A 2-million-year-old microbial and viral communities from the Kap København Formation in North Greenland

Antonio Fernandez-Guerra[1,2,*], Guillaume Borrel[3], Tom O Delmont[4], Bo Elberling5, A. Murat Eren[6,7,8], Simonetta Gribaldo[3], Annika Jochheim[9], Rasmus Amund Henriksen[2], Kai-Uwe Hinrichs[10], Thorfinn S. Korneliussen[1,2], Mart Krupovic[3], Nicolaj K. Larsen[1,2], Rafael Laso-Pérez[11], Mikkel Winther Pedersen[1,2], Vivi K. Pedersen[12], Karina K. Sand[1,2], Martin Sikora[1,2], Martin Steinegger[13], Iva Veseli[14], Lars Wörmer[10], Lei Zhao[2], Marina Žure[1,2], Kurt Kjær[1,2], Eske Willerslev[1,2,10,15,16,*]

[1]Centre for Ancient Environmental Genomics, Globe Institute, University of Copenhagen, Copenhagen, Denmark; [2]Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark; [3]Institut Pasteur, Université Paris Cité, CNRS UMR6047, Evolutionary Biology of the Microbial Cell Unit, Paris, France; [4]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Université Paris-Saclay, Evry, France; [5]Center for Permafrost (CENPERM), Department of Geosciences and Natural Resource Management, University of Copenhagen, DK1350, Copenhagen, Denmark; [6]Helmholtz Institute for Functional Marine Biodiversity, Oldenburg, Germany; [7]Max Planck Institute for Marine Microbiology,Bremen, Germany; [8]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany; [9]Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany; [10]MARUM Center for Marine Environmental Sciences and Faculty of Geosciences, University of Bremen, Bremen, Germany; [11]Biogeochemistry and Microbial Ecology Department, Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain; [12]Department of Geoscience, Aarhus University, Denmark; [13]School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea; [14]Biophysical Sciences Program, The University of Chicago, Chicago, IL 60637, USA; [15]University of Cambridge, Cambridge, UK; [16]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK

*Corresponding authors: antonio.fernandez.guerra@sund.ku.dk, ewillerslev@sund.ku.dk

Summary

Using ancient environmental DNA (eDNA)[1] we reconstructed microbial and viral communities from the Kap København Formation in North Greenland[2]. We find pioneer microbial communities, along with likely dormant methanogens from the permafrost's seed bank. Our findings reveal that at the time of the formation, the terrestrial input of the Kap København site originated from a palustrine wetland, suggesting non-permafrost conditions. During this time, detection of methanogenic archaea and carbon processing pathways suggests a moderate strengthening of methane emissions through the northward expansion of wetlands. Intriguingly, we discover a remarkable sequence similarity (>98%) between pioneer methanogens and present-day thawing permafrost counterparts. This suggests that not all microbes respond uniformly to environmental change over geological timescales, but that some microbial taxa's adaptability and resilience remain constant over time. Our findings further suggest that the composition of microbial communities is changing prior to plant communities as a result of global warming.

Main

In recent years, ancient environmental DNA (eDNA)[1] has transformed our understanding of ecosystem dynamics and how ecosystems have evolved in response to environmental changes[3–5]. Using eDNA preserved in the Kap København Formation, Kjær et al.[2] reconstructed the plant and animal communities that inhabited North Greenland ~2 million years ago (Fig. 1A), and found a complex and diverse species assemblage that has no modern analogue. However, this study do not consider the microbes which play fundamental roles in the maintenance of biogeochemical processes that shape the community structure of multicellular organisms[6–8].

The characterisation of the microbial eDNA component is crucial for insights into the complex functioning of past ecosystems. In fact, most sedimentary ancient DNA studies typically neglect microbial communities[8], primarily due to their dynamic nature, which distinguishes them from plant and animal communities whose genetic signals are not overprinted by continued activity within the deposit. The Kap København Formation consists of a c. 100 m thick succession of shallow marine sediments that were primarily deposited during a c. 20,000 years long interglacial period c. 2 million years ago[2,9]. These sediments serve as an archive capturing the community composition upstream from the depositional sites, as they were fluvially transported to the foreshore and concentrated as organic detritus mixed into sandy near-shore sedimentsl[2].

Reconstructing pioneer microbial communities that inhabited the Kap København Formation is challenging due to their complexity, namely, intermixing of microbial communities present initially at the depositional site and those deposited over the course of the following 20,000 years[10,11]. Moreover, the composition of these communities and their descendants (Fig. 1C) continues to evolve in response to local changes like nutrient availability, pH, oxygen levels, and other factors within the depositional site[12,13]. Additionally, the presence of allochthonous microorganisms that colonized the site via motility through the porous sediments also plays a role in shaping the community composition over time[14–17]. Despite these challenges, the unique geological and climatic features of the Kap København Formation provide an exceptional opportunity to investigate the microbial communities that inhabited this ecosystem about two million years ago.

After the Kap København Formation was deposited in a delta and shallow marine environment c. 2 million years ago, it was uplifted due to the solid Earth's flexural isostatic response to the erosional unloading of the fjord incision[18] (Fig. 1B). We estimate an uplift rate of 0.0933 mm/yr, which elevated these shallow marine deposits to their current elevation (Fig. 1D). Based on this

rate, we infer that the four sampled localities (localities 50, 75, 119, 69) (Fig. 1B) across three stratigraphic units (B1, B2 and B3) were exposed subaerially between 0.8-1.2 million years ago (Fig. 1D). Sediment records from below the present ice sheet indicate that the Greenland Ice Sheet was already established at this time[19], although Greenland may have been nearly ice free for shorter time periods since[72]. It is reasonable to assume that in this interval as the sediments emerged subaerially, they gradually became part of the permafrost, creating a distinction between the pioneer communities and the microbial communities that have inhabited the permafrost sediments up to the present day (Fig. 1D).

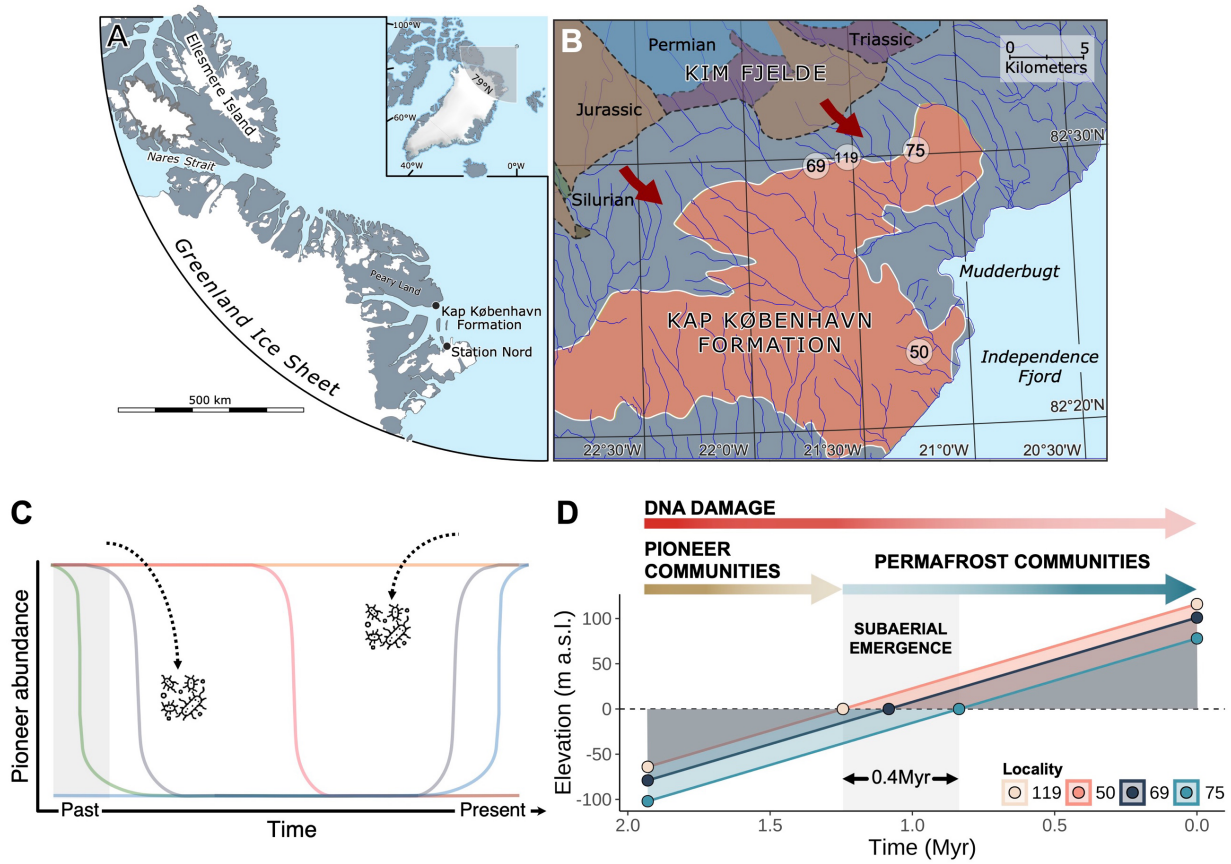


Fig. 1 | Geographical location and uplift model A, Map indicating the location of Kap København Formation in North Greenland at the entrance to the Independence Fjord (82° 24′ N 22° 12′ W). B, Spatial distribution of the ~100-m thick succession of shallow marine near-shore sediments between Mudderbugt and the low mountains towards the north, circles highlight the localities used in the manuscript. C, Schematic diagram illustrating the dynamics of pioneer community composition since the deposition event, depicted as a grey-shaded area. Each coloured line represents the relative abundances of a certain microbe from past to present and

how these relative abundances evolve across time. The “microbial icons” represent allochthonous microorganisms that arrived at the original site, i.e., motility through porous sediments. D, Simple uplift model of the Kap København Formation. The DNA damage arrow depicts taxonomic groups, with darker colours indicating greater damage, while the arrows for the pioneer and permafrost microbial communities show abundance levels, with darker colours indicating higher abundance.

Pioneer and permafrost microbial communities

We designed a strategy for inferring the microbial community composition in deep-time samples from ultra-short and damaged DNA sequencing reads (Extended Data Fig. 1). Our approach involved three main components. Firstly, we created a comprehensive genomic database with a common taxonomic framework (Supplementary Information 1) covering Archaea, Bacteria, Viruses and Eukarya (only chloroplasts and mitochondria). Secondly, we performed a sensitive search using bowtie2[20] and applied stringent filtering criteria that leveraged read patterns across the references, allowing us to confidently retrieve low-coverage genomes, even with as little as 1% of their genome present (hereafter, we will refer to the breadth of coverage as detection). We inferred their relative abundance by normalising the estimated TAD80 (truncated average depth) number of reads to the reference length (Supplementary Information 2). Lastly, we performed large-scale damage estimation on the filtered results to identify those references that are potentially members of the pioneer microbial communities (Supplementary Information 3). We excluded eukaryotic hits from subsequent analyses since the eukaryotic references were only intended to be used as bait in the competitive mapping. We then focused our analyses on 22 samples that met two criteria: (1) having more than 10 million unique reads and (2) where the combined relative abundance of references exhibiting damage greater than 0.11 accounted for at least 50% (Supplementary Information 3). For damage estimate, we used the background-subtracted damage frequency at the first position provided by metaDMG[21], where a value of 1 represents the maximum level. We mapped an average of 7.2% of unique reads per sample with an average of 1,950 references (Fig. 2A). Finally, we aggregated our results by calculating the geometric mean of taxonomic abundances and averaging the inferred damage at the species level. On average, the damaged fraction composed 84% of the estimated relative abundances in each sample, revealing Archaea as the dominant pioneer group (Fig. 2A).

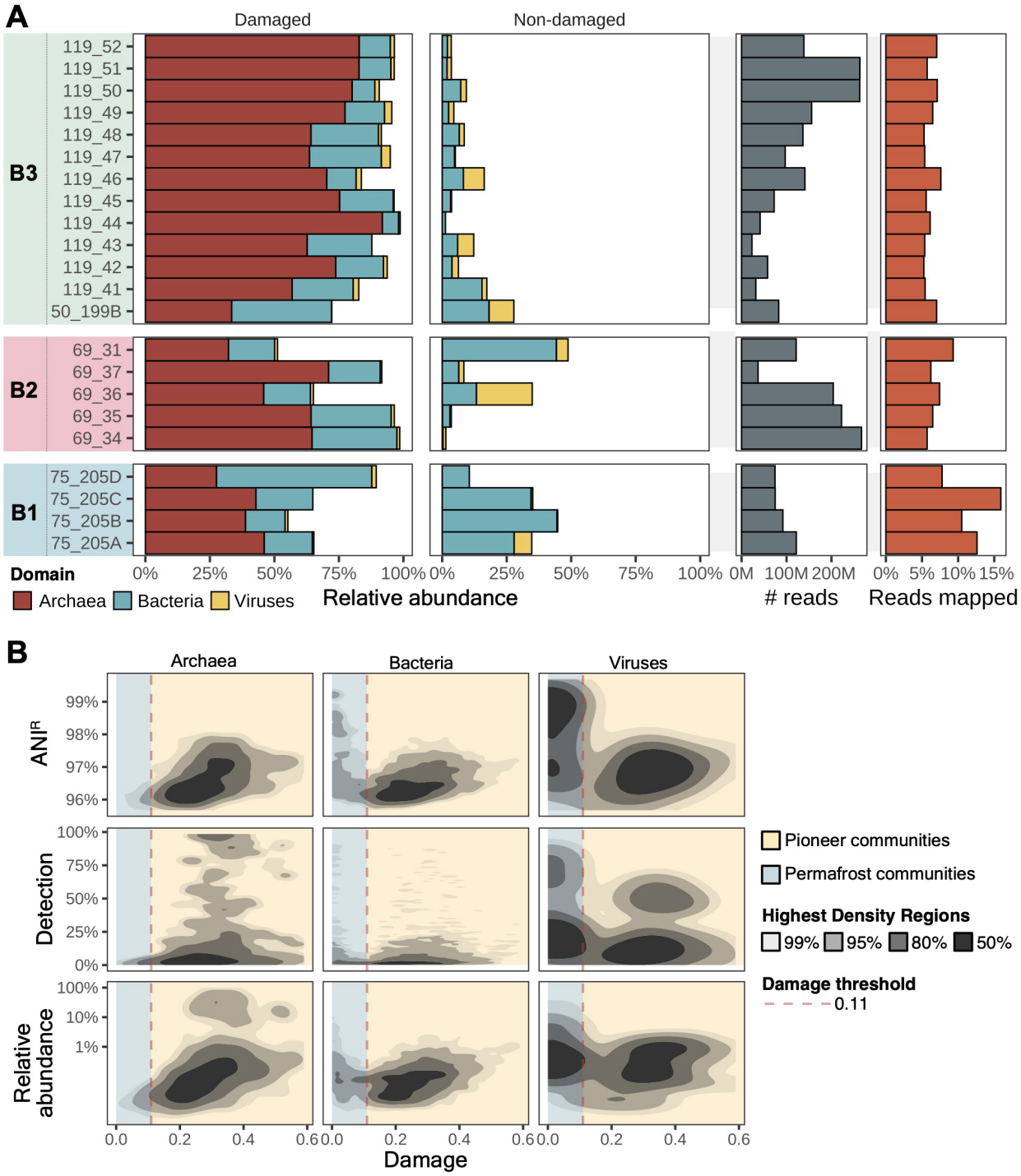


Fig. 2 | Damage and non-damage distribution patterns A, Proportion of taxa identified as damaged and non-damaged on the Kap København Formation based on the estimated taxonomic abundances. The two rightmost panels show the number of de-replicated reads utilised for mapping and the percentage of reads that were successfully mapped to any of the references within the database employed for taxonomic profiling. B, Highest density regions plots showing the average nucleotide identity, detection and relative abundance distribution as a function of the

damage for each of the three domains. The dashed red line sets the damage threshold we inferred from the Eukaryotic data. The potential fraction of pioneer communities is highlighted in yellow, while the potential communities inhabiting the permafrost are marked in blue

By integrating damage estimates, relative abundances, detection, and mean read Average Nucleotide Identity estimates (ANI^R), we were able to differentiate between the initial pioneer communities and those communities that inhabited the permafrost after the subaerial emergence. Our results indicate a limited presence of Archaea in the permafrost communities (blue area in Fig. 2B), which is consistent with the low abundance of archaeal taxa typically found in the cryosphere[22,23]. On the contrary, the high ANI^R values to modern references, along with the detection and relative abundance of the bacterial and viral fraction, demonstrate the distinct characteristics of the present-day communities inhabiting the permafrost. Despite the significant time gap between the modern reference genomes and the sequences identified in the pioneer communities, we observed elevated ANI^R values even when considering the extent of damage present in the sequences. In some cases, the ANI^R values approached 98%, indicating a relatively high level of identity. It is particularly remarkable that our samples show high detection values for Archaea and certain bacterial taxa, where some of them have more than 75% of the genome recovered.

