N/A" for the Gene ID. Also reported is the number of reads containing the observation, the overall depth and the depth (at a Phred score ≥ 15) at the position, and the Phred scores of the calls. Note that the depth is reported with respect to the Sanger reads only, whereas the consensus sequence is derived from both Sanger and 454 sequence, so low-Sanger-depth positions with mutations can occur.

Consensus base	Mutation	Depth at position	Depth Phred≥15	Number of obs.	Phred scores	Codon change	AA change	Gene ID	Notes and gene description
a	t	7	7	2	15,23	N/A	N/A	N/A	N/A
g	a	8	7	2	50,51	ggt:gtt	G:V	Daud0308	COG3808 V-type H(+)-translocating pyrophosphatase
g	c	10	8	2	20,15	cag:cag	Q:Q	Daud0481	COG2239, Mg/Co/Ni transporter MgtE
g	a	10	10	2	30,55	gag:tag	E:*	Daud0551	gltA COG493 NADPH-dependent glutamate synthase beta chain
g	a	4	4	2	22,18	aga:ata	R:I	Daud0679	Response regulator PF01966 Metal-dependent phosphohydrolase, HD region
g	a	6	6	2	22,19	aga:ata	R:I	Daud0679	Response regulator PF01966 Metal-dependent phosphohydrolase, HD region
a	c	18	16	2	23,16	N/A	N/A	N/A	N/A
t	a	9	9	2	51,16	cag:ctg	Q:L	Daud1430	COG3879 Uncharacterized protein conserved in bacteria
c	g	13	11	3	21,29,27	N/A	N/A	N/A	N/A
c	a	12	10	2	16,16	gtg:ttg	V:L	Daud1587	PF03193 GTPase EngC, ribosome SSU dependent
c	g	5	4	2	43,21	ccg:ccc	P:P	Daud1974	ORFan
g	a	5	4	3	52,31,52	ctc:ctt	L:L	Daud1974	ORFan
a	g	5	5	2	44,42	ggt:ggc	G:G	Daud1974	ORFan
t	c	5	5	2	41,50	cca:ccg	P:P	Daud1974	ORFan
g	c	5	4	2	41,38	gtc:gtg	V:V	Daud1974	ORFan

t	g	5	2	2	35,44	agg:cgg	R:R	Daud1974	ORFan
t	c	5	3	2	43,40	gca:gcg	A:A	Daud1974	ORFan
a	g	4	3	2	29,33	tct:ttc	S:F	Daud1974	ORFan; 2078100 and 2078101 affect same codon
g	a	4	3	2	29,23	tct:ttc	S:F	Daud1974	ORFan; 2078100 and 2078101 affect same codon
a	t	4	3	2	27,43	ttg:atg	L:M	Daud1974	ORFan
g	a	4	4	2	27,37	tcc:tct	S:S	Daud1974	ORFan
c	t	4	3	2	26,43	ggt:agt	G:S	Daud1974	ORFan
g	c	4	3	2	38,38	N/A	N/A	N/A	N/A
c	t	4	3	2	38,43	N/A	N/A	N/A	N/A
c	t	4	3	2	29,37	N/A	N/A	N/A	N/A
c	a	9	7	2	23,27	gac:gat	D:D	Daud1982	COG747, ABC-type dipeptide transport system, periplasmic component
a	g	14	12	2	33,17	aac:cac	N:H	Daud1989	hypothetical protein
c	t	15	11	2	50,23	tac:taa	Y:*	Daud1989	hypothetical protein
c	a	14	13	2	23,16	aac:aat	N:N	Daud2001	ORFan
t	c	14	13	2	29,22	ttg:gtg	L:V	Daud2001	ORFan
c	g	20	18	2	15,16	N/A	N/A	N/A	N/A
c	t	14	12	2	48,51	gcc:acc	A:T	Daud2092	rfe COG472 Glycosyl transferase, family 4

Table S8. Functional RNA genes.

Genes coding for functional RNA. Of note are the duplication of tRNA for Met (bacterial start codon) and insertion of tRNA for Ala and Ile into second rRNA operon. Also of interest is the unusual "Ornate, Large, Extremophilic" RNA ("OLE" RNA) (*55*). A complete set of tRNA genes is present, including SeC. The two SSU rRNA (16S) genes are 100% identical to one another.

Gene	Name	Description	Operon	Strand	Start	Len	Notes
DaudR0057	oleRNA	Ornate large extremophilic (OLE) RNA		+	1074679	582	
DaudR0058	ffs	4.5S RNA component of the SRP		+	38628	269	
DaudR0015	ssrA	tmRNA (transfer messenger RNA or 10Sa RNA)		+	275529	354	
DaudR0018	rnpB	RnpB RNA: catalytic subunit of RNase P		+	536618	358	
DaudR0025	ydaO/yuaA	ydaO/yuaA element as predicted by Rfam (RF00379)		+	1122618	138	
DaudR0016	yybP-ykoY	yybP-ykoY element as predicted by Rfam (RF00080)		+	348454	124	
DaudR0047	yybP-ykoY	yybP-ykoY element as predicted by Rfam (RF00080)		-	1963552	129	
DaudR0030	rrfB	5S ribosomal RNA	RIB2	-	1439006	114	
DaudR0031	rrlB	23S ribosomal RNA	RIB2	-	1439160	3357	
DaudR0032	tRNA-Ala	tRNA with anticodon TGC for Ala	RIB2	-	1442590	75	
DaudR0033	tRNA-Ile	tRNA with anticodon GAT for Ile	RIB2	-	1442802	77	
DaudR0034	rrsB	16S ribosomal RNA	RIB2	-	1442921	1692	
DaudR0007	rrsA	16S ribosomal RNA	RIB1	+	142718	1692	
DaudR0008	rrlA	23S ribosomal RNA	RIB1	+	144532	3357	
DaudR0009	rrfA	5S ribosomal RNA	RIB1	+	147929	114	
DaudR0029	tRNA-Leu	tRNA with anticodon GAG for Leu		+	1307354	85	
DaudR0035	tRNA-Val	tRNA with anticodon GAC for Val		-	1482017	75	
DaudR0036	tRNA-Lys	tRNA with anticodon TTT for Lys		-	1569502	76	
DaudR0037	tRNA-Gln	tRNA with anticodon TTG for Gln		-	1569623	76	
DaudR0038	tRNA-His	tRNA with anticodon GTG for His		-	1569707	77	

DaudR0039	tRNA-Leu	tRNA with anticodon TAG for Leu		+	1570390	85	
DaudR0040	tRNA-Arg	tRNA with anticodon TCT for Arg		-	1581835	76	
DaudR0041	tRNA-Gly	tRNA with anticodon TCC for Gly		-	1584258	74	
DaudR0042	tRNA-Arg	tRNA with anticodon CCG for Arg		-	1605731	75	
DaudR0043	tRNA-Glu	tRNA with anticodon CTC for Glu		-	1612824	76	
DaudR0044	tRNA-Gln	tRNA with anticodon CTG for Gln		-	1612917	74	
DaudR0045	tRNA-Thr	tRNA with anticodon CGT for Thr		+	1935182	75	
DaudR0046	tRNA-Val	tRNA with anticodon CAC for Val		-	1958919	75	
DaudR0048	tRNA-Ala	tRNA with anticodon GGC for Ala		-	1967347	75	
DaudR0049	tRNA-Leu	tRNA with anticodon TAA for Leu		-	2020979	88	
DaudR0050	tRNA-Cys	tRNA with anticodon GCA for Cys		-	2058346	76	
DaudR0051	tRNA-Asp	tRNA with anticodon GTC for Asp		+	2077429	78	
DaudR0052	tRNA-Phe	tRNA with anticodon GAA for Phe		+	2077515	76	
DaudR0053	tRNA-Gly	tRNA with anticodon GCC for Gly		+	2077605	75	
DaudR0054	tRNA-Ala	tRNA with anticodon CGC for Ala		+	2093802	76	
DaudR0055	tRNA-Gly	tRNA with anticodon CCC for Gly		-	2322214	75	
DaudR0056	tRNA-Glu	tRNA with anticodon TTC for Glu		-	2322550	76	
DaudR0001	tRNA-Ser	tRNA with anticodon TGA for Ser		+	16098	91	
DaudR0002	tRNA-Ser	tRNA with anticodon GCT for Ser		+	16321	95	
DaudR0003	tRNA-Arg	tRNA with anticodon ACG for Arg		+	16551	76	
DaudR0004	tRNA-Arg	tRNA with anticodon CCT for Arg		+	16805	77	
DaudR0005	tRNA-Ser	tRNA with anticodon CGA for Ser		+	17472	96	
DaudR0006	tRNA-Ser	tRNA with anticodon GGA for Ser		+	38521	90	
DaudR0010	tRNA-Asn	tRNA with anticodon GTT for Asn		+	191655	75	
DaudR0011	tRNA-Thr	tRNA with anticodon TGT for Thr		+	220154	75	
DaudR0012	tRNA-Met	tRNA with anticodon CAT for Met		+	220296	76	duplicate
DaudR0013	tRNA-Thr	tRNA with anticodon GGT for Thr		+	220386	76	
DaudR0014	tRNA-Met	tRNA with anticodon CAT for Met		+	220469	77	duplicate

DaudR0017	tRNA-Val	tRNA with anticodon TAC for Val		+	533273	75	
DaudR0019	tRNA-Leu	tRNA with anticodon CAG for Leu		-	909888	87	
DaudR0020	tRNA-Lys	tRNA with anticodon CTT for Lys		-	921840	77	
DaudR0021	tRNA-Tyr	tRNA with anticodon GTA for Tyr		-	921925	85	
DaudR0022	tRNA-Pro	tRNA with anticodon TGG for Pro		-	925167	78	
DaudR0023	tRNA-SeC(p)	tRNA with anticodon TCA for SeC (selenocys)		-	1006319	90	
DaudR0024	tRNA-Leu	tRNA with anticodon CAA for Leu		-	1116444	88	
DaudR0026	tRNA-Pro	tRNA with anticodon GGG for Pro		+	1134210	78	
DaudR0027	tRNA-Pro	tRNA with anticodon CGG for Pro		+	1169824	78	
DaudR0028	tRNA-Trp	tRNA with anticodon CCA for Trp		+	1288721	76	

Table S9(a,b,c). Potential genomic determinants of hyperthermophily.

We investigated the presence of 58 COGs determined by Makarova, et al. (*56*) as possibly playing a role in hyperthermophily based on their distribution in extremophilic Archaea and the bacteria *Thermoanaerobacter tengcongensis*, *Thermus thermophilus*, *Thermotoga maritima*, and *Aquifex aeolicus*. The signature hyperthermophile gene "reverse gyrase" (*57*), which has an N-terminal helicase domain and a C-terminal topoisomerase I domain, is not found as a complete gene in *D. audaxviator*, but may not be absolutely essential for hyperthermophily (*58*). *D. audaxviator* does possess a gene that is similar to the topoisomerase I domain of *Thermoanaerobacter tengcongensis* reverse gyrase (Table S9b) as well as helicase encoding genes (although none closely resemble the helicase domain of *T. tengcongensis* reverse gyrase).

Table S9(a). Presence/absence of potential hyperthermophile COGs in relevant organisms.

Presence and absence of 50 hyperthermophilic COGs in *D. audaxviator*, and other relevant bacteria (hyperthermophilic archaea not included for clarity). Since the study by Makarova, et al. (*56*) was a "guilt by association" study, some genes are undoubtedly incorrectly implicated as having a direct role in hyperthermophily, when in fact they may be playing other roles (e.g. CRISPR-associated genes or Carbon monoxide dehydrogenase, some of which are also reported in Table S10 as horizontal transfers between archaea and *D. audaxviator*). Other genes may be necessary for hyperthermophily, but only because they are putatively thermostable

forms or otherwise enzymatically functional at high temperature of essential proteins (e.g. the xenologous replacement of fructose-1,6,bisphosphatase in the gluconeogenic pathway).

Key

+	gene present by match to COG
-	gene absent
?	incomplete genome, absence of gene indeterminate
B0	homolog detected by BLASTp to *T.tengcongensis* representative (TTE1745)
B1	homolog detected in *D.audaxviator* by BLASTp to *D.reducens* representative (VIMSS1188125)
B2	homolog detected in *D.audaxviator* by BLASTp to *P.thermopropionicum* representative (VIMSS1359824)
B3	homolog detected in *M.thermoacetica* by BLASTp to *P.thermopropionicum* representative (VIMSS1360512)
B4	homolog detected in *D.audaxviator* by BLASTp to *P.thermopropionicum* representative (VIMSS1359650)

Species Abbreviations

Daudax	*Desulforudis audaxviator*	Ctetani	*Clostridium tetani* E88
Ptherm	*Pelotomaculum thermopropionicum* SI	Cacet	*Clostridium acetobutylicum* ATCC824
Dred	*Desulfotomaculum reducens* MI-1	Gkaus	*Geobacillus kaustophilus* HTA426
Chyd	*Carboxydothermus hydrogenoformans* Z-2901	Dethen	*Dehalococcoides ethenogenes* 195
Mtherm	*Moorella thermoacetica* ATCC39073	Gmetal	*Geobacter metallireducens* GS-15
DhafDCB2	*Desulfitobacterium hafniense* DCB-2	Gsulf	*Geobacter sulfurreducens* PCA
DhafY51	*Desulfitobacterium hafniense* Y51	Tmari	*Thermotoga maritima*
Stherm	*Symbiobacterium thermophilum* IAM 14863	Ttherm	*Thermus thermophilus* HB8
Tteng	*Thermoanaerobacter tengcongensis* MB4T	Aaeol	*Aquifex aeolicus* VF5

	D	P	D	C	M	D	D	S	T	C	C	G	D	G	G	T	T	A		
	a	t	r	h	t	h	h	t	t	a	t	k	e	m	s	m	t	a		
	u	h	e	y	h	a	a	h	e	c	e	a	t	e	u	a	h	e		
	d	e	d	d	e	f	f	e	n	e	t	u	h	t	l	r	e	o		
	a	r			r	D	Y	r	g	t	a	s	e	a	f	i	r	l		
	x	m			m	C	5	m			n		n	l			m			
						B	1				i									
COG						2													Gene Name	Notes
1110	B 0	B 0	B 0	B 0	B 0	B 0	B 0	-	+	B 0	B 0	B 0	B 0	-	-	+	+	+	N-term: helicase, C-term: DNA Topo I (B0: C-term matches)	
2250	+	+	+	+	+	-	+	-	+	-	-	-	-	+	-	+	+	-	Related to C-terminal domain of eukaryotic chaperone, SACSIN	
1980	+	+	?	+	+	-	-	-	+	-	-	-	+	-	-	-	+	+	Archaeal fructose 1,6-bisphosphatase	us Daud1840 ppa (inorganic pyrophosphatase, not H+ translocating type)
1688	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	Predicted to be involved in DNA repair (RAMP superfamily)	
1618	-	?	?	-	-	+	+	-	-	-	-	-	-	-	-	+	-	+	Predicted nucleotide kinase	
1313	+	+	?	+	+	-	-	-	+	+	+	-	+	+	+	+	-	+	Fe-S protein PflX, homolog of pyruvate formate lyase activating proteins	
1468	+	+	+	+	+	+	+	-	+	-	+	-	-	+	-	+	+	+	RecB family exonuclease	CAS 1,2B (CRISPR)
3635	+	+	?	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	Predicted phosphoglycerate mutase, AP superfamily	us pfkA,pyk,DedA, ds sodB
1318	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	Predicted transcriptional regulators	
1350	+	+	?	+	-	+	+	-	-	-	-	-	+	+	+	+	-	+	Predicted alternative tryptophan synthase beta-subunit (paralog of TrpB)	ds lysA
1353	+	+	+	+	-	-	-	-	+	-	-	-	-	-	-	+	+	+	Predicted hydrolase of the HD superfamily (permuted catalytic motifs)	CAS 2B (CRISPR)

1144	B l	?	+	+	+	+	+	-	+	+	+	-	+	+	+	+	-	+	Pyruvate:ferredoxin oxidoreductase, delta subunit	COG1145 in Daud
1578	+	?	?	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	Predicted acyl binding protein	
1149	+	+	+	-	+	+	+	-	+	-	-	-	-	-	-	+	-	-	MinD superfamily P-loop ATPase containing an inserted ferredoxin domain	
1237	+	?	+	-	+	+	+	-	+	+	+	-	-	+	+	+	-	-	Metal-dependent hydrolases of the beta-lactamase superfamily II	
1583	+	+	+	+	+	+	+	-	+	-	+	-	-	-	-	+	-	+	Predicted to be involved in DNA repair (RAMP superfamily)	CAS 2B (CRISPR)
1568	-	?	?	+	-	-	-	-	+	-	-	-	-	-	-	-	+	+	Predicted methyltransferases	
1906	-	?	?	-	-	+	+	-	+	-	-	-	-	+	+	+	-	-	Predicted membrane transporter	
2152	+	?	?	-	-	-	-	+	+	-	-	+	-	-	-	+	+	+	Predicted glycosylase	
1336	+	+	?	-	-	-	-	-	+	-	-	-	-	-	-	+	+	+	Predicted to be involved in DNA repair (RAMP superfamily)	CAS 2B (CRISPR)
1148	+	+	+	+	+	-	-	-	-	-	-	-	-	+	+	-	-	-	Heterodisulfide reductase, subunit A and related polyferredoxins	hdrA
2516	-	?	?	+	+	-	-	-	-	-	-	-	-	-	-	+	-	+	Biotin synthase-related enzyme	
1856	-	?	?	+	+	-	-	-	-	-	-	-	-	-	-	+	-	+	Biotin synthase-related enzyme	
1604	+	+	?	-	-	-	-	-	+	-	-	-	-	-	-	+	+	+	Predicted to be involved in DNA repair (RAMP superfamily)	CAS 2B (CRISPR)
2406	-	?	?	-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	Protein distantly related to bacterial ferritins	
1059	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	Thermostable 8-oxoguanine DNA glycosylase	
1769	+	+	?	-	-	-	-	-	+	-	-	-	-	-	-	+	+	+	Predicted to be involved in DNA repair (RAMP superfamily)	CAS 2B (CRISPR)
1542	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	Contains coiled-coil domains	
1367	+	?	?	-	-	-	-	-	+	-	-	-	-	-	-	+	+	+	Predicted to be involved in DNA repair (RAMP superfamily)	CAS 2B (CRISPR)
2000	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	Predicted Fe-S protein, hydrogenase subunit	
1913	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	Predicted Zn-dependent proteases	

1857	+	?	?	-	-	-	-	-	+	-	+	-	-	-	-	-	-	+	Predicted to be involved in DNA repair	CAS 1,2B (CRISPR)
2112	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	Predicted Ser/Thr protein kinase	
1423	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	ATP-dependent DNA ligase, homolog of eukaryotic ligase III	
1458	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	Predicted DNA-binding protein containing PIN domain	
1992	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	Uncharacterized conserved protein	
1839	-	?	?	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	Uncharacterized conserved protein	
2044	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	Predicted peroxiredoxins	
1743	+	+	?	-	+	+	-	-	-	-	-	-	-	-	-	+	+	-	Adenine-specific DNA methylase containing a Zn-ribbon	
1483	B2	+	?	-	-	+	-	-	-	-	-	-	-	-	-	+	+	-	Predicted ATPase (AAA+ superfamily)	2 genes in D.audax.
1672	-	+	?	+	B3	+	+	-	-	-	-	-	-	+	-	+	-	-	Predicted ATPase (AAA+ superfamily)	
2245	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	Predicted membrane protein	
1337	+	?	+	+	-	-	-	-	-	-	-	-	-	-	-	+	+	-	Predicted to be involved in DNA repair (RAMP superfamily)	3 genes in CAS 2A (CRISPR)
1150	+	?	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	+	Heterodisulfide reductase, subunit C	hdrC
1517	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	Predicted to be involved in DNA repair	
4089	B4	+	+	+	+	-	-	-	+	-	-	-	-	-	-	-	-	-	Predicted membrane protein	
1431	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	Contains piwi/argonaute domain homolog, possible role in translation	
4353	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	Uncharacterized conserved protein	
1851	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	Uncharacterized conserved protein	
1421	-	?	?	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	Predicted to be involved in DNA repair	
1895	-	?	+	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	Related to C-terminal domain of eukaryotic chaperone, SACSIN	
3374	-	?	?	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	Predicted membrane protein	
3044	-	+	?	+	-	-	-	-	+	-	-	+	-	-	-	-	+	-	Predicted ATPase of the ABC class	

4756	-	?	?	-	-	-	-	+	+	-	-	-	-	-	-	-	-	Predicted cation transporter	
1152	+	?	?	-	-	-	-	-	?	-	-	-	-	-	-	-	-	CO dehydrogenase/acetyl-CoA synthase alpha subunit	
2043	-	?	+	-	+	+	+	+	+	-	-	-	-	-	-	-	-	Distant homolog of plant chalcone/stilbene synthases, possible acyltransferase	
4911	-	?	?	-	-	-	-	+	-	-	-	-	-	-	-	+	+	Uncharacterized conserved protein	
1751	+	?	+	+	+	-	-	-	-	-	+	-	-	-	-	+	-	-	Uncharacterized conserved protein	us spoIID,spoIIID,mreB; ds hyp.pro., COG4113 (nuc.acid.binding,PIN), Peptidase S8 and S53 with S-layer homology region

Table S9(b). Detection of *Thermoanaerobacter tengcongensis* Reverse gyrase homologs with tBLASTn.

Table S9(b)-i. tBLASTn search with full-length *Thermoanaerobacter tengcongensis* MB4T reverse gyrase gene, N-terminal Helicase + C-terminal Topoisomerase I domains (residues 1-1117)

Bacteria?	Species	Bit Score	E-value	Notes
*	Thermoanaerobacter tengcongensis MB4T	1996	0	
*	Thermotoga maritima	1058	0	
	Pyrococcus abyssi	652	0	
	Pyrococcus furiosus DSM 3638	635	1.00E-180	
*	Aquifex aeolicus	618	1.00E-175	
	Archaeoglobus fulgidus	600	1.00E-170	
	Sulfolobus solfataricus	579	1.00E-163	
	Sulfolobus tokodaii	578	1.00E-163	
	Sulfolobus acidocaldarius DSM 639	487	1.00E-154	
	Pyrobaculum aerophilum	538	1.00E-151	
	Aeropyrum pernix	520	1.00E-145	
*	Thermus thermophilus HB8	491	1.00E-137	
	Pyrococcus horikoshii	462	1.00E-128	

	Thermococcus kodakaraensis KOD1	457	1.00E-126	
	Methanocaldococcus jannaschii	437	1.00E-121	
	Nanoarchaeum equitans Kin4-M	425	1.00E-117	
	Methanopyrus kandleri AV19	272	4.00E-71	(drops with just C-term)
*	Carboxydothermus hydrogenoformans Z-2901	214	1.00E-53	
*	Clostridium tetani E88	211	8.00E-53	
*	Clostridium acetobutylicum ATCC824	208	6.00E-52	
*	Desulfotomaculum reducens MI-1	201	8.00E-50	
*	Desulforudis audaxviator	197	1.00E-48	
*	Geobacillus kaustophilus HTA426	197	1.00E-48	
*	Pelotomaculum thermopropionicum SI	196	4.00E-48	
*	Dehalococcoides ethenogenes 195	195	7.00E-48	
*	Desulfitobacterium hafniense DCB-2	193	2.00E-47	
*	Desulfitobacterium hafniense Y51	193	2.00E-47	

Table S9(b)-ii. tBLASTn search with *Thermoanaerobacter tengcongensis* MB4T reverse gyrase gene C-terminal: residues 553-1117 (residues selected by match to *D. audaxviator*) Topoisomerase I domain

Bacteria?	Species	Bit Score	E-value	Notes
*	Thermoanaerobacter tengcongensis MB4T	1069	0	
*	Thermotoga maritima	629	1E-179	
	Pyrococcus abyssi	437	1E-121	
	Nanoarchaeum equitans Kin4-M	422	1E-116	(score drops for full length)
	Pyrococcus furiosus DSM 3638	420	1E-116	
	Archaeoglobus fulgidus	416	1E-114	
	Sulfolobus tokodaii	406	1E-112	
	Sulfolobus acidocaldarius DSM 639	402	1E-110	
*	Aquifex aeolicus	397	1E-109	
	Sulfolobus solfataricus	394	1E-108	

	Aeropyrum pernix	349	1E-94	
	Pyrobaculum aerophilum	338	2E-91	
*	Thermus thermophilus HB8	321	4E-86	
	Thermococcus kodakaraensis KOD1	252	2E-65	
	Pyrococcus horikoshii	246	2E-63	
	Methanocaldococcus jannaschii	218	5E-55	
*	Carboxydothermus hydrogenoformans Z-2901	214	5E-54	
*	Clostridium tetani E88	211	3E-53	
*	Clostridium acetobutylicum ATCC824	208	3E-52	
*	Desulfotomaculum reducens MI-1	201	4E-50	
*	Desulforudis audaxviator	197	5E-49	
*	Geobacillus kaustophilus HTA426	197	6E-49	
*	Pelotomaculum thermopropionicum SI	195	2E-48	
*	Dehalococcoides ethenogenes 195	195	3E-48	
*	Desulfitobacterium hafniense DCB-2	193	9E-48	
*	Desulfitobacterium hafniense Y51	193	9E-48	

Table S9(b)-iii. tBLASTn search with *Thermoanaerobacter tengcongensis* MB4T reverse gyrase gene N-terminal: residues 1-552, Helicase domain (does not match to *D. audaxviator*)

Bacteria?	Species	Bit Score	E-value	Notes
*	Thermoanaerobacter tengcongensis MB4T	931	0	
*	Thermotoga maritima	439	1E-122	
	Nanoarchaeum equitans Kin4-M	263	1E-68	(score drops for full length)
	Pyrococcus horikoshii	236	2E-60	
	Pyrococcus furiosus DSM 3638	234	4E-60	
*	Aquifex aeolicus	233	8E-60	
	Pyrococcus abyssi	232	2E-59	

	Methanocaldococcus jannaschii	231	3E-59	
	Thermococcus kodakaraensis KOD1	224	5E-57	
	Pyrobaculum aerophilum	220	7E-56	
	Sulfolobus solfataricus	202	2E-50	
	Aeropyrum pernix	197	5E-49	
	Archaeoglobus fulgidus	197	5E-49	
	Sulfolobus tokodaii	192	3E-47	
*	Thermus thermophilus HB8	182	2E-44	
	Sulfolobus acidocaldarius DSM 639	142	6E-39	
	Methanopyrus kandleri AV19	138	5E-31	

Table S9(c). Putative hyperthermophile COGs (from Makarova, et al.) present in *D. audaxviator*.

Gene	Name	Description	Operon	Len	CH id	CH species	Notes
Daud0600	topA	COG550 (not COG1110): Topoisomerase IA		712	62.8	D. reducens	Related to T.tengcongensis Reverse gyrase (C-terminal topoisomerase domain)
Daud0133		COG2250: Related to C-terminal domain of eukaryotic chaperone, SACSIN		130	50.79	P. thermopropionicum	near 23S
Daud1839	fbp	COG1980: Archaeal fructose 1,6-bisphosphatase		371	79.12	P. thermopropionicum	us Daud1840 (inorganic pyrophosphatase, not H+ translocating type)
Daud1072	pflX	COG1313: Fe-S protein PflX, homolog of pyruvate formate lyase activating proteins		307	59.59	P. thermopropionicum	
Daud1287	cas4	COG1468: RecB family exonuclease	CAS1	191	34.76	A. fulgidus	
Daud1812	cas4	COG1468: RecB family exonuclease	CAS2B	179	66.47	T. tengcongensis	
Daud1058	apgM	COG3635: Predicted phosphoglycerate mutase, AP superfamily		406	59.2	P. thermopropionicum	
Daud1205	trpB	COG1350: Predicted alternative tryptophan synthase beta-subunit (paralog of TrpB)		452	69.47	P. thermopropionicum	

Daud1823	crm2 (cmr?)	COG1353: Predicted hydrolase of the HD superfamily (permuted catalytic motifs)	CAS2B	608	72.12	P. thermopropionicum	
Daud0910		COG1145 (not COG1144): Pyruvate: ferredoxin oxidoreductase, delta subunit		136	55.3	D. reducens	
Daud1372		COG1578: Predicted acyl-binding protein	TPT6	276	35.71	P. abyssi	us mutS
Daud1537		COG1149: MinD superfamily P-loop ATPase containing an inserted ferredoxin domain		288	53.85	P. carbinolicus	
Daud1155		COG1237: Metal-dependent hydrolases of the beta-lactamase superfamily II		283	45.99	G. sulfurreducens	ds DNA polIV (family X) Daud1154 which is most similar to T. thermophilus HB8
Daud1817	cas6	COG1583: Predicted to be involved in DNA repair (RAMP superfamily)	CAS2B	265	68.83	P. thermopropionicum	
Daud0864		COG2152: Predicted glycosylase		490	50	Pirellula sp. 1	
Daud1820	cmr4	COG1336: Predicted to be involved in DNA repair (RAMP superfamily)	CAS2B	289	72.05	P. thermopropionicum	
Daud1884	hdrA/qmoA	COG1148: Heterodisulfide reductase, subunit A and related polyferredoxins	SR8	420	58.31	C. chlorochromatii	
Daud0092	hdrA	COG1148: Heterodisulfide reductase, subunit A and related polyferredoxins	SR1	1014	55.77	D. reducens	
Daud0565	hdrA	COG1148: Heterodisulfide reductase, subunit A and related polyferredoxins	SR3	662	68.58	C. hydrogenoformans	
Daud0877	hdrA-like	COG1148: Heterodisulfide reductase, subunit A and related polyferredoxins	CODH2	996	45.08	D. reducens	part of CODH?
Daud1818	cmr6	COG1604: Predicted to be involved in DNA repair (RAMP superfamily)	CAS2B	324	73.15	P. thermopropionicum	
Daud1822	cmr3	COG1769: Predicted to be involved in DNA repair (RAMP superfamily)	CAS2B	392	73.05	P. thermopropionicum	
Daud1821	cmr1	COG1367: Predicted to be involved in DNA repair (RAMP superfamily)	CAS2B	453	40.62	T. thermophilus HB27	
Daud1292	cst2	COG1857: Predicted to be involved in DNA repair	CAS1	356	42.48	M. acetivorans	

Daud1815	devR	COG1857: Predicted to be involved in DNA repair	CAS2B	325	75	T. tengcongensis	
Daud0506		COG1743: Adenine-specific DNA methylase containing a Zn-ribbon		1004	35.15	T. thermophilus HB8	
Daud0361		(not COG1483) Predicted ATPase (AAA+ superfamily)		954	32	T. thermophilus HB8	
Daud0508		(not COG1483) Predicted ATPase (AAA+ superfamily)		925	29.11	T. thermophilus HB8	
Daud1800	csx7A	COG1337: Predicted to be involved in DNA repair (RAMP superfamily)	CAS2A	300	40.07	Syn. sp. JA-3	
Daud1801	csx7B	COG1337: Predicted to be involved in DNA repair (RAMP superfamily)	CAS2A	283	38.57	Syn. sp. JA-3	
Daud1804	csx7C	COG1337: Predicted to be involved in DNA repair (RAMP superfamily)	CAS2A	241	30.29	Syn. sp. JA-2	
Daud0563	hdrC/qmoC?	COG1150: Heterodisulfide reductase, subunit C	SR3	212	47.03	C. hydrogenoformans	
Daud1177		(not COG4089) Predicted membrane protein		239	56.71	P. thermopropionicum	
Daud0105	cdhA/cooS	COG1152: CODH/acetyl-CoA synthase, alpha subunit	CODH1	767	52.37	M. thermautotrophicus	
Daud2132		COG1751: Uncharacterized conserved protein, putative pyruvate kinase domain		182	56.91	D. reducens	us spoIID,spoIIID,mreB

Table S10. Horizontally transferred genes shared between clade and archaea.

Horizonatally transferred genes that are shared between the clade and archaea (not including horizontal transfers that happened prior to divergence or potentially also happened between archaea and other bacterial clades) were identified by finding homologous genes from archaeal genomes that have a BLASTp bit score greater than 1.2 times higher than any other homologous gene from all bacteria that are not a member of the clade (where the clade consisted of *Pelotomaculum thermopropionicum*, *Desulfotomaculum reducens*, *Carboxydothermus hydrogenoformans*, *Moorella thermoacetica*, *Desulfobacterium hafniense* (both Y51 and DCB-2), *Symbiobacterium thermophilum*, and *Thermoanaerobacter tengcongensis*). The clade members in which homologous genes are found is reported. Because the genomes are incomplete, absence of homologous genes from *P. thermopropionicum* and/or *D. reducens* does not necessarily indicate a most recent acquisition by *D. audaxviator*.

Gene Daud0895 may in fact be succinyl-CoA synthetase, as the acetyl-CoA synthetase α subunit family resembles that of succinyl-CoA synthetase α subunit (although *D. audaxviator* does not appear to possess the succinyl-CoA synthetase β subunit). Another potential xenologous replacement is for pckA (phosphoenolpyruvate carboxykinase) of the reverse TCA pathway. Some genes that are likely transferred from archaea are not present in this table even though they have a closest hit to an archaeal homolog, due to our strict bit score separation requirement with respect to non-clade bacteria (e.g. the H+ translocating pyrophosphatase Daud0308). Some of the putatively horizontally transferred genes that are not included in this table also have additional support for horizontal transfer indicated by the adjacent presence of clearly transferred genes (e.g. CRISPR-associated genes Daud1287 and Daud1289 within the CAS1 operon are missing from this table, even though they have archaeal closest hits and their adjoining genes did meet the bit score separation criterion for inclusion).

Clade members with homologous genes are listed under "Notes", with the following abbreviations: "P. thermo.": *Pelotomaculum thermopropionicum* SI, "D. red.": *Desulfotomaculum reducens* MI-1, "M. therm.": *Moorella thermoacetica* ATCC 39073, "C. hyd.": *Carboxydothermus hydrogenoformans* Z-2901, "T. teng.": *Thermoanaerobacter tengcongensis* MB4T, "S. therm.": *Symbiobacterium thermophilum* IAM 14863, "D.haf.Y51": *Desulfitobacterium hafniense* Y51, "D.haf.DCB2": *Desulfitobacterium hafniense* DCB-2.

Gene	Name	Description	Operon	Len	Archaeal CH id	Archaeal CH species	Notes and Clade homologs
Daud0105	cdhA/cooS	COG1152: CODH/acetyl-CoA synthase alpha subunit	CODH1	767	52.37	M. thermautotrophicus	
Daud0106	cdhB/acsE	COG1880: CODH/acetyl-CoA synthase	CODH1	185	38.18	M. maripaludis	

		epsilon subunit					
Daud0111	cdhC/acsB	COG1614: CODH/acetyl-CoA synthase beta subunit	CODH1	400	58.52	M. jannaschii	C.hydro., M.therm., D.haf.Y51, D.haf.DCB2
Daud0137		COG1600: Uncharacterized Fe-S protein		252	45.42	M. acetivorans	C.hydro.
Daud0143	nifH	COG1348: Nitrogenase subunit NifH (ATPase)	NIF1	281	79.27	M. thermautotrophicus	D.red.
Daud0144	nifI1	COG0347: Nitrogen regulatory protein PII	NIF1	106	64	M. maripaludis	D.red., D.haf.Y51, D.haf.DCB2
Daud0145	nifI2	COG0347: Nitrogen regulatory protein PII	NIF1	121	51.64	M. acetivorans	D.red., M.therm., D.haf.Y51, D.haf.DCB2
Daud0146	nifD	COG2710: Nitrogenase molybdenum-iron protein, alpha and beta	NIF1	483	46.04	M. thermautotrophicus	M.therm.
Daud0147	nifK	COG2710: Nitrogenase molybdenum-iron protein, alpha and beta	NIF1	491	46.55	M. maripaludis	
Daud0148	nifE	COG2710, [TIGR01283: Nitrogenase MoFe cofactor biosyn. protein NifE]	NIF1	460	38.86	M. thermautotrophicus	M.thermoautotrophicus annot.
Daud0167	dsrE	COG1553, PF02635: DsrE-like protein	DSR2	109	39.42	M. mazei	
Daud0194		COG0491: Zn-dependent hydrolases, including glyoxylases		529	28.2	M. mazei	
Daud0318		COG0778: Nitroreductase [NAD(P)H-flavin oxidoreductase]		240	54.63	T. kodakaraensis	T.kodakaraensis annot.
Daud0387	nikC	COG1173, TIGR02790: nickel ABC transporter, permease subunit NikC		310	59.55	M. acetivorans	C.hydro., S.therm.
Daud0408		[PF01850: PilT protein, N-terminal]		137	33.04	M. hungatei	M.thermoacetica annot; P.therm., D.red., M.therm.
Daud0442		COG0863: DNA modification methylase (Adenine specific?)		259	30.08	T. volcanium	
Daud0473	glnB/K	PF00543: Nitrogen regulatory protein P-II (GlnB, GlnK)	NIF2	118	54.21	M. hungatei	C.hydro., D.haf.Y51, D.haf.DCB2
Daud0474		PF07556: DUF1538 [putative membrane protein]	NIF2	246	51.49	M. mazei	C.hydro., D.haf.Y51, D.haf.DCB2
Daud0475		PF07556: DUF1538 [putative membrane protein]	NIF2	231	49.78	M. mazei	C.hydro.

