Systematic and Applied Microbiology 42 (2019) 54–66

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

Discovery and ecogenomic context of a global Caldiserica-related phylum active in thawing permafrost, *Candidatus* Cryosericota phylum nov., *Ca.* Cryosericia class nov., *Ca.* Cryosericales ord. nov., Ca. Cryosericaceae fam. nov., comprising the four species Cryosericum septentrionale gen. nov. sp. nov., *Ca.* C. hinesii sp. nov., *Ca.* C. odellii sp. nov., *Ca.* C. terrychapinii sp. nov.

Miguel A. Martinez[a,1], Ben J. Woodcroft[b], Julio C. Ignacio Espinoza[a,2], Ahmed A. Zayed[a], Caitlin M. Singleton[b], Joel A. Boyd[b], Yueh-Fen Li[a], Samuel Purvine[c], Heather Maughan[d], Suzanne B. Hodgkins[a,e], Darya Anderson[f], Maya Sederholm[f], Ben Temperton[g], Benjamin Bolduc[a], IsoGenie Project Coordinators[3], Scott R. Saleska[h], Gene W. Tyson[b], Virginia I. Rich[a,f,*]

[a] *Department of Microbiology, The Ohio State University, Columbus, OH 43210, United States*
[b] *Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, University of Queensland, Australia*
[c] *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99354, United States*
[d] *Ronin Institute, Montclair, NJ 07043, United States*
[e] *Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL 32306, United States*
[f] *Department of Soil, Water and Environmental Sciences, University of Arizona, Tucson, AZ 85716, United States*
[g] *School of Biosciences, University of Exeter, Exeter, EX4 4QD, UK*
[h] *Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85716, United States*

ARTICLE INFO

Article history:
Received 30 April 2018
Received in revised form 5 December 2018
Accepted 5 December 2018

Keywords:
Metagenome-assembled genome
Caldiserica
Ca. Cryosericota
Ca. Cryosericum
Permafrost
Stordalen Mire

ABSTRACT

The phylum Caldiserica was identified from the hot spring 16S rRNA gene lineage 'OP5' and named for the sole isolate *Caldisericum exile*, a hot spring sulfur-reducing chemoheterotroph. Here we characterize 7 Caldiserica metagenome-assembled genomes (MAGs) from a thawing permafrost site in Stordalen Mire, Arctic Sweden. By 16S rRNA and marker gene phylogenies, and average nucleotide and amino acid identities, these Stordalen Mire Caldiserica (SMC) MAGs form part of a divergent clade from *C. exile*. Genome and meta-transcriptome and proteome analyses suggest that unlike *Caldisericum*, the SMCs (i) are carbohydrate- and possibly amino acid fermenters that can use labile plant compounds and peptides, and (ii) encode adaptations to low temperature. The SMC clade rose to community dominance within permafrost, with a peak metagenome-based relative abundance of ~60%. It was also physiologically active in the upper seasonally-thawed soil. Beyond Stordalen Mire, analysis of 16S rRNA gene surveys indicated a global distribution of this clade, predominantly in anaerobic, carbon-rich and cold environments. These findings establish the SMCs as four novel phenotypically and ecologically distinct species within a single novel genus, distinct from *C. exile* clade at the phylum level. The SMCs are thus part of a novel cold-habitat phylum for an understudied, globally-distributed superphylum encompassing the Caldiserica. We propose the names *Candidatus* Cryosericota phylum nov., *Ca.* Cryosericia class nov., *Ca.* Cryosericales ord. nov., *Ca.* Cryosericaceae fam. nov., *Ca.* Cryosericum gen. nov., *Ca.* Cryosericum septentrionale sp. nov., *Ca.* C. hinesii sp. nov., *Ca.* C. odellii sp. nov., and *Ca.* C. terrychapinii sp. nov.

* Corresponding author at: The Ohio State University, 105 Biological Sciences Building, 484 W. 12th Ave, Columbus OH, 43210, United States.
E-mail addresses: rich.270@osu.edu (V.I. Rich).
[1] Current address: Department of Marine Biotechnology, Ensenada Center for Scientific Research and Higher Education, CICESE, Ensenada-Tijuana highway, Ensenada 22860, Mexico.
[2] Current address: Department of Biology, University of Southern California, Los Angeles CA 90089.
[3] A list of authors and affiliations appears in the Supplementary Information.

https://doi.org/10.1016/j.syapm.2018.12.003

Introduction

Originally detected in Yellowstone hot springs as the 16S rRNA-based lineage 'OP5' [21], the phylum Caldiserica (*caldus* denoting the hot habitat, *sericum* indicating the silk-like filaments it forms) has a single cultured isolate from a Japanese hot spring, *Caldisericum exile* [48]. *C. exile* is a hyperthermophilic chemoheterotrophic bacterium that contributes to sulfur cycling via the reduction of sulfur compounds (though not sulfate) [47]. Marker gene surveys have revealed that this phylum inhabits a wide range of environments: hot springs and vents [75], mesic and cryic habitats [sediment [20], permafrost [13] and active layer (i.e. the seasonally-thawed ground overlying permafrost) [38], and ice [76], as well as engineered and human-impacted habitats such as waste bioreactors [44] and hydrocarbon-contaminated soils [70]. These surveys also expanded Caldiserica's known phylogenic breadth from the original isolate family *Caldisericaceae* to 5 additional unofficial families [56]: 08D2Z23, TTA-B15, TTA-B1, LF045, WCHB1-02.

Further insights into Caldiserica's metabolism have been advanced by the recovery of 7 genomes from large-scale sequencing efforts: a single-cell assembled genome (SAG) from a terephthalate-degrading reactor [61], a metagenome-assembled genome (MAG) CG2 from a cold-water geyser in Utah [56], and 5 MAGs (from [54], names prefixed with UBA for Uncultivated Bacteria and Archaea) mainly from hydrocarbon-contaminated environments (Supplementary Table 1). These genomes revealed broader metabolic functionality than that present in *C. exile*, including sulfate reduction [61], syntrophic or secondary degradation [50], amino acid degradation [61], ammonia assimilation, and formate and acetate production [56].

As part of ongoing research at Stordalen Mire in subarctic Sweden, where permafrost thaw is altering habitats and plant and microbial communities [16,19,23,45,51,67,74] and increasing the release of the potent greenhouse gas methane [3,12,42], we identified Caldiserica 16S rRNA gene sequences in the rim and hole of a collapsed palsa feature, from a lineage which rose to high abundance within the permafrost layer. Metagenomes from nearby sites (permafrost and active layer) were then examined for Caldiserica, and 7 MAGs were assembled. Herein, we relate these Stordalen Mire Caldiserica (SMC) MAGs to previously-known Caldiserica members, infer their metabolism and geographic distribution, use *in situ* expression data to support their status as active members of post-thaw microbial communities, and discuss their role in the Stordalen Mire ecosystem. Phylogenetic, identity, and metabolic analyses showed these MAGs to be a distinct lineage within the Caldiserica and yet novel at the phylum level, such that Caldiserica is a superphylum.

Materials and methods

Sample collection and sequencing

The collection, extraction, and sequencing of samples from a collapsed palsa feature, used for 16S rRNA amplicon-based microbial community characterizations, are detailed in the Supplementary Information. Samples for metagenomes were collected, extracted and metagenomically sequenced as described in Woodcroft and Singleton et al. [74]. Briefly, metagenomes were sequenced from 214 samples collected across Stordalen Mire palsa, bog and fen environments [74], which represent a permafrost thaw gradient (the palsas are underlain by intact permafrost). TruSeq Nano (Illumina, San Diego, CA) paired-end libraries were prepared from 100 ng of starting material and sequenced on 1/12th of an Illumina HiSeq200 lane (100 bp), or 1/24th of an Illumina NextSeq lane (150 bp), yielding an average of 6.3 Gb per sample, with a subset of samples undergoing deeper sequencing [74]. Sequencing depth details are available in Supplementary Table 2.

Genome binning

Stordalen Mire Caldiserica MAG 1 (SMC1)

Reads from several palsa metagenomes (Supplementary Table 2) were co-assembled using CLC Genomics Workbench version 4.4 (CLC Genomics) after adapter clipping and quality trimming with SeqPrep (https://github.com/jstjohn/SeqPrep) and Nesoni (https://github.com/Victorian-Bioinformatics-Consortium/nesoni). Binning was performed by mapping reads back to the assemblies with BamM 'make' and then 'parse' using the trimmed mean (tpmean) coverage (https://github.com/Ecogenomics/BamM), and then manually inspecting R plots [59] to ensure that binned contigs satisfied manually determined cutoffs for GC content (<0.625) and coverage (determined for the 2 samples in which SMC1 was most abundant, P19.201205 and P28.201205, to be $275 < X < 375$ and <7, respectively). Completeness and contamination of each MAG bin was determined using the CheckM v0.9.4 [53] lineage_wf pipeline that identifies and quantifies single-copy marker genes, making use of pplacer 1.1 alpha 16 [41]. SMC1 was selected based on completeness (>80%) and contamination (<5%). For context on this tool's performance on the closest isolate genome for these MAGs, CheckM assessed the *C. exile* genome as 98.21% complete and 0% contaminated. The SMC1 16S rRNA gene identified it as belonging to Caldiserica.

SMC2-7

Reads from each sample (Supplementary Table 2) were individually assembled with an estimated insert size of 50–500 bp in CLC Genomics Workbench version 4.4 (CLC Genomics) [74]. Binning based on differential coverage was done by mapping all reads from each sampled site (palsa, bog or fen) to all assemblies of that site, using BamM's 'make' function (Imelfort and Lamberton et al., unpublished, http://ecogenomics.github.io/BamM/) version 1.3.8–1.5.0, BWA 0.7.12 [36], samtools [35], and GNU parallel [69]. Scaffolds from each sample were binned using MetaBAT v3127e20aa4e7 [26] using sample contigs and BAM files to identify differential coverage. Each MAG bin was examined using the CheckM v1.0.4 [53] lineage_wf pipeline using the same completeness and contamination thresholds as above. The SMC2-7 genomes were identified as Caldiserica by placement in a genome tree created from the concatenated alignment of 120 single-copy marker genes using GTDB (http://gtdb.ecogenomic.org/), as described below and in Woodcroft et al. [74]. The SMC5 16S rRNA gene also placed it within Caldiserica.

