Genomic comparisons of a bacterial lineage that inhabits both marine and terrestrial deep subsurface systems

Sean P Jungbluth Corresp., [1, 2] , **Tijana Glavina del Rio** [3] , **Susannah G Tringe** [3] , **Ramunas Stepanauskas** [4] , **Michael S Rappé** Corresp. [5]

[1] Department of Oceanography, University of Hawaii at Manoa, Honolulu, HI, United States
[2] Center for Dark Energy Biosphere Investigations, University of Southern California, Los Angeles, CA, United States
[3] DOE Joint Genome Institute, Walnut Creek, CA, United States
[4] Single Cell Genomics Center, Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, United States
[5] Hawaii Institute of Marine Biology, University of Hawaii at Manoa, Kaneohe, HI, United States

Corresponding Authors: Sean P Jungbluth, Michael S Rappé
Email address: jungbluth.sean@gmail.com, rappe@hawaii.edu

It is generally accepted that diverse, poorly characterized microorganisms reside deep within Earth's crust. One such lineage of deep subsurface-dwelling Bacteria is an uncultivated member of the *Firmicutes* phylum that can dominate molecular surveys from both marine and continental rock fracture fluids, sometimes forming the sole member of a single-species microbiome. Here, we reconstructed a genome from basalt-hosted fluids of the deep subseafloor along the eastern Juan de Fuca Ridge flank and used a phylogenomic analysis to show that, despite vast differences in geographic origin and habitat, it forms a monophyletic clade with the terrestrial deep subsurface genome of "*Candidatus* Desulforudis audaxviator" MP104C. While a limited number of differences were observed between the marine genome of "*Candidatus* Desulfopertinax cowenii" modA32 and its terrestrial relative that may be of potential adaptive importance, here it is revealed that the two are remarkably similar thermophiles possessing the genetic capacity for motility, sporulation, hydrogenotrophy, chemoorganotrophy, dissimilatory sulfate reduction, and the ability to fix inorganic carbon via the Wood-Ljungdahl pathway for chemoautotrophic growth. Our results provide insights into the genetic repertoire within marine and terrestrial members of a bacterial lineage that is widespread in the global deep subsurface biosphere, and provides a natural means to investigate adaptations specific to these two environments.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2592v1 | CC BY 4.0 Open Access | rec: 14 Nov 2016, publ:

Genomic comparisons of a bacterial lineage that inhabits both marine and terrestrial deep subsurface systems

Sean P. Jungbluth[1,2,*], Tijana Glavina del Rio[3], Susannah G. Tringe[3], Ramunas Stepanauskas[4], and Michael S. Rappé[5]*

[1]Department of Oceanography, SOEST, University of Hawaii, Honolulu, HI

[2]Center for Dark Energy Biosphere Investigations, University of Southern California, Los Angeles, CA

[3]DOE Joint Genome Institute, Walnut Creek, CA

[4]Single Cell Genomics Center, Bigelow Laboratory for Ocean Sciences, East Boothbay, ME

[5]Hawaii Institute of Marine Biology, SOEST, University of Hawaii, Kaneohe, HI

*Corresponding authors: jungbluth.sean@gmail.com or rappe@hawaii.edu

Running title: Deep subsurface Firmicutes

KEY WORDS: deep subsurface · microorganisms · Firmicutes · Juan de Fuca Ridge · chemoautotrophy · basement biosphere · sulfate reduction · sporulation

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2592v1 | CC BY 4.0 Open Access | rec: 14 Nov 2016, publ:

Abstract

It is generally accepted that diverse, poorly characterized microorganisms reside deep within Earth's crust. One such lineage of deep subsurface-dwelling Bacteria is an uncultivated member of the *Firmicutes* phylum that can dominate molecular surveys from both marine and continental rock fracture fluids, sometimes forming the sole member of a single-species microbiome. Here, we reconstructed a genome from basalt-hosted fluids of the deep subseafloor along the eastern Juan de Fuca Ridge flank and used a phylogenomic analysis to show that, despite vast differences in geographic origin and habitat, it forms a monophyletic clade with the terrestrial deep subsurface genome of "*Candidatus* Desulforudis audaxviator" MP104C. While a limited number of differences were observed between the marine genome of "*Candidatus* Desulfopertinax cowenii" modA32 and its terrestrial relative that may be of potential adaptive importance, here it is revealed that the two are remarkably similar thermophiles possessing the genetic capacity for motility, sporulation, hydrogenotrophy, chemoorganotrophy, dissimilatory sulfate reduction, and the ability to fix inorganic carbon via the Wood-Ljungdahl pathway for chemoautotrophic growth. Our results provide insights into the genetic repertoire within marine and terrestrial members of a bacterial lineage that is widespread in the global deep subsurface biosphere, and provides a natural means to investigate adaptations specific to these two environments.

Introduction

Recent progress in understanding the nature of microbial life inhabiting the sediment-buried oceanic crust has been made through the use of ocean drilling program borehole observatories as platforms to successfully sample fluids that percolate through the subseafloor basement (Wheat et al., 2011). In 2003, a pioneering study by Cowen and colleagues used a passive-flow device to collect microbial biomass from fluids emanating out of an over-pressured borehole that originated from deep with the igneous basement of the eastern flank of the Juan de Fuca Ridge in the Northeast Pacific Ocean (Cowen et al., 2003). Ribosomal RNA (rRNA) gene cloning and sequencing from the crustal fluids led to the first confirmation of microbial life in the deep marine igneous basement and revealed the presence of diverse Bacteria and Archaea. Discovered in this initial survey was an abundant, uniquely branching lineage within the bacterial phylum *Firmicutes* that was only distantly related to its closest known relative at the time, a thermophilic nitrate-reducing chemoautotroph isolated from a terrestrial volcanic hot spring, *Ammonifex degensii* (Huber et al., 1996).

Subsequent molecular surveys within both the terrestrial and marine deep subsurface revealed the presence of microorganisms related to the original marine firmicutes lineage (Lin et al., 2006; Jungbluth et al., 2013). In the deep subseafloor basement, this lineage has been recovered in high abundance (up to nearly 40%) from basaltic crustal fluids collected from a borehole nearby the initial location sampled ten years previously by Cowen and colleagues, as well as from multiple boreholes spaced up to ~70 km apart in the same region of the Northeast Pacific Ocean seafloor (Jungbluth et al., 2013; Jungbluth et al., 2014). In a surprising discovery, a single ecotype closely related to this firmicutes lineage was discovered in deep terrestrial subsurface fracture water of South Africa and found to be widespread (Magnabosco et al., 2014), where it sometimes made up an extremely high proportion of microorganisms *in situ* (Chivian et al., 2008). This lineage has since been found in other terrestrial habitats such as the Fennescandian Shield in Finland (Itävaara et al., 2011), a saline geothermal aquifer in Germany (Lerm et al., 2013), and an alkaline aquifer in Portugal (Tiago & Veríssimo, 2013). Based on ribosomal RNA sequence analyses, most of the terrestrial and marine lineages form a monophyletic clade of predominantly subsurface origin but do not partition into subclades of exclusively terrestrial and marine origin, suggesting that there may have been multiple transitions between the terrestrial and marine deep subsurface environments (Jungbluth et al., 2013).

In 2008, Chivian and colleagues reconstructed the first complete genome from a terrestrial member of this firmicutes lineage, provisionally named "*Candidatus* Desulforudis audaxviator" MP104C, via metagenome sequencing of a very low diversity sample from a deep gold mine in South Africa (Chivian et al., 2008). The "*Ca.* D. audaxviator" genome revealed a motile, sporulating, thermophilic chemolithoautroptroph genetically capable of dissimilatory sulfate reduction, hydrogenotrophy, nitrogen fixation, and carbon fixation via the reductive acetyl-coenzyme A (Wood-Ljungdahl) pathway (Chivian et al., 2008). Thus, "*Ca.* D. audaxviator" appears well suited for an independent lifestyle within the deep continental subsurface environment. "*Ca.* D. audaxviator" and close relatives have continued to be recovered in subsequent metagenomes sequenced from the South African subsurface (Lau et al., 2014). Recently, five flow-sorted and single amplified genomes related to "*Ca.* D. audaxviator" were sequenced from the terrestrial subsurface of South Africa, revealing significant genotypic variation with the terrestrial genomes and providing evidence for horizontal gene transfer and viral infection in the terrestrial subsurface environment (Labonté et al., 2015). To date, knowledge regarding marine members of this deep subsurface firmicutes lineage has been limited to phylogenetic (rRNA) and functional (dsr) gene surveys (Jungbluth et al., 2013; Robador et al., 2015).

In this study, we sought to improve understanding of the functional and evolutionary attributes of microorganisms inhabiting the deep subseafloor basement by sequencing the environmental DNA from two basement fluid samples from Juan de Fuca Ridge flank boreholes U1362A and U1362B, generating the first metagenomes from this environment. Binning of the resulting sequence data led to the reconstruction of a nearly complete genome closely related to "*Ca.* D. audaxviator". This genome has allowed us to compare the functional composition of members of a microbial lineage that spans the terrestrial and marine deep subsurface, investigate its evolutionary history, and determine its prevalence within a globally-distributed assemblage of metagenomes.

Materials and Methods

Borehole fluid sampling

The methods used to collect samples during R/V Atlantis cruise ATL18_07 (28 June 2011 – 14 July 2011) are described elsewhere (Jungbluth et al., 2016). Briefly, basement crustal fluids were collected from CORK observatories located in 3.5 million-year-old ocean crust east of the Juan de Fuca spreading center in the Northeast Pacific Ocean. Basement fluids were collected from the polytetrafluoroethylene (PTFE) lined fluid delivery lines associated with the lateral CORKs (L-CORKs) at boreholes U1362A (47°45.6628'N, 127°45.6720'W) and U1362B (47°45.4997'N, 127°45.7312'W). These lines extend to 200 and 30 meters below the sediment-basement interface, respectively. Fluids were filtered *in situ* via a mobile pumping system

(Cowen et al., 2012) through Steripak-GP20 filter cartridges (Millipore, Billerica, MA, USA) containing 0.22 µm pore-sized polyethersulfone membranes. A filtration rate of 1 liter min^{-1} was calculated from laboratory tests, indicating that ~124 liters (U1362A) and ~70 liters (U1362B) of deep subsurface crustal fluids were filtered.

Metagenomic DNA sequencing

Borehole fluid nucleic acids were extracted using a modified phenol/chloroform lysis and purification method and is described in detail elsewhere (Jungbluth et al., 2016) (samples SSF21-22 and SSF23-24). Library preparation and sequencing was conducted by the Joint Genome Institute as part of the Community Sequencing Program. A total of 100 ng (U1362A) or 5 ng (U1362B) of DNA was sheared using a focused-ultrasonicator (Covaris, Woburn, MA, USA). The sheared DNA fragments were size selected using SPRI beads (Beckman Coulter, Brea, CA, USA). The selected fragments from U1362A were then end-repaired, A-tailed, and ligated of Illumina compatible adapters (Integrated DNA Technologies, Coralville, IA, USA) using KAPA-Illumina library creation kit (KAPA Biosystems, Wilmington, MA, USA). The selected fragments from U1362B were treated with end repair, ligation of adapters and 9 cycle of PCR on the Mondrian SP+ Workstations (Nugen, San Carlos, CA, USA) using the Ovation SP+ Ultralow DR Multiplex System kit (Nugen).

The library was quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a LightCycler 480 real-time PCR instrument (Roche, Basel, Switzerland). The quantified U1362A library was then prepared for sequencing on the HiSeq sequencing platform (Illumina, San Diego, CA, USA) utilizing a TruSeq paired-end cluster kit, v3, and Illumina's cBot instrument to generate clustered flowcell for sequencing. The U1362B library was prepared for sequencing in the same manner except the library was multiplexed with another sample library for a pool of 2 prior to use of the TruSeq kit. Sequencing of the flowcell was performed on the Illumina HiSeq2000 sequencer using a TruSeq SBS sequencing kit 200 cycles, v3, following a 2x150 indexed run recipe.