Daud0489	wrbA	COG0655: Multimeric flavodoxin WrbA		207	40.41	M. thermautotrophicus	D.red., C.hydro., M.therm.
Daud0520		PF03681: Protein of unknown function UPF0150		75	49.21	M. barkeri	paralog to Daud0768; M.therm.
Daud0526		PF07833: Copper amine oxidase-like, N-terminal		457	39.8	M. maripaludis	P.therm., D.red., M.therm., T.teng, S.therm.
Daud0537		COG1099, PF01026: TatD-related deoxyribonuclease		263	35.43	M. kandleri	D.red., C.hydro.
Daud0541		COG1691: NCAIR mutase (PurE)-related proteins		267	46.25	M. jannaschii	M.therm.
Daud0544	TF	COG4978, PF06445: Bacterial transcription activator, effector binding		161	29.49	M. thermautotrophicus	us Rubrerythrin Daud0543
Daud0560		COG0388, PF00795: hydratase & apolipoprotein N-acyltransferase		273	47.35	M. thermautotrophicus	D.red.
Daud0647	mtaP	COG0005, TIGR01694: Methylthioadenosine phosphorylase	HGT2	293	58.24	T. kodakaraensis	P.therm., C.hydro., M.therm., D.haf.Y51, D.haf.DCB2, S.therm.
Daud0651		COG0402: Cytosine deaminase and related metal-dependent hydrolases	HGT2	431	52.42	M. mazei	P.therm., D.red., C.hydro., M.therm., T.teng, D.haf.Y51, D.haf.DCB2, S.therm.
Daud0686		PF07900: DUF1670		275	26.07	M. acetivorans	1 copy in D.audaxviator, 16 copies in M.acetivorans C2A
Daud0711		PF08378: nuclease-related domain (NERD)		186	35.62	M. thermautotrophicus	T.teng
Daud0723		COG0811: Biopolymer transport [MotA/TolQ/ExbB proton channel]	HGT3	227	46.43	M. thermautotrophicus	M.thermautotrophicus annot.; D.haf.Y51
Daud0724		COG4744: Uncharacterized conserved protein	HGT3	117	52.59	M. maripaludis	
Daud0773	ppc	COG1892, TIGR02751: archaeal-type PEP carboxylase		490	55.17	M. thermautotrophicus	no bacterial form of ppc in genome
Daud0808		COG0595: Predicted hydrolase, metallo-beta-lactamase family		520	38.61	P. abyssi	P.therm.
Daud0859	thi4	COG1635, TIGR00292: Thiamine biosynthesis Thi4 protein		260	57.75	M. thermautotrophicus	
Daud0895	sucA?	COG1042, TIGR02717: Acetyl-CoA		701	47.65	M. stadtmanae	Succinyl-CoA synthetase?; ds

		synthetase (ADP forming), alpha					phosphotransacetylase (HGT?); P.therm., D.red., C.hydro.
Daud0904		PF03681: Protein of unknown function UPF0150		135	45.28	M. acetivorans	
Daud0983		COG1487, PF01850: PilT protein, N-terminal		134	36.36	S. tokodaii	
Daud0987	aroE	COG0169, TIGR00507: Quinate/Shikimate 5-dehydrogenase	ARO1	290	51.76	M. kandleri	P.therm., D.red., C.hydro., M.therm., S.therm.
Daud1047		COG3387: Glucoamylase and related glycosyl hydrolases		664	37.19	T. acidophilum	C.hydro., M.therm.
Daud1092	cbiK	COG5266: ABC-type Co2+ transport system, periplasmic component		227	28.12	M. acetivorans	
Daud1095		hypothetical protein		157	35.92	M. acetivorans	P.therm., D.red., M.therm.
Daud1104		COG2191: Formylmethanofuran dehydrogenase subunit E		198	40.1	M. barkeri	C.hydro., M.therm., T.teng
Daud1123		hypothetical protein		344	28.25	A. fulgidus	
Daud1125		COG1682: ABC-type polysaccharide phosphate export, permease		261	43.82	M. hungatei	OR TIGR01247: Daunorubicin resistance ABC transporter (drrB); us ATPase; P.therm., C.hydro., M.therm.
Daud1144		COG4880: Secreted protein, C-term. b-propeller domain		663	36.18	M. hungatei	P.therm., D.haf.Y51, D.haf.DCB2
Daud1216	fepB	COG0614: ABC-type Fe3+-hydroxamate transport system, periplasmic	SID1	479	36.25	M. hungatei	other siderophore ABC transport components ds
Daud1217	fepD	COG0609: ABC-type Fe3+-siderophore transport system, permease	SID1	363	51.74	M. hungatei	M.therm., T.teng, D.haf.DCB2
Daud1218	fepB	COG0614: ABC-type Fe3+-hydroxamate transport system, periplasmic	SID1	359	30.79	M. acetivorans	
Daud1219	cobN	COG1429: Cobalamin biosynthesis CobN and related Mg-chelatases	SID1	1286	57.56	M. mazei	
Daud1238		COG3063: Tfp pilus assembly protein PilF		378	31.27	M. thermautotrophicus	D.red., M.therm., S.therm.
Daud1259		COG0714: MoxR-like ATPases	HGT1	413	37.75	P. aerophilum	
Daud1260		COG3864: Uncharacterized, conserved in	HGT1	416	32.23	P. aerophilum	

		bacteria					
Daud1286	csa1	COG4343, CAS AF1879 family	CAS1	312	37.83	A. fulgidus	
Daud1288	cas1	COG1343: Predicted DNA repair	CAS1	107	36.49	M. mazei	
Daud1290	cas3	COG1203: Predicted helicases	CAS1	796	31.23	M. barkeri	T.teng
Daud1291	cas5t	COG1688: Predicted DNA repair (RAMP)	CAS1	768	37.65	M. acetivorans	
Daud1292	cst2	COG1857, TIGR02585: CAS regulatory DevR	CAS1	356	42.48	M. acetivorans	T.teng
Daud1293	cst1	[TIGR01908: CAS, CXXC_CXXC region]	CAS1	519	22.74	M. barkeri	M.barkeri annot.
Daud1309		COG1809, PF02679: (2R)-phospho-3-sulfolactate synthase, ComA		263	43.03	M. thermautotrophicus	P.therm., D.red., M.therm., S.therm.
Daud1329		COG1647: Esterase/lipase	BIO1	224	33.33	M. barkeri	Biotin synthase operon
Daud1372		COG1578: Uncharacterized conserved protein		276	35.71	P. abyssi	
Daud1374		COG0826: Collagenase and related proteases		839	39.79	M. mazei	D.red., C.hydro., M.therm., D.haf.Y51, D.haf.DCB2
Daud1485	gvpA	PF00741: Gas vesicle protein GvpA	GVP1	117	53.33	M. barkeri	ds tNRA-Leu
Daud1486	gvpL	PF06386: Gas vesicle protein GvpL/GvpF	GVP1	369	32.54	M. barkeri	paralog to Daud1491
Daud1487	gvpK	PF05121: Gas vesicle protein GvpK	GVP1	106	58.59	M. barkeri	
Daud1489	gvpA	PF00741: Gas vesicle protein GvpA	GVP1	127	53.78	M. barkeri	Daud1488 is prob. also HGT [COG71: Molecular chaperone (small heat shock protein)]
Daud1563	pckA	COG1866: Phosphoenolpyruvate carboxykinase (ATP)		522	35.28	A. pernix	P.therm.
Daud1641	uspA	COG0589: Universal stress protein & related nucleotide-binding		295	28.33	M. acetivorans	
Daud1677		COG5559: Uncharacterized conserved small protein		72	57.89	A. fulgidus	
Daud1679		COG2445: Uncharacterized conserved protein		144	42.24	T. kodakaraensis	P.therm., C.hydro., M.therm.
Daud1708	abrB	TIGR01439: Transcriptional regulator AbrB (SpoVT/AbrB family)		93	54.55	T. kodakaraensis	

Daud1796		[TIGR02710: Cas02710]	CAS2A	315	28.48	M. thermautotrophicus	M. thermoautotrophicus annot.
Daud1813	cas3	COG1203, TIGR01587: helicase Cas3, core	CAS2B	730	29.27	P. horikoshii	T.teng
Daud1814	cas5	[TIGR02593: Cas5]	CAS2B	239	28	M. thermautotrophicus	T.tengcongensis annot.; T.teng
Daud1815	devR	COG1857, TIGR02585: CAS regulatory DevR	CAS2B	325	32.28	P. horikoshii	T.teng
Daud1911		COG2452: Predicted site-specific integrase-resolvase	TPS18	217	49.04	M. jannaschii	T.teng
Daud1977		COG5421: Transposase	TPS21	556	22.89	M. acetivorans	P.therm., T.teng, D.haf.DCB2
Daud1980		COG1945, TIGR00286: Pyruvoyl-dependent arginine decarboxylase		153	53.33	M. jannaschii	operon with spermidine synthase; D.red., C.hydro.
Daud1991		COG1533: DNA repair photolyase		295	43.68	A. fulgidus	M.therm.
Daud2113	abrB	TIGR01439: Transcriptional regulator AbrB (SpoVT/AbrB family)		88	37.97	T. kodakaraensis	
Daud2118		hypothetical protein		107	34.78	T. kodakaraensis	M.therm.
Daud2127		TIGR01439: Transcriptional regulator AbrB (SpoVT/AbrB family)		91	45.21	S. solfataricus	M.therm.
Daud2184		COG0863: DNA modification methylase (Adenine specific?)		372	44.97	T. acidophilum	C.hydro., M.therm., S.therm.

Figure S4(a,b). Archaeal-type molybdenum nitrogenase.

Phylogenetic tree based on nitrogenase nifH protein sequence (and nitrogenase-like sequences) from both sequenced organisms and environmental isolates used by Mehta and Baross (*59*). Sequences aligned with MUSCLE (*46*). Tree determined by maximum likelihood with PHYML (*47*) using JTT substitution model (*48*). High bootstrap value supported nodes are indicated by circles. FS406-22 nifH1 has been identified as functional at 92 ^{o}C (*59*). Truncated environmental clones were not included in (a) to allow for better resolution of the tree. While the nifH possessed by *D. audaxviator* is closest to the high temperature archaeal cluster, low bootstrap supported nodes with short branch lengths do not permit its confident phylogenetic placement. However, these sequences are sufficient to determine that the nifH possessed by *D. audaxviator* is not related by vertical decent to that possessed by *D. reducens*.

Figure S4(a). NifH tree.
Tree built from multiple sequence alignment over 242 positions, not including truncated environmental clones.

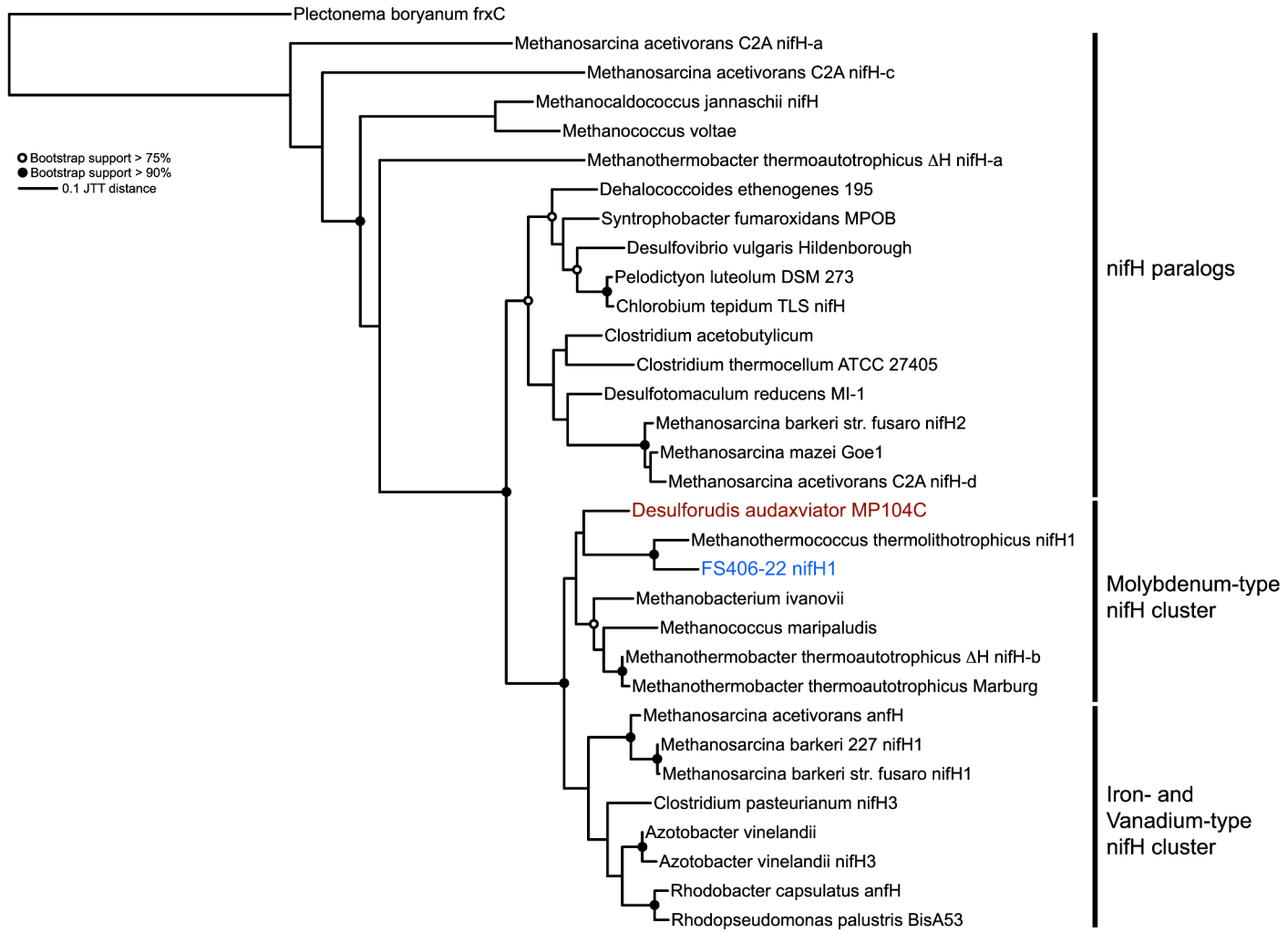


Figure S4(b). NifH tree, with truncated environmental clones of Mehta and Baross.
Tree built from multiple sequence alignment over 127 positions, including truncated environmental clones.

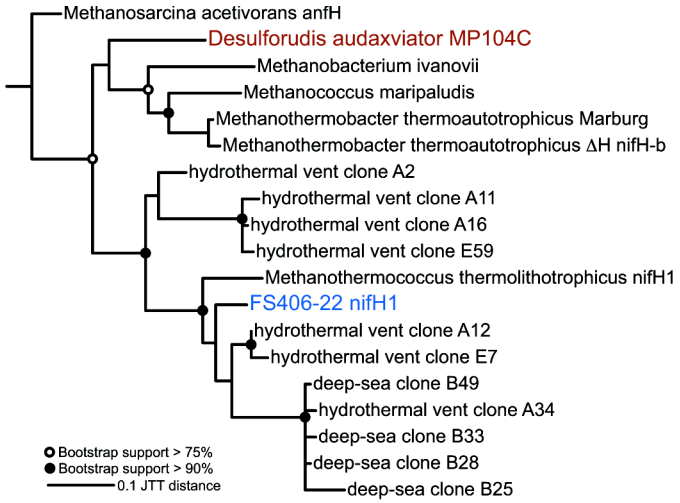


Table S11. Transposons, Integrases, and phage-associated genes.

D. audaxviator possesses a number of transposon insertion sites (83 sites with 30 types), some degenerate, that were identified by homology to known transposon sequence families or by their repetition in the genome. Several of the transposons with multiple sites have very high identity to one another, suggesting recent activity (e.g. TPN5, TPN7, TPN8, TPN10, TPN11, TPN12, TPN16, and TPN30). Additionally, some of these appear to be highly active, with numerous copies (e.g. TPN11, TPN12, TPN16, and TPN30). Many of the transposons are quite distant from the closest detected homolog in another species (e.g. TPN1, TPN2, TPN4, TPN9, TPN11, TPN20, TPN21, TPN29, TPN30). The frequent presence of the transposons adjacent to horizontally transferred genes suggests continued roles in genetic rearrangement and potentially transfer that contribute to adaptive flexibility, or perhaps such regions simply present targets that are more amenable to destructive insertions.

Genes with the closest match to an archaeal homolog and those that only match the N-terminal or C-terminal portion of the full transposon are indicated in the Notes column. Genes that appear to be pseudogenes are indicated by "*", and have lengths measured in base pairs rather than amino acids.

Gene	Name	Description	Group	Len	id of most distant paralog	CH id	CH species	Notes
Daud1448*		COG675 Transposase, IS605 OrfB	TPN1	1314*	96.74	27.17	N. pharaonis	Archaea
Daud0713		COG675 Transposase, IS605 OrfB	TPN1	461	96.74	35.29	N. pharaonis	Archaea
Daud1582		COG675 Transposase, IS605 OrfB	TPN2	441	92.27	27.51	S. ruber	
Daud1730		COG675 Transposase, IS605 OrfB	TPN2	444	92.27	27.25	S. ruber	
Daud0784		COG675 Transposase, IS605 OrfB	TPN2	441	92.27	26.99	S. ruber	
Daud0762		COG675 Transposase, IS605 OrfB	TPN2	441	92.27	28.11	S. ruber	
Daud1583		COG1943 Transposase, IS200 like	TPN3	132	98.47	67.69	T. maritima	
Daud1731		COG1943 Transposase, IS200 like	TPN3	132	98.47	67.69	T. maritima	
Daud0888		COG1943 Transposase, IS200 like	TPN3	132	98.47	67.69	T. maritima	
Daud1711*		COG675 Transposase, IS605 OrfB	TPN4	1266*	N/A	26.67	P. thermopropionicum	
Daud1955		COG5421 Transposase, (IS4?)	TPN5	579	99.31	56.37	G. kaustophilus	
Daud0721		COG5421 Transposase, (IS4?)	TPN5	579	99.31	56.91	G. kaustophilus	
Daud0958		COG5421 Transposase, (IS4?)	TPN5	579	99.31	56.55	G. kaustophilus	
Daud2153		COG1943 Transposase, DUF1568	TPN6	222	N/A	44.39	Pir. sp. 1	
Daud0201		COG3436 Transposase, IS66	TPN7	524	99.62	75.13	M. thermoacetica	
Daud0704		COG3436 Transposase, IS66	TPN7	524	99.62	75.38	M. thermoacetica	
Daud0202		COG3436 Transposase, IS66 Orf2 like	TPN8	119	100	76.47	M. thermoacetica	
Daud0703		COG3436 Transposase, IS66 Orf2 like	TPN8	119	100	76.47	M. thermoacetica	
Daud0774		COG3666 Transposase, IS4	TPN9	78	91.55	46.27	Jann. sp. CCS1	N-terminal
Daud0792		COG3666 Transposase, IS4	TPN9	578	91.55	27.59	T. tengcongensis	
Daud0714		COG675 Transposase, IS891/IS1136/IS1341, IS605 OrfB	TPN10	382	99.74	49.44	P. haloplanktis	

Daud0845		COG675 Transposase, IS891/IS1136/IS1341, IS605 OrfB	TPN10	382	99.74	49.16	P. haloplanktis	
Daud1502		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud1503		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud1735		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud1752		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud2194		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud0159		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud0173		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud0215		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud0218		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud0257		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud0399		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud0428		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud0432		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud0587		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud1246		PF01609 Transposase, IS4	TPN11	444	99.55	34.48	L. sakei	
Daud1807		PF01609 Transposase, IS4	TPN12	465	100	44.4	P. thermopropionicum	
Daud1808		PF01609 Transposase, IS4	TPN12	465	100	44.4	P. thermopropionicum	
Daud1809		PF01609 Transposase, IS4	TPN12	465	100	44.4	P. thermopropionicum	
Daud0695		PF01609 Transposase, IS4	TPN12	465	100	44.4	P. thermopropionicum	
Daud0420		COG3039 Transposase, IS5	TPN12	247	74.5	52.87	M. thermoacetica	N-terminal
Daud0197*		COG3039 Transposase, IS5	TPN12	480*	78.4	57.52	M. thermoacetica	N-terminal
Daud2240*		PF00872 Transposase, mutator type	TPN13	99*	N/A	68.75	P. thermopropionicum	
Daud0741*		PF00872 Transposase, mutator type	TPN14	652*	N/A	63.64	P. thermopropionicum	
Daud0744		PF01527 Transposase IS3/IS911	TPN15	76	N/A	40.28	M. thermoacetica	
Daud1501		PF00665 Integrase, catalytic region	TPN16	443	99.55	44.08	P. thermopropionicum	
Daud1862		PF00665 Integrase, catalytic region	TPN16	443	99.55	43.8	P. thermopropionicum	

Daud0321		PF00665 Integrase, catalytic region	TPN16	463	99.55	43.8	P. thermopropionicum	
Daud0950		PF00665 Integrase, catalytic region	TPN16	443	99.55	44.08	P. thermopropionicum	
Daud0405*		Putative transposase y4bF	TPN16	1328*	98.32	34.87	K. pneumoniae	N-terminal
Daud0582*		Putative transposase y4bF	TPN16	426*	97.48	35.71	V. splendidus	N-terminal
Daud1707		TF01784 CHP, (Transposase, ISNCY?)	TPN17	281	N/A	52.92	P. thermopropionicum	
Daud1710		COG2452 Site-specific integrase-resolvase, Orf in partial transposon ISC1904	TPN18	223	38.76	47.62	M. jannaschii	Archaea
Daud1911		COG2452 Site-specific integrase-resolvase, Orf in partial transposon ISC1904	TPN18	217	38.76	49.04	M. jannaschii	Archaea
Daud1833		(COG5421 Transposase)	TPN19	159	N/A	57.01	P. thermopropionicum	
Daud1910		(PF07282 Transposase, IS605 OrfB)	TPN20	508	N/A	27.56	T. tengcongensis	
Daud1977		(COG5421 Transposase)	TPN21	556	N/A	22.89	M. acetivorans	Archaea
Daud1992		(COG5421 Transposase)	TPN22	92	N/A	68.89	D. hafniense Y51	
Daud0315*		PF00665 Integrase, catalytic region	TPN23	279*	70	80	M. thermoacetica	
Daud0743		PF00665 Integrase, catalytic region	TPN23	173	59.62	51.14	M. thermoacetica	
Daud0786		COG3344 Retron-type reverse transcriptase	TPN23	310	59.62	81.62	P. thermopropionicum	
Daud0317		(COG2801 Transposase)	TPN24	142	N/A	33.58	A. tumefaciens	
Daud0322*		(COG3547 Transposase)	TPN25	362*	N/A	57.5	S. thermophilum	split gene call?
Daud0322*		(COG3547 Transposase)	TPN25	362*	N/A	55.56	C. hydrogenoformans	split gene call?
Daud0404*		(COG4584 Transposase)	TPN26	141*	N/A	56.41	M. aqueolei	
Daud0765*		(COG1484 DNA replication, IS21 transposase)	TPN27	375*	N/A	52.05	S. thermophilum	
Daud0785*		PF00665 Integrase, catalytic region	TPN28	204*	N/A	72.31	M. thermoacetica	
Daud0819		COG1961 Site-specific recombinases, DNA invertase Pin	TPN29	541	36.17	31.3	P. thermopropionicum	
Daud1245		PF00239 Resolvase, N-terminal	TPN29	128	36.17	36.19	P. thermopropionicum	N-terminal
Daud0770		Putative transposase	TPN30	473	99.15	32.41	M. thermoacetica	
Daud0981		Putative transposase	TPN30	473	99.36	32.62	M. thermoacetica	
Daud1109		Putative transposase	TPN30	473	99.15	32.62	M. thermoacetica	

Daud1922		Putative transposase	TPN30	473	99.58	32.62	M. thermoacetica	
Daud0622		Putative transposase	TPN30	473	99.58	32.62	M. thermoacetica	
Daud0030		Putative transposase	TPN30	473	99.36	32.62	M. thermoacetica	
Daud1527		Putative transposase	TPN30	473	99.15	32.62	M. thermoacetica	
Daud2239		Putative transposase	TPN30	473	99.15	32.62	M. thermoacetica	
Daud0780		Putative transposase	TPN30	473	99.15	32.2	M. thermoacetica	
Daud1954		Putative transposase	TPN30	472	99.15	32.48	M. thermoacetica	
Daud0817		Putative incomplete transposase	TPN30	400	99.24	33.99	M. thermoacetica	
Daud0573		Putative incomplete transposase	TPN30	326	99.42	31.82	M. thermoacetica	
Daud0326*		Putative incomplete transposase	TPN30	318*	96.34	N/A	NONE_DETECTED	
Daud0764		COG4679 Phage-related protein		118	N/A	40.37	P. luteolum	
Daud0763		PF01381 Putative phage-like regulator		92	N/A	46.51	M. magneticum	
Daud1110	hflX	COG2262 GTP-binding protease for phage lambda cII repressor		423	N/A	62.11	P. thermopropionicum	
Daud1981		COG648 Endonuclease IV		288	N/A	46.62	T. tengcongensis	
Daud1282	xerD	COG4974 Site-specific recombinase XerD	XERD	296	28.62	56.27	P. thermopropionicum	
Daud0498	xerD	COG4974 Site-specific recombinase XerD	XERD	391	26.74	27.21	X. campestris	
Daud0499	xerD	PF00589 Phage integrase, catalytic core	XERD	286	25.74	38.97	M. thermoacetica	
Daud0519	xerD	COG4974 Site-specific recombinase XerD	XERD	301	27.27	36.54	Dehalo. sp. CBDB1	
Daud0775	xerD	COG4974 Site-specific recombinase XerD	XERD	287	28.52	37.11	D. reducens	
Daud0795	xerD	COG4974 Site-specific recombinase XerD	XERD	284	25.74	39.73	D. reducens	
Daud0684*	xerD	PF00589 Phage integrase, catalytic core	XERD	210*	80.6	36.9	P. furiosus	Archaea; like Daud0775, Daud0795
Daud1530		COG1573, TF00758 Phage SPO1 DNA polymerase-related	DPOL	238	41.13	35.98	D. geothermalis	
Daud0290		COG1573, TF00758 Phage SPO1 DNA polymerase-related	DPOL	208	41.13	63.68	P. thermopropionicum	
Daud0289		COG1532 Predicted RNA-binding protein		63	N/A	60.66	C.	precedes Daud0290

							hydrogenoformans	
Daud0794		COG1525 prophage LambdaCh01, nuclease domain protein	TNUC	328	53.11	47.31	C. hydrogenoformans	
Daud0844		COG1525 prophage LambdaCh01, nuclease domain protein	TNUC	219	53.11	51.24	C. hydrogenoformans	between recA and Daud0845
Daud0710		COG1974 putative prophage LambdaCh01, repressor protein		246	N/A	30.54	C. hydrogenoformans	
Daud1912		COG1974 putative prophage repressor of the SOS regulon		197	N/A	34.62	N. farcinica	

Table S12(a,b,c). CRISPR sequences and CRISPR-associated genes.

The genome possesses two "clustered regularly interspaced short palindromic repeat" (CRISPR) regions, and several CRISPR-associated proteins (CAS) (*60*), cst1, cst2, cas5t, and cas3 (the first three of which currently only have BLAST-detectable homologs in *Thermosinus carboxydivorans*) appear in the same linear order as their archaeal homologs, suggesting their transfer as a cassette. The viral defense role of CRISPRs has recently been confirmed (*61*), and appears to employ an RNAi-like approach (*62*) with variable sequences that contain viral antisense nucleotides between the CRISPR sequences (*61*), called “spacers”. We did not find similarity between these variable sequences and the unassembled reads, although the 0.2 μm filter pore size used to collect bacterial cells would have prohibited capture of external viral particles. We also found no significant hits to known protein sequences, but viral sequence is notoriously fast-evolving and vastly under-sampled, so we cannot rule out a viral defense role for the CRISPR regions. The extremophile association of some of the genes in the CRISPR-associated genes of region 1 suggests uncharacterized viral types may inhabit this high temperature environment.

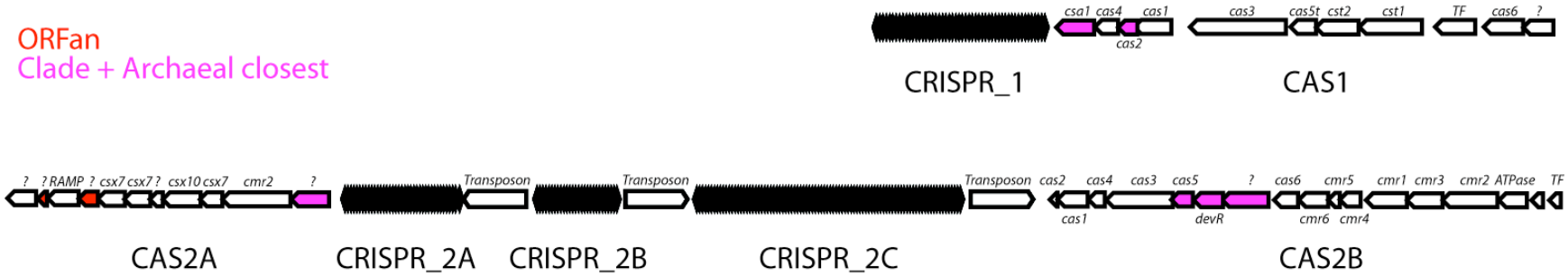
Region 1 (from position 1355523 to 1359321 in the genome) has 52 instances (51 of which are perfectly identical) of the 34 base repeat sequence CTTTCAGTCCCCTTTTCGT[C51,T1]GGGTCGGTCGCTGA, with intervening variable sequences of length 36 to 43 bases. Region 2 (from position 1898565 to 1912072 in the genome) has 157 instances (all of which are perfectly identical, but two of which are truncated to 21 bases just before a transposon) of the 30 base repeat sequence GTTTCAATCCCTCGTAGGTAGGCTGGAAAC. Region 2 can be divided into 3 sub-regions with 3 transposons (Daud1807, Daud1808, and Daud1809, all transposons of the TPN12 group), with 34 instances of the repeat sequence in CRISPR_2A (from position 1898565 to 1900798) with intervening variable sequences of length 35 to 40 bases, 24 instances in CRISPR_2B (from

position 1902591 to 1904342) with intervening variable regions of length 35 to 38 bases, and 90 instances in CRISPR_2C (from position 1906154 to 1912072) with intervening variable regions of length 35 to 40 bases.

The intervening variable sequences were scanned with BLASTx against the non-redundant (NR) sequence database from the NCBI, both as separate sequences and as a concatenated sequence for each sub-region. No significant matches were found in the NR. Searches for similarity between intervening variable sequences and the Sanger and 454 reads that did not match the assembly using BLASTn also did not yield significant hits. The intervening variable sequences were also scanned with BLASTn against the genome of *D. audaxviator*. There were no strong matches for region 1 or the first sub-region of region 2. A single strong match was found in the second sub-region of region 2 for the 36 base sequence ACACTCTACCCTGGATGTACTGGGCCTTCTTCCGCC (positions 1903352 to 1903387) to a perfect complement from the third sub-region of region 2 (positions 1906842 to 1906877). The third sub-region of region 2 possesses a group of three identical 36 base intervening sequences (positions 1909564 to 1909599, 1909960 to 1909995, and 1910092 to 1910127) with sequence CTGCGCTTCCCCAGCAGTACCCCCGCTTGTCTCCAG, a pair of identical 36 base intervening sequences (positions 1910026 to 1910061 and 1910158 to 1910193) with sequence TTTTGCAAAGTGAGTTGAGCAACTTAATGTCCCGAA, a pair of identical 36 base intervening sequences (positions 1910817 to 1910852 and 1911020 to 1911055) with sequence CACCCCAACCCCTCCGGGAGTAAAACCTACGGAGGG, a pair of identical 37 base intervening sequences (positions 1910883 to 1910919 and 1911086 to 1911122) with sequence GTCAATACAACAGAATAAAATTCGCCGAGATTCGGCA. Other, less strong, matches from the third sub-region of region 2 include the incomplete match of a variable intervening region (only 27 of 40 bases contiguously perfect) of sequence TTCTTTACTTCTTCCTGCCGGGATTTA (positions 1909440 to 1909466 with 1909903 to 1909929), and the internal palindromic match of 20 of 36 bases of sequence AGTTTCTACATGTAGAAACT to itself (positions 1911686 to 1911705).

Region 1 has an adjoining downstream collection of CRISPR-associated (Cas) genes (*60*), whereas region 2 has both upstream and downstream Cas genes. Several of these genes have closest homologs in clade members (mostly *Thermosinus carboxydivorans* and *Thermoanaerobacter tengcongensis*) or archaea, including several in a row in region 1 that suggest their conservation as a cassette. We have grouped these genes into 3 putative operons, CAS1 (downstream of region 1), CAS2A (downstream of region 2), and CAS2B (upstream of region 2). All three operons are on the (-) strand.