Modern analogues to the Kap København Formation

After disentangling the pioneer from the permafrost microbial communities, we investigated potential modern analogues of the pioneer microbial communities. Firstly, we employed meta-Sourcetracker[24], trained on over 1,003 modern metagenomes, to identify the contribution of 33 different biomes to our samples (Supplementary Information 4, Supplementary Table 1). We further supplemented our analysis by employing a k-mer similarity method[25] on the reads retrieved from taxa identified as part of the pioneer communities. In both cases, the Aquatic:Freshwater:Sediment biome was the most significant contributor to all samples, with proportions as high as 80% in some cases (Fig. 3A and Supplementary Information 4, Supplementary Table 1). The second most significant contributor was Terrestrial:Source:Permafrost, with proportions higher than 25% in some samples. The biome Terrestrial:Soil:Forest soil was the third most significant contributor, with some samples having a proportion of 10%. While meta-Sourcetracker failed to detect the marine signal present in some of our locations[2], the k-mer-based method showed higher sensitivity, identifying a 10%

contribution from the biome Environmental:Aquatic:Estuary in a sample from location 75 (Supplementary Information 4, Supplementary Table 1).

Next, we explored the distribution of the reads extracted from the pioneer communities in each sample using the Genomes from Earth's Microbiomes (GEM) catalogue[26], which covers diverse habitats spanning all of Earth's continents and oceans and the glacier microbiome catalogue[22]. We used the estimated damage, ANI^R and detection to select those biomes that recruited more non-unique references with a detection higher than 50%. The biomes with the highest number of references (Fig. 3B) were consistent with the results from the microbial source tracking methods, with archaea having a detection rate of over 75% and a high ANI^R, despite high levels of the estimated damage. The Arctic peat soil biomes and soil had the highest number of references, which supports our efforts to recover the pioneer microbial communities. To increase the resolution of our analysis, we determined where our samples would fall on a permafrost thaw gradient. To do so, we utilised the metagenome-assembled genomes recovered from the Stordalen Mire[27] to recruit the reads from the pioneer communities. Our results show that fens, which are peat-forming wetlands, recruit the largest number of archaeal MAGs, with detection above 50% (Fig. 3C). We also found references with moderate detection in palsa, a permanently frozen peat (Fig. 3C), which aligns with the permafrost signal detected by the microbial source tracking methods and the genome collections read recruitment.

It is worth noting that the samples collected from the Kap København Formation aggregate diverse microbial communities sourced from upstream locations as well as in-situ micro-environments within the delta as known from nowadays delta systems[28]. These communities have been deposited over a period of 20,000 years, resulting in a blend of local and upstream contributions and the formation of a composite estimate that reflects the depositional history of the formation. This composite estimate serves as a representation of the main dominant biome during the Early Pleistocene period. Furthermore, the vegetation recovered on the site by Kjær et al.[2] reinforces our findings that the region experienced prolonged periods with annual mean temperatures well above freezing and consequently without permafrost. Among the abundant tree species identified by the ancient eDNA in the Kap København Formation were poplars (*Populus*). While the natural range of Balsam poplars (*Populus balsamifera*) commonly are below the treeline, it can today be found on the northern slope of the Brooks Range, an area characterised by continuous permafrost. However, these poplar groves are confined to river channels within a limited range of a few hundred meters. These areas often exhibit a 'thaw bulb,' a depression in

the permafrost table[29]. Other genera recovered from the samples, such as *Crataegus*, *Rhamnus*, *Thuja*, and *Kalmia*, have a distribution in northern temperate regions but are not adapted to boreal environments or capable of surviving, even as clones, on permafrost. In particular, *Kalmia angustifolia* and *Thuja occidentalis* are typically found in mixed deciduous forests of the Appalachia region, extending north to the southern boreal margin in Quebec. They are unable to withstand arctic winters or grow on permafrost. Hence, these genera most have colonised northern Greenland, by dispersing through Baffin and Ellesmere islands, during a period when the region was free of continuous permafrost. This suggests that they may have migrated north-eastward during a relatively warmer interval.

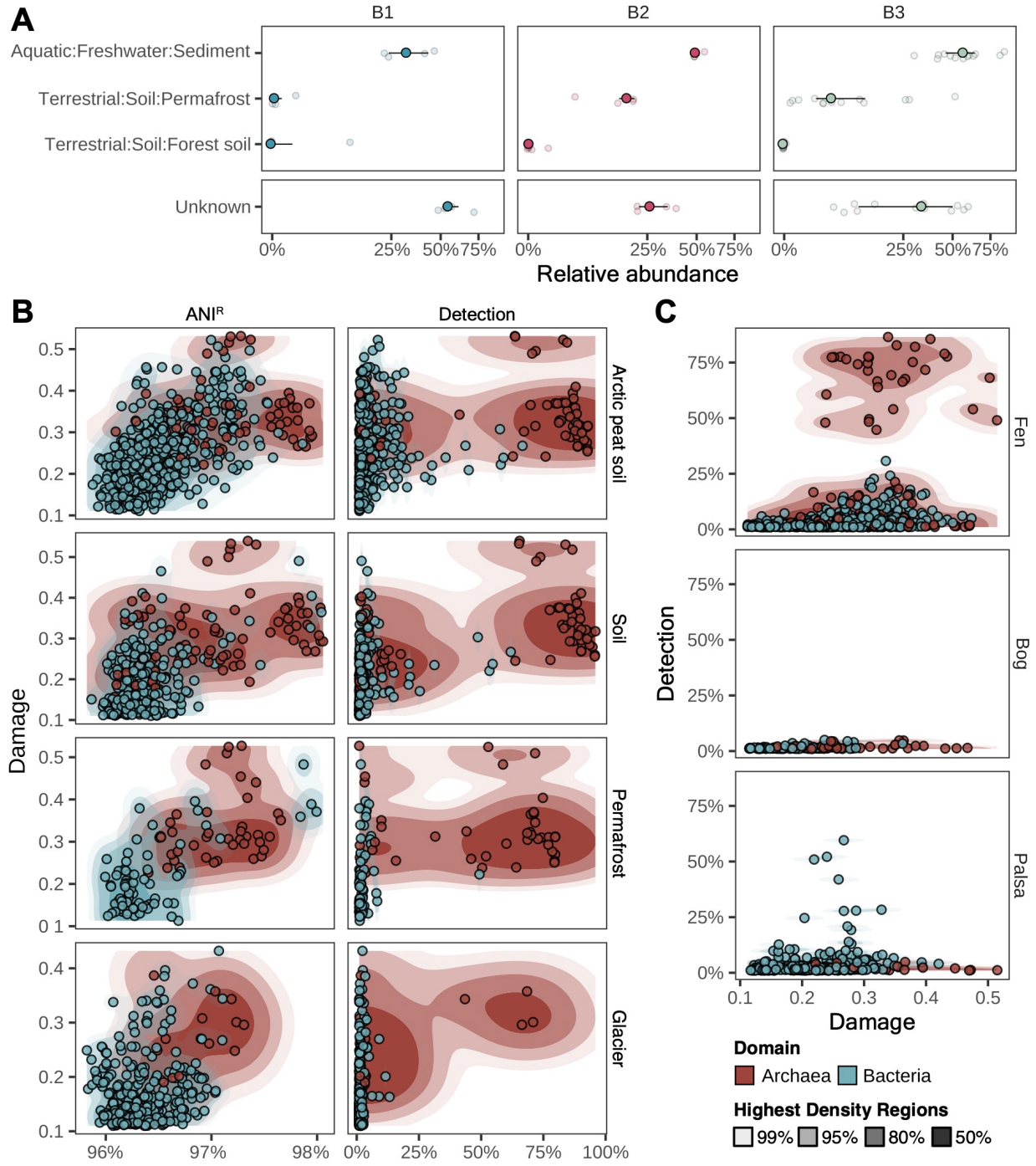


Fig. 3 | Kap København Formation modern analogue biomes A, Boxplot representing the estimated proportions by meta-SourceTracker from each "source" or biome for a single Kap København Formation sample or "sink". B, Highest density plots depicting the relationship between damage and the average nucleotide identity, as well as the detection (the proportion of the reference that is covered) after mapping reads from the damaged taxa against metagenome-assembled genomes recovered from the biomes described in Nayfach et al. 2020[26] and Liu et al. 2022[22]. C, Detection and damage of the recruited genomes through mapping the reads from

damaged taxa against metagenome-assembled genomes obtained from a thawing permafrost gradient.

Community structure of a 2-million-year-old microbiome

After establishing that the pioneer microbial communities from two million years ago share similarities with modern communities in palustrine wetlands, we analysed their structure and composition. Reconstructing and estimating the abundance of such an ancient microbiome poses a significant challenge due to various taphonomic and biological processes that the DNA has undergone over the last two million years[30,31]. Nevertheless, we were able to identify a diverse set of archaeal and bacterial taxa from our dataset. While our initial findings suggest that methanogenic archaea were the dominant group 2 million years ago at the Kab København Formation (Fig. 4), such a high relative abundance (averaging over 60% in most samples) of methanogens in the ancient microbiome is unlikely. While methanogens can exhibit higher abundances relative to other archaea in a thawing permafrost environment[32], it is important to note that in typical anaerobic microbial communities supporting methanogenesis, the abundance of methanogens is only 1-5%[33–37] of the overall community composition. Methanogens rely on the by-products of bacterial metabolism for energy, but bacteria consume most of the energy available in organic matter, leaving only a small fraction for methanogenesis[38]. This highlights the potential biases inherent in analysing deep-time ancient DNA. Nonetheless, considering the warmer conditions of the Early Pleistocene[2], it is reasonable to hypothesise the presence of methane-producing microbial communities resembling those found in wetlands or thawing permafrost upstream of the deposition site 2 million years ago.

We identified the presence of methane-related archaea from various genera with detection values of 75% or higher. These include the CO_2-reducing hydrogenotrophic g__Bog-38 (*Methanoflorens*), *Methanoregula*, and *Methanobacterium_A*. We also detected the acetoclastic methanogen *Methanothrix* and the versatile methanogen *Methanosarcina* (Fig. 4). Additionally, we observed the methanotroph g__Kmv04 (*Methanocomedenaceae*). We also found members of the bacterial phyla *Atribacterota* (g__CG2-30-33-13) and *Aerophobota* (g__AE-B3A), which are commonly found in methane-rich sediments across various environments such as temperate soils, deep marine sediments, and permafrost[39,40]. We detected several methanotrophic bacteria with high detection, including high-affinity methanotrophs like *Methyloceanibacter*[41–43] and genera from *Methylomonadaceae* such as *g__KS41*, which are commonly found in acidic forest soils[44], and *Methyloglobulus*, found in lake sediments[45]. Microorganisms involved in carbon and nitrogen

cycling processes were also identified. These include members of the families *Bacteroidaceae* (g__UBA1794 and *Bacteroides*) and *Clostridiaceae* (*Clostridium_AD*); the phylum *Myxococcota* (g__Fen-1088, found in symbiotic relationships with the arbuscular mycorrhizal fungal hyphae[46]), and members of the genus *Propionospira*, which produce propionate, acetate, and CO_2 as the main products of their carbohydrate fermentation[47,48]. Additionally, we detected members of the family f__UBA5619 (g__UBA5619), which could provide fermentation products to *Methanothrix* while removing H_2 and formate from the environment[49]. *Eubacteriaceae* (g__UBA1792), capable of degrading substrates such as fructose and mannose (Supplementary Table 2), and g__JADGOW01, possessing the necessary gene repertoire for acetogenesis (Supplementary Table 2), were also identified. We also recovered the signal of taxa involved in syntrophic relationships, such as members from the family *MBNT15* (g__CG2-30-66-27), which are commonly found in peatlands where they act as scavengers by fully mineralising small organic molecules produced during the microbial breakdown of complex polymeric compounds[50,51]. We detected ammonia-oxidizing archaea (AOA) from the *Nitrososphaeraceae* (g__UBA10452 and g__TA-21), which are mainly found in acidic polar and alpine soils[52] and a high-arctic glacier[53], respectively. Alongside the AOA, we also identified bacterial taxa involved in nitrogen fixation, including *Pseudolabrys* (73% detection rate) and *Bradyrhizobium*, as well as commamox bacteria from *Nitrospiraceae* (g__Palsa-1315) capable of complete ammonia oxidation[54]. Finally, we identified the archaeal *Thermoproteota* genus g__AD8-1 (70% detection rate) from the class *Bathyarchaeia* and the *Asgardarchaeota* genus g__6H3-1 from the class *Lokiarchaeia*, both of which inhabit marine sediments. The closest reference to our data of g__AD8-1 was reconstructed from permafrost samples with marine influence[55], while the closest reference from g__6H3-1 was isolated from deep-sea sediment samples of the Hikurangi Subduction Margin[56].

The preservation disparity between archaea and bacteria is striking, with archaea exhibiting preferred preservation over bacteria. Archaea, particularly methanogens, rely on specific compounds like CO_2, H_2, formate, acetate, and methanol supplied by fermenting bacteria, often in syntrophy. However, while methanogens are preserved, the supporting bacteria are not (Fig. 4). Multiple factors might contribute to the enhanced preservation of archaea. Firstly, environmental conditions, such as differences between local oxic and anoxic environments, could play a role. Secondly, most archaea possess distinct cell envelopes consisting of a lipid membrane externally covered by a paracrystalline surface (S-) layer composed of one or two proteins with strong intersubunit interactions[57]. Several features of the S-layers could be responsible for the enhanced preservation, namely, the S-layers can repair themselves upon

damage and are also found to serve as nucleation sites for mineral formation, resulting in encrustation[58]. Third, differences in the local chemical environment, influenced by the metabolic activity of archaea, could also contribute to preservation disparities. Finally, differences in chromatin structure could also have an influence on the differences in preservation we observe. DNA is complexed with histones in many archaea, including methanogens, making them among the most abundant proteins in *Methanobacteriales*[59]. This distinct chromatin organisation in archaea compared to bacteria may also be another factor that contributes to the observed differences in conservation patterns.