Insert size analysis was performed at JGI using bbmerge to pair overlapping reads and, with sufficient coverage, non-overlapping reads using gapped kmers. The "percentage reads joined" was calculated by (number of joined reads/total number of reads × 100). Raw reads were used for the insert size calculation (no trimming or filtering). Insert size statistics for the U1362A metagenome were: 68.342% reads joined, 216.60 bp average read length, 37.40 bp standard deviation read length, and 215.00 bp mode read length. Insert size statistics for the U1362B metagenome were: 50.40% reads joined, 210.80 bp average read length, 39.70 bp standard deviation read length, and 196.00 mode read length.

Metagenome quality control, read trimming and assembly

Assembly was performed by the JGI; corresponding JGI assembly identifications are 1020465 (U1362A) and 1020462 (U1362B). Raw Illumina metagenomic reads were screened against Illumina artifacts with a sliding window with a kmer size of 28, step size of 1. Screened read portions were trimmed from both ends using a minimum quality cutoff of 3, reads with 3 or more 'Ns' or with average quality score of less than Q20 were removed. In addition, reads with a minimum sequence length of <50 bp were removed. Trimmed, screened, paired-end Illumina reads were assembled using SOAPdenovo v1.05 (Luo et al., 2012) with default settings (options: -K 81, -p 32, -R, -d 1) and a range of Kmers (81, 85, 89, 93, 97, 101). Contigs generated by each assembly (six contig sets in total) were de-replicated using JGI in-house perl scripts. Contigs were then sorted into two pools based on length. Contigs smaller than 1800 bp were assembled using Newbler (Life Technologies, Carlsbad, CA, USA) in an attempt to generate larger contigs (flags: -tr, -rip, -mi 98, -ml 80). All assembled contigs larger than 1800 bp were combined with the contigs generated from the final Newbler run using minimus2 (flags: -D MINID=98 –D OVERLAP=80) (Treangen et al., 2011). JGI-reported read depths available in IMG were estimated based on read mapping with JGI in-house mapping programs.

Gene prediction and annotation

All aspects of metagenome annotation performed at JGI can be found at img.jgi.doe.gov/m/doc/MetagenomeAnnotationSOP.pdf (Huntemann et al., 2016). Briefly, metagenome sequences were preprocessed to resolve ambiguities, trim low-quality regions and trailing 'N's using LUCY (Chou & Holmes, 2001), masked for low-complexity regions using DUST (Morgulis et al., 2006), and dereplicated (95% threshold). Genes were predicted in the following order: CRISPRs, non-coding RNA genes, protein-coding genes. CRISPR elements were identified by concatenating the results from the programs CRT (Bland et al., 2007) and PILER-CR (Edgar, 2007). tRNAs were predicted using tRNA scan SE-1.23 (Lowe & Eddy, 1997) three times using each of the domains of life (Bacteria, Archaea, Eukaryota) as the parameter required; the best scoring predictions were selected. Fragmented tRNAs were identified by comparison to a database of tRNAs identified in isolate genomes. Ribosomal RNA genes were predicted using JGI-developed rRNA models (SPARTAN: SPecific & Accurate rRNA and tRNA ANnotation). Protein-coding genes were identified using a majority rule-based decision schema using four different gene callings tools: prokaryotic GeneMark (hmm version 2.8) (Lukashin & Borodovsky, 1998) Metagene Annotator v1.0 (Noguchi, Park & Takagi, 2006), Prodigal v2.5 (Hyatt et al., 2012) and FragGeneScan v1.16 (Rho, Tang & Ye, 2010). When there was no clear decision, the selection was based on preference order of gene callers determined by JGI-based runs on simulated metagenomic datasets [GeneMark > Prodigal > Metagenome > FragGeneScan].

Predicted CDSs were translated and associated with Pfams COGs, KO terms, EC numbers, and phylogeny. Genes were associated with Pfam-A using hmmsearch (Durbin et al., 1998). Genes were associated with COGs by comparing protein sequences with the database of PSSMs for COGs downloaded from NCBI; rpsblast v2.26 (Marchler-Bauer et al., 2003) was used to find hits. Assignments of KO terms, EC numbers, and phylogeny were made using similarity searches to reference databases constructed by starting with the set of all non-redundant sequences taken from public genomes in IMG. Sequences from the KEGG database that were not present in IMG were added and all data was merged to related gene IDs to taxa, KO terms, and EC numbers. USEARCH (Edgar, 2010) was used to compare predicted protein-coding genes to genes in this database and the top five hits for each gene were retained. Phylogenetic assignment was based on the top hit only; for assignment of KO terms, the top 5 hits to genes in the KO index were used. A hit resulted in an assignment if there was at least 30% identity and greater than 70% of the query protein sequence or the KO gene sequence were covered by the alignment.

Genomic bin identification and reconstruction

All metagenomic scaffolds greater than 200 basepairs (bp) from U1362A (n=137,672 contigs) and U1362B (n=212,542 contigs) were binned separately with MaxBin v1.4 (Wu et al., 2014) using the 40 marker gene set universal among bacteria and archaea (Wu, Jospin & Eisen, 2013), minimum contig length of 1000 bp, and default parameters. Contig coverage from each metagenome was estimated using the quality control-filtered raw reads as input for mapping using Bowtie2 v2.1.0 (Langmead & Salzberg 2012) via MaxBin. The genomic bins were screened and analyzed for completeness, contamination, and assigned taxonomic identifications using CheckM v1.0.5 (Parks et al., 2015) with default parameters.

Raw quality control-filtered sequence reads from the U1362A and U1362B metagenomes related to "*Ca.* D. audaxviator" were identified by mapping to three sources: (1) a single genomic bin from U1362A related to "*Ca.* D. audaxviator" identified via CheckM (bin A32), (2) the "*Ca.* D. audaxviator" genome, (3) and all "*Ca.* D. audaxviator"-related contigs > 200 bp from the U1362A and U1362B metagenome assemblies generated by the Joint Genome Institute. Mapping was performed independently for the U1362A and U1362B metagenomes using both the bbmap v34.25 (http://sourceforge.net/projects/bbmap/) and Bowtie2 v2.1.0 (Langmead & Salzberg 2012) software packages with default parameters and the paired-end read-mapping feature (Supplementary Table 1). All reads from the U1362A metagenome mapping to any of the three sources (1,785,284 sequences) were assembled using SPAdes v3.5.0 (Bankevich et al., 2012) with options –k: 21,33,55,77, --careful –pe1-12 and default parameters. Contaminating contigs in the assembly were screened and removed using the JGI ProDeGe web portal v2.0 (https://prodege.jgi-psf.org/) on April 10, 2015, using default parameters with the following taxonomy specified: "Bacteria; Firmicutes; Clostridia" (Tennessen et al., 2016). Contigs remaining following the use of ProDeGe comprise the genome bin henceforth named "*Ca.* Desulfopertinax cowenii" modA32 and were screened using CheckM as described above.

Genome annotation and analysis

The modified genome bin resulting from the pipeline described above ("Ca. D. cowenii" modA32) was annotated via the Joint Genome Institute's Integrated Microbial Genomes-Expert Review (IMG-ER) web portal (Markowitz et al., 2014; Huntemann et al., 2015). Annotations in the IMG-ER web portal served as the source of reported genome characteristics and reported genes and their assignment to COGs. Phylogenetically informative marker genes from "Ca. D. cowenii" were identified and extracted using the 'tree' command in CheckM. In CheckM, open reading frames were called using prodigal v2.6.1 (Hyatt et al., 2012) and a set of 43 lineage-specific marker genes, similar to the universal set used by PhyloSift (Darling et al., 2014), were identified and aligned using HMMER v3.1b1 (Eddy, 2011). Initial phylogenetic analysis used pplacer (v1.1.alpha16-1-gf748c91) (Matsen, Kodner & Armbrust, 2010) to place sequences into a CheckM tree/database (version 0.9.7) composed of 2052 finished and 3604 draft genomes (Markowitz et al., 2012).

An alignment 6988 amino acids in length corresponding to the 43 concatenated marker genes from "*Ca.* D. cowenii", "*Ca.* D. audaxviator", other *Firmicutes,* and *Actinobacteria* were used for additional phylogenetic analysis. The concatenated amino acid alignment was used to generate a phylogeny using FastTree v2.1.9 (Price, Dehal & Arkin, 2010) with the WAG amino acid substitution model. The dendogram was visualized using iTOL v3 (Letunic and Bork, 2016).

Average nucleotide identity (ANI) was computed in IMG-ER using pairwise bidirectional best nSimScan hits of genes having 70% or more identity and at least 70% coverage of the shorter gene. The "*Ca.* D. cowenii" → [other genome] values are reported. Protein-coding genes in "*Ca.* D. cowenii" with and without homologs in "*Ca.* D. audaxviator", and vice versa, were identified and percent similarity estimated using the "Phylogenetic Profiler" tool in IMG-ER with default parameters (max e-value: 10e^{-5}; minimum identity: 30%). Average amino acid identity (AAI) was computed for pairs of genomes closely related to "*Ca.* D. cowenii" with an online web tool (http://enve-omics.ce.gatech.edu/aai/) using default parameters. All non-RNA genes at least 100 amino acids in length were used in this analysis. Two-way average amino acid identity scores are reported and the percent shared genes were calculated as follows: 100 × (2 × [number of proteins used for two-way AAI analysis]) / ([total number of amino acids ≥ 100 from genome A] + [total number of amino acids ≥ 100 from genome B]). Estimates of transposase and integrase abundance were derived in IMG using a functional profile of 100 pfams and COG functions selected searching for keywords "transposase" and "integrase".

Genome and scaffold visualizations

Global genome comparisons were visualized in Circos v0.67-5 (Krzywinski et al., 2009). Links between genomic regions of "*Ca.* D. cowenii" and "*Ca.* D. audaxviator" represent best reciprocal BLAST hits, which were generated using the blast_rbh.py script (https://github.com/peterjc/galaxy_blast/tree/master/tools/blast_rbh) with blastn v2.2.29 (Altschul et al., 1990) and default parameters. Links between genomic regions from the single amplified genomes of Labonté et al. (Labonté et al., 2015) represent BLAST hits that were generated using blastn with default parameters and using "*Ca.* D. cowenii" and "*Ca.* D. audaxviator" as reference databases.

Selected scaffold regions were visualized with Easyfig v2.2.2 (Sullivan, Petty & Beatson, 2011). Similarity between regions was assessed using BLAST wrapped within Easyfig using default parameters and task: blastn; minimum hit length: 50; max e-value: 0.001; minimum identity value: 50. In all instances of blast, contigs from "*Ca.* D. cowenii" were used as the query and "*Ca.* D. audaxviator" was used as the reference, with the exception of the single three-scaffold comparison where "*Ca.* D. audaxviator" was used as the query and "*Ca.* D. cowenii" Ga007115_16 used as the reference.

Metagenome fragment recruitment

Quality-filtered raw reads from the U1362A and U1362B metagenomes were mapped to the six scaffolds that make up the "*Ca.* D. cowenii" genome bin and the "*Ca.* D. audaxviator" genome. Recruitment was performed using FR-HIT v0.7.1 (Niu et al., 2011) with default parameters (minimum sequence similarity 75%) and reporting a single best top hit for each read (-r 1).

Analysis of metagenome-derived SSU rRNA genes

Full length SSU rRNA genes from the raw quality-filtered U1362A metagenome reads were assembled using EMIRGE (Miller et al., 2011) with default parameters and -a 20, -i 270, -s 100, -l 150, -j 1.0, --phred33, and using the SILVA SSURef_Nr99 version 119 database that was prepared using the fix_nonstandard_chars.py script supplied on the EMIRGE website (https://github.com/csmiller/EMIRGE). Out of 1951 near full-length SSU rRNA sequences constructed after 67 iterations of EMIRGE, a single sequence related to the "*Ca.* D. audaxviator" lineage was identified through the SILVA online portal (Pruesse, Peplies & Glöckner, 2012). The sequence was aligned using the SINA online aligner and manually curated in ARB (Ludwig et al., 2004). Ambiguous and mis-aligned positions were excluded from further analysis.