ORFan
Clade + Archaeal closest



Gene	Name	Description	Operon	Len	CH id	CH species	Notes
Daud1286	csa1	COG4343, CAS AF1879 family	CAS1	312	37.83	A. fulgidus	Archaea
Daud1287	cas4	COG1468, TIGR00372: Cas4	CAS1	191	34.76	A. fulgidus	Archaea
Daud1288	cas2	COG1343: Predicted DNA repair	CAS1	107	36.49	M. mazei	Archaea
Daud1289	cas1	COG1518, TIGR00287: Cas1	CAS1	307	40.07	A. fulgidus	Archaea
Daud1290	cas3	COG1203: Predicted helicases	CAS1	796	31.23	M. barkeri	Archaea
Daud1291	cas5t	COG1688: Predicted DNA repair (RAMP)	CAS1	768	37.65	M. acetivorans	Archaea
Daud1292	cst2	COG1857, TIGR02585: CAS regulatory DevR	CAS1	356	42.48	M. acetivorans	Archaea
Daud1293	cst1	[TIGR01908: CAS, CXXC_CXXC region]	CAS1	519	22.74	M. barkeri	Archaea; M.barkeri annot.
Daud1294	TF	COG2378: Predicted transcriptional regulator	CAS1	337	37.42	P. luteolum	
Daud1295	cas6	[COG5551, TIGR01877: Cas6]	CAS1	332	32.57	M. succiniciproducens	A. fulgidus annot.
Daud1296		hypothetical protein	CAS1	256	46.27	Arthrobacter sp. FB24	Arthrobacter only homolog
Daud1796		[TIGR02710: Cas02710]	CAS2A	315	28.48	M. thermautotrophicus	Archaea; M. thermoautotrophicus annot.
Daud1797		hypothetical protein	CAS2A	105	N/A	ORFan	
Daud1798	RAMP	[COG1337: Predicted DNA repair (RAMP)]	CAS2A	340	31.56	Syn. sp. JA-3-3	Synechococcus annot.
Daud1799		hypothetical protein	CAS2A	215	N/A	ORFan	

Daud1800	csx7A	COG1337 Predicted DNA repair (RAMP)	CAS2A	300	40.07	Syn. sp. JA-3-3	
Daud1801	csx7B	COG1337, TIGR02581: CAS RAMP SSO1426	CAS2A	283	38.57	Syn. sp. JA-3-3	
Daud1802		hypothetical protein	CAS2A	142	N/A	ORFan	
Daud1803	csx10	[TIGR02674: CAS RAMP Csx10]	CAS2A	424	30.5	Syn. sp. JA-2-3	Synechococcus annot.
Daud1804	csx7C	COG1337: CAS RAMP, [SSO1426 family]	CAS2A	241	30.29	Syn. sp. JA-2-3	Synechococcus annot.
Daud1805	crm2 (cmr?)	[COG1353, TIGR02577: CAS Crm2]	CAS2A	760	26.51	Syn. sp. JA-3-3	Synechococcus annot.
Daud1806	RE?	Restriction endonuclease?	CAS2A	424	25.76	V. splendidus	
Daud1810	cas2	COG1343, TIGR01573: Cas2	CAS2B	88	80.46	T. tengcongensis	
Daud1811	cas1	COG1518, TIGR00287: Cas1	CAS2B	332	72.81	T. tengcongensis	
Daud1812	cas4	COG1468, TIGR00372: Cas4	CAS2B	179	66.47	T. tengcongensis	
Daud1813	cas3	COG1203, TIGR01587: helicase Cas3, core	CAS2B	730	58.4	T. tengcongensis	
Daud1814	cas5	[TIGR02593: Cas5]	CAS2B	239	63.27	T. tengcongensis	T. tengcongensis annot.
Daud1815	devR	COG1857, TIGR02585: CAS regulatory DevR	CAS2B	325	75	T. tengcongensis	
Daud1816		hypothetical protein	CAS2B	515	65.2	T. tengcongensis	
Daud1817	cas6	COG1583, TIGR01877: Cas6	CAS2B	265	68.83	P. thermopropionicum	
Daud1818	cmr6	COG1604, TIGR01898: CAS RAMP Cmr6	CAS2B	324	73.15	P. thermopropionicum	
Daud1819	cmr5	COG3337, TIGR01881: CAS Cmr5	CAS2B	126	67.29	P. thermopropionicum	
Daud1820	cmr4	COG1336, TIGR02580: CAS RAMP Cmr4	CAS2B	289	72.05	P. thermopropionicum	
Daud1821	cmr1	COG1367, TIGR01894: CAS RAMP Cmr1	CAS2B	453	40.62	T. thermophilus HB27	
Daud1822	cmr3	COG1769, [TIGR01888: CAS Cmr3]	CAS2B	392	73.05	P. thermopropionicum	B. halodurans annot.
Daud1823	crm2 (cmr?)	COG1353, TIGR02577: CAS Crm2	CAS2B	608	72.12	P. thermopropionicum	
Daud1824	TF	COG2865: Predicted transcriptional regulator (HTH domain + unc. domain)	CAS2B	305	34.68	T. tengcongensis	Daud1825 paralog over C-term
Daud1825		hypothetical protein	CAS2B	127	62.7	T. tengcongensis	Daud1824 paralog
Daud1827	RR	COG2197: Response regulator (CheY-like domain + HTH DNA-binding domain)	CAS2B	282	67.74	P. thermopropionicum	

Table S13, Figures S5 and S6. Sulfate and sulfite reduction genes.

Sulfate and sulfite reducing and related genes were identified by membership in known sequence families (e.g. COG, TIGRFAM, and Pfam) or by gene context (proximity and/or presence in operons with other identified sulfate and sulfite reducing genes). Annotation was by protein family, or if no protein family could be assigned with confidence, by the protein family assignment of the nearest homolog (such annotations are indicated with square brackets, with the source organism provided in the notes).

Consistent with the thermodynamic evaluation (*3*) that SO_4^{2-} offers the most energetically favorable electron acceptor, the genome possesses the capacity for dissimilatory SO_4^{2-}reduction (DSR) with a gene repertoire like that of other SO_4^{2-} reducing microorganisms (*63*). Access to extracellular SO_4^{2-} is provided by a Na^+/SO_4^{2-} symporter. The SO_4^{2-} is activated by Sat (sulfate adenylyltransferase), three putative copies of which exist in the genome. Two of the Sat genes are in a cluster (in SR7 of Fig. 2), the first of which has orthologs within *P. thermopropionicum*, *D. reducens*, and *C. hydrogenoformans*. The second Sat gene, which is very close the first Sat gene, follows a proline tRNA gene (a common insertion point for horizontal transfers (*64*)) and a methyl-accepting chemotaxis protein (MCP), and has orthologs primarily among archaea (with the exception of one other bacterial genome at the time of this writing, *Mycobacterium avium* 104), suggesting the collective acquisition of a set of useful genes. The third putative Sat (in SR8) has only ~30-35% amino acid identity to the nearest homologs, and may be involved in either assimilatory or dissimilatory sulfate reduction. The genome also contains a H^+-translocating pyrophosphatase for utilization of pyrophosphate released by the activation of SO_4^{2-} by Sat to further enhance the H^+ gradient (Fig. 3). Interestingly, this gene appears to have been horizontally acquired and is one of the few genes showing a non-synonymous SNP in the population (Table S7).

In dissimilatory sulfate reduction, the activated SO_4^{2-} is then reduced to sulfite (SO_3^{2-}) by AprAB (adenylylsulfate reductase), of which there are three instances, two of which are proximal (SR9A and SR9B) and separated only by an uncharacterized gene on the opposite strand. Lastly, the SO_3^{2-} is converted into hydrogen sulfide (H_2S) by DsrAB (dissimilatory sulfite reductase), one copy of which occurs in the genome (in SR11). The cytoplasmic components HmeD/DsrK, QmoA, and QmoB of the membrane-associated Hdr-like menaquinol-oxidizing enzyme ("Hme", also called "DsrMKJOP") and quinone-interacting membrane-bound oxidoreductase ("Qmo") complexes that contribute electrons and H^+ extrusion are found, as are the membrane-bound components HmeC/DsrM and two putative, domain-split, copies of QmoC. Other missing components may have their functionality provided by the frh-type hydrogenase (Coenzyme F420-reducing hydrogenase) and numerous heterodisulfide-like reductases (labeled "hdrA" and "hdrX") found in operons with other DSR genes (e.g. SR4 of Fig. 2). Alternatively, these uncharacterized components may instead form a novel complex that could play a role in SO_4^{2-} reduction. Other genes may play slightly different roles depending on conditions, such

as the genes for PAPS-reductase (cysH), which may also act as APS-reductase, and could be active in either assimilatory or dissimilatory pathways.

Table S13. Sulfate and sulfite reduction genes.

Gene	Name	Description	Operon	Len	CH id	CH species	Notes
Daud0092	hdrA	COG1148: Heterodisulfide reductase, subunit A and related polyferredoxins	SR1	1014	55.77	D. reducens	
Daud0093	hmcF	COG0247: Fe-S oxidoreductase [HmcF, 52.7 kd protein in hmc operon]	SR1	424	42.23	D. reducens	DvH annot.
Daud0167	dsrE	COG1553, PF02635: DsrE-like protein	SR2	109	39.42	M. mazei	HGT
Daud0307	secG	TIGR00810 Preprotein translocase SecG subunit	SR3 / TPT29 / GLY1	77	67.11	D. reducens	hh on GLY1?
Daud0308	hppA	COG3808, TIGR01104: V-type H(+)-translocating pyrophosphatase	SR3 / TPT29 / GLY1	683	61.01	M. acetivorans	proton pump; probable HGT; high SNP count; hh on GLY1?
Daud0563	hdrC / qmoC?	COG1150: Heterodisulfide reductase, subunit C	SR4	212	47.03	C. hydrogenoformans	incomplete match to DvH qmoC
Daud0564	hdrB / qmoC?	COG2048: Heterodisulfide reductase, subunit B	SR4	284	55.96	C. hydrogenoformans	
Daud0565	hdrA	COG1148: Heterodisulfide reductase, subunit A and related polyferredoxins	SR4	662	68.58	C. hydrogenoformans	
Daud0566	frhD	COG1908: Coenzyme F420-reducing hydrogenase, delta subunit	SR4	145	70.63	C. hydrogenoformans	
Daud0567	hdrX	COG1139, PF00037: 4Fe-4S ferredoxin	SR4	329	46.13	P. thermopropionicum	
Daud0568	hdrX	PF00037: 4Fe-4S ferredoxin, [putative anaerobic sulfite reductase, A subunit]	SR4	343	54.12	P. thermopropionicum	C.hydrogenoformans annot.
Daud0569	hdrX	COG0543, PF00970: Oxidored. FAD-binding & PF00175: Oxidored. FAD/NAD(P)-binding	SR4	282	64.13	C. hydrogenoformans	
Daud0787	Na+/SO42-	COG0471: Di- and tricarboxylate transporters, PF00939: Sodium/sulphate symporter	SR5	492	36.29	B. xenovorans	likely Pfam correct based on gene context
Daud0788	aprA /	COG1053, [TIGR02061: Adenylylsulfate reductase,	SR5	580	42.04	C. acetobutylicum	A. fulgidus annot.

	sdhA	alpha subunit]					
Daud0789	aprB	PF00037, [TIGR02060: Adenylylsulfate reductase, beta subunit]	SR5	108	44.76	A. fulgidus	A. fulgidus annot.
Daud0790	hdrA / frhD	COG1148: Heterodisulfide reductase, subunit A and related polyferredoxins	SR5	701	31.3	C. tepidum	
Daud0799	cysH / sat2?	COG0175: PAPS reductase & COG1148: Heterodisulfide reductase, subunit A	SR6	756	27.22	R. albus	
Daud1076	sat	COG2046, TIGR00339: ATP-sulfurylase	SR7	421	55.31	C. tepidum	
Daud1077	MCP	COG0840: Methyl-accepting chemotaxis protein	SR7	370	37.38	D. reducens	
Daud1078	sat	COG2046, TIGR00339: ATP-sulfurylase	SR7	421	54.65	M. avium	HGT
Daud1079		hypothetical protein	SR7	102	47.25	M. avium	HGT
Daud1080	FD	COG1141: Ferredoxin	SR7	63	62.3	P. thermopropionicum	
Daud1701	sat?	COG3119: Arylsulfatase A and related, PF00884: Sulfatase	SR8	484	30.21	H. marismortui	
Daud1702	cysH	COG0175, PF01507: PAPS reductase	SR8	229	39.19	D. vulgaris DP4	
Daud1703	wcaJ	COG1086: Sugar epimerases & COG2148: Sugar transferases lipopolysaccharide syn.	SR8	512	46.85	D. reducens	
Daud1881	hdrB / qmoC?	COG2048: Heterodisulfide reductase, subunit B	SR9A	296	48.12	D. reducens	
Daud1882	hdrC / qmoC?	COG1150: Heterodisulfide reductase, subunit C	SR9A	196	38.34	D. reducens	
Daud1883	hdrA / qmoB	COG1148: Heterodisulfide reductase A & COG1908: Coenz. F420-reducing hydro. delta	SR9A	736	64.86	D. reducens	ortholog of DvH qmoB
Daud1884	hdrA / qmoA	COG1148: Heterodisulfide reductase, subunit A and related polyferredoxins	SR9A	420	58.31	C. chlorochromatii	ortholog of DvH qmoA
Daud1885	aprA	COG1053, TIGR02061: Adenylylsulfate reductase, alpha subunit	SR9A	630	76.63	D. reducens	
Daud1886	aprB	PF00037, TIGR02060: Adenylylsulfate reductase, beta subunit	SR9A	146	78.42	D. reducens	
Daud1887		hypothetical protein		140	29.81	D. reducens	
Daud1888	aprB	[TIGR02060: Adenylylsulfate reductase, beta subunit]	SR9B	115	36.7	D. reducens	A. fulgidus annot.
Daud1889	aprA /	COG1053, [TIGR02061: Adenylylsulfate reductase,	SR9B	592	38.02	A. fulgidus	HGT?

	sdhA	alpha subunit]					
Daud2187	nfi	COG1515: Deoxyinosine 3'endonuclease (endonuclease V)	SR10	227	51.8	D-monas spp.	
Daud2188	cutA	COG1324, PF03091: CutA1 divalent ion tolerance protein	SR10	108	42.42	L. interrogans	
Daud2189	cbiA (dsrN?)	COG1797, TIGR00379: Cobyrinic acid a,c-diamide synthase CbiA [DsrN?]	SR10	466	54.18	D. reducens	C.tepidum annot.
Daud2190	dsrC	COG2920: Dissimilatory sulfite reductase (desulfoviridin), gamma subunit	SR10	106	64.76	D. reducens	
Daud2191	dsrK / hmeD?	COG0247, PF00037: 4Fe-4S ferredoxin [DsrK?]	SR10	462	62.55	D. reducens	ortholog of DvH dsrK; C.tepidum annot.
Daud2192	dsrM / hmeC?	COG2181: Nitrate reductase gamma subunit [DsrM?]	SR10	256	49.02	D. reducens	ortholog of DvH dsrM; COG:narI?; C.tepidum annot.
Daud2193	dsrA/B? (hdrA?)	COG2221: Dsr (desulfoviridin) A&B [heterodisulfide reductase (hdrA-2)]	SR10	61	56.45	D. reducens	truncated compared with dsrA or dsrB; A. fulgidus annot.
Daud2197		hypothetical protein	SR11	208	36.27	D. desulfuricans G20	
Daud2198	pilF	COG3063: Tfp pilus assembly protein PilF	SR11	210	62.93	M. thermoacetica	
Daud2199	dsrD	[dissimilatory sulfite reductase D]	SR11	78	56.16	M. thermoacetica	DvH annot.
Daud2200	dsrB	COG2221, TIGR02066: Sulfite reductase, dissimilatory-type beta subunit	SR11	397	74.37	M. thermoacetica	
Daud2201	dsrA	COG2221, TIGR02064: Sulfite reductase, dissimilatory-type alpha subunit	SR11	474	73.15	M. thermoacetica	

Figure S5. Sat phylogenetic genome context analysis.

Desulforudis audaxviator has a gene cluster than includes two copies of sulfate adenylyltransferase (Sat), one of which (Daud1076) resembles that of its clade relatives (in the blue box), while the other (Daud1078) has primarily been found in archaea (in the red box), with the exception of its presence in *Mycobacterium avium* 104. Figure S5's gene context analysis coupled with phylogenetic analysis does not reveal much of the history of the Sat genes in D. audaxviator, except to reveal that the "bacterial version" (Daud1076, in the blue box) has not retained the gene order of sat, aprB, aprA, hdrA, frhD that other bacteria (*D. reducens, C. chlorochromatii, S.*

fumaroxidans) have either vertically inherited obtained as a cassette via horizontal gene transfer. Additionally, the sat gene appears to be quite mobile, with the gene phylogeny not closely corresponding to the species phylogeny. The tree below is from a multiple sequence alignment of the Sat protein sequence built using MUSCLE (*46*), and determined by maximum likelihood by PHYML (*47*) with 100 replicates for bootstrapping (sampling with replacement), using the JTT amino acid substitution model (*48*). The Sat gene is indicated by the gray arrow, with the archaeal-type gene context in the red box and the bacterial-type context in the blue box. Archaeal species names are blue. The gap between the *D. audaxviator* gene regions is zero bases.

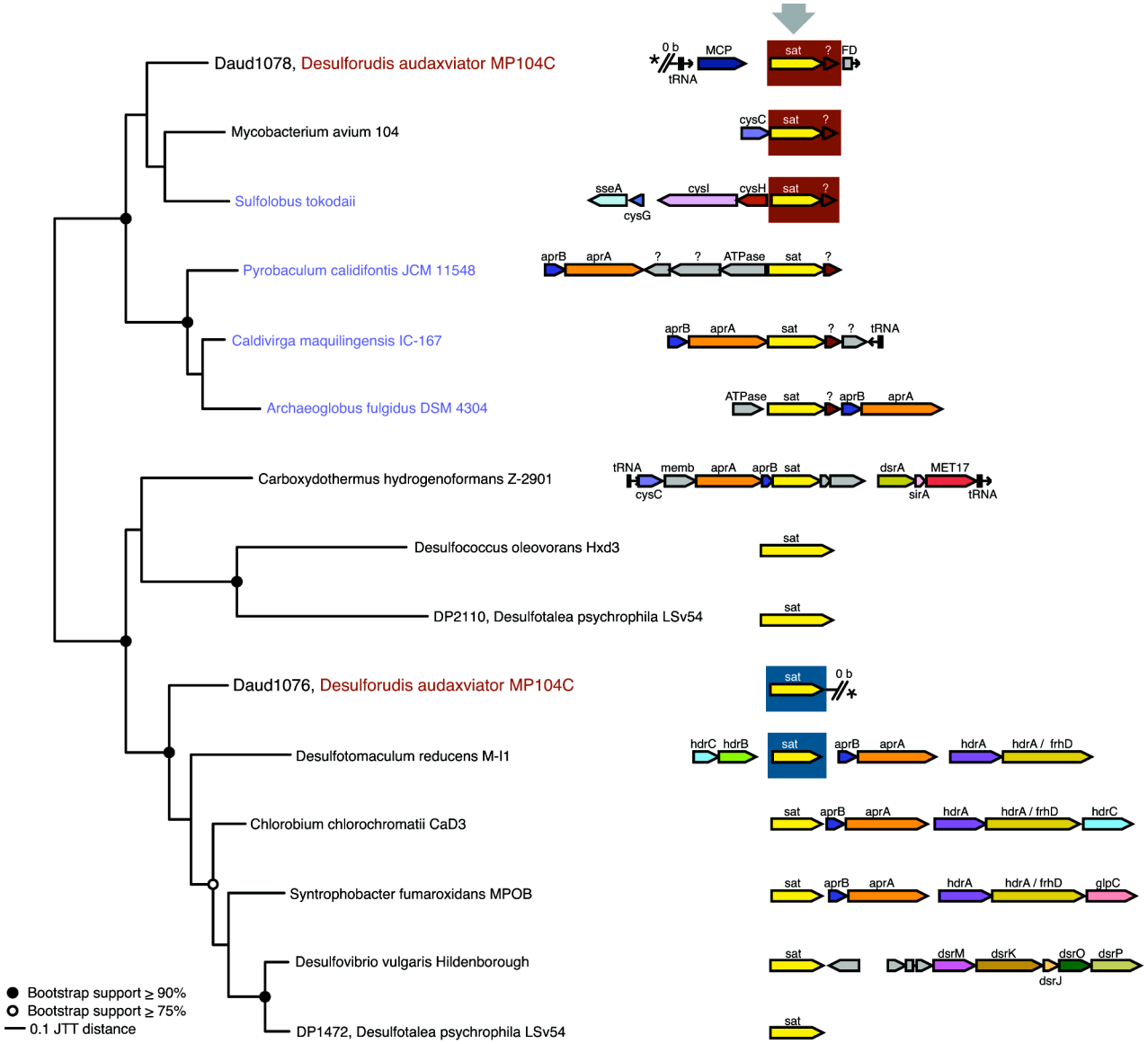

Daud1078, Desulforudis audaxviator MP104C
Mycobacterium avium 104
Sulfolobus tokodaii
Pyrobaculum calidifontis JCM 11548
Caldivirga maquilingensis IC-167
Archaeoglobus fulgidus DSM 4304
Carboxydothermus hydrogenoformans Z-2901
Desulfococcus oleovorans Hxd3
DP2110, Desulfotalea psychrophila LSv54
Daud1076, Desulforudis audaxviator MP104C
Desulfotomaculum reducens M-I1
Chlorobium chlorochromatii CaD3
Syntrophobacter fumaroxidans MPOB
Desulfovibrio vulgaris Hildenborough
DP1472, Desulfotalea psychrophila LSv54
Bootstrap support ≥ 90%
Bootstrap support ≥ 75%
0.1 JTT distance

Figure S6. DsrAB phylogenetic genome context analysis.

The dsrAB gene cluster has been shown to be subject to horizontal gene transfer, even between archaea and bacteria (*65*). Figure S6's gene context analysis coupled with phylogenetic analysis of the dsrAB genes (dsrA: Daud2201, dsrB: Daud2200) in *D. audaxviator* and other dsrAB containing bacteria and archaea reveals that *D. audaxviator* and *Moorella thermoacetica* have received the form of the dsrAB genes (shown in red boxes below) that resembles that found in *Desulfovibrio vulgaris*. This acquisition appears to have occurred after the divergence of *D. audaxviator* from *Desulfotomaculum reducens*. *D. audaxviator* does not possess additional copies of dsrAB and retains the context shared with *Desulfotomaculum reducens* of the ferredoxin(FD)-dsrMKCN operon and the other non-sulfate reducing genes shown below in dark gray and black within blue and green boxes. Interestingly, *M. thermoacetica* shows a phage-like gene immediately upstream of the dsr operon, whereas *D. audaxviator* has a transposon just downstream. *M. thermoacetica* additionally has another dsrAB-like cluster only 20 kb upstream that does not resemble any of the dsrAB genes in other sequenced organisms. The tree below is from a concatenated multiple sequence alignment of dsrA and dsrB built using MUSCLE (*46*), and determined by maximum likelihood by PHYML (*47*) with 100 replicates for bootstrapping (sampling with replacement), using the JTT amino acid substitution model (*48*).

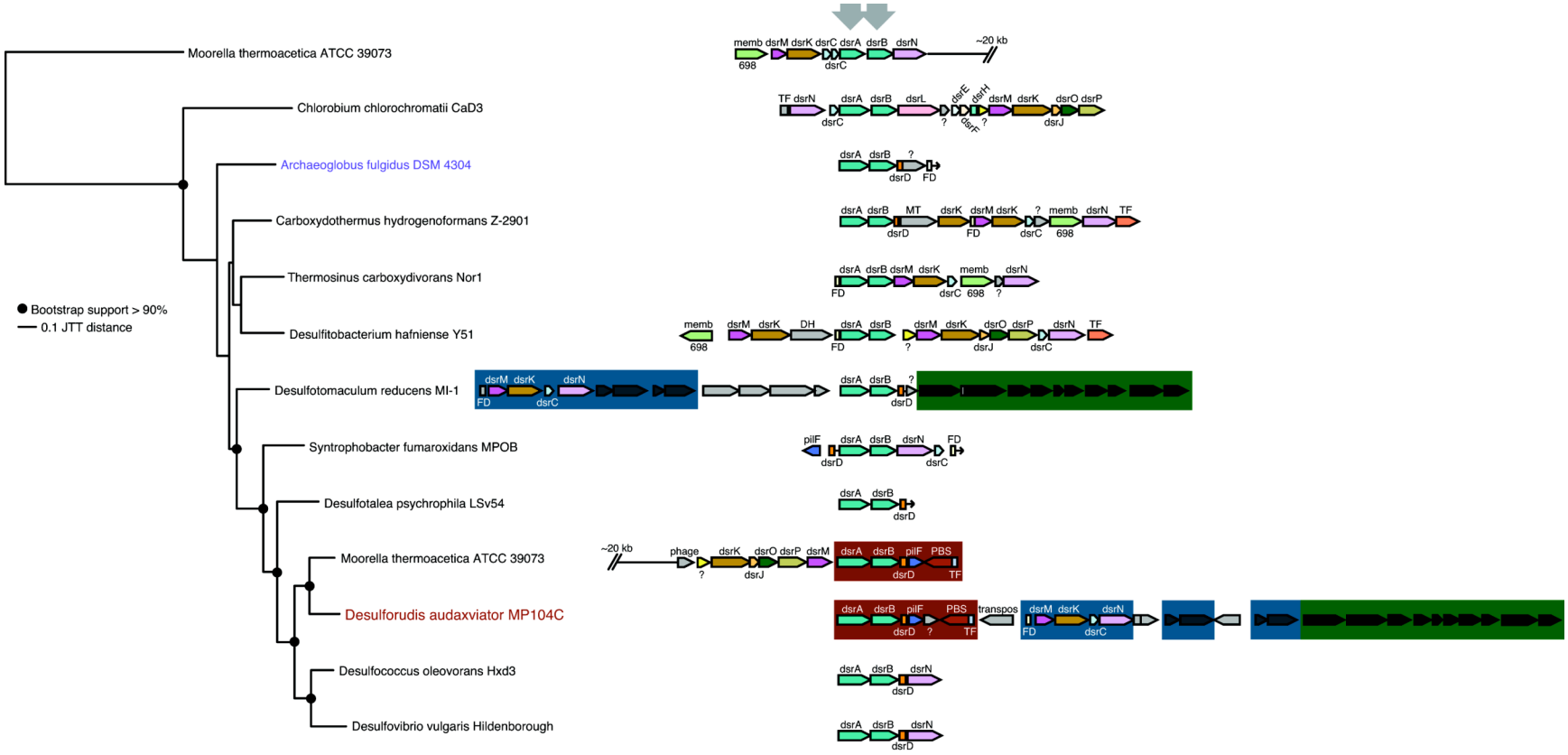

Moorella thermoacetica ATCC 39073
Chlorobium chlorochromatii CaD3
Archaeoglobus fulgidus DSM 4304
Carboxydothermus hydrogenoformans Z-2901
Thermosinus carboxydivorans Nor1
Desulfitobacterium hafniense Y51
Desulfotomaculum reducens MI-1
Syntrophobacter fumaroxidans MPOB
Desulfotalea psychrophila LSv54
Moorella thermoacetica ATCC 39073
Desulforudis audaxviator MP104C
Desulfococcus oleovorans Hxd3
Desulfovibrio vulgaris Hildenborough
Bootstrap support > 90%
0.1 JTT distance
~20 kb
transpos

Table S14 and Figure S7. Acetyl-CoA synthesis (Wood-Ljungdahl) and related carbon fixation genes.

Genes for assimilation of carbon from inorganic carbon (formate, CO, CO_2, bicarbonate, and carbonate) *via* the carbon monoxide dehydrogenase (CODH) / acetyl-CoA synthesis (Wood-Ljungdahl) pathway (*51*) were identified by membership in known sequence families (e.g. COG, TIGRFAM, and Pfam) or by gene context (proximity and/or presence in operons with other identified formate utilization and CODH/acetyl-CoA genes). Annotation was by protein family, or if no confident protein family could be assigned, by the protein family assignment of the nearest homolog (such annotations are indicated with square brackets, with the source organism provided in the notes).

D. audaxviator appears to have two CODH systems, one in operon CF2 that is similar to the CODH-III carbon fixation system of *C. hydrogenoformans* (*52*), and another system present in operon CF1 with formate dehydrogenase that resembles archaeal CODH (see Figure S7 below).

To determine whether the Wood-Ljungdahl pathway was functional the free energy of formation for the acetyl-CoA dehydrogenase synthase complex reported by Grahame and DeMoll (*66*) was used to calculate the free energy for acetyl-CoA synthesis from H_2 and CO_2, from CO and from formate and H^+ for the observed concentrations reported for the fracture environment and assuming an intracellular pH of 8.5. These calculations indicate that a H_2 partial pressure of ~0.1 atm is required for net synthesis of acetyl-CoA. This condition is met in the environments where *D. audaxviator* is prevalent. Uncertain is whether under low pH_2, the reverse reaction transfers electrons from acetate decomposition to sulfate reduction as hypothesized by Dai, et al. (*67*) for *A. fulgidus*. This favorable result is also dependent upon the intracellular pH as no gene for carbonic anhydrase has been detected in the genome and is dependent on the equilibrium conversion of CO_3^{2-} to CO_2. The free energy for synthesis of acetyl-CoA from CO was -240 to -270 kJ mol^{-1}. The free energy for synthesis of acetyl-CoA from formate was 3 to -21 kJ mol^{-1}, but is sensitive to the intracellular pH and formate concentrations, which are not known. Application of the Wood-Ljungdahl pathway may have the added benefit of Na^+ export to aid in maintaining the Na^+ gradient utilized by the Na^+ antiporters and symporters, including the Na^+/H^+ antiporter that could aid in driving ATP synthase (H^+-dependent). Na^+ could potentially be used by ATP synthase in very alkaline conditions, but it is not known whether the ATP synthase possessed by *D. audaxviator* is of the type that may use Na^+.

Table S14. CODH genes.

Gene	Name	Description	Operon	Len	CH id	CH species	Notes
Daud0103	folD	COG0190: Methenyl tetrahydrofolate cyclohydrolase	CF1	295	62.59	M. thermoacetica	
Daud0104	fdhA	COG3383, TIGR01591: Formate dehydrogenase,	CF1	725	46.92	C. hydrogenoformans	

		alpha subunit					
Daud0105	cdhA / cooS	COG1152, TIGR00314: CODH/acetyl-CoA synthase complex alpha subunit	CF1	767	52.37	M. thermautotrophicus	
Daud0106	cdhB / acsE	COG1880: TIGR00315: CODH beta subunit/acetyl-CoA synthase epsilon subunit	CF1	185	38.18	M. maripaludis	
Daud0107	cooC	COG3640: CODH maturation factor, PF01656: Cobyrinic acid a,c-diamide synthase	CF1	252	54.98	P. thermopropionicum	
Daud0108	nuoE	COG1810: Conserved in archaea & COG1905: NADH:ubiquinone oxidoreductase	CF1	378	53.15	C. hydrogenoformans	
Daud0109	nuoF	COG1894: NADH:ubiquinone oxidoreductase, NADH-binding subunit	CF1	657	60.1	T. tengcongensis	
Daud0110	fdhA-like	COG3383, [TIGR01591: Formate dehydrogenase, alpha subunit]	CF1	999	42.61	D. psychrophila	D.psychrophila annot.
Daud0111	cdhC / acsB	COG1614, TIGR00316: CODH/acetyl-CoA synthase complex beta subunit	CF1	400	58.52	M. jannaschii	
Daud0112	cdhE / acsC	COG1456: CODH/acetyl-CoA synthase gamma subunit	CF1	453	43.57	D. ethenogenes	
Daud0113		COG1410: Methionine synthase I, cobalamin-binding domain	CF1	291	40.93	C. hydrogenoformans	
Daud0114	cdhD / acsD	COG2069: CODH/acetyl-CoA synthase delta subunit	CF1	299	42.11	D. ethenogenes	
Daud0115	frhD	COG1908: Coenzyme F420-reducing hydrogenase, delta subunit	CF1	174	44.81	M. thermoacetica	
Daud0116		Zinc-finger protein?	CF1	222	50.23	P. thermopropionicum	
Daud0117	metF	COG0685: 5,10-methylenetetrahydrofolate reductase	CF1	307	56.77	M. thermoacetica	
Daud0118	fhs	COG2759: Formyltetrahydrofolate synthetase	CF1	558	64.99	C. hydrogenoformans	
Daud0119	RR-lytR	COG3279: Response regulator of the LytR/AlgR family	CF1	250	49.4	C. hydrogenoformans	
Daud0868		COG3937: Uncharacterized conserved protein	CF2	119	30.77	C. hydrogenoformans	us tRNA-pro
Daud0869		COG1524: Uncharacterized proteins of the AP superfamily	CF2	378	36.94	D. reducens	
Daud0870	cooS	COG1151, TIGR01702: CODH, catalytic subunit	CF2	629	51.26	M. mazei	
Daud0871	cdhC /	COG1152: CODH alpha subunit & COG1614,	CF2	736	52.45	D. ethenogenes	

	acsB	TIGR00316: CODH beta subunit					
Daud0872	cdhE / acsC	COG1456: CODH/acetyl-CoA synthase gamma subunit	CF2	444	60.22	C. hydrogenoformans	
Daud0873		COG3894: Uncharacterized metal-binding protein	CF2	631	50.55	P. thermopropionicum	
Daud0874	cooC	COG3640: CODH maturation factor, PF01656: Cobyrinic acid a,c-diamide synthase	CF2	259	51.23	C. hydrogenoformans	
Daud0875	cdhD / acsD	COG2069: CODH/acetyl-CoA synthase delta subunit	CF2	327	61.54	P. thermopropionicum	
Daud0876		COG1410: Methionine synthase I, cobalamin-binding domain	CF2	282	69.47	C. hydrogenoformans	acsE? in C.hydrogenoformans
Daud0877	hdrA-like	COG1148: Heterodisulfide reductase, subunit A and related polyferredoxins	CF2	996	45.08	D. reducens	
Daud0878	frhD	COG1908: Coenzyme F420-reducing hydrogenase, delta subunit	CF2	146	53.24	M. thermoacetica	
Daud0879		Zinc-finger protein?	CF2	222	52.58	P. thermopropionicum	
Daud0880	metF	COG0685: 5,10-methylenetetrahydrofolate reductase	CF2	315	62.33	M. thermoacetica	
Daud1098	echA / mnhA	COG1009: NADH:ubiquinone oxidoreductase subunit 5 (chain L)/Multisubunit Na+/H+ antiporter, MnhA subunit	CF3	654	45.37	D. ethenogenes	Na+/H+ antiport?
Daud1099	echB / hyfC / cook	COG650: Formate hydrogenlyase subunit 4	CF3	291	51.62	D. ethenogenes	Na+/H+ antiport? Ortholog of C. hydrogenoformans cooK (CODH-I)
Daud1100	echC / cooL	COG3260: Ni,Fe-hydrogenase III small subunit	CF3	142	72.46	Dehalo. sp. CBDB1	Na+/H+ antiport? Ortholog of C. hydrogenoformans cooL (CODH-I)
Daud1101	echD	PIRSF036585 [NiFe]-hydrogenase-3-type complex Ech, subunit EchD	CF3	121	36.54	D. ethenogenes	Na+/H+ antiport?
Daud1102	hycE / cooH	COG3261: Ni,Fe-hydrogenase III large subunit	CF3	359	64.62	D. ethenogenes	Na+/H+ antiport? Ortholog of C. hydrogenoformans cooH (CODH-I)
Daud1103	echF /	COG1143: Formate hydrogenlyase subunit	CF3	127	41.44	T. tengcongensis	Na+/H+ antiport?

	nuoI	6/NADH:ubiquinone oxidoreductase 23 kD subunit (chain I)					
Daud1104	fmdE	COG1405: Formylmethanofuran dehydrogenase subunit E	CF3	198	42.56	M. thermoacetica	
Daud1105		hypothetical protein	CF3	281	25.16	C. hydrogenoformans	
Daud1569		hypothetical protein	CF4A	55	N/A	ORFan	
Daud1570	Na^+ symport	COG4147, PF00474 Na^+/solute symporter	CF4A	694	34.68	S. avermitilis	
Daud1571		hypothetical protein	CF4A	123	32.1	N. farcinica	
Daud1572	acsA	COG0365, TIGR02188: Acetate--CoA ligase	CF4A	661	66.51	C. hydrogenoformans	
Daud1573	paaK-1	COG1541: Coenzyme F390 synthetase [phenylacetate-coenzyme A ligase]	CF4B	432	63.72	D. reducens	C.hydrogenoformans annot.
Daud1574	paaK-2	COG1541: Coenzyme F390 synthetase [phenylacetate-coenzyme A ligase]	CF4B	442	67.36	M. thermoacetica	C.hydrogenoformans annot.
Daud1575		COG4747: ACT domain-containing protein	CF4B	144	66.43	M. thermoacetica	
Daud1576	dfx?	TIGR00319: Desulfoferrodoxin Dfx (short?)	CF4B	40	61.11	Dehalo. sp. CBDB1	truncated?
Daud1577	fdhE	COG3058: Involved in formate dehydrogenase formation	CF4C	283	39.23	P. thermopropionicum	
Daud1578	hybB	COG5557, [probable cytochrome Ni/Fe component of hydrogenase-2]	CF4C	395	45.08	P. thermopropionicum	2 paralogs: Daud1044 (97.5%) Daud0160 (80.8%); E.coli K12 annot.
Daud1579	fdoH / hybA	COG0437, [formate dehydrogenase beta subunit]	CF4C	263	49.61	D. reducens	S.thermophilum annot.
Daud1580	fdoG	COG0243, [TIGR01553: Formate dehydrogenase, alpha]	CF4C	809	61.62	P. thermopropionicum	S.thermophilum annot.
Daud1581	fdoG-2	COG0243, TIGR01409: Twin-arginine translocation pathway signal	CF4C	217	49.07	P. thermopropionicum	
Daud2010	fdhD	COG1526, TIGR00129: Formate dehydrogenase, subunit FdhD	CF5	249	38.08	P. thermopropionicum	
Daud2158	mnhD-1	COG0651: Formate hydrogenlyase subunit 3/Multisubunit Na+/H+ antiporter, MnhD	CF6	590	55.17	R. etli	Na+/H+ antiport?
Daud2159		hypothetical protein	CF6	75	43.28	R. rubrum	TF?

