Two-million-year-old microbial communities from the Kap København Formation in North Greenland

Antonio Fernandez-Guerra[1,2,3,*], Lars Wörmer[4], Guillaume Borrel[5], Tom O Delmont[6], Bo Elberling[7], Marcus Elvert[4], A. Murat Eren[8,9,10,11], Simonetta Gribaldo[5], Rasmus Amund Henriksen[2], Kai-Uwe Hinrichs[4], Annika Jochheim[12], Thorfinn S. Korneliussen[1,2,3], Mart Krupovic[13], Nicolaj K. Larsen[1,2,3], Rafael Perez-Laso[14], Mikkel Winther Pedersen[1,2,3], Vivi K. Pedersen[15], Anthony H. Ruter[1,2,3], Karina K. Sand[1,2,3], Martin Sikora[1,2,3], Martin Steinegger[16], Iva Veseli[8,10], Yucheng Wang[1,2,3,17,18], Lei Zhao[2], Marina Žure[1,2,3], Kurt H. Kjær[1,2,3], Eske Willerslev[1,2,3,4,18*]

[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]Ancient Environmental Genomics Initiative for Sustainability, Globe Institute, University of Copenhagen, Copenhagen, Denmark;
[4]MARUM – Center for Marine Environmental Sciences and Faculty of Geosciences, University of Bremen, Bremen, Germany;
[5]Institut Pasteur, Université Paris Cité, CNRS UMR6047, Evolutionary Biology of the Microbial Cell Unit, Paris, France;
[6]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Université Paris-Saclay, Evry, France;
[7]Center for Permafrost (CENPERM), Department of Geosciences and Natural Resource Management, University of Copenhagen, DK1350, Copenhagen, Denmark;
[8]Helmholtz Institute for Functional Marine Biodiversity, Oldenburg, Germany;
[9]Max Planck Institute for Marine Microbiology, Bremen, Germany;
[10]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany;

[11]Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, Oldenburg, Germany;
[12]Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany;
[13]Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, Paris 75015, France
[14]Biogeochemistry and Microbial Ecology Department, Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain;
[15]Department of Geoscience, Aarhus University, Denmark;
[16]School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea;
[17]Group of Alpine Paleoecology and Human Adaptation (ALPHA), State Key Laboratory of Tibetan Plateau Earth System, Resources and Environment (TPESRE), Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, 100101, China;
[18]Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK;

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

Summary

Environmental DNA (eDNA) from the 2-million-year-old Kap København Formation of northern Greenland has revealed an ecosystem of plants and animals with no contemporary analogue[1]. Here, we reconstruct the microbial (bacterial, archaeal, and viral) communities that thrived at the site during this time. By leveraging a novel analytical framework that integrates taxonomic profiling, DNA damage estimates, and functional reconstructions, we identify and distinguish pioneer microbial communities from later permafrost microbial assemblages. We show that at the time of sediment deposition, the terrestrial input at the Kap København site originated from a palustrine wetland, suggesting warmer, non-permafrost conditions. During this period, the detection of methanogenic archaea and signals of their carbon metabolism is consistent with Kap København and similar northern ecosystems contributing moderate methane emissions. Intriguingly, we discover a remarkable nucleotide sequence similarity—exceeding 98%—between pioneer methanogens and present-day analogues in thawing permafrost. This aligns with the concept of "time-traveling" microbes[2] surviving across geological time and waiting for conditions to turn favourable rather than evolving to adapt to changing conditions. Importantly, in contrast to the plant and animal communities of the Kap København, a striking similarity in microbial composition to that of a contemporary thawing Arctic suggests that microbial communities may serve as the first indication of broader climate-driven ecosystem disruptions.

Main

Understanding how global warming will impact Earth's biodiversity remains a subject of great debate. Some argue for tipping points where nature responds in drastic ways, reshuffling species distributions[3], while others suggest a more gradual mode of change[4]. The predictability of biodiversity across time and space, i.e., whether species distributions, their compositions and ecosystem functions may repeat themselves across similar geological climate events, also remains contentious[5]. Addressing these issues is crucial for predicting future ecosystem transformations in response to climate change.

Of Earth's many geological Epochs, the Pliocene to early Pleistocene (5 to 2 million years ago) is argued to be among the best analogues of a future world under global warming[6,7]. However, because of its time depth, the current understanding of biodiversity from this period is limited primarily to pollen[8] and macrofossils[9]. Recent advantages in eDNA have pushed back ancient DNA retrieval from about 1 to 2 million years at the Kap København Formation of northern Greenland, a time when the average temperature was 10°C warmer than today[10]. The diverse plant and animal eDNA uncovered shows an environment with no modern equivalent, composed of taxa whose contemporary relatives are found in both Arctic and temperate environments[1]., implying a high degree of unpredictability in future species distributions under climate change.

Importantly, the eDNA study on the Kap København Formation did not consider the microbial communities. Microbes play a fundamental role in maintaining biogeochemical processes that influence microbial and non-microbial life[11–13] composition and dynamics. Thus, characterising the ancient microbial eDNA component can provide a comprehensive view of the complex functioning of past ecosystems. Nevertheless, most ancient eDNA studies neglect the microbial communities because sustained microbial activity in sediments often disturbs and obscures the original microbial signatures[13]. Over time, the genetic signals of pioneering microbes—those present at the time of sediment deposition—become mixed with DNA from microbes that continue to thrive in the sediment, making it challenging to separate ancient microbial DNA from later microbial activity when reconstructing past ecological conditions and their impact on broader biogeochemical processes. As a result, ancient sediments contain two putative ancient DNA fractions[14]. One originates from organisms that died long ago and is highly fragmented and

contains post-mortem modifications such as deamination damage at its termini [15]. The other is DNA from organisms that have either remained dormant or survived with reduced metabolic activity for decades to hundreds of years (seed bank microbes) or for thousands to millions of years ("time-travelling" microbes). While seed bank microbes are considered part of modern ecosystems, acting as a genetic reservoir, "time-travelling" microbes can reintroduce ancient genetic traits into present-day ecosystems[16].