Phylogenetic reconstructions

Full-length 16S rRNA sequences

16S rRNA genes (n = 722) were obtained from SILVA [57] using the query "Caldiserica" (April 2015). Sequences were clustered using uclust [15] at 97% identity with the following settings "uclust –maxrejects 500 –id 0.97 –w 12 –stepwords 20 –usersort –maxaccepts 20 –stable_sort -uc", and for each cluster, only full- or near full-length 16S seed sequences were retained (>1400 nt, n = 58). HMTAb62 and HMTb66 [50] and SMC1's 16S rRNA sequences were added to these reference sequences, and were aligned with MAFFT [27] with settings "–genafpair –maxiterate 16 –phylip out -reorder". The phylogenetic tree was built with RAxML [68] with model GTR + GAMMA4 +I as specified by the following settings "-f a -p 12345 -s <ALIGNMENT> -x 12345 -n 100 -m GTRGAMMAI -n <OUTPUT>". The resulting tree recapitulated, with high bootstrap confidence (>75%), the 6 family-level clades of Rinke et al. [61]. Trees were visualized in the interactive Tree of Life (iTOL v3 [34]). Short 16S rRNA sequences (<1400 nt) were then

integrated into the tree using the maximum likelihood insertion algorithm in pplacer (see below).

Marker genes

A set of 120 bacterial single-copy marker genes were used to calculate maximum likelihood phylogenetic trees, using NCBI reference sequences, sequences from the recently published UBA genomes [54], and this study's 7 MAGs. (The SAG genome was not included in this analysis due to its low completeness (~25% [61])). The alignments from all marker genes were concatenated using HMMER v3.1.b2 [14], and the resulting alignment was used for tree building in FastTree v2.1.9 [55] with the WAG + GAMMA model and the approximately maximum likelihood method. Bootstrap values were estimated using genometreetk v0.0.35 (Parks, unpublished, https://github.com/dparks1134/GenomeTreeTk) after performing 100 FastTree iterations. Taxonomies were obtained using NCBI annotations and added to the tree using tax2tree [43]. Trees were visualized in ARB v6.0.6 [39], and refined in iTOL [34] before assembling panels and improving aesthetics in Inkscape.

Genetic discreteness: ANI and AAI calculation

Average nucleotide identity (ANI) [18] was calculated between all genome pairs in Jspecies v1.2.1 [60] with BLAST 2.2.28+ [1,10] using default parameters. ANI values <70% were not reported. To calculate average amino acid identity (AAI) [30], protein sequences were predicted for UBA genomes (UBA4822, UBA2182, UBA4770, UBA646, and UBA6126) with prokka 1.12 [65] in metagenomic mode with 'Bacteria' lineage. Protein sequences for SMCs were those obtained during functional annotation as described below. AAI was then calculated between pairs of protein sequence sets using script 'aai.rb' within the enveomics collection [62].

Phenotypic discreteness: genome annotation and metabolic pathway prediction

Feature prediction and initial functional annotation of population genomes was based on prokka v1.12 [65] using default parameters. Due to the novelty of these genomes, we then applied additional generalized, and specific, annotation approaches to maximize metabolic pathway predictions. Annotations based on RAST (parameters: classic annotation, RAST gene caller, FIGfam release 70) [2,52] and KEGG Orthology assignations through ghostKOALA (database genus_prokaryotes) [24,25] were compared to those of prokka, using Artemis [11]. Predicted pathways were visualized through KEGG Mapper v2.6. Genes and proteins of interest were then manually inspected by BLAST+ [10] searches against NCBI's non-redundant database (e-value $<1\times10^{-5}$, or $<1\times10^{-10}$ if there were numerous high-quality results), and potential operons were examined by comparing SMC MAGs to other genomes of interest using the doubleACT v2 webserver [http://www.hpa-bioinfotools.org.uk/pise/double_actv2.html], visualized in the Artemis Comparison Tool [11]. Successful alignments of the SMC MAGs to the *C. exile* genome region containing the *mbx* cluster suggested that its absence in the SMC MAGs was less likely to be due to incomplete assembly or binning. Functional profiles of genomes of interest were analyzed through principal components analysis of KEGG-KO annotations in a presence-absence matrix for all Caldiserica genomes or in both abundance and presence-absence matrix for SMCs (Supplementary Table 6).

Carbohydrate-active enzymes

Families of enzymes associated with glycosidic bond processing were identified in predicted protein sequences from each MAG using HMM models of carbohydrate-active enzymes (CAZy) families from the dbCAN database [79]. Profile searches were performed with HMMER 3.1b2 [14] and results were analyzed using thresholds recommended by dbCAN authors (e-value $<1\times10^{-3}$ or $<1\times01^{-5}$ and coverage >0.35 or >0.5 for small (<80 aa) and large (>80 aa) models, respectively).

Experimental validation

Metatranscriptome sequencing and analysis

Metatranscriptome sequencing and analyses were conducted as described previously [74] with minor adjustments. Briefly, select samples from 2010, 2011 and 2012 (Supplementary Table 3) underwent library preparation using the ScriptSeq Complete (Bacterial) low-input kit (Epicentre, Madison WA) with 240 ng of RNA as input. DNAse I (Roche, Pleasanton CA) was used to remove residual DNA from the RNA after extractions. Quality of RNA and libraries (during processing) was assessed with Agilent 2100 Bioanalyzer and Agilent 2200 Tapestation (Agilent Technologies, Santa Clara CA), and quantities were assessed by QubitR (ThermoFisher Scientific, Waltham MA). These samples were sequenced on 1/8th of a NextSeq (Illumina, San Diego CA) lane, following initial exploratory shallow runs on 1/11th of HiSeq (Illumina) and MiSeq (Illumina) lanes, producing an average of 18.1 Gb per sample (detailed in Supplementary Table 3).

SeqPrep (https://github.com/jstjohn/SeqPrep) was used to trim adaptor sequences, and contaminating PhiX sequences were removed by mapping and then removing reads aligning to the PhiX genome using BamM (Imelfort and Lamberton et al., http://ecogenomics.github.io/BamM/). SortMeRNA [33] facilitated the removal of non-coding RNA sequences (tRNA, 5S, 16S, 18S, 23S, 28S, tmRNA). The cleaned reads were then mapped against the 647 MAGs set from Woodcroft and Singleton et al. [74], with the addition of the SMC1 MAG from the earlier assembly. BamM 'filter' was used to ensure mapping identities and alignments of at least 95% and 75%, respectively. DirSeq (https://github.com/wwood/dirseq, internally using bedtools [58]) was used to calculate the counts of reads mapping to each gene feature, determined by Prokka [65]. Potential DNA contamination was accounted for by determining which genes had significant mapping in the sense direction compared to the antisense direction (p-value <0.05 in a one-sided binomial test) [74]. These genes were considered expressed, and read numbers were normalized by then subtracting the antisense reads from the number of sense reads. The normalized expression counts were used to determine relative expression of the genes as transcripts per million reads mapped (TPM [72]) to the protein coding genes (CDS regions, as determined by Prokka) (summarized in Supplementary Table 5).

Metaproteome characterization and analysis

Sixteen metaproteomes were previously generated from 22 samples collected at the central palsa-bog-fen thaw gradient sites in 2012 (three of the 16 were generated from pooled triplicate cores due to material limitation; Supplementary Table 3; [74]). Protein extraction, purification, and digestion as well as peptide fractionation and mass spectroscopy were previously described for these samples, which were used for a distinct community-level analysis in [74]. Here, the 16 metaproteome spectral files were searched to specifically examine potential SMC expression, as follows. Spectral input files were converted from .RAW to .dta using MSConvert (v3.0.9490 (2016-3-22)), wherein the correct monoisotope was attempted to be assigned to the mass spectra files. Spectral files were used as queries in a MSGFPlus [29] (v2018.01.30 (2018-1-30)) search against a comprehensive protein database containing predicted protein sequences from the seven SMC MAGs, all the metatranscriptomes, and the metagenomes from the same samples as the metaproteomes (detailed in Supplementary Table 3). Predicted proteins were dereplicated at 100% amino

acid identity using Protein Digestion Simulator (v2.2.6638 (2018-5-3), https://omics.pnl.gov/software/protein-digestion-simulator). MSGFPlus searches were performed with the following parameters: ±20 ppm parent mass tolerance; fully tryptic enzyme settings due to search FASTA file size; no additional modifications due to search size considerations; 25-part splitting of search FASTA for parallel processing, followed by merging the results using the best MSGF Spectral Probability (smallest value) as the arbiter of reported result; and decoy mode to allow FDR calculations to be performed, as described in [74]. An FDR cutoff of 1% was applied, leading to the recovery of peptide-spectrum matches with Q-values ≤ 0.00478.

In order to assess the likelihood of peptides arising from the SMC lineages, we processed initial search results with Protein Cover Summarizer (v1.3.6635 March 2, 2018, https://omics.pnl.gov/software/protein-coverage-summarizer) to detect all proteins in the reference database that shared a specific peptide. A peptide was assigned to SMC if (a) it matched solely to SMC MAG predicted proteins (n = 13 proteins, matched by 5 peptides; Supplementary Table 4), or (b) the number of times it was seen in the proteins predicted from SMC MAGs was at least equal to the number of times it was seen in the proteins predicted from the metagenomes (n = 27 proteins, matched by 11 peptides; Supplementary Table 4); this was done to account for conserved protein regions for which the peptides were likely (see below) to represent SMC (and *without* explicitly considering population heterogeneity of SMC lineages in these metagenomes). Of the total peptides shared between SMC MAGs and the metagenomes, ~16.4% (11 out of 67) occurred as frequently or more frequently in the SMC MAGs, and were thus assigned as SMC (Chi-Square test p-value = 4×10^{-89}). The detection of a SMC peptide in at least one sample was considered evidence of expression of a specific enzyme. A breakdown of the number of unique peptides matching to each protein assigned to SMC using this overall fairly conservative approach is provided in Supplementary Table 4.

Distribution of Caldiserica

Global distribution based on 16S rRNA genes

Sequences recovered from SILVA (see above) were inserted into the reference tree generated from full-length 16S sequences using Maximum Likelihood in pplacer v2.2.alpha17 as described above. Study metadata associated with each corresponding sequence (when available, see Supplementary Table 7) was then used to plot the temperature range and geographic distribution of these lineages using the R (v3.2.4) package ggplot2 v2.1.0.

Mire habitats

Relative abundances of SMC genomes in metagenomes from Stordalen Mire were calculated by competitively mapping the raw metagenome reads of each sample to the SMC MAGs and the dereplicated genomes (n = 647, 97% ANI dereplication) from Woodcroft and Singleton et al. [74]. Reads were mapped using BamM v1.7.3 'make' (https://github.com/Ecogenomics/BamM), which used BWA MEM [36] and samtools [35] to generate a BAM file. CoverM v142494d (Woodcroft unpublished, https://github.com/wwood/CoverM) was then used in 'genome' mode to calculate per-genome coverage. Reads were discarded for pairs with less than 90 bases mapping, with percent identity to the mapped region <97%, or where both reads in the pair were not mapped in a proper pair to the same contig. Genomes with <10% of their genome having 1× coverage were marked as having no coverage; otherwise coverage was calculated as the average number of reads mapped to each base in the genome, i.e the mean per-base coverage. To calculate a relative abundance for each genome in each metagenome amongst all cells (including those for which no genome was recovered), its mean per-base coverage was scaled to the fraction of total reads in that metagenome that mapped to all genomes.