Daud2160	mnhD-2	COG0651: Formate hydrogenlyase subunit 3/Multisubunit Na+/H+ antiporter, MnhD	CF6	555	46.96	H. marismortui	Na+/H+ antiport?
Daud2161	mnhD-3	COG0651: Formate hydrogenlyase subunit 3/Multisubunit Na+/H+ antiporter, MnhD	CF6	502	41.88	D-monas spp.	Na+/H+ antiport?
Daud2162	mnhC	COG1006: Multisubunit Na+/H+ antiporter, MnhC subunit	CF6	131	44.35	D-monas spp.	Na+/H+ antiport?
Daud2163	mnhB-1	PF04039: Na+/H+ antiporter MnhB subunit-related protein	CF6	144	38.24	S. frigidimarina	Na+/H+ antiport?
Daud2164	mnhB-2	[COG2111: Multisubunit Na+/H+ antiporter, MnhB subunit]	CF6	98	48.86	D-monas spp.	Na+/H+ antiport?; D-monas spp. annot.
Daud2165	mnhB-3	[COG1563: Predicted subunit of the Multisubunit Na+/H+ antiporter]	CF6	80	37.33	D-monas spp.	Na+/H+ antiport?; D-monas spp. annot.
Daud2166	mnhG	COG1320: Multisubunit Na+/H+ antiporter, MnhG subunit, TIGR01300	CF6	103	45.19	D-monas spp.	Na+/H+ antiport?
Daud2167	mnhF	[COG2212: Multisubunit Na+/H+ antiporter, MnhF subunit]	CF6	95	42.17	D-monas spp.	Na+/H+ antiport?; P.horikoshii annot.
Daud2168	mnhE	COG1863: Multisubunit Na+/H+ antiporter, MnhE subunit	CF6	203	40.88	D-monas spp.	Na+/H+ antiport?

Figure S7. CODH catalytic subunit phylogenetic genome context analysis.

Desulforudis audaxviator has two CODH gene clusters, each with a phylogenetically distinct catalytic subunit. The first, with Daud0870, corresponds to COG 1151 (cooS) and is in the acetogenic CODH-III family of *Carboxydothermus hydrogenoformans* (*52*). COG 1151 is a fairly broad family but, with the additional requirement that cdhC/acsB was found in the neighborhood, the CODH-III type of cooS was only found in the sequenced bacterial genomes at the time of this analysis. The other CODH gene cluster, with Daud0105, corresponds to COG 1152 (cdhA) and, other than *D. audxaviator*, was only found in the sequenced archaeal genomes included at the time of this analysis. Daud0870 and Daud0105 are distantly related, showing sequence identity of ~27-30% and can be aligned with other members of COG1151 and COG1152 to make the tree of Figure S7. Other genes in the cluster that are closest to archaeal homologs are shown in the red boxes, whereas those genes that are closest to bacterial homologs are shown in green and blue boxes. Interestingly, the archaeal-type gene cluster includes some genes that more closely resemble their counterparts in the bacterial cluster (cooC, cdhE/acsC, acsE, frhD, Zn-finger, and metF). The tree below is from a multiple sequence alignment of COG1151 and COG1152 using MUSCLE (*46*), and determined by maximum likelihood by PHYML (*47*) with 100 replicates for bootstrapping

(sampling with replacement), using the JTT amino acid substitution model (*48*). The cooS and cdhA genes are indicated by the gray arrow. Archaeal species names are blue. The gap between the *Methanopyrus kandleri* cdhA gene duplication is zero bases. The *C. hydrogenoformans* protein sequence used was as determined after removal of the pseudogene-inducing frame shift present in the sequenced strain.

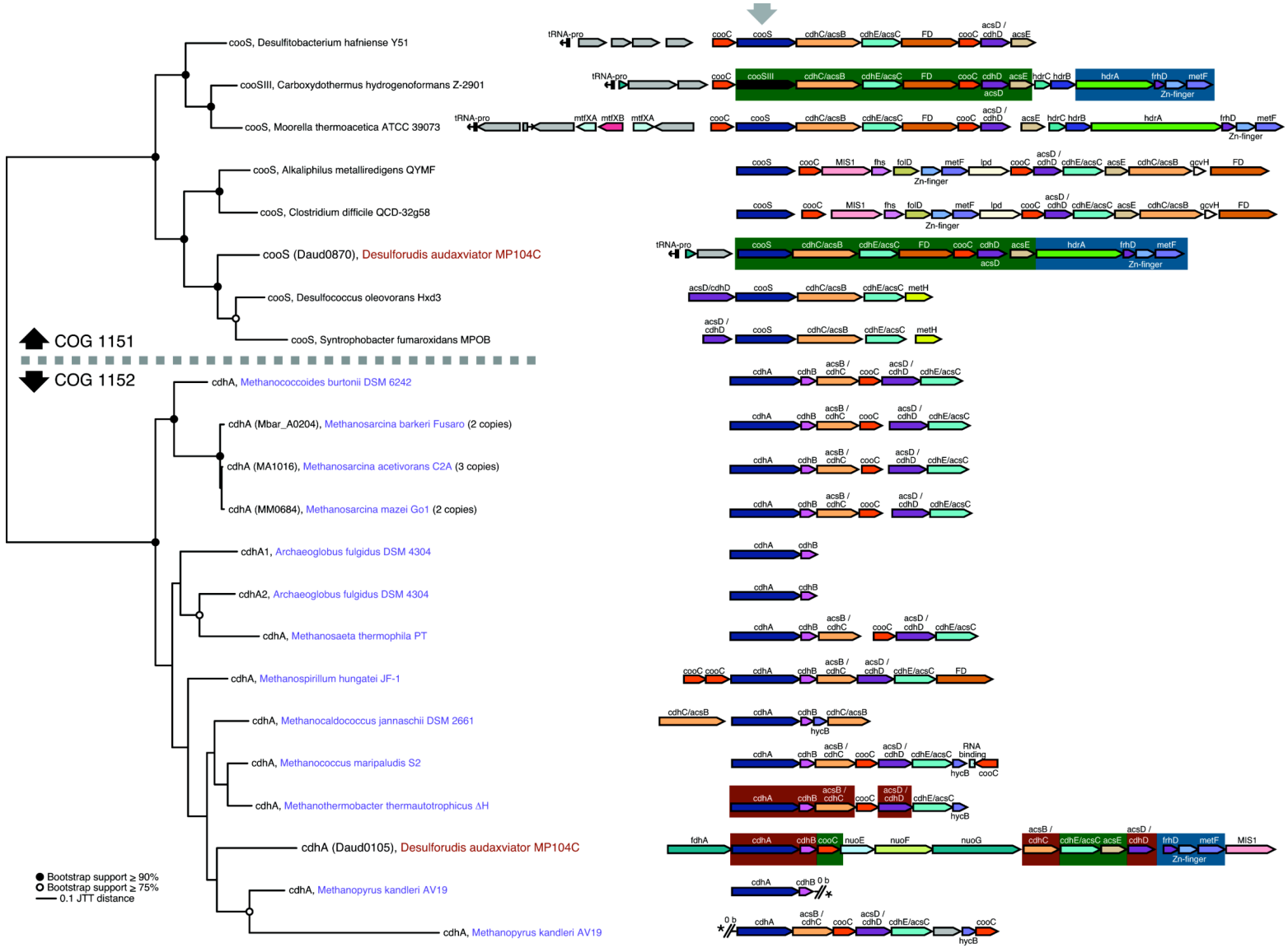

cooS, Desulfitobacterium hafniense Y51
cooSIII, Carboxydothermus hydrogenoformans Z-2901
cooS, Moorella thermoacetica ATCC 39073
cooS, Alkaliphilus metalliredigens QYMF
cooS, Clostridium difficile QCD-32g58
cooS (Daud0870), Desulforudis audaxviator MP104C
cooS, Desulfococcus oleovorans Hxd3
cooS, Syntrophobacter fumaroxidans MPOB
COG 1151
COG 1152
cdhA, Methanococcoides burtonii DSM 6242
cdhA (Mbar_A0204), Methanosarcina barkeri Fusaro (2 copies)
cdhA (MA1016), Methanosarcina acetivorans C2A (3 copies)
cdhA (MM0684), Methanosarcina mazei Go1 (2 copies)
cdhA1, Archaeoglobus fulgidus DSM 4304
cdhA2, Archaeoglobus fulgidus DSM 4304
cdhA, Methanosaeta thermophila PT
cdhA, Methanospirillum hungatei JF-1
cdhA, Methanocaldococcus jannaschii DSM 2661
cdhA, Methanococcus maripaludis S2
cdhA, Methanothermobacter thermautotrophicus ΔH
cdhA (Daud0105), Desulforudis audaxviator MP104C
cdhA, Methanopyrus kandleri AV19
cdhA, Methanopyrus kandleri AV19
Bootstrap support ≥ 90%
Bootstrap support ≥ 75%
0.1 JTT distance

Table S15 and Figure S8. Nitrogen fixation genes.

Genes for assimilation of nitrogen from N_2 and ammonia were identified by membership in known sequence families (e.g. COG, TIGRFAM, and Pfam) or by gene context (proximity and/or presence in operons with other identified nitrogen fixation genes). Annotation was by protein family, or if no confident protein family could be assigned, by the protein family assignment of the nearest homolog (such annotations are indicated with square brackets, with the source organism provided in the notes).

Parts of the NF1 nitrogenase operon appears to have been horizontally transferred from archaea, and based on the phylogeny of the nifH subunit (Fig. S4), groups with Molybdenum-containing nitrogenases that can sometimes function at quite high temperatures(*59*). The maintenance of nitrogenase function in the presence of high CO concentrations (CO, like O_2, inhibits the functioning of nitrogenase) may be assited by CODH-mediated removal of CO within the cell.

Table S15. Nitrogen fixation genes.

Gene	Name	Description	Operon	Len	CH id	CH species	Notes
Daud0141	glnB	COG0347: Nitrogen regulatory protein PII	NF1	112	64.22	D. hafniense Y51	
Daud0142	amt	COG0004, TIGR00836: Ammonium permease	NF1	465	58.64	D. hafniense Y51	
Daud0143	nifH	COG1348: Nitrogenase subunit NifH (ATPase)	NF1	281	79.27	M. thermautotrophicus	HGT; high temp?
Daud0144	nifI1	COG0347: Nitrogen regulatory protein PII	NF1	106	64	M. maripaludis	HGT; high temp?
Daud0145	nifI2	COG0347: Nitrogen regulatory protein PII	NF1	121	51.64	M. acetivorans C2A	HGT; high temp?
Daud0146	nifD	COG2710: Nitrogenase molybdenum-iron protein, alpha and beta chains	NF1	483	67.86	M. thermoacetica	HGT?; high temp?
Daud0147	nifK	COG2710: Nitrogenase molybdenum-iron protein, alpha and beta chains	NF1	491	46.55	M. maripaludis	HGT; high temp?
Daud0148	nifE	COG2710, [TIGR01283: Nitrogenase MoFe cofactor biosynthesis protein NifE]	NF1	460	38.86	M. thermautotrophicus	HGT; M.thermoautotrophicus annot.
Daud0149	nifB	COG0535, [Nitrogenase cofactor biosynthesis protein NifB, putative]	NF1	286	54.68	G. sulfurreducens	D.ethenogenes annot.
Daud0150	nadE	COG0388: Amidohydrolase & COG0171, TIGR00552: NAD+ synthase	NF1	544	55.83	M. thermoacetica	
Daud0473	glnB/K	PF00543: Nitrogen regulatory protein P-II (GlnB, GlnK)	NF2	118	54.21	M. hungatei	

Daud0474	MEMB	PF07556: DUF1538 [putative membrane protein]	NF2	246	58.75	C. hydrogenoformans	
Daud0475	MEMB	PF07556: DUF1538 [putative membrane protein]	NF2	231	55.46	C. hydrogenoformans	
Daud1060	nifU	COG0694, PF01106: Nitrogen-fixing NifU, C-terminal	NF3	44	74.36	Bacillus clausii	
Daud2023	glnA	COG0174, TIGR00653: Glutamine synthetase type I	NF4	444	74.22	D. reducens	
Daud2024	gltB-3	COG0070: Glutamate synthase domain 3	NF4	247	64.08	C. hydrogenoformans	
Daud2025	gltB-2	COG0069: Glutamate synthase domain 2	NF4	531	67.11	C. hydrogenoformans	
Daud2026	gltB-1	COG0067: Glutamate synthase domain 1	NF4	360	64.04	C. hydrogenoformans	
Daud2027	glnA	COG0174, TIGR00653: Glutamine synthetase type I	NF4	444	72.52	D. reducens	
Daud2028	amiR-2	COG3707: Response regulator with putative antiterminator output domain	NF4	193	50.8	D. reducens	
Daud2029	gltB-2	COG0069: Glutamate synthase domain 2	NF4	526	73.9	P. thermopropionicum	
Daud2030	glnA	COG0174: Glutamine synthetase	NF4	446	56.39	P. thermopropionicum	
Daud2031	amiR-1	COG3707: Response regulator with putative antiterminator output domain	NF4	199	46.49	P. thermopropionicum	
Daud2032	ilvE	COG0115: Branched-chain amino acid aminotransferase (class IV)	NF4	286	34.74	M. kandleri	

Figure S8. NifH phylogenetic and genome context analysis.
Desulforudis audaxviator has one nif nitrogen fixation gene cluster. The genes in this cluster do not possess sufficient homology at the per-gene level to confidently place them in a nif subfamily or gene context similarity to a known nif cassette. While its relative, *Desulfotomaculum reducens*, also has a nif gene cluster, only some of the genes in this cluster appear to be related by vertical decent to those in *D. audaxviator* (labeled with green and blue boxes). The other nif genes in *D. audaxviator* (nifD, nif K, nifE, nifB) are quite distant from their homologs in the other genomes cannot be placed clearly within a phylogenetic context. These genes probably represent new subfamilies, ones that may not require nifN as it is not found in the *D. audaxviator* genome. The other nif genes (nifH, nifI1, nifI2), while also difficult to place phylogenetically, do appear to group with archaeal-type nitrogenases (labeled with red boxes). The tree below is from a multiple sequence alignment of nifH (with anfH and vnfH) using MUSCLE (*46*), and determined by maximum likelihood by PHYML (*47*) with 100 replicates for bootstrapping (sampling with replacement), using the JTT amino acid substitution model (*48*). The nifH gene is indicated by the gray arrow. Archaeal species names are blue. While the nifH possessed by *D. audaxviator* is closest to the high temperature archaeal cluster, low bootstrap supported nodes with short branch lengths do not permit its confident phylogenetic placement. However, these sequences are sufficient to determine that the nifH possessed by *D. audaxviator* is not related by vertical decent to that possessed by *D. reducens*.

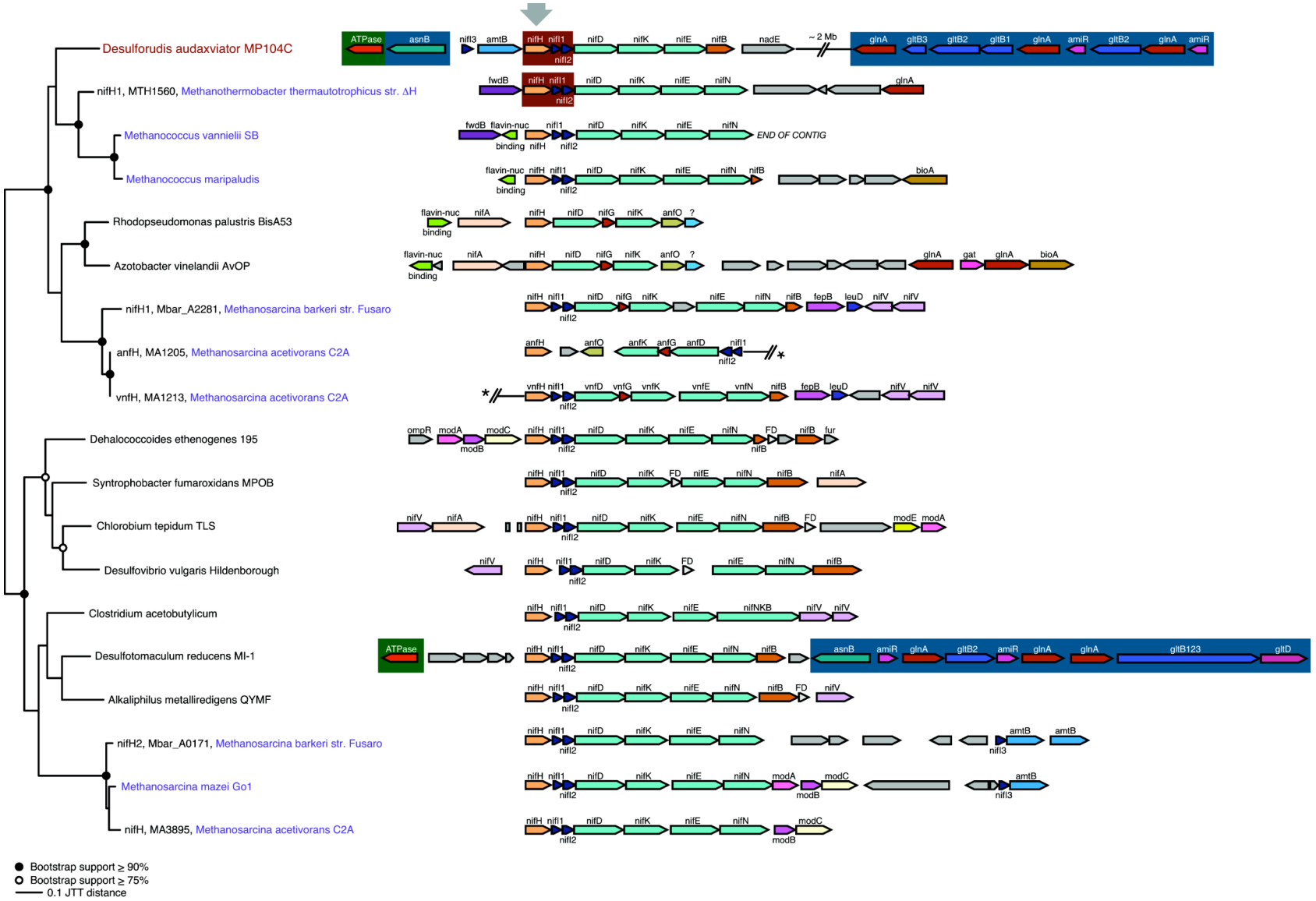

Desulforudis audaxviator MP104C
nifH1, MTH1560, Methanothermobacter thermautotrophicus str. ΔH
Methanococcus vannielii SB
Methanococcus maripaludis
Rhodopseudomonas palustris BisA53
Azotobacter vinelandii AvOP
nifH1, Mbar_A2281, Methanosarcina barkeri str. Fusaro
anfH, MA1205, Methanosarcina acetivorans C2A
vnfH, MA1213, Methanosarcina acetivorans C2A
Dehalococcoides ethenogenes 195
Syntrophobacter fumaroxidans MPOB
Chlorobium tepidum TLS
Desulfovibrio vulgaris Hildenborough
Clostridium acetobutylicum
Desulfotomaculum reducens MI-1
Alkaliphilus metalliredigens QYMF
nifH2, Mbar_A0171, Methanosarcina barkeri str. Fusaro
Methanosarcina mazei Go1
nifH, MA3895, Methanosarcina acetivorans C2A
~ 2 Mb
END OF CONTIG
Bootstrap support ≥ 90%
Bootstrap support ≥ 75%
0.1 JTT distance

Table S16. Sporulation and germination genes.

Sporulation and germination genes identified as orthologs (by reciprocal best BLASTp hit) to *C. hydrogenoformans* sporulation and germination genes. Sporulation and germination genes in *C. hydrogenoformans* were identified by Wu, *et al*. (*52*) using orthology to known spore forming genes in *B. subtils* (CHY_1978 to CHY_0424 in below table) and by phenotype footprint technique to identify genes associated with spore formers and not associated with non-spore formers (CHY_0020 to CHY_2676). Names taken from closest *B. subtilis* homolog (not necessarily an ortholog). "N/A": no homolog was detected in *B. subtilis* from which to derive the name. "N/D": no ortholog was detected within the genome. Most names and descriptions taken directly from Wu, *et al*. Additional putative sporulation and germination genes in *D. audaxviator* that are not orthologous to *C. hydrogenoformans* sporulation and germination genes are not reported.

Name	C.hyd gene	B.sub gene	D.aud gene	Description
spo0A	CHY_1978	Bsu2420	Daud1615	Stage 0 sporulation protein A
spo0J	CHY_0010	Bsu4093	Daud2229	Stage 0 sporulation protein J
obg	CHY_0370	Bsu2788	Daud1873	spo0B-associated GTP-binding protein
soj	CHY_0009	Bsu4094	Daud2230	Sporulation initiation inhibitor protein soj
spoIIAB	CHY_1960	Bsu2345	Daud1225	AB Anti-sigma F factor
spoIID	CHY_2541	Bsu3673	Daud2135	Stage II sporulation protein D
spoIID	CHY_1517	N/D	N/D	Putative stage II sporulation protein D
spoIIE	CHY_0212	Bsu0064	Daud0085	Putative stage II sporulation protein E
spoIIGA	CHY_2057	N/D	Daud1427	Putative sporulation specific protein SpoIIGA
spoIIM	CHY_1965	Bsu2352	N/D	Putative stage II sporulation protein M
spoIIP	CHY_1923	N/D	Daud1176	Putative stage II sporulation protein P
spoIIP	CHY_0408	N/D	N/D	Putative sporulation protein
spoIIR	CHY_2054	N/D	N/D	Stage II sporulation protein R
spoIID	CHY_0206	N/D	Daud1268	Putative stage II sporulation protein D
spoIIIAA	CHY_2007	Bsu2441	Daud1007	Putative sporulation protein
spoIIIAB	CHY_2006	Bsu2440	Daud1008	Putative sporulation protein
spoIIIAC	CHY_2005	Bsu2439	Daud1009	Putative sporulation protein
spoIIIAD	CHY_2004	Bsu2438	Daud1010	Putative sporulation protein

spoIIIAE	CHY_2003	Bsu2437	Daud1011	Putative sporulation protein
spoIIIAG	CHY_2001	N/D	Daud1013	Putative sporulation protein
spoIIID	CHY_2534	Bsu3640	Daud2134	Stage III sporulation protein D
spoIIIE	CHY_1159	Bsu1681	Daud0837	DNA translocase FtsK
spoIIIJ	CHY_0004	Bsu4101	N/D	Sporulation associated-membrane protein
spoIVA	CHY_1916	Bsu2279	Daud0897	Stage IV sporulation protein A
spoIVB	CHY_1979	Bsu2421	Daud1616	Putative stage IV sporulation protein B
spoVAC	CHY_1957	Bsu2341	Daud1222	Stage V sporulation protein AC
spoVAD	CHY_1956	Bsu2340	Daud1221	Stage V sporulation protein AD
spoVAE	CHY_1955	N/D	Daud1220	Stage V sporulation protein AE
spoVB	CHY_0960	Bsu2763	Daud1230	Stage V sporulation protein B
spoVFA	CHY_1152	Bsu1674	Daud0943	Dipicolinate synthase, A subunit
spoVFB	CHY_1153	Bsu1675	Daud0944	Dipicolinate synthase, B subunit
spoVK	CHY_1391	Bsu1743	N/D	Stage V sporulation protein K
spoVR	CHY_1202	Bsu0940	Daud0593	Stage V sporulation protein R
spoVS	CHY_1171	Bsu1699	Daud1063	Stage V sporulation protein S
spoVT	CHY_0202	Bsu0056	Daud0073	Stage V sporulation protein T
cotJC	CHY_2272	N/D	N/D	cotJC protein
cotJC	CHY_0786	N/D	N/D	cotJC protein
sspD	CHY_1463	N/D	N/D	Small acid-soluble spore protein
sspD	CHY_1464	Bsu1349	N/D	Small acid-soluble spore protein
sspF	CHY_1175	N/D	N/D	Small acid-soluble spore protein
N/A	CHY_1465	N/D	N/D	Putative small acid-soluble protein
spmA	CHY_1941	Bsu2317	Daud1198	Spore maturation protein A
spmB	CHY_1940	Bsu2316	Daud1197	Spore maturation protein B
N/A	CHY_0958	N/D	N/D	Small acid-soluble spore protein
sleB	CHY_1160	N/D	Daud2186	Putative spore cortex-lytic enzyme
sleB	CHY_1756	N/D	N/D	Putative spore cortex-lytic enzyme
gerKA	CHY_0336	Bsu0371	Daud1936	Spore germination protein GerKA
gerKB	CHY_1404	Bsu0373	N/D	Spore germination protein

gerKC	CHY_0337	Bsu0372	Daud1935	Spore germination protein
gerM	CHY_0305	N/D	N/D	Putative germination protein GerM
yndE	CHY_1950	Bsu1777	Daud1937	Putative spore germination protein
sigX	CHY_0143	N/D	N/D	RNA polymerase sigma factor
sigE	CHY_2056	Bsu1534	Daud1426	RNA polymerase sigma-E factor
sigF	CHY_1959	Bsu2344	Daud1224	RNA polymerase sigma-F factor
sigG	CHY_2055	Bsu1535	Daud1425	RNA polymerase sigma-G factor
sigH	CHY_2333	Bsu0098	Daud0192	RNA polymerase sigma-H factor
sigK	CHY_0617	N/D	N/D	RNA polymerase sigma-K factor
gpr	CHY_1462	Bsu2550	Daud0631	Spore protease
bofA	CHY_2672	N/D	N/D	Sigma-K processing regulatory protein BofA
yqfD	CHY_0424	Bsu2531	Daud2049	Putative sporulation protein
yyaC	CHY_0020	Bsu4092	Daud2222	conserved hypothetical protein
ykvI	CHY_0021	Bsu1373	Daud2221	putative membrane protein
yybS	CHY_0038	Bsu4049	Daud2213	putative membrane protein
yvjD	CHY_0171	Bsu3519	Daud0288	putative membrane protein
yabP	CHY_0207	Bsu0060	Daud0077	conserved hypothetical protein
ypjA	CHY_0278	Bsu2252	N/D	putative membrane protein
lonB	CHY_0329	Bsu2817	Daud1477	putative ATP-dependent protease La
yqfC	CHY_0423	N/D	Daud2050	conserved hypothetical protein
yqzB	CHY_0441	Bsu2521	Daud0486	CBS domain protein
yoaR	CHY_0544	N/D	Daud0917	vanW domain protein
yqfV	CHY_0651	Bsu2507	N/D	transcription regulator, Fur family
ypfP	CHY_1043	Bsu2193	N/D	putative glycosyl transferase
yerC	CHY_1082	Bsu0659	N/D	conserved hypothetical protein
dapG	CHY_1155	Bsu1677	N/D	aspartate kinase, monofunctional class
ypeB	CHY_1161	Bsu2291	Daud2185	conserved hypothetical protein
ydbG	CHY_1367	Bsu0446	Daud1257	C4-dicarboxylate response regulator
ynbB	CHY_1390	Bsu1745	Daud0558	conserved hypothetical protein

ylpC	CHY_1452	Bsu1589	N/D	conserved hypothetical protein
ylbJ	CHY_1457	Bsu1505	Daud0635	putative membrane protein
yloH	CHY_1487	Bsu1570	Daud1596	RpoZ DNA-directed RNA polymerase, omega subunit
ylzA	CHY_1489	Bsu1568a	Daud1598	conserved hypothetical protein
ykoZ	CHY_1519	Bsu1347	Daud2038	RNA polymerase sigma facter
yviA	CHY_1529	Bsu3545	N/D	degV family protein
yunB	CHY_1560	Bsu3232	Daud1367	conserved hypothetical protein
ytxC	CHY_1589	Bsu2892	N/D	conserved hypothetical protein
ylaJ	CHY_1593	N/D	Daud1400	putative lipoprotein
ytaF	CHY_1648	N/D	Daud1405	putative membrane protein
glpP	CHY_1843	Bsu0927	N/D	glycerol uptake operon antiterminator regulator
pheB	CHY_1913	Bsu2787	N/D	ACT domain protein PheB
ytfJ	CHY_1943	Bsu2946	Daud1199	conserved hypothetical protein
ywcB	CHY_2034	Bsu3819	N/D	conserved hypothetical protein
ylmC	CHY_2053	Bsu1538	Daud1423	PRC-barrel domain protein
cwlD	CHY_2271	Bsu0153	Daud0332	N-acetylmuramoyl-L-alanine amidase
yacK	CHY_2346	Bsu0088	Daud0183	putative DNA-binding protein
yacI	CHY_2349	Bsu0085	Daud0180	ATP:guanido phosphotransferase domain protein
yacH	CHY_2350	Bsu0084	Daud0179	uvrB/uvrC motif domain protein
ywqE	CHY_2481	Bsu3622	Daud1719	putative phosphoesterase
ykwD	CHY_2600	N/D	N/D	SCP-like extracellular protein
yabG	CHY_2611	Bsu0043	Daud0048	YabG peptidase, U57 family
yabE	CHY_2617	N/D	Daud0044	conserved hypothetical protein
abrB	CHY_2622	Bsu0037	Daud0040	transcriptional regulator, AbrB family
ylxP	CHY_2676	Bsu1665	Daud0887	conserved hypothetical protein

Table S17. Pilus genes.

Genes for pilus formation were identified by membership in known sequence families (e.g. COG, TIGRFAM, and Pfam) or by gene context (proximity and/or presence in operons with other identified pilus genes). Annotation was by protein family, or if no confident protein family could be assigned, by the protein family assignment of the nearest homolog (such annotations are indicated with square brackets, with the source organism provided in the notes). The majority of matched protein families indicate type IV pili ("Tfp"), but such protein families also often include type II-like genes. One pilus assembly gene (Daud2198) is found in a sulfate reduction operon.

Gene	Name	Description	Operon	Len	CH id	CH species	Notes
Daud2198	pilF	COG3063? Type IV pilus (Tfp) assembly?	SR11	210	62.93	M. thermoacetica	Sulfate reduction?
Daud0951	pilT	COG2805 Tfp pilus assembly, pilus retraction ATPase	P1	373	68.38	P. thermoproprionicum	us transposon
Daud0952	pulF / pilC	COG1459 Type II secretory (PulF), type IV pilus biogenesis (PilC)	P1	403	53.88	D. reducens	
Daud0953	pilM	COG4972 Tfp pilus assembly protein, ATPase	P1	346	43.91	P. thermoproprionicum	
Daud0954	pilN	COG3166 Tfp pilus assembly protein	P1	201	31.31	P. thermoproprionicum	
Daud0955	pilO	COG3167 Tfp pilus assembly protein	P1	193	35.93	P. thermoproprionicum	
Daud0956	pulE / pilB	COG2804 Type II secretory, ATPase (PulE), Tfp pilus assembly, ATPase (PilB)	P1	562	64.98	P. thermoproprionicum	
Daud0957	pilE	COG4968 Tfp pilus assembly protein	P1	138	32.81	V. vulnificus	ds transposon
Daud0961		TF02532 Prepilin-type cleavage/methylation, SSF54523 Pili subunits	P2	551	28.92	P. thermoproprionicum	
Daud0962		PF07963 Prepilin-type cleavage/methylation, SSF54523 Pili subunits	P2	119	N/A	ORFan	
Daud0963		hypothetical protein	P2	133	37.8	P. thermoproprionicum	
Daud0965	hofQ / pilQ	COG4796 Type II secretory (HofQ), Tfp assembly protein (PilQ)	P3	416	53.04	P. thermoproprionicum	
Daud0966		hypothetical protein	P3	182	30.07	D. reducens	
Daud0967	pulO / pilD	COG1989 Type II secretory, prepilin signal peptidase (PulO), Tfp prepilin leader peptidase (PilD)	P4	255	57.02	D. reducens	
Daud0968	pulE / pilB	COG2804 Type II secretory, ATPase (PulE), Tfp pilus assembly, ATPase (PilB)	P4	562	52.6	M. thermoacetica	

Daud0969	pilT	COG2805 Tfp pilus assembly, pilus retraction ATPase	P4	366	68.32	C. hydrogenoformans	
Daud0970	pulF / pilC	COG1459 Type II secretory (PulF), type IV pilus biogenesis (PilC)	P4	418	44.86	M. thermoacetica	
Daud0971	fimT / gspH	COG4970 Tfp pilus assembly protein (FimT), General secretion pathway protein H (GspH)	P4	174	29.41	M. thermoacetica	
Daud0972		PF07963 Prepilin-type cleavage/methylation, SSF54523 Pili subunits	P4	148	38.52	M. thermoacetica	
Daud0973		PF07963 Prepilin-type cleavage/methylation, SSF54523 Pili subunits	P4	157	30.52	M. thermoacetica	
Daud0974		hypothetical protein	P4	410	25.29	D. reducens	
Daud0975	pilM	COG4972 Tfp pilus assembly protein, ATPase	P4	375	32.94	C. hydrogenoformans	
Daud0976	pilN?	PF05137 Fimbrial assembly, COG3166 Tfp pilus assembly protein?	P4	222	26.34	D. reducens	
Daud0977	pilO?	COG3167 Tfp pilus assembly protein?	P4	303	26.94	D. reducens	
Daud0978		PF07833 Cu amine oxidase? Cation transport?	P4	142	30.77	D. reducens	
Daud0979		RR with CheY and HD-GYP domains (GAF,GGDEF?)	P4	375	44.69	D. reducens	ds transposon
Daud0991	pulE / pilB	COG2804 Type II secretory, ATPase (PulE), Tfp pilus assembly, ATPase (PilB)	P5	565	53.57	C. hydrogenoformans	hh on aro operon
Daud0992	pulF / pilC	COG1459 Type II secretory (PulF), type IV pilus biogenesis (PilC)	P5	434	44.47	M. thermoacetica	hh on aro operon
Daud0993	pulG / gspG	COG2165 Type II secretory, pseudopilin (PulG), General secretion pathway protein G (GspG)	P5	185	31.54	C. hydrogenoformans	hh on aro operon
Daud0994	fimT / gspH	COG4970 Tfp pilus assembly protein (FimT), General secretion pathway protein H (GspH)	P5	160	34.64	P. thermoproprionicum	hh on aro operon
Daud0995		TF02532 Prepilin-type cleavage/methylation, SSF54523 Pili subunits	P5	124	27.73	C. hydrogenoformans	hh on aro operon
Daud0996		SSF54523 Pili subunits	P5	198	N/A	ORFan	hh on aro operon
Daud0997		hypothetical protein	P5	400	20.45	D. reducens	hh on aro operon
Daud0998	pilM	COG4972 Tfp pilus assembly protein, ATPase	P5	362	31.43	D. reducens	hh on aro operon
Daud0999	pilN?	PF05137 Fimbrial assembly, COG3166 Tfp pilus assembly protein?	P5	197	26.26	P. thermoproprionicum	hh on aro operon
Daud1000	pilO	COG3167 Tfp pilus assembly protein PilO	P5	234	30.33	C. hydrogenoformans	hh on aro operon

Table S18. Flagellar genes.