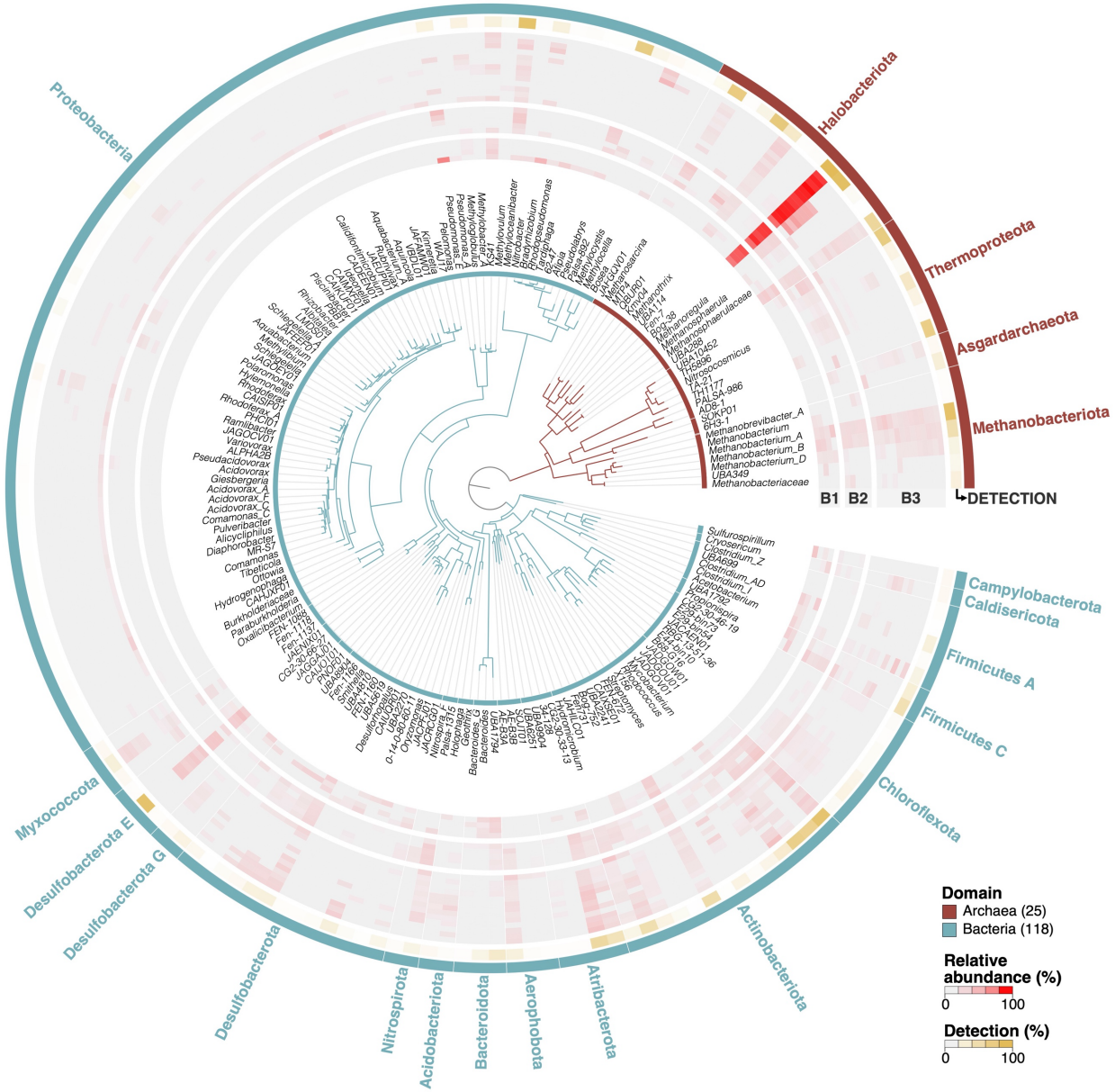


Fig. 4 | Phylogenomic distribution of the archaeal and bacterial damaged taxa Taxonomic

profile of the damaged identified taxa, which have been mapped into the GTDB phylogenomic tree at the genus level. Only genera that make up at least 1% of the taxonomic abundance at the family level are displayed. Each layer of the circle corresponds to a different sample grouped by member unit. The coloured sections of the circle indicate the relative abundance of each genus in the sample, with the intensity of the colour representing the transformed square root of the proportion. The outer layer corresponds to the maximum detection value for the references in a genus across samples.

Microbial carbon processing

Our taxonomic profiling analysis (Fig. 4) shows that the dominant group in our samples consists of CO_2-reducing hydrogenotrophic members along with other versatile methanogenic archaea. Furthermore, our results revealed the presence of high affinity methanotrophs in the samples. We complemented these taxonomic analyses with ecosystem-wide metabolic reconstructions[60] to identify the primary metabolisms associated with methanogenesis two million years ago. Peatlands store vast amounts of carbon, which is derived from plant polymers such as cellulose and hemicellulose. These polymers are degraded by microorganisms, forming glucose, xylose, and N-acetylglucosamine, producing low-molecular-weight alcohols and organic acids such as ethanol, propionate, acetate, and lactate, as well as H_2 and CO_2 via fermentation. These products then serve as substrates for hydrogenotrophic and acetoclastic methanogens, which utilise H_2/CO_2 and acetate, respectively[61].

We compiled a collection of standard and custom KEGG modules and specific CAZymes[27] (Carbohydrate-active enzymes) to reveal the different steps involved in carbon processing by the pioneer communities. Our analysis revealed evidence of all carbon processing steps in our samples (Fig. 5), including when we only used reads that map to the pioneer communities (Fig. 5, tiles with thicker borders). We used a highly conservative approach, starting from the translated searches using ultra-short damaged reads[62] to the identification of those modules with completion over 80%. In agreement with the taxonomic profiling, we identify pathways for CO_2-reducing hydrogenotrophic, acetoclastic, and methylotrophic methanogenesis, as well as methanotrophy (Fig. 5). The co-occurrence of these pathways suggests that a significant proportion of CH_4 might have been oxidised, thereby limiting emissions to the atmosphere two million years ago[43,63]. It should be noted that our results may underestimate the true functional potential due to the presence of cytosine deaminations, which can introduce stop codons and non-synonymous substitutions in the predicted open reading frames used in our translated searches[62]. Our stringent pathway identification thresholds, coupled with the presence of deaminations, likely contributed

to the failure to identify methane pathways in B1. However, we were able to detect methane pathways in some sites from B1 (acetoclastic methanogenesis in 75_205B and hydrogenotrophic methanogenesis in 75_205A) by lowering the pathway identification threshold to 0.7. Moreover, the average damage values for the methanogenic archaea in B1 were 0.48 ± 0.05, which are amongst the highest values estimated in our dataset, a pattern also observed in the eukaryotic fraction.

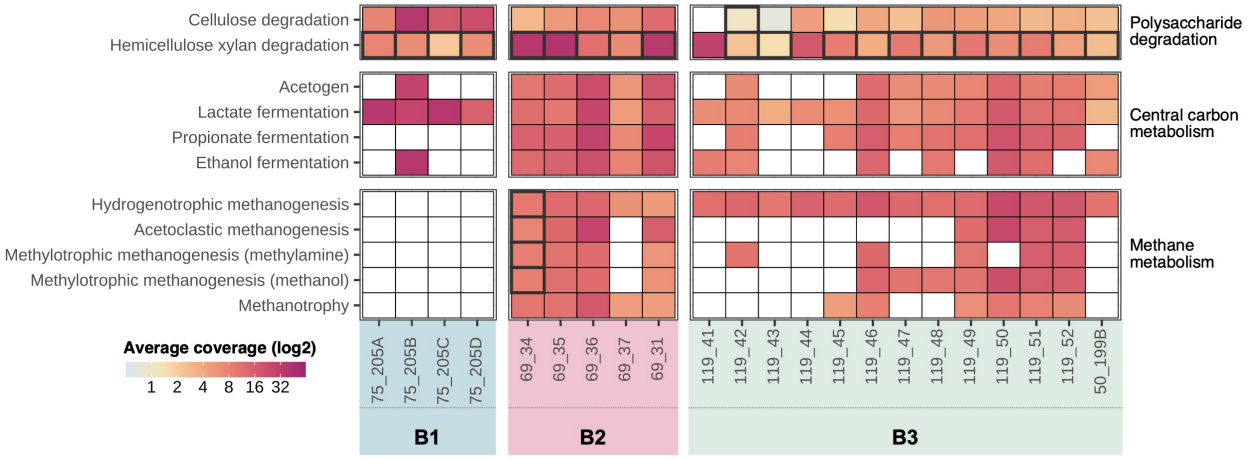


Fig. 5 | Carbon metabolism in the Kap København Formation Functional profile depicting pathways involved in the carbon metabolism in the permafrost. Tiles with thicker borders represent samples where the pathways were also detected using only reads that map to damaged references. White tiles indicate samples where no pathways have been detected.
1

The Kap København Formation virome

As viruses play a major role in structuring microbial communities across ecosystems[64], we further explored the community structure of the formation by characterising its virome. From the taxonomic profiling, we recovered the reads that mapped to references from the IMG/VR included in our database. To increase the confidence of the results, we only considered the references with a detection above 10%. We recovered 36 references, 15 damaged and 21 non-damaged (Fig. 6A), all belonging to the class *Caudoviricetes* (phylum *Uroviricota*), the dominant group of prokaryotic viruses. To further expand the reach of the homology detection, we searched the translated reads against the proteomes of viruses from the IMG/VR included in our database as well as a curated collection of mobile genetic elements (MGE) associated with methanogenic archaea[65]. After the search step, we identified which of the proteins have a higher likelihood to be present in a sample, and then selected those genomes that have over 20% of their proteins detected. This approach led to the detection of 231 reference viruses, expanding our view of the

two million-years virome. Although members of the *Caudoviricetes* again represented the dominant (70%) component of the virome (Fig. 6B), a considerable fraction could be affiliated with the realms *Monodnaviria* (viruses with ssDNA genomes) and *Varidnaviria* (non-tailed icosahedral viruses with double jelly-roll capsid proteins). Notably, the identified monodnaviruses belong to three different phyla specific to eukaryotic (*Cressdnaviricota*) and bacterial (*Hofneiviricota* [filamentous phages] and *Phixviricota* [phiX174-like phages]) hosts. Members of the *Cressdnaviricota* represent some of the simplest viruses associated with diverse eukaryotic hosts, including protists, fungi, plants and animals[66]. Finally, *Varidnaviria* was represented by viruses of the class *Megaviricetes* (phylum *Nucleocytoviricota*), which includes environmentally widespread eukaryotic viruses with big and giant dsDNA genomes. Finally, the identified methanogenic MGE, which include members of the *Caudoviricetes*, archaea-specific lemon-shaped viruses and several unclassified MGE (Fig. 6C), are associated with *Methanosarcinales*, *Methanobacteriales* and *Methanomicrobiales*, all of which were detected in our samples (Fig. 4).

On average, the viral fragments we retrieved exhibited an amino acid similarity (PAM30) of 84% to the references (Fig. 6D). Among them, the members of *Cressdnaviricota* displayed the greatest divergence from modern references, with similarities as low as 60%. These similarity values indicate a significant divergence of the viruses found in the formation when compared to modern references. These viruses would have gone undetected if only nucleotide assignment methods were employed. Both approaches identified 14 *Caudoviricetes* references (Fig. 6A), with three of them showing substantial detection (>50%) and damage (>0.3). These references, considering the amount of damage, show a high identity to the reference, with ANI^R exceeding 96%. Particularly noteworthy is the similarity to the reference IMGVR_UViG_2617271244_000001 (detection: 64%, damage: 0.34) with an ANI^R of 97% and an amino acid similarity of 95%. This virus is predicted to infect *Protochlamydia naegleriophila*[67], an obligate intracellular symbiont of the free-living amoeba *Naegleria*, commonly found in warm freshwater[68]. *Caudoviricetes*, *Cressdnaviricota* and the methanogenic viruses have the highest relative abundance, with certain samples surpassing 20% of the total viral fraction (Fig. 6D, Supplementary Table 3).

The amino acid searches also provided information about the functional potential of the viruses we recruited. Although a large proportion (71%) of the recruited viral proteins lack functional annotation, the remaining 29% of proteins validated the taxonomic assignment of the corresponding reads and provided valuable insights into the functional capacity of the virus community (Fig, 6D, Supplementary Table 3). In particular, the two largest functional categories

corresponded to taxon-specific virus structural proteins (16%) and proteins involved in DNA/RNA transactions (genome replication, transcription, recombination; 9.5%), respectively. We also identified proteins responsible for host cell lysis as well as toxin-antitoxin and restriction-modification systems. Finally, viruses encoded diverse auxiliary metabolic genes involved in various metabolic pathways (phosphoheptose isomerase, UDP-glucose dehydrogenase, phosphoadenosine phosphosulfate reductase, sulfatase-modifying factor enzyme, deoxyribonucleoside 5' monophosphate phosphatase, nitroreductase, glutamine amidotransferase, cation efflux protein), enzymes for modification of various substrates (acetyltransferase, nucleotide kinase, serine-threonine kinase, GtrB glycosyltransferase for O-antigen conversion), hydrolases (glycoside hydrolase, esterase/lipase, proteases) as well as proteins responsible for modulation of host responses (MazF-like growth inhibitor, sporulation stage III protein D, translational regulator).

We leveraged the ecosystem information associated with each reference in the IMG/VR data to gain insights into the environmental origins of the references identified through both approaches (Fig. 6E). Consistent with the source tracking analyses, most of the references originate from aquatic freshwater biomes. Notably, the references with the highest ANI^R and amino acid similarity are derived from environments that resemble permafrost conditions. The biome labelled Aquatic;Marine;Coastal corresponds to samples obtained from arctic subzero sea-ice, and cryopeg brines[69], while the Non-marine Saline and Alkaline;Saline biome represents samples collected from Lost Hammer Spring in Axel Heiberg Island situated in the Arctic Ocean, as well as various lakes in Antarctica.

Overall, these results suggest that viruses were an active component in the pioneer and permafrost microbial communities, and their composition in the 2-million-year-old sample has similarities to viral lineages that are found in contemporary environments.