Accession numbers

The SMC MAGs are available under NCBI accession numbers QXIS00000000, QXIT00000000, QXIU00000000, QXIV00000000, QXIW00000000, QXIX00000000, QXIY00000000, and metagenomes and metatranscriptomes used in this study are available under NCBI BioProject accession number PRJNA386568. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [71] partner repository with the dataset identifier PXD009096 and 10.6019/PXD009096.

Results and discussion

Microbial community profiling of a collapsed palsa feature in Stordalen Mire revealed abundant Caldiserica phylotypes (further detailed in *Ecological Characteristics*, below), distinct from *C. exile* and other described Caldiserica genomes. We therefore sought and recovered Caldiserica assemblies from a large metagenome-sequencing effort focused on the active layer of an adjacent thaw gradient [74]. Below we discuss and elaborate on each of the criteria proposed for description of uncultivated Bacteria and Archaea [32] for these Stordalen Mire Caldiserica (SMC) genomes.

Genome completeness and novelty

Seven high-quality (>80% complete, <5% contaminated; Supplementary Table 8) SMC MAGs were recovered, one from co-assembly ('SMC1'; unique to this study) and 6 from single-assemblies (and distinct from SMC1; SMCs 2–7; reported but not analyzed in [74]) (Supplementary Tables 2 and 8). SMC1 and SMC5 contained near full-length 16S rRNA sequences, that were 99.9% identical (Supplementary Fig. 1B).

The phylogenetic relationship of SMC1 to publicly available Caldiserica was inferred from aligned full-length and near-full-length 16S rRNA gene sequences (Fig. 1 A and Supplementary Fig. 2). SMC1 clustered with publicly-available environmental 16S rRNA sequences as a clade separate from (but monophyletic with) that of *C. exile*. The Stordalen Mire 16S rRNA amplicon OTUs were most closely related to the SMC1-containing clade, based on maximum likelihood insertion of the 6 distinct phylogenetic groups represented by the 74 collapsed palsa 16S rRNA amplicon OTUs (pie charts in Fig. 1A; Supplementary Fig. 2). Analysis of intra- and inter-clade variation in full-length 16S gene sequence identities (Fig. 1B) indicated the SMC1 and *C. exile* clades were appreciably distinct; average intra-clade identities were 92% and 89.3%, respectively, while the average inter-clade identity was ~79% (range 76–82%). Thus, although these sister clades were located internally within the larger Caldiserica phylogeny, they represented distinct phyla by the standard of <83% 16S rRNA identity [32], and corroborated below by additional criteria. Similarly, the low inter-clade identities of SMC to the next most related Caldiserica clades (73.7% to LF045, and 75.6% to WCHB1-02) supported SMC's phylum-level distinctness. The overall Caldiserica intra-clade identity averaged 78.5%, suggesting it might be considered a super-phylum.

To examine the divergence and phylogenetic placement of the SMC MAGs lacking 16S rRNA genes, a Caldiserica phylogeny was built using 120 bacterial single-copy marker genes (see Methods; Fig. 1C) from the 14 genomes available (the 7 SMC MAGs and 7 others). The resulting tree topology was consistent with the 16S rRNA-based tree, and all 7 SMC MAGs were monophyletic, and in a clade separate from *C. exile*, with strong bootstrap support (>95%). Of note, one UBA MAG (UBA4822) was a basal member of

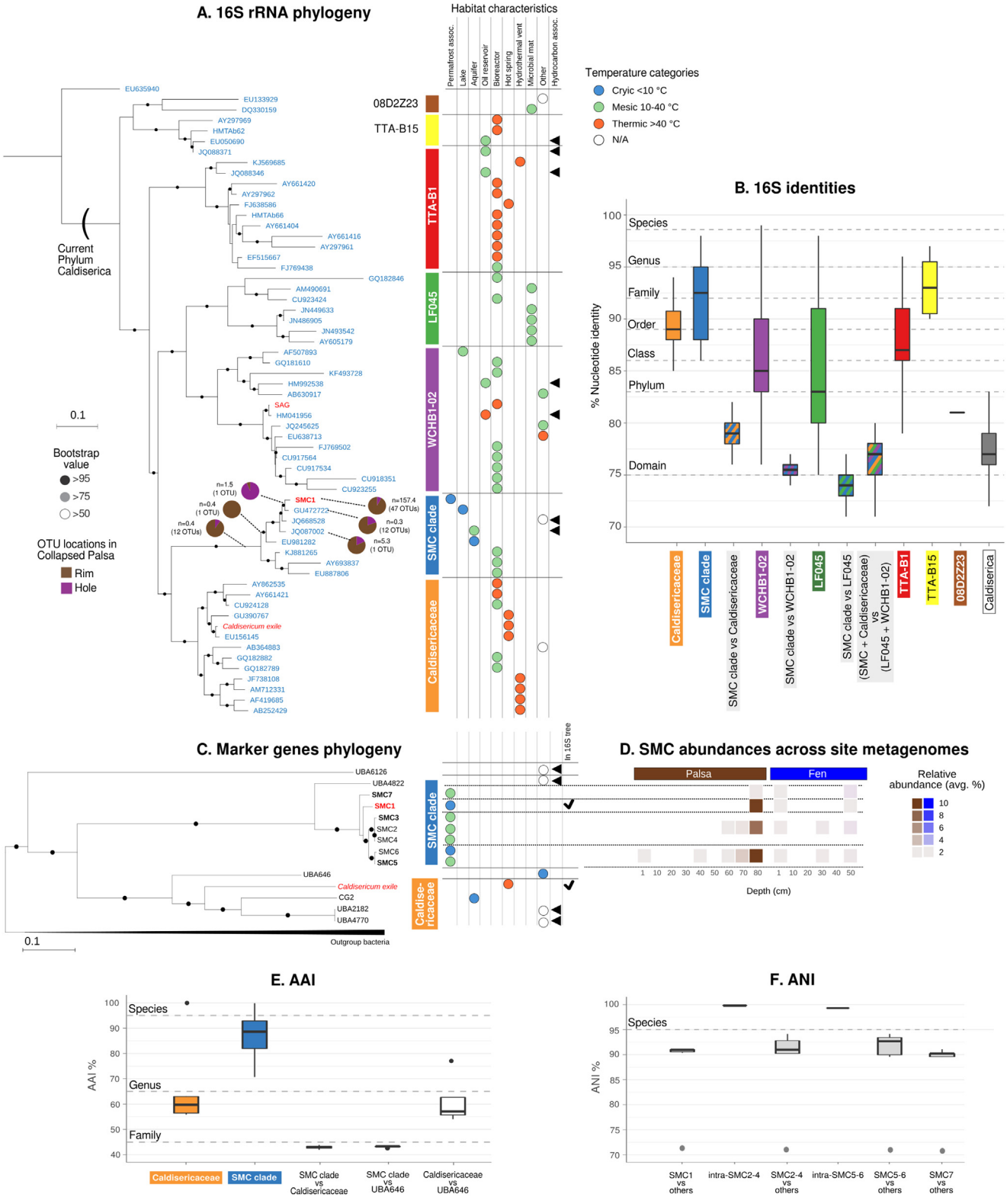

Fig. 1. Ecogenomic context of SMC MAGs.
(A) Maximum likelihood phylogenetic tree containing Caldiserica sequences from other 16S rRNA environmental surveys (blue text), and those from genomes (red text; isolate, single cell or metagenome-assembled genomes; the latter, bolded, is SMC1 from this survey). Overall Caldiserica topology was re-rooted to match that presented in Rinke et al. [61]. Current unofficial family-level designations from the SILVA database [57] are indicated to the right. Bootstrap values are shown for nodes with >50% support (100 replicates; black circles = >95, gray = >75, and white = >50). 16S rRNA gene amplicon sequences from this study were then aligned and inserted in this reference tree, and placement of their phylogenetic groups is denoted by dashed lines. The habitat distributions (brown for the palsa rim of the collapse feature, purple for the adjacent hole) of the amplicon phylogenetic groups are shown by the inset pie charts, with the normalized number of sequences and 97% OTUs comprising each phylogenetic group indicated adjacent. To the right of the tree are markers characterizing the origin habitat of the reference tree sequences; columns describe basic habitat types, circle colors denote temperature (orange = thermic, >40 °C; green = mesic, 10–40 °C; blue = cryic, <10 °C; white = N/A), and black triangles indicate hydrocarbon-associated habitats. (B) Pairwise 16S rRNA gene identities (y-axis) within and between each unofficial family in (A). Intra-clade identity bars are colored according to the arbitrary family colors used

the SMC clade. UBA4822 was assembled and binned as part of a large effort spanning >1500 public metagenomes [54]; its source was a metagenome from tailing pond sample in Alberta, Canada (SRA accession: SRX559915), and its phylogenetic distinctiveness was not established in Parks et al. [54] due the broad perspective of that study.

Average nucleotide identity (ANI) and amino acid identity (AAI) [30] were calculated for all genome pairs (Fig. 1E and F and Supplementary Fig. 1A), and further supported the distinctness of the SMC MAGs. Their AAIs of ~42–44% to the other genomes indicated greater than family-level divergence [31], with the exception of UBA4822, which at AAI ~70–72% appeared to derive from a distinct genus [31]. The SMC MAGs themselves comprised a single genus, based on intra-group AAI values of 87–99%, and 4 distinct species-level groups (with >95% AAI and >95% ANI denoting the same species [31]): SMC1, SMC2-4, SMC5-6, and SMC7 (Fig. 1E, F). This highlights the usefulness of genomic context, since at 99.9% 16S rRNA identity SMC1 and SMC5 (the only SMC MAGs with 16S rRNA genes; Supplementary Fig. 1B) would have been designated the same species, while their 90% ANI and 88% AAI clearly indicate they are not (Supplementary Fig. 1A). Below we describe the metabolism and ecology for the four SMC lineages as inferred via the SMC MAGs, 16S rRNA gene amplicons, and Stordalen Mire metaomes, in further support of their description as species.

Phenotypic discreteness

The metabolism and physiology of SMC were predicted (Fig. 2A) based on MAG annotations (Supplementary Table 9), accomplished in multiple ways due to the novelty of the genomes. Inferred metabolism was contrasted to that of *C. exile*, as the type species and only complete genome for Caldiserica, as discussed below. The distinctness of the SMC clade's metabolic potential was further supported by comparison to the 6 available incomplete Caldiserica genome assemblies and *C. exile*, via their KEGG-orthologous group (KO) profiles (Supplementary Fig. 3 and Supplementary Table 6). A caveat to the analysis of SMC's phenotypic discreteness is the incompleteness of their genomes (which are 83–90% complete); the strongest inferences of discreteness are therefore those of gained functions rather than lost ones, with respect to *C. exile*. The activity of this lineage (and further evidence for specific predicted metabolisms) was validated by examining expression data (metatranscriptomes and metaproteomes; Fig. 2B, Supplementary Fig. 7 and Supplementary Tables 4 and 5) using a conservative approach of only considering transcripts and peptides that uniquely matched one of the SMC MAGs (see Methods).

Predicted central carbon metabolism indicated that SMCs are anaerobic heterotrophs, like *Caldisericum*. The SMCs encoded a broad set of sugar transporters, nearly all glycolysis genes, and the non-oxidative pentose phosphate pathway (PPP) (Fig. 2, Supplementary Table 9). Metatranscriptomic evidence was found for two ABC transporters, two PTS enzymes and glucose-6-phosphate isomerase of glycolysis, and transketolase of PPP (Supplementary Table 5). Metaproteomic evidence was found for one ABC transporter and for fructose-bisphosphate aldolase (Supplementary Table 4). The absence of genes encoding the tricarboxylic acid cycle support an anaerobic lifestyle, consistent with the deeper depths at which this lineage was often found (see *Ecological Characteristics* section below). Also in accordance with an anaerobic metabolism are the use of ferredoxin for electron exchange, and the pyruvate processing to acetyl-CoA through anaerobic enzymes (por and pfl) instead of pyruvate dehydrogenase (pdh), which is absent. In fact, two aldehyde ferredoxin oxidoreductases were found in the metatranscriptomes (Supplementary Table 5).

Unlike *Caldisericum*, in which fermentation has not been observed nor its genetic potential encoded [47], the SMC clade is predicted to ferment acetyl-CoA to acetate and CO_2 using Ack-Pta enzymes (Fig. 2; Supplementary Table 9). Fermentation by SMCs may produce propanoate or ethanol, based on the presence of enzymes that convert acetyl-CoA to acetate and may be able to convert propanoyl-CoA to propanoate, and the presence of aldehyde/alcohol dehydrogenases (which were also found in the metatranscriptomes; Supplementary Table 5). The potential for generating these fermentation products makes SMC a candidate provider of metabolites to other community members, including methanogens.

SMC MAGs encode enzymes for interconverting several amino acids or their degradation products to tricarboxylic acid cycle intermediates (fumarate, oxaloacetate, 2-oxoglutarate), and others (pyruvate, 3-phosphoglycerate) that can be used for biosynthesis or used as substrates for fermentation (Fig. 2, Supplementary Table 9), enhancing the importance of pyruvate in this organism. The presence of transporters for amino acids and oligopeptides (eight were detected in the metatranscriptomes and one in the metaproteomes; Supplementary Tables 4 and 5), along with the absence of enzymes for the tricarboxylic acid cycle, support this degradation pathway. Due to the presence of a similar theoretical interconnection of pathways, amino acid fermentation has been suggested as a life strategy for other members of the phylum as part of large-scale MAG analyses [40,50].

SMC MAGs encoded the glycine cleavage system that converts glycine to 5,10-methylene-tetrahydrofolate (Supplementary Table 9). This latter compound is an intermediary of the Wood-Ljungdahl pathway for carbon fixation, however the critical enzymes for that pathway were not present in SMC MAGs. Instead, the pathway could work in the reverse direction (oxidative) to degrade 5,10-methylene-tetrahydrofolate to formate (generating 1 ATP molecule) and ultimately to hydrogen and CO_2, which would then be excreted (and could then be substrates for hydrogenotrophic methanogenesis). SMC may be processing glycine, and via glycine also serine and threonine, through this pathway; enzymes for those transformations were also present.

Given the predicted fermentative metabolism and the availability of sugar transporters, we analyzed SMC's potential for degrading plant compounds and oligosaccharides by annotating the carbohydrate-active enzymes (CAZys). On average, 60 protein sequences per MAG had a predicted CAZy domain, 14–26% of which were annotated as glycoside hydrolases (Fig. 2C, Supple-

in (A), while inter-clade identity bars are hatched using the colors of the compared families; gray was used for the identity bar spanning the entire phylum. The currently accepted 16S rRNA gene identity thresholds for each taxonomic level are indicated [32]. (C) A phylogenetic tree of 120 concatenated bacterial genes was created with the genomes shown (designated by their internal "UBA_XXX" IDs for UBA and our SMC MAGs) as well as 94 outgroup lineages from different bacterial phyla. Circles indicate 100% confidence bootstraps. Clade relationships are indicated to the right, as in (A). To the right of the tree are markers characterizing the origin habitat of the sequences, as in (A). (D) The relative abundances of the dereplicated SMC MAGs (using one MAG for each species-level grouping) in 214 Stordalen Mire metagenomes reported in [74], based on competitive read recruitment to all dereplicated MAGs from that paper and SMC1 (which was not in that paper). Color intensity indicates an average relative abundance (see Methods) for each sampled depth and habitat (brown = palsa, blue = fen), with an absence of color indicating no reads recruited; SMCs were considered absent from the bog habitat. (E) Pairwise average amino acid identity (AAI; y-axis) within the SMC clade and *Caldisericaceae* (blue and orange respectively), and between several clades (white). (F) Pairwise average nucleotide identity (ANI) comparisons among SMC clade genomes (including UBA4822) support differentiation of four cohesive species within SMC MAGs. ANI values against UBA4822 MAG are depicted as grey points.

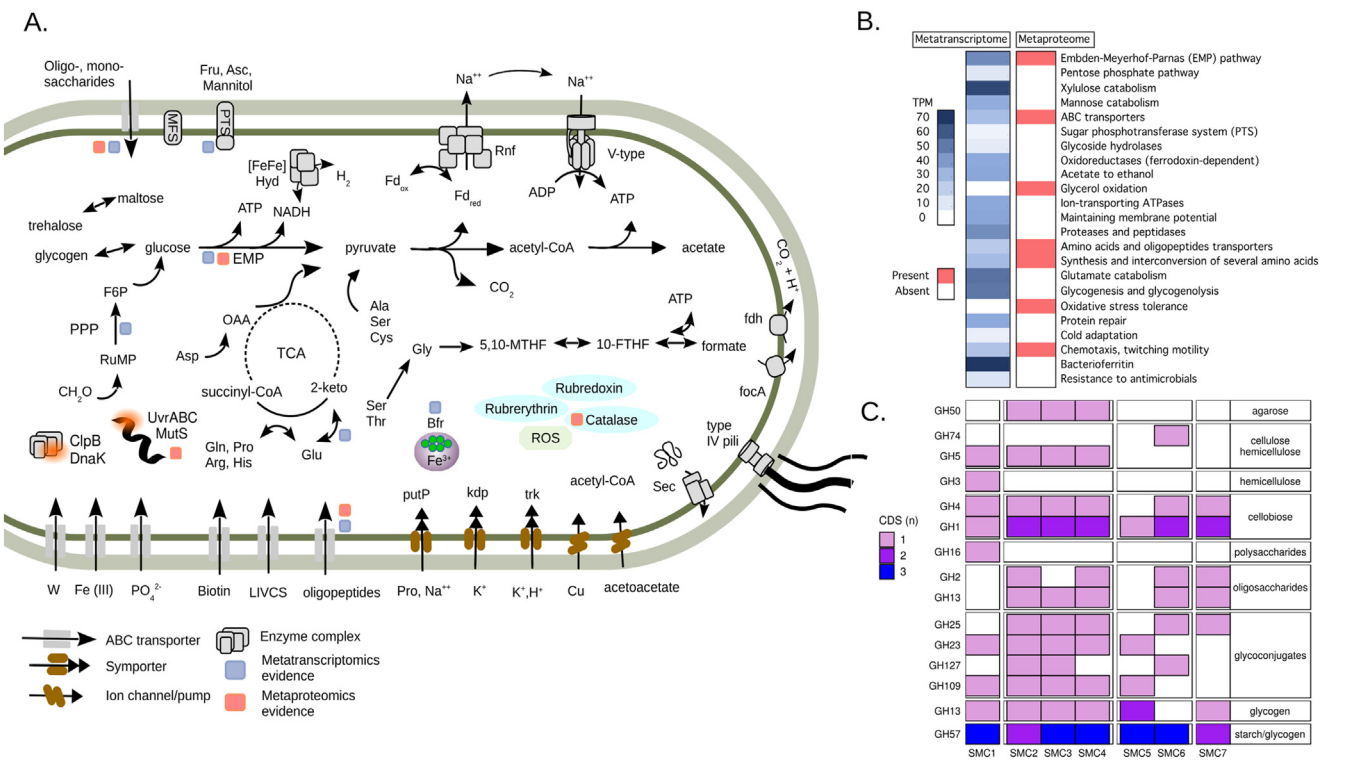

Fig. 2. The metabolism and activity of SMC.
(A) Cell cartoon of key SMC metabolisms based on genomic inferences. Evidence of expression of these metabolisms from metatranscriptomes and metaproteomes is denoted by blue and red squares respectively, indicating at least one expressed gene in the pathway, per Methods, and see panels B and C. Arrows represent reactions, modules or pathways, and dashed lines indicate absent enzymes. DNA replication and information processing have been omitted for clarity. Fru, fructose; Asc, ascorbate; EMP, Embden–Meyerhoff pathway; PTS, sugar phosphotransferase system; MFS, Major Facilitator Superfamily; F6P, fructose-6-biphosphate; RuMP, ribulose monophosphate; PPP, non-oxidant reactions of the pentose phosphate pathway; CoA, coenzyme A; OAA, oxaloacetate; LIVCS, branched chain amino acid transporter; Fd, ferredoxin; TCA, tricarboxylic acid cycle; 5,10-MTHF, 5,10-methylene tetrahydrofolate; 10-FTHF, 10-formyltetrahydrofolate; ROS, reactive oxygen species; Bfr, bacterioferritin. (B) Expression of the major metabolisms depicted in (A) and discussed in the text. Magnitude of expression in the transcriptome is indicated by color intensity corresponding to transcripts per million reads mapped (TPM, averaged across genes in pathway), and is shown solely for the fen habitat as that was the only one in which SMC transcripts were observed. Metaproteome evidence of pathway expression is indicated as presence/absence, and combined across all three habitats (palsa, bog and fen), due to the limited number of SMC peptides identified. (C) Glycoside hydrolase (GH) profile of the seven SMCs, characterized by the number of CDSs in each GH family; broad substrates category for each GH are shown at the right.

mentary Table 10). These had a variety of labile or non-recalcitrant substrates like cellulose, hemicellulose, cellobiose and oligosaccharides, while CAZys for processing recalcitrant substrates (e.g. lignocellulose) were not found. Metatranscriptomic evidence was found for beta-glucosidase A (Supplementary Table 5).

While *C. exile* is non-motile [48], the gene content of the SMCs hint at the potential for motility. Three of the 4 SMC clades encoded the gliding motility protein MglA (absent in *C. exile*), and all SMCs encoded the major type IV pilin protein PilA (which *C. exile* lacks), and many encoded PilB, PilT, and other pilus-associated proteins. The type IV pilus can be used for attachment, biofilm formation, DNA uptake, extracellular electron transfer, and twitching motility, a non-flagellar form of microbial movement. In addition, some evidence for twitching motility/chemotaxis was seen in the metatranscriptomes and metaproteomes (*cheR2* and ChvE, respectively).

C. exile requires the reduction of thiosulfate, sulfite or elemental sulfur for growth [47]; while there are several pathways that can confer these abilities, their genetic basis in *C. exile* is unknown. It has been proposed to be linked to the mbx-cluster [49], which performs sulfur reduction-related electron transport in the archaeon *Pyrococcus furiosus* [64] and is present in *C. exile* [49]. No known gene for reducing sulfur compounds, nor any of the 13-gene mbx cluster, was found in SMC MAGs (Supplementary Table 9). In addition, sulfate concentrations in the collapsed palsa habitat are very low (0.03–0.05 μM, data not shown), while *C. exile* was isolated from a hot spring with 0.6–0.8 mM sulfate, and was not found in a nearby site depleted in sulfate (0.005–0.15 mM) [47].

Electron transport in SMC was encoded by an Rnf complex, which couples ferredoxin and NAD^+ oxidoreduction to Na^+ or H^+ translocation, which is then used to drive ATP synthesis through a V/A-ATPase [6]. The metatranscriptomes showed evidence for two additional ATPases as well as two enzymes involved in spermidine biosynthesis (S-adenosylmethionine decarboxylase and polyamine aminopropyltransferase; Supplementary Table 5) which synchronizes Na+ ATPases to maintain membrane potential and control intracellular pH. Gene organization of the Rnf complex has been found in two operon structures: *rnfABCDE* in the N-fixer *Rhodobacter* where it was first described [63], and *rnfCDGEAB* in fermentative *Clostridium* strains and in SMC [5,9]. SMC MAGs encode another protein involved in oxidoreduction and electron flow, a non-membrane-bound Hnd hydrogenase identified as [FeFe] Group A3, capable of NAD^+ reduction or, in the opposite direction, of hydrogen production.

SMC also encodes several cryoprotection strategies, as expected given its higher relative abundance at colder depths (see *Ecological Characteristics* below). Psychrophiles encode a variety of cold adaptations, with no single strategy shared by all [7,22]. One challenge associated with cold lifestyles is nutrient and energy storage, since low temperatures slow biochemical processes, including nutrient uptake. SMC encoded enzymes for the synthesis and degradation of glycogen, a glucose-based polymer commonly used by microbes for carbon and energy storage. Freezing temperatures also make reactions difficult, by limiting water availability and concentrating salts, decreasing overall fluidity. Cellular osmoprotection can be achieved by the accumulation of compatible solutes, such as

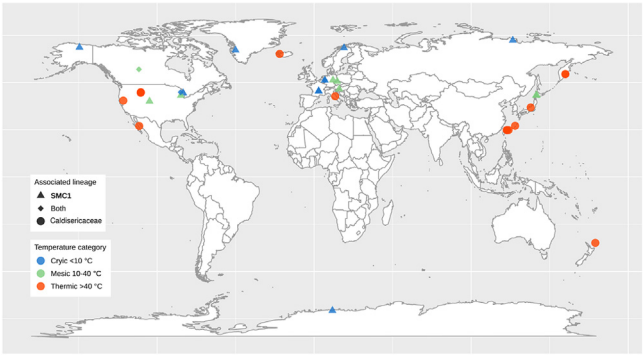

Fig. 3. Global distribution of SMC and Caldisericaceae lineages.
Locations and characteristics of study sites and sample sources that produced 16S rRNA sequences assigned to *Caldisericaceae* (circles), SMC (triangles), or both lineages (diamond) (see Methods), from cryic (<10 °C, blue), mesic (10–40 °C, green) and thermic (>40 °C, red) environments. Locations and sample characteristics were recovered from information in Genbank sequence records and associated publications (see Supplementary Table 7 for details).

sugars, polyols, amino acids or betaine. The SMCs encode differing pathways for compatible solutes, including trehalose, proline, and potassium. Another challenge of life at low temperature is the higher solubility of oxygen, leading to reactive oxygen species (ROS) generation. The enzymes rubrerythrin, rubredoxin, and catalase C are used by anaerobes to protect against ROS [8,64], and were encoded by SMC, but not *C. exile*. Metaproteomic evidence was found for catalase C.

Furthermore, the genetic divergence *among* the 4 SMC lineages (Fig. 1) appears not to be functionally neutral. The SMC lineages had distinct carbon substrate utilization potentials based on their CAZy profiles (Fig. 3C), with the highest diversity of overall glycoside hydrolases present in SMC2-4 and the lowest in SMC7, while the diversity of potential cellulose utilization peaked in SMC1, then decreased from SMC2-4 to SMC5-6 and SMC7. SMC2-4 were also unique in encoding agarose utilization, with a GH that matches the β-agarase of an experimentally verified agarose-degrading bacteria (*Saccharophagus degradans* 2–40; nr-blastp, qcov = 40, e-val <1e–15, iden 26%) [28], as well as that of Antarctic sediment psychrophile (*Pseudoalteromonas* sp. NJ21, nr-blastp, qcov 46, eval = 5e–23, ident 25%) [37]. While agarose is mainly derived from marine red algae, there could be a source in this environment, or the GH could be degrading another sugar with a similar β-1,4-linkage. Moreover, SMC2-4 encoded proteins with the potential to eliminate formaldehyde (HxlA and HxlB), a highly toxic byproduct of glycolysis (and of glycine degradation) that reacts irreversibly with proteins and DNA. HxlA and HxlB allow conversion of formaldehyde for use in the Ribulose Monophosphate Pathway, which eventually leads its products to glycolysis.

In addition to their distinct carbon substrate utilization potentials, the SMC lineages have further inferred phenotypic distinctions (Supplementary Figs. 3B and C). Of note among these are differences in cellular osmoprotection strategies relevant to life at extreme temperatures. SMC1 encodes TreS for the production of trehalose (a disaccharide, and a compatible solute) from maltose (*C. exile* in contrast encodes trehalose production directly from glucose). SMC1 also encodes a Na+/proline symporter, which can also indirectly be considered an osmoprotective strategy since proline is a compatible solute, and has been seen in several psychrophilic genomes [7]. In contrast, the other 3 SMC lineages encoded alternate potential osmoprotective strategies such as potassium uptake (Trk) and potassium-transporting ATPase (Kdp). Of additional relevance to life at low temperature, SMC5-6 encodes and expresses catalase C (which protects against ROS, in addition to the rubrerythrin and rubredoxin encoded by all SMCs). Lastly SMC5-6 and SMC7 also encode the CheBR fusion protein involved in chemotaxis and aerotaxis. Chemosensing is consistent with the potential motility of the SMC clade, and negative aerotaxis would be beneficial for them as anaerobes. A major caveat to identifying potential metabolic differences among the 4 SMC lineages is that all the MAGs, while high quality (estimated >80% completeness, <5% contamination), are incomplete. Furthermore, genome alignments (which would aid significantly in assessing the robustness of gene-content differences, specifically for gene absences) are precluded by the large number of contigs that comprise each MAG. This highlights the usefulness of follow-up studies to close the genomes and/or isolate these lineages for direct phenotypic characterization.

Collectively, annotations and gene expression data suggest that the SMC clade is an anaerobic heterotrophic group that is phenotypically discrete from *Caldisericum* by predictions of (i) fermentation, including the predicted production of acetate, ethanol, and possibly propanoate; (ii) a lack of sulfur reduction; (iii) a saccharolytic lifestyle; (iv) distinct adaptations to extreme temperature; and (v) possible motility. The 4 SMC lineages are distinguished by carbon utilization, cryoprotection, detoxification of formaldehyde and ROS, and possible chemosensing.

Ecological characteristics

The distribution of SMC lineages in Stordalen Mire microbial communities was examined in two complementary approaches: by 16S rRNA gene amplicon-based characterization of a palsa permafrost collapse feature (Supplementary Methods), and by metagenome-based characterization of an adjacent permafrost thaw gradient from palsa to bog to fen habitats, primarily of the active layer (the seasonally-thawed ground overlying the permafrost). The palsa is a dry bulk-aerobic habitat through the active layer (likely with anaerobic microsites as in most soil habitats), while the palsa collapse feature and bog have fluctuating often mid-depth water tables, and the fen is fully inundated.

The 74 recovered Caldiserica OTUs in the 16S amplicon dataset were all closely related to the SMC MAGs, and 60 were closely related to SMC1 (Fig. 1A). These Caldiserica lineages rose to high abundances in the palsa Rim of the feature, in the permafrost depths (Supplementary Figs. 4 and 5), to dominate the community composition at up to 70% relative abundance. Notably, while such high Caldiserica abundances are atypical in other reported observations of them, a new study from a nearby mire found similar dominance – up to 60% relative abundance – of Caldiserica in the permafrost [46]. Although that study also used 16S rRNA gene amplicon sequencing, it employed different primers targeting a different variable region, suggesting that the high Caldiserica abundances observed in the two datasets are not due to primer bias. The sole other report of such high relative abundances of Caldiserica was up to 83% in an Antarctic lake sediment [66].

Caldiserica were absent from the active layer of the palsa rim of the collapse feature, which helped drive the differentiation of permafrost from active layer communities (Supplementary Fig. 6) along with more Firmicutes in the permafrost and fewer Planctomycetes, Dormibacteria (formerly AD3) and Eremiobactereaeota (formerly WPS-2) (all but the latter being known permafrost or active-layer clades; [4,17,22,78]). Post-thaw, Caldiserica became much less abundant in the former permafrost depths in the thawed deep hole than rim, as were Firmicutes, replaced by active layer phyla. Overall however, in spite of the thaw-associated penetrance of active layer phyla deeper into the peat, at finer phylogenetic resolution, recent thaw did not result in the deeper peat communities more closely resembling those of the overlying active layer (Supplementary Fig. 6). The environmental drivers of the community shifts were pH (R^2 = 0.93, p-value <0.001), %C (R^2 = 0.76, p-value <0.001), %N (R^2 = 0.47, p-value <0.001) (Supplementary Fig. 6). These results

suggest that microbes are differentially constrained depending on the layer they inhabit: the active layer community may be more limited by the lower pH at this depth (which limits acid-intolerant microbes), while the permafrost community dominated by Caldiserica may be more constrained by organic matter availability, as the higher mineral content at this depth results in relative depletion and possible mineral adsorption of carbon and nitrogen [19,73,77]. SMC1's possession of a greater diversity of cellulose-utilization genes than other SMCs is consistent with this.

In an adjacent permafrost thaw gradient from palsa to bog to fen, across 214 almost exclusively active-layer metagenomes (from 2010 to 2012, and reported in [74]), SMC was a sporadic community member. SMC MAGs conservatively occurred in 19 palsa and fen metagenomes, ranging in collective relative abundance from <0.1% to ~59% in a single sample based on recruitment analysis (Fig. 1D, which presents averages of sample replicates at each depth). Notably, the clade's peak metagenome-based relative abundance occurred in the permafrost (where it was also more consistently present than in the active layer), mirroring its distribution in the collapse palsa feature 16S rRNA amplicon profiles, and was virtually identical to the peak amplicon-based relative abundance of ~56% at the same depth (80 cm; it rose to ~73% at 90 cm, a depth lacking a metagenome). In addition, the MAGs used to assess clade abundance in the metagenome may represent only a subset of the local amplicon-observed SMC genetic diversity (Fig. 1A, Supplementary Figs. 4 and 5; only 2 SMC MAGs contained 16S rRNA genes), such that additional unmapped metagenomic reads could have belonged to the SMC clade. Across 24 metatranscriptomes, SMC transcripts were only observed in the fen samples; from 16 metaproteomes, SMC peptides were observed in all 3 habitats but at low abundances. Notably, community expression profiles were only available for active layer samples, so while they demonstrate that SMC is active at this site, they do not clarify its activity at the depths where it achieves community dominance.

The four SMC lineages had distinct ecological profiles. SMC5-6 had the broadest distribution, appearing sporadically across the depth profile of the palsa and fen; this might be aided by their encoding and expressing catalase C for increased ROS protection.

Table 1
Digital Protologue table for Stordalen Mire Caldiserica (SMC) 4 proposed novel Species and 1 novel Genus.

Taxonumber	CA00044	CA00030	CA00031	CA00032	CA00033
Species name (give the binomial species name)		Ca. Cryosericum septentrionale	Ca. Cryosericum hinesii	Ca. Cryosericum odellii	Ca. Cryosericum terrychapinii
Genus name	Cryosericum	Cryosericum	Cryosericum	Cryosericum	Cryosericum
Specific epithet		septentrionale	hinesii	odellii	terrychapinii
Genus etymology	Cryosericum (Cry.o.se'ri.cum. Gr. neut. n. kryos, icy cold, frost; L. neut. n. sericum, silk; N.L. neut. n. Cryosericum, silk from a cold environment).	Cryosericum (Cry.o.se'ri.cum. Gr. neut. n. kryos, icy cold, frost; L. neut. n. sericum, silk; N.L. neut. n. Cryosericum, silk from a cold environment).			
Type species of the genus	Ca. Cryosericum septentrionale	Ca. Cryosericum septentrionale	Ca. Cryosericum septentrionale	Ca. Cryosericum septentrionale	Ca. Cryosericum septentrionale
Taxonumber of the type species	CA00030	CA00030	CA00030	CA00030	CA00030
Genus status	Gen. nov.	Gen. nov.			
Species etymology		septentrionale (sep.ten.tri.o.na'le. L. neut. adj. septentrionale, northern)	hinesii (hines'i.i. N.L. gen. n. hinesii, in honor of biogeochemist and IsoGenie project member Mark E. Hines, 1950–2018).	odellii (o.dell'i.i. N.L. gen. n. odellii, in honor of cell, developmental and computational biologist and polymath Garrett M. Odell, 1943–2018).	terrychapinii (ter.ry.cha.pi'ni.i. N.L. gen. n. terrychapinii, in honor of ecosystem ecologist F. Stuart "Terry" Chapin, III).
Species status		Sp. nov.	Sp. nov.	Sp. nov.	Sp. nov.
Authors	Martinez Miguel A., Woodcroft Ben J., Ignacio Espinoza Julio C, Zayed Ahmed A., Singleton Caitlin M., Boyd Joel A., Li Yueh-Fen, Purvine Samuel O., Maughan Heather, Hodgkins Suzanne B., Anderson Darya, Sederholm Maya, Temperton Ben, Bolduc Benjamin, IsoGenie Project Coordinators, Saleska Scott R., Tyson Gene W., Rich Virginia I.				
Title	Discovery and ecogenomic context of a global Caldiserica-related phylum active in thawing permafrost, Candidatus Cryosericа phylum nov., Ca. Cryosericia class nov., Ca. Cryosericales ord. nov., Ca. Cryosericaceae fam. nov., comprising the four species Cryosericum septentrionale gen. nov. sp. nov., Ca. C. hinesii sp. nov., Ca. C. odellii sp. nov., Ca. C. terrychapinii sp. nov.				
Journal	Systematics and Applied Microbiology				
Corresponding author	Rich Virginia I	Rich Virginia I	Rich Virginia I	Rich Virginia I	Rich Virginia I
E-mail of the corresponding author	rich.270@osu.edu	rich.270@osu.edu	rich.270@osu.edu	rich.270@osu.edu	rich.270@osu.edu
Submitter	Miguel Martinez-Mercado	Miguel Martinez-Mercado	Miguel Martinez-Mercado	Miguel Martinez-Mercado	Miguel Martinez-Mercado
E-mail of the submitter	marmigues@gmail.com	marmigues@gmail.com	marmigues@gmail.com	marmigues@gmail.com	marmigues@gmail.com
Designation of the type MAG	SMC1	SMC1	SMC3	SMC5	SMC7
Metagenome accession number	MAG	MAG	MAG	MAG	MAG
MAG/SAG Accession number [RefSeq]	QXIY00000000	QXIY00000000	QXIW00000000	QXIU00000000	QXIS00000000
Genome status		Draft	Draft	Draft	Draft
Genome size	1400–2100	2095.932	1936.32	1985.68	1487.05
GC mol %	56.5–57.4	57.4	57.2	56.9	57.3
Country of origin	Sweden	Sweden	Sweden	Sweden	Sweden
Region of origin	Abisko	Abisko	Abisko	Abisko	Abisko
Source of sample	Soil (ENVO: 00001998)	Soil (ENVO: 00001998)	Soil (ENVO: 00001998)	Soil (ENVO: 00001998)	Soil (ENVO: 00001998)
Sampling date	2011-06-15	2011-06-15	2011-08-16	2011-08-16	2012-08-23

Table 1 (*Continued*)

Taxonumber	CA00044	CA00030	CA00031	CA00032	CA00033
Geographic location	Stordalen Mire	Stordalen Mire	Stordalen Mire	Stordalen Mire	Stordalen Mire
Latitude	68°21'12.6"N	68°21'12.6"N	68°21'12.6"N	68°21'11.9"N	68°21'12.2"N
Longitude	19°02'49.6"E	19°02'49.6"E	19°02'48.1"E	19°02'47.8"E	19°02'48.5"E
Altitude	363	363	363	363	363
Temperature of the sample [in celsius degrees]	<0–14	<0–14	10.4	10.4	10.8
pH of the sample	3.3–6.1	3.3–5.9	5.5	5.5	6.1
Salinity of the sample [in percentage %]	0	0	0	0	0
Relationship to O2	Anaerobe	Anaerobe	Anaerobe	Anaerobe	Anaerobe
Energy metabolism	Chemoorganotroph	Chemoorganotroph	Chemoorganotroph	Chemoorganotroph	Chemoorganotroph
DNA extraction method		PowerMax Soil DNA Isolation kit (MoBio) as described in R. Mondav, B.J. Woodcroft, et al. Discovery of a novel methanogen prevalent in thawing permafrost, Nat Commun, 5 (2014) 3212.	PowerMax Soil DNA Isolation kit (MoBio) as described in R. Mondav, B.J. Woodcroft, et al. Discovery of a novel methanogen prevalent in thawing permafrost, Nat Commun, 5 (2014) 3212.	PowerMax Soil DNA Isolation kit (MoBio) as described in R. Mondav, B.J. Woodcroft, et al. Discovery of a novel methanogen prevalent in thawing permafrost, Nat Commun, 5 (2014) 3212.	PowerMax Soil DNA Isolation kit (MoBio) as described in R. Mondav, B.J. Woodcroft, et al. Discovery of a novel methanogen prevalent in thawing permafrost, Nat Commun, 5 (2014) 3212.
Assembly		Replicate different samples	1 sample	1 sample	1 sample
Sequencing technology		Illumina HiSeq200 (100 bp), or Illumina NextSeq lane (150 bp)	Illumina HiSeq200 (100 bp), or Illumina NextSeq lane (150 bp)	Illumina HiSeq200 (100 bp), or Illumina NextSeq lane (150 bp)	Illumina HiSeq200 (100 bp), or Illumina NextSeq lane (150 bp)
Binning software used		BamM (https://github.com/Ecogenomics/BamM)	MetaBAT (Kang DD, et al., 2015. PeerJ)	MetaBAT (Kang DD, et al., 2015. PeerJ)	MetaBAT (Kang DD, et al., 2015. PeerJ)
Assembly software used		CLC Genomics Workbench version 4.4 (CLC Genomics)	CLC Genomics Workbench version 4.4 (CLC Genomics)	CLC Genomics Workbench version 4.4 (CLC Genomics)	CLC Genomics Workbench version 4.4 (CLC Genomics)
Habitat	Palsa (ENVO: 00000489), permafrost (ENVO: 00000134), thermokarst depression (ENVO: 03000084), thermokarst (ENVO: 03000085), permafrost thawing process (ENVO: 03000086), fen (ENVO: 00000232), wetland ecosystem (ENVO: 01001209), freshwater wetland ecosystem (ENVO: 00000243), peatland (ENVO: 00000044), palsa mire (ENVO: 00000188)	Palsa (ENVO: 00000489), permafrost (ENVO: 00000134), wetland ecosystem (ENVO: 01001209), freshwater wetland ecosystem (ENVO: 00000243), peatland (ENVO: 00000044), palsa mire (ENVO: 00000188)	Thermokarst depression (ENVO: 03000084), thermokarst (ENVO: 03000085), permafrost thawing process (ENVO: 03000086), fen (ENVO: 00000232), wetland ecosystem (ENVO: 01001209), freshwater wetland ecosystem (ENVO: 00000243), peatland (ENVO: 00000044), palsa mire (ENVO: 00000188)	Thermokarst depression (ENVO: 03000084), thermokarst (ENVO: 03000085), permafrost thawing process (ENVO: 03000086), fen (ENVO: 00000232), wetland ecosystem (ENVO: 01001209), freshwater wetland ecosystem (ENVO: 00000243), peatland (ENVO: 00000044), palsa mire (ENVO: 00000188)	Thermokarst depression (ENVO: 03000084), thermokarst (ENVO: 03000085), permafrost thawing process (ENVO: 03000086), fen (ENVO: 00000232), wetland ecosystem (ENVO: 01001209), freshwater wetland ecosystem (ENVO: 00000243), peatland (ENVO: 00000044), palsa mire (ENVO: 00000188)
Biotic relationship	Free-living	Free-living	Free-living	Free-living	Free-living
Known pathogenicity	None	None	None	None	None

SMC2-4 was somewhat more restricted, still spanning the fen's depths but occurring only in the permafrost of the palsa. SMC1 and SMC7 had the most restricted distributions, occurring only in the deepest palsa permafrost, and bimodally in the surface and deepest fen; unlike SMC1 which reached high abundance in the permafrost, SMC7 was always at low abundance. Interestingly, SMC5-6 and SMC2-4 were also at high abundance in the deepest palsa permafrost. The SMC lineages' varying cryoprotection strategies may contribute to their prevalence in permafrost, and to their niche differentiation in this harsh environment.

Beyond Stordalen Mire, the global geographic distribution of SMC was explored, and compared to that of *C. exile*, by retrieving all available environmental sources and geopositions for 16S sequences (partial or full) within the *C. exile* and SMC clades. This revealed that SMC has been found in a variety of environments around the world (at <10% of 16S amplicon-based abundances except for an Antarctic lake [66], where it became dominant a meter below the sediment surface), with a distinct preference for anaerobic environments with abundant carbon and low temperatures (Fig. 3 and Supplementary Table 7), consistent with our sampling site. SMC-like sequences have not been found in extremely hot environments like hydrothermal vents or terrestrial hot springs [47,48]. These habitat observations confirm the apparent hyperthermophilic adaptation of the *C. exile* clade, and support our predictions that SMC degrades and ferments organic compounds such as oligosaccharides or peptides.

Conclusion

The use of high-throughput sequencing and genome analysis tools will further enable discovery of microbial lineages to reveal the true diversity currently hidden as microbial dark matter. As technologies continue to improve with longer read lengths and bioinformatic predictions, we can begin to more system-

atically explore the microbial drivers of all ecosystems. Within the Stordalen Mire peatland ecosystem, permafrost thaw is dramatically altering the landscape, shifting plant and microbial communities as well as belowground carbon composition and net carbon gas emissions. While studying these climate-driven microbial processes, we have discovered seven MAGs from a poorly-described clade, which genetic analysis indicates to be a novel phylum. Although we lack a cultivated member, the assembly of MAGs has provided a view into the metabolism and ecology of this underexplored clade, which as a saccharolytic fermenter may help channel plant-derived carbon – including old forms preserved in permafrost – to other microbial community members, including methanogens. A key next step will be targeted investigation of the gene expression of permafrost communities, and of *in situ* substrate diversity, to solve the mystery of this novel phylum's remarkably high dominance in the permafrost of this site.

Based on the genetic and phenotypic distinctness presented above, we propose the Candidate phylum Cryosericota, which is differentiated from its sister clade Caldiserica by ∼79% 16S identity and ∼43% AAI, and from the other major Caldiserica clades by even lower 16S identities. Within this phylum, we propose the Candidate order Cryosericales ord. nov., Candidate family Cryosericaceae fam. nov., and Candidate genus Cryosericum gen. nov. comprising all the 16S sequences recovered from Stordalen Mire and 4 new Candidate species: the type species Cryosericum septentrionale sp. nov. for SMC1, Cryosericum hinesii sp. nov. for SMC2-4, Cryosericum odellii sp. nov. for SMC5-6, and Cryosericum terrychapinii sp. nov. for SMC7. The Digital Protologues for formal proposals of these taxa are given below (Table 1), with the corresponding Taxon numbers CA00030-CA00033.

Description of **Cryosericaceae fam. nov.**

Cryosericaceae (Cry.o.se.ri.ca'ce.ae. N.L. neut. n. *Cryosericum* a candidate genus; -aceae ending to denote a family; N.L. fem. pl. n. *Cryosericaceae* the *Cryosericum* candidate family).

Bacteria in the family Cryosericaceae are anaerobic heterotrophs living mainly in cryic (<10 °C) or mild mesic environents. The type genus is Cryosericum.

Description of **Cryosericales ord. nov.**

Cryosericales (Cry.o.se.ri.ca'les. N.L. neut. n. *Cryosericum* a candidate genus; -ales ending to denote an order; N.L. fem. pl. n. *Cryosericales* the *Cryosericum* candidate order).

The description is the same as that for the family Cryosericaceae. The type genus is Cryosericum.

Description of **Cryosericia class nov.**

Cryosericia (Cry.o.se.ri'ci.a. N.L. neut. n. *Cryosericum* a candidate genus; -icia ending to denote a class; N.L. fem. pl. n. *Cryosericia* the *Cryosericum* candidate class).

The description is the same as that for the family Cryosericaceae. The type order is Cryosericales.

Description of **Cryosericota phylum nov.**

Cryosericota (Cry.o.se'.ri.co.ta. N.L. neut. n. *Cryosericum* a candidate genus; -ota ending to denote a phylum; N.L. fem. pl. n. *Cryosericota* the *Cryosericum* candidate phylum).

The phylum Cryosericota is defined based on phylogenetic analysis of 16S rRNA gene and a set of 120 marker genes. The type class is Cryosericia.

Acknowledgements

This study was funded by the Genomic Science Program of the United States Department of Energy Office of Biological and Environmental Research, grants DE-SC0004632, DE-SC0010580, and DE-SC0016440. A portion of the research was performed under the Facilities Integrating Collaborations for User Science (FICUS) initiative, with resources at both the DOE Joint Genome Institute (JGI) and the Environmental Molecular Sciences Laboratory (EMSL), under EMSL Proposal ID 49521. A portion was additionally performed through an EMSL Science Theme Award, Proposal ID 48467. JGI and EMSL are DOE Office of Science User Facilities, sponsored by the Office of Biological and Environmental Research and operated under Contract Nos. DE-AC02-05CH11231 (JGI) and DE-AC05-76RL01830 (EMSL). A portion of the computational support was provided by an award from the Ohio Supercomputer Center (OSC; www.osc.edu) to Matthew B. Sullivan. We thank the IsoGenie1 and IsoGenie2 Project Teams for scientific discussions about these results, and the 2010–2012 field teams for sample collection, particularly Tyler Logan, as well as the Abisko Scientific Research Station for sampling infrastructure and support. We thank Eun-Hae Kim and Robert Jones for DNA extraction of the palsa collapse feature samples, Sarah Owens at Argonne National Labs for assistance with amplicon sequencing, and Margretta Murphy for early assistance with the amplicon data. We gratefully acknowledge the reviews of Dr. Ramon Rossello-Mora and two anonymous reviewers, whose thorough and thoughtful feedback improved the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.syapm.2018.12.003.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
[2] Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O. (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9, 75.
[3] Bäckstrand, K., Crill, P.M., Jackowicz-Korczyński, M., Mastepanov, M., Christensen, T.R., Bastviken, D. 2010 Annual carbon gas budget for a subarctic peatland, northern Sweden Meddelanden fran Lunds Universitets Geografiska Institutioner, Avhandlingar, pp. , 53–65.
[4] Bakermans, C., Skidmore, M.L., Douglas, S., McKay, C.P. (2014) Molecular characterization of bacteria from permafrost of the Taylor Valley, Antarctica. FEMS Microbiol. Ecol. 89, 331–346.
[5] Biegel, E., Schmidt, S., Muller, V. (2009) Genetic, immunological and biochemical evidence for a Rnf complex in the acetogen *Acetobacterium woodii*. Environ. Microbiol. 11, 1438–1443.
[6] Biegel, E., Schmidt, S., Gonzalez, J.M., Muller, V. (2011) Biochemistry, evolution and physiological function of the Rnf complex, a novel ion-motive electron transport complex in prokaryotes. Cell. Mol. Life Sci. 68, 613–634.
[7] Bowman, J.P. 2008 Genomic Analysis of Psychrophilic Prokaryotes. Psychrophiles: From Biodiversity to Biotechnology, pp. , 265–284.
[8] Briukhanov, A.L., Thauer, R.K., Netrusov, A.I. (2002) Catalase and superoxide dismutase in the cells of strictly anaerobic microorganisms. Mikrobiologiia 71, 330–335.
[9] Brüggemann, H., Baümer, S., Fricke, W.F., Wiezer, A., Liesegang, H., Decker, I., Herzberg, C., Martinez-Arias, R., Merkl, R., Henne, A., Gottschalk, G. (2003) The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. Proc. Natl. Acad. Sci. U. S. A. 100, 1316–1321.
[10] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10, 421.
[11] Carver, T., Harris, S.R., Berriman, M., Parkhill, J., McQuillan, J.A. (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28, 464–469.
[12] Christensen, T.R., Johansson, T., Åkerman, H.J., Mastepanov, M., Malmer, N., Friborg, T., Crill, P., Svensson, B.H. (2004) Thawing sub-arctic permafrost: effects on vegetation and methane emissions. Geophys. Res. Lett. 31.
[13] Coolen, M.J., van de Giessen, J., Zhu, E.Y., Wuchter, C. (2011) Bioavailability of soil organic matter and microbial community dynamics upon permafrost thaw. Environ. Microbiol. 13, 2299–2314.
[14] Eddy, S.R. (2011) Accelerated profile HMM searches. PLoS Comput. Biol. 7, e1002195.

[15] Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461.
[16] Emerson, J.B., Roux, S., Brum, J.R., Bolduc, B., Woodcroft, B.J., Jang, H.B., Singleton, C.M., Solden, L.M., Naas, A.E., Boyd, J.A., Hodgkins, S.B., Wilson, R.M., Trubl, G., Li, C., Frolking, S., Pope, P.B., Wrighton, K.C., Crill, P.M., Chanton, J.P., Saleska, S.R., Tyson, G.W., Rich, V.I., Sullivan, M.B. (2018) Host-linked soil viral ecology along a permafrost thaw gradient. Nat. Microbiol. 3, 870–880.
[17] Gilichinsky, D., Vishnivetskaya, T., Petrova, M., Spirina, E., Mamykin, V., Rivkina, E. 2008 Bacteria in Permafrost. Psychrophiles: From Biodiversity to Biotechnology, pp. , 83–102.
[18] Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M. (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. 57, 81–91.
[19] Hodgkins, S.B., Tfaily, M.M., McCalley, C.K., Logan, T.A., Crill, P.M., Saleska, S.R., Rich, V.I., Chanton, J.P. (2014) Changes in peat chemistry associated with permafrost thaw increase greenhouse gas production. Proc. Natl. Acad. Sci. U. S. A. 111, 5819–5824.
[20] Hori, T., Aoyagi, T., Itoh, H., Narihiro, T., Oikawa, A., Suzuki, K., Ogata, A., Friedrich, M.W., Conrad, R., Kamagata, Y. (2015) Isolation of microorganisms involved in reduction of crystalline iron(III) oxides in natural environments. Front. Microbiol. 6, 386.
[21] Hugenholtz, P., Pitulle, C., Hershberger, K.L., Pace, N.R. (1998) Novel division level bacterial diversity in a Yellowstone hot spring. J. Bacteriol. 180, 366–376.
[22] Jansson, J.K., Tas, N. (2014) The microbial ecology of permafrost. Nat. Rev. Microbiol. 12, 414–425.
[23] Johansson, T., Malmer, N., Crill, P.M., Friborg, T., ÅKerman, J.H., Mastepanov, M., Christensen, T.R. (2006) Decadal vegetation changes in a northern peatland, greenhouse gas fluxes and net radiative forcing. Glob. Change Biol. 12, 2352–2369.
[24] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M. (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–D462.
[25] Kanehisa, M., Sato, Y., Morishima, K. (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J. Mol. Biol. 428, 726–731.
[26] Kang, D.D., Froula, J., Egan, R., Wang, Z. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3, e1165.
[27] Katoh, K., Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.
[28] Kim, H.T., Lee, S., Lee, D., Kim, H.S., Bang, W.G., Kim, K.H., Choi, I.G. (2010) Overexpression and molecular characterization of Aga50D from *Saccharophagus degradans* 2-40: an exo-type beta-agarase producing neoagarobiose. Appl. Microbiol. Biotechnol. 86, 227–234.
[29] Kim, S., Pevzner, P.A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. 5, 5277.
[30] Konstantinidis, K.T., Tiedje, J.M. (2005) Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. 187, 6258–6264.
[31] Konstantinidis, K.T., DeLong, E.F. (2008) Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J. 2, 1052–1065.
[32] Konstantinidis, K.T., Rossello-Mora, R., Amann, R. (2017) Uncultivated microbes in need of their own taxonomy. ISME J. 11, 2399–2406.
[33] Kopylova, E., Noe, L., Touzet, H. (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28, 3211–3217.
[34] 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 Res. 44, W242–W245.
[35] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.
[36] Li, H. 2013 Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM, pp. , 1–3, arXiv: 1303.3997 [q-bio.GN] 00.
[37] Li, J., Hu, Q., Li, Y., Xu, Y. (2015) Purification and characterization of cold-adapted beta-agarase from an Antarctic psychrophilic strain. Braz. J. Microbiol. 46, 683–690.
[38] Liebner, S., Harder, J., Wagner, D. (2008) Bacterial diversity and community structure in polygonal tundra soils from Samoylov Island, Lena Delta, Siberia. Int. Microbiol. 11, 195–202.
[39] Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Forster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., Konig, A., Liss, T., Lussmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K.H. (2004) ARB: a software environment for sequence data. Nucleic Acids Res. 32, 1363–1371.
[40] Lykidis, A., Chen, C.L., Tringe, S.G., McHardy, A.C., Copeland, A., Kyrpides, N.C., Hugenholtz, P., Macarie, H., Olmos, A., Monroy, O., Liu, W.T. (2011) Multiple syntrophic interactions in a terephthalate-degrading methanogenic consortium. ISME J. 5, 122–130.
[41] Matsen, F.A., Kodner, R.B., Armbrust, E.V. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11, 538.
[42] McCalley, C.K., Woodcroft, B.J., Hodgkins, S.B., Wehr, R.A., Kim, E.H., Mondav, R., Crill, P.M., Chanton, J.P., Rich, V.I., Tyson, G.W., Saleska, S.R. (2014) Methane dynamics regulated by microbial community response to permafrost thaw. Nature 514, 478–481.
[43] McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., Hugenholtz, P. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 6, 610–618.
[44] Meng, L.-W., Li, X.k., Wang, K., Ma, K.-L., Zhang, J. (2015) Influence of the amoxicillin concentration on organics removal and microbial community structure in an anaerobic EGSB reactor treating with antibiotic wastewater. Chem. Eng. J. 274, 94–101.
[45] Mondav, R., McCalley, C.K., Hodgkins, S.B., Frolking, S., Saleska, S.R., Rich, V.I., Chanton, J.P., Crill, P.M. (2017) Microbial network, phylogenetic diversity and community membership in the active layer across a permafrost thaw gradient. Environ. Microbiol. 19, 3201–3218.
[46] Monteux, S., Weedon, J.T., Blume-Werry, G., Gavazov, K., Jassey, V.E.J., Johansson, M., Keuper, F., Olid, C., Dorrepaal, E. (2018) Long-term in situ permafrost thaw effects on bacterial communities and potential aerobic respiration. ISME J. 12, 2129–2141.
[47] Mori, K., Sunamura, M., Yanagawa, K., Ishibashi, J., Miyoshi, Y., Iino, T., Suzuki, K., Urabe, T. (2008) First cultivation and ecological investigation of a bacterium affiliated with the candidate phylum OP5 from hot springs. Appl. Environ. Microbiol. 74, 6223–6229.
[48] Mori, K., Yamaguchi, K., Sakiyama, Y., Urabe, T., Suzuki, K. (2009) *Caldisericum exile* gen. nov., sp. nov., an anaerobic, thermophilic, filamentous bacterium of a novel bacterial phylum, *Caldiserica* phyl. nov., originally called the candidate phylum OP5, and description of *Caldisericaceae* fam. nov., *Caldisericales* ord. nov. and *Caldisericia* classis nov. Int. J. Syst. Evol. Microbiol. 59, 2894–2898.
[49] Mori, K., Fujita, N. (2014) The family *Caldisericaceae*. In: Rosenberg, E., DeLong, E.F., Lory, S., Stackebrandt, E., Thompson, F. (Eds.), The Prokaryotes, Springer, Berlin, Heidelberg.
[50] Nobu, M.K., Narihiro, T., Rinke, C., Kamagata, Y., Tringe, S.G., Woyke, T., Liu, W.T. (2015) Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. ISME J. 9, 1710–1722.
[51] Normand, A.E., Smith, A.N., Clark, M.W., Long, J.R., Reddy, K.R. (2017) Chemical composition of soil organic matter in a subarctic peatland: influence of shifting vegetation communities. Soil Sci. Soc. Am. J. 81, 41–49.
[52] Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A.R., Xia, F., Stevens, R. (2014) The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res. 42, D206–D214.
[53] Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055.
[54] Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., Tyson, G.W. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. 2, 1533–1542.
[55] Price, M.N., Dehal, P.S., Arkin, A.P. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490.
[56] Probst, A.J., Castelle, C.J., Singh, A., Brown, C.T., Anantharaman, K., Sharon, I., Hug, L.A., Burstein, D., Emerson, J.B., Thomas, B.C., Banfield, J.F. (2017) Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO_2 concentrations. Environ. Microbiol. 19, 459–474.
[57] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glockner, F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596.
[58] Quinlan, A.R., Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.
[59] R Core Team, 2017 R: A Language and Environment for Statistical Computing , https://www.R-project.org.
[60] Richter, M., Rossello-Mora, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U. S. A. 106, 19126–19131.
[61] Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P., Woyke, T. (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431–437.
[62] Rodriguez-R, L.M., Konstantinidis, K.T. (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes microbial genomes and metagenomes. PeerJ Preprints 4, e1900v1.
[63] Schmehl, M., Jahn, A., Meyer zu Vilsendorf, A., Hennecke, S., Masepohl, B., Schuppler, M., Marxer, M., Oelze, J., Klipp, W. (1993) Identification of a new class of nitrogen fixation genes in Rhodobacter capsulatus: a putative membrane complex involved in electron transport to nitrogenase. Mol. Gen. Genet. 241, 602–615.
[64] Schut, G.J., Bridger, S.L., Adams, M.W. (2007) Insights into the metabolism of elemental sulfur by the hyperthermophilic archaeon *Pyrococcus furiosus*: characterization of a coenzyme A- dependent NAD(P)H sulfur oxidoreductase. J. Bacteriol. 189, 4431–4441.
[65] Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069.
[66] Shivaji, S., Kumari, K., Kishore, K.H., Pindi, P.K., Rao, P.S., Radha Srinivas, T.N., Asthana, R., Ravindra, R. (2011) Vertical distribution of bacteria in a

lake sediment from Antarctica by culture-independent and culture-dependent approaches. Res. Microbiol. 162, 191–203.
[67] Singleton, C.M., McCalley, C.K., Woodcroft, B.J., Boyd, J.A., Evans, P.N., Hodgkins, S.B., Chanton, J.P., Frolking, S., Crill, P.M., Saleska, S.R., Rich, V.I., Tyson, G.W. (2018) Methanotrophy across a natural permafrost thaw environment. ISME J. 12, 2544–2558, http://dx.doi.org/10.1038/s41396-018-0065-5.
[68] Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
[69] Tange, O. (2011) Gnu parallel-the command-line power tool. USENIX Mag., 42–47.
[70] Tischer, K., Kleinsteuber, S., Schleinitz, K.M., Fetzer, I., Spott, O., Stange, F., Lohse, U., Franz, J., Neumann, F., Gerling, S., Schmidt, C., Hasselwander, E., Harms, H., Wendeberg, A. (2013) Microbial communities along biogeochemical gradients in a hydrocarbon-contaminated aquifer. Environ. Microbiol. 15, 2603–2615.
[71] Vizcaino, J.A., Csordas, A., Del-Toro, N., Dianes, J.A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q.W., Wang, R., Hermjakob, H. (2016) 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 44, 11033.
[72] Wagner, G.P., Kin, K., Lynch, V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 131, 281–285.
[73] Ward, C.P., Cory, R.M. (2015) Chemical composition of dissolved organic matter draining permafrost soils. Geochim. Cosmochim. Acta 167, 63–79.
[74] Woodcroft, B.J., Singleton, C.M., Boyd, J.A., Evans, P.N., Emerson, J.B., Zayed, A.A.F., Hoelzle, R.D., Lamberton, T.O., McCalley, C.K., Hodgkins, S.B., Wilson, R.M., Purvine, S.O., Nicora, C.D., Li, C., Frolking, S., Chanton, J.P., Crill, P.M., Saleska, S.R., Rich, V.I., Tyson, G.W. (2018) Genome-centric view of carbon processing in thawing permafrost. Nature 560, 49–54.
[75] Yanagawa, K., Morono, Y., de Beer, D., Haeckel, M., Sunamura, M., Futagami, T., Hoshino, T., Terada, T., Nakamura, K., Urabe, T., Rehder, G., Boetius, A., Inagaki, F. (2013) Metabolically active microbial communities in marine sediment under high-CO(2) and low-pH extremes. ISME J. 7, 555–567.
[76] Yde, J.C., Finster, K.W., Raiswell, R., Steffensen, J.P., Heinemeier, J., Olsen, J., Gunnlaugsson, H.P., Nielsen, O.B. (2010) Basal ice microbiology at the margin of the Greenland ice sheet. Ann. Glaciol. 51, 71–79.
[77] Ye, R., Jin, Q., Bohannan, B., Keller, J.K., McAllister, S.A., Bridgham, S.D. (2012) pH controls over anaerobic carbon mineralization, the efficiency of methane production, and methanogenic pathways in peatlands across an ombrotrophic–minerotrophic gradient. Soil Biol. Biochem. 54, 36–47.
[78] Yergeau, E., Hogues, H., Whyte, L.G., Greer, C.W. (2010) The functional potential of high Arctic permafrost revealed by metagenomic sequencing, qPCR and microarray analyses. ISME J. 4, 1206–1214.
[79] Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., Xu, Y. (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 40, W445–W451.