nature SUPPLEMENTARY INFORMATION
https://doi.org/10.1038/s41586-018-0338-1

In the format provided by the authors and unedited.

Genome-centric view of carbon processing in thawing permafrost

Ben J. Woodcroft[1,10], Caitlin M. Singleton[1,10], Joel A. Boyd[1], Paul N. Evans[1], Joanne B. Emerson[2,9], Ahmed A. F. Zayed[2], Robert D. Hoelzle[1], Timothy O. Lamberton[1], Carmody K. McCalley[3], Suzanne B. Hodgkins[4], Rachel M. Wilson[4], Samuel O. Purvine[5], Carrie D. Nicora[5], Changsheng Li[6], Steve Frolking[6], Jeffrey P. Chanton[4], Patrick M. Crill[7], Scott R. Saleska[8], Virginia I. Rich[2] & Gene W. Tyson[1]*

[1]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, Queensland, Australia. [2]Department of Microbiology, The Ohio State University, Columbus, OH, USA. [3]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY, USA. [4]Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL, USA. [5]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA. [6]Earth Systems Research Center, Institute for the Study of Earth, Oceans and Space, University of New Hampshire, Durham, NH, USA. [7]Department of Geological Sciences and Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden. [8]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. [9]Present address: Department of Plant Pathology, University of California, Davis, CA, USA. [10]These authors contributed equally: Ben J. Woodcroft, Caitlin M. Singleton. *e-mail: g.tyson@uq.edu.au

Supplementary Notes

Supplementary Note 1

To conservatively estimate the number of strains present at the site, raw metagenomes were processed with the SingleM v0.2.1 command 'pipe' using default parameters (https://github.com/wwood/singlem). The pipe command finds the abundances of discrete sequence-based operational taxonomic units (OTUs) directly from shotgun metagenome data, by aligning translated reads to profile HMMs representing conserved ribosomal proteins. Reads completely covering a conserved 20 amino acid stretch of the HMM are back-translated into their respective 60 bp nucleotide sequences and compared with each other via sequence similarity. Based on analysis of mock metagenome data comprised of known compositions of fully sequenced genomes[92], roughly 10% of sequences derived from Illumina HiSeq 2000 data contain sequencing errors. The number of distinct OTU sequences detected amongst the Stordalen Mire reads from each of the 15 ribosomal proteins was 28,610-34,700 (comprised of 75,540-85,090 total reads). Assuming that 10% of reads are erroneous and all those reads are singletons, the number of distinct lineages detected at the site was estimated to be the number of different sequences (OTUs) detected minus 10% of the total sequence count. The median of this estimate across the 15 ribosomal genes was calculated as 24,300 (21,060-26,190). After removal, the large number of singletons that remained (average 19%, range 17-21%) suggests that many low abundance populations not detected by SingleM are present at the site and that 24,300 is an underestimate of the number of OTUs present at the site.

Supplementary Note 2

To determine the fraction of the community represented in the Stordalen genomes, SingleM 'pipe' was used as in Supplementary Note 1. OTU sequences derived from the reads were compared to those derived from the genomes. A read-derived OTU sequence was estimated to be from the same genus as a recovered MAG if it shared at least 89% global sequence identity with one or more genome-derived OTUs, calculated using VSEARCH[93]. The 89% similarity cutoff was inferred through an analysis of IMG 4.1 isolate genomes[94], where on average two species in the same genus were 89% identical, in line with previous studies on average nucleotide identity[95]. In a minority of cases, a ribosomal protein was found multiple times in the same MAG due to contamination resulting from imperfect assembly and/or binning. To remove the possibility of these contaminants inflating the overall estimate of community recovery, genes found more than once in a genome were excluded. The specific command used was "singlem appraise --imperfect --sequence_identity 0.89 --metagenome_otu_tables <metagenome_otu_table> --genome_otu_tables <genome_otu_table>", where <metagenome_otu_table> and <genome_otu_table> files were the output of "singlem pipe --otu_table" for the raw reads and MAGs, respectively.

Supplementary Note 3

Bulk density measurements in a palsa core taken from July 2013 showed an increase in grams per millilitre of sample from surface to deep. Wet density increased from 0.30 g/mL in the surface sample to 0.50 g/mL in the deep, and dry density increased from 0.05 g/mL to 0.16 g/mL in the deep.

Supplementary Note 4

Monosaccharide degradation pathways were observed in many recovered genomes, including those for glucose (34%, 36% and 29% of the recovered palsa, bog and fen communities), lactose (28%, 42% and 47%), galactose (36%, 41% and 32%), fructose (35%, 21% and 20%) and mannose (0%, 0.1% and 7%). An overview of these degradation pathways is presented in Extended Data Figure 5.

Supplementary Note 5

Six pathways are known to mediate the degradation of xylose (Extended Data Figure 6a). The 'isomerase' pathway is found in bacteria, the Weimberg and Dahms (WD) pathways found in bacteria and archaea (Jackson and Nicolson 2002, Weimberg 1961, Dahms 1974), the 'oxidoreductase' pathways are found in fungi and the "xylonate dehydratase" pathway characterised in *Gluconobacter oxydans*[20]. A smaller number of Acidobacteria encoded xylonate dehydratase but no other steps in the WD pathways, suggesting they degrade xylose through a membrane bound glucose dehydrogenase as previously observed in *Gluconobacter oxydans*[20].

Concordant with the high prevalence of genomes encoding xylan degrading enzymes (37.0% of genomes), at least one of the degradation pathways for xylose is encoded by 42.5% of all MAGs. Of the MAGs with xylose degrading potential, 69.8% exclusively encode the isomerase pathway, 6.8% encode only the oxidoreductase pathway and 17.7% encode both (Extended Data Figure 6b). While only two genomes encoded either of the WD pathways, 42 (2.7%) genomes were found to encode the gene for xylonate dehydratase, the last common reaction shared by the Weimberg and Dahms pathways. The oxidoreductase xylose degradation pathway was abundant within the Stordalen Mire MAGs and so may be an important pathway for xylose metabolism in thawing permafrost. Most MAGs that encode the oxidoreductase pathway also encode the isomerase pathway (Extended Data Figure 6b), and utilize enzymes that act more generally on monosaccharide sugars. In both the bog and fen, genomes encoding the oxidoreductase pathway increased in abundance with depth (Fig. 2). While MAGs from a range of phyla encoded the canonical xylose isomerase degradation pathway, the oxidoreductase pathway was found predominantly in the Acidobacteria (69 of 364 genomes), Actinobacteria (65 of 385 genomes) and the Chloroflexi (20 of 66 genomes) (Fig. 2).

In eight instances, genomes encoding both the isomerase pathway and the oxidoreductase pathway only expressed the oxidoreductase pathway, as evidenced by their expression of xylulokinase and at least one of aldehyde reductase, L-iditol 2-dehydrogenase, xylose reductase or xylulose reductase, but not expression of xylose isomerase. This suggests that the expression of xylose degradation pathways is likely to be environment-specific.

Supplementary Note 6

The CH_4 produced by methanogens is released to the atmosphere or oxidized to CO_2 by methanotrophs. While anaerobic methanotrophs (e.g. ANME) were not detected, MAGs of eleven aerobic methanotrophs belonging to the families *Methylocystaceae* (7 genomes), *Beijerinckiaceae* (2) and *Methylococcaceae* (2) were recovered (Extended Data Fig. 1). At the mire, members of the *Methylocystaceae* and *Beijerinckiaceae* were predominantly found in the bog, and *Methylococcaceae* populations were found in the fen (Fig. 2 'MAG abundances'), following the known environmental distributions of these lineages[96]. Consistent with CH_4 availability, methanotrophs were present at most depths in the bog, and predominantly the surface and mid depths in the fen (Fig. 2 'distribution boxplots'). Expression of methanotrophy transcripts was greatest in the fen, but also high in the palsa and bog, and protein expression was detected in all three environments (Fig. 2 'pathway expression'). The presence of methanotrophs in the deep region of the peat columns supports recent findings that microaerobic environments have sufficient oxygen for the growth of specialised methanotrophs[97]. In contrast to the bog and fen, the palsa is predominantly aerobic with minimal methane production, and appears to only support a very low abundance of methanotrophs (Fig. 2 'MAG abundances'). The high relative abundance of methanogens to methanotrophs in the fen (24:1 per core) and their likely production and consumption rates suggest that methanotrophs mitigate only a small amount of CH_4 before it is released to the atmosphere, consistent with observed methane production from this fen[33]. A lower CH_4 flux in the bog corresponds to a lower abundance of methanogens, and a lower ratio to methanotrophs (7:1 per core), which suggests that the amount of mitigated CH_4 may be significant in this environment (Extended Data Figure 8b).

Supplementary Note 7

To determine the taxonomic diversity of specific lineages average amino acid identities (AAIs) between MAGs were calculated. The AAI between *Ca*. 'Acidiflorens stordalenmirensis' and the fen and palsa clades was determined to be 80-85%, where AAIs between MAGs within these clades were >95%. These AAI values indicate that these clades form species and strains within the *Ca.* 'Acidiflorens' respectively[95]. Likewise, AAI values between the *Ca*. 'Methanoflorens stordalenmirensis' and *Ca.* 'M. crillii' 97% ANI dereplicated representative genomes was 89%, indicating these lineages form distinct species within the genus *Ca.* 'Methanoflorens'. Lineages of the candidate phylum AD3 were analysed similarly, with all AAIs between clades 50-63% indicating that they were divergent at least at the family level.

The dereplicated *Ca.* 'Changshengia' MAGs from Stordalen share AAIs of 86-91.4%, indicating they belong to the same genus. These populations have AAIs of 77-78.6% with the two most closely related publicly available MAGs, UBA8260 and UBA12275 (Fig. 2), binned from peat bog samples (SRA experiments SRX461673 and SRX461727)[62], suggesting that these MAGs also fall within the *Ca.* 'Changshengia'.

Supplementary Note 8

There were 11 clades detected within the *Ca.* 'Acidiflorens' genus. Predicted substrates for these clades include plant materials such as xylan and fructose, aromatic compounds, oxalic acid, fatty acids, fermentation end products like glycerol, lactate and formate, and also nitrate, nitrite and sulphate. Spatial separation of the 11 *Ca*. 'Acidiflorens' groupings was observed across the thawing stages and sample depths.

Five of the 11 clades (clades 1, 9-11 and *Ca*. 'Acidiflorens' sp. 3) were observed primarily in palsa samples, varying in relative abundance with depth. Four of these five 'palsa' clades (clades 1, 9, 10 and 11) appear to have metabolisms including β-oxidation, aromatic compound degradation and oxalate oxidation. *Ca*. 'Acidiflorens' sp. 3 appears to have the capability for nitrate (*narGHI*) and nitrite (*nirAB*) reduction.

In the bog and fen, *Ca*. 'Acidiflorens' clades 2, 3, 7, *Ca*. 'Acidiflorens' sp. 2 and *Ca*. 'Acidiflorens stordalenmirensis' were observed. Both clades 2 and 7 encode a β-oxidation pathway, but where clade 2 is observed at high relative abundance in the surface and mid-depth samples of the bog, clade 7 was found more abundant in deep and extra deep samples. Clade 3 genomes encoded arabinose and xylose fermentation pathways and were more abundant in the surface and mid-depths of the bog and fen. *Ca*. 'Acidoflorens' sp. 2 and *Ca*. 'A. stordalenmirensis' both utilise fumarate addition for degradation of aromatic compounds (Fig. 3). Clade 8 is present in deeper samples at lower relative abundance across all three thaw stages and is predicted to use nitrate as an electron acceptor.

Ca. 'A. stordalenmirensis' and other *Ca*. 'Acidiflorens' clades have the ability to degrade xylan, based on the presence of β-1,4-xylanase genes, and utilise xylose as an energy source (*xylAB*, *ack*, *xpkA*) forming glyceraldehyde-3-phosphate which feeds into the glycolytic pathway. All *Ca*. 'Acidiflorens' clades encode cyctochrome-c (*cytABCD*) and high affinity cytochrome-bd oxidases (*cbdAB*) and are predicted to exist in aerobic and microaerobic conditions. *Ca*. 'A. stordalenmirensis', the *Ca*. 'Acidiflorens' clade with the strongest correlation in relative abundance with *Ca*. 'Methanoflorens stordalenmirensis', is the only Acidiflorens clade predicted to metabolise fructose. It is likely that *Ca*. 'Methanoflorens stordalenmirensis' consumes hydrogen produced by NADH-oxidising NiFe (Complex IV and *hox*) hydrogenases present in the *Ca*. 'A. stordalenmirensis' genomes. Other hydrogenases found in the *Ca*. 'A. stordalenmirensis' genomes include hydrogen formate lyase and formate dehydrogenase. Surprisingly, *Ca*. 'A. stordalenmirensis' and five of the other *Ca*. 'Acidiflorens' clades have dissimilatory sulphite reductase genes (*dsrAB*) that cluster with genes from the environmental supercluster 1[98]. The presence of genes for *apr*, *sat*, and *dsrMJKOP* suggests that sulphur metabolism is linked to electron transport in these populations. Although *qmoABC* genes are not present, genes for heterodisulphide reductase (*hdrABC*) may substitute this complex, given its similarities in structure, to reduce sulphate to sulphite providing an electron sink under anaerobic conditions.

Supplementary Note 9

The high abundances of Dormibacteraeota in the mid and deep layers of the palsa and bog facilitated recovery of 47 MAGs, including 14 from *Ca.* 'Changshengia' (Fig. 1), and expanded genomic representation of the Dormibacteraeota phylum by three fold. Members of Dormibacteraeota have previously been detected in 16S rRNA gene sequences from subsurface soil and permafrost-associated environments[99-101], with three genomes most recently recovered from Antarctic soil[102]. Metabolic reconstruction of the Dormibacteraeota genomes revealed diverse metabolic traits (Extended Data Figure 9b), including aromatic degradation, sugar fermentation, fumarate reduction and nitrate reduction. Genes for the oxidation of glycerol from the *Ca.* 'Changshengia' MAGs were also found to be highly expressed in the metatranscriptomes of deeper bog layers (Extended Data Figure 9b).

Supplementary References

20 Zhang, M. *et al.* Genetic analysis of D-xylose metabolism pathways in Gluconobacter oxydans 621H. *Journal of Industrial Microbiology & Biotechnology* **40**, 379-388 (2013).

33 McCalley, C. K. *et al.* Methane dynamics regulated by microbial community response to permafrost thaw. *Nature* **514**, 478-481 (2014).

62 Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature microbiology* (2017).

92 Shakya, M. *et al.* Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental microbiology* **15**, 1882-1899 (2013).

93 Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

94 Markowitz, V. M. *et al.* IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic acids research*, gkt963 (2013).

95 Konstantinidis, K. T. & Tiedje, J. M. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current opinion in microbiology* **10**, 504-509 (2007).

96 Knief, C. Diversity and habitat preferences of cultivated and uncultivated aerobic methanotrophic bacteria evaluated based on pmoA as molecular marker. *Frontiers in microbiology* **6**, 1346 (2015).

97 Kits, K. D., Klotz, M. G. & Stein, L. Y. Methane oxidation coupled to nitrate reduction under hypoxia by the Gammaproteobacterium Methylomonas denitrificans, sp. nov. type strain FJG1. *Environmental microbiology* **17**, 3219-3232 (2015).

98 Müller, A. L., Kjeldsen, K. U., Rattei, T., Pester, M. & Loy, A. Phylogenetic and environmental diversity of DsrAB-type dissimilatory (bi) sulfite reductases. *The ISME journal* **9**, 1152-1165 (2015).

99 Zhou, J. *et al.* Bacterial phylogenetic diversity and a novel candidate division of two humid region, sandy surface soils. *Soil Biology and Biochemistry* **35**, 915-924 (2003).

100 Taş, N. *et al.* Impact of fire on active layer and permafrost microbial communities and metagenomes in an upland Alaskan boreal forest. *The ISME journal* **8**, 1904-1919 (2014).

101 Ji, M. *et al.* Microbial diversity at Mitchell Peninsula, Eastern Antarctica: a potential biodiversity “hotspot”. *Polar Biology* **39**, 237-249 (2016).

102 Ji, M. *et al.* Atmospheric trace gases support primary production in Antarctic desert surface soil. *Nature* **552**, 400 (2017).

Supplementary Information Guide to Supplementary Data Files 10-14

The following Supplementary Data Files are available on figshare.

Supplementary Data File 10

Bacterial genome tree (newick format) created from the concatenated alignment of 120 Bacteria specific single copy marker genes derived from NCBI bacterial genomes, UBA genomes, and the Stordalen bacterial MAGs, bootstrapped 100×. This tree is available on figshare under DOI: 10.6084/m9.figshare.6233660. The multiple sequence alignment used to generate this tree is presented in Supplementary Data File 13.

Supplementary Data File 11

Archaeal genome tree (newick format) created from the concatenated alignment of 122 Archaea specific single copy marker genes derived from NCBI archaeal genomes, UBA genomes, and the Stordalen archaeal MAGs, bootstrapped 100×. This tree is available on figshare under DOI: 10.6084/m9.figshare.6233660. The multiple sequence alignment used to generate this tree is presented in Supplementary Data File 12.

Supplementary Data File 12

Multiple sequence alignment used to generate the archaeal phylogenetic tree presented in Supplementary Data File 11. This alignment file is available on figshare under DOI: 10.6084/m9.figshare.6233660.

Supplementary Data File 13

Multiple sequence alignment used to generate the bacterial phylogenetic tree presented in Supplementary Data File 10. This alignment file is available on figshare under DOI: 10.6084/m9.figshare.6233660.

Supplementary Data File 14

Proteins used in the proteomics search. After the contaminant entries, each entry in the FASTA file consists of the name of each protein, where the first 3 elements in snake case refer to the genome, and the final number refers to the protein ID within that genome. Where proteins were identical after conversion of isoleucine residues to leucine, each protein is identified in the entry's header, space separated. This protein fasta file is available on figshare under DOI: 10.6084/m9.figshare.6233660.