A base SSU rRNA gene phylogenetic tree was reconstructed in ARB from 36 sequences and an alignment of 797 nucleotide positions using RAxML v7.72 (Stamatakis, 2006) with default parameters, the GTR+G+I nucleotide substitution model identified via JModelTest v2.1.1 (Darriba et al., 2012), and selecting the best tree from 100 iterations. Bootstrapping was performed in ARB using the RAxML tool with 2000 replicates (Stamatakis, Hoover & Rougemont, 2008). Sequences of short length, including a masked version of the "*Ca.* D. audaxviator" -related SSU rRNA gene found here, were added to the phylogeny using the parsimony insertion tool in ARB and a filter containing 363 nucleotide positions.

Phylogenetic analysis of dsrAB gene sequences

DNA sequences corresponding to dissimilatory sulfite reductase subunits alpha and beta (*dsrAB*) were aligned in ARB using the 'integrated aligners' tool and a previously published database of aligned *dsrAB* sequences (Loy et al., 2009). Additional sequences were identified and included via BLAST search of the non-redundant NCBI database using megablast and blastn with default parameters. Phylogenetic analyses were performed individually for *dsrA* and *dsrB* using RAxML with the GTR model of nucleotide substitution under the gamma- and invariable-models of rate heterogeneity, identified via jModelTest. The tree with the highest negative log-likelihood score was selected from performing 100 iterations using RAxML with default parameters. Phylogenies for the base trees were derived from partial length *dsrA* and *dsrB* alignments (545 and 303 nucleotides, respectively) and bootstrapping was performed in ARB using the RAxML rapid bootstrap analysis algorithm with 2000 bootstraps.

Analysis of global distribution patterns

All protein-coding genes corresponding to the genomes of "*Ca.* D. cowenii" (1782 genes) and "*Ca.* D. audaxviator" (2239 genes) were used to generate a profile against 489 globally-distributed metagenomes from marine subsurface fluids, the terrestrial subsurface, terrestrial hot springs, marine sediments, and seawater (Supplementary Table 2). In IMG-ER, the "Profile & Alignment" tool was used to query assembled metagenomes using genes corresponding to the two genomes, a maximum e-value of 10^{-5}, and a minimum similarity of 70%. The number of gene hits was converted to a relative frequency and the location of hits was visualized in R v3.1.2 (R Core Team, 2015) using latitude and longitude information provided as metadata and the R maps package (version 2.3-10).

Fragment recruitment was subsequently used in effort to discriminate between the distribution of the marine ("*Ca.* D. cowenii' modA32A) and terrestrial (*"Ca.* D. audaxviator") genomes of this *Firmicutes* lineage. Raw reads corresponding to IMG-ER metagenomes with the highest hit frequencies in the profiles generated in IMG, and additional unamplified metagenomes from the marine and terrestrial subsurface available only via NCBI sequence read archive and MG-RAST, were used as references for mapping to the genomes of "*Ca.* D. cowenii" and "*Ca.* D. audaxviator" (Supplementary Table 3). In order to determine a % similarity cutoff that can discriminate between the two targets, the two genomes were cut into non-overlapping 150 bp fragments to simulate the most common sequence read length in current metagenome projects, and mapped back to the intact "*Ca.* D. cowenii" and *"Ca.* D. audaxviator" genomes using FR-HIT with default parameters, restricting matches to the single top best hit. Percent similarities ranging from 70-100% were tested in one percent increments in order to quantify the frequency that the fragmented genomes map to their source genome. A 96% similarity level was ultimately used because it restricted spurious matches (i.e. reads mapping from one genome to the other) to a frequency of ~1% (Supplementary Figure 1). The ratio of reads mapping to "*Ca.* D. cowenii" or "*Ca.* D. audaxviator" was calculated and visualized using Circos.

Sample access and affiliated information

The annotated draft genome of "*Ca.* D. cowenii" modA32 is available via the IMG web portal under Taxon ID number 2615840622 (Gold Analysis Project ID: Ga0071115). The U1362A and U1362B metagenomes are available via the IMG-M web portal under Taxon ID numbers 330002481 and 3300002532, respectively. Gold Analysis Project ID numbers are Ga0004278 (U1362A) and Ga0004277 (U1362B). Sample metadata can be accessed using the BioProject identifier PRJNA269163. The NCBI BioSamples used here are SAMN03166137 (U1362A) and SAMN03166138 (U1362B). Raw sequence data can be accessed using NCBI SRA identifiers SRR3723048 (U1362A) and SRR3732688 (U1362B).

Results and Discussion

Bin identification and refinement

Of 60 and 41 genome bins representing diverse groups of uncultivated bacteria and archaea reconstructed from the U1362A and U1362B metagenomes, respectively, one that comprised a nearly complete genome from U1362A (bin A32) was preliminarily identified as related to "*Ca.* D. audaxviator" by phylogenetic analyses of a set of concatenated single copy marker genes. In order to maximize genome recovery while minimizing potential contamination, contigs within genome bin A32, the "*Ca.* D. audaxviator" genome, and scaffolds related to "*Ca.* D. audaxviator" that were assembled directly from the U1362A and U1362B metagenomes were used as references for mapping raw sequence reads from the U1362A and U1362B metagenomes via several read mapping methods. Sequence mate pairs from the U1362A metagenome that mapped to these templates were pooled and reassembled (Supplementary Table 1). Following subsequent screening and removal of contaminating sequences (Supplementary Table 4), six genomic scaffolds totaling 1,778,734 base pairs (bp) in length were identified that correspond to the draft "*Ca.* D. cowenii" modA32 genome described here (Table 1). The purity of the modified genomic bin was supported by results generated using CheckM (Parks et al., 2015) (Table 2), congruent phylogenetic analyses of concatenated marker genes (Figure 1A) and *dsrB* (Figure 2A) and *dsrA* genes (Supplementary Figure 2), and a high percent of shared genes and gene synteny between the six genomic scaffolds of "*Ca.* D. cowenii" and the "*Ca.* D. audaxviator" genome (Figures 1B and 3A).

The 1.78 Mbp "*Ca.* D. cowenii" modA32 genome is 98-99% complete based on separate analyses of tRNA and other marker gene content specific to the phylum *Firmicutes* (Table 1). A phylogenomic analysis of 43 conserved marker genes confirmed a monophyletic relationship between "*Ca.* D. cowenii" and "*Ca.* D. audaxviator" within the *Firmicutes* (Figure 1A), a relationship that was also supported by analyses of both *dsrA* (Supplementary Figure 2) and *dsrB* genes (Figure 2A). While no small-subunit (SSU) rRNA genes were identified in the "*Ca.* D. cowenii" genome bin, a single full-length SSU rRNA gene related to "*Ca.* D. audaxviator" was reconstructed from raw U1362A metagenome reads. Phylogenetic analyses revealed this gene to form a tight cluster with SSU rRNA genes recovered previously from the deep subseafloor along the Juan de Fuca Ridge flank and, more broadly, a monophyletic lineage with "*Ca.* D. audaxviator" within the phylum *Firmicutes* (Figure 2B). Consistent with previous studies (Jungbluth et al., 2014; Jungbluth et al., 2016), oceanic crustal fluid SSU rRNA gene clones formed at least two independent sub-lineages within this clade (Figure 2B).

Comparative genomics

The genomes of "*Ca.* D. cowenii" and "*Ca.* D. audaxviator" share an average nucleotide identity of 76.9%, which is almost 7% higher than "*Ca.* D. cowenii" shares with the next most similar firmicutes genome, *Desulfovirgula thermocuniculi*. Similarly, the genomes of "*Ca.* D. cowenii" and "*Ca.* D. audaxviator" share an average amino acid identity of 74.2%, a value that is almost 18% higher than "*Ca.* D. cowenii" shares with its next most similar genome, the firmicute *Desulfotomaculum kuznetsovii* DSM 6115 (Figure 1B). A similar result was obtained by quantifying the proportion of genes shared between "*Ca.* D. cowenii" and "*Ca.* D. audaxviator" (73.2%) (Figure 1B).

Compared to the genomes of its closest relatives, the 1.78 Mbp genome harbored by "*Ca.* D. cowenii" is small (Figure 1B). Despite the smaller size of the "*Ca.* D. cowenii" genome compared to the 2.35 Mbp genome of "*Ca.* D. audaxviator", the two share similar coding density (89.8% vs. 87.6%), resulting in 451 fewer genes in "*Ca.* D. cowenii" (1842 vs. 2293) (Table 1). Compared to other firmicutes, the predicted genome size of "*Ca.* D. cowenii" is among the smallest for members of the Class *Clostridia* with an elevated %GC (Figure 4). The smaller genome of "*Ca.* D. cowenii" shares 1514 of its 1782 (85.0%) protein coding genes with "*Ca.* D. audaxviator". Despite the lower gene content overall, "*Ca.* D. cowenii" harbors a similar number of protein coding genes with a predicted function as the genome of "*Ca.* D. audaxviator" (1518 vs. 1587) (Table 1). In addition to a smaller genome and fewer genes, "*Ca.* D. cowenii" also contained fewer pseudogenes (0 vs. 82) and paralogs (137 vs. 265) in comparison to "*Ca.* D. audaxviator" (Table 1), which together suggest some form of streamlining of the "*Ca.* D. cowenii" genome. Compared to "*Ca.* D. audaxviator", the genome of "*Ca.* D. cowenii" contains fewer CRISPR elements, integrases and transposases, and phage-related genes, which suggests lower viral infection and less horizontal gene transfer in the marine lineage.

Extensive gene synteny between "*Ca.* D. cowenii" and "*Ca.* D. audaxviator" was revealed by comparing locations of homologs (Figures 3A and 3B). Aligning the genome of "*Ca.* D. cowenii" with five incomplete (3.6-7.8% complete) single amplified genomes (SAGs) isolated from the terrestrial South Africa subsurface and related to "*Ca.* D. audaxviator" (Labonté et al., 2015) revealed that all of the SAGs were more similar to "*Ca.* D. audaxviator" than "*Ca.* D. cowenii" (Figure 5).

Similarities in functional gene complement

Comparisons of predicted proteins assigned to clusters of orthologous groups (COGs) revealed a markedly similar distribution within the "*Ca.* D. cowenii" and "*Ca.* D. audaxviator" genomes (Figure 3C). A detailed description of these shared features is included in Supplementary Table 5.

The genome of "*Ca.* D. cowenii" reveals a microorganism that is functionally similar to "*Ca.* D. audaxviator": an independent lifestyle consisting of a motile, sporulating, thermophilic, anaerobic chemolithoautroptroph genetically capable of dissimilatory sulfate reduction, hydrogenotrophy, carbon fixation via the reductive acetyl-coenzyme A (Wood-Ljungdahl) pathway, and synthesis of all amino acids. The genome of "*Ca.* D. cowenii" also indicates a chemoorganotroph that possesses abundant sugar transporters and is capable of glycolysis, which is somewhat surprising given the low dissolved organic carbon concentrations in this system (Lin et al., 2012). Similar to "*Ca.* D audaxviator", hydrogenases were abundant in "*Ca.* D. cowenii", which is consistent with the availability of hydrogen in basement fluids of the Juan de Fuca

Ridge flank (Lin et al., 2014). Altogether, the shared features between "*Ca.* D. cowenii" and "*Ca.* D. audaxviator" help to explain the wide distribution of this lineage in the global deep subsurface.