For this reason, DNA from seed bank and "time-travelling" microbes is challenging to distinguish from modern contaminant DNA. To be able to analyse ancient microbial diversity from eDNA records, we developed a new analytical framework that integrates taxonomic profiling, DNA damage estimates, and functional reconstructions and applied it to the Kap København Formation.

The Kap København Formation

The Kap København Formation consists of a c. 100 m thick succession of shallow marine sediments primarily accumulated during a ~20,000 years long interglacial period about two million years ago[1,17]. These sediments serve as an archive for 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 sediments. After the Kap København Formation was deposited, it was uplifted due to the solid Earth's flexural isostatic response to erosional unloading by fjord incision[18] (Fig. 1B). We estimate an uplift rate of ~0.09 mm/yr, which raised these shallow marine deposits to their current elevation (Fig. 1C; Supplementary Table S3). 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; from older to younger) were exposed subaerially between 0.8-1.2 million years ago (Fig. 1C). Sediment records from below the present ice sheet suggest that the Greenland Ice Sheet was already established at the time of subaerial exposure 1.2 million years ago[19]. It is reasonable to assume that during the period when 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. 1C).

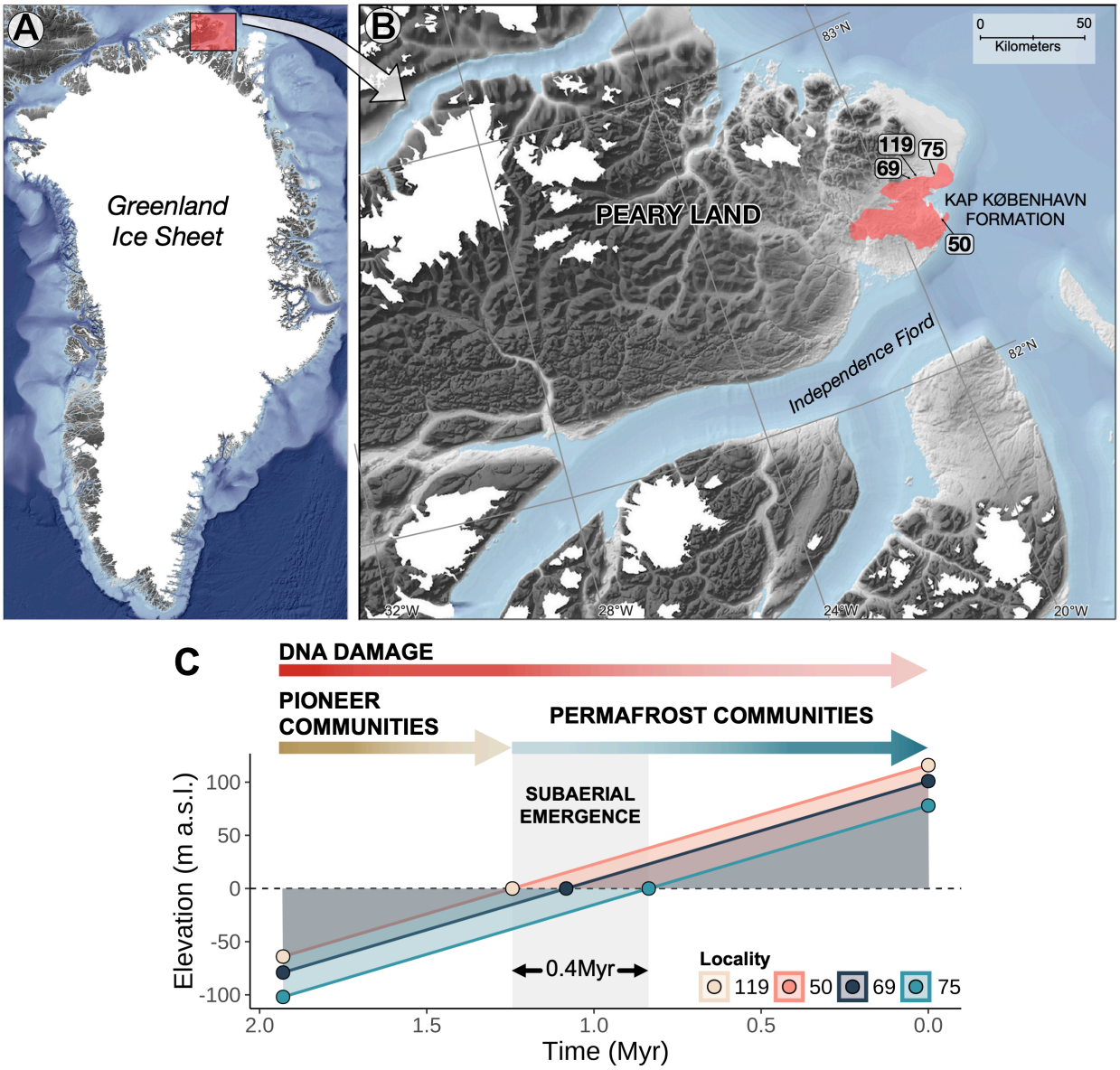


Fig. 1 | Geographical location and uplift model A, Map indicating the location of Kap København Formation in North Greenland at the entrance to the Independence Fjord (82° 24′ N 22° 12′ W). B. Spatial distribution of the ~100-m thick succession of shallow marine near-shore sediments between Mudderbugt and the low mountains towards the north, circles highlight the localities used in the study. C, Simple uplift model of the Kap København Formation. The darker colour in the DNA damage arrow indicates higher damage. In contrast, the arrows for the pioneer and permafrost microbial communities show abundance levels, with darker colours indicating higher abundance. Localities 119 and 50 are superimposed.