Genes for chemotactic motility were identified by membership in known sequence families (e.g. COG, TIGRFAM, and Pfam) or by gene context (proximity and/or presence in operons with other identified flagellar genes). Annotation was by protein family, or if no confident protein family could be assigned, by the protein family assignment of the nearest homolog (such annotations are indicated with square brackets, with the source organism provided in the notes). Chemotactic signal transduction genes are only listed when present within a flagellar operon (see Table S19 for the full list signal transduction genes).

Gene	Name	Description	Operon	Len	CH id	CH species	Notes
Daud1734*	fliN	flagellar motor switch protein (short)	F1	183*	78	P. thermopropionicum	ds transposon A; pseudogene?
Daud1736	fliY / cheC	COG1776 Chemotaxis, inhibitor of MCP methylation	F2	370	45.55	P. thermopropionicum	us transposon A
Daud1737	fliM	COG1868 Flagellar motor switch	F2	333	58.66	P. thermopropionicum	
Daud1738	cheY	COG784 FOG: CheY-like receiver	F2	123	80.99	D. reducens	
Daud1739	cheC	COG1776 Chemotaxis, inhibitor of MCP methylation	F2	206	51.28	P. thermopropionicum	
Daud1740	cheR	COG1352 Methylase of chemotaxis methyl-accepting proteins	F2	270	58.59	P. thermopropionicum	
Daud1741	cheD	COG1871 Chemotaxis, stimulates methylation of MCP proteins	F2	172	58.02	P. thermopropionicum	
Daud1742	flgG	COG4786 Flagellar basal body rod protein	F2	247	39.76	M. thermoacetica	
Daud1743		hypothetical protein	F2	146	N/A	ORFan	
Daud1744	fliA	COG1191 RNA polymerase sigma 28 (flagellar biosynthesis)	F2	258	53.04	P. thermopropionicum	
Daud1745	ypfA / jofA	COG5581 Predicted glycosyltransferase	F2	216	29.63	P. thermopropionicum	
Daud1746	fleN	COG455 ATPases chrom. partitioning, flag. syn. reg. (FleN)	F2	291	52.22	P. thermopropionicum	
Daud1747	flhF	COG1419 Flagellar GTP-binding protein	F2	414	40.97	P. thermopropionicum	
Daud1748	flhA	COG1298 Flagellar biosyn.	F2	693	68.37	P. thermopropionicum	
Daud1749	flhB	COG1377 Flagellar biosyn.	F2	360	45.48	D. reducens	
Daud1750	fliR	COG1684 Flagellar biosyn.	F2	254	41.06	D. reducens	
Daud1751	fliQ	COG1987 Flagellar biosyn.	F2	88	52.87	P. carbinolicus	ds transposon B

Daud1753	fliP	COG1338 Flagellar biosyn.	F3	261	61.7	D. reducens	us transposon B
Daud1754	fliO	COG3190 Flagellar biogenesis	F3	131	52.44	P. thermopropionicum	
Daud1755	fliN	COG1886 Flagellar motor switch/type III secretory	F3	128	49.58	P. thermopropionicum	
Daud1756	fliL	COG1580 Flagellar basal body-associated protein (PF03748)	F3	162	35.44	P. thermopropionicum	
Daud1757	flgG / flgE	COG4786 Flagellar basal body rod (FliG), Flag. hook (FlgE)	F3	311	42.17	C. hydrogenoformans	
Daud1758		TF02530 Putative flagellar hook associated protein	F3	124	64.95	P. thermopropionicum	
Daud1759	flgD	PF03963 Flagellar hook capping, basal-body rod modification	F3	135	38.64	P. thermopropionicum	
Daud1760	fliK	PF02120 Flagellar hook-length control (COG3144)	F3	518	28.57	A. vinelandii	
Daud1761	fliJ	TF02473 Flagellar export (COG2882 Flag. biosyn. chaperone)	F3	154	30.94	P. thermopropionicum	
Daud1762	fliI / yscN	COG1157 Flagellar biosyn./type III secretory ATPase	F3	455	67.28	P. thermopropionicum	
Daud1763	fliH	COG1317 Flagellar biosyn./type III secretory	F3	232	36.36	M. thermoacetica	
Daud1764	fliG	COG1536 Flagellar motor switch	F3	337	61.7	P. thermopropionicum	
Daud1765	fliF	COG1766 Flagellar biosyn. lipoprot., TF00206 Flag. M-ring	F3	519	47.67	P. thermopropionicum	
Daud1766	fliE	COG1677 Flagellar hook-basal body complex	F3	103	61.97	D. reducens	
Daud1767	flgC	COG1558 Flagellar basal body rod	F3	140	54.68	T. maritima	
Daud1768	flgB	COG1815 Flagellar basal body rod	F3	133	46.46	P. thermopropionicum	ds sporulation
Daud1777	flmD / spsG	COG3980 Spore coat polysaccharide biosyn., putative flagellin modification (FlmD)	F4	354	50.45	N. punctiforme	us sporulation
Daud1778	flmA / wcaG	COG451 Nucleoside-diphosphate-sugar epimerases, putative flagellar glycosylation (FlmA)	F4	351	59.82	M. magneticum	hh wcaG, ubiG
Daud1779		hypothetical protein	F5	183	N/A	ORFan	
Daud1780	maf3	COG2604 Unchar., put. motility accessory factor (Maf3)	F5	627	32.81	M. thermoacetica	
Daud1781	fliC / flaB / hag	COG1344 Flagellin and related hook-associated proteins	F6	421	44.96	D. hafniense DCB-2	
Daud1782	flaG	COG1334 Uncharacterized flagellar protein	F6	124	51.39	T. tengcongensis	
Daud1783	fliD	COG1345 Flagellar hook-associated capping protein	F6	607	32.23	D. reducens	
Daud1784	fliS	COG1516 Flagellin-specific chaperone	F6	135	45.83	B. subtilis	
Daud1785		hypothetical protein	F6	162	27.52	P. thermopropionicum	

Daud1786		hypothetical protein	F7	117	N/A	ORFan	
Daud1787	flgL	COG1344 Flagellar biosynthesis 3; hook-filament junction	F7	311	30.16	M. thermoacetica	
Daud1788	flgK	COG1256 Flagellar hook-associated protein 1	F7	533	32.77	D. reducens	
Daud1789	flgN	COG3418 Flagellar biosyn./type III secretory chaperone	F7	162	N/A	ORFan	
Daud1790	flgM	PF04316 negative reg. of flagellin syn., anti-sigma 28	F7	99	39.13	P. thermopropionicum	
Daud1791	cheY + cheB	COG2201 RR: CheY-like receiver + CH3-esterase	F7	351	50.69	D. reducens	
Daud1792	cheA	COG643 Chemotaxis protein histidine kinase	F7	702	57.18	P. thermopropionicum	
Daud1793	cheW	COG835 Positive regulator of CheA protein activity	F7	168	67.32	P. thermopropionicum	
Daud1794	motA	COG1291 Proton conductor component of motor	F8	273	46.9	P. thermopropionicum	
Daud1795	motB	COG1360 Enables flag. motor rotation, links torque mach. to cell wall	F8	258	54.55	P. thermopropionicum	
Daud1835		hypothetical protein	F9	314	N/A	ORFan	
Daud1836	fhlB?	COG2257 Homolog of the cytoplasmic domain of FhlB	F9	118	51.65	P. thermopropionicum	hh rnhA

Table S19. Signal transduction genes.

Genes for signal transduction were identified by membership in known sequence families (e.g. COG, TIGRFAM, and Pfam). Annotation was by protein family, or if no confident protein family could be assigned, by the protein family assignment of the nearest homolog (such annotations are indicated with square brackets, with the source organism provided in the notes). Non-signaling genes found in operons with signaling genes are sometimes included in the table as they suggest possible roles for the signaling proteins. The operon name is also used to indicate such relationships. Examples of such context-derived candidate roles include **phosphate:** SIG11 operon genes; **sporulation:** SIG5A, SIG5B, SIG14, and SIG25 operon genes**; carbon assimilation:** Daud0119; **aromatic amino acids:** SIG10 operon genes. Putative pseudogenes are indicated with "*" and have lengths in nucleotides instead of amino acids.

Abbreviations used in "Notes" column include "RR": response regulator, "TF": transcription factor, "WH": winged helix transcription factor domain, "UNK": unknown domain, "Y": cheY-like receiver domain, "cNMP": cyclic nucleotide monophosphate binding domain, "GGDEF": GGDEF motif containing domain (likely diguanylate cyclase activity), "PAS": PAS domain (ligand and cofactor

binding), "GAF": GAF domain (cyclic GMP-specific phosphodiesterase), "HD" and "HD-GYP": metal dependent phosphohydrolase domain, "B": cheB-like methylesterase domain, "HK": histidine kinase domain, "ANTAR": AmiR and NasR transcription antitermination regulators (RNA-binding domain), "LytTr": LytTr-type winged helix DNA binding domain, "SENS": ligand sensing domain, "HAMP": HAMP linker region, "NUC": nucleotide binding domain, "ATP": ATP binding domain, "MEMB": membrane associated domain, "EAL": EAL motif containing domain (likely diguanylate phosphodiesterase activity), "CBS": cystathionine-beta synthase domain, "IMPDH": inosine-5'-monophosphate dehydrogenase domain, "EPP": exopolyphosphatase domain, and "PAP": polyA polymerase domain.

Gene	Name	Description	Operon	Len	CH id	CH species	Notes
Daud0119	lytT	COG3279: RR: LytR/AlgR family	CODH1	250	49.4	C. hydrogenoformans	RR: Y+LytTr
Daud0138		COG4936: Predicted sensor domain, TIGR00229: PAS, TIGR00254: GGDEF, PF01966: Metal-dependent phosphohydrolase, HD region, subdomain		1339	48.42	P. thermopropionicum	SENS+PAS+PAS+ PAS+PAS+PAS+PAS+ GGDEF+HD
Daud0264		COG517: FOG: CBS domain		154	41.61	N. punctiforme	
Daud0271	uspA	COG589: Universal stress protein UspA and related nucleotide-binding proteins	TPT27	251	34.97	M. jannaschii	
Daud0272	dhlC	COG4147, TIGR00813: Na+/solute symporter	TPT27	537	36.1	G. kaustophilus	Na+/solute symport
Daud0377	pleD	COG3706: RR: CheY-like + GGDEF domains		315	35.88	D. reducens	RR: Y+GGDEF
Daud0417		COG3322: Predicted periplasmic ligand-binding sensor domain, PF00672: Histidine kinase, HAMP region, TIGR00229: PAS, TIGR00254: GGDEF, PF01966 Metal-dependent phosphohydrolase, HD region		854	63.98	P. thermopropionicum	SENS+HAMP+PAS+ GGDEF+HD
Daud0446	agrB	COG4512: Membrane protein putatively involved in post-translational modification of the autoinducing quorum-sensing peptide	SIG1	198	29.03	M. thermoacetica	us Daud0445 COG4632 Exopolysaccharide biosynthesis protein
Daud0447		[COG0642: Signal transduction histidine kinase]	SIG1	155	24.11	D. hafniense Y51	S. wolfei annot.
Daud0448	ntrB	COG3852: Signal transduction histidine kinase, nitrogen specific	SIG1	520	36.03	D. reducens	SENS domain? Nitrogen specific?
Daud0468	lytS	COG3275: Putative regulator of cell autolysis	SIG2	451	59.95	P. thermopropionicum	
Daud0469	lytT	COG3279: RR: LytR/AlgR family	SIG2	255	65.28	P. thermopropionicum	RR: Y+LytTr
Daud0486	yqzB	PF08279 Helix-turn-helix, type 11, COG517: FOG:	SIG3	214	70.62	P. thermopropionicum	us Daud0485 glyS

		CBS domain					
Daud0487	yqfL	COG1806: Unc. protein conserved in bacteria	SIG3	270	70.79	D. reducens	
Daud0530	ntrB	COG3852: Signal transduction histidine kinase, nitrogen specific, TIGR00229: PAS, PF00512: Histidine kinase A, N-terminal, PF02518: ATP-binding region, ATPase-like		580	47.7	D. reducens	PAS+PAS+HK+NUC
Daud0538		TIGR00229: PAS, TIGR00254: GGDEF, COG2203: FOG: GAF, PF01966: Metal-dependent phosphohydrolase, HD region, subdomain		861	60.1	M. thermoacetica	PAS+GGDEF+GAF+ GAF+HD ; ds Daud0539 COG0286 Type I restriction-modification system methyltransferase subunit
Daud0585		PF01966: Metal-dependent phosphohydrolase, HD region, subdomain		421	45.27	D. reducens	HD+HD
Daud0675	ntrB	COG3852: Signal transduction histidine kinase, nitrogen specific, SSF55781: GAF domain-like, TIGR00229: PAS, PF00512: Histidine kinase A, N-terminal, PF02518: ATP-binding region, ATPase-like	SIG4 / NIF?	856	36.49	G. sulfurreducens	SENS?+GAF+PAS+ HK+ATP; nitrogen specific?
Daud0676	vicK	IPR001789 + COG3604 + TIGR00229 + COG5000	SIG4 / NIF?	671	26.43	E. coli K12	RR: Y+GAF+PAS+HK
Daud0677		COG3437: RR: CheY-like receiver + HD-GYP domains	SIG4 / NIF?	344	41.49	D. ethenogenes	RR: Y+HD
Daud0678	nfnB	COG778: Nitroreductase	SIG4 / NIF?	243	42.69	P. carbinolicus	
Daud0679	pocR	COG4936: Predicted sensor domain, COG2203: FOG: GAF domain, PF01966 Metal-dependent phosphohydrolase, HD region, subdomain	SIG4 / NIF?	572	47.3	M. thermoacetica	SENS+GAF+HD
Daud0680	tar	COG840: Methyl-accepting chemotaxis protein, PF00672: Histidine kinase, HAMP region, PF00015: Bacterial chemotaxis sensory transducer		680	31.75	D. hafniense DCB-2	SENS?+HAMP+UNK ; facing SIG20
Daud0688		hypothetical protein	SIG5A/ SPO1A	123	39.78	C. hydrogenoformans	hh?
Daud0689	kaiC	COG467: RecA-superfamily ATPases implicated in signal transduction, PF06745: Circadian clock protein KaiC	SIG5A/ SPO1A	468	56.85	C. hydrogenoformans	ATPase+ATPase

Daud0690		SSF55781: GAF domain-like, COG2199: FOG: GGDEF domain	SIG5A/ SPO1A	472	31.41	G. metallireducens	GAF+GGDEF
Daud0691		COG2203: FOG: GAF domain, PF01966: Metal-dependent phosphohydrolase, HD region, subdomain	SIG5A/ SPO1A	393	54.38	M. thermoacetica	GAF+HD
Daud0692	rsbR / spoIIAA	COG1366: Anti-anti-sigma regulatory factor (antagonist of anti-sigma factor)	SIG5A/ SPO1A	253	53.36	T. tengcongensis	
Daud0693	rsbS / spoIIAA	COG1366: Anti-anti-sigma regulatory factor (antagonist of anti-sigma factor)	SIG5A/ SPO1A	121	62.28	T. tengcongensis	
Daud0694*	rsbT	[Anti-sigma regulatory factor (Ser/Thr protein kinase)]	SIG5A/ SPO1A	192*	60.71	T. tengcongensis	T. tengcongensis annot.
Daud0696	rsbX	PF07228: Sporulation stage II, protein E C-terminal	SIG5B / SPO1B	209	33.5	T. tengcongensis	SIG5B broken from SIG5A by transposon Daud0696
Daud0697		COG2199: FOG: GGDEF domain	SIG5B / SPO1B	265	51.13	T. tengcongensis	
Daud0698		COG5001: Predicted signal transduction protein containing a membrane domain, an EAL and a GGDEF domain	SIG5B / SPO1B	575	36.49	M. magneticum	MEMB+GGDEF+EAL
Daud0745	baeS	COG642: Signal transduction histidine kinase	SIG6	461	31.05	D. hafniense DCB-2	SENS?+UNK+ UNK+ATP
Daud0746	lytT	COG3279: RR: LytR/AlgR family	SIG6	261	37.04	D. hafniense DCB-2	RR: Y+LytTr
Daud0771		TIGR00229: PAS, TIGR00254: GGDEF, COG2203: FOG: GAF domain, PF01966: Metal-dependent phosphohydrolase, HD region, subdomain		881	60.16	M. thermoacetica	PAS+GGDEF+GAF+ GAF+HD
Daud0783	crp-like	COG664: cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases, Crp/Fnr family		218	48.36	D. reducens	cNMP+WH; paralog of Daud1482 ; ds Daud0784 transposase
Daud0821		hypothetical protein	SIG7	124	65.77	D. reducens	
Daud0822		COG2524: Predicted transcriptional regulator, contains C-terminal CBS domains, COG2199: FOG: GGDEF domain	SIG7	288	54.64	D. reducens	CBS+GGDEF
Daud0823		TIGR00229: PAS, TIGR00254: GGDEF, PF01966: Metal-dependent phosphohydrolase, HD region, subdomain	SIG8	756	45.94	D. reducens	SENS?+PAS+PAS+ GGDEF+HD

Daud0824		COG3287, PF08495: Domain of unknown function DUF1745	SIG8	380	53.99	P. thermopropionicum	
Daud0855	tar	COG840: Methyl-accepting chemotaxis protein, PF00672: Histidine kinase, HAMP region, PF00015: Bacterial chemotaxis sensory transducer		551	21.49	D. reducens	SENS?+HAMP+UNK
Daud0979		COG2203: FOG: GAF domain, PF01966: Metal-dependent phosphohydrolase, HD region, subdomain		375	44.69	D. reducens	GAF+HD
Daud1069		hypothetical protein [response regulator receiver protein]	SIG9	88	56.72	M. thermoacetica	S. wolfei annot.
Daud1070	guaB	COG0516: IMP dehydrogenase/GMP reductase, COG0517: FOG: CBS domain	SIG9	486	75.72	P. thermopropionicum	IMPDH+CBS+IMPDH
Daud1077	tar	COG840: Methyl-accepting chemotaxis protein, PF00672: Histidine kinase, HAMP region, PF00015: Bacterial chemotaxis sensory transducer	SR6	370	37.38	D. reducens	HAMP+UNK
Daud1192		COG2206: HD-GYP domain	SIG10 / ARO4	365	46.28	D. reducens	paralog Daud1193
Daud1193		COG2206: HD-GYP domain	SIG10 / ARO4	374	41.14	D. reducens	paralog Daud1192 ; us Daud1194 aroH
Daud1203		COG618: Exopolyphosphatase-related proteins, COG2524: Predicted transcriptional regulator, contains C-terminal CBS domains, COG0617: tRNA nucleotidyltransferase/poly(A) polymerase		874	51.1	D. reducens	EPP+CBS+PAP
Daud1257	citB	COG4565: Response regulator of citrate/malate metabolism		225	41.74	P. thermopropionicum	RR: Y+UNK
Daud1415	phoU	COG704: Phosphate uptake regulator	SIG11 / PHO1	219	57.8	D. reducens	
Daud1416	pstB	COG1117: ABC-type phosphate transport system, ATPase component	SIG11 / PHO1	252	78.71	P. thermopropionicum	PO4 transport
Daud1418	phoR	COG5002: Signal transduction histidine kinase	SIG11 / PHO1	445	48.33	P. thermopropionicum	
Daud1419	phoB	COG745: RR CheY-like + winged-helix DNA-binding domains	SIG11 / PHO1	232	61.21	D. reducens	RR: Y+WH; operon also contains Daud1420, Daud1421
Daud1482	crp-like	COG664: cAMP-binding proteins - catabolite gene		236	57.75	D. reducens	cNMP+WH; paralog of

		activator and regulatory subunit of cAMP-dependent protein kinases, Crp/Fnr family					Daud0783
Daud1560	cheW	COG835: Chemotaxis signal transduction protein	SIG12	147	55.64	D. reducens	
Daud1561	tar	COG840: Methyl-accepting chemotaxis protein	SIG12	529	35.99	P. thermopropionicum	
Daud1567	dnaQ	COG847: DNA polymerase III, epsilon subunit and related 3'-5' exonucleases	SIG13	237	38.54	C. hydrogenoformans	
Daud1568		COG2905: Predicted signal-transduction protein containing cAMP-binding and CBS domains	SIG13	635	37.38	C. hydrogenoformans	
Daud1613		hypothetical protein	SIG14 / SPO2	289	43.94	D. reducens	ds Daud1612 COG463 Glycosyltransferases involved in cell wall biogenesis
Daud1614		COG4825: Unc. membrane-anchored protein, PF04263 Thiamin pyrophosphokinase, catalytic region	SIG14 / SPO2	376	56.99	D. reducens	
Daud1615	spo0A	COG2197: RR: CheY-like receiver + HTH DNA-binding domains	SIG14 / SPO2	257	81.64	P. thermopropionicum	RR: Y+WH
Daud1616	spoIVB	TIGR02860: Peptidase S55, sporulation stage IV, protein B	SIG14 / SPO2	441	47.73	P. thermopropionicum	
Daud1723	pleD	COG3706: RR: CheY-like + GGDEF domains	SIG15	291	25.91	N. punctiforme	RR: Y+GGDEF
Daud1724		hypothetical protein, SSF46785: ""Winged helix"" DNA-binding domain	SIG15	233	39.91	M. thermoacetica	
Daud1728		TIGR00229: PAS, COG2203: FOG: GAF, PF01966: Metal-dependent phosphohydrolase, HD region, subdomain	SIG16	1172	39.54	T. thermophilus HB27	PAS+GAF+PAS+ PAS+GAF+GAF+HD
Daud1729*		COG3706: RR: CheY-like receiver + GGDEF domain (pseudogene?)	SIG16	392*	52.08	A. vinelandii	RR: Y
Daud1736	fliY / cheC	COG1776: Chemotaxis protein CheC, inhibitor of MCP methylation	SIG17 / F2	370	45.55	P. thermopropionicum	
Daud1737	fliM	COG1868: Flagellar motor switch protein	SIG17 / F2	333	58.66	P. thermopropionicum	
Daud1738	cheY	COG784: RR: CheY-like receiver	SIG17 / F2	123	80.99	D. reducens	RR: Y

Daud1739	cheC	COG1776: Chemotaxis protein CheC, inhibitor of MCP methylation	SIG17 / F2	206	51.28	P. thermopropionicum	
Daud1740	cheR	COG1352: Methylase of chemotaxis methyl-accepting proteins	SIG17 / F2	270	58.59	P. thermopropionicum	
Daud1741	cheD	COG1871: Chemotaxis protein; stimulates methylation of MCP proteins	SIG17 / F2	172	58.02	P. thermopropionicum	us Daud1742 flgG
Daud1764	fliG	COG1536: Flagellar motor switch protein		337	61.7	P. thermopropionicum	
Daud1791	cheY + cheB	COG2201: RR: CheY-like receiver + methylesterase domain	SIG18 / F7	351	50.69	D. reducens	RR: Y+B
Daud1792	cheA	COG643: Chemotaxis protein histidine kinase and related kinases	SIG18 / F7	702	57.18	P. thermopropionicum	
Daud1793	cheW	COG835: Chemotaxis signal transduction protein	SIG18 / F7	168	67.32	P. thermopropionicum	
Daud1843		COG517: FOG: CBS domain, PF00571: Cystathionine-beta-synthase	SIG19	182	51.94	P. carbinolicus	
Daud1844		COG2206: HD-GYP domain, TIGR00277: uncharacterized domain HDIG	SIG19	214	36.52	P. thermopropionicum	
Daud1845		hypothetical protein	SIG20	225	N/A	ORFan	paralog Daud1834
Daud1846	pleD	COG3706: RR: CheY-like + GGDEF domains	SIG20	315	43.81	G. sulfurreducens	RR: Y+GGDEF
Daud1896	cpxP	[P pilus assembly/Cpx signaling pathway, periplasmic inhibitor/zinc-resistance associated protein]	SIG21	166	28.57	N. oceani	N. oceani annot.
Daud1897	ompR	COG745: RR: CheY-like receiver + winged-helix DNA-binding domains	SIG21	228	67.7	P. thermopropionicum	RR: Y+WH
Daud1898	vicK	COG5002: Signal transduction histidine kinase	SIG21	572	47.09	P. thermopropionicum	
Daud1908	baeS	COG642: Signal transduction histidine kinase + COG4753: RR: CheY-like receiver domain		773	35.19	S. ruber	RR: GAF?+HK+Y
Daud1949		COG2206: HD-GYP domain	SIG22	227	60.23	M. thermoacetica	
Daud1950	cheY?	COG4378: Unc. protein, [Predicted diverged CheY-domain]	SIG22	109	44.32	C. acetobutylicum	RR: Y?; C. acetobutylicum annot.
Daud2002		COG2905: Predicted signal-transduction protein containing cAMP-binding and CBS domains	SIG23	642	35.14	C. hydrogenoformans	
Daud2003	dnaQ	COG847: DNA polymerase III, epsilon subunit and	SIG23	248	40.61	C. hydrogenoformans	

		related 3'-5' exonucleases					
Daud2028	amiR	COG3707: RR with putative antiterminator output domain	SIG24 / NIF4	193	50.8	D. reducens	RR: Y+ANTAR
Daud2031	amiR	COG3707: RR with putative antiterminator output domain	SIG24 / NIF4	199	46.49	P. thermopropionicum	RR: Y+ANTAR
Daud2083		COG2203: FOG: GAF domain, COG2199: FOG: GGDEF domain		569	37.16	D. reducens	GAF+GGDEF
Daud2176		hypothetical protein	SIG25 / SPO3	221	41.95	P. thermopropionicum	
Daud2177	spo0F	COG784: FOG: CheY-like receiver	SIG25 / SPO3	121	53.51	P. thermopropionicum	RR: Y
Daud2178		hypothetical protein	SIG25 / SPO3	219	42.92	P. thermopropionicum	
Daud2207		hypothetical protein	SIG26	50	64.44	R. albus	TF?
Daud2208		COG517: FOG: CBS domain	SIG26	222	53.81	M. thermoacetica	us tRNA-Gly

Table S20. Transport genes.

Genes encoding transport proteins were identified by membership in known sequence families (e.g. COG, TIGRFAM, and Pfam). Annotation was by protein family, or if no confident protein family could be assigned, by the protein family assignment of the nearest homolog (such annotations are indicated with square brackets, with the source organism provided in the notes). Associated non-transport genes found in operons with transport genes are included in the table as they suggest possible roles for the transporters. Putative pseudogenes are indicated with "*" and have lengths in nucleotides instead of amino acids.

Wanger, *et al.* detected 50 μm spatial variations in adsorbed Fe, S and exopolysaccharide-type organic species (consistent with the polysaccharide ABC exporter and exopolysaccharide synthesis genes found in the genome of *D. audaxviator*) on a surface from this fracture zone(*68*). These variations in adsorbed species also produce gradients in surface charges that in turn may lower the pH close to the mineral surfaces and perhaps alleviate the impact of the high pH in the fracture fluid on the ability of *D. audaxviator* to maintain a H^+ gradient across the cell membrane. Whether due to an advantageous pH or because of increased access to nutrients, *D. audaxviator* does appear to colonize nutrient-rich mineral surfaces (*68*).

Gene	Name	Description	Operon	Len	CH id	CH species	Notes
Daud0057		COG428: Predicted divalent heavy-metal cations transporter		236	48.28	P. thermopropionicum	Zn/Fe permease?
Daud0128	yfkE	COG530: Ca2+/Na+ antiporter		331	41.34	W. succinogenes	Ca+/Na+ antiport; K+ dependent?
Daud0141	glnB/K	COG347: Nitrogen regulatory protein PII	TPT1	112	64.22	D. hafniense Y51	
Daud0142	amtB	COG4: Ammonia permease, TIGR00836 Ammonium transporter	TPT1	465	58.64	D. hafniense Y51	ammonium
Daud0175	mgtA	COG474: Cation transport ATPase		909	61.22	D-monas spp.	divalent cations ATPase
Daud0209	secE	COG690: Preprotein translocase subunit SecE		114	41.23	P. thermopropionicum	
Daud0245	secY	COG201: Preprotein translocase subunit SecY		425	75.84	D. reducens	
Daud0265	trkG	COG168: Trk-type K+ transport systems, membrane components	TPT2	467	56.16	D. reducens	K+
Daud0266	trkA	COG569: K+ transport systems, NAD-binding component	TPT2	446	60.14	D. reducens	K+
Daud0271	uspA	COG589: Universal stress protein UspA and related nucleotide-binding proteins	TPT3	251	34.97	M. jannaschii	signaling
Daud0272	dhlC	COG4147, TIGR00813 Na+/solute symporter	TPT3	537	36.1	G. kaustophilus	Na+/solute symport
Daud0273	cbiM	COG310: ABC-type Co2+ transport system, permease component	TPT4	232	43.15	T. denticola	Co2+ permease; TPT27?
Daud0274	cbiQ	COG619: ABC-type cobalt transport system, permease component CbiQ and related transporters	TPT4	266	31.98	P. thermopropionicum	Co2+ permease; TPT27?
Daud0275	cbiO	COG1122: ABC-type cobalt transport system, ATPase component	TPT4	271	53.28	C. hydrogenoformans	Co2+ ATPase; TPT27?
Daud0292	rbsB	COG1879: ABC-type sugar transport system, periplasmic component		426	44.55	P. thermopropionicum	sugar periplasmic
Daud0307	secG	TIGR00810 Preprotein translocase SecG subunit	TPT5 / GLY1	77	67.11	D. reducens	hh on GLY1?
Daud0308	hppA	COG3808, TIGR01104: V-type H(+)-translocating pyrophosphatase	TPT5 / GLY1	683	61.01	M. acetivorans	proton pump; probable HGT; high SNP count; hh on GLY1?

Daud0309		COG730: Predicted permeases		320	57.01	D. vulgaris DP4	
Daud0310	modA	COG725: ABC-type molybdate transport system, periplasmic component	TPT6	276	52.43	D. reducens	SO4/MoO4 periplasmic
Daud0311		COG1695: Predicted transcriptional regulators	TPT6	139	52.88	D. reducens	TF
Daud0312	arsA	TIGR00345 Anion-transporting ATPase		398	48.47	B. cereus	anion (arsenite?) efflux
Daud0336	lysE	COG1280: Putative threonine efflux, PF01810 Lysine exporter protein (LYSE/YGGA)		226	61.74	P. thermopropionicum	AA export
Daud0349		COG4392, PF05437: Branched-chain amino acid transport	TPT7	107	52.48	D. reducens	AA
Daud0350	azlC	COG1296: Predicted branched-chain amino acid permease	TPT7	239	42.59	D. reducens	AA
Daud0373	lldP	COG1620: L-lactate permease		505	43.38	R. palustris	L-lactate
Daud0374*	rarD	COG2962: Predicted permeases	TPT8	872*	55.38	B. halodurans	Pseudogene?
Daud0386	nikB / dppB	COG601: ABC-type dipeptide/oligopeptide/nickel transport systems, permease components	TPT9	313	50.83	B. licheniformis	Ni or peptide permease
Daud0387	nikC / dppC	COG1173: ABC-type dipeptide/oligopeptide/nickel transport systems, permease components	TPT9	310	59.55	M. acetivorans	Ni or peptide permease
Daud0388	nikD / dppD	COG444: ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component	TPT9	325	55.41	S. thermophilum	Ni or peptide ATPase
Daud0389	appF / dppF	COG4608: ABC-type oligopeptide transport system, ATPase component	TPT9	322	57.23	B. japonicum	Ni or peptide ATPase
Daud0390	lmbE	COG2120: Uncharacterized proteins, LmbE homologs	TPT9	250	50.46	D. geothermalis	synteny with C. tepidum
Daud0391	dpdA / ddpA	COG747: ABC-type dipeptide transport system, periplasmic component	TPT9	523	31.7	M. acetivorans	Ni or peptide periplasmic; synteny with C. tepidum
Daud0410	modA	COG725: ABC-type molybdate transport system, periplasmic component	TPT10	297	65.13	P. thermopropionicum	SO4/MoO4 periplasmic; us LeuA Daud0409
Daud0411	modC	COG4149: ABC-type molybdate transport system, permease component	TPT10	228	81.94	P. thermopropionicum	SO4/MoO4 permease
Daud0412	cysA	COG1118: ABC-type sulfate/molybdate transport	TPT10	220	63.08	P. thermopropionicum	SO4/MoO4 ATPase

		systems, ATPase component					
Daud0439	rhaT	COG697: Permeases of the drug/metabolite transporter (DMT) superfamily		326	38.35	R. eutropha	
Daud0440		COG730: Predicted permeases	TPT11	257	45.27	P. carbinolicus	
Daud0441	gntR	COG1802: Transcriptional regulators	TPT11	234	59.81	D. reducens	TF
Daud0476	araJ	COG2814: Arabinose efflux permease		475	55.61	M. thermoacetica	sugar/drug export
Daud0481	mgtE	COG2239: Mg/Co/Ni transporter MgtE (contains CBS domain)		451	48.95	S. thermophilum	divalent cations
Daud0532*		[COG1682 ABC-type polysaccharide/polyol phosphate export systems, permease]		225*	56.76	P. thermopropionicum	truncated pseudogene? P. thermopropionicum annot.
Daud0533		COG4619: ABC-type uncharacterized transport system, ATPase component	TPT12	227	47.96	D. reducens	
Daud0534		COG390: ABC-type uncharacterized transport system, permease component	TPT12	268	70.94	D. reducens	
Daud0545	trkG	COG168: Trk-type K+ transport systems, membrane components		459	52.75	C. hydrogenoformans	K+
Daud0612	chrA	COG2059: Chromate transport protein ChrA	TPT13	410	64.19	O. iheyensis	CrO4
Daud0613		COG1695: Predicted transcriptional regulators	TPT13	102	60	B. thuringiensis	TF
Daud0614	perM	COG628: Predicted permease		359	40.35	D. reducens	
Daud0662	lepB	COG681: Signal peptidase I		175	60.71	D. reducens	protein sec
Daud0668	livK	COG683: ABC-type branched-chain amino acid transport systems, periplasmic component	TPT14	393	48.17	G. kaustophilus	AA periplasmic
Daud0669	livH	COG559: Branched-chain amino acid ABC-type transport system, permease components	TPT14	295	50	G. kaustophilus	AA permease
Daud0670	livM	COG4177: ABC-type branched-chain amino acid transport system, permease component	TPT14	336	51.31	G. kaustophilus	AA permease
Daud0671	livG	COG411: ABC-type branched-chain amino acid transport systems, ATPase component	TPT14	253	56.63	T. thermophilus HB27	AA ATPase
Daud0672	livF	COG410: ABC-type branched-chain amino acid transport systems, ATPase component	TPT14	240	55.08	G. kaustophilus	AA ATPase
Daud0722		hypothetical protein	TPT15	182	N/A	ORFan	us transposase; replaces MTH672 of M.