Differences in functional gene complement

Despite highly similar genomes overall, comparisons of predicted proteins assigned to clusters of orthologous groups (COGs) revealed unique genes in "*Ca.* D cowenii" that were not found in "*Ca.* D. audaxviator" (Figure 3D; also see Supplementary Tables 6 and 7). These genes are likely locations to uncover features that differentiate the marine versus terrestrial members of this lineage. While most unique genes in the "*Ca.* D. cowenii" genome have general functional characterizations only (COG category R), the largest fraction of unique genes in the "*Ca.* D. cowenii" versus "*Ca.* D. audaxviator" genome are found within COG category M (Cell wall/membrane/envelop biogenesis) and include nucleoside-diphosphate-sugar epimerases (e.g. *galE*) and glycosyltransferases (e.g. *treT*) involved in cell wall biosynthesis, and possibly in the production of exopolysaccharides involved with biofilm formation. Defense mechanisms (COG category V) contained the highest ratio of unique genes in the "*Ca.* D. cowenii" genome compared to "Ca D. audaxviator" and includes genes related to ABC-type multidrug transport systems, multidrug resistance efflux pumps (*hylD*), and a class-A beta-lactamase. The marine genome has numerous monosaccharide transporters not present in the terrestrial genome, including those encoding for components of ribose/xylose, arabinose, methyl-galactoside, xylose, allose, and rhamnose transport. Thus, potential differences in organic carbon substrate specificity are evident.

Though the genome of "*Ca.* D. cowenii" is incomplete, within assembled contigs there are a small number of large indels that are also potential sources of functional differentiation between "*Ca*. D cowenii" and "*Ca.* D. audaxviator". An indel present in "*Ca.* D. audaxviator" but lacking in "*Ca.* D. cowenii" includes a nitrogenase operon as well as genes for ammonium transport and nitrogen regulation (Figure 6). While the genes for glutamine synthetase and glutamate synthase within the genome of "*Ca.* D. cowenii" suggest that it obtains its nitrogen from the abundant ammonia in Juan de Fuca Ridge flank crustal fluids (Lin et al., 2012), it appears to be unable to fix inorganic dinitrogen. Another indel suggests that "*Ca.* D. cowenii" lacks the capacity to produce cobalamin (Figure 6). Moreover, a large cassette of genes present in the "*Ca.* D. audaxviator" genome that is related to gas vesicle production (and flanked by an integrase and two transposases) is missing in "*Ca.* D. cowenii". Finally, CRISPR-CAS gene arrays and CRISPR elements were distinct between the two genomes (Figure 6), with the genome of "*Ca.* D. cowenii" encoding 14 CRISPR-associated proteins versus 25 in "*Ca.* D. audaxviator".

Distribution

The Desulfopertinax/Desulforudis lineage was detected in metagenomic data generated from the terrestrial subsurface of Mt. Terri, Switzerland and the Coast Range Ophiolite, California, USA (Figure 7A; see also Supplementary Table 2). It was also found within marine sediments from the coastal Atlantic and Pacific, a Yellowstone National Park hot spring, and the terrestrial subsurface in Ontario, Canada, but never identified in seawater worldwide. Mapping raw metagenome reads in a lineage-specific manner that discriminated between reads mapping to "*Ca.* D. audaxviator" and "*Ca.* D. cowenii" revealed partitioning of these genomes between terrestrial and marine environments, respectively (Figure 7B; see also Supplementary Table 3). Surprisingly, the ratio of mapped reads from "*Ca.* D. cowenii" to "*Ca.* D. audaxviator" was, highest (18.9) in a sample from the terrestrial subsurface. The next largest ratios were from the U1362A metagenome (7.3), three serpentinite groundwater metagenomes (1.7-1.6), and the U1362B metagenome (1.4). The ratio of "*Ca.* D. audaxviator" to "*Ca.* D. cowenii" reads was highest (up to ~165) in samples collected from the terrestrial subsurface of Witwatersrand Basin, South Africa, although this lineage also appears present in serpentinite fluids from the terrestrial subsurface. Thus, it appears that the Desulfopertinax/Desulforudis lineage has a cosmopolitan distribution throughout the global subsurface environment, as indicated by mapping reads from 489 metagenomes from the terrestrial and marine subsurface to the genomes of "*Ca.* D. cowenii" and "*Ca.* D. audaxviator", as well as gene clones identified in published SSU rRNA surveys (Figure 7; see also Figure 2B and Supplementary Tables 2 and 3).

Conclusions

Crustal fluids within the terrestrial and marine deep subsurface contain microbial life living at the biosphere's limit; globally, deep subsurface biosphere is thought be one of the largest reservoirs for microbial life on our planet. This study takes advantage of new sampling technologies and couples them with improvements to DNA sequencing and associated informatics tools in order to reconstruct the genome of an uncultivated *Firmicutes* bacterium from fluids collected deep within the subseafloor of the Juan de Fuca Ridge flank that has previously been documented within both the terrestrial and marine subsurface. Based on our analyses, the capacity for both autotrophic and heterotrophic lifestyles combined with motility and sporulation confers upon "*Ca.* D. cowenii" and "*Ca.* D. audaxviator" the ability to colonize the global deep biosphere. We believe this to be the only microbial lineage known to inhabit both marine and terrestrial deep subsurface systems, providing a unique opportunity to advance our understanding of subsurface microbiology. By comparing the genome of this microorganism to a terrestrial counterpart, we reveal a high and unsuspected degree of functional similarity spanning the marine and terrestrial members of this lineage. Based on the predicted ability to reduce sulfate for energy generation, the persistent detection of this lineage in deep marine biosphere studies, and its initial discovery by deep subseafloor pioneer James Cowen, we propose the name "Desulfopertinax cowenii" for this candidatus taxon.

Acknowledgements

We thank the captain and crew, A. Fisher, K. Becker, C. G. Wheat, and other members of the science teams on board R/V Atlantis cruise AT18-07. We thank Beth Orcutt for facilitating metagenome sequencing. We also thank the pilots and crew of remote-operated vehicle *Jason II.* We thank Brian Foster and Alex Copeland of the JGI for initial assembly of the metagenomes. We thank Sean Cleveland and the Hawaii HPC facility. This research was supported by funding from National Science Foundation grants MCB06-04014 and OCE-1260723 (to MSR) and OCE-1136488 (to RS), the Center for Dark Energy Biosphere Investigations, a National Science Foundation-funded Science and Technology Center of Excellence (NSF award OCE-0939564), and from Department of Energy Joint Genome Institute Community Sequencing Award 987 (to RS). The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This study used samples and data provided by the Integrated Ocean Drilling Program. This is SOEST contribution XXXX, HIMB contribution XXXX, and C-DEBI contribution XXXX.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* 215:403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. DOI: 10.1089/cmb.2012.0021.

Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, *et al.* (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8:209. DOI: 10.1186/1471-2105-8-209.

Chivian D, Brodie EL, Alm EJ, Culley DE, Dehal PS, DeSantis TZ, *et al.* (2008). Environmental genomics reveals a single-species ecosystem deep within Earth. *Science* 322:275–278. DOI: 10.1126/science.1155495.

Chou HH, Holmes MH. (2001). DNA sequence quality trimming and vector removal. *Bioinformatics* 17:1093–1104. DOI: 10.1093/bioinformatics/17.12.1093.

Cowen JP, Copson DA, Jolly J, Hsieh C-C, Lin H-T, Glazer BT, *et al.* (2012). Advanced instrument system for real-time and time-series microbial geochemical sampling of the deep (basaltic) crustal biosphere. *Deep Sea Res Pt I* 61:43–56. DOI: 10.1016/j.dsr.2011.11.004.

Cowen JP, Giovannoni SJ, Kenig F, Johnson HP, Butterfield D, Rappé MS, *et al.* (2003). Fluids from aging ocean crust that support microbial life. *Science* 299:120–123. DOI: 10.1126/science.1075653.

Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2:e243. DOI: 10.7717/peerj.243.

Darriba D, Taboada GL, Doallo R, Posada D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 9:772. DOI: 10.1038/nmeth.2109.

Durbin R, Eddy SR, Krogh A, Mitchison G. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press.

Eddy SR. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol* 7:e1002195. DOI: 10.1371/journal.pcbi.1002195.

Edgar RC. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* 8:18. DOI: 10.1186/1471-2105-8-18.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461. DOI: 10.1093/bioinformatics/btq461.

Huber R, Rossnagel P, Woese CR, Rachel R, Langworthy TA, Stetter KO. (1996). Formation of ammonium from nitrate during chemolithoautotrophic growth of the extremely thermophilic bacterium *Ammonifex degensii* gen. nov. sp. nov. *Syst Appl Microbiol* 19:40–49. DOI: 10.1016/S0723-2020(96)80007-5.

Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Palaniappan K, *et al.* (2015). The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Stand Genomic Sci* 10:86. DOI: 10.1186/s40793-015-0077-y.

Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Tennessen K, *et al.* (2016). The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). *Stand Genomic Sci* 11:17. DOI: 10.1186/s40793-016-0138-x.

Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* 28:2223–2230. DOI: 10.1093/bioinformatics/bts429.

Itävaara M, Nyyssönen M, Kapanen A, Nousiainen A, Ahonen L, Kukkonen I. (2011). Characterization of bacterial diversity to a depth of 1500 m in the Outokumpu deep borehole, Fennoscandian Shield. *FEMS Microbiol Ecol* 77:295–309. DOI: 10.1111/j.1574-6941.2011.01111.x.

Jungbluth SP, Bowers R, Lin H-T, Cowen JP, Rappé MS. (2016). Novel microbial assemblages inhabiting crustal fluids within mid-ocean ridge flank subsurface basalt. *ISME J* 10:2033–2047. DOI: 10.1038/ismej.2015.248.

Jungbluth SP, Grote J, Lin H-T, Cowen JP, Rappé MS. (2013). Microbial diversity within basement fluids of the sediment-buried Juan de Fuca Ridge flank. *ISME J* 7:161–172. DOI: 10.1038/ismej.2012.73.

Jungbluth SP, Lin H-T, Cowen JP, Glazer BT, Rappé MS. (2014). Phylogenetic diversity of microorganisms in subseafloor crustal fluids from Holes 1025C and 1026B along the Juan de Fuca Ridge flank. *Front Microbiol* 5:119. DOI: 10.3389/fmicb.2014.00119.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, *et al.* (2009). Circos: an information aesthetic for comparative genomics. *Genome Res* 19:1639–1645. DOI: 10.1101/gr.092759.109.

Labonté JM, Field EK, Lau M, Chivian D, van Heerden E, Wommack KE, *et al.* (2015). Single cell genomics indicates horizontal gene transfer and viral infections in a deep subsurface Firmicutes population. *Front Microbiol* 6:349. DOI: 10.3389/fmicb.2015.00349.

Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. DOI: 10.1038/nmeth.1923.

Lau MCY, Cameron C, Magnabosco C, Brown CT, Schilkey F, Grim S, *et al.* (2014). Phylogeny and phylogeography of functional genes shared among seven terrestrial subsurface metagenomes reveal N-cycling and microbial evolutionary relationships. *Front Microbiol* 5:531. DOI: 10.3389/fmicb.2014.00531.

Lerm S, Westphal A, Miethling-Graff R, Alawi M, Seibt A, Wolfgramm M, *et al.* (2013). Thermal effects on microbial composition and microbiologically induced corrosion and mineral precipitation affecting operation of a geothermal plant in a deep saline aquifer. *Extremophiles* 17:311–327. DOI: 10.1007/s00792-013-0518-8.

Letunic, I, Bork, P (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees, *Nucleic Acids Research* 44 (W1): W242-W245. DOI: 10.1093/nar/gkw290.

Lin H-T, Cowen JP, Olson EJ, Amend JP, Lilley MD. (2012). Inorganic chemistry, gas compositions and dissolved organic carbon in fluids from sedimented young basaltic crust on the Juan de Fuca Ridge flanks. *Geochim Cosmochim Acta* 85:213–227. DOI: 10.1016/j.gca.2012.02.017.

Lin H-T, Cowen JP, Olson EJ, Lilley MD, Jungbluth SP, Wilson ST, *et al.* (2014). Dissolved hydrogen and methane in the oceanic basaltic biosphere. *Earth Planet Sci Lett* 405:62–73. DOI: 10.1016/j.epsl.2014.07.037.

Lin L-H, Wang P-L, Rumble D, Lippmann-Pipke J, Boice E, Pratt LM, *et al.* (2006). Long-term sustainability of a high-energy, low-diversity crustal biome. *Science* 314:479–482. DOI: 10.1126/science.1127376.