Pioneer and permafrost microbial communities

Most endogenous DNA in ancient sediments is highly fragmented and subject to deamination damage at the termini. Therefore, we designed a novel approach for inferring microbial community

92 composition in deep-time samples from ultra-short and damaged DNA sequencing reads. Our
93 approach involved three main components. Firstly, we created a comprehensive genomic database
94 with a common taxonomic framework (Extended Data Fig. 1) covering Archaea, Bacteria, Viruses
95 and eukaryotic organelles. Secondly, in our workflow (Extended Data Fig. 2), we performed a
96 sensitive search using Bowtie2[20] and applied stringent filtering criteria that leveraged read patterns
97 across the references, allowing us to confidently retrieve low-coverage genomes, even with as little
98 as 1% of their genome present (hereafter, we will refer to the "breadth of coverage" as detection,
99 the proportion of a given genome that is covered at least 1X). We inferred their relative abundance
100 by normalising the estimated TAD80[21] (truncated average depth) number of reads to the reference
101 length. Lastly, we performed large-scale damage estimation on the filtered results to identify those
102 references that are potentially members of the pioneer microbial communities. We excluded taxa
103 identified in the controls and blanks (Supplementary Table S2; Extended Data Fig. 3), as well as
104 the eukaryotic hits from subsequent analyses, since the eukaryotic references were intended to be
105 used as bait in the competitive mapping and as a reference to select the damage threshold
106 (Extended Data Fig. 4A). We then focused our analyses on 22 samples across the stratigraphic
107 units (B1, B2, and B3) (Supplementary Table S1) that met two criteria: (1) having more than 10
108 million unique reads and (2) where the combined relative abundance of references exhibiting
109 damage greater than 0.11 (we used the background-subtracted damage frequency at the first
110 position of the reads provided by metaDMG[22], where one represents the maximum level)
111 accounted for at least 50% (Extended Data Fig. 4B). We mapped an average of 7.2% of unique
112 reads per sample with an average of 1,950 references (Fig. 2A; Supplementary Table S4). Finally,
113 we aggregated our results by calculating the geometric mean of taxonomic abundances and
114 averaging the inferred damage at the species level. On average, the damaged fraction composed
115 84% of the estimated relative abundances in each sample, revealing Archaea as the dominant
116 pioneer group (Fig. 2A; Supplementary Table S4).

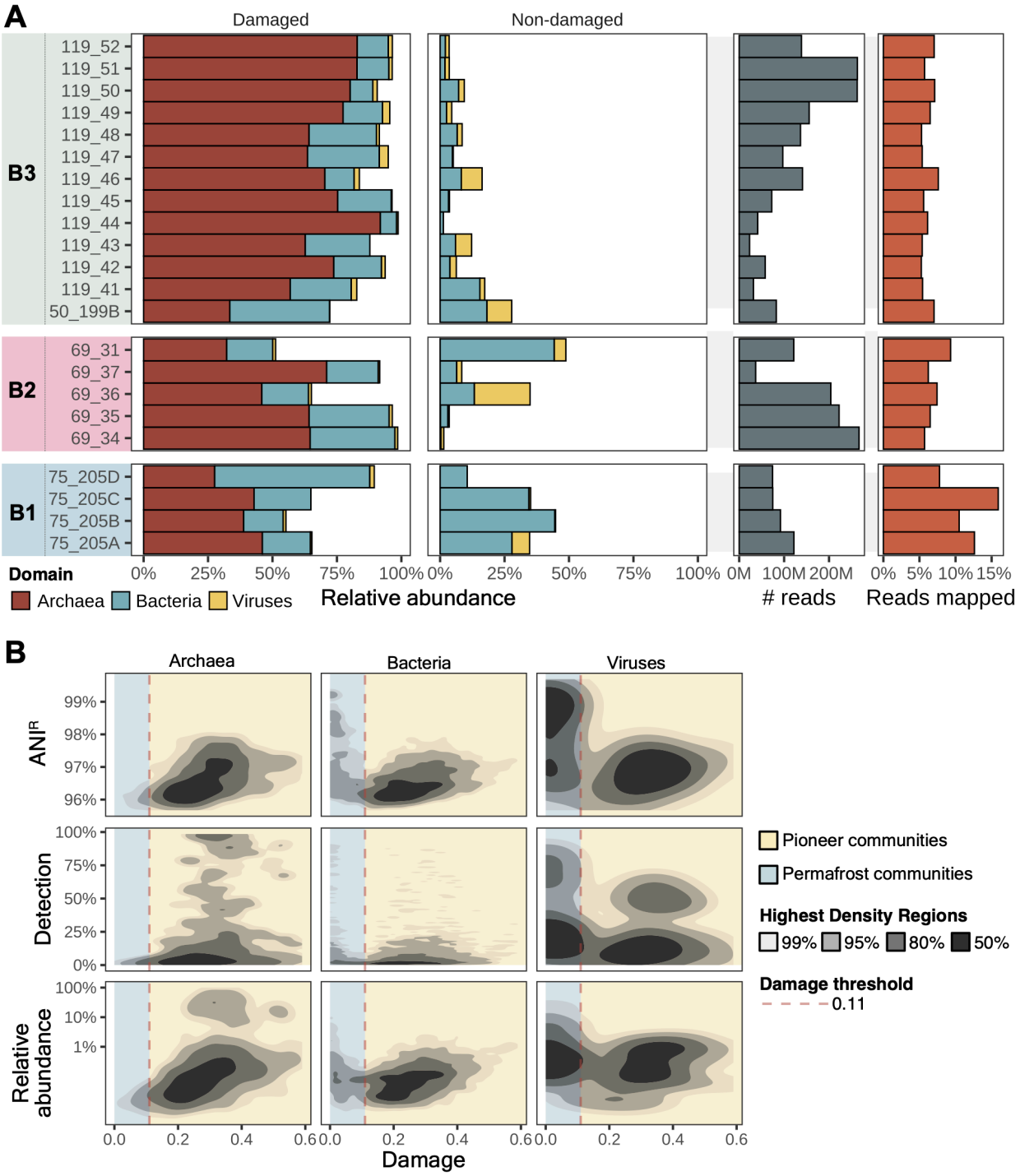


Fig. 2 | Damage and non-damage distribution patterns A, Proportion of taxa identified as damaged and non-damaged in the three stratigraphic units B1 to B3 of the Kap København Formation based on the estimated taxonomic abundances. The two rightmost panels show the number of dereplicated reads used for mapping and the percentage of reads successfully mapped to any references within the database employed for taxonomic profiling. B, The Highest Density Regions plots show the average nucleotide identity, detection, and relative abundance distribution as a function of the damage for Bacteria, Archaea

and Viruses. 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 contemporary 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 differentiate between the initial pioneer communities and those 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), consistent with the low abundance of archaeal taxa typically found in the cryosphere[23,24]. On the contrary, the high values for ANI^R, detection and relative abundance of the bacterial and viral fraction indicate the distinct composition of contemporary permafrost communities. 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 in the sequences. In some cases, the ANI^R values approached 98%, indicating a relatively high level of identity. Remarkably, our samples show high detection values for Archaea and some bacterial taxa, where some of them have more than 75% of the genome recovered (Fig. 2B; Supplementary Table S4).