							thermoautotrophicus (COG4827 Predicted transporter) which was in TPT15 operon
Daud0723	tolQ	COG811: Biopolymer transport proteins [PF01618 MotA/TolQ/ExbB proton channel]	TPT15	227	46.43	M. thermautotrophicus	H+ or Na+ channel?; M. thermoautotrophicus annot.
Daud0724		COG4744: Uncharacterized conserved protein	TPT15	117	52.59	M. maripaludis	
Daud0787	citT	COG471: Di- and tricarboxylate transporters, PF00939: Na/SO4 symporter		492	36.29	B. xenovorans	C4 di- and tri-carboxylate permease or Na/SO4 symport?
Daud0807	chaC	COG3703: Uncharacterized protein involved in cation transport		158	44.67	B. pseudomallei	
Daud0831		COG432, TIGR00149: Protein of unknown function UPF0047	TPT16	138	80.15	P. thermopropionicum	
Daud0832		hypothetical protein	TPT16	134	40.15	T. thermophilus HB8	
Daud0833	lraI	COG803: ABC-type metal ion transport system, periplasmic component/surface adhesin	TPT16	346	44.23	M. thermoacetica	metai ion periplasmic
Daud0848		COG1811: Uncharacterized membrane protein, possible Na+ channel or pump		230	55.9	D. hafniense Y51	Na+ channel
Daud0860	yrvC	COG490: Putative regulatory, ligand-binding protein related to C-terminal domains of K+ channels	TPT17	164	44.03	A. aeolicus	K+ channel
Daud0861	kefB	COG475: Kef-type K+ transport systems, membrane components	TPT17	396	34.25	A. aeolicus	K+ channel
Daud0948	lspA	COG597: Lipoprotein signal peptidase	TPT18	154	44.03	M. thermoacetica	lipoprotein sec
Daud0949	rluA	COG564: Pseudouridylate synthases, 23S RNA-specific	TPT18	310	64.24	P. thermopropionicum	
Daud1092	cbiK	COG5266: ABC-type Co2+ transport system, periplasmic component		227	28.12	M. acetivorans	Co2+ periplasmic
Daud1098	echA / mnhA	COG1009: NADH:ubiquinone oxidoreductase subunit 5 (chain L)/Multisubunit Na+/H+ antiporter, MnhA subunit	TPT19 / FHL2	654	45.37	D. ethenogenes	Na+/H+ antiport?
Daud1099	echB / hyfC /	COG650: Formate hydrogenlyase subunit 4	TPT19 / FHL2	291	51.62	D. ethenogenes	Na+/H+ antiport? Ortholog of C. hydrogenoformans cooK

	cooK						(CODH-I)
Daud1100	echC / cooL	COG3260: Ni,Fe-hydrogenase III small subunit	TPT19 / FHL2	142	72.46	Dehalo. sp. CBDB1	Na+/H+ antiport? Ortholog of C. hydrogenoformans cooL (CODH-I)
Daud1101	echD	PIRSF036585 [NiFe]-hydrogenase-3-type complex Ech, subunit EchD	TPT19 / FHL2	121	36.54	D. ethenogenes	Na+/H+ antiport?
Daud1102	hycE / cooH	COG3261: Ni,Fe-hydrogenase III large subunit	TPT19 / FHL2	359	64.62	D. ethenogenes	Na+/H+ antiport? Ortholog of C. hydrogenoformans cooH (CODH-I)
Daud1103	echF / nuoI	COG1143: Formate hydrogenlyase subunit 6/NADH:ubiquinone oxidoreductase 23 kD subunit (chain I)	TPT19 / FHL2	127	41.44	T. tengcongensis	Na+/H+ antiport?
Daud1117	tatA	COG1826: Sec-independent protein secretion pathway components	TPT20	116	37.11	D. geothermalis	protein sec
Daud1118	tatA	COG1826: Sec-independent protein secretion pathway components	TPT20	85	53.52	P. thermopropionicum	protein sec
Daud1119	tatA	COG1826: Sec-independent protein secretion pathway components	TPT20	84	50	P. thermopropionicum	protein sec
Daud1120	tatC	COG805: Sec-independent protein secretion pathway component TatC	TPT20	257	54.84	P. thermopropionicum	protein sec
Daud1125	tagG	COG1682: ABC-type polysaccharide/polyol phosphate export systems, permease component	TPT21	261	48.88	P. thermopropionicum	export permease; paralog of Daud1206, Daud1211, Daud1212
Daud1126	ccmA	COG1131: ABC-type multidrug transport system, ATPase component	TPT21	327	57.23	M. thermoacetica	export ATPase
Daud1127	livF	COG410: ABC-type branched-chain amino acid transport systems, ATPase component	TPT22	236	67.95	D. reducens	AA ATPase
Daud1128	livG	COG411: ABC-type branched-chain amino acid transport systems, ATPase component	TPT22	257	64.71	D. reducens	AA ATPase
Daud1129	livM	COG4177: ABC-type branched-chain amino acid transport system, permease component	TPT22	300	65.77	D. reducens	AA permease
Daud1130	livH	COG559: Branched-chain amino acid ABC-type transport system, permease components	TPT22	294	69.69	D. reducens	AA permease
Daud1131	livK	COG683: ABC-type branched-chain amino acid	TPT22	390	65.06	D. reducens	AA periplasmic

		transport systems, periplasmic component					
Daud1132	rarD	COG2962: Predicted permeases		311	64.24	D. reducens	paralog of Daud1134
Daud1133		hypothetical protein		119	N/A	ORFan	
Daud1134	rhaT	COG697: Permeases of the drug/metabolite transporter (DMT) superfamily		320	53.66	M. thermoacetica	paralog of Daud1132
Daud1135	hisJ	COG834: ABC-type amino acid transport, periplasmic component	TPT23	269	51.48	M. thermoacetica	AA periplasmic
Daud1136	hisM	COG765: ABC-type amino acid transport system, permease component	TPT23	236	62.34	M. thermoacetica	AA permease
Daud1137	glnQ	COG1126: ABC-type polar amino acid transport system, ATPase component	TPT23	241	76.67	M. thermoacetica	AA ATPase
Daud1145	salX	COG1136: ABC-type antimicrobial peptide transport system, ATPase component	TPT24	237	50.65	A. fulgidus	export ATPase
Daud1146	lolE	COG4591: ABC-type transport system, involved in lipoprotein release, permease component	TPT24	788	30.26	C. tepidum	export permease
Daud1147	emrA	COG1566: Multidrug resistance efflux pump	TPT24	396	33.33	T. denitrificans	export
Daud1148	ubiE	COG2226: Methylase involved in ubiquinone/menaquinone biosynthesis	TPT25	216	44.56	M. thermoacetica	
Daud1149	znuB	COG1108: ABC-type Mn2+/Zn2+ transport systems, permease components	TPT25	274	65.25	M. thermoacetica	divalent cations permease
Daud1150	znuC	COG1121: ABC-type Mn2+/Zn2+ transport systems, ATPase component	TPT25	273	43.9	D. vulgaris HB	divalent cations ATPase
Daud1151	lraI	COG803: ABC-type metal ion transport system, periplasmic component/surface adhesin	TPT25	292	37.05	D. vulgaris DP4	divalent cations periplasmic
Daud1162	potE	COG531: Amino acid transporters		519	59.21	M. thermoacetica	AA permease
Daud1206	tagG-like	COG842: ABC-type multidrug transport system, permease component	TPT26	244	61.37	P. thermopropionicum	ABC-2 export permease; paralog of Daud1125, Daud1211, Daud1212
Daud1207	ccmA	COG1131: ABC-type multidrug transport system, ATPase component	TPT26	280	66.55	P. thermopropionicum	ABC-2 export ATPase; paralog of Daud1213
Daud1208	chlI	COG1239: Mg-chelatase subunit ChlI	TPT26	373	63.61	M. barkeri	Mg2+/Co2+ cheletase
Daud1209	chlI	COG1239: Mg-chelatase subunit ChlI	TPT26	671	51.27	M. barkeri	Mg2+/Co2+ cheletase

Daud1210	cobN	COG1429: Cobalamin biosynthesis protein CobN and related Mg-chelatases	TPT26	1211	46.17	M. thermautotrophicus	Mg2+/Co2+ cheletase?
Daud1211	tagG-like	COG842: ABC-type multidrug transport system, permease component	TPT26	253	33.62	P. thermopropionicum	ABC-2 export permease; paralog of Daud1125, Daud1206, Daud1212
Daud1212	tagG-like	COG842: ABC-type multidrug transport system, permease component	TPT26	260	35	P. thermopropionicum	ABC-2 export permease; paralog of Daud1125, Daud1206, Daud1211
Daud1213	ccmA	COG1131: ABC-type multidrug transport system, ATPase component	TPT26	310	61.68	P. thermopropionicum	ABC-2 export ATPase; paralog of Daud1207
Daud1214	fepC	COG1120: ABC-type cobalamin/Fe3+-siderophores transport systems, ATPase components	TPT26	275	49.61	M. barkeri	hydroxamate / siderophore ATPase
Daud1215		hypothetical protein	TPT26	785	30.15	M. thermoacetica	
Daud1216	fepB	COG614: ABC-type Fe3+-hydroxamate transport system, periplasmic component	TPT26	479	36.25	M. hungatei	hydroxamate / siderophore periplasmic
Daud1217	fepD	COG609: ABC-type Fe3+-siderophore transport system, permease component	TPT26	363	51.74	M. hungatei	hydroxamate / siderophore permease
Daud1218	fepB	COG614: ABC-type Fe3+-hydroxamate transport system, periplasmic component	TPT26	359	30.79	M. acetivorans	hydroxamate / siderophore periplasmic
Daud1219	cobN	COG1429: Cobalamin biosynthesis protein CobN and related Mg-chelatases	TPT26	1286	57.56	M. mazei	Mg2+/Co2+ cheletase
Daud1255		COG4666: TRAP-type uncharacterized transport system, fused permease components	TPT27	663	55.75	B. halodurans	C4 dicarboxylate permease; ds ttdA L-tartrate dehydratase Daud1254
Daud1256	imp	COG2358: TRAP-type uncharacterized transport system, periplasmic component	TPT27	345	49.86	B. halodurans	C4 dicarboxylate periplasmic; us citB citrate/malate regulator Daud1257
Daud1264	nptA	COG1283: Na+/phosphate symporter		559	58	D. reducens	Na/PO4 symp.
Daud1266	feoB	COG370: Fe2+ transport system protein B	TPT28	690	38.82	P. gingivalis	Fe2+
Daud1267	feoA	COG1918: Fe2+ transport system protein A	TPT28	84	40.74	D-monas spp.	Fe2+
Daud1307		hypothetical protein		88	67.44	P. thermopropionicum	TF for TPT3?

Daud1308	araJ	COG2814: Arabinose efflux permease	TPT29	405	57.22	D. reducens	sugar/drug export
Daud1309		COG1809: Uncharacterized conserved protein	TPT29	263	55.56	D. reducens	
Daud1310		COG2707: Predicted membrane protein	TPT29	153	55.1	D. reducens	
Daud1315	fepC	COG1120: ABC-type cobalamin/Fe3+-siderophores transport systems, ATPase components	TPT30	268	51.37	S. thermophilum	siderophore ATPase
Daud1316	fepD	COG609: ABC-type Fe3+-siderophore transport system, permease component	TPT30	339	48.66	Dehalo. sp. CBDB1	siderophore permease
Daud1317	fepB	COG614: ABC-type Fe3+-hydroxamate transport system, periplasmic component	TPT30	308	38.59	Dehalo. sp. CBDB1	siderophore periplasmic
Daud1355	secF	COG341: Preprotein translocase subunit SecF	TPT31	294	55.24	C. hydrogenoformans	protein sec
Daud1356	secD	COG342: Preprotein translocase subunit SecD	TPT31	411	55.04	D. reducens	protein sec
Daud1357		COG3294: Uncharacterized conserved protein	TPT31	220	76.61	D. reducens	protein sec
Daud1358	yajC	COG1862: Preprotein translocase subunit YajC	TPT31	91	69.88	P. thermopropionicum	protein sec
Daud1370	sufB	COG719: ABC-type transport system involved in Fe-S cluster assembly, permease component	TPT32	395	50.91	Dehalo. sp. CBDB1	
Daud1371	sufC	COG396: ABC-type transport system involved in Fe-S cluster assembly, ATPase component	TPT32	252	54.69	D. ethenogenes	
Daud1372		COG1578: Uncharacterized conserved protein	TPT32	276	35.71	P. abyssi	
Daud1380	trkA	COG569: K+ transport systems, NAD-binding component		216	68.4	P. thermopropionicum	K+
Daud1388	livF	COG410: ABC-type branched-chain amino acid transport systems, ATPase component	TPT33	237	72.34	P. thermopropionicum	AA ATPase
Daud1389	livG	COG411: ABC-type branched-chain amino acid transport systems, ATPase component	TPT33	257	66.01	P. thermopropionicum	AA ATPase
Daud1390	livM	COG4177: ABC-type branched-chain amino acid transport system, permease component	TPT33	335	65.51	P. thermopropionicum	AA permease
Daud1391	livH	COG559: Branched-chain amino acid ABC-type transport system, permease components	TPT33	295	74.15	P. thermopropionicum	AA permease
Daud1392	livK	COG683: ABC-type branched-chain amino acid transport systems, periplasmic component	TPT33	387	60.05	P. thermopropionicum	AA periplasmic
Daud1415	phoU	COG704: Phosphate uptake regulator	TPT34	219	57.8	D. reducens	PO4 TF

Daud1416	pstB	COG1117: ABC-type phosphate transport system, ATPase component	TPT34	252	78.71	P. thermopropionicum	PO4 ATPase
Daud1418	vicK	COG5002: Signal transduction histidine kinase	TPT34	445	48.33	P. thermopropionicum	PO4 HK
Daud1419	ompR	COG745: Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	TPT34	232	61.21	D. reducens	PO4 RR
Daud1471	yrbG	COG530: Ca2+/Na+ antiporter		331	50	M. thermoacetica	Ca+/Na+ antiport; K+ dependent?
Daud1474	tauA	COG715: ABC-type nitrate/sulfonate/bicarbonate transport systems, periplasmic components		333	34.34	D. reducens	bicarbonate? Periplasmic
Daud1540	pstB	COG1117: ABC-type phosphate transport system, ATPase component	TPT35	245	42.55	M. mazei	P04/SO4/WO4/MoO4 ATPase
Daud1541	tupA	COG4662: ABC-type tungstate transport system, periplasmic component	TPT35	218	45.75	A. variabilis	P04/SO4/WO4/MoO4 periplasmic
Daud1542	tupB	COG2998: ABC-type tungstate transport system, permease component	TPT35	284	56.07	D-monas spp.	P04/SO4/WO4/MoO4 permease
Daud1544	mgtA	COG474: Cation transport ATPase		830	45.05	P. carbinolicus	divalent cations, related to TPT10?; ds ATP-dependent protease LonB
Daud1545*		Pseudogene [COG1376, PF03734: ErfK/YbiS/YcfS/YnhG]	TPT36	410*	49.15	C. acetobutylicum	C. acetobutylicum annot.
Daud1546	znuC	COG1121: ABC-type Mn2+/Zn2+ transport systems, ATPase component	TPT36	249	40.93	B. clausii	divalent cations ATPase
Daud1547	znuB	COG1108: ABC-type Mn2+/Zn2+ transport systems, permease components	TPT36	280	53.26	M. thermoacetica	divalent cations permease
Daud1550	feoB	COG370: Fe2+ transport system protein B	TPT37	664	63.89	D. reducens	Fe2+
Daud1551	feoA	PF04023: Fe2+ transport system protein FeoA	TPT37	75	68.06	D. reducens	Fe2+
Daud1564	araJ	COG2814: Arabinose efflux permease		401	39.01	G. sulfurreducens	sugar/drug export
Daud1569		hypothetical protein	TPT38	55	N/A	ORFan	
Daud1570	ywcA	COG4147, PF00474: Na+/solute symporter	TPT38	694	34.68	S. avermitilis	Na+/solute symport
Daud1571		hypothetical protein	TPT38	123	32.1	N. farcinica	
Daud1639	cbiK	COG5266: ABC-type Co2+ transport system, periplasmic component		315	24.88	M. acetivorans	Co2+ periplasmic

Daud1856	cbiD	COG1903: Cobalamin biosynthesis protein CbiD	TPT39	370	46.59	M. thermoacetica	
Daud1857	cbiX	COG2138, PF01903: Cobalamin biosynthesis CbiX	TPT39	160	37.74	M. thermoacetica	
Daud1858	cbiO	COG1122: ABC-type cobalt transport system, ATPase component	TPT39	542	55.76	M. thermoacetica	Co2+ ATPase
Daud1859	cbiQ	COG619: ABC-type cobalt transport system, permease component CbiQ and related transporters	TPT39	258	39.68	M. thermoacetica	Co2+ permease
Daud1860	cbiN	COG1930: ABC-type cobalt transport system, periplasmic component	TPT39	97	58.82	M. maripaludis	Co2+ periplasmic
Daud1861	cbiM	COG310: ABC-type Co2+ transport system, permease component	TPT39	225	66.67	D. aromatica	Co2+ permease
Daud1863	mgtA	COG474: Cation transport ATPase		894	61.7	G. metallireducens	divalent cations ATPase; ds transposon and tRNA
Daud1876		COG1283: Na+/phosphate symporter		290	37.92	D-monas spp.	Na+/PO4 symport (short?)
Daud1878		hypothetical protein	TPT40	96	N/A	ORFan	TF?
Daud1879	zntA	COG2217: Cation transport ATPase	TPT40	837	52.88	E. faecalis	cation export?
Daud1899		COG701: Predicted permeases	TPT41	351	52.65	D. reducens	
Daud1900	arsR	COG640: Predicted transcriptional regulators	TPT41	106	54.55	P. thermopropionicum	TF?
Daud1909		COG1079: Unc. ABC-type transport system, permease component		306	71.15	D. reducens	
Daud1948*		COG2217 Cation transport ATPase		282*	43.28	C. glutamicum	cation ATPase; pseudogene?
Daud1951	feoB	COG370: Fe2+ transport system protein B	TPT42	681	61.55	T. tengcongensis	Fe2+
Daud1952	feoA	COG1918: Fe2+ transport system protein A	TPT42	80	58.75	M. thermoacetica	Fe2+
Daud1953	feoA	COG1918: Fe2+ transport system protein A	TPT42	81	50.63	D. ethenogenes	Fe2+
Daud1956	pstS	COG226: ABC-type phosphate transport system, periplasmic component	TPT43	284	59.22	D. reducens	PO4 periplasmic
Daud1957		COG622: Predicted phosphoesterase	TPT43	241	46.41	D. reducens	PO4
Daud1958	pstA	COG581: ABC-type phosphate transport system, permease component	TPT43	288	58.06	D. reducens	PO4 permease
Daud1959	pstC	COG573: ABC-type phosphate transport system, permease component	TPT43	291	63.35	D. reducens	PO4 permease

Daud1960		hypothetical protein	TPT43	73	N/A	ORFan	TF?
Daud1969	sbtA?	COG3329: Predicted permease, [putative sodium-dependent bicarbonate transporter]		335	45.02	Jann. sp. CCS1	Bicarbonate?; Prochlorococcus marinus MIT9313 annot.
Daud1975		COG730: Predicted permeases	TPT44	395	52.51	D. vulgaris DP4	
Daud1976		hypothetical protein	TPT44	65	N/A	ORFan	TF?; paralog of
Daud1982	ddpA	COG747: ABC-type dipeptide transport system, periplasmic component		509	39.44	D. reducens	dipeptide periplasmic
Daud1999		COG5413: Uncharacterized integral membrane protein	TPT45	179	28.33	Syn. sp. PCC 6803	
Daud2000		COG53: Predicted Co/Zn/Cd cation transporters	TPT45	286	56.2	P. thermopropionicum	cation diffusion
Daud2004	yugS	COG1253: Hemolysins and related proteins containing CBS domains		432	42.41	S. thermophilum	
Daud2034	araJ	COG2814: Arabinose efflux permease		387	29.18	D. hafniense DCB-2	sugar/drug export
Daud2079	prfB	COG1186: Protein chain release factor B	TPT46	331	71.56	P. thermopropionicum	protein sec
Daud2080	secA	COG653: Preprotein translocase subunit SecA (ATPase, RNA helicase)	TPT46	904	68.63	M. thermoacetica	protein sec
Daud2081	yvyD	COG1544: Ribosome-associated protein Y (PSrp-1)	TPT46	175	56.82	M. thermoacetica	
Daud2082	comFC	COG1040: Predicted amidophosphoribosyltransferases	TPT46	215	48.37	P. thermopropionicum	
Daud2099	acrB	COG841: Cation/multidrug efflux pump	TPT47	1044	45.6	M. thermoacetica	efflux
Daud2100	acrA	COG845: Membrane-fusion protein	TPT47	380	42.86	P. thermopropionicum	efflux?
Daud2101	marR	COG1846: Transcriptional regulators	TPT47	164	41.89	P. thermopropionicum	TF
Daud2120	tolC	COG1538: Outer membrane protein		375	46.54	P. thermopropionicum	efflux?
Daud2136	atpC	COG355: F0F1-type ATP synthase, epsilon subunit	TPT48	138	57.46	P. thermopropionicum	ATP synthase
Daud2137	atpD	COG55: F0F1-type ATP synthase, beta subunit	TPT48	473	85.93	P. thermopropionicum	ATP synthase
Daud2138	atpG	COG224: F0F1-type ATP synthase, gamma subunit	TPT48	296	61.82	P. thermopropionicum	ATP synthase
Daud2139	atpA	COG56: F0F1-type ATP synthase, alpha subunit	TPT48	508	79.4	C. hydrogenoformans	ATP synthase

Daud2140	atpH	COG712: F0F1-type ATP synthase, delta subunit	TPT48	183	47.73	D. reducens	ATP synthase
Daud2141	atpF	COG711: F0F1-type ATP synthase, subunit b	TPT48	164	48.77	D. reducens	ATP synthase
Daud2142	atpE	TIGR01260 ATPase, F0 complex, subunit c	TPT48	77	63.16	D. reducens	ATP synthase
Daud2143	atpB	COG356: F0F1-type ATP synthase, subunit a	TPT48	258	49.17	P. thermopropionicum	ATP synthase
Daud2144		hypothetical protein	TPT48	134	23.62	P. thermopropionicum	ATP synthase?
Daud2145		COG5336 Unc. protein conserved in bacteria	TPT48	100	45.71	P. thermopropionicum	ATP synthase?
Daud2158	mnhD-1	COG0651: Formate hydrogenlyase subunit 3/Multisubunit Na+/H+ antiporter, MnhD	TPT49 / FHL1	590	55.17	R. etli	Na+/H+ antiport?
Daud2159		hypothetical protein	TPT49 / FHL1	75	43.28	R. rubrum	TF?
Daud2160	mnhD-2	COG0651: Formate hydrogenlyase subunit 3/Multisubunit Na+/H+ antiporter, MnhD	TPT49 / FHL1	555	46.96	H. marismortui	Na+/H+ antiport?
Daud2161	mnhD-3	COG0651: Formate hydrogenlyase subunit 3/Multisubunit Na+/H+ antiporter, MnhD	TPT49 / FHL1	502	41.88	D-monas spp.	Na+/H+ antiport?
Daud2162	mnhC	COG1006: Multisubunit Na+/H+ antiporter, MnhC subunit	TPT49 / FHL1	131	44.35	D-monas spp.	Na+/H+ antiport?
Daud2163	mnhB-1	PF04039: Na+/H+ antiporter MnhB subunit-related protein	TPT49 / FHL1	144	38.24	S. frigidimarina	Na+/H+ antiport?
Daud2164	mnhB-2	[COG2111: Multisubunit Na+/H+ antiporter, MnhB subunit]	TPT49 / FHL1	98	48.86	D-monas spp.	Na+/H+ antiport?; D-monas spp. annot.
Daud2165	mnhB-3	[COG1563: Predicted subunit of the Multisubunit Na+/H+ antiporter]	TPT49 / FHL1	80	37.33	D-monas spp.	Na+/H+ antiport?; D-monas spp. annot.
Daud2166	mnhG	COG1320: Multisubunit Na+/H+ antiporter, MnhG subunit, TIGR01300	TPT49 / FHL1	103	45.19	D-monas spp.	Na+/H+ antiport?
Daud2167	mnhF	[COG2212: Multisubunit Na+/H+ antiporter, MnhF subunit]	TPT49 / FHL1	95	42.17	D-monas spp.	Na+/H+ antiport?; P.horikoshii annot.
Daud2168	mnhE	COG1863: Multisubunit Na+/H+ antiporter, MnhE subunit	TPT49 / FHL1	203	40.88	D-monas spp.	Na+/H+ antiport?
Daud2235	yidC	COG706: Preprotein translocase subunit YidC	TPT50	301	58.26	D. reducens	
Daud2236		COG759: Uncharacterized conserved protein	TPT50	72	62.12	Nos. sp. PCC 7120	Alpha-hemolysin?

Table S21. Amino acid synthesis genes.

Genes for amino acid synthesis were identified by membership in known sequence families (e.g. COG, TIGRFAM, and Pfam) or by gene context (proximity and/or presence in operons with other identified amino acid synthesis genes). Annotation was by protein family, or if no confident protein family could be assigned, by the protein family assignment of the nearest homolog (such annotations are indicated with square brackets, with the source organism provided in the notes).

Gene	Name	Description	Operon	Len	CH id	CH species	Notes
tRNA synthetases							
Daud0013	serS	COG172: Seryl-tRNA synthetase	SER1	426	67.61	D. reducens	ds tRNA-Ser, tRNA-Ser, tRNA-Arg, tRNA-Arg
Daud0041	metG	COG143: Methionyl-tRNA synthetase		517	73.29	P. thermopropionicum	us Daud0040 TF?; ds Daud0042 tatD Mg-dependent DNase
Daud0131	lysU	COG1190: Lysyl-tRNA synthetase (class II)		491	63.3	C. hydrogenoformans	us Daud0130 greA; ds DaudR0007 16S RNA
Daud0188	cysS	COG215: Cysteinyl-tRNA synthetase	CYS1	479	64.54	P. thermopropionicum	
Daud0484	glyQ	COG752: Glycyl-tRNA synthetase, alpha subunit	TRNA-G	296	80.56	P. thermopropionicum	
Daud0485	glyS	COG751: Glycyl-tRNA synthetase, beta subunit	TRNA-G	689	52.54	D. reducens	
Daud0618	proS	COG442: Prolyl-tRNA synthetase		567	66.84	P. thermopropionicum	us Daud0617 ispG; ds Daud0619 nfeD membrane-bound serine protease
Daud0905	hisS	COG124: Histidyl-tRNA synthetase	TRNA-HDN	421	61.89	D. reducens	
Daud0906	aspS	COG173: Aspartyl-tRNA synthetase	TRNA-HDN	602	69.86	P. thermopropionicum	
Daud0912	alaS	COG13: Alanyl-tRNA synthetase		881	60.67	P. thermopropionicum	
Daud1201	trpS	COG180: Tryptophanyl-tRNA synthetase		328	62.62	D. reducens	us Daud1202 Zn-dependent protease
Daud1368	tyrS	COG162: Tyrosyl-tRNA synthetase		433	66.17	D. reducens	
Daud1377	pheT	COG72: Phenylalanyl-tRNA synthetase beta	TRNA-F	801	55.71	P. thermopropionicum	ds Daud1376 TF?

		subunit					
Daud1378	pheS	COG16: Phenylalanyl-tRNA synthetase alpha subunit	TRNA-F	340	67.65	D. reducens	
Daud1384	thrS	COG441: Threonyl-tRNA synthetase		636	71.61	P. thermopropionicum	
Daud1408	ileS	COG60: Isoleucyl-tRNA synthetase		931	66.17	P. thermopropionicum	
Daud1473	valS	COG525: Valyl-tRNA synthetase		882	67.5	D. reducens	
Daud1531	glnS	COG8: Glutamyl- and glutaminyl-tRNA synthetases		502	57.84	M. thermoacetica	ds tRNA-Gln, tRNA-Glu
Daud1865	leuS	COG495: Leucyl-tRNA synthetase		827	66.31	M. thermoacetica	
Daud2181	argS	COG18: Arginyl-tRNA synthetase		564	61.07	D. reducens	
Ala							
Daud0472	alr	COG787: Alanine racemase		381	58.26	P. thermopropionicum	us Daud0471 sugar kinase
Daud0943	spoVFA / ald?	COG0686 Alanine dehydrogenase?	LYS1	311	49.66	P. thermopropionicum	spoVFA; paralogous to canonical alaDH; very distant from other clade members, which almost all have canonical form of alaDH
Arg							
Daud0339	argC	COG2: Acetylglutamate semialdehyde dehydrogenase	ARG1	347	65.22	P. thermopropionicum	
Daud0340	argJ	COG1364: N-acetylglutamate synthase (N-acetylornithine aminotransferase)	ARG1	408	59.95	P. thermopropionicum	
Daud0341	argB	COG548: Acetylglutamate kinase	ARG1	306	70.45	P. thermopropionicum	
Daud0342	argD	COG4992: Ornithine/acetylornithine aminotransferase	ARG1	418	64.81	P. thermopropionicum	
Daud0343	argF	COG78: Ornithine carbamoyltransferase	ARG1	310	66.56	P. thermopropionicum	
Daud0344	argG	COG137: Argininosuccinate synthase	ARG1	400	74.06	P. thermopropionicum	
Daud0345	argH	COG165: Argininosuccinate lyase	ARG1	460	73.9	P. thermopropionicum	

Daud0653	argD	COG4992: Ornithine/acetylornithine aminotransferase		468	54.1	T. tengcongensis	
Asn, Asp							
Daud0140	asnB	COG367: Asparagine synthase (glutamine-hydrolyzing)		615	68.46	P. thermopropionicum	
Daud0162	aspA	COG1027: Aspartate ammonia-lyase		485	67.74	P. thermopropionicum	
Cys							
Daud1311	cysK	COG31: Cysteine synthase		303	73.84	P. thermopropionicum	
Daud0187	cysE	COG1045: Serine acetyltransferase	CYS1	239	74.07	P. thermopropionicum	
Gln, Glu							
Daud2023	glnA	COG174: Glutamine synthetase	GLX1	444	74.22	D. reducens	
Daud2024	gltB	COG70: Glutamate synthase domain 3	GLX1	247	64.08	C. hydrogenoformans	
Daud2025	gltB	COG69: Glutamate synthase domain 2	GLX1	531	67.11	C. hydrogenoformans	
Daud2026	gltB	COG67: Glutamate synthase domain 1	GLX1	360	64.04	C. hydrogenoformans	
Daud2027	glnA	COG174: Glutamine synthetase	GLX1	444	72.52	D. reducens	
Daud2028	amiR	COG3707: Response regulator with putative antiterminator output domain	GLX1	193	50.8	D. reducens	RR
Daud2029	gltB	COG69: Glutamate synthase domain 2	GLX1	526	73.9	P. thermopropionicum	
Daud2030	glnA	COG174: Glutamine synthetase	GLX1	446	56.39	P. thermopropionicum	
Daud2031	amiR	COG3707: Response regulator with putative antiterminator output domain	GLX1	199	46.49	P. thermopropionicum	RR
Daud0550	ubiB	COG543: 2-polyprenylphenol hydroxylase and related flavodoxin oxidoreductases	GLX2	280	65.59	P. thermopropionicum	
Daud0551	gltD	COG493: NADPH-dependent glutamate synthase beta chain and related oxidoreductases	GLX2	480	66.01	M. thermoacetica	
Gly							

Daud2147	glyA	COG112: Glycine/serine hydroxymethyltransferase		416	68.05	D. reducens	
His							
Daud1449	his2	COG1387: Histidinol phosphatase and related hydrolases of the PHP family	HIS1	268	54.17	S. thermophilum	
Daud1450	metE?	hypothetical protein	HIS1	362	56.45	P. thermopropionicum	similar to (28%) [PF01717 Methionine synthase, vitamin-B12 independent] of Mycobacterium sp. MCS
Daud1621	hisI	COG139: Phosphoribosyl-AMP cyclohydrolase	HIS2	245	51.43	D. reducens	
Daud1622	hisF	COG107: Imidazoleglycerol-phosphate synthase	HIS2	253	78.97	P. thermopropionicum	
Daud1623	hisA	COG106: Phosphoribosylformimino-5-aminoimidazole carboxamide ribonucleotide (ProFAR) isomerase	HIS2	263	59.92	P. thermopropionicum	
Daud1624	hisH	COG118: Glutamine amidotransferase	HIS2	206	65	P. thermopropionicum	
Daud1625	hisB	COG131: Imidazoleglycerol-phosphate dehydratase	HIS2	197	64.06	P. thermopropionicum	
Daud1626	hisC	COG79: Histidinol-phosphate/aromatic aminotransferase and cobyric acid decarboxylase	HIS2	356	55.2	P. thermopropionicum	
Daud1627	hisD	COG141: Histidinol dehydrogenase	HIS2	440	62	P. thermopropionicum	
Daud0083	hisG	COG40, TIGR00070 Histidine biosynthesis HisG: ATP phosphoribosyltransferase		290	60.14	C. tepidum	
Daud1184	hisC	COG79: Histidinol-phosphate/aromatic aminotransferase and cobyric acid decarboxylase	ARO3	363	54.9	M. thermoacetica	
Lys							
Daud0942	dapB	COG289: Dihydrodipicolinate reductase	LYS1	255	57.65	P. thermopropionicum	
Daud0943	spoVFA / ald	COG0686 Alanine dehydrogenase	LYS1	311	49.66	P. thermopropionicum	see Ala

Daud0944	spoVFB	COG1036: Archaeal flavoproteins	LYS1	196	65.46	P. thermopropionicum	
Daud0945	asd	COG136: Aspartate-semialdehyde dehydrogenase	LYS1	337	69.79	C. hydrogenoformans	
Daud0946	dapA	COG329: Dihydrodipicolinate synthase/N-acetylneuraminate lyase	LYS1	295	71.13	P. thermopropionicum	
Daud0947	ykqC	COG595: Predicted hydrolase of the metallo-beta-lactamase superfamily	LYS1	555	74.86	P. thermopropionicum	ds tRNA-SeC
Daud1602		COG436: Aspartate/tyrosine/aromatic aminotransferase	LYS2 / TCA4	398	60.71	D. reducens	
Daud1603	dapL2	COG436: Aspartate/tyrosine/aromatic aminotransferase	LYS2 / TCA4	393	72.49	C. hydrogenoformans	
Daud1604	dapF	COG253: Diaminopimelate epimerase	LYS2 / TCA4	279	61.45	D. reducens	
Daud1204	lysA	COG19: Diaminopimelate decarboxylase		441	65.06	P. thermopropionicum	
Met							
Daud1450	metE?	hypothetical protein	HIS1	362	56.45	P. thermopropionicum	similar to (28%) [PF01717 Methionine synthase, vitamin-B12 independent] of Mycobacterium sp. MCS
Daud1690	metW?	PF08241 Methyltransferase type 11		236	42.31	F. tularensis	
Daud1996	metE	COG620: Methionine synthase II (cobalamin-independent)		755	52.37	B. halodurans	
Phe, Tyr, Trp							
Daud1904	pheA	COG77: Prephenate dehydratase	ARO1	373	44.3	P. thermopropionicum	
Daud1905	aroA	COG2876: 3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase	ARO1	322	60.64	D. reducens	
Daud2032	ilvE	COG115: Branched-chain amino acid aminotransferase/4-amino-4-deoxychorismate lyase	AA1	286	34.74	M. kandleri	
Daud2033	trpE	COG147: Anthranilate/para-aminobenzoate synthases component I	AA1	510	45.16	M. thermoacetica	

Daud2182		COG436: Aspartate/tyrosine/aromatic aminotransferase	AA2	406	64.4	D. reducens	
Daud2183	lrp	COG1522: Transcriptional regulators	AA2	163	61.78	P. thermopropionicum	
Daud0306	eno	COG148: Enolase	GLY1	428	72.6	P. thermopropionicum	
Daud0986	yqeG	COG2179: Predicted hydrolase of the HAD superfamily	ARO2	178	48.8	P. thermopropionicum	
Daud0987	aroE	COG169: Shikimate 5-dehydrogenase	ARO2	290	53.17	P. thermopropionicum	
Daud0988	aroC	COG82: Chorismate synthase	ARO2	391	66.41	P. thermopropionicum	
Daud0989	aroK	COG703: Shikimate kinase	ARO2	190	61.63	P. thermopropionicum	
Daud0990	aroB	COG337: 3-dehydroquinate synthetase	ARO2	357	54.83	D. reducens	
Daud1003	aroQ	COG757: 3-dehydroquinate dehydratase II	ARO2	151	62.42	S. thermophilum	
Daud1181	aroA	COG128: 5-enolpyruvylshikimate-3-phosphate synthase	ARO3	430	64.19	P. thermopropionicum	
Daud1182	tyrA	COG287: Prephenate dehydrogenase	ARO3	365	54.82	P. thermopropionicum	
Daud1183	aroA	COG2876: 3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase	ARO3	338	70.62	P. thermopropionicum	
Daud1184	hisC	COG79: Histidinol-phosphate/aromatic aminotransferase and cobyric acid decarboxylase	ARO3	363	54.9	M. thermoacetica	
Daud1185	trpA	COG159: Tryptophan synthase alpha chain	ARO3	268	49.62	D. reducens	
Daud1186	trpB	COG133: Tryptophan synthase beta chain	ARO3	395	71.91	P. thermopropionicum	
Daud1187	trpF	COG135: Phosphoribosylanthranilate isomerase	ARO3	205	52.43	D. psychrophila	
Daud1188	trpC	COG134: Indole-3-glycerol phosphate synthase	ARO3	276	52.55	M. thermoacetica	
Daud1189	trpD	COG547: Anthranilate phosphoribosyltransferase	ARO3	347	59.28	P. thermopropionicum	
Daud1190	pabA	COG512: Anthranilate/para-aminobenzoate synthases component II	ARO3	198	67.91	M. thermoacetica	
Daud1191	trpE	COG147: Anthranilate/para-aminobenzoate synthases component I	ARO3	488	54.73	D. reducens	
Daud1192		COG2206: HD-GYP domain	ARO4	365	46.28	D. reducens	ARO4 adjoins ARO3

Daud1193		COG2206: HD-GYP domain	ARO4	374	41.14	D. reducens	
Daud1194	aroH	COG4401: Chorismate mutase	ARO4	122	58.47	D. reducens	
Daud1195		hypothetical protein	ARO4	154	55.07	D. reducens	
Daud1205	trpB-like	COG1350: Predicted alternative tryptophan synthase beta-subunit (paralog of TrpB)		452	69.47	P. thermopropionicum	
Pro							
Daud1412	proC	COG345: Pyrroline-5-carboxylate reductase		276	51.13	A. dehalogenans	
Daud1870	proA	COG14: Gamma-glutamyl phosphate reductase	PRO1	420	69.71	M. thermoacetica	inserted gene between proA and proB
Daud1872	proB	COG263: Glutamate 5-kinase	PRO1	375	67.83	P. thermopropionicum	
Ser							
Daud0011		COG75: Serine-pyruvate aminotransferase/archaeal aspartate aminotransferase	SER1	385	60.16	C. hydrogenoformans	
Daud0012	serA	COG111: Phosphoglycerate dehydrogenase and related dehydrogenases	SER1	527	61.29	D. reducens	
Daud0825	serB	COG560: Phosphoserine phosphatase		226	35.75	M. thermautotrophicus	
Thr							
Daud0281	thrC	COG498: Threonine synthase		499	59.56	D. reducens	
Daud0945	asd	COG136: Aspartate-semialdehyde dehydrogenase		337	69.79	C. hydrogenoformans	
Daud1073	thrA	COG460: Homoserine dehydrogenase	THR1	432	66.59	P. thermopropionicum	
Daud1074	thrB	COG83: Homoserine kinase	THR1	314	52.19	P. thermopropionicum	
Daud1075	lysC	COG527: Aspartokinases	THR1	411	66	P. thermopropionicum	
Val, Leu, Ile							
Daud1279	leuB	COG473: Isocitrate/isopropylmalate		365	58.5	P. abyssi	

		dehydrogenase					
Daud2032	ilvE	COG115: Branched-chain amino acid aminotransferase/4-amino-4-deoxychorismate lyase	AA1	286	34.74	M. kandleri	
Daud0066	ilvA	COG1171: Threonine dehydratase		359	55.06	D. reducens	
Daud0351	ilvE	COG115: Branched-chain amino acid aminotransferase/4-amino-4-deoxychorismate lyase	AA3	294	79.18	C. hydrogenoformans	
Daud0352	ilvB	COG28: Thiamine pyrophosphate-requiring enzymes	AA3	557	69.51	P. thermopropionicum	
Daud0353	ilvH	COG440: Acetolactate synthase, small (regulatory) subunit	AA3	166	75	C. hydrogenoformans	
Daud0354	ilvC	COG59: Ketol-acid reductoisomerase	AA3	342	71.65	D. reducens	
Daud0355	leuA	COG119: Isopropylmalate/homocitrate/citramalate synthases	AA3	510	70.44	P. thermopropionicum	
Daud0356	leuC	COG65: 3-isopropylmalate dehydratase large subunit	AA3	421	76.79	D. reducens	
Daud0357	leuD	COG66: 3-isopropylmalate dehydratase small subunit	AA3	165	70.99	P. thermopropionicum	
Daud0358	leuA	COG119: Isopropylmalate/homocitrate/citramalate synthases	AA3	531	66.34	D. reducens	
Daud0540	ilvD	COG129: Dihydroxyacid dehydratase/phosphogluconate dehydratase	AA3	570	60.97	C. chlorochromatii	
Daud1035	leuA	COG119: Isopropylmalate/homocitrate/citramalate synthases	AA4	387	80.42	P. thermopropionicum	
Daud1036	leuC	COG65: 3-isopropylmalate dehydratase large subunit	AA4	419	73.21	D. reducens	
Daud1037	leuD	COG66: 3-isopropylmalate dehydratase small subunit	AA4	167	76.51	P. thermopropionicum	
Daud1038	leuB	COG473: Isocitrate/isopropylmalate dehydrogenase	AA4	337	75	P. thermopropionicum	

Table S22. Vitamin and cofactor synthesis genses.