Lowe TM, Eddy SR. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25:955–964. DOI: 10.1093/nar/25.5.0955.

Loy A, Duller S, Baranyi C, Mussmann M, Ott J, Sharon I, *et al.* (2009). Reverse dissimilatory sulfite reductase as phylogenetic marker for a subgroup of sulfur-oxidizing prokaryotes. *Environ Microbiol* 11:289–299. DOI: 10.1111/j.1462-2920.2008.01760.x.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* 32:1363–1371. DOI: 10.1093/nar/gkh293.

Lukashin AV, Borodovsky M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* 26:1107–1115. DOI: 10.1093/nar/26.4.1107.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, *et al.* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18. DOI: 10.1186/2047-217X-1-18.

Magnabosco C, Tekere M, Lau MCY, Linage B, Kuloyo O, Erasmus M, *et al.* (2014). Comparisons of the composition and biogeographic distribution of the bacterial communities occupying South African thermal springs with those inhabiting deep subsurface fracture water. *Front Microbiol* 5:679. DOI: 10.3389/fmicb.2014.00679.

Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, *et al.* (2003). CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res* 31:383–387. DOI: 10.1093/nar/gkg087.

Markowitz VM, Chen I-MA, Chu K, Szeto E, Palaniappan K, Pillay M, *et al.* (2014). IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res* 42:D568–73. DOI: 10.1093/nar/gkt919.

Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, *et al.* (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* 40:D115–22. DOI: 10.1093/nar/gkr1044.

Matsen FA, Kodner RB, Armbrust EV. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11:538. DOI: 10.1186/1471-2105-11-538.

Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF. (2011). EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol* 12:R44. DOI: 10.1186/gb-2011-12-5-r44.

Morgulis A, Gertz EM, Schäffer AA, Agarwala R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* 13:1028–1040. DOI: 10.1089/cmb.2006.13.1028.

Niu B, Zhu Z, Fu L, Wu S, Li W. (2011). FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics* 27:1704–1705. DOI: 10.1093/bioinformatics/btr252.

Noguchi H, Park J, Takagi T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* 34:5623–5630. DOI: 10.1093/nar/gkl723.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25:1043–1055. DOI: 10.1101/gr.186072.114.

Price, MN, Dehal, PS, Arkin, AP. (2010). FastTree 2- Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. DOI: 10.1371/journal.pone.0009490.

Pruesse E, Peplies J, Glöckner FO. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28:1823–1829. DOI: 10.1093/bioinformatics/bts252.

R Core Team. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria. ISBN: 3-900051-07-0.

Rho M, Tang H, Ye Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* 38:e191. DOI: 10.1093/nar/gkq747.

Robador A, Jungbluth SP, LaRowe DE, Bowers RM, Rappé MS, Amend JP, *et al.* (2015). Activity and phylogenetic diversity of sulfate-reducing microorganisms in low-temperature subsurface fluids within the upper oceanic crust*.* *Front Microbiol* 5:748. DOI: 10.3389/fmicb.2014.00748.

Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690. DOI: 10.1093/bioinformatics/btl446.

Stamatakis A, Hoover P, Rougemont J. (2008). A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol* 57:758–771. DOI: 10.1080/10635150802429642.

Sullivan MJ, Petty NK, Beatson SA. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27:1009–1010. DOI: 10.1093/bioinformatics/btr039.

Tennessen K, Andersen E, Clingenpeel S, Rinke C, Lundberg DS, Han J, *et al.* (2016). ProDeGe: a computational protocol for fully automated decontamination of genomes. *ISME J* 10:269–272. DOI: 10.1038/ismej.2015.100.

Tiago I, Veríssimo A. (2013). Microbial and functional diversity of a subterrestrial high pH groundwater associated to serpentinization. *Environ Microbiol* 15:1687–1706. DOI: 10.1111/1462-2920.12034.

Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M. (2011). Next generation sequence assembly with AMOS. *Curr Protoc Bioinformatics* Chapter 11:11.8. DOI: 10.1002/0471250953.bi1108s33.

Wheat CG, Jannasch HW, Kastner M, Hulme S, Cowen J, Edwards KJ, *et al.* (2011). Fluid sampling from oceanic borehole observatories: design and methods for CORK activities (1990-2010). *In* Proceedings of the Integrated Ocean Drilling Program Vol. 327 (eds. A. T. Fisher, T. Tsuji, K. Petronotis, & Expedition 327 Scientists) 1-36 (Integrated Ocean Drilling Program Management International, Inc., 2011). DOI: 10.2204/iodp.proc.327.109.2011.

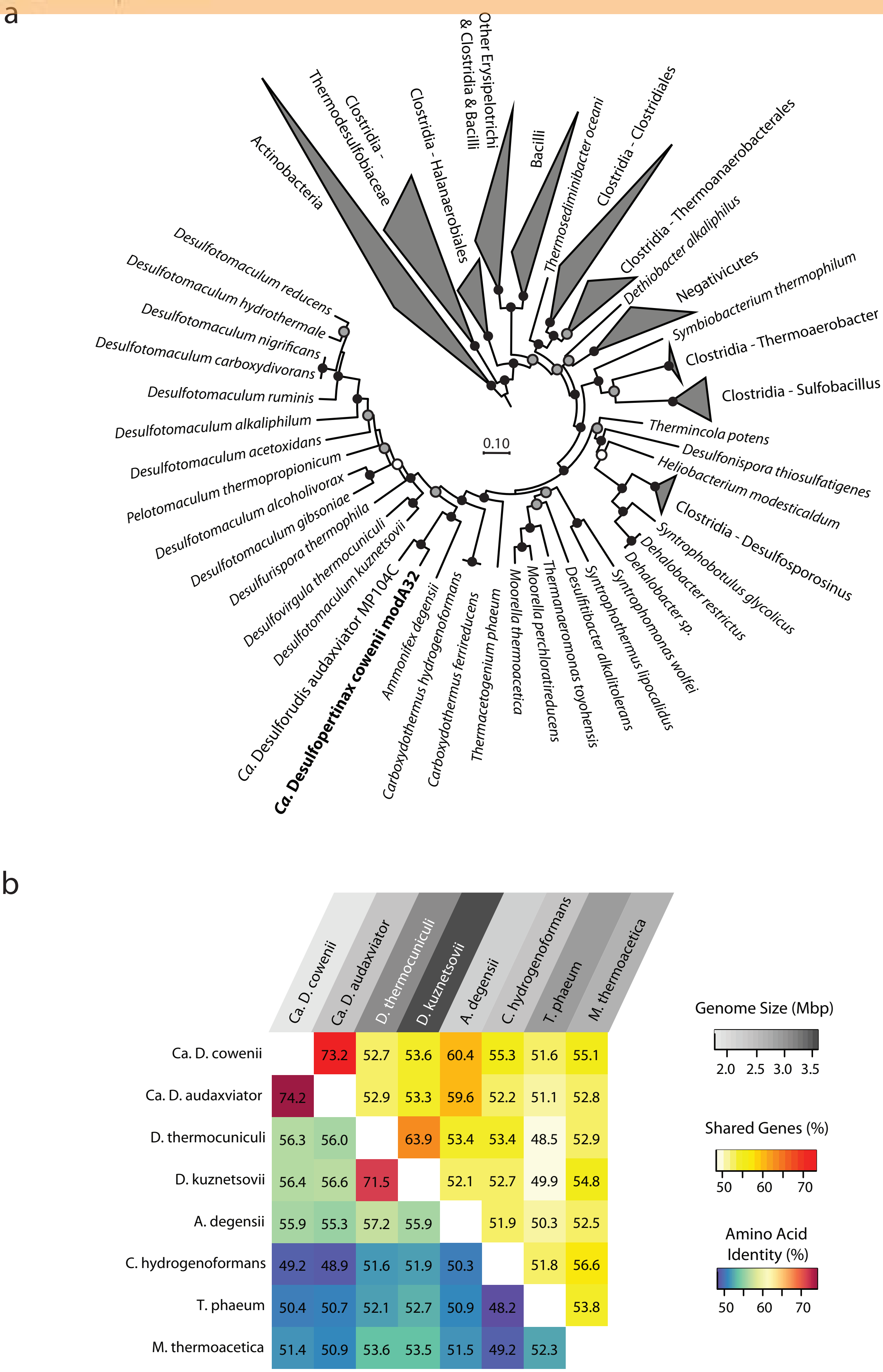
Wu D, Jospin G, Eisen JA. (2013). Systematic identification of gene families for use as “markers” for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. *PLoS ONE* 8:e77033. DOI: 10.1371/journal.pone.0077033.

Wu Y-W, Tang Y-H, Tringe SG, Simmons BA, Singer SW. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2:26. DOI: 10.1186/2049-2618-2-26.

Figure 1(on next page)

Phylogenomic and shared gene content analysis of “*Ca.* Desulfopertinax cowenii”, “*Ca.* Desulforudis audaxviator” and other Firmicutes.

Analysis of phylogenomic relationships, percent shared genes, and average amino-acid identity between “*Ca.* Desulfopertinax cowenii” modA32 and “*Ca.* Desulforudis audaxviator” MP104C reveal two lineages similar to each other and distinct from other *Firmicutes*. (A) Phylogenomic relationships between “*Ca.* D. cowenii”, “*Ca.* D. audaxviator”, and other *Firmicutes* based on a concatenated amino acid alignment. Black (100%), gray (>80%), and white (>50%) circles indicate nodes with high local support values, from 1000 replicates. Actinobacteria (n=687) were used as an outgroup. The scale bar corresponds to 0.10 substitutions per amino acid position. (B) Percent shared genes (upper right) and average amino-acid identity (lower left) between “*Ca.* D. cowenii”, “*Ca.* D. audaxviator”, and six closely related *Firmicutes* lineages from panel (A). The grey scale distinguishing horizontal axis labels corresponds to genome size.