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[25], trained on over 1,003 modern metagenomes, to identify the contribution of 33 different biomes to our samples (Supplementary Table S8). We further supplemented our analysis by employing a k-mer similarity method[26] 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 Table S9). The second most important 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 in some of our locations, 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 Table S9).

155

156 Next, we explored the distribution of the reads extracted from the pioneer communities in each
157 sample using the Genomes from Earth's Microbiomes (GEM) catalogue[27], which covers diverse
158 habitats of Earth's continents and oceans and the glacier microbiome catalogue[23]. We used the
159 estimated damage, $ANI^{R,}$ and detection to select the biomes that recruited multiple non-unique
160 references with a detection rate higher than 50%. The biomes with the highest number of references
161 (Fig. 3B) were consistent with the results from the microbial source tracking methods, with archaea
162 having a detection rate of over 75% and a high ANI^{R} despite high levels of estimated damage
163 (Supplementary Table S10). The Arctic peat soil biomes and soil had the highest number of
164 references, indicating the successful recovery of the pioneer microbial communities. To increase
165 the resolution of our analysis, we calculated where our samples would fall on a permafrost thaw
166 gradient (Supplementary Table S11). We used the metagenome-assembled genomes recovered
167 from the Stordalen Mire[28] to recruit the reads from the pioneer communities. Our results show that
168 fens, which are peat-forming wetlands, recruit the largest number of archaeal MAGs with detection
169 above 50% (Fig. 3C). We also found microbial genomes with moderate detection in palsa (Fig.
170 3C), which aligns with the permafrost signal detected by both the microbial source tracking
171 methods and the read recruitment to genome collections.
172 It is worth noting that the samples collected from the Kap København Formation aggregate diverse
173 microbial communities sourced from upstream locations mixed with *in-situ* micro-environments
174 within the delta, as observed in extant delta systems[29]. These communities have been deposited
175 over 20,000 years, resulting in a blend of local and upstream contributions reflecting the
176 depositional history of the formation. This composite estimate serves as a representation of the
177 dominant biome during the Early Pleistocene period. Furthermore, the vegetation recovered on the
178 site reinforces our findings that the region experienced prolonged periods with annual mean
179 temperatures well above freezing and consequently without permafrost (Supplementary
180 Information, section 1).