Genes for vitamin and other cofactor synthesis were identified by membership in known sequence families (e.g. COG, TIGRFAM, and Pfam) or by gene context (proximity and/or presence in operons with other identified cofactor synthesis genes). Annotation was by protein family, or if no confident protein family could be assigned, by the protein family assignment of the nearest homolog (such annotations are indicated with square brackets, with the source organism provided in the notes).

While *D. audaxviator* has genes that appear to be coenzyme F420 dependent, it does not appear to have the canonical F420 synthetic pathway, lacking easily recognizable forms of cofC, cofD, cofE, cofG, and cofH, suggesting that either it possesses an alternate pathway for the synthesis of F420 or that the genes that appear to belong to F420-dependent families instead employ other cofactors. Additionally, *D. audaxviator* appears to be missing the canonical form of the pyrroloquinoline quinone synthesis genes (pqqABCDEF), with the possible exception of pqqF (also called pqqL), although the match between Daud0936 and the known pqqF in *Klebsiella pneumoniae* is weak (~29% identity) and only covers the N-terminal ¼ of the latter gene.

Gene	Name	Description	Operon	Len	CH id	CH species	Notes
Biotin and Thiamine							
Daud0161		hypothetical protein	BIO1	83	69.51	M. thermoacetica	TF?
Daud0162	aspA	COG1027: Aspartate ammonia-lyase	BIO1	485	67.74	P. thermopropionicum	
Daud0163	thiH	COG1060: Thiamine biosynthesis enzyme ThiH and related uncharacterized enzymes	BIO1	492	52.41	T. tengcongensis	
Daud0164		hypothetical protein	BIO1	129	N/A	ORFan	
Daud0165		COG1160: Predicted GTPases	BIO1	428	65.35	P. thermopropionicum	
Daud0166	bioB	COG0502 Biotin synthase and related enzymes	BIO1	371	57.85	P. thermopropionicum	
Daud0167	dsrE	COG1553, PF02635: DsrE-like protein	BIO1	109	39.42	M. mazei	ds tRNA-Asn; HGT?
Daud2017	thiL	COG611: Thiamine monophosphate kinase	THI1	338	43.88	D. reducens	
Daud2018	thiC	COG422: Thiamine biosynthesis protein ThiC	THI1	433	69.21	M. thermoacetica	
Daud2019	thiE	COG352: Thiamine monophosphate synthase	THI1	353	40.7	T. elongatus	
Daud0163	thiH	COG1060: Thiamine biosynthesis enzyme ThiH and	THI2	492	52.41	T. tengcongensis	

		related uncharacterized enzymes					
Daud0164		hypothetical protein	THI2	129	N/A	ORFan	
Daud0277	thiF	COG476: Dinucleotide-utilizing enzymes involved in molybdopterin and thiamine biosynthesis family 2		244	45.19	P. thermopropionicum	
Daud0479	thiM	COG2145: Hydroxyethylthiazole kinase, sugar kinase family		274	51.13	D. psychrophila	
Daud0859	thi4	COG1635, TIGR00292 Thiamine biosynthesis Thi4 protein		260	57.75	M. thermautotrophicus	
Daud1089	thiH	COG1060: Thiamine biosynthesis enzyme ThiH and related uncharacterized enzymes	THI3	369	66.58	C. hydrogenoformans	
Daud1090		COG1427: Predicted periplasmic solute-binding protein	THI3	275	50.18	D. reducens	
Daud1091	thiH	COG1060: Thiamine biosynthesis enzyme ThiH and related uncharacterized enzymes	THI3	357	63.46	D. reducens	
Daud1326	bioF	COG156: 7-keto-8-aminopelargonate synthetase and related enzymes	BIO2	385	47.76	G. sulfurreducens	
Daud1327	bioD	COG132: Dethiobiotin synthetase	BIO2	264	44.92	G. sulfurreducens	
Daud1328	bioA	COG161: Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	BIO2	454	54.32	G. metallireducens	
Daud1329		COG1647: Esterase/lipase	BIO2	224	33.33	M. barkeri	
Daud1330		COG4106 Trans-aconitate methyltransferase	BIO2	249	34.35	M. magneticum	
Daud1331	bioB	COG502: Biotin synthase and related enzymes	BIO2	329	55.24	P. thermopropionicum	
Daud2086	bioY	COG1268, PF02632 BioY protein		198	38.81	D. ethenogenes	
CoA							
Daud0632		COG742: N6-adenine-specific methylase	COA1	188	44.81	S. thermophilum	
Daud0633	coaD	COG669: Phosphopantetheine adenylyltransferase	COA1	164	62.73	C. hydrogenoformans	
Daud0634		[Archaeal/vacuolar-type H+-ATPase subunit H]	COA1	153	56.76	P. thermopropionicum	T. tengcongensis annot.
Daud1406	coaE	COG237: Dephospho-CoA kinase		198	47.69	P. thermopropionicum	
Daud1595	coaBC	COG452: Phosphopantothenoylcysteine synthetase/decarboxylase		411	54.66	P. thermopropionicum	

Cobalamin and Heme							
Daud0043	cobO / cobP / btuR	COG2109: ATP:corrinoid adenosyltransferase		177	64.16	P. thermopropionicum	
Daud1210	cobN	COG1429: Cobalamin biosynthesis protein CobN and related Mg-chelatases	TPT26	1211	46.17	M. thermautotrophicus	
Daud1219	cobN	COG1429: Cobalamin biosynthesis protein CobN and related Mg-chelatases	TPT26	1286	57.56	M. mazei	
Daud1312	cobU	COG2087: Adenosyl cobinamide kinase/adenosyl cobinamide phosphate guanylyltransferase	COB1	201	43.62	Dehalo. sp. CBDB1	
Daud1313	cobS	COG368: Cobalamin-5-phosphate synthase	COB1	253	38.46	Dehalo. sp. CBDB1	
Daud1314	cobT	COG2038: NaMN:DMB phosphoribosyltransferase	COB1	354	53.41	D. ethenogenes	
Daud1340	hemL	COG1: Glutamate-1-semialdehyde aminotransferase	COB2	433	66.35	M. thermoacetica	
Daud1341	lrp	COG1522: Transcriptional regulators	COB2	156	51.3	P. thermopropionicum	TF?
Daud1342	lrp	COG1522: Transcriptional regulators	COB2	158	55.26	D. reducens	TF?
Daud1343		COG535: Predicted Fe-S oxidoreductases	COB2	335	68.4	D. reducens	
Daud1344	hemB	COG113: Delta-aminolevulinic acid dehydratase	COB2	323	66.03	M. thermoacetica	
Daud1345		COG535: Predicted Fe-S oxidoreductases	COB2	395	63.14	D. reducens	
Daud1346	cysG	COG7: Uroporphyrinogen-III methylase	COB2	512	64.62	P. thermopropionicum	
Daud1347	hemC	COG181: Porphobilinogen deaminase	COB2	310	61.44	P. thermopropionicum	
Daud1348	hemA	COG373: Glutamyl-tRNA reductase	COB2	447	61.76	P. thermopropionicum	
Daud1349	cysG	COG1648: Siroheme synthase (precorrin-2 oxidase/ferrochelatase domain)	COB2	214	55.5	P. thermopropionicum	
Daud1350	ispA	COG142: Geranylgeranyl pyrophosphate synthase	COB2	319	58.12	D. reducens	
Daud1845		hypothetical protein	COB3	225	N/A	ORFan	
Daud1846	pleD	COG3706: Response regulator containing a CheY-like receiver domain and a GGDEF domain	COB3	315	43.81	G. sulfurreducens	RR
Daud1847	cbiB	COG1270: Cobalamin biosynthesis protein CobD/CbiB	COB3	320	46.1	C. tepidum	
Daud1848	cobQ	COG1492: Cobyric acid synthase	COB3	497	50.99	M. thermoacetica	

Daud1849	hisC	COG79: Histidinol-phosphate/aromatic aminotransferase and cobyric acid decarboxylase	COB3	347	38.35	M. thermoacetica	
Daud1850	cobB	COG1797: Cobyrinic acid a,c-diamide synthase	COB3	455	43.5	Dehalo. sp. CBDB1	
Daud1851	cobJ	COG1010: Precorrin-3B methylase	COB3	221	58.09	G. metallireducens	
Daud1852	cbiG	COG2073: Cobalamin biosynthesis protein CbiG	COB3	280	44.8	P. carbinolicus	
Daud1853	cobM	COG2875: Precorrin-4 methylase	COB3	248	61.76	M. thermoacetica	
Daud1854	cobF	COG2243: Precorrin-2 methylase	COB3	244	41.38	M. thermoacetica	
Daud1855	cobL	COG2241: Precorrin-6B methylase 1	COB3	217	49.01	P. carbinolicus	
Daud1856	cbiD	COG1903: Cobalamin biosynthesis protein CbiD	COB3	370	46.59	M. thermoacetica	
Daud1858	cbiO	COG1122: ABC-type cobalt transport system, ATPase component	COB3	542	55.76	M. thermoacetica	
Daud1859	cbiQ	COG619: ABC-type cobalt transport system, permease component CbiQ and related transporters	COB3	258	39.68	M. thermoacetica	
Daud1860	cbiN	COG1930: ABC-type cobalt transport system, periplasmic component	COB3	97	58.82	M. maripaludis	
Daud1861	cbiM	COG310: ABC-type Co2+ transport system, permease component	COB3	225	66.67	D. aromatica	
NAD/NADP							
Daud0120	panD?	hypothetical protein	NAD1	74	N/A	ORFan	truncated (C-term) paralog (83%) of Daud0102: panD Aspartate 1-decarboxylase
Daud0121	nadA	COG379: Quinolinate synthase	NAD1	304	63.42	P. thermopropionicum	
Daud0122	nadB	COG29: Aspartate oxidase	NAD1	539	57.44	P. thermopropionicum	
Daud0123	nadC	COG157: Nicotinate-nucleotide pyrophosphorylase	NAD1	285	60	D. reducens	
Daud0150	nadE	COG171: NAD synthase		544	55.83	M. thermoacetica	
Daud1029	yjbN	COG61, PF01513 ATP-NAD/AcoX kinase		284	39.08	D. reducens	
Daud1366	nadE	COG171: NAD synthase		242	58.8	C. hydrogenoformans	
Daud1868	nadD	COG1057: Nicotinic acid mononucleotide		211	62.37	D. reducens	

		adenylyltransferase					
Riboflavin							
Daud0623	ribD	COG117: Pyrimidine deaminase	RBF1	373	56.99	D. reducens	
Daud0624	ribC	COG307: Riboflavin synthase alpha chain	RBF1	216	64.35	P. thermopropionicum	
Daud0625	ribB	COG108: 3,4-dihydroxy-2-butanone 4-phosphate synthase	RBF1	403	75.88	P. thermopropionicum	
Daud0626	ribH	COG54: Riboflavin synthase beta-chain	RBF1	159	75.48	P. thermopropionicum	
Daud0930	ribF	COG196: FAD synthase		309	56.48	P. thermopropionicum	
S-adenosylmethionine							
Daud0649	ahcY	COG499: S-adenosylhomocysteine hydrolase		418	69.68	P. thermopropionicum	
Daud1085	speD	COG1586: S-adenosylmethionine decarboxylase	UBI1	130	76.98	D. reducens	
Daud2084	metK	COG192: S-adenosylmethionine synthetase		396	76.21	P. thermopropionicum	
Tetrahydrofolate (THF)							
Daud0095	folP	COG294: Dihydropteroate synthase and related enzymes	FOL1	397	64.45	P. thermopropionicum	
Daud0096	folB	COG1539: Dihydroneopterin aldolase	FOL1	124	56.78	M. thermoacetica	
Daud0097	folK	COG801: 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase	FOL1	170	62.42	P. thermopropionicum	
Daud1233	folE	COG302: GTP cyclohydrolase I		187	74.86	C. hydrogenoformans	
Daud1472	folC	COG285: Folylpolyglutamate synthase		426	51.05	P. thermopropionicum	
Ubiquinone							
Daud0168	ubiD	COG43: 3-polyprenyl-4-hydroxybenzoate decarboxylase and related decarboxylases		590	51.2	C. muridarum	
Daud0443	ubiE	COG2226: Methylase involved in ubiquinone/menaquinone biosynthesis		209	30.61	Silic. sp. TM1040	
Daud0482	ubiX	COG163: 3-polyprenyl-4-hydroxybenzoate decarboxylase		208	62.09	P. thermopropionicum	

Daud0673	ubiE	COG2226 Methylase involved in ubiquinone/menaquinone biosynthesis		253	35.21	N. punctiforme	
Daud0856	aarF?	COG661: Predicted unusual protein kinase [Ubiquinone biosynthesis protein AarF, putative]		556	36.91	D-monas spp.	G. metallireducens annot.
Daud1086		COG0142 Geranylgeranyl pyrophosphate synthase	UBI1	250	33.06	D. reducens	
Daud1087	ubiE	COG2226: Methylase involved in ubiquinone/menaquinone biosynthesis	UBI1	249	63.48	P. thermopropionicum	
Daud1088	ubiA	COG382: 4-hydroxybenzoate polyprenyltransferase and related prenyltransferases	UBI1	285	60.07	D. reducens	
Daud1089	thiH	COG1060: Thiamine biosynthesis enzyme ThiH and related uncharacterized enzymes	UBI1	369	66.58	C. hydrogenoformans	
Daud1090		COG1427: Predicted periplasmic solute-binding protein	UBI1	275	50.18	D. reducens	
Daud1091	thiH	COG1060: Thiamine biosynthesis enzyme ThiH and related uncharacterized enzymes	UBI1	357	63.46	D. reducens	
Daud1148	ubiE	COG2226: Methylase involved in ubiquinone/menaquinone biosynthesis		216	44.56	M. thermoacetica	
Daud1458	ubiE?	COG500: SAM-dependent methyltransferases, PF01209 UbiE/COQ5 methyltransferase (weak)		266	36.15	M. thermoacetica	PFAM hit is weak
Daud1733	aarF?	COG661: Predicted unusual protein kinase [Ubiquinone biosynthesis protein AarF, putative]		558	45.54	M. thermoacetica	G. metallireducens annot.
Daud1775	ubiG	COG2227: 2-polyprenyl-3-methyl-5-hydroxy-6-metoxy-1,4-benzoquinol methylase		299	37.76	S. baltica	
Daud1985	ubiE?	COG2230: Cyclopropane fatty acid synthase and related methyltransferases, PF01209 UbiE/COQ5 methyltransferase (weak)		262	40.84	M. bovis	PFAM hit is weak
Daud2102	ubiA	COG109: Polyprenyltransferase (cytochrome oxidase assembly factor), PF01040 UbiA prenyltransferase		318	42.91	S. thermophilum	

Table S23. Glycolysis/Gluconeogenesis and TCA cycle genes.

Genes for in the glycolytic, gluconeogenic, and transcarboxylic acid cycle pathways were identified by membership in known sequence families (e.g. COG, TIGRFAM, and Pfam) or by gene context (proximity and/or presence in operons with other identified central metabolism genes). Annotation was by protein family, or if no confident protein family could be assigned, by the protein family assignment of the nearest homolog (such annotations are indicated with square brackets, with the source organism provided in the notes). Putative pseudogenes are denoted with "*", and have length indicating number of nucleotides rather than amino acid length.

D. audaxviator is missing easily recognizable forms of succinyl-CoA synthetase, aconitase, and citrate synthase genes in the reverse transcarboxylic acid (TCA) cycle for assimilation of CO_2. *D. audaxviator* does have a gene (Daud0895) shared with archaea that may substitute for succinyl-CoA synthetase and may have other non-standard forms of genes that complete the TCA pathway (*69*), making it impossible to rule out its functionality.

Gene	Name	Description	Operon	Len	CH id	CH species	Notes
Glycolysis							
Daud0297	nagC	COG1940: Transcriptional regulator/sugar kinase	GLY1	322	51.26	C. hydrogenoformans	
Daud0298	gde	COG3408: Glycogen debranching enzyme	GLY1	669	45.58	M. barkeri	
Daud0299	pgi	COG166: Glucose-6-phosphate isomerase	GLY1	356	40.4	M. thermoacetica	
Daud0300		COG1660: Predicted P-loop-containing kinase	GLY1	295	62.15	D. reducens	
Daud0301		COG1481: Uncharacterized protein conserved in bacteria	GLY1	316	53.5	D. reducens	
Daud0302	gapA	COG57: Glyceraldehyde-3-phosphate dehydrogenase	GLY1	346	60.47	P. thermopropionicum	
Daud0303	pgk	COG126: 3-phosphoglycerate kinase	GLY1	394	58.67	C. hydrogenoformans	
Daud0304	tpiA	COG149: Triosephosphate isomerase	GLY1	258	66.53	D. reducens	
Daud0305	gpmI	COG696: Phosphoglyceromutase	GLY1	510	67.06	P. thermopropionicum	
Daud0306	eno	COG148: Enolase	GLY1	428	72.6	P. thermopropionicum	
Daud0307	secG	TIGR00810: Preprotein translocase SecG subunit	GLY1	77	67.11	D. reducens	hitchhiking?

Daud0308	hppA	COG3808, TIGR01104: V-type H(+)-translocating pyrophosphatase	GLY1	683	61.01	M. acetivorans	hitchhiking?
Daud0309		COG730: Predicted permeases	GLY1	320	57.01	D. vulgaris DP4	hitchhiking?
Daud0881	gapA	COG57: Glyceraldehyde-3-phosphate DH		359	63.83	C. hydrogenoformans	
Daud1051		PF02081: Tryptophan RNA-binding attenuator protein	GLY2	80	77.61	C. hydrogenoformans	
Daud1052		hypothetical protein	GLY2	68	57.14	D. reducens	
Daud1053	accD	COG777: Acetyl-CoA carboxylase beta subunit	GLY2	276	64.06	D. reducens	
Daud1054	accA	COG825: Acetyl-CoA carboxylase alpha subunit	GLY2	316	61.36	D. reducens	
Daud1055	pfkA	COG205: 6-phosphofructokinase	GLY2	321	70.94	P. thermopropionicum	
Daud1056	pykF	COG469: Pyruvate kinase	GLY2	586	60.85	P. thermopropionicum	
Daud1057	dedA	COG586: Uncharacterized membrane-associated protein	GLY2	196	48.69	P. thermopropionicum	
Daud1058		COG3635: Predicted phosphoglycerate mutase, AP superfamily	GLY2	406	59.2	P. thermopropionicum	
Daud1067	gpmB	COG406: Fructose-2,6-bisphosphatase		203	49.5	P. thermopropionicum	
Daud1116	pfkA	COG205: 6-phosphofructokinase		371	70.68	P. thermopropionicum	
Daud1158	acyP	COG1254: Acylphosphatases		95	54.44	P. thermopropionicum	
Daud1270	gapA	COG57: Glyceraldehyde-3-phosphate dehydrogenase		333	58.13	T. tengcongensis	extra copy
Daud2175	fba	COG191: Fructose/tagatose bisphosphate aldolase		289	67.72	C. hydrogenoformans	
Gluconeogenesis							
Daud1838		hypothetical protein	NEO1	82	N/A	ORFan	
Daud1839	fbp	COG1980: Archaeal fructose 1,6-bisphosphatase	NEO1	371	79.12	P. thermopropionicum	thermophile associated
Daud1840	ppa	COG221: Inorganic pyrophosphatase	NEO1	172	57.24	S. thermophilum	
Daud1926	ppsA	COG574: Phosphoenolpyruvate synthase/pyruvate phosphate dikinase		891	75.51	P. thermopropionicum	
Daud1946	ppsA	COG574: Phosphoenolpyruvate synthase/pyruvate phosphate dikinase		1116	39.2	Desulfuromonas spp.	

Daud2089	manB	COG1109: Phosphomannomutase		468	62.82	P. thermopropionicum	
L-lactate dehydrogenase							
Daud1522	mdh	COG39: Malate/lactate dehydrogenases		311	57.28	P. thermopropionicum	
Pyruvate decarboxylase							
Daud1015		COG5016: Pyruvate/oxaloacetate carboxyltransferase	PYR1	615	70.36	C. hydrogenoformans	
Daud1016		COG4770: Acetyl/propionyl-CoA carboxylase, alpha subunit	PYR1	447	71.11	P. thermopropionicum	
Pyruvate dehydrogenase							
Daud0280	hcaD	COG446: Uncharacterized NAD(FAD)-dependent dehydrogenases		568	54.76	P. thermopropionicum	
Daud2077		COG3958: Transketolase, C-terminal subunit	PYR2	309	74.51	P. thermopropionicum	
Daud2078		COG3959: Transketolase, N-terminal subunit	PYR2	272	78.95	P. thermopropionicum	
Acetate-CoA ligase							
Daud1572	acs	COG365: Acyl-coenzyme A synthetases/AMP-(fatty) acid ligases		661	66.51	C. hydrogenoformans	
Transcarboxylic acid (TCA) cycle							
Daud0773	ppc	COG1892, TIGR02751: archaeal-type phosphoenolpyruvate carboxylase		490	55.17	M. thermautotrophicus	no bacterial form of ppc in genome
Daud0895		COG1042: Acyl-CoA synthetase (NDP forming)	TCA1	701	62.46	P. thermopropionicum	candidate atypical succinyl-CoA synthetase; matches alpha subunit in P. thermopropionicum, but no ortholog for P. thermopropionicum beta subunit

Daud0896	pta	COG857: BioD-like N-terminal domain of phosphotransacetylase	TCA1	352	53.71	P. thermopropionicum	
Daud1035	leuA	COG0119: Isopropylmalate/homocitrate/citramalate synthases, TIGR02660: Homocitrate synthase NifV-type	TCA2	386	80.42	P. thermopropionicum	candidate atypical citrate synthase ; weak match to confirmed atypical citrate synthase from D. vulgaris Hildenborough (Dvu_0398)
Daud1036	leuC	COG0065: 3-isopropylmalate dehydratase large subunit, TIGR01343: Cis-homoaconitase	TCA2	418	73.86	P. thermopropionicum	candidate atypical aconitate hydratase
Daud1037	leuD	COG0066, TIGR02087: 3-isopropylmalate dehydratase, small subunit, archaeal like, PF00694: Aconitate hydratase, C-terminal	TCA2	166	76.51	P. thermopropionicum	candidate atypical aconitate hydratase
Daud1038	leuB	COG473: Isocitrate/isopropylmalate dehydrogenase	TCA2	337	75	P. thermopropionicum	
Daud1048		COG1832: Predicted CoA-binding protein		140	45.93	D. reducens	candidate atypical succinyl-CoA synthetase
Daud1121	mdh	COG39: Malate/lactate dehydrogenases		310	71.15	P. thermopropionicum	
Daud1249	sdhA	COG1053: Succinate dehydrogenase/fumarate reductase, flavoprotein subunit	TCA3	533	50.47	G. metallireducens	TCA?
Daud1250		COG4656: Predicted NADH:ubiquinone oxidoreductase, subunit RnfC	TCA3	228	35.24	G. metallireducens	TCA?
Daud1251	hdrB	COG2048: Heterodisulfide reductase, subunit B	TCA3	283	37.32	C. hydrogenoformans	TCA?
Daud1252	hdrC	COG1150: Heterodisulfide reductase, subunit C	TCA3	161	50	C. hydrogenoformans	TCA?
Daud1253	fumA	COG1838: Tartrate dehydratase beta subunit/Fumarate hydratase class I, C-terminal domain	TCA3	187	60.89	P. thermopropionicum	TCA?
Daud1254	ttdA	COG1951: Tartrate dehydratase alpha subunit/Fumarate hydratase class I, N-terminal domain	TCA3	295	63.83	D. reducens	TCA?
Daud1563	pckA	COG1866: Phosphoenolpyruvate carboxykinase (ATP)		522	62.99	P. thermopropionicum	
Daud1606	porG	COG1014: Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, gamma subunit	LYS2 / TCA4	193	56.35	P. thermopropionicum	putative 2-oxoglutarate synthase ; paralog Daud2073

Daud1607	porB	COG1013: Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, beta subunit	LYS2 / TCA4	222	65.9	P. thermopropionicum	putative 2-oxoglutarate synthase ; paralogs Daud1961, Daud1965
Daud1608	porA	COG674: Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, alpha subunit	LYS2 / TCA4	368	64.55	P. thermopropionicum	putative 2-oxoglutarate synthase ; paralogs Daud2075, Daud1962, Daud1966
Daud1609		COG1149: MinD superfamily P-loop ATPase containing an inserted ferredoxin domain	LYS2 / TCA4	81	54.29	P. thermopropionicum	paralog Daud2072
Daud1926	ppsA	COG574: Phosphoenolpyruvate synthase/pyruvate phosphate dikinase	TCA5	891	75.51	P. thermopropionicum	
Daud1927	ttdA	COG1951: Tartrate dehydratase alpha subunit/Fumarate hydratase class I, N-terminal domain	TCA5	281	58.21	D. reducens	
Daud1928	fumA	COG1838: Tartrate dehydratase beta subunit/Fumarate hydratase class I, C-terminal domain	TCA5	188	51.38	D. reducens	
Daud1929	fdrC	[fumarate reductase cytochrome B subunit]	TCA5	213	37.98	P. thermopropionicum	Shewanella oneidensis MR-1 annot.
Daud1930	sdhA	COG1053: Succinate dehydrogenase/fumarate reductase, flavoprotein subunit	TCA5	596	60.5	P. thermopropionicum	
Daud1931	frdB	COG479: Succinate dehydrogenase/fumarate reductase, Fe-S protein subunit	TCA5	248	64.29	P. thermopropionicum	
Daud2072		COG1146: Ferredoxin	TCA6	93	50	P. thermopropionicum	paralog Daud1609
Daud2073	porG	COG1014: Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, gamma subunit	TCA6	196	68.45	C. hydrogenoformans	putative 2-oxoglutarate synthase ; paralog Daud1606
Daud2074*	porB	COG1013: Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, beta subunit	TCA6	779*	67.46	P. thermopropionicum	putative 2-oxoglutarate synthase ; Pseudogene?
Daud2075	porA	COG674: Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, alpha subunit	TCA6	381	66.94	D. reducens	putative 2-oxoglutarate synthase ; paralogs Daud1608, Daud1962, Daud1966
Daud2076		hypothetical protein	TCA6	238	61.01	D. reducens	

Table S24. Hydrogenases, dehydrogenases, and other oxidoreductases.

Genes for oxidoreductase activity (that are not already reported in the preceeding tables) were identified by membership in known sequence families (COG). Annotation was by protein family. Putative pseudogenes are denoted with "*", and have length indicating number of nucleotides rather than amino acid length.

Gene	Name	Description	Opero	Len	CH id	CH species	Notes
Daud0025	mviM	COG806: Predicted dehydrogenases and related proteins		334	44.33	V. vulnificus	
Daud0129		COG871: Predicted dehydrogenase		372	63.94	C. hydrogenoformans	
Daud0134	hybA	COG1521: Fe-S-cluster-containing hydrogenase components 1	OR1	162	48.2	M. acetivorans	
Daud0135		COG507: Aldehyde:ferredoxin oxidoreductase	OR1	579	48.45	M. acetivorans	
Daud0152		COG473: Iron only hydrogenase large subunit, C-terminal domain		581	63.76	M. thermoacetica	
Daud0156	hyaD	COG1305: Ni,Fe-hydrogenase maturation factor	OR2A	178	25.15	A. dehalogenans	
Daud0157		COG506: Iron only hydrogenase large subunit, C-terminal domain	OR2A	524	69.65	P. thermopropionicum	
Daud0158*	hycB	COG1444: Fe-S-cluster-containing hydrogenase components 2	OR2A	561*	51.4	P. thermopropionicum	ds transposon; pseudogene?
Daud0160		COG914: Polysulphide reductase	OR2B	403	43	P. thermopropionicum	
Daud0276	fabG	COG634: Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)		261	60.7	T. thermophilus HB8	
Daud0279		COG166: Aldehyde:ferredoxin oxidoreductase		580	56.3	G. metallireducens	
Daud0318	nfnB	COG578: Nitroreductase		240	54.63	T. kodakaraensis	
Daud0369	hybA	COG1007: Fe-S-cluster-containing hydrogenase components 1		249	30.34	Shewanella sp ANA-3	
Daud0413		COG1152: Iron only hydrogenase large subunit, C-terminal domain		486	64.92	D. reducens	
Daud0521		COG1582: Predicted oxidoreductases of the aldo/keto reductase family		295	43.88	Nos. sp. PCC 7120	
Daud0559	trxB	COG1086: Thioredoxin reductase		305	50.33	P. carbinolicus	
Daud0615	dxr	COG1060: 1-deoxy-D-xylulose 5-phosphate reductoisomerase		385	56.05	D. reducens	
Daud0642	fabG	COG1233: Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)		248	60	G. kaustophilus	

Daud0914	tas	COG1302: Predicted oxidoreductases (related to aryl-alcohol dehydrogenases)		317	55.21	C. hydrogenoformans	
Daud1041	hyaA	COG1031: Ni,Fe-hydrogenase I small subunit	OR3	319	60.45	P. thermopropionicum	
Daud1042	hyaB	COG1002: Ni,Fe-hydrogenase I large subunit	OR3	486	65	P. thermopropionicum	
Daud1043	hybA	COG1116: Fe-S-cluster-containing hydrogenase components 1	OR3	281	42.41	D. reducens	
Daud1044		COG1550: Polysulphide reductase	OR3	402	43.65	P. thermopropionicum	
Daud1139	trxB	COG219: Thioredoxin reductase		294	46.48	D. hafniense Y51	
Daud1156		COG168: Predicted Fe-S oxidoreductase		308	49.83	Gloeobacter violaceus	
Daud1166	eutG	COG1189: Alcohol dehydrogenase, class IV	OR4	383	42.54	P. thermopropionicum	
Daud1167	nfnB	COG332: Nitroreductase	OR4	175	47.37	A. fulgidus	
Daud1168	gpsA	COG1138: Glycerol-3-phosphate dehydrogenase	OR4	352	65.37	P. thermopropionicum	
Daud1173		COG745: Fe-S oxidoreductase, related to NifB/MoaA family		452	50.69	D. reducens	
Daud1300	pyrD	COG235: Dihydroorotate dehydrogenase	OR5	308	67.21	P. thermopropionicum	
Daud1301	ubiB	COG1276: 2-polyprenylphenol hydroxylase and related flavodoxin oxidoreductases	OR5	271	43.66	D. reducens	
Daud1337		COG1555: Iron only hydrogenase large subunit, C-terminal domain	OR6	591	65.16	M. thermoacetica	
Daud1338	nuoF	COG404: NADH:ubiquinone oxidoreductase, NADH-binding (51 kD) subunit	OR6	573	65.19	D. reducens	
Daud1339	nuoE	COG515: NADH:ubiquinone oxidoreductase 24 kD subunit	OR6	156	54.61	P. thermopropionicum	
Daud1434	murB	COG1358: UDP-N-acetylmuramate dehydrogenase		301	52.03	P. thermopropionicum	
Daud1455		COG751: Fe-S oxidoreductase		620	61.8	P. thermopropionicum	
Daud1507	hgdB	COG1454: Benzoyl-CoA reductase/2-hydroxyglutaryl-CoA dehydratase subunit, BcrC/BadD/HgdB		327	68.11	P. thermopropionicum	
Daud1548		COG935: Iron only hydrogenase large subunit, C-terminal domain		616	60.82	P. thermopropionicum	
Daud1642	hypE	COG1213: Hydrogenase maturation factor	OR7	339	70.75	M. thermoacetica	
Daud1643	hypD	COG1445: Hydrogenase maturation factor	OR7	374	63.2	M. thermoacetica	
Daud1644	hypC	COG601: Hydrogenase maturation factor	OR7	81	50.63	M. thermoacetica	
Daud1645	hypF	COG998: Hydrogenase maturation factor	OR7	774	64.68	M. thermoacetica	
Daud1648	hypB	COG856: Ni^{2+}-binding GTPase involved in regulation of	OR7	224	68.66	M. thermoacetica	

		expression and maturation of urease and hydrogenase					
Daud1649	hybF	COG342: Zn finger protein HypA/HybF (possibly regulating hydrogenase expression)	OR7	137	37.17	M. thermoacetica	
Daud1650	hyaD	COG1044: Ni,Fe-hydrogenase maturation factor	OR7	162	41.83	D-monas spp.	
Daud1651	frhA	COG275: Coenzyme F420-reducing hydrogenase, alpha subunit	OR7	486	41.5	A. dehalogenans	
Daud1652	frhG	COG548: Coenzyme F420-reducing hydrogenase, gamma subunit	OR7	318	39.56	A. dehalogenans	
Daud1686	rfbD	COG398: dTDP-4-dehydrorhamnose reductase		286	65.33	P. thermopropionicum	
Daud1720	nfnB	COG321: Nitroreductase		197	57.45	P. thermopropionicum	
Daud1725		COG633: Predicted Fe-S oxidoreductases		461	73.5	C. hydrogenoformans	
Daud1923	nuoF	COG918: NADH:ubiquinone oxidoreductase, NADH-binding (51 kD) subunit	OR8	540	58.44	P. thermopropionicum	
Daud1924	nuoE	COG722: NADH:ubiquinone oxidoreductase 24 kD subunit	OR8	185	46.79	D. reducens	
Daud1939	hyaA	COG1292: Ni,Fe-hydrogenase I small subunit	OR9	376	44.48	D. hafniense Y51	
Daud1940	hyaB	COG336: Ni,Fe-hydrogenase I large subunit	OR9	476	45	C. hydrogenoformans	
Daud1941		COG148: Thiosulfate reductase cytochrome B subunit (membrane anchoring protein)	OR9	187	36.22	C. hydrogenoformans	
Daud1944	nfnB	COG540: Nitroreductase		190	53.72	D-monas spp.	
Daud1961	porB-1	COG1013 Pyruvate:ferredoxin oxidoreductase and related, beta subunit	OR10	287	48.77	A. aeolicus	HGT; paralog Daud1965
Daud1962	porA-1	COG0674 Pyruvate:ferredoxin oxidoreductase and related, alpha subunit	OR10	395	56.53	A. aeolicus	HGT; paralog Daud1966
Daud1963	porD	COG1144 Pyruvate:ferredoxin oxidoreductase and related, delta subunit	OR10	65	48.21	A. fulgidus	HGT
Daud1964	porG	COG1014 Pyruvate:ferredoxin oxidoreductase and related, gamma subunit	OR10	224	52.22	A. aeolicus	HGT
Daud1965	porB-2	COG1013 Pyruvate:ferredoxin oxidoreductase and related, beta subunit	OR10	291	54.55	A. aeolicus	HGT; paralog Daud1961
Daud1966	porA-2	COG0674 Pyruvate:ferredoxin oxidoreductase and related, alpha subunit	OR10	381	54.52	A. aeolicus	HGT; paralog

							Daud1962
Daud1967	cdhB / acsE	PF02552: CODH beta subunit/acetyl-CoA synthase epsilon subunit	OR10	206	60.67	A. aeolicus	HGT
Daud1968	glnB-like	SSF54913 GlnB-like, Nitrogen regulatory protein P-II	OR10	106	33.03	A. aeolicus	HGT; regulator?
Daud1969	sbtA?	COG3329: Predicted permease, [putative sodium-dependent bicarbonate transporter]	OR10	335	45.02	Jann. sp. CCS1	Bicarbonate ?; Prochloroc occus marinus MIT9313 annot.
Daud1986		COG788: Predicted Fe-S oxidoreductases		454	45.27	M. bovis	
Daud1995	nrdD	COG239: Oxygen-sensitive ribonucleoside-triphosphate reductase		710	59.94	T. tengcongensis	
Daud2043	eutG	COG485: Alcohol dehydrogenase, class IV		859	62.56	P. thermopropionicum	
Daud2157	nirD	COG604: Ferredoxin subunits of nitrite reductase and ring-hydroxylating dioxygenases		183	34.52	D-monas spp.	
Daud2204		COG512: Cytochrome c biogenesis factor		192	30.69	C. hydrogenoformans	
Daud2206	aslB	COG1585: Arylsulfatase regulator (Fe-S oxidoreductase)		457	49.45	D. reducens	
Daud2210		COG831: Uncharacterized FAD-dependent dehydrogenases		459	66.16	C. hydrogenoformans	

Table S25. Oxygen tolerance.

Fracture environments at this depth are anoxic (*3*). Accordingly, the *D. audaxviator* genome lacks obvious functional homologs of catalase, peroxidase, and superoxide reductase, but does possess Mn/Fe superoxide dismutase that converts O_2^- to H_2O_2. It also lacks obvious full-length homologs to most of the rubredoxin / rubrerythrin O_2 tolerance system, with the exception of rubrerythrin which allows it to convert the H_2O_2 produced by superoxide dismutase, or from radiolytic reactions (*3*), to H_2O. A very truncated pseudogene for catalase was found, as was a pseudogene for another instance of rubrerythrin. The loss of most of the O_2 tolerance systems suggests the long-term sequestration from O_2 and isolation from the surface, and has likely contributed to the failure to isolate *D. audaxviator*.

Genes for oxygen tolerance were identified by membership in known sequence families (COG). Annotation was by protein family. Putative pseudogenes are denoted with "*", and have length indicating number of nucleotides rather than amino acid length.

Gene	Name	Description	Group	Len	CH id	CH species	Notes
Daud0372*	katE*	COG753 Catalase (N-terminal)	PG3	161*	72	S. wittichii	very short pseudogene
Daud0543	rbr	COG1592 Rubrerythrin		178	70.86	D. reducens	incomplete P. thermo. genome
Daud0583*	rbr*	COG1592 Rubrerythrin	PG7	456*	73.27	D. audaxviator	short split pseudogene
Daud1059	sodA	COG605 Superoxide dismutase		197	73.58	D. reducens	incomplete P. thermo. genome

Table S26. Pseudogenes.

Pseudogenes (protein coding genes that are no longer functional due to early stop codons or are otherwise truncated, split, or frameshifted) were identified at ORNL. The *D. audaxviator* genome possessed 83 pseudogenes, more than 48 of *D. reducens*, the 25 of *C. hydrogenoformans,* and the 58 of *M. thermoacetica* (pseudogene counts were not available for *P. thermopropionicum*), all larger genomes than that of *D. audaxviator*. The relatively large number of pseudogenes corresponded with the large number of transposons, not surprising given that many of the pseudogenes themselves represent transposon "scars" or were likely caused by adjacent transposons.

We classified pseudogenes by BLASTx (translating BLAST) by similarity to known protein coding genes in D. audaxviator or other organisms. These pseudogenes did not themselves possess full-length open reading frames due to transposon invasion, truncation, early stop codons, or frameshifts. Classification was by membership of the closest homolog (CH) in known sequence families (COG, PFAM, or TIGRFAM). Regions with a high density of pseudogenes are indicated by assigning a "PG" group (Group). Lengths indicate number of nucleotides rather than amino acid length. The percentage of the matched functional gene is reported (CH cov %) as is the amino acid identity over that match (CH ident), and the VIMSS gene ID of the match (CH VIMSS ID). Many of the pseudogenes either represented remanants of transposons or may have been caused by proximal or interrupting transposon activity and are indicated, as are other reasons for the cause of the pseudogene in the pseudogene character field (ψgene character). Immediate sequence upstream (us) and downstream (ds) of matched pseudogene sequence up until the next gene or pseudogene were also scanned with BLASTx against known functional proteins to allow for additional classification, primarily for genes interrupted by transposons (e.g. Daud0158, which is interrupted by the transposon Daud0159). Functional homologs present in *D. audaxviator* are

also reported in the Notes column. While the majority of the pseudogenes represent derivatives of transposon activity, of note are the small remaining piece of the missing catalase gene (Daud0372) and the broken duplicate copy of rubrerythrin (Daud0583). Many of the other interesting pseudogenes are redox proteins (Daud0153, Daud0154, Daud0155, Daud0158, Daud0577, Daud0739, Daud0791, Daud0834, Daud1097, Daud1646, Daud1828, Daud2074) or transport proteins (Daud0374, Daud0532, Daud1417, Daud1513, Daud1545, Daud1830, Daud1948) that are either difficult to specifically classify or have duplicate functional versions present in the genome that may take the role of the lost proteins, making it difficult to reliably infer the loss of a capability.

Gene	Name	Description	Group	Len	CH len	CH cov (%)	CH ident	CH VIMSS ID	CH species	ψgene character	Notes
Daud0055		COG3591 Putative secreted protein		240	1014	23	64.56	1366212	D. audaxviator	short	
Daud0153	nuoF	COG1894 NADH:ubiquinone oxidoreductase, NADH-binding	PG1	300	1833	14	35.63	764555	M. capsulatus	short	other nuoF present
Daud0154us		COG4624 Hydrogenases, Fe-only	PG1	579	1743	31	70.22	1365961	D. audaxviator	split	hom. Daud0152
Daud0154		COG4624 Hydrogenases, Fe-only	PG1	1748	1743	67	64.66	1365961	D. audaxviator	split	hom. Daud0152
Daud0155us	nuoF	COG1894 NADH:ubiquinone oxidoreductase	PG1	463	1773	20	44.17	1902758	S. wolfei	split	hom. Daud0109
Daud0155	frhD	COG1908 Coenzyme F420-reducing hydrogenase, delta subunit	PG1	819	441	87	68.75	1367395	D. audaxviator	split	hom. Daud1653
Daud0158us	hybA	COG437 Fe-S-cluster-containing hydrogenase components 1	PG1	449	786	35	44.09	2809954	D. reducens	interrupted	hom. Daud1043
Daud0158		PF01609 Transposase, IS4-like	PG1	2571	1332	100	100	1367922	D. audaxviator	transposon	is Daud0159
Daud0158ds	hybA	COG437 Fe-S-cluster-containing hydrogenase components 1	PG1	849	843	68	69.79	1366796	D. audaxviator	interrupted	hom. Daud1043

Daud0195us	pcrA	COG210 Superfamily I DNA and RNA helicases	PG2	641	2121	5	62.86	2265173	C. thermocellum	short	hom. Daud1619
Daud0195		PF04986 Transposase, IS801/IS1294	PG2	213	1542	12	52.38	966774	K. pneumoniae	transposon, short	
Daud0195ds		TIGR01439 transcriptional regulator, AbrB family	PG2	663	267	53	81.25	1366171	D. audaxviator	short	hom. Daud0376
Daud0197		PF01609 Transposase, IS4-like	PG2	480	741	50	87.2	1366209	D. audaxviator	transposon, short	hom. Daud0695
Daud0204		TIGR01439 transcriptional regulator, AbrB family		247	366	25	70.97	1362309	P. thermopropionicum	short? split?	
Daud0204ds		conserved hypothetical protein		281	261	34	86.67	1367652	D. audaxviator	short? split?	hom. Daud1918
Daud0315us		COG1293 Predicted RNA-binding protein		415	1761	5	82.14	1367049	D. audaxviator	short	hom. Daud1297
Daud0315		COG2801 Transposase and inactivated derivatives		279	807	15	82.5	3191939	B. coagulans	transposon, short	hom. Daud0786
Daud2240		COG3328 Transposase and inactivated derivatives		99	1218	8	71.9	3286398	Leptosp. Group II UBA	transposon, short	
Daud0322		COG3547 Transposase and inactivated derivatives		362	1284	14	62.3	238950	T. tengcongensis	transposon, short	
Daud0322ds		COG3547 Transposase and inactivated derivatives		341	1284	11	60.42	1082757	C. hydrogenoformans	transposon, short	
Daud0326		PF06782 protein of unknown function UPF0236		318	1419	20	98.96	1366861	D. audaxviator	transposon, short	hom. Daud1109
Daud0365		PF05016 plasmid stabilization system	PG3	201	333	53	78.33	1366516	D. audaxviator	short	hom. Daud0750
Daud0371		COG1708 Predicted nucleotidyltransferases (DNA polymerase, beta-like region)	PG3	893	489	56	40.22	3050001	R. castenholzii	split	
Daud0371ds		COG2445 Uncharacterized conserved protein	PG3	500	420	57	33.33	1366162	D. audaxviator	split	hom. Daud0363
Daud0372	katE	COG753 Catalase (N-terminal)	PG3	161	1515	5	72	3129626	S. wittichii	short	CATALASE!

Daud0374	rarD	COG2962 Predicted permeases	PG3	872	927	66	54.85	66447	B. halodurans	split	
Daud0374ds	rarD	COG2962 Predicted permeases	PG3	409	942	21	51.52	2841552	B. weihenstephanensis	split	
Daud0393	hepA	COG553 Superfamily II DNA/RNA helicases, SNF2 family	PG4	2930	2877	42	31.28	1116365	N. pharaonis	split	
Daud0393ds	hepA	COG553 Superfamily II DNA/RNA helicases, SNF2 family	PG4	1592	2940	50	26.43	70314	Halobact. sp. NRC-1	split	note: Daud0394 hom. N. pharaonis 1116366
Daud0396		COG1205 Distinct helicase family	PG4	5248	5304	47	28.25	1116367	N. pharaonis		note: Daud0394 hom. N. pharaonis 1116366
Daud0396ds		COG1205 Distinct helicase family	PG4	2874	5304	46	29.2	1116367	N. pharaonis		
Daud0397		COG4637 Predicted ATPase	PG4	489	1230	31	46.51	2087354	A. avenae	short	
Daud0400us	cheC	COG1776 Chemotaxis protein CheC, inhibitor of MCP methylation	PG4	408	1110	9	100	1367473	D. audaxviator	short	hom. Daud1736
Daud0400		ATPase associated with various cellular activities, AAA_5	PG4	1299	2103	42	42.76	2850751	M. flavescens		note: Daud0399 is a transposon
Daud0404		PF00665 Integrase, catalytic core	PG4	141	1572	6	61.76	3113849	G. lovleyi	transposon	
Daud0405		PF00665 Integrase, catalytic core	PG4	1328	1329	55	99.59	1367596	D. audaxviator	transposon, frameshift?	hom. Daud1862
Daud0405ds		PF00665 Integrase, catalytic core	PG4	959	1389	46	99.53	1366122	D. audaxviator	transposon, frameshift?	hom. Daud1862
Daud0426		PF01909 DNA polymerase, beta-like region, PF05168 nucleotidyltransferase/HEPN domain protein	PG5	237	858	25	60.27	1081965	C. hydrogenoformans	short	
Daud0427		COG1315 Predicted polymerase	PG5	276	1101	21	37.97	1361526	P. thermopropionicum	short	note: Daud0428 is a transposon

Daud0516		hypothetical protein	PG6	216	N/A	N/A	N/A	N/A	ORFan		
Daud0517		hypothetical protein	PG6	96	N/A	N/A	N/A	N/A	ORFan		
Daud0518		COG2856 Predicted Zn peptidase	PG6	327	792	38	32.69	2536490	A. metalliredigenes	short, split with Daud0516 and Daud0517?	note: Daud0519 phage integrase XerD
Daud0523		PF01909 DNA polymerase, beta-like region		180	315	46	69.39	1112540	M. thermoacetica	short	
Daud0532		COG1682 ABC-2 transporter		225	777	28	56.76	1360468	P. thermopropionicum	short	
Daud0577		COG778 Nitroreductase	PG7	229	591	24	68.09	1367459	D. audaxviator	short	hom. Daud1720
Daud0577ds		COG778 Nitroreductase	PG7	97	591	14	78.57	1367459	D. audaxviator	short	hom. Daud1720
Daud0578		COG1680 Beta-lactamase class C (penicillin-binding protein)	PG7	978	1089	62	39.42	2722213	L. blandensis		
Daud0580		PF07454 stage II sporulation protein P	PG7	658	885	38	50	2843607	H. orenii	split	hom. Daud1176
Daud0580ds		PF07454 stage II sporulation protein P	PG7	431	1128	11	44.19	239375	T. tengcongensis	split	hom. Daud1176
Daud0581		hypothetical protein	PG7	626	N/A	N/A	N/A	N/A	ORFan		
Daud0582		PF00665 Integrase, catalytic core	PG7	426	1389	35	83.23	1366122	D. audaxviator	transposon, short	hom. Daud0321
Daud0583		COG1592 Rubrerythrin	PG7	456	537	56	73.27	1366325	D. audaxviator	short, split	hom. Daud0543, RUBRERYTHRIN
Daud0583ds		COG1592 Rubrerythrin	PG7	122	594	13	77.78	2266818	C. thermocellum	short, split	hom. Daud0543, RUBRERYTHRIN
Daud0586		PF06782 protein of unknown function UPF0236	PG7	186	1401	12	59.32	3176620	T. carboxydivorans	likely transposon, short	note: Daud0587 is a transposon
Daud0586ds		PF06782 protein of unknown function UPF0236	PG7	528	315	51	75.93	1367195	D. audaxviator	likely transposon, short	hom. Daud1447

Daud0684		COG4974 Site-specific recombinase XerD (phage integrase)	PG8	210	861	20	81.36	1366538	D. audaxviator	transposon, short	hom. Daud0775
Daud0685		hypothetical protein	PG8	249	1044	18	78.12	1360492	P. thermopropionicum	short	
Daud0694us	rsbW2	COG2172 Anti-sigma regulatory factor (Ser/Thr protein kinase)		384	435	38	60.71	238847	T. tengcongensis	interrupted	
Daud0694		PF01609 Transposase, IS4-like		2163	1395	100	100	1367546	D. audaxviator	transposon	Daud0695
Daud0694ds	rsbW2	COG2172 Anti-sigma regulatory factor (Ser/Thr protein kinase)		405	435	51	58.97	238847	T. tengcongensis	interrupted	
Daud0699		hypothetical protein		506	3315	9	38.46	2910353	P. aeruginosa	short	
Daud0732		hypothetical protein	PG9	168	321	43	53.19	1114632	N. pharaonis	short	
Daud0733	mecR1	COG4219 Antirepressor regulating drug resistance, predicted signal transduction	PG9	786	2589	28	71.14	238873	T. tengcongensis	short, but not uncommonly so	
Daud0737		PF01609 Transposase, IS4-like	PG9	195	1395	12	78.57	1367546	D. audaxviator	transposon, short	hom. Daud1809
Daud0739us		COG0535 Predicted Fe-S oxidoreductases	PG9	516	1014	29	55.67	182187	P. aerophilum	split	
Daud0739		COG0535 Predicted Fe-S oxidoreductases	PG9	958	1014	63	38.43	182187	P. aerophilum	split	
Daud0741us		COG3328 Transposase and inactivated derivatives (mutator type)	PG9	1180	1227	13	62.26	3177990	T. carboxydivorans	transposon	
Daud0741		COG3328 Transposase and inactivated derivatives (mutator type)	PG9	652	1227	23	63.54	3177990	T. carboxydivorans	transposon	
Daud0756		COG2801 Transposase and inactivated derivatives		336	858	35	52.94	2089653	A. avenae	transposon, short	
Daud0761		COG2405 Predicted nucleic acid-binding protein,		444	471	78	36.51	3439	A. pernix	broken by transposon?	note: Daud0762 is a transposon

		contains PIN domain									
Daud0765	dnaC	COG1484 DNA replication protein		375	483	46	53.33	2267629	C. thermocellum	short?	
Daud0782		COG3436 Transposase and inactivated derivatives, IS66		240	1572	15	77.5	1366475	D. audaxviator	transposon, short	hom. Daud0704 ; note: duplicate Crp (Daud0783) downstream, followed by transposon (Daud0784)
Daud0785		COG2801 Transposase and inactivated derivatives		204	873	22	72.31	1113213	M. thermoacetica	transposon, short	piece of Daud0786
Daud0791	hdrB	COG2048 Heterodisulfide reductase, subunit B		402	1494	17	43.53	2954694	C. ferrooxidans	short	end of SR operon SR5; note: Daud0792 is a transposon
Daud0798		hypothetical protein		668	672	72	46.58	1780122	Mesorhizobium sp. BNC1	split	
Daud0798ds		hypothetical protein		1438	672	25	46.43	1780122	Mesorhizobium sp. BNC1	split	
Daud0805us		hypothetical protein	PG10	381	198	72	47.92	2429490	M. tuberculosis		
Daud0805		COG4584 Transposase and inactivated derivatives	PG10	402	2085	9	43.94	2429489	M. tuberculosis	transposon, short	
Daud0806		hypothetical protein	PG10	336	1932	17	49.09	3047741	R. castenholzii	short	
Daud0811	nusG	COG250 Transcription antiterminator		420	555	85	31.01	940811	T. thermophilus		
Daud0834		COG3481 Predicted HD-superfamily hydrolase		198	942	20	57.81	1360199	P. thermopropionicum	short	
Daud0850		COG296 1,4-alpha-glucan branching enzyme		192	1929	9	57.89	1908055	S. fumaroxidans	short	note: piece of Daud0849?
Daud0959		COG1943, PF07605 Transposase and inactivated derivatives		203	837	20	39.29	2533682	A. metalliredigenes	transposon, short	note: Daud0958 is a transposon
Daud1097		COG2768 Uncharacterized		381	1113	28	60.95	1360090	P.	short	

		Fe-S center protein							thermopropionicum		
Daud1319		PF07670 Nucleoside recognition		303	516	44	50.65	2808405	D. reducens	short	
Daud1332	hmsR	COG1215 Glycosyltransferases, probably involved in cell wall biogenesis		171	600	19	66.67	2810127	D. reducens	short	
Daud1409		COG1872 Uncharacterized conserved protein		300	288	70	41.18	1361359	P. thermopropionicum		
Daud1417	pstA	COG581 ABC-type phosphate transport system, permease component		252	858	24	53.52	1081032	C. hydrogenoformans	short	note: Daud1416 is pstB, Daud1958 is pstA
Daud1448		COG675 Transposase and inactivated derivatives, TIGR01766 IS605 OrfB family		1314	1383	95	96.57	1366484	D. audaxviator	transposon	hom. Daud0713
Daud1513		PF07690 Major facilitator superfamily MFS_1 (transporter)	PG11	168	1194	14	50.91	22235	A. fulgidus	short	
Daud1516		hypothetical protein	PG11	613	528	24	53.49	1360381	P. thermopropionicum		
Daud1520		hypothetical protein	PG11	268	273	58	79.25	1111285	M. thermoacetica	split	
Daud1520ds		hypothetical protein	PG11	150	273	33	58.06	1111285	M. thermoacetica	split	
Daud1545us		COG474 Cation transport ATPase		533	2667	6	39.29	2007174	T. pendens	short	
Daud1545		COG1376, PF03734 Uncharacterized protein, ErfK family		410	690	25	49.15	120237	C. acetobutylicum		
Daud1545ds		COG1376, PF03734 Uncharacterized protein, ErfK family		190	705	13	67.74	1868827	L. casei		
Daud1646us	hybB	COG378, TIGR00073 Hydrogenase accessory protein HypB		116	672	13	73.33	1367390	D. audaxviator	short	hom. Daud1648

Daud1646		COG2189 Adenine specific DNA methylase Mod		213	3087	6	72.13	3047518	R. castenholzii	short	hom. Daud0200
Daud1676		COG3593 Predicted ATP-dependent endonuclease of the OLD family		477	2013	23	36.59	822883	B. clausii	short	
Daud1691		COG438 Glycosyltransferase		277	1041	15	54.9	3285813	A. ferrooxidans	short, split	note: Daud1692 same class, perhaps extension
Daud1691ds		COG438 Glycosyltransferase		115	1056	8	71.43	1105751	M. barkeri	short, split	note: Daud1692 same class, perhaps extension
Daud1711		TIGR01766 transposase, IS605 OrfB family		1266	1704	79	21.62	1865068	L. delbrueckii	transposon	
Daud1711ds		COG675, TIGR01766 transposase, IS605 OrfB family		862	1146	12	97.83	1366600	D. audaxviator	transposon	hom. Daud0845
Daud1729		COG3437 Response regulator containing a CheY-like receiver domain and an HD-GYP domain		392	1044	14	52.08	2259701	A. vinelandii	short	note: Daud1730 is a transposon
Daud1734	cheC / fliN	COG1776 Chemotaxis protein CheC, TIGR02480 flagellar motor switch protein FliN		183	1269	12	78	1360147	P. thermopropionicum	short	note: Daud1735 is a transposon
Daud1826		putative restriction endonuclease, CAS?	PG12	174	837	20	73.68	1359873	P. thermopropionicum	short	hom. Daud1806 (putative restriction endonuclease in CAS operon)
Daud1828		COG348 Polyferredoxin	PG12	231	1065	17	45.31	2405452	P. phaeoclathratiforme	short	
Daud1830	arsA	COG3, PF02374 Anion-transporting ATPase involved in chromosome partitioning	PG12	498	1194	41	62.8	1366114	D. audaxviator	short	hom. Daud0312

Daud1925		COG501 Zn-dependent protease with chaperone function		240	882	14	72.09	1360239	P. thermopropionicum	short	
Daud1948		COG2217 Cation (heavy metal) transport ATPase		282	2316	8	47.54	3051778	S. proteamaculans	short	
Daud2022		COG5577 Spore coat protein Coat F		225	573	32	47.54	1112997	M. thermoacetica	short	
Daud2074	porB	COG1013 Pyruvate:ferredoxin oxidoreductase and related, beta subunit		779	777	55	68.75	1359463	P. thermopropionicum	split	hom. Daud1607
Daud2074ds	porB	COG1013 Pyruvate:ferredoxin oxidoreductase and related, beta subunit		416	777	42	81.65	1359463	P. thermopropionicum	split	hom. Daud1607
Daud2109		PF01909 DNA polymerase, beta-like region	PG13	386	354	38	40	3048785	R. castenholzii	split	
Daud2110		PF05168 HEPN	PG13	183	432	25	44.4	3047571	R. castenholzii	split	

V. DATA AVAILABILITY

The genome sequence reported in this study has been deposited in GenBank under accession number CP000860. The metagenomic data is available from the Joint Genome Institute (http://www.jgi.doe.gov/) under project number 4000602. The annotated *D. audaxviator* genome is accessible *via* MicrobesOnline (http://www.microbesonline.org). The clone library sequences have been submitted to GenBank with accession numbers EU730965 - EU731008. The traces from the reads for the clone library sequences have been submitted to the NCBI trace archive and may be accessed by searching for 'CENTER_NAME = "JGI" and SEQ_LIB_ID = "SGNY"' or 'CENTER_NAME = "JGI" and SEQ_LIB_ID = "SGNX"'.

VI. AUTHOR CONTRIBUTIONS

LHL, GS, and GW collected the "Massive filter" sample used for the environmental genomics. LHL collected microscopy sample #1. DPM collected microscopy sample #2. DEC and FJB extracted the DNA from the filter. AL and SRL sequenced and assembled the *D. audaxviator* genome. DC, EJA, and APA performed the annotation and analysis of the *D. audaxviator* genome. DC and PSD analyzed the reads present in the metagenome. ELB, TZD, GLA performed the PhyloChip analysis. ELB, TZD, GLA, and DC performed the 16S rRNA gene analysis. GS and GW performed the electron microscopy. DPM and Jim Bruckner performed the DAPI stain fluorescence microscopy. ELB performed the 16S CARD-FISH microscopy. TCO performed the chemical speciation and thermodynamic calculations. TCH and PR coordinated the sequencing. FJB postulated environmental sequencing to potentially produce a closed genome sequence. TCO, DC, APA, FJB, TCH, GLA, GS, and LMP guided the project. DC and TCO wrote the manuscript, with significant contributions from EJA, ELB, PSD, LHL, DPM, LMP, FJB, and APA, as well as input from all authors.

VII. REFERENCES

1. R. G. Murray, E. Stackebrandt, *Int J Syst Bacteriol* 45, 186 (1995).
2. D. P. Moser *et al.*, *Appl Environ Microbiol* 71, 8773 (2005).
3. L. H. Lin *et al.*, *Science* 314, 479 (2006).
4. K. Takai, D. P. Moser, M. DeFlaun, T. C. Onstott, J. K. Fredrickson, *Appl Environ Microbiol* 67, 5750 (2001).
5. T. M. Gihring *et al.*, *Geomicrobiology Journal* 23, 415 (2006).
6. M. F. Deflaun *et al.*, *Syst Appl Microbiol* 30, 152 (2007).
7. A. Bonin, Portland State University (2005).
8. T. M. Gihring, J. K. Fredrickson, in *The 103rd General Meeting of the American Society for Microbiology (ASM)*. (Washington, D.C., 2003).
9. T. L. Kieft *et al.*, *Appl Environ Microbiol* 65, 1214 (1999).
10. K. Takai *et al.*, *Int J Syst Evol Microbiol* 51, 1245 (2001).
11. G. Omar, T. C. Onstott, J. Hoek, *Geofluids* 3, 69 (2003).
12. L. H. Lin *et al.*, *Geochemistry Geophysics Geosystems* 6, 10.1029/2004GC000907 (2005).
13. L. Lefticariu, L. M. Pratt, E. M. Ripley, *Geochimica. Cosmochim. Acta* 70, 4889 (2006).
14. P. Chomczynski, K. Mackey, R. Drews, W. Wilfinger, *Biotechniques* 22, 550 (1997).

15. F. Sanger, S. Nicklen, A. R. Coulson, *Proc Natl Acad Sci U S A* 74, 5463 (1977).
16. M. Margulies *et al.*, *Nature* 437, 376 (2005).
17. B. Ewing, P. Green, *Genome Res* 8, 186 (1998).
18. B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res* 8, 175 (1998).
19. E. J. Alm *et al.*, *Genome Res* 15, 1015 (2005).
20. J. H. Badger, G. J. Olsen, *Mol Biol Evol* 16, 512 (1999).
21. A. L. Delcher, D. Harmon, S. Kasif, O. White, S. L. Salzberg, *Nucleic Acids Res* 27, 4636 (1999).
22. T. M. Lowe, S. R. Eddy, *Nucleic Acids Res* 25, 955 (1997).
23. S. F. Altschul *et al.*, *Nucleic Acids Res* 25, 3389 (1997).
24. E. Camon *et al.*, *Genome Res* 13, 662 (2003).
25. R. D. Finn *et al.*, *Nucleic Acids Res* 34, D247 (2006).
26. D. H. Haft, J. D. Selengut, O. White, *Nucleic Acids Res* 31, 371 (2003).
27. R. L. Tatusov *et al.*, *BMC Bioinformatics* 4, 41 (2003).
28. M. N. Price, K. H. Huang, E. J. Alm, A. P. Arkin, *Nucleic Acids Res* 33, 880 (2005).
29. W. Ludwig *et al.*, *Nucleic Acids Res* 32, 1363 (2004).
30. P. Hugenholtz, G. W. Tyson, L. L. Blackall, *Methods Mol Biol* 179, 29 (2002).
31. T. Z. DeSantis *et al.*, *Appl Environ Microbiol* 72, 5069 (2006).
32. R. Sekar *et al.*, *Appl Environ Microbiol* 69, 2928 (May, 2003).
33. E. L. Brodie *et al.*, *Appl Environ Microbiol* 72, 6288 (2006).
34. E. L. Brodie *et al.*, *Proc Natl Acad Sci U S A* 104, 299 (2007).
35. T. Z. DeSantis *et al.*, *Microbial Ecology* 53, 371 (2007).
36. T. Z. DeSantis, Jr. *et al.*, *Nucleic Acids Res* 34, W394 (2006).
37. J. Felsenstein, *Cladistics* 5, 3 (1989).
38. D. J. Lane, in *Nucleic Acid Techniques in Bacterial Systematics* E. Stackebrandt, M. Goodfellow, Eds. (Wiley, New York, 1991), vol. 1, pp. 115-175.
39. P. D. Schloss, J. Handelsman, *Microbiol Mol Biol Rev* 68, 686 (2004).
40. A. E. Magurran, *Ecological diversity and its measurement.* (Princeton University Press, Princeton, N.J., 1988).
41. A. Chao, *Scand. J. Stat.* 11, 265 (1984).
42. P. D. Schloss, J. Handelsman, *Appl Environ Microbiol* 71, 1501 (2005).
43. I. Letunic *et al.*, *Nucleic Acids Res* 34, D257 (2006).
44. D. Wilson, M. Madera, C. Vogel, C. Chothia, J. Gough, *Nucleic Acids Res* 35, D308 (2007).

45. F. D. Ciccarelli *et al.*, *Science* 311, 1283 (2006).
46. R. C. Edgar, *Nucleic Acids Res* 32, 1792 (2004).
47. S. Guindon, O. Gascuel, *Syst Biol* 52, 696 (2003).
48. D. T. Jones, W. R. Taylor, J. M. Thornton, *Comput Appl Biosci* 8, 275 (1992).
49. H. Imachi *et al.*, *Int J Syst Evol Microbiol* 52, 1729 (2002).
50. B. M. Tebo, A. Y. Obraztsova, *FEMS Microbiology Letters* 162, 193 (1998).
51. H. L. Drake, S. L. Daniel, *Research in Microbiology* 155, 869 (2005).
52. M. Wu *et al.*, *PLoS Genet* 1, e65 (2005).
53. M. Hasegawa, H. Kishino, T. Yano, *J Mol Evol* 22, 160 (1985).
54. G. W. Tyson *et al.*, *Nature* 428, 37 (2004).
55. E. Puerta-Fernandez, J. E. Barrick, A. Roth, R. R. Breaker, *Proc Natl Acad Sci U S A* 103, 19490 (2006).
56. K. S. Makarova, Y. I. Wolf, E. V. Koonin, *Trends Genet* 19, 172 (2003).
57. P. Forterre, *Trends Genet* 18, 236 (2002).
58. H. Atomi, R. Matsumi, T. Imanaka, *J Bacteriol* 186, 4829 (2004).
59. M. P. Mehta, J. A. Baross, *Science* 314, 1783 (2006).
60. D. H. Haft, J. Selengut, E. F. Mongodin, K. E. Nelson, *PLoS Comput Biol* 1, e60 (2005).
61. R. Barrangou *et al.*, *Science* 315, 1709 (2007).
62. K. S. Makarova, N. V. Grishin, S. A. Shabalina, Y. I. Wolf, E. V. Koonin, *Biol Direct* 1, 7 (2006).
63. M. Mussmann *et al.*, *J Bacteriol* 187, 7126 (2005).
64. D. P. Brown, K. B. Idler, L. Katz, *J Bacteriol* 172, 1877 (1990).
65. V. Zverlov *et al.*, *J Bacteriol* 187, 2203 (2005).
66. D. A. Grahame, E. DeMoll, *Biochemistry* 34, 4617 (1995).
67. Y. R. Dai *et al.*, *Arch Microbiol* 169, 525 (1998).
68. G. Wanger, T. C. Onstott, G. Southam, *Geomicrobiology Journal* 23, 443 (2006).
69. K. S. Makarova, E. V. Koonin, *FEMS Microbiol Lett* 227, 17 (2003).