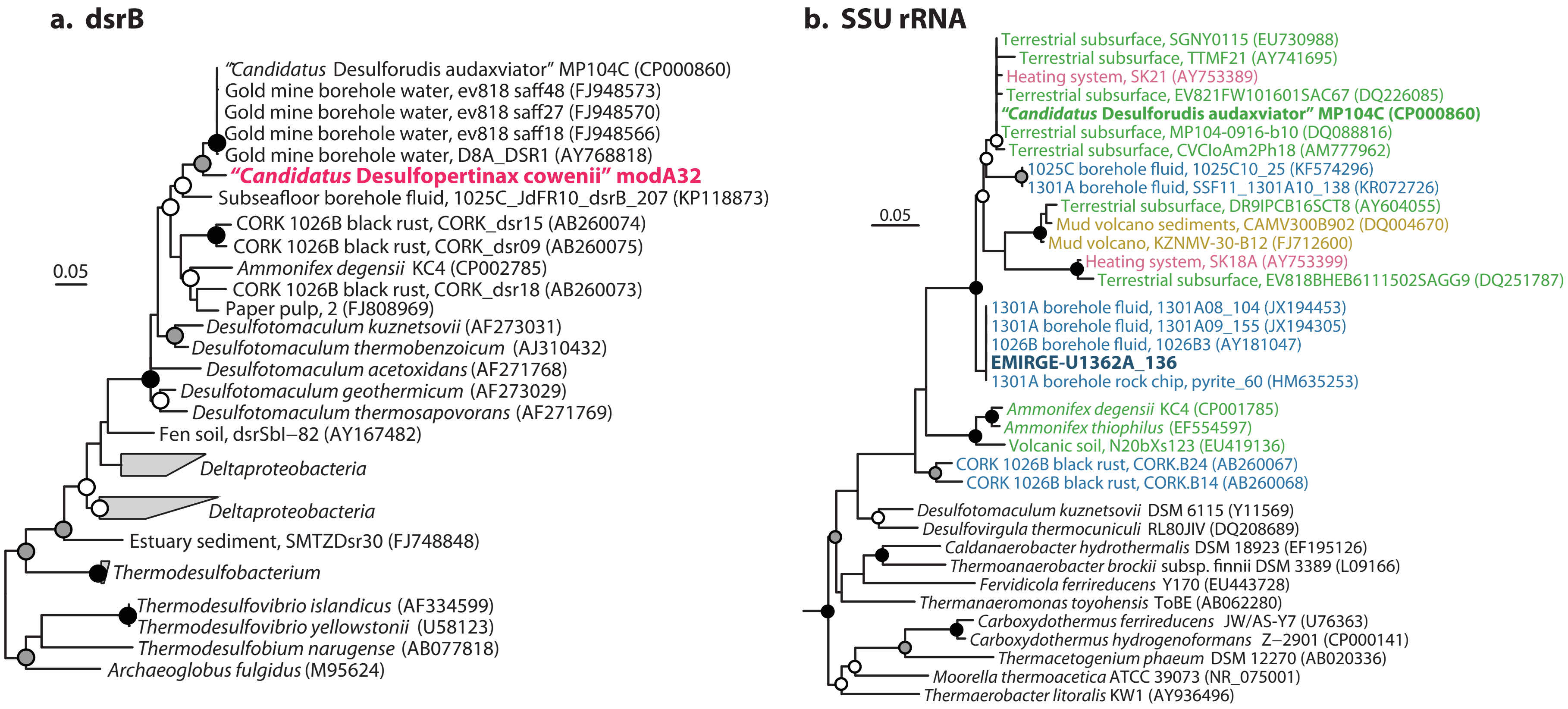


	Ca. D. cowenii	Ca. D. audaxviator	D. thermocuniculi	D. kuznetsovii	A. degensii	C. hydrogenoformans	T. phaeum	M. thermoacetica
Ca. D. cowenii		73.2	52.7	53.6	60.4	55.3	51.6	55.1
Ca. D. audaxviator	74.2		52.9	53.3	59.6	52.2	51.1	52.8
D. thermocuniculi	56.3	56.0		63.9	53.4	53.4	48.5	52.9
D. kuznetsovii	56.4	56.6	71.5		52.1	52.7	49.9	54.8
A. degensii	55.9	55.3	57.2	55.9		51.9	50.3	52.5
C. hydrogenoformans	49.2	48.9	51.6	51.9	50.3		51.8	56.6
T. phaeum	50.4	50.7	52.1	52.7	50.9	48.2		53.8
M. thermoacetica	51.4	50.9	53.6	53.5	51.5	49.2	52.3	

Figure 2(on next page)

Phylogenetic analysis of “*Ca.* Desulfopertinax cowenii”, “*Ca.* Desulforudis audaxviator” and other closely related *dsrB* and SSU rRNA genes.

Phylogenetic relationships between “*Ca.* Desulfopertinax cowenii”, “*Ca.* Desulforudis audaxviator”, and closely related *dsrB* genes (A) and a SSU rRNA gene related to “*Ca.* D. audaxviator” reconstructed from the U1362A metagenome via EMIRGE (B) lend additional support to a shared evolutionary history between “*Ca.* D. cowenii” and “*Ca.* D. audaxviator”. Black (100%), gray (≥80%), and white (≥50%) circles indicate nodes with bootstrap support, from 2000 replicates. The scale bars correspond to 0.05 substitutions per nucleotide position.



a. dsrB

"Candidatus Desulforudis audaxviator" MP104C (CP000860)
Gold mine borehole water, ev818 saff48 (FJ948573)
Gold mine borehole water, ev818 saff27 (FJ948570)
Gold mine borehole water, ev818 saff18 (FJ948566)
Gold mine borehole water, D8A_DSR1 (AY768818)
"Candidatus Desulfopertinax cowenii" modA32
Subseafloor borehole fluid, 1025C_JdFR10_dsrB_207 (KP118873)
CORK 1026B black rust, CORK_dsr15 (AB260074)
CORK 1026B black rust, CORK_dsr09 (AB260075)
Ammonifex degensii KC4 (CP002785)
CORK 1026B black rust, CORK_dsr18 (AB260073)
Paper pulp, 2 (FJ808969)
0.05
Desulfotomaculum kuznetsovii (AF273031)
Desulfotomaculum thermobenzoicum (AJ310432)
Desulfotomaculum acetoxidans (AF271768)
Desulfotomaculum geothermicum (AF273029)
Desulfotomaculum thermosapovorans (AF271769)
Fen soil, dsrSbl–82 (AY167482)
Deltaproteobacteria
Deltaproteobacteria
Estuary sediment, SMTZDsr30 (FJ748848)
Thermodesulfobacterium
Thermodesulfovibrio islandicus (AF334599)
Thermodesulfovibrio yellowstonii (U58123)
Thermodesulfobium narugense (AB077818)
Archaeoglobus fulgidus (M95624)

b. SSU rRNA

Terrestrial subsurface, SGNY0115 (EU730988)
Terrestrial subsurface, TTMF21 (AY741695)
Heating system, SK21 (AY753389)
Terrestrial subsurface, EV821FW101601SAC67 (DQ226085)
"Candidatus Desulforudis audaxviator" MP104C (CP000860)
Terrestrial subsurface, MP104-0916-b10 (DQ088816)
Terrestrial subsurface, CVCloAm2Ph18 (AM777962)
1025C borehole fluid, 1025C10_25 (KF574296)
1301A borehole fluid, SSF11_1301A10_138 (KR072726)
0.05
Terrestrial subsurface, DR9IPCB16SCT8 (AY604055)
Mud volcano sediments, CAMV300B902 (DQ004670)
Mud volcano, KZNMV-30-B12 (FJ712600)
Heating system, SK18A (AY753399)
Terrestrial subsurface, EV818BHEB6111502SAGG9 (DQ251787)
1301A borehole fluid, 1301A08_104 (JX194453)
1301A borehole fluid, 1301A09_155 (JX194305)
1026B borehole fluid, 1026B3 (AY181047)
EMIRGE-U1362A_136
1301A borehole rock chip, pyrite_60 (HM635253)
Ammonifex degensii KC4 (CP001785)
Ammonifex thiophilus (EF554597)
Volcanic soil, N20bXs123 (EU419136)
CORK 1026B black rust, CORK.B24 (AB260067)
CORK 1026B black rust, CORK.B14 (AB260068)
Desulfotomaculum kuznetsovii DSM 6115 (Y11569)
Desulfovirgula thermocuniculi RL80JIV (DQ208689)
Caldanaerobacter hydrothermalis DSM 18923 (EF195126)
Thermoanaerobacter brockii subsp. finnii DSM 3389 (L09166)
Fervidicola ferrireducens Y170 (EU443728)
Thermanaeromonas toyohensis ToBE (AB062280)
Carboxydothermus ferrireducens JW/AS-Y7 (U76363)
Carboxydothermus hydrogenoformans Z–2901 (CP000141)
Thermacetogenium phaeum DSM 12270 (AB020336)
Moorella thermoacetica ATCC 39073 (NR_075001)
Thermaerobacter litoralis KW1 (AY936496)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2592v1 | CC BY 4.0 Open Access | rec: 14 Nov 2016, publ:

Figure 3(on next page)

Analysis of genome alignment and shared and unique gene inventories in "*Ca.* Desulfopertinax cowenii" and "*Ca.* Desulforudis audaxviator".

Multiple genome alignment and analysis of shared and unique gene inventories reveal key conserved and variable features of "*Ca.* Desulfopertinax cowenii" and "*Ca.* Desulforudis audaxviator". (A) Comparison of the "*Ca.* D. cowenii" genome scaffolds with "*Ca.* D. audaxviator" based on reciprocal best BLAST. From innermost to outermost, concentric circles show: nucleotide positions of genomes and scaffolds, percent GC content using a 100 bp sliding window, similarity of mapped U1362A reads. Links connecting circles are colored according to "*Ca.* D. cowenii" scaffold origin [Ga007115_(11-16)] and the degree of shading represents similarities (minimum similarity 70%) based on BLAST comparisons using < 75% (light shade), ≥ 75% (dark shade) nucleic acid identity thresholds. (B) Frequency of reciprocal best BLAST hits (n=1364) by percent similarity. Percent similarity histogram bins are in 2% increments and the dashed lines indicate average nucleotide identity (red) and average amino acid identity (blue) between "*Ca.* D. cowenii" and "*Ca.* D. audaxviator". Relative abundance of shared (C) and unique (D) genes in the "*Ca.* D. cowenii" and "*Ca.* D. audaxviator" genomes, sorted by annotated COG categories.