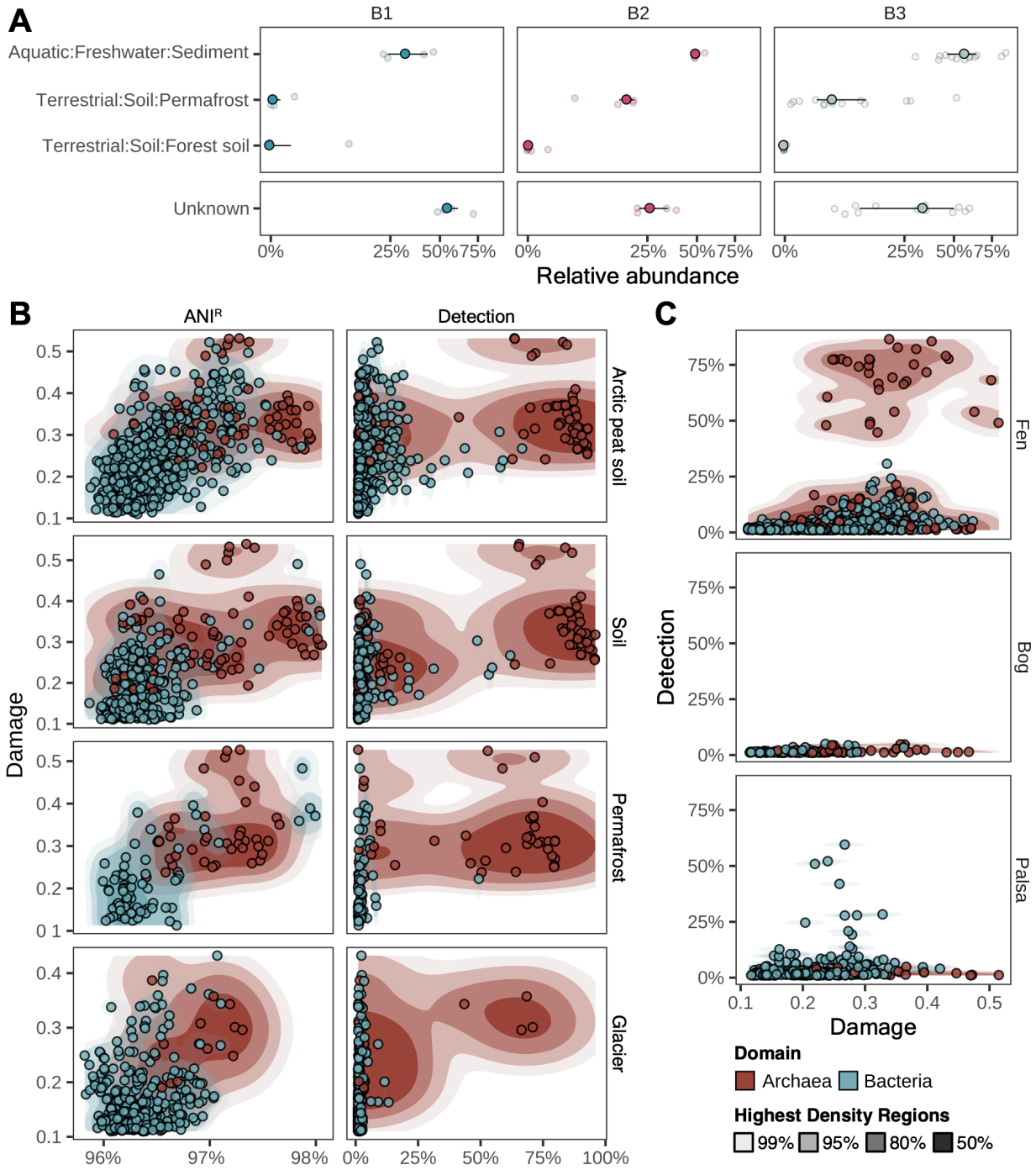
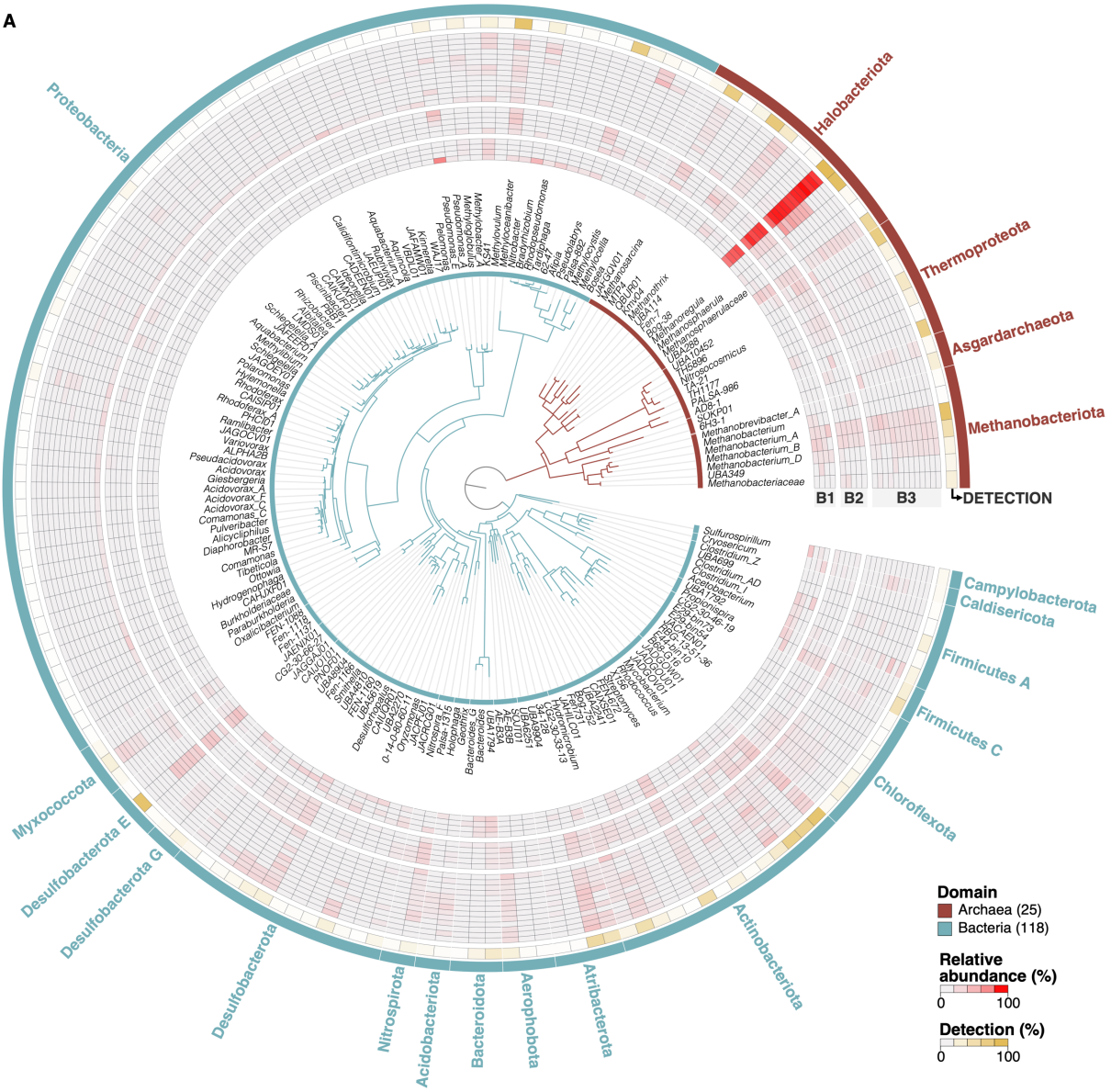


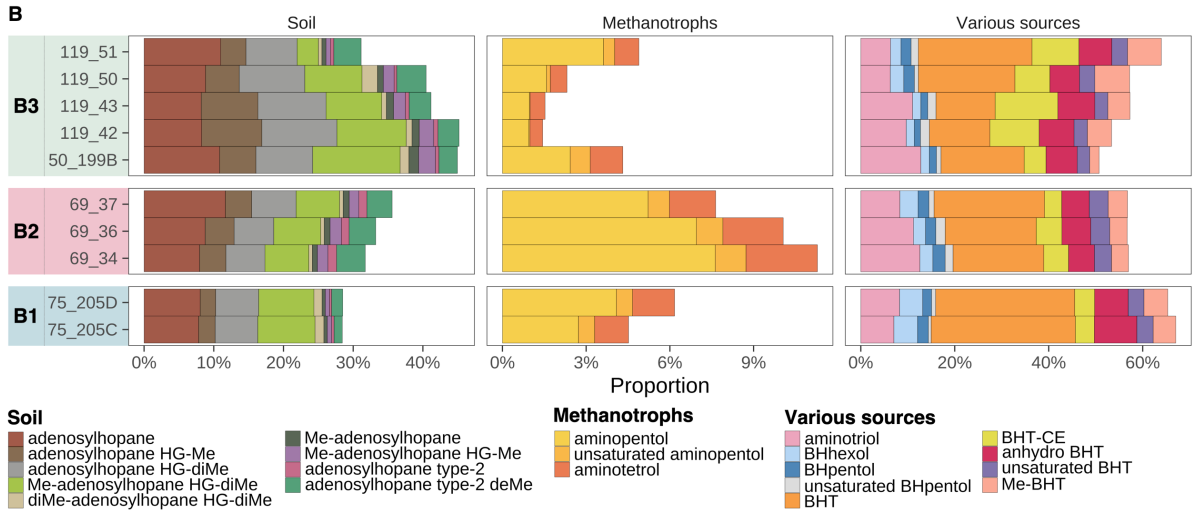
Fig. 3 | Kap København Formation modern analogue biomes A, Boxplot representing the estimated proportions by meta-SourceTracker from each "source" or biome for a single Kap København Formation sample or "sink". Colours correspond to the three stratigraphic units, B1, B2 and B3. B, Highest Density Regions 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[27] and Liu et al. 2022[23]. 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 two-million-year-old microbiome