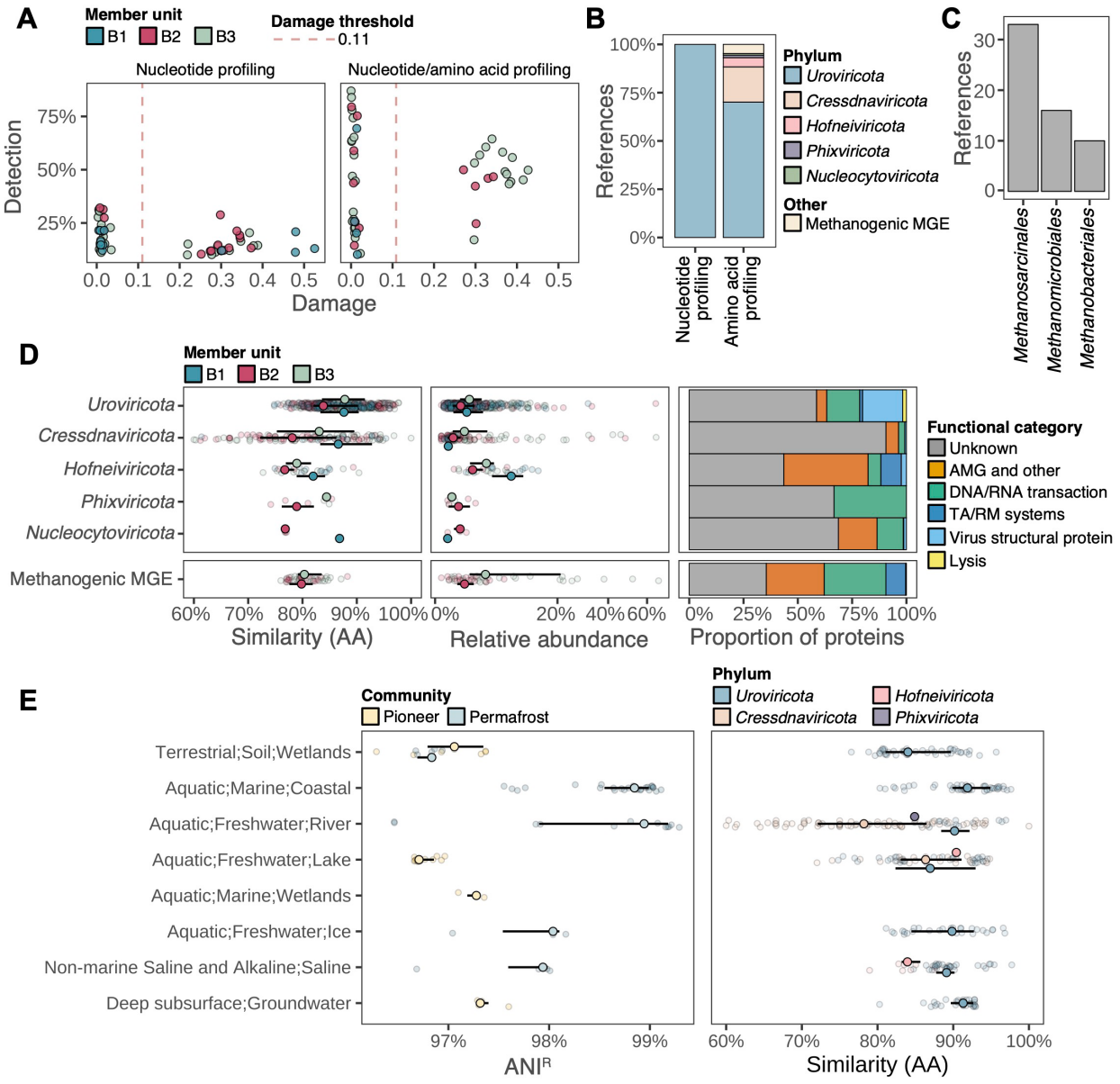
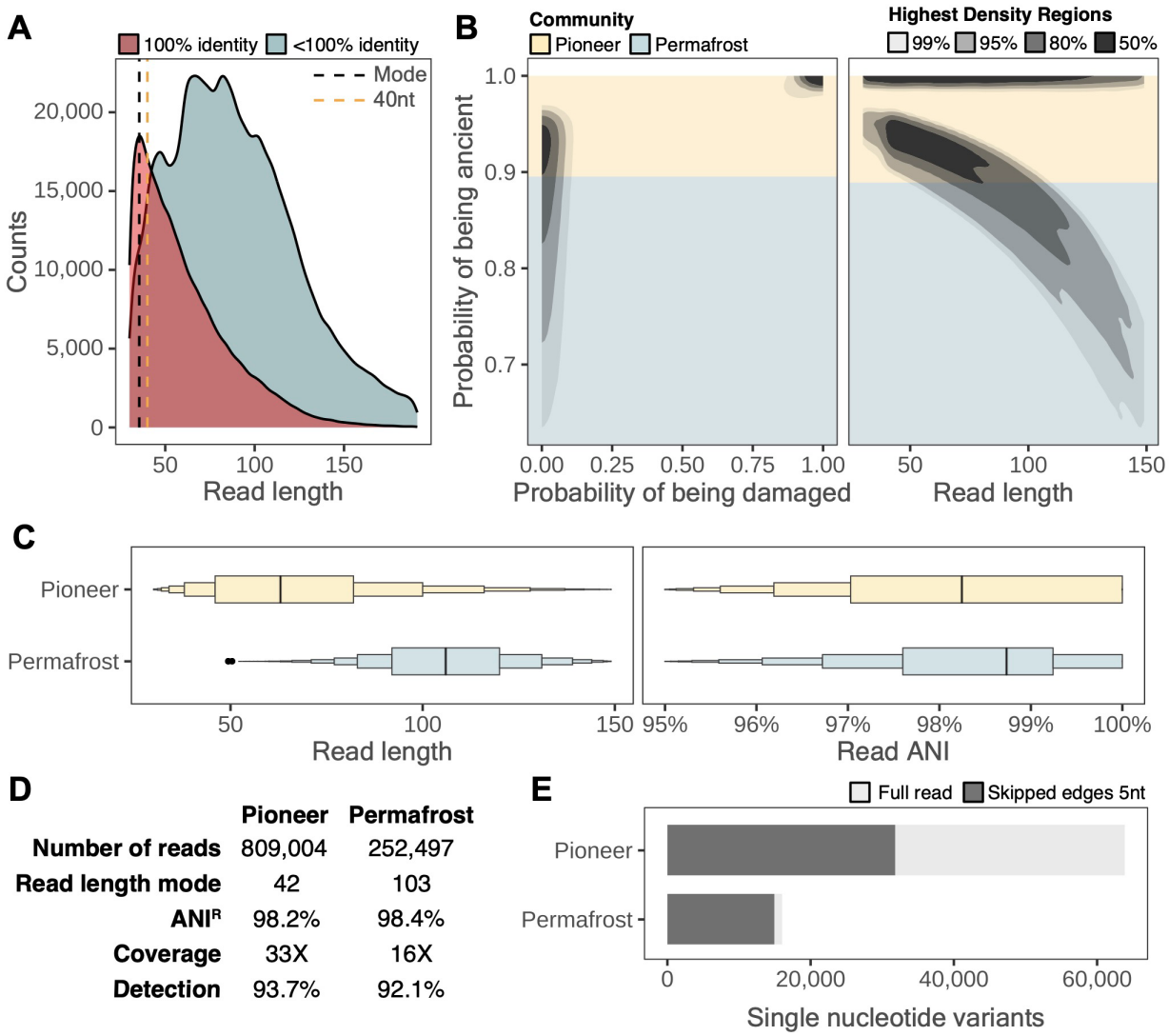


Fig. 6 | The Kap København Formation virome A, Each point depicts a viral reference identified in a sample by nucleotide profiling on the left panel, and by nucleotide and amino acid profiling on the right panel. Red dashed line depicts the damage threshold we used to delimit the pioneer from the permafrost communities. B, Proportion of references we detected in the whole dataset by phylum and by methanogenic mobile genetic elements. C, Number of references associated to each predicted host for the methanogenic MGEs. D, The two first panels show the distribution of the amino acid similarity values and inferred abundances for each viral reference identified in each sample. The third panel corresponds to the function categories of the proteins annotated to each group, phylum or methanogenic MGE. E, Environmental distribution of the different references we recruited provided by the ecosystem information in the IMG/VR v4 database[70]

The Kap København microbial seed bank

The remarkable sequence similarity between the pioneer microbial communities in Kap København and modern references retrieved from permafrost thawing gradients offers an exceptional opportunity to gain ecological and evolutionary insights from these data. Members of the genus *Methanoflorens* (g__Bog-38), which exhibited the highest recruitment of reads in our study, represents a particularly interesting case. *Methanoflorens*, CO_2-reducing methanogens, are widely distributed in high methane-flux habitats[32]. In the context of thawing permafrost, it exhibits a high abundance compared to other methanogenic microorganisms and emerges as a significant contributor to methane production[32]. The reference 3300025461_7, which recruited the majority of the reads in all samples of our study for *Methanoflorens*, is a metagenome-assembled genome derived from a drained thaw lake basin in the Arctic coastal plain near Barrow, Alaska[71]. To disentangle the signal of the pioneer community and the seed bank (surviving in the permafrost), we analysed the data collected from sample 119_50, where 3300025461_7 successfully recruited 1,114,032 reads (estimated damage: 0.26, detection: 94% and ANI^R: 98.2%). For an in-depth analysis of these reads we used a novel method[21] that allows us to calculate the probability of the read being ancient based on the information on nucleotide changes and the fragment length. To ensure accurate inference of the probabilities of being ancient, we initially excluded reads from our analysis that exhibited a 100% identity to the reference and had a read length smaller than 40 nt (modal read length + 5nt). This precautionary step minimised potential interferences caused by a skewed distribution of identical reads (Fig. 7A). An assumption of the method is the requirement for prior knowledge of the modern contamination rate in the sample. Given the inherent difficulty in estimating this rate for our sample type, we adopted a pragmatic approach. We employed modern contamination rates of 5% and 10% and then calculated the average of the resulting probabilities. The results of the inference highlight that we have a group of reads that have a high probability of being ancient (AN) and damaged (PD) and a second group that has a low probability of being damaged but a high probability of being ancient due to their read length (Fig. 7A, left panel). For a more formal characterization of the deamination patterns observed in each ancient DNA strand for each fraction data, here we estimated the Briggs parameters[72], namely $\boldsymbol{\delta}_d$ (deamination rates within the double-stranded regions), $\boldsymbol{\delta}_s$ (deamination rates within the single-stranded regions), $\boldsymbol{\nu}$ (nick occurring rate in the sample), and $\boldsymbol{\lambda}$ (overhang length distribution). The pioneer fraction exhibited $\boldsymbol{\lambda}$=0.44, $\boldsymbol{\delta}_d$=0.029, $\boldsymbol{\delta}_s$=1 and $\boldsymbol{\nu}$=0.11. On the other hand, the surviving fraction in the permafrost had $\boldsymbol{\lambda}$=0.23, $\boldsymbol{\delta}_d$=0, $\boldsymbol{\delta}_s$=0, $\boldsymbol{\nu}$=0.000000. The larger values of $\boldsymbol{\delta}_d$, $\boldsymbol{\delta}_s$, and $\boldsymbol{\nu}$ and smaller values of $\boldsymbol{\lambda}$ typically indicates a more

severe level of DNA damage. Thus, Briggs parameters suggest that the pioneer fraction exhibits clear indications of damage, unlike the permafrost fraction, confirming the ancient nature of reads from the sample 119_50 that matched to the reference 3300025461_7 with high levels of identity. We also investigated the association between the probability of a given read to be ancient and its length, where a gradual decrease in probability as a function of increasing read lengths is expected. To establish the ANI threshold for distinguishing read fractions associated to the pioneer or permafrost communities, we selected the lowest limit within the 50% highest density region, considering both PD and read length (Supplementary Table 4). We added the reads that were excluded due to their 100% identity to the reference to the pioneer read fraction, it is a common feature in ancient metagenomic studies that many short reads are 100% identical to the reference (Supplementary Table 5). The reads assigned to pioneer communities exhibited a significantly smaller read length (mode: 42 nt) than those assigned to permafrost communities (mode: 103 nt). Despite these differences, the ANI^R values remained remarkably consistent (Fig. 7C). In both cases, we recovered more than 92% of the modern reference genome, with coverage of 33X and 16X for the pioneer and permafrost fractions, respectively. Although, differences in coverage are unlikely quantitative since our extraction protocols prioritize the recruitment of short DNA fragments, which reduces DNA recovery from living organisms. Lastly, we implemented a new profiling method in anvi'o[73] to exclude the edges of short reads during the calculation of single-nucleotide variants (SNVs) for a more accurate characterization of genuine variants in ancient DNA by minimizing the impact of post-mortem changes. Using this approach, we re-profiled each fraction by excluding 5 nts from the read edges (Fig. 7D), and found 31,802 SNVs in the pioneer community, two times more than in the permafrost (14,933 SNVs). The higher number of SNVs in the pioneer community reflects the result of the deposition over 20,000 years, resulting in a blend of local and upstream organisms, while the reduction of observed SNVs in the permafrost fraction demonstrates how the community is shaped by selection for survival and growth under permafrost conditions.



	Pioneer	Permafrost
Number of reads	809,004	252,497
Read length mode	42	103
ANI^R	98.2%	98.4%
Coverage	33X	16X
Detection	93.7%	92.1%

Fig. 7 | The Kap København microbial seed bank A, Read length distribution of the reads mapping to the reference 3300025461_7 from the genus g__Bog-38 in sample 119_50. The orange dashed line depicts the threshold we used to exclude the sequences mapping to the reference at 100%. B, Highest Density Regions plot showing the relationship between the probability of being ancient and the probability of being damaged and the read length. We used a probability of being ancient of 0.89 as the threshold to split the reads between pioneer and permafrost communities. C, Letter-value plots depicting the read length and read Average Nucleotide Identity to the MAG 3300025461_7 for both fractions. D, Table summarising the statistics of the mapping of each of the fraction to 3300025461_7. E, Number of single nucleotide variants for each fraction as inferred by anvi'o, including (light grey) and excluding 5nt on the read's edges (dark grey).

Conclusion and discussion

Our study addresses some of the key challenges associated with reconstructing microbial communities using ancient eDNA, providing unprecedented insights into the structure and function of the microbial and viral communities that existed 2 million years ago in the Kap København Formation. By distinguishing the ancient pioneer microbial communities from the permafrost microbial communities, we have strengthened the evidence of a largely ice-free North Greenland without extensive permafrost during that time period, complementing existing climatic, plant and animal records[2,9,74]. Furthermore, the detection of methanogenic archaea and the reconstruction of carbon processing pathways suggest that high-latitude ecosystems may have contributed to a moderate strengthening of methane emissions through the northward expansion of wetlands (Supplementary Information 5). This observation aligns with ice core data from the Early Pleistocene, indicating that atmospheric methane concentrations and their glacial and interglacial variability were not significantly higher than those observed after the mid-Pleistocene transition[75] (1.2 – 0.8 Myr). Importantly, the potential emergence of wetlands in the Arctic tundra has implications for accounting for additional sources of greenhouse gases in the future.

The recovery of pioneer microbial communities, along with the plant and animal communities that existed before permafrost formation, offers a rare opportunity to observe the complete picture of an environment that flourished two million years ago. The study of plant and animal eDNA from the Kap København Formation[2] shows an environment consisting of a mixture of temperate and arctic taxa. Thus, very distinct from what is seen in the Arctic today despite of global warming. This, contrast our findings of the pioneer microbial community showing high similarities to that of thawing arctic sites suggesting that the microbial composition is the first to change as a consequence of rising temperatures. This is particularly significant given the potential for a resurgence of these past environments in the near future due to warming. Our results provide support for the re-emergence of certain pioneer methanogens that are currently present in the seed bank within permafrost. These methanogens, that have remained inactive but alive, have the potential to become active and contribute to methane production if the environmental conditions become favourable as has been already observed in multiple locations[27,71,76,77]. Our detailed analysis to distinguish between the sequences originating from pioneer methanogens and those found in the seed banks within the permafrost reveals a strikingly high level of identity between the two fractions and a modern reference, as evidenced by an ANI^R exceeding 98%. Additionally, the permafrost fraction exhibits a significantly lower number of SNVs compared to the pioneer fraction providing compelling evidence for the impact of selection driven by the unique

conditions of the permafrost environment and leading to a preservation of genetic traits that are advantageous for survival and adaptation in the frozen environment[78]. The remarkable similarity observed of the pioneer and permafrost fraction to a modern reference separated by a time span of two million years and a geographical distance of 3,000 km, as this metagenome-assembled genome was recovered from a drained thaw lake basin in Alaska[71], is an observation that supports a conceivable idea that has been elusive to demonstrate for microbes in geological timescales: environmental change does not trigger a response from every member of a given habitat the same way, and in fact for some taxa, the aspects of the environment they were adapted to survive is not changing, at all. These analyses demonstrate the opportunities hidden within the evolutionary signal that is present in the microbial fraction of ancient metagenomes to identify genetic markers of the rate and direction of change across time to complement conventional ecological parameters. Our observations on these methanogens also suggest the existence of the so called time-travelling microbes[79,80] that appear and disappear in sync with the glacial and interglacial periods, operating over million-year time scale. The pioneer methanogens we have recovered, dating back two million years, appeared during the period when the Kap København Formation was deposited, in a largely ice-free Greenland without extensive permafrost. During this time, the landscape was predominantly occupied by palustrine wetlands and boreal forests. These favourable conditions persisted until the onset of the next glacial period but may also have existed for shorter time periods sporadically since the sediments became exposed subaerially after 0.8-1.2 million years ago (Fig. 1D), if warm long-duration interglacials resulted in substantial permafrost thawing reaching the sample sites. Currently, as the Arctic gets warmer and permafrost thaws, these time-travelling microbes are reawakening from their dormant state, traversing through time, and resurfacing in the present.

Our study raise captivating questions concerning the mechanisms of community assembly and the roles played by dispersal, genetic drift, and evolutionary diversification[81]. By further investigating the intricate dynamics of these ancient microbial communities and their adaptations to changing environments, we may gain a deeper understanding of the interplay between climate change, microbial life and shed light on the intricate web of interactions that shape our planet's past, present, and future.

Material and methods

Metagenomic data

We retrieved the shotgun metagenomic data from Kjær et al.[2] (ENA project accession PRJEB55522), and we pre-processed it according to the same protocol described in Kjaer et al.. We analysed 41 samples (53 metagenomic libraries) distributed across three stratigraphic units B1, B2 and B3, spanning 5 different localities (50, 69, 74a, 74b and 119). We created a short version of the sample names used by Kjaer et al.[2] (Supplementary Table 5)

Post-depositional uplift history of Kap København Formation

To determine the timing of when the shallow marine Kap København Formation emerged above sea level and became permafrozen, we conducted a post-depositional uplift history modelling. We adopted a steady uplift rate assumption for the broader region, which resulted from the flexural isostatic response of the solid Earth to the erosional unloading caused by fjord incision[18]. According to our model, the upper undisturbed sediments of the Kap København Formation were deposited at approximately 15 meters of water depth around 1.93 million years ago[2] and subsequently uplifted to their present elevation of about 165 meters above sea level, with the different sample sites being exposed subaerially between 0.8-1.2 million years ago.

Taxonomic database generation

We utilised a simplified version of the taxdb-integration workflow (https://github.com/chassenr/taxdb-integration) to compile a unified taxonomy with ten levels that combined NCBI and GTDB taxonomies for our genomic references. We sourced data from various genomic datasets, including NCBI[82], PhyloNorway plastid sequences[5], complete GTDB v207[83], and high-quality IMG/VR v4 genomic data[70]. We also incorporated metagenome-assembled genomes from TARA Oceans[84], Genomes from Earth's Microbiomes[26], the glacier microbiome catalogue[22], and those recovered from a permafrost study[27].

To ensure consistency with GTDB v207, we re-annotated the bacterial and archaeal metagenome-assembled genomes using gtdb-tk v2[85]. Subsequently, we utilised derep-genomes[62] (https://github.com/aMG-tk/derep-genomes) to de-replicate the assemblies at the species or vOTU level for viruses. We then employed the viral, archaeal, bacterial, and organelle dereplicated data to construct a Bowtie2 database[20]. Prior to the database construction, we concatenated the contigs of each assembly with stretches of 50Ns.

Functional database generation

To generate a non-redundant version of the KEGG GENES[86] database (downloaded in February 2022), we clustered the amino acid sequences of each KEGG orthologous entry using MMseqs2 (version b0b8e85f3b8437c10a666e3ea35c78c0ad0d7ec2). We used the following parameters for clustering: -c 0.8, --min-seq-id 0.9, --cov-mode 0, and --cluster-mode 2. To ensure reproducibility, we developed a Snakemake workflow[87], which is available at https://github.com/aMG-tk/kegg-db-setup.

Taxonomic profiling

To recover the taxonomic profiles from ancient metagenomic data, we employed the taxonomic module of aMAW, an ancient metagenomics analysis workflow[62]. Briefly, we converted the reads into super-reads using Tadpole, a kmer-based assembler included in the BBTools software suite[88], with strict parameters. The resulting super-reads were then dereplicated using VSEARCH[89] –fastx-uniques. We used the de-replicated reads for functional and taxonomic profiling. Then we mapped the super-read dereplicated reference sequences to the original quality-controlled ancient reads and using Bowtie2, aMAW mapped the reads against the database we generated for this study. We removed mapping duplicates using the MarkDuplicates program from Picard[90], and filtered the BAM files with the filterBAM program[62] (https://github.com/aMG-tk/bam-filter) using the following parameters: -N -g auto -e auto -n 100 -b 0.75 -c 0.4 -A 94 -a 90 --read-length-freqs --read-hits-count --include-low-detection --min-breadth 0.01. Finally, we estimated the post-mortem damage using Bayesian estimates in metaDMG[21] in local and LCA modes. For the specific versions of the program used check the files in {ZENODO_LINK}

Best search parameters estimation

To determine the optimal parameters for different searches, we utilised a workflow (https://github.com/aMG-tk/aMGSIM-smk) that combined the aMAW and aMGSIM (https://github.com/aMG-tk/aMGSIM) tools to generate synthetic metagenomes. We used ten samples from the Kap Kobenhavn formation (Supplementary Table 6) to model the community composition (bacteria, archaea, and viruses), the fragment length distribution, and the damage patterns for each reference. Each synthetic metagenome contained a maximum of 1000 damaged references (determined using Bayesian damage estimates), 500 non-damaged references, and 10 million reads. We then annotated these synthetic metagenomes using the aMAW taxonomy module, exploring different bowtie2 parameters (-N, -k [100, 250, 500, 750, 1000]), read ANI

filtering thresholds (92%, 93%, 94%, 95%, 96%), and breadth filtering. We evaluated the sensitivity and specificity of the taxonomic profiling, the abundance estimations, and the damage estimations using precision, recall, F1, and F05[91] for classification problems, and Spearman correlation and median absolute error for quantitative comparisons.

In addition, we utilised the aMGSIM tool to track damage at the codon level to benchmark the MMseqs2 translated searches.

Abundance table generation

We used filterBAM to estimate the taxonomic abundances[92] based on the 80% truncated average depth (TAD80). Briefly, we estimated the number of reads in the TAD80 region using the formula N = (G * C) / L, where G stands for the length of the region where we estimated the TAD80, C for the TAD80 coverage value and L for the average read length mapped to the reference. We then estimated the taxonomic abundance by normalising the number of reads by the length of the region where we estimated the TAD80 and scaled by 1 million.

As we used the *--include-low-detection* parameter in filterBAM, there will be cases when the TAD80 estimated taxonomic abundance will be 0. In this case, we used the taxonomic abundances estimated by the number of reads mapped normalised by the length of the reference and scaled by 1 million.

Damage threshold estimation

To determine the intervals of postmortem damage and the minimum number of reads required for reliable damage estimates of non-eukaryotic references, we utilised metaDMG to estimate damage in local mode and with a significance threshold of at least 2, focusing on the estimates for chloroplasts and mitochondria references. We analysed the lower tails of the damage estimates distribution for all samples and references combined using letter values[93]. Subsequently, we only considered those references that were classified as damaged in all downstream analyses.

Contamination identification and removal

We followed the same processing procedure for the 13 negative extraction- and library controls as for the Kap Kobenhavn Formation samples. For each reference found in a control, we assessed the number of mapped reads and their damage patterns and excluded them from further analysis using a similar approach to that of Kjær et al.[2].

Microbial source tracking analyses

To create the source dataset for the source tracking analyses, we used getBiomes[62] (https://github.com/aMG-tk/get-biomes) to retrieve the biome information and associated raw sequence data from MGnify[94] and ENA[95]. We ran getBiomes with the parameters *--ena-filter {library_layout: PAIRED, library_strategy: WGS, library_source: METAGENOMIC, library_selection: RANDOM, read_count: 10000000}' --combine --exclude-terms human,16S*. Raw sequences were then processed using the aMAW-qc pipeline with modern metagenomics settings. This involved using fastp[96] to filter low-quality reads and selecting the forward pair for any read pairs that did not merge. The resulting metagenomes were then taxonomically annotated using aMAW with modern metagenomics settings, which disabled the read-extension step and set the read ANI to 95%.

We used the same database and procedure that we used to annotate, filter, and estimate the taxonomic abundances of the sinks. We merged the taxonomic annotations of the sources and sinks and filtered the merged dataset by removing samples with less than 10,000 counts and species that were observed less than three times in at least 20% of the samples, with a coefficient of variation smaller than 3 and a mean proportion over all samples smaller than 1e-5. This filtering step was performed to ensure that only species that were consistently present and abundant in the samples were included in the final dataset. The resulting tables were exported into BIOM objects that were used as input for meta-Sourcetracker[24] using the parameters *--sink_rarefaction_depth 0 --source_rarefaction_depth 0 --per_sink_feature_assignments --restarts 100 --draws_per_restart 5 --diagnostics*.

We used decOM[25] with default parameters as a complement to meta-Sourcetracker. To create a custom k-mer matrix, we employed the kmtricks[97] pipeline, using the same sources as meta-Sourcetracker. The parameters we used for kmtricks[97] were *--kmer-size 29 --mode kmer:pa:bin --nb-partitions 20000 --restrict-to-list 1000 and --recurrence-min 3*.

Extraction of reads from damaged references

We used dReads[62] (https://github.com/aMG-tk/dmg-reads) to separate the reads into different domains (Eukarya, Bacteria, Archaea, and Viruses) and classify them based on their damage status. For the read extraction, we used the parameters *-f '{ damage: 0.11, significance: 2}' --fb-filter '{breadth: 0.01, n_reads: 100}' --rank '{domain:[d__Bacteria, d__Archaea, d__Viruses, d__Eukaryota]}'*. The number of reads and damage values were estimated based on our analysis of Eukaryotic damage estimates.

Biome-associated metagenomic read-recruitment analyses

To investigate the distribution of reads found in damaged references across multiple biomes, we constructed two additional bowtie2 databases. The first database consisted of MAGs from the Genomes from Earth's Microbiomes catalogue[26] and the glacier microbiome catalogue [22]. For the second database, we aimed to obtain a comprehensive overview of a permafrost thaw gradient (palsa, bog, fern) by utilising MAGs recovered from Woodcroft et al.[27]. We employed the same exact mapping, filtering (expected breadth ratio >= 0.25), and damage estimation strategy used in our previous analyses.

Functional profiling

We utilised the functional profiling module of aMAW to annotate the de-replicated super-reads with MMseqs2[98], using fine-tuned parameters for ancient DNA amino acid searches (*--comp-bias-corr 0 --mask 0 -e 1e-5 --exact-kmer-matching 1 --sub-mat PAM30.out -s 3 -k 6 --spaced-kmer-mode 1 --spaced-kmer-pattern 11011101 --min-length 15 --format-mode 2 -c 0.8 --cov-mode 2 --min-seq-id 0.6*). To perform the annotation, we utilised the non-redundant KEGG GENES database[86] and dbCAN2[99] v11 gene sequences (CAZyDB.08062022). We applied xFilter[62], that implements a modified version of FAMLI[100] to identify the KEGG/dbCAN2 genes most likely present in the samples with the parameters *-n 25 -b 20 -e 1e-5 --breadth-expected-ratio 0 -f depth_evenness --depth-evenness 1*. For KEGG-related results, xFilter produced output that was compatible with the enzymes-txt mode of anvi-estimate-metabolism in anvi'o[73].

We expanded the standard KEGG modules collection in anvi'o using anvi-setup-user-modules to include a set of custom modules (Supplementary Data X) designed to explore carbon metabolism in permafrost[27]. We ran anvi-estimate-metabolism with the parameters *--add-coverage --output-modes modules,module_paths,module_steps,hits --include-kos-not-in-kofam --user-modules*, and we parsed the output to identify pathways with 100% completeness.

Furthermore, we extracted hits to CAZy families related to cellulose degradation (GH5, GH9, 3.2.1.4, GH51, GH6, GH7, GH48, 3.2.1.91), xylan degradation (GH5, GH8, GH10, GH11, GH43, 3.2.1.8, GH3, GH30, GH39, GH52, GH54, GH116, GH120, 3.2.1.37, GH67, GH115, 3.2.1.139, CE1, CE2, CE3, CE4, CE5, CE6, CE7, CE12, 3.1.1.72), and xylose degradation (3.2.1.37, GH3, GH30, GH39, GH10, GH43).

Reference metabolism estimation

We used the program *anvi-estimate-metabolism* from anvi'o[60] with default parameters to estimate the metabolisms of the archaeal and bacterial references that recruited reads in our study

(Supplementary Table 7). Note that we used a version of KEGG downloaded in January 2023 (for reproducibility, the hash of the KEGG snapshot available via `anvi-setup-kegg-kofams` is d20a0dcd2128)

Virome characterisation

To increase the sensitivity of the viral taxonomic classification we envisioned an approach to identify which viral reference is most likely to be present in a sample. We created a protein database containing 15,948,451 sequences from the viral sequences from IMG/VR v4 present in the database we used for the nucleotide-based taxonomic profiling and we enriched it with 16,441 proteins from a curated collection of mobile genetic elements (MGE) associated with methanogenic archaea. We used MMseqs2 with the fine-tuned parameters for ancient DNA amino acid searches (*--comp-bias-corr 0 --mask 0 -e 1e-5 --exact-kmer-matching 1 --sub-mat PAM30.out -s 3 -k 6 --spaced-kmer-mode 1 --spaced-kmer-pattern 11011101 --min-length 15 --format-mode 2 -c 0.8 --cov-mode 2 --min-seq-id 0.6*) and then filtered the results with xFilter[62] to identify the most likely proteins present in the samples with the parameters *-n 25 -b 20 -e 1e-5 --breadth-expected-ratio 0 -f depth_evenness --depth-evenness 1*. Then we estimated the proteome completion ratio for each reference and selected all the ones with a value over the 10%. We functionally annotated the selected proteins using HHblits using two iterations in combination with PFAM v35[101] and PHROG v4[102]. We followed a similar approach than in Vanni et al.[103] to remove overlapping domains and select those hits with a probability over 90% and with a target or query coverage larger than the 40%. We manually curated the results and classified them in six functional categories: “Unknown”, “Virus structural protein”, “DNA/RNA transactions/processing”, “Lysis”, “TA/RM systems” and “AMG and other”.

Estimation of probabilities of being ancient and being damaged

First, we extracted reads from damaged references and aligned them against the taxonomic profiling database. Unlike previous alignments, we did not use the -k option in bowtie2 and reported the best alignment. The filtering and damage estimation steps were conducted as described earlier. We then used the program getRPercId[62](https://github.com/aMG-tk/get-read-percid) to subset the BAM files to the reference 3300025461_7 and calculate for each the read length and average nucleotide identity (ANI). After exploring the read length and ANI, we subset the BAM files to the reference 3300025461_7 and calculate the read length and average nucleotide identity (ANI) for each read. We used this BAM to infer the probability of each read to be ancient and damaged. Briefly, the method is based on the counts of nucleotides changes

(including but not limited to C to T and G to A) at the first and last 15 cyclic positions of the strands and the position-specific error rates, the tool conducts a multinomial regression to estimate the four parameters that describe the deamination pattern of the ancient DNA samples (similar parameters as mentioned in Briggs et al.[72]). Under the assumption that the length distributions of the ancient indigenous strands and modern contaminate strands are distinguishable, the tool will then go through each strand within the length range [30,150), and estimate the best-fit ancient and modern length distributions if the prior knowledge of the modern contamination rate of the sample is provided. The previous four estimates can help to calculate the deaminated probability conditioned on the focal strand is ancient (AN), and the posterior probability of being ancient for each focal strand given its information on nucleotide changes and the fragment length (PD). We used the following parameters for the inference -isrecal 1 -model b -eps 0.05 and -eps 0.1. After identifying the pioneer and permafrost fractions, we estimated the Briggs parameters. Lastly, we used a new profile mode in anvi'o especially designed for ancient DNA to infer the SNVs in each fraction where we exclude a certain number of positions from the read edges using the option --skip-edges 5.