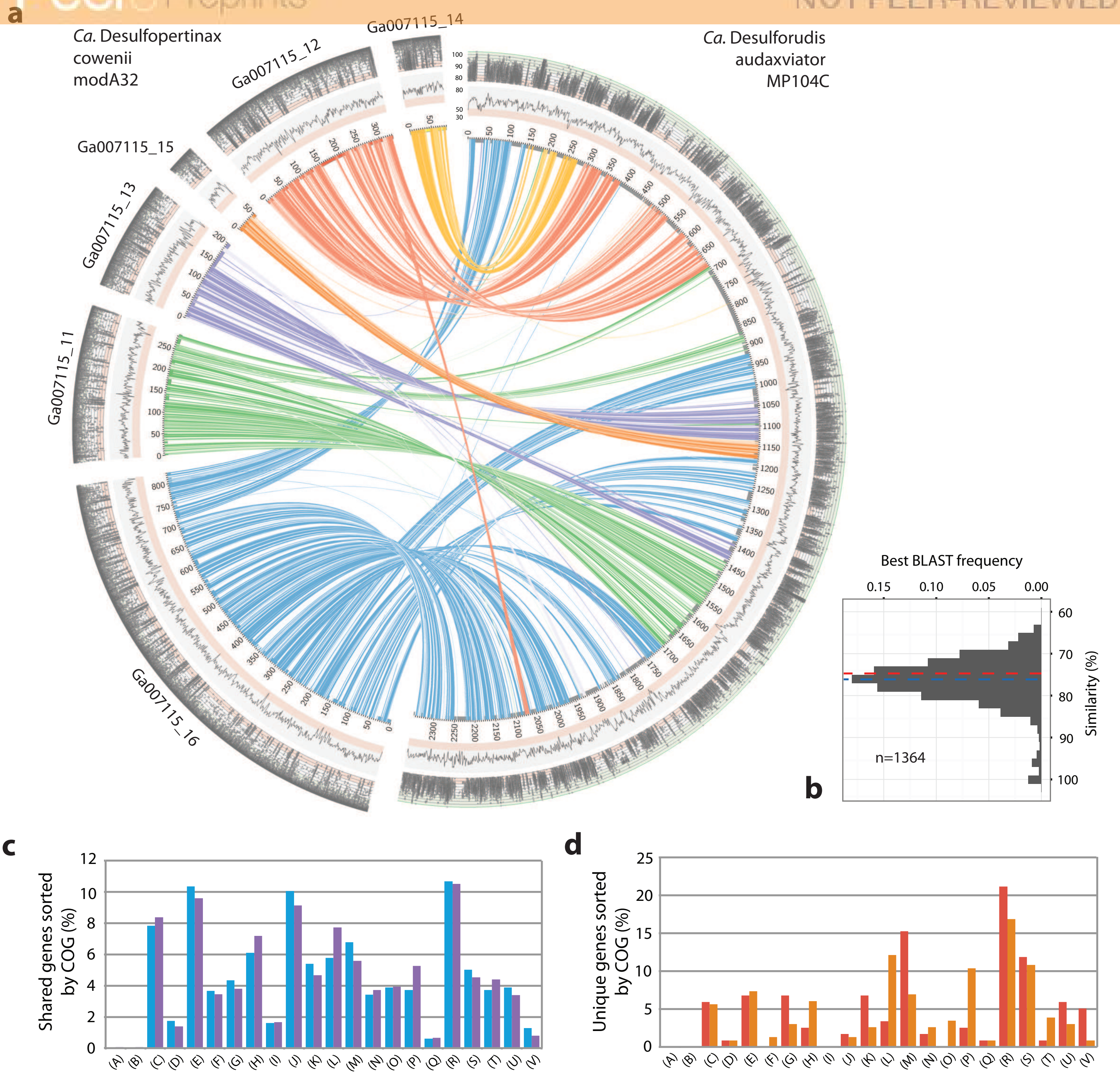
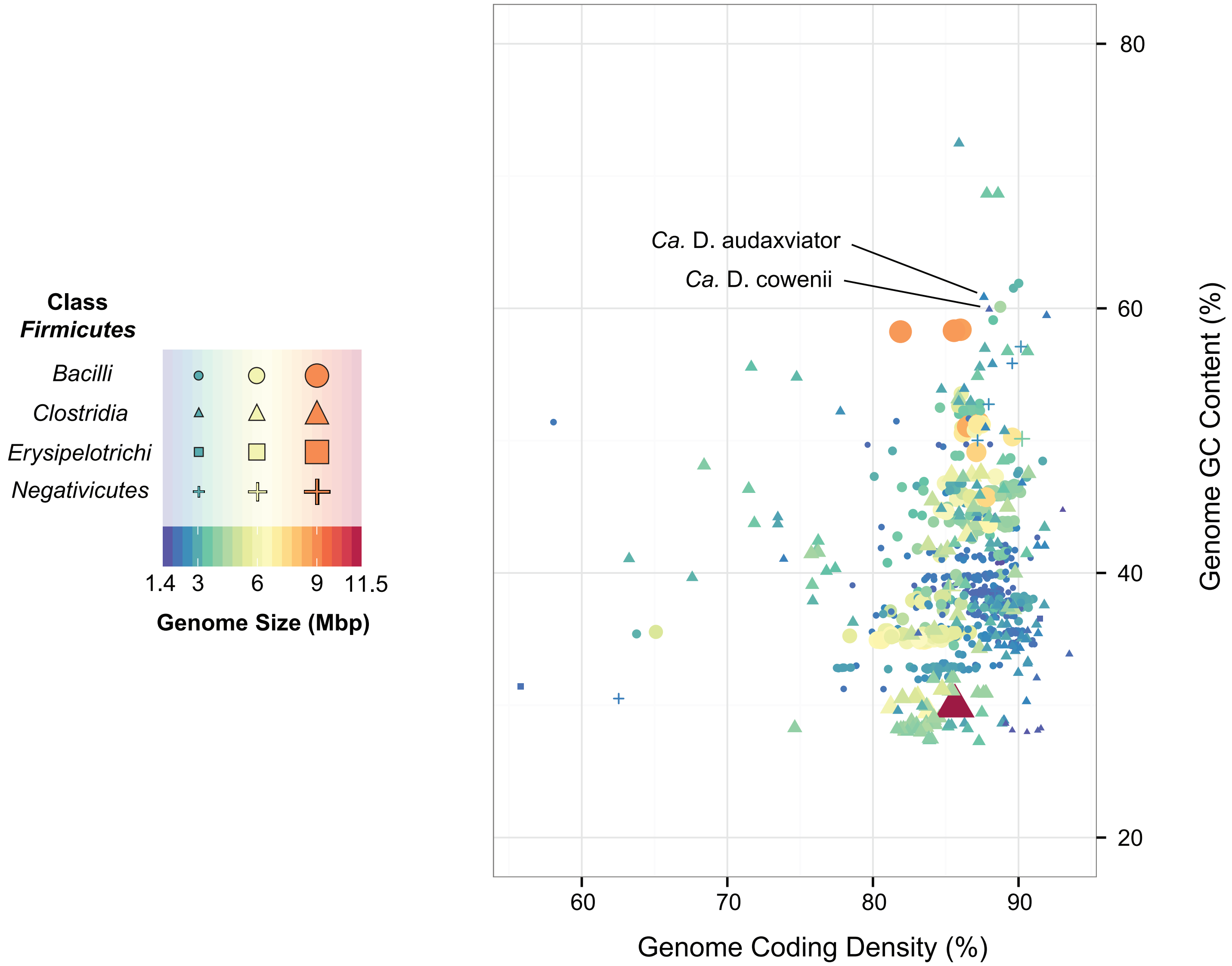

a
Ca. Desulfopertinax cowenii modA32
Ca. Desulforudis audaxviator MP104C
Ga007115_14
Ga007115_12
Ga007115_15
Ga007115_13
Ga007115_11
Ga007115_16
b
Best BLAST frequency
0.15
0.10
0.05
0.00
60
70
80
90
100
Similarity (%)
n=1364
c
Shared genes sorted by COG (%)
0
2
4
6
8
10
12
(A) (B) (C) (D) (E) (F) (G) (H) (I) (J) (K) (L) (M) (N) (O) (P) (Q) (R) (S) (T) (U) (V)
d
Unique genes sorted by COG (%)
0
5
10
15
20
25
(A) (B) (C) (D) (E) (F) (G) (H) (I) (J) (K) (L) (M) (N) (O) (P) (Q) (R) (S) (T) (U) (V)

Figure 4(on next page)

Survey of Firmicutes genome characteristics.

Survey of *Firmicutes* genome size, genome GC content, and coding density separated by different classes (*Bacilli*, *Clostridia*, *Erysipelotrichi*, *Negativicutes*). Only complete genomes and genomes with GC content >20% were used (n=909). The genome size of "*Ca.* Desulfopertinax cowenii" was estimated by assuming the current genome length (1.78 Mbp) was 98% the total genome length. Classes are distinguished by shape, while genome size is indicated by shape size and color. All genomes were downloaded from IMG on December 13, 2015.

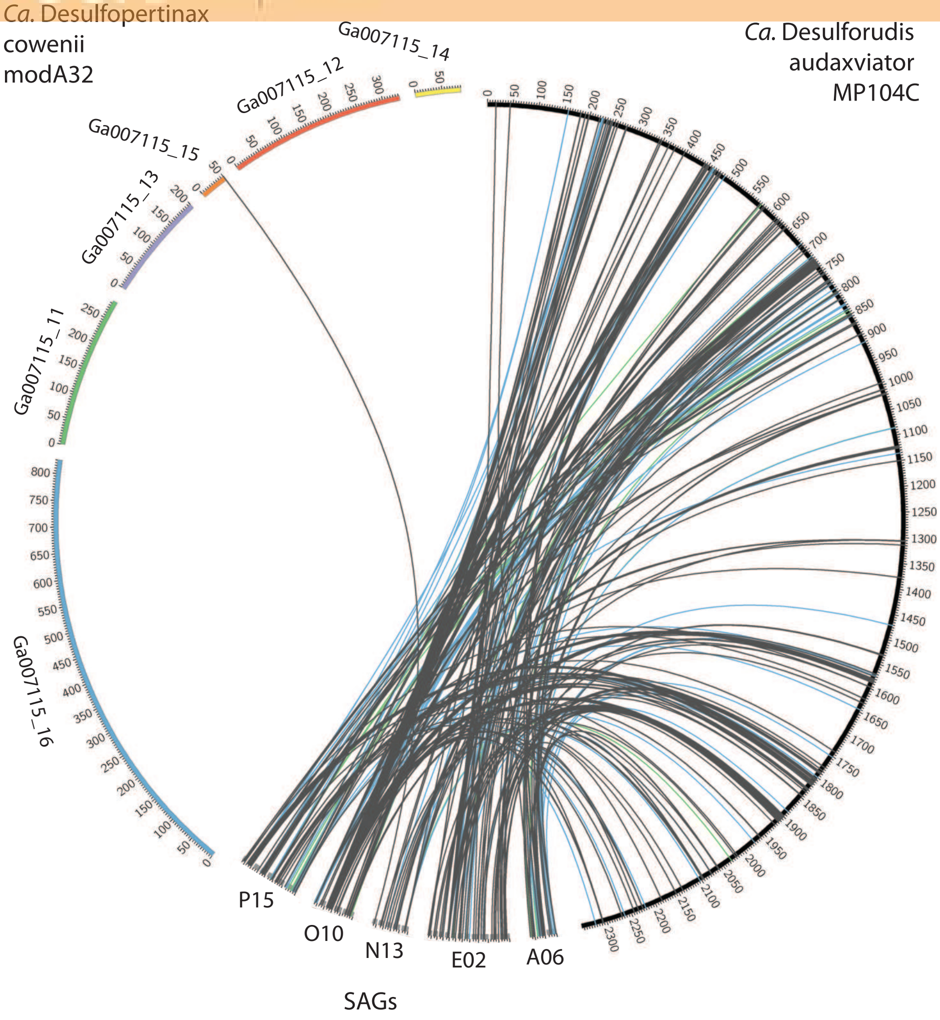


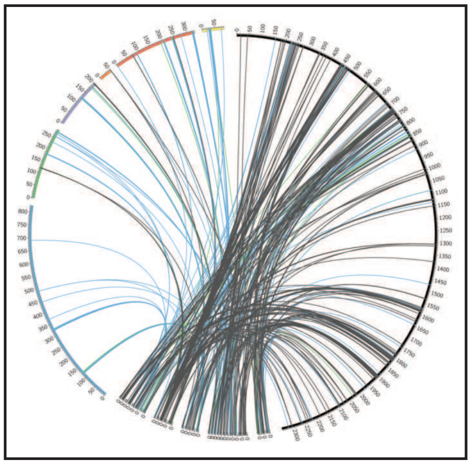
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2592v1 | CC BY 4.0 Open Access | rec: 14 Nov 2016, publ:

Figure 5(on next page)

Analysis of genome alignment between “*Ca.* Desulfopertinax cowenii”, “*Ca.* Desulforudis audaxviator” and five closely related single-cell genomes.

Comparison of terrestrial deep subsurface SAGs AC-310-P15, O10, N13, E02, and A06 with the genomes of “*Ca.* Desulfopertinax cowenii” and “*Ca.* Desulforudis audaxviator”. Links connecting colored circles represent similarities based on blastn comparisons allowing a maximum of two best hit and using 75 – 80% (green), 80 – 85% (blue), > 85% (grey) nucleic acid identity thresholds. Inset plot indicates blastn comparisons allowing a maximum of a two best hits.





PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2592v1 | CC BY 4.0 Open Access | rec: 14 Nov 2016, publ:

Figure 6(on next page)

Comparative analysis of genomic organization in “*Ca.* Desulfopertinax cowenii” and “*Ca.* Desulforudis audaxviator”.

Comparison of genomic organization in “*Ca.* Desulfopertinax cowenii” with “*Ca.* Desulforudis audaxviator” highlighting regions with large, internal insertion/deletion events containing no homologous genes in the opposing genome. (A) nitrogen-fixation operon, (B) vitamin B12 synthesis, (C) gas vesicle production, (D) a CRISPR-CAS array. Genes are colored according to COG categories and BLAST similarity between regions is indicated by shading intensity.

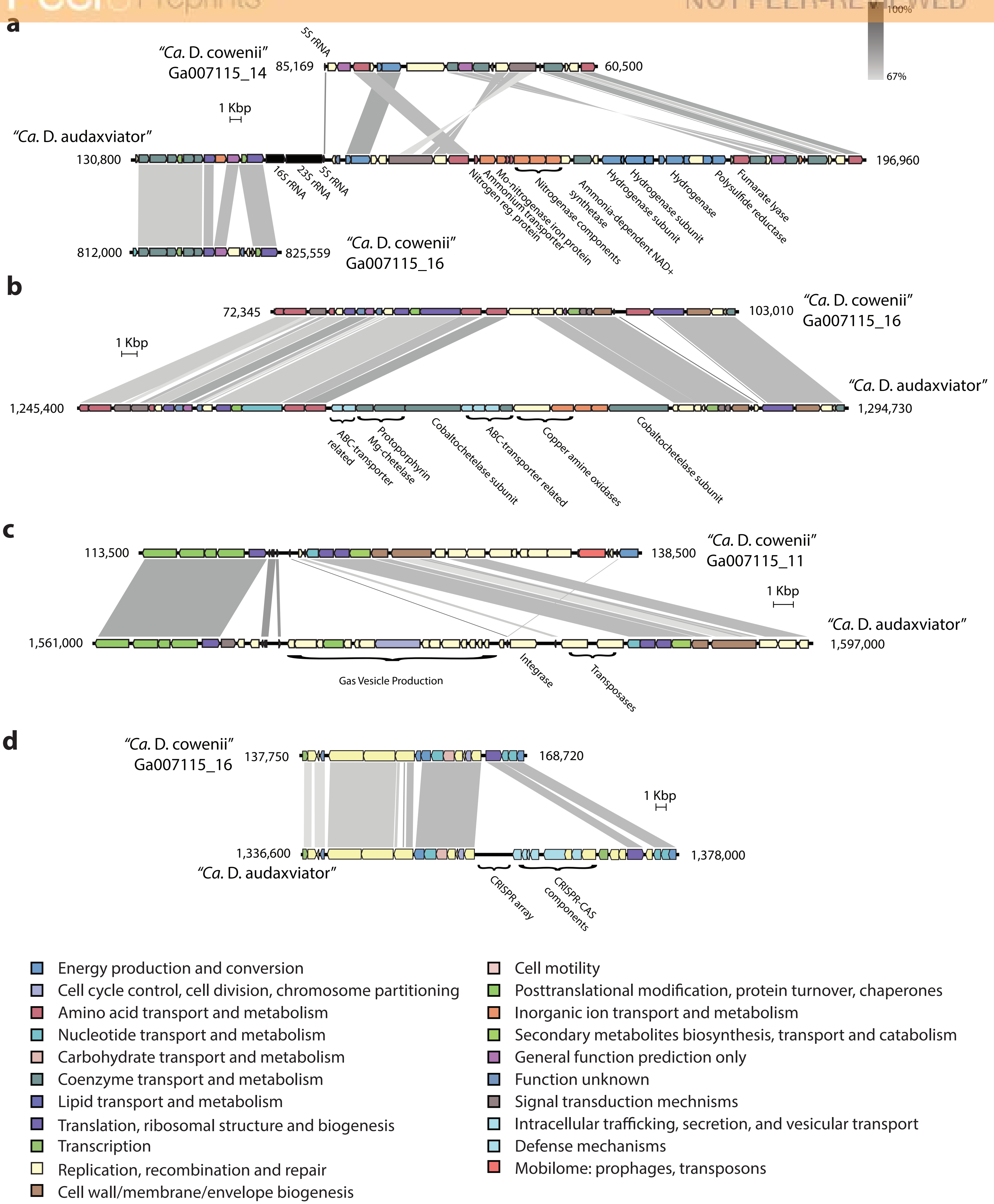

a
100%
67%
"Ca. D. cowenii"
Ga007115_14
85,169
5S rRNA
60,500
1 Kbp
"Ca. D. audaxviator"
130,800
16S rRNA
23S rRNA
5S rRNA
Nitrogen reg. protein
Ammonium transporter
Mo-nitrogenase iron protein
Nitrogenase components
Ammonia-dependent NAD+ synthetase
Hydrogenase subunit
Hydrogenase subunit
Hydrogenase
Polysulfide reductase
Fumarate lyase
196,960
812,000
825,559
"Ca. D. cowenii"
Ga007115_16
b
72,345
103,010
"Ca. D. cowenii"
Ga007115_16
1 Kbp
"Ca. D. audaxviator"
1,245,400
1,294,730
ABC-transporter related
Mg-chetelase
Protoporphyrin
Cobaltochetelase subunit
ABC-transporter related
Copper amine oxidases
Cobaltochetelase subunit
c
113,500
138,500
"Ca. D. cowenii"
Ga007115_11
1 Kbp
"Ca. D. audaxviator"
1,561,000
1,597,000
Gas Vesicle Production
Integrase
Transposases
d
"Ca. D. cowenii"
Ga007115_16
137,750
168,720
1 Kbp
1,336,600
1,378,000
"Ca. D. audaxviator"
CRISPR array
CRISPR-CAS components
Energy production and conversion
Cell cycle control, cell division, chromosome partitioning
Amino acid transport and metabolism
Nucleotide transport and metabolism
Carbohydrate transport and metabolism
Coenzyme transport and metabolism
Lipid transport and metabolism
Translation, ribosomal structure and biogenesis
Transcription
Replication, recombination and repair
Cell wall/membrane/envelope biogenesis
Cell motility
Posttranslational modification, protein turnover, chaperones
Inorganic ion transport and metabolism
Secondary metabolites biosynthesis, transport and catabolism
General function prediction only
Function unknown
Signal transduction mechnisms
Intracellular trafficking, secretion, and vesicular transport
Defense mechanisms
Mobilome: prophages, transposons

Figure 7(on next page)

Analysis of the global distribution of “*Ca.* Desulfopertinax cowenii” and ”*Ca.* Desulforudis audaxviator”.

“*Ca.* Desulfopertinax cowenii” and ”*Ca.* Desulforudis audaxviator” are globally-distributed in the deep subsurface. (A) Ellipse sizes correspond to the frequency of mapped reads from environmental metagenomes to “*Ca.* D. cowenii” and ”*Ca.* D. audaxviator” genomes. Triangles indicate locations where a lineage has been detected in SSU rRNA gene surveys. The average frequency of reads mapped to “*Ca.* D. coweii” and “*Ca.* D. audaxviator” are shown for all metagenomes listed in Supplementary Table 2 with >50000 genes. (B) Graphical representation of the frequency of environmental genome reads mapping to the “*Ca.* D. cowenii” and “*Ca.* D. audaxviator” genomes using a 96% read similarity score. Environmental metagenomes with the highest ratio of reads mapped to “*Ca.* D. cowenii” vs. “*Ca.* D. audaxviator” and having an average frequency of ≥0.00025 mapped reads are ordered in clockwise fashion from highest to lowest (Supplementary Table 2). MG-RAST metagenome 4440282 was retained solely because it had the highest ratio of reads mapped to “*Ca.* D. cowenii”:”*Ca.* D. audaxviator”. Links are colored according to the environmental source of each metagenome, while link sizes are proportional to the frequency of a read from a metagenome to map to one genome or the other. The log of metagenome size (number of reads) was used to create the relative length of the outer edges of the circle, which coarsely divide the environments into marine versus terrestrial. The “*Ca.* D. cowenii” genome is sized 2.2x the largest displayed metagenome and “*Ca.* D. audaxviator” is 1.32x (ratio of genome sizes) larger than the “*Ca.* D. cowenii” genome.

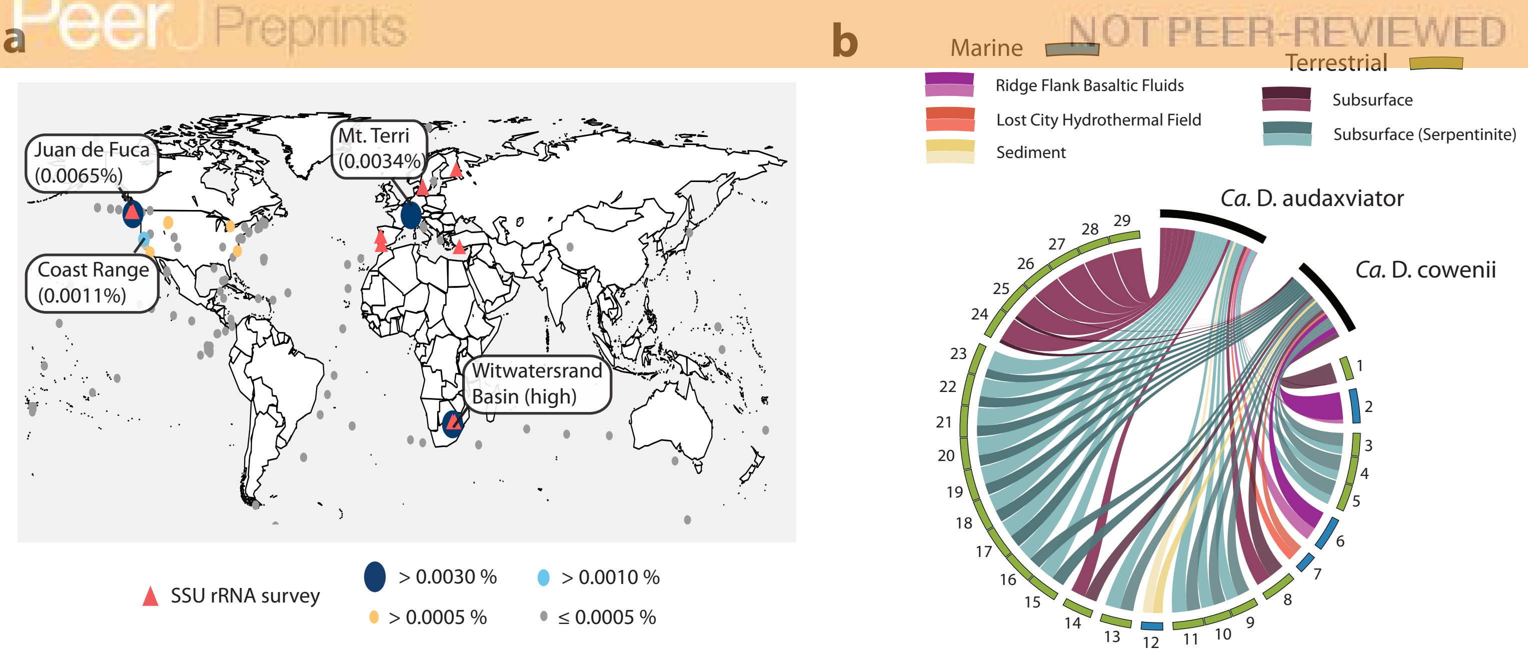

a
Juan de Fuca
(0.0065%)
Mt. Terri
(0.0034%)
Coast Range
(0.0011%)
Witwatersrand
Basin (high)
▲ SSU rRNA survey
● > 0.0030 %
● > 0.0010 %
● > 0.0005 %
● ≤ 0.0005 %
b
Marine
Terrestrial
Ridge Flank Basaltic Fluids
Lost City Hydrothermal Field
Sediment
Subsurface
Subsurface (Serpentinite)
Ca. D. audaxviator
Ca. D. cowenii
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

Table 1(on next page)

Genome characteristics of “*Ca.* Desulfopertinax cowenii” modA32 and “*Ca.* Desulforudis audaxviator” MP104C.

	"Ca. D. cowenii"	*"Ca. D. audaxviator"*
Percent complete	98-99% (6 scaffolds)	100% (closed)
Genome size (bp)	1,778,734	2,349,476
Percent coding	89.8%	87.6%
GC content	60.2%	60.9%
Total no. of genes	1842	2293
No. of protein coding genes	1782 (96.7%)	2239 (97.6%)
With function prediction	1518 (85.2%)	1587 (70.9%)
Without function prediction	264 (14.8%)	652 (29.1%)
Shared	1514 (85.0%)	1606 (71.7%)
Paralogs	137	265
Pseudogenes	n.d.[a]	82
rRNA genes	2	6
5S rRNA	2	2
16S rRNA	n.d.	2
23S rRNA	n.d.	2
tRNA genes	44	45
CRISPR elements	1	4
Mobile elements (integrases/transposons)	6/7	23/81

[a] n.d. – not detected

Table 2(on next page)

“*Ca.* Desulforudis audaxviator” MP104C-related genome bins from the U1362A metagenome, analyzed by CheckM.

Bin_ID	Total contigs/ N50 (Kbp)/ longest contig (Kbp)	Completeness (%)	Contamination (%)	Strain Heterogeneity (%)	Total Bases (Mbp)
D. audaxviator	--	98.09	0.32	0	2.35
1362A_maxbin32	50/112/179	97.61	5.10	100	1.87
1362A_maxbin32 (ProDeGe filtered)	31/112/179	95.70	5.10	100	1.81
"*Ca.* D. cowenii" modA32 (SPAdes reassembly, ProDeGe filtered)	6/332/826	97.61	0	0	1.78