The structure and composition of microbial communities preserved in sediments from two million years ago revealed remarkable similarities to those in modern palustrine wetlands. Reconstructing the abundance of such ancient microbiomes presents significant challenges due to the taphonomic and biological processes that DNA undergoes over geological timescales[30,31]. Nevertheless, we identified a diverse array of archaeal and bacterial taxa from our dataset (Supplementary Table S4). Our analyses suggest that methanogenic archaea were the dominant group at the Kap København site two million years ago (Fig. 4A). Considering the warmer conditions of the Early Pleistocene, the presence of methane-producing microbial communities upstream of the deposition site two million years ago resembling those found in modern wetlands or thawing permafrost, is conceivable. However, their unexpectedly high relative abundance (averaging over 60% in most samples) is unlikely. While methanogens can show higher abundances than other archaea in a thawing permafrost environment[32], they typically account for only 1-5%[33–37] of the overall community in present-day settings. 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]. To discard potential DNA extraction biases, we validated our results using a new optimised extraction protocol in a subset of our samples (Supplementary Information, section 2; Supplementary Table S5-S6; Extended Data Fig. 5). Multiple factors might contribute to the increased preservation of archaea. Firstly, differences in the local chemical environment, influenced by the metabolic activity of archaea, could contribute to preservation disparities. 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 inter-subunit interactions[39]. Several features of the S-layers could be responsible for the increased preservation; the S-layers can repair themselves against damage and are also found to serve as nucleation sites for mineral formation, resulting in encrustation[40]. Finally, differences in chromatin structure could also influence preservation. DNA is complexed with histones in many archaea, including methanogens, making them among the most abundant proteins in *Methanobacteriales*[41]. This distinct chromatin organisation in archaea compared to bacteria may be yet another factor contributing to the observed differences in preservation patterns.


A
Proteobacteria
Halobacteriota
Thermoproteota
Asgardarchaeota
Methanobacteriota
B1 B2 B3
DETECTION
Campylobacterota
Caldisericota
Firmicutes A
Firmicutes C
Chloroflexota
Actinobacteriota
Atribacterota
Aerophobota
Bacteroidota
Acidobacteriota
Nitrospirota
Desulfobacterota
Desulfobacterota G
Desulfobacterota E
Myxococcota
Domain
Archaea (25)
Bacteria (118)
Relative abundance (%)
0 100
Detection (%)
0 100


B
Soil
Methanotrophs
Various sources
B3
119_51
119_50
119_43
119_42
50_199B
B2
69_37
69_36
69_34
B1
75_205D
75_205C
0% 10% 20% 30% 40%
0% 3% 6% 9%
0% 20% 40% 60%
Proportion
Soil
adenosylhopane
adenosylhopane HG-Me
adenosylhopane HG-diMe
Me-adenosylhopane HG-diMe
diMe-adenosylhopane HG-diMe
Me-adenosylhopane
Me-adenosylhopane HG-Me
adenosylhopane type-2
adenosylhopane type-2 deMe
Methanotrophs
aminopentol
unsaturated aminopentol
aminotetrol
Various sources
aminotriol
BHhexol
BHpentol
unsaturated BHpentol
BHT
BHT-CE
anhydro BHT
unsaturated BHT
Me-BHT

219

Fig. 4 | Community composition of the archaeal and bacterial taxa A) Taxonomic profile of the damaged identified taxa, 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. B) The relative abundance of bacteriohopanepolyol derivatives is expressed in percentages of the total. Adenosylhopanes are named after Hopmans et al.[42].

We identified a variety of methane-related archaea across multiple genera, each showing detection rates 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*, as well as the versatile methanogen *Methanosarcina* (Fig. 4A). Additionally, we observed the archaeal methanotroph g__Kmv04 (*Methanocomedenaceae*). Our analyses also revealed the presence of bacterial phyla such as *Atribacterota* (g__CG2-30-33-13) and *Aerophobota* (g__AE-B3A), which are commonly found in methane-rich environments ranging from temperate soils and deep marine sediments to permafrost[43,44]. Moreover, our study identified a range of methanotrophic bacteria exhibiting high detection rates. Among these, we found high-affinity methanotrophs such as *Methyloceanibacter*[45–48], as well as genera within the *Methylomonadaceae,* such as *g__KS41*, commonly found in acidic forest soils[49], and *Methyloglobulus*, usually found in lake sediments[50].

We further supported the patterns recovered from ancient DNA by analysing microbial lipid biomarkers and their carbon isotopic compositions for a subset of our samples. This data provided additional compelling evidence for a dynamic microbial turnover of methane by detecting lipids from methanogens and methanotrophs. Among the archaeal lipids, we identified a predominance of archaeol and hydroxy- archaeol[51] (Supplementary Table S7), which are specifically abundant in methanogenic archaea, particularly *Methanothrix* and *Methanosarcina*[51]. Carbon isotopic analysis for archaeol in sample 69_34 revealed a $\delta^{13}C$ value of -21.6‰, consistent with CO_2 and/or acetate as a carbon source for lipid biosynthesis (Supplementary Information, section 6). The detection of GDGT-0 further corroborates the presence of hydrogenotrophic methanogens, aligning with the DNA-based results.