Methane budget estimation

TBA

Biomarker analyses

TBA

Acknowledgements

The authors thankfully acknowledge the computer resources and the technical support provided by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B). R.L-P. was funded by a Ramón y Cajal grant (RyC2021-031775-I) from the Spanish Ministerio de Ciencia e Innovación (MCIN/AEI/10.13039/501100011033) and the European Union («NextGenerationEU»/PRTR). EW thanks the Lundbeck Foundation, the Carlsberg Foundation, the Danish National Research Foundation, the Novo Nordisk Foundation, and the Wellcome Trust for financial support and the St. John's College, Cambridge for providing an environment for scientific discussion and thought. IV acknowledges support from the National Science Foundation Graduate Research Fellowship under Grant No. 1746045. GB was supported by the French National Agency for Research Grants (Methevol, ANR-19-CE02-0005-01). M.S. acknowledges support from the National Research Foundation of Korea (grants 2021R1C1C102065), the Samsung DS Research Fund, and the Creative-Pioneering Researchers Program through Seoul National University. VKP is supported by a research grant (15467) from the Danish foundation VILLUM FONDEN. Additional support was provided by Germany's Excellence Strategy (EXC-2077), project 390741603 "The Ocean Floor – Earth's Uncharted Interface". The authors thank Rayan Chikhi and Camila Duitama González for their insightful discussions on effectively utilizing decOM.

References

1. Willerslev, E. *et al.* Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* **300**, 791–795 (2003).
2. Kjær, K. H. *et al.* A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. *Nature* **612**, 283–291 (2022).
3. Willerslev, E. *et al.* Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature* **506**, 47–51 (2014).
4. Pedersen, M. W. *et al.* Postglacial viability and colonization in North America's ice-free corridor. *Nature* **537**, 45–49 (2016).
5. Wang, Y. *et al.* Late Quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature* 1–7 (2021).
6. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).
7. Singh, B. K., Bardgett, R. D., Smith, P. & Reay, D. S. Microorganisms and climate change: terrestrial feedbacks and mitigation options. *Nat. Rev. Microbiol.* **8**, 779–790 (2010).
8. Capo, E. *et al.* Environmental paleomicrobiology: using DNA preserved in aquatic sediments to its full potential. *Environ. Microbiol.* **24**, 2201–2209 (2022).
9. Funder, S. V. *et al.* Late Pliocene Greenland - The Kap København Formation in North Greenland. *Bull. Geol. Soc. Den.* (2001).
10. Willerslev, E. *et al.* Long-term persistence of bacterial DNA. *Curr. Biol.* **14**, R9-10 (2004).
11. Johnson, S. S. *et al.* Ancient bacteria show evidence of DNA repair. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 14401–14405 (2007).
12. Vass, M. & Langenheder, S. The legacy of the past: effects of historical processes on microbial metacommunities. *Aquat. Microb. Ecol.* **79**, 13–19 (2017).
13. Abramov, A., Vishnivetskaya, T. & Rivkina, E. Are permafrost microorganisms as old as permafrost? *FEMS Microbiol. Ecol.* **97**, (2021).
14. Dehkharghani, A., Waisbord, N. & Guasto, J. S. Self-transport of swimming bacteria is impaired by porous microstructure. *Communications Physics* **6**, 1–9 (2023).
15. Fenchel, T. Motility of bacteria in sediments. *Aquat. Microb. Ecol.* **51**, 23–30 (2008).
16. Mitchell, J. G. & Kogure, K. Bacterial motility: links to the environment and a driving force for microbial physics. *FEMS Microbiol. Ecol.* **55**, 3–16 (2006).
17. Bhattacharjee, T. & Datta, S. S. Bacterial hopping and trapping in porous media. *Nat. Commun.* **10**, 2075 (2019).
18. Pedersen, V. K., Larsen, N. K. & Egholm, D. L. The timing of fjord formation and early

glaciations in North and Northeast Greenland. *Geology* **47**, 682–686 (2019).

19. Christ, A. J. *et al.* A multimillion-year-old record of Greenland vegetation and glacial history preserved in sediment beneath 1.4 km of ice at Camp Century. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
20. Longmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie2. *Nat. Methods* **9**, 357–359 (2012).
21. Michelsen, C. *et al.* metaDMG – A Fast and Accurate Ancient DNA Damage Toolkit for Metagenomic Data. *bioRxiv* 2022.12.06.519264 (2022) doi:10.1101/2022.12.06.519264.
22. Liu, Y. *et al.* A genome and gene catalog of glacier microbiomes. *Nat. Biotechnol.* **40**, 1341–1348 (2022).
23. Margesin, R. & Collins, T. Microbial ecology of the cryosphere (glacial and permafrost habitats): current knowledge. *Appl. Microbiol. Biotechnol.* **103**, 2537–2549 (2019).
24. McGhee, J. J. *et al.* Meta-SourceTracker: application of Bayesian source tracking to shotgun metagenomics. *PeerJ* **8**, e8783 (2020).
25. González, C. D. *et al.* decOM: Similarity-based microbial source tracking of ancient oral samples using k-mer-based methods. *bioRxiv* 2023.01.26.525439 (2023) doi:10.1101/2023.01.26.525439.
26. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0718-6.
27. Woodcroft, B. J. *et al.* Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**, 49–54 (2018).
28. Schneider, J., Grosse, G. & Wagner, D. Land cover classification of tundra environments in the Arctic Lena Delta based on Landsat 7 ETM+ data and its application for upscaling of methane emissions. *Remote Sens. Environ.* **113**, 380–391 (2009).
29. Bockheim, J. G., O'Brien, J. D., Munroe, J. S. & Hinkel, K. M. Factors affecting the distribution of Populus balsamifera on the north slope of Alaska, U.s.a. *Arct. Antarct. Alp. Res.* **35**, 331–340 (2003).
30. Freeman, C. L. *et al.* Survival of environmental DNA in sediments: Mineralogic control on DNA taphonomy. *bioRxiv* 2020.01.28.922997 (2023) doi:10.1101/2020.01.28.922997.
31. Wasmund, K. *et al.* Genomic insights into diverse bacterial taxa that degrade extracellular DNA in marine sediments. *Nat Microbiol* **6**, 885–898 (2021).
32. Mondav, R. *et al.* Discovery of a novel methanogen prevalent in thawing permafrost. *Nat. Commun.* **5**, 3212 (2014).
33. Borrel, G. *et al.* Stratification of Archaea in the deep sediments of a freshwater meromictic

lake: vertical shift from methanogenic to uncultured archaeal lineages. *PLoS One* **7**, e43346 (2012).

34. Schwarz, J. I. K., Eckert, W. & Conrad, R. Community structure of Archaea and Bacteria in a profundal lake sediment Lake Kinneret (Israel). *Syst. Appl. Microbiol.* **30**, 239–254 (2007).
35. Purdy, K. J., Nedwell, D. B. & Embley, T. M. Analysis of the sulfate-reducing bacterial and methanogenic archaeal populations in contrasting Antarctic sediments. *Appl. Environ. Microbiol.* **69**, 3181–3191 (2003).
36. Zepp Falz, K. *et al.* Vertical distribution of methanogens in the anoxic sediment of Rotsee (Switzerland). *Appl. Environ. Microbiol.* **65**, 2402–2408 (1999).
37. Prasitwuttisak, W., Hoshiko, Y., Maeda, T., Haraguchi, A. & Yanagawa, K. Microbial Community Structures and Methanogenic Functions in Wetland Peat Soils. *Microbes Environ.* **37**, (2022).
38. Chan, O. C. *et al.* Vertical distribution of structure and function of the methanogenic archaeal community in Lake Dagow sediment. *Environ. Microbiol.* **7**, 1139–1149 (2005).
39. Glass, J. B. *et al.* Microbial metabolism and adaptations in Atribacteria-dominated methane hydrate sediments. *Environ. Microbiol.* **23**, 4646–4660 (2021).
40. Sipes, K. *et al.* Eight Metagenome-Assembled Genomes Provide Evidence for Microbial Adaptation in 20,000- to 1,000,000-Year-Old Siberian Permafrost. *Appl. Environ. Microbiol.* **87**, e0097221 (2021).
41. Yasuda, S. *et al.* Identification of a Metagenome-Assembled Genome of an Uncultured Methyloceanibacter sp. Strain Acquired from an Activated Sludge System Used for Landfill Leachate Treatment. *Microbiol Resour Announc* **9**, (2020).
42. Qi, Q. *et al.* Microbially enhanced methane uptake under warming enlarges ecosystem carbon sink in a Tibetan alpine grassland. *Glob. Chang. Biol.* **28**, 6906–6920 (2022).
43. Rusley, C. *et al.* Metagenome-Assembled Genome of USCα AHI, a Potential High-Affinity Methanotroph from Axel Heiberg Island, Canadian High Arctic. *Microbiol Resour Announc* **8**, (2019).
44. Nguyen, N.-L. *et al.* Genomic Insights Into the Acid Adaptation of Novel Methanotrophs Enriched From Acidic Forest Soils. *Front. Microbiol.* **9**, 1982 (2018).
45. Deutzmann, J. S., Hoppert, M. & Schink, B. Characterization and phylogeny of a novel methanotroph, Methyloglobulus morosus gen. nov., spec. nov. *Syst. Appl. Microbiol.* **37**, 165–169 (2014).
46. Nuccio, E. E. *et al.* HT-SIP: a semi-automated stable isotope probing pipeline identifies cross-kingdom interactions in the hyphosphere of arbuscular mycorrhizal fungi. *Microbiome*

10, 199 (2022).

47. Schink, B., Thompson, T. E. & Zeikus, J. G. Characterization of Propionispira arboris gen. nov. sp. nov., a Nitrogen-fixing Anaerobe Common to Wetwoods of Living Trees. *Microbiology* **128**, 2771–2779 (1982).
48. Horn, M. A., Matthies, C., Küsel, K., Schramm, A. & Drake, H. L. Hydrogenotrophic methanogenesis by moderately acid-tolerant methanogens of a methane-emitting acidic peat. *Appl. Environ. Microbiol.* **69**, 74–83 (2003).
49. Langwig, M. V. *et al.* Large-scale protein level comparison of Deltaproteobacteria reveals cohesive metabolic groups. *ISME J.* **16**, 307–320 (2022).
50. Begmatov, S., Beletsky, A. V., Dedysh, S. N., Mardanov, A. V. & Ravin, N. V. Genome analysis of the candidate phylum MBNT15 bacterium from a boreal peatland predicted its respiratory versatility and dissimilatory iron metabolism. *Front. Microbiol.* **13**, 951761 (2022).
51. Seidel, L., Broman, E., Turner, S., Ståhle, M. & Dopson, M. Interplay between eutrophication and climate warming on bacterial communities in coastal sediments differs depending on water depth and oxygen history. *Sci. Rep.* **11**, 23384 (2021).
52. Pessi, I. S., Rutanen, A. & Hultman, J. Candidatus Nitrosopolaris, a genus of putative ammonia-oxidizing archaea with a polar/alpine distribution. *FEMS Microbes* **3**, (2022).
53. Zeng, Y. *et al.* Potential Rhodopsin- and Bacteriochlorophyll-Based Dual Phototrophy in a High Arctic Glacier. *MBio* **11**, (2020).
54. Overholt, W. A. *et al.* Carbon fixation rates in groundwater similar to those in oligotrophic marine systems. *Nat. Geosci.* **15**, 561–567 (2022).
55. Liang, R. *et al.* Genomic reconstruction of fossil and living microorganisms in ancient Siberian permafrost. *Microbiome* **9**, 110 (2021).
56. Sun, J. *et al.* Recoding of stop codons expands the metabolic potential of two novel Asgardarchaeota lineages. *ISME Commun* **1**, 30 (2021).
57. Rodrigues-Oliveira, T., Belmok, A., Vasconcellos, D., Schuster, B. & Kyaw, C. M. Archaeal S-Layers: Overview and Current State of the Art. *Front. Microbiol.* **8**, 2597 (2017).
58. Kish, A. *et al.* Preservation of Archaeal Surface Layer Structure During Mineralization. *Sci. Rep.* **6**, 26152 (2016).
59. Stevens, K. M. & Warnecke, T. Histone variants in archaea - An undiscovered country. *Semin. Cell Dev. Biol.* **135**, 50–58 (2023).
60. Veseli, I. *et al.* Microbes with higher metabolic independence are enriched in human gut microbiomes under stress. *bioRxiv* (2023) doi:10.1101/2023.05.10.540289.

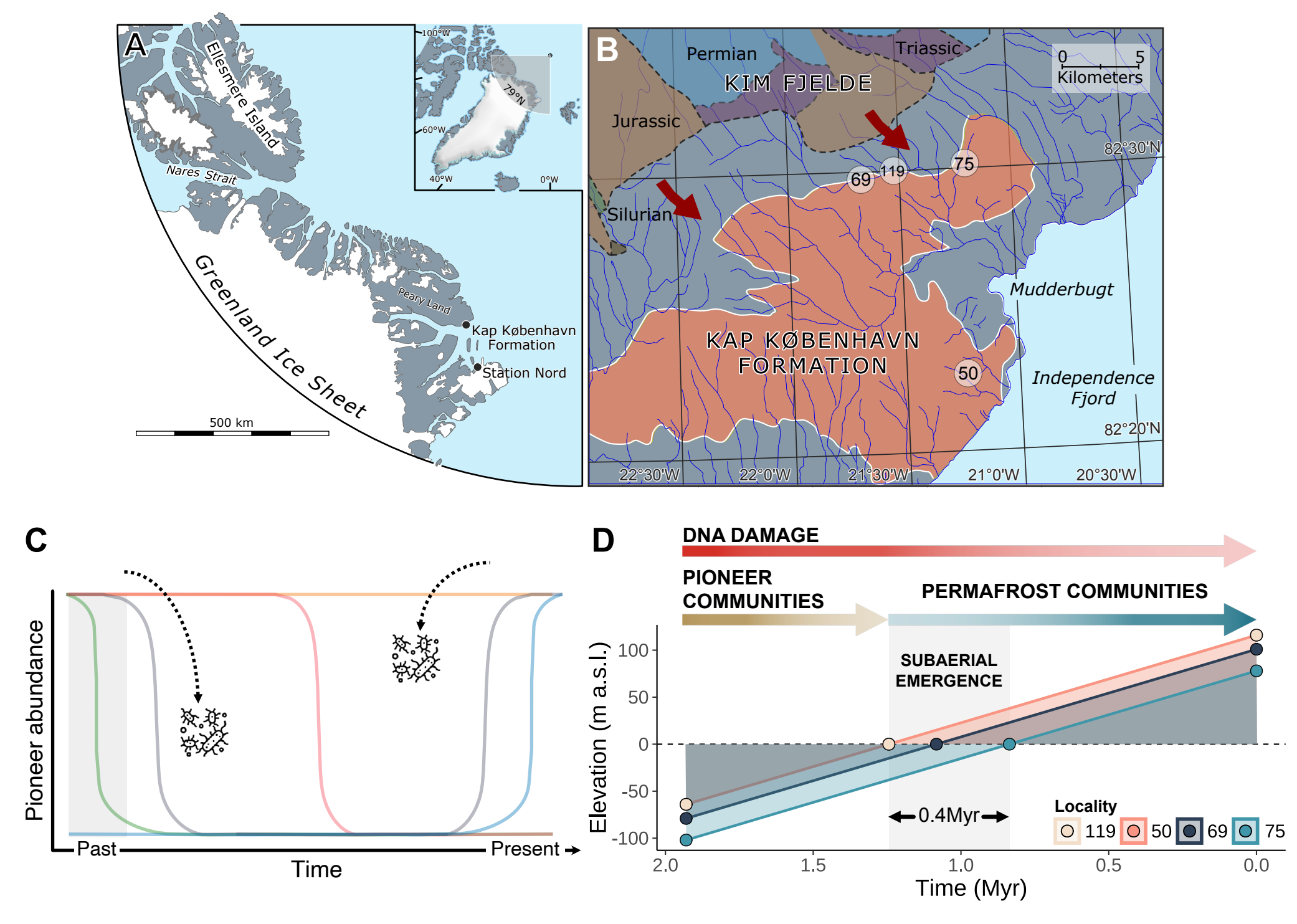
61. Tveit, A., Schwacke, R., Svenning, M. M. & Urich, T. Organic carbon transformations in high-Arctic peat soils: key functions and microorganisms. *ISME J.* **7**, 299–311 (2013).
62. Fernandez-Guerra, A. A metagenomic toolkit to reconstruct million-year-old microbial communities. *Manuscript in preparation* (2023).
63. Oh, Y. *et al.* Reduced net methane emissions due to microbial methane oxidation in a warmer Arctic. *Nat. Clim. Chang.* **10**, 317–321 (2020).
64. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
65. Medvedeva, S., Borrel, G., Krupovic, M. & Gribaldo, S. A global virome of methanogenic archaea highlights novel diversity and adaptations to the gut environment. (2023) doi:10.21203/rs.3.rs-2539466/v1.
66. Krupovic, M. *et al.* Cressdnaviricota: a Virus Phylum Unifying Seven Families of Rep-Encoding Viruses with Single-Stranded, Circular DNA Genomes. *J. Virol.* **94**, (2020).
67. Casson, N., Michel, R., Müller, K.-D., Aubert, J. D. & Greub, G. Protochlamydia naegleriophila as etiologic agent of pneumonia. *Emerg. Infect. Dis.* **14**, 168–172 (2008).
68. Stahl, L. M. & Olson, J. B. Environmental abiotic and biotic factors affecting the distribution and abundance of Naegleria fowleri. *FEMS Microbiol. Ecol.* **97**, (2020).
69. Rapp, J. Z., Sullivan, M. B. & Deming, J. W. Divergent Genomic Adaptations in the Microbiomes of Arctic Subzero Sea-Ice and Cryopeg Brines. *Front. Microbiol.* **12**, 701186 (2021).
70. Camargo, A. P. *et al.* IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.* **51**, D733–D743 (2023).
71. Lipson, D. A. *et al.* Metagenomic insights into anaerobic metabolism along an Arctic peat soil profile. *PLoS One* **8**, e64659 (2013).
72. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 14616–14621 (2007).
73. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* **6**, 3–6 (2021).
74. Schaefer, J. M. *et al.* Greenland was nearly ice-free for extended periods during the Pleistocene. *Nature* **540**, 252–255 (2016).
75. Yan, Y. *et al.* Two-million-year-old snapshots of atmospheric gases from Antarctic ice. *Nature* **574**, 663–666 (2019).
76. Knoblauch, C., Beer, C., Liebner, S., Grigoriev, M. N. & Pfeiffer, E.-M. Methane production as key to the greenhouse gas budget of thawing permafrost. *Nat. Clim. Chang.* **8**, 309–312

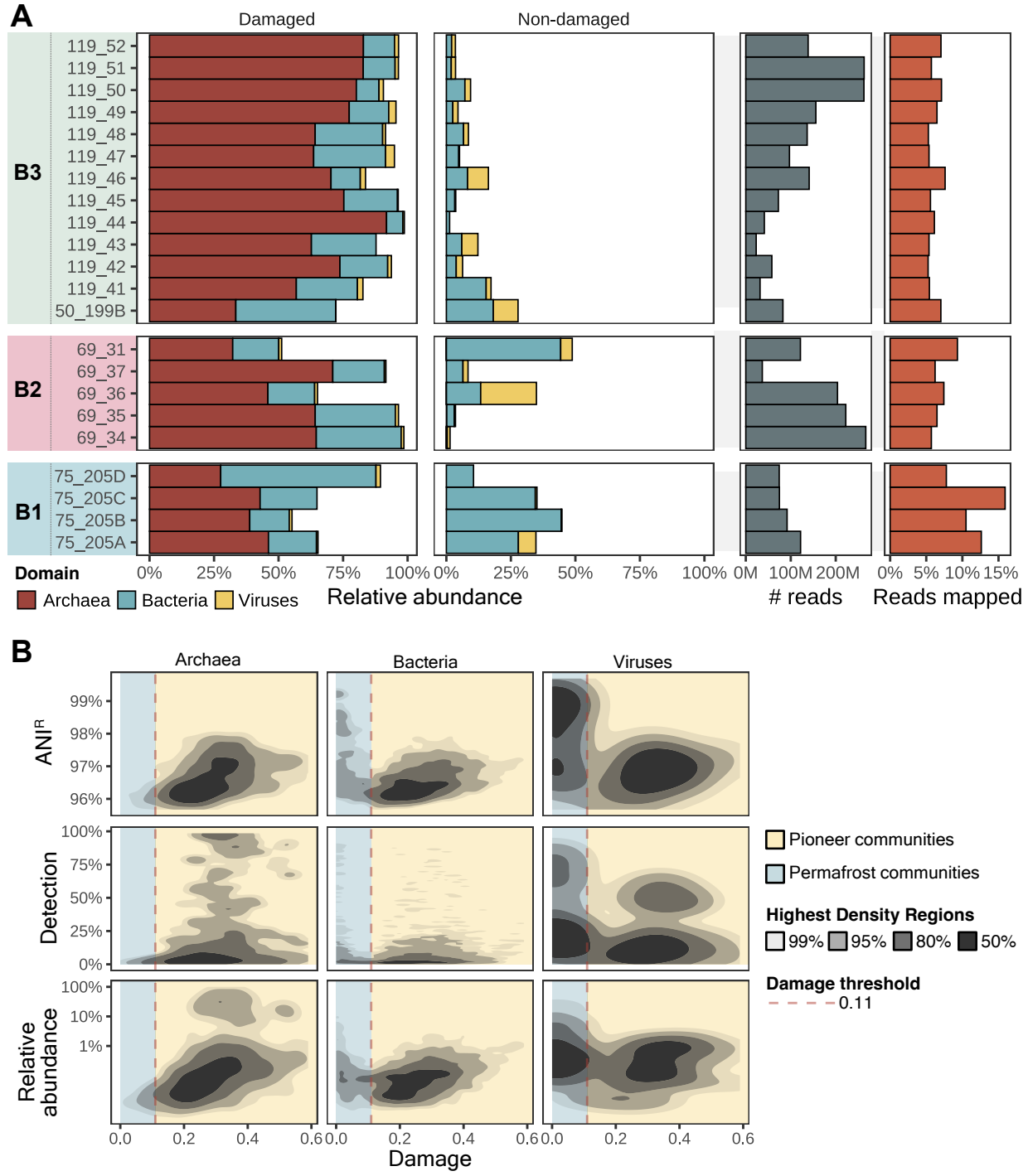
(2018).

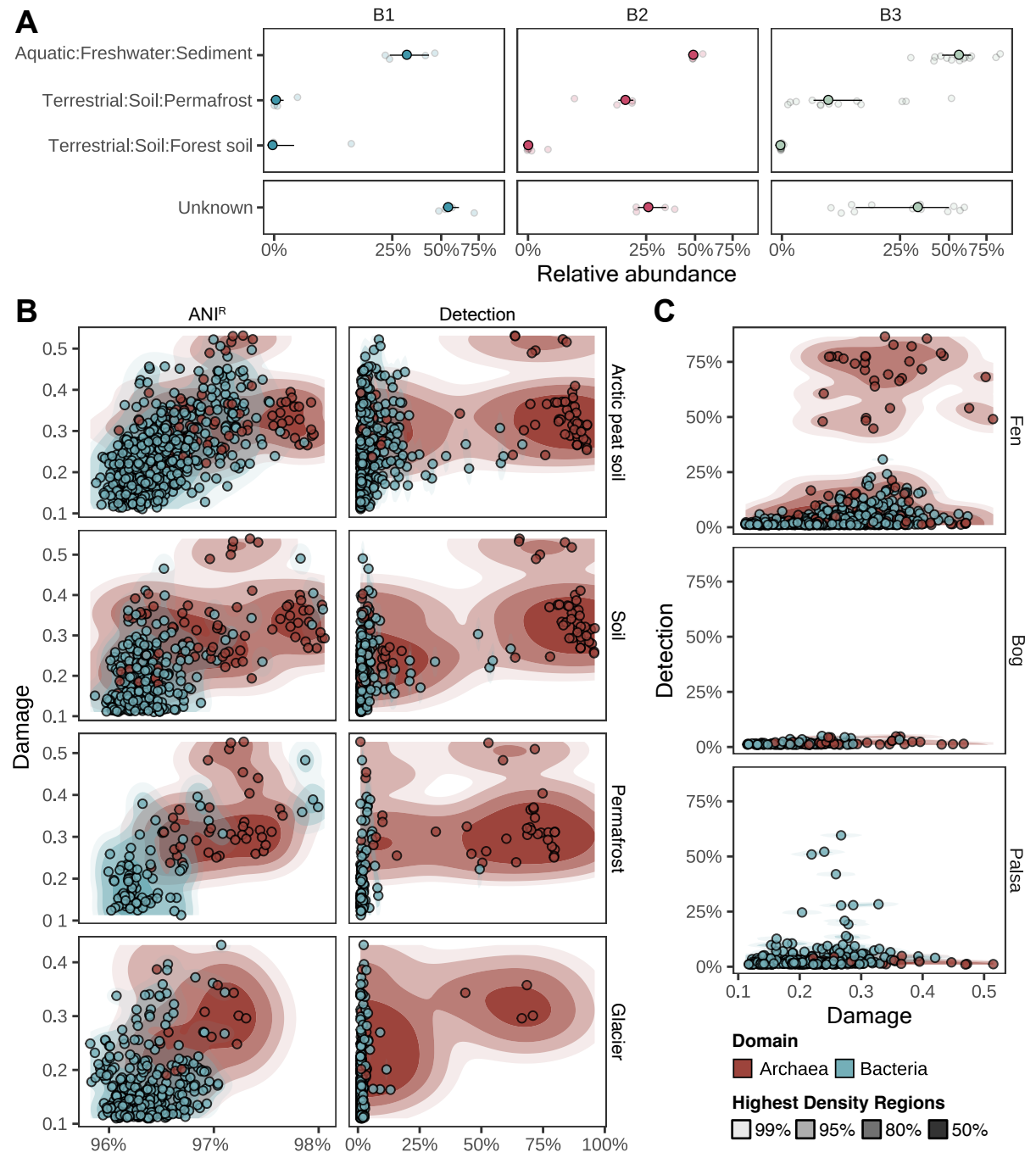
77. Ernakovich, J. G. *et al.* Microbiome assembly in thawing permafrost and its feedbacks to climate. *Glob. Chang. Biol.* **28**, 5007–5026 (2022).
78. Ayala-del-Río, H. L. *et al.* The genome sequence of Psychrobacter arcticus 273-4, a psychroactive Siberian permafrost bacterium, reveals mechanisms for adaptation to low-temperature growth. *Appl. Environ. Microbiol.* **76**, 2304–2312 (2010).
79. Brennan, G. L. & Logares, R. Tracking contemporary microbial evolution in a changing ocean. *Trends Microbiol.* **31**, 336–345 (2023).
80. Pedrós-Alió, C. Time travel in microorganisms. *Syst. Appl. Microbiol.* **44**, 126227 (2021).
81. Nemergut, D. R. *et al.* Patterns and processes of microbial community assembly. *Microbiol. Mol. Biol. Rev.* **77**, 342–356 (2013).
82. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).
83. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
84. Delmont, T. O. Discovery of nondiazotrophic Trichodesmium species abundant and widespread in the open ocean. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
85. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
86. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199-205 (2014).
87. Köster, J. Reproducible data analysis with Snakemake. *F1000Res.* **7**, (2018).
88. Bushnell, B., Rood, J. & Singer, E. BBTools software package. Preprint at (2014).
89. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
90. Toolkit, P. Picard toolkit. *Broad Institute, Github Repository* (2019).
91. Portik, D. M., Brown, C. T. & Pierce-Ward, N. T. Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics* **23**, 541 (2022).
92. Sun, Z. *et al.* Challenges in benchmarking metagenomic profilers. *Nat. Methods* **18**, 618–626 (2021).
93. Hofmann, H., Wickham, H. & Kafadar, K. Letter-Value Plots: Boxplots for Large Data. *J. Comput. Graph. Stat.* **26**, 469–477 (2017).
94. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*

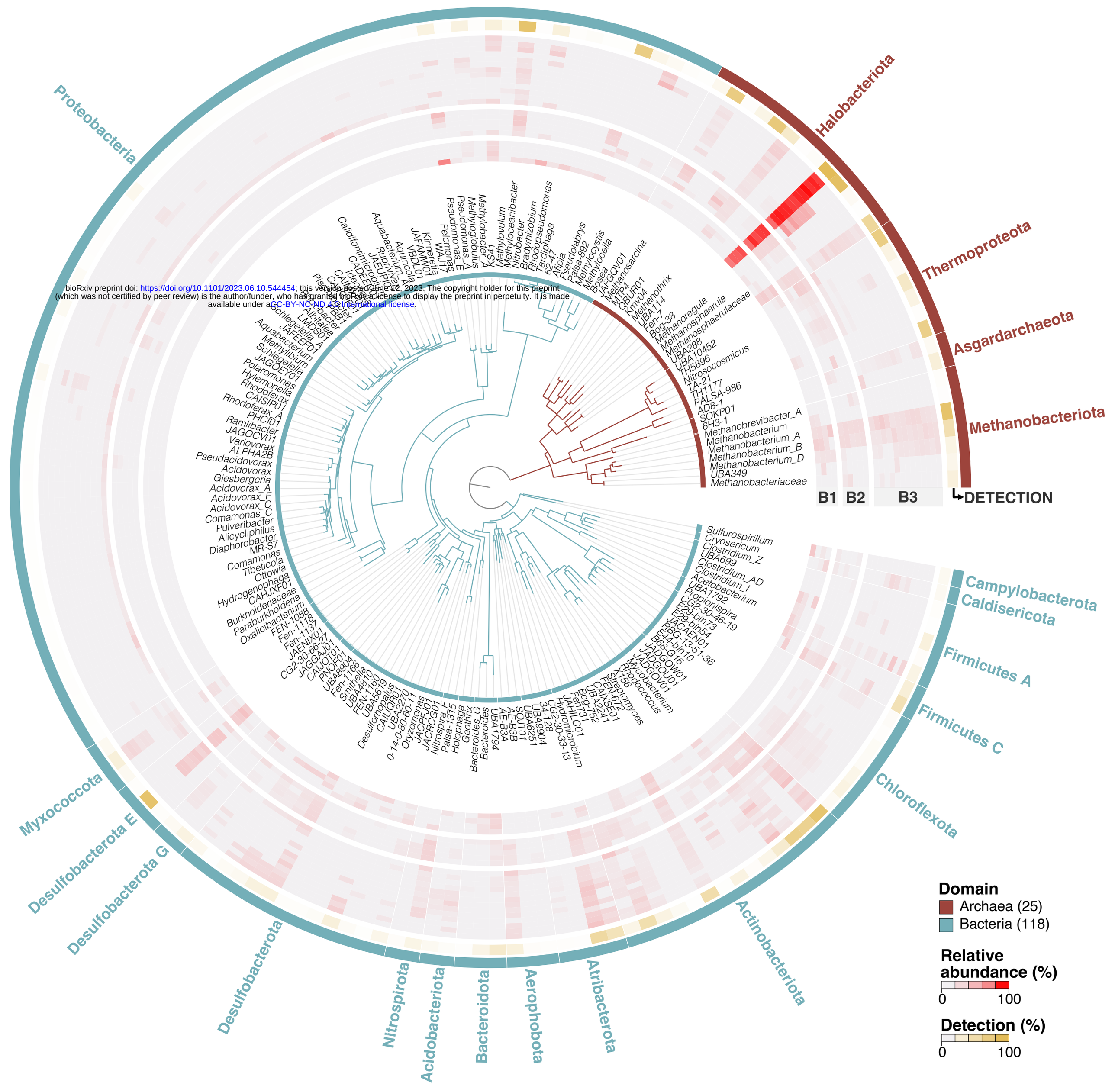
48, D570–D578 (2020).

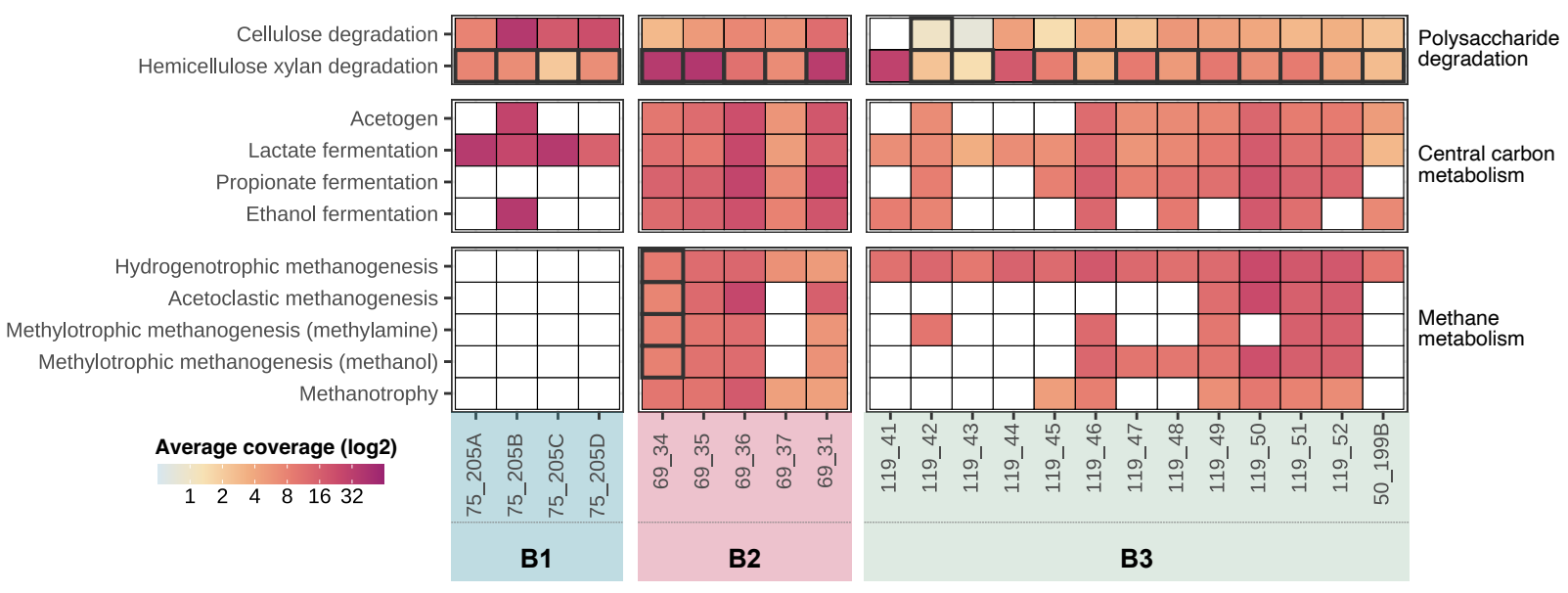
95. Harrison, P. W. *et al.* The European Nucleotide Archive in 2020. *Nucleic Acids Res.* **49**, D82–D85 (2021).
96. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
97. Lemane, T., Medvedev, P., Chikhi, R. & Peterlongo, P. kmtricks: efficient and flexible construction of Bloom filters for large sequencing data collections. *Bioinform Adv* **2**, vbac029 (2022).
98. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
99. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).
100. Golob, J. L. & Minot, S. S. In silico benchmarking of metagenomic tools for coding sequence detection reveals the limits of sensitivity and precision. *BMC Bioinformatics* **21**, 459 (2020).
101. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
102. Terzian, P. *et al.* PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genom Bioinform* **3**, lqab067 (2021).
103. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *Elife* **11**, (2022).

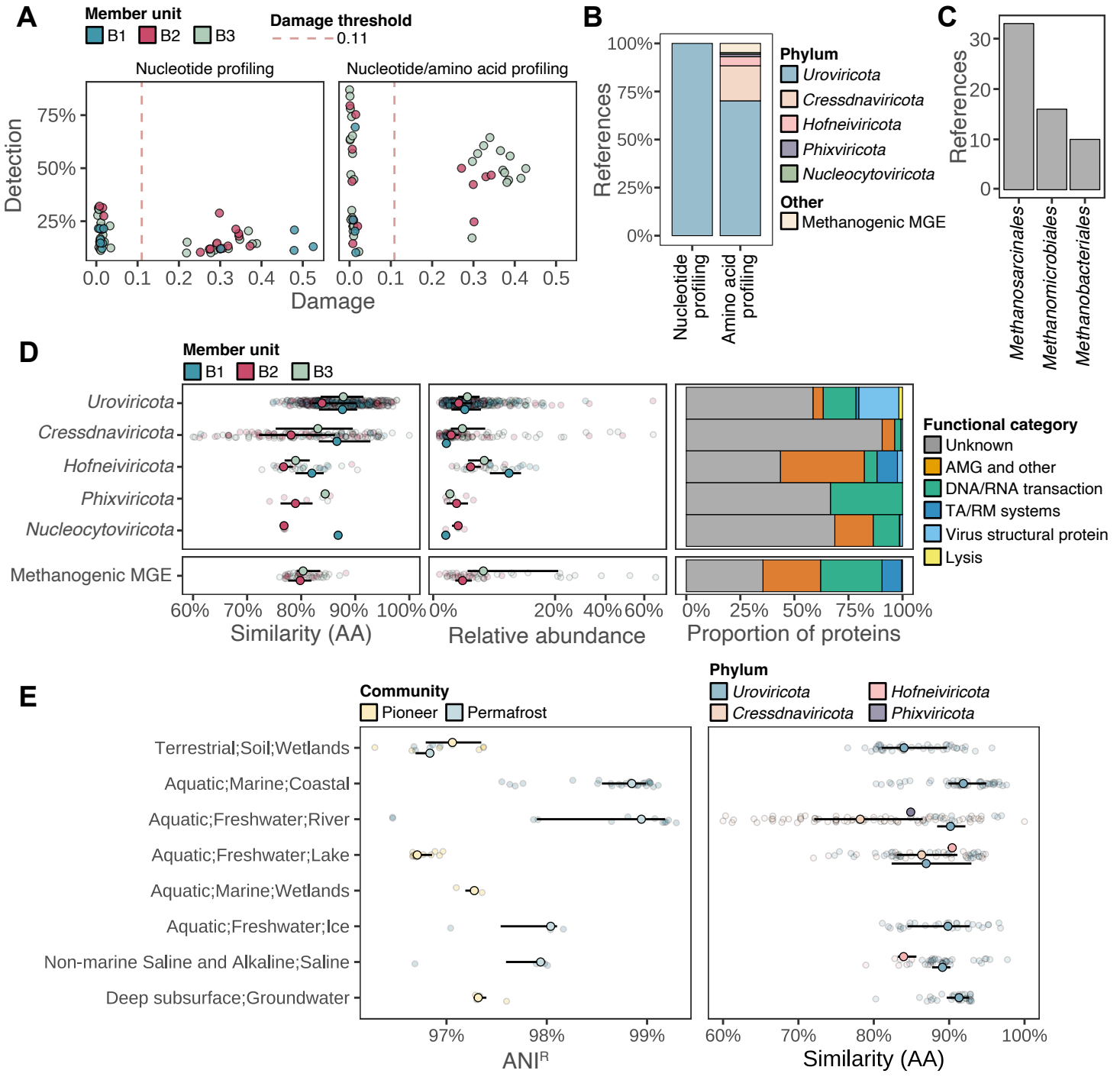

A
Ellesmere Island
Nares Strait
Greenland Ice Sheet
Peary Land
Kap København Formation
Station Nord
500 km
100°W
60°W
40°W
0°W
79°N
B
Permian
Triassic
KIM FJELDE
Jurassic
Silurian
0
5
Kilometers
82°30'N
69
119
75
KAP KØBENHAVN FORMATION
50
Mudderbugt
Independence Fjord
82°20'N
22°30'W
22°0'W
21°30'W
21°0'W
20°30'W
C
Pioneer abundance
Past
Present
Time
D
DNA DAMAGE
PIONEER COMMUNITIES
PERMAFROST COMMUNITIES
SUBAERIAL EMERGENCE
Elevation (m a.s.l.)
100
50
0
-50
-100
0.4Myr
Locality
119
50
69
75
2.0
1.5
1.0
0.5
0.0
Time (Myr)

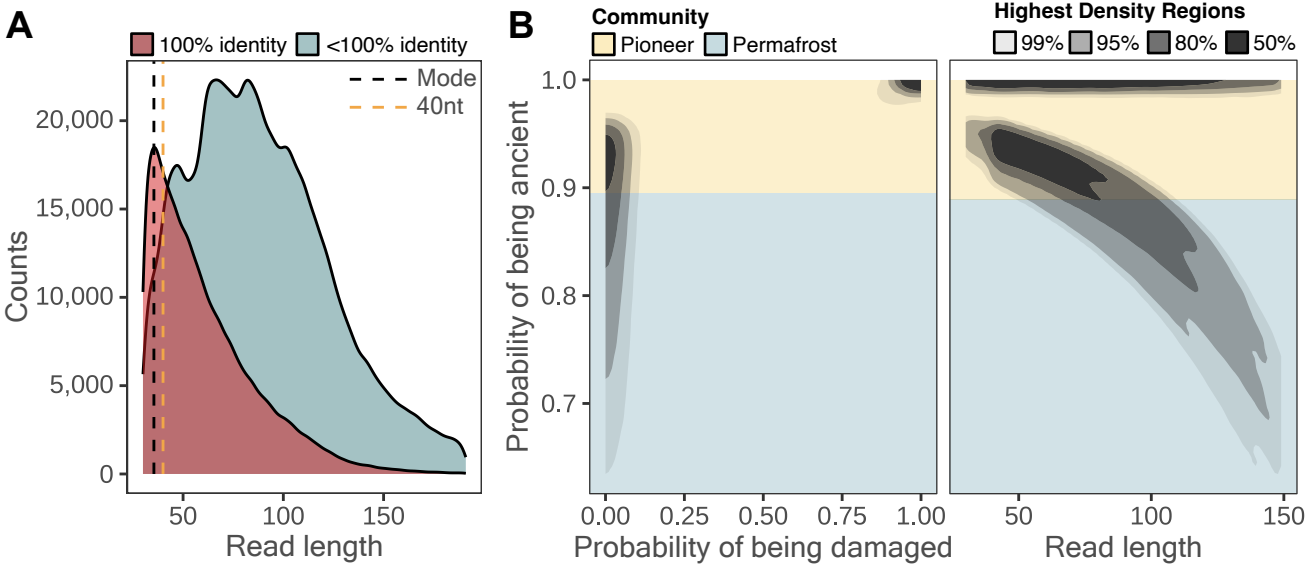

A
Damaged
Non-damaged
B3
119_52
119_51
119_50
119_49
119_48
119_47
119_46
119_45
119_44
119_43
119_42
119_41
50_199B
B2
69_31
69_37
69_36
69_35
69_34
B1
75_205D
75_205C
75_205B
75_205A
Domain
Archaea
Bacteria
Viruses
0%
25%
50%
75%
100%
Relative abundance
0M
100M
200M
reads
0%
5%
10%
15%
Reads mapped
B
Archaea
Bacteria
Viruses
ANIR
99%
98%
97%
96%
Detection
100%
75%
50%
25%
0%
Relative abundance
100%
10%
1%
0.0
0.2
0.4
0.6
Damage
Pioneer communities
Permafrost communities
Highest Density Regions
99%
95%
80%
50%
Damage threshold
0.11

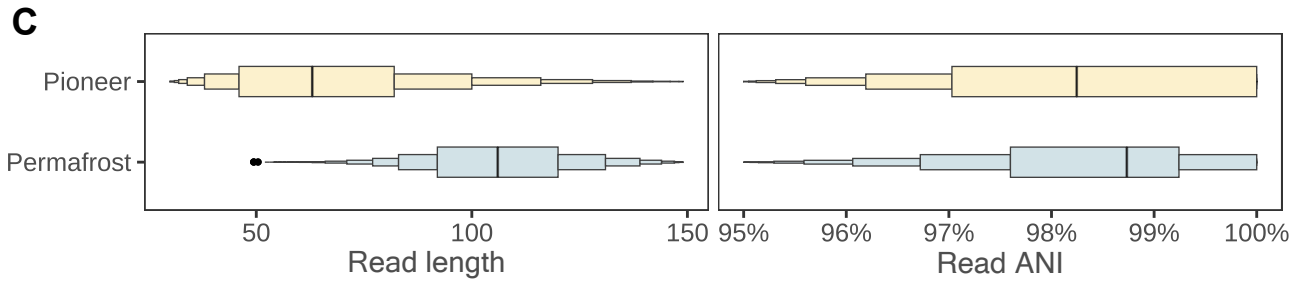

A
B1
B2
B3
Aquatic:Freshwater:Sediment
Terrestrial:Soil:Permafrost
Terrestrial:Soil:Forest soil
Unknown
0%
25%
50%
75%
Relative abundance
B
ANIR
Detection
Arctic peat soil
Soil
Permafrost
Glacier
Damage
96%
97%
98%
100%
C
Fen
Bog
Palsa
Detection
Damage
Domain
Archaea
Bacteria
Highest Density Regions
99%
95%
80%
50%




Cellulose degradation
Hemicellulose xylan degradation
Polysaccharide degradation
Acetogen
Lactate fermentation
Propionate fermentation
Ethanol fermentation
Central carbon metabolism
Hydrogenotrophic methanogenesis
Acetoclastic methanogenesis
Methylotrophic methanogenesis (methylamine)
Methylotrophic methanogenesis (methanol)
Methanotrophy
Methane metabolism
Average coverage (log2)
1 2 4 8 16 32
75_205A
75_205B
75_205C
75_205D
69_34
69_35
69_36
69_37
69_31
119_41
119_42
119_43
119_44
119_45
119_46
119_47
119_48
119_49
119_50
119_51
119_52
50_199B
B1
B2
B3


A
Member unit
B1
B2
B3
Damage threshold
0.11
Nucleotide profiling
Nucleotide/amino acid profiling
Detection
75%
50%
25%
0.0
0.1
0.2
0.3
0.4
0.5
Damage
B
References
100%
75%
50%
25%
0%
Nucleotide profiling
Amino acid profiling
Phylum
Uroviricota
Cressdnaviricota
Hofneiviricota
Phixviricota
Nucleocytoviricota
Other
Methanogenic MGE
C
References
30
20
10
0
Methanosarcinales
Methanomicrobiales
Methanobacteriales
D
Member unit
B1
B2
B3
Uroviricota
Cressdnaviricota
Hofneiviricota
Phixviricota
Nucleocytoviricota
Methanogenic MGE
60%
70%
80%
90%
100%
Similarity (AA)
0%
20%
40%
60%
Relative abundance
0%
25%
50%
75%
100%
Proportion of proteins
Functional category
Unknown
AMG and other
DNA/RNA transaction
TA/RM systems
Virus structural protein
Lysis
E
Community
Pioneer
Permafrost
Phylum
Uroviricota
Cressdnaviricota
Hofneiviricota
Phixviricota
Terrestrial;Soil;Wetlands
Aquatic;Marine;Coastal
Aquatic;Freshwater;River
Aquatic;Freshwater;Lake
Aquatic;Marine;Wetlands
Aquatic;Freshwater;Ice
Non-marine Saline and Alkaline;Saline
Deep subsurface;Groundwater
97%
98%
99%
ANIR
60%
70%
80%
90%
100%
Similarity (AA)





D

	Pioneer	Permafrost
Number of reads	809,004	252,497
Read length mode	42	103
ANI^R	98.2%	98.4%
Coverage	33X	16X
Detection	93.7%	92.1%