Additionally, the lipid pool showed a notable abundance of bacteriohopanepolyols (BHPs) derived from soil bacteria, increasing from 28% in B1 to 41% in B3 (Fig. 4B). This is an indicator of a

marked input of soil organic matter, consistent with fluvial deposits originating from permafrost environments as well as forest soils[52]. In unit B2 (Fig. 4B), the high relative proportion of typical methanotrophic BHPs, such as aminotetrol and aminopentol, combined with a distinctly ^{13}C-depleted signal observed in the hopanoid diplopterol (hopan-22-ol) with a $\delta^{13}C$ value of -61.6‰, points towards localised hotspots of methane oxidation by aerobic methanotrophic bacteria upstream. Such methanotrophic activities potentially contributed to reducing methane emissions in high-latitude ecosystems (Supplementary Information, section 3) despite the northward expansion of wetlands.

We also identified microorganisms involved in carbon and nitrogen cycling processes. 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[53]), and members of the genus *Propionospira*, which produce propionate, acetate, and CO_2 as the main products of their carbohydrate fermentation[54,55]. We also detected predicted homoacetogenic members of the family f__UBA5619 (g__UBA5619), which may contribute fermentation products to *Methanothrix* while removing H_2 and formate from the environment[56]. *Eubacteriaceae* (g__UBA1792), capable of degrading substrates such as fructose and mannose (Supplementary Table S4), and g__JADGOW01, possessing the necessary gene repertoire for acetogenesis (Supplementary Table S4), were also identified. We additionally observed members from the family *MBNT15* (g__CG2-30-66-27), commonly found in peatlands, where they act as scavengers by fully mineralising small organic molecules produced during the microbial breakdown of complex polymeric compounds[57,58]. We detected ammonia-oxidising archaea (AOA) from the *Nitrososphaeraceae* (g__UBA10452 and g__TA-21), which are found in acidic polar and alpine soils[59] and a High-Arctic glacier[60], respectively, and may indicate the presence of aerobic niches. Besides the AOA, we also identified bacterial taxa involved in nitrogen fixation, including *Pseudolabrys* (73% detection rate) and *Bradyrhizobium*, as well as comammox bacteria from *Nitrospiraceae* (g__Palsa-1315) capable of complete ammonia oxidation[61]. 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 from *g__AD8-1* to our data was reconstructed from permafrost

284 samples with marine influence[62]. In contrast, the closest reference from g__6H3-1 was isolated
285 from deep-sea sediment samples of the Hikurangi Subduction Margin[63].

286 Microbial carbon processing

287 Our taxonomic profiling analysis (Fig. 4A) shows that the dominant group consists of CO_2-
288 reducing hydrogenotrophic members and other versatile methanogenic archaea. Furthermore, our
289 results revealed the presence of high-affinity methanotrophic bacteria in the samples. We
290 complemented these taxonomic analyses with ecosystem-wide metabolic reconstructions[64] to
291 identify the primary metabolisms associated with methanogenesis two million years ago.
292 Peatlands store vast amounts of carbon from plant polymers such as cellulose and hemicellulose.
293 In present-day peatlands, these polymers are degraded by microorganisms, forming glucose,
294 xylose, and N-acetylglucosamine. These compounds are further converted into low-molecular-
295 weight alcohols and organic acids such as ethanol, propionate, acetate, and lactate, as well as H_2
296 and CO_2 via fermentation. These fermentation products serve as substrates for CO_2-reducing
297 hydrogenotrophic and acetoclastic methanogens, which utilise H_2/CO_2 and acetate, respectively[65].
298 Given the ubiquity of these processes in modern peatlands, we expected to observe evidence of
299 similar metabolic pathways in our ancient samples. We compiled a collection of standard and
300 custom KEGG modules and specific CAZymes[28] (Carbohydrate-active enzymes) to reveal the
301 different steps involved in carbon processing by the pioneer communities. Our analysis showed
302 evidence of all carbon processing steps in our samples (Fig. 5, Supplementary Table S12-13), even
303 when only using reads that map to the pioneer communities (Fig. 5, tiles with thicker borders;
304 Supplementary Table S14-15). We applied a highly conservative approach by using translated
305 searches with ultra-short damaged reads and only considering those modules with completion over
306 80%. In agreement with the taxonomic profiling, we identified pathways for CO_2-reducing
307 hydrogenotrophic, acetoclastic, and methylotrophic methanogenesis, as well as aerobic
308 methanotrophy (Fig. 5). The co-occurrence of this pathway suggests that a significant proportion
309 of CH_4 might have been oxidised, thereby limiting emissions to the atmosphere two million years
310 ago[47,66]. Our results may underestimate the true functional potential of the ancient microbial
311 community due to cytosine deaminations, which can introduce stop codons and non-synonymous
312 substitutions in the predicted open reading frames used in our translated searches. Our stringent
313 pathway identification thresholds, coupled with the presence of deaminations, likely contributed

to the failure to identify methane pathways in B1. Indeed, the average damage values for the methanogenic archaea in B1 were 0.48 ± 0.05, amongst the highest values estimated in our dataset, a pattern also observed in the eukaryotic fraction. However, we detected methane pathways in some sites from B1 (acetoclastic methanogenesis in 75_205B and CO_2-reducing hydrogenotrophic methanogenesis in 75_205A) by lowering the pathway identification threshold to 0.7.

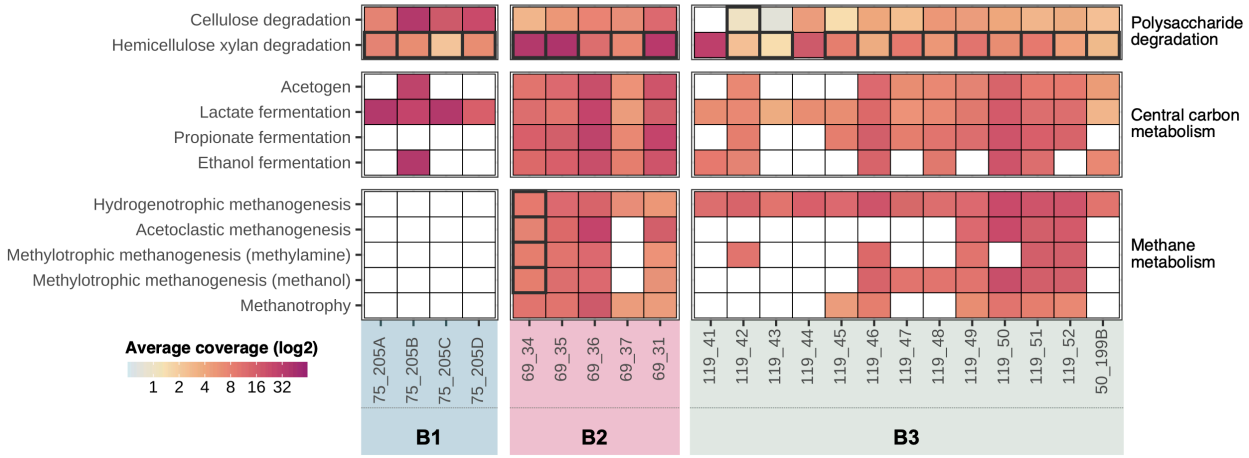


Fig. 5 | Carbon metabolism in the Kap København Formation Functional profile depicting pathways involved in the carbon metabolism in the permafrost. Tiles with thicker borders represent samples where the pathways were also detected using only reads that map to damaged references. White tiles indicate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

The Kap København Formation virome

Viruses play a major role in structuring microbial communities across ecosystems[67]; hence, we further explored the community structure of the Kap København Formation by characterising its DNA virome. We recovered the reads mapping to references from the IMG/VR included in our database from the taxonomic profiling. To ensure the accuracy and reliability of the results, we only considered the references with a detection rate 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 (Supplementary Tables S4, S16). 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 and a curated collection of mobile genetic elements (MGE) associated with methanogenic archaea[68]. After the search step, we

337 identified which proteins were more likely to be present in a sample and then selected those
338 genomes with over 20% of their proteins detected. This approach detected 231 reference viruses,
339 establishing a detailed view of the two-million-year-old virome (Supplementary Tables S4, S16).
340 Although members of the *Caudoviricetes* again represented the dominant (70%) component of the
341 virome (Fig. 6B), a considerable fraction could be affiliated with the realms *Monodnaviria* (viruses
342 with ssDNA genomes) and *Varidnaviria* (non-tailed icosahedral viruses with double jelly-roll
343 capsid proteins). Notably, the identified monodnaviruses belong to three different phyla specific
344 to eukaryotic (*Cressdnaviricota*) and bacterial (*Hofneiviricota* [filamentous phages] and
345 *Phixviricota* [phiX174-like phages]) hosts. Members of the *Cressdnaviricota* represent some of
346 the simplest viruses associated with diverse eukaryotic hosts, including protists, fungi, plants and
347 animals. Additionally, *Varidnaviria* was represented by viruses of the class *Megaviricetes* (phylum
348 *Nucleocytoviricota*), which includes environmentally widespread eukaryotic viruses with big and
349 giant dsDNA genomes. Finally, the identified methanogenic MGE, which include members of the
350 *Caudoviricetes*, archaea-specific lemon-shaped viruses and several unclassified MGE (Fig. 6C),
351 are associated with *Methanosarcinales*, *Methanobacteriales* and *Methanomicrobiales*, all of which
352 were detected in our samples (Fig. 4A).
353
354 On average, the viral fragments we retrieved exhibited an amino acid similarity (PAM30) of 84%
355 to the references (Fig. 6D). Among them, the members of *Cressdnaviricota* displayed the largest
356 divergence from modern references, with similarities as low as 60%. These similarity values
357 indicate a significant divergence of the viruses found in the formation compared to modern
358 references. These viruses would have gone undetected if only nucleotide assignment methods were
359 employed. Both approaches identified 14 *Caudoviricetes* references (Fig. 6A), with three of them
360 showing substantial detection (>50%) and damage (>0.3). Considering the amount of damage,
361 these virus genomes show a high identity to the reference, with ANI^R exceeding 96%. Particularly
362 noteworthy is the similarity to the reference IMGVR_UViG_2617271244_000001 (detection:
363 64%, damage: 0.34) with an ANI^R of 97% and an amino acid similarity of 95%. This virus is
364 predicted to infect *Protochlamydia naegleriophila*[69], an obligate intracellular symbiont of the free-
365 living amoeba *Naegleria*, commonly found in warm freshwater[70]. *Caudoviricetes*,
366 *Cressdnaviricota* and the methanogenic viruses have the highest relative abundance, with certain
367 samples surpassing 20% of the total viral fraction (Fig. 6D, Supplementary Tables S4, S16).

368

369 The amino acid searches also provided information about the functional potential of the viruses
370 we recruited (Supplementary Table S17). Although a large proportion (71%) of the recruited viral
371 proteins lack functional annotation, the remaining 29% of proteins validated the taxonomic
372 assignment of the corresponding reads and provided valuable insights into the functional capacity
373 of the virus community (Fig. 6D, Supplementary Table S17). In particular, the two largest
374 functional categories corresponded to taxon-specific virus structural proteins (16%) and proteins
375 involved in DNA/RNA transactions (genome replication, transcription, recombination; 9.5%),
376 respectively. We also identified proteins responsible for host cell lysis and toxin-antitoxin and
377 restriction-modification systems. Finally, viruses encoded diverse auxiliary metabolic genes
378 involved in various metabolic pathways (phosphoheptose isomerase, UDP-glucose
379 dehydrogenase, phosphoadenosine phosphosulfate reductase, sulfatase-modifying factor enzyme,
380 deoxyribonucleoside 5' monophosphate phosphatase, nitroreductase, glutamine amidotransferase,
381 cation efflux protein), enzymes for modification of various substrates (acetyltransferase,
382 nucleotide kinase, serine-threonine kinase, GtrB glycosyltransferase for O-antigen conversion),
383 hydrolases (glycoside hydrolase, esterase/lipase, proteases) as well as proteins responsible for
384 modulation of host responses (MazF-like growth inhibitor, sporulation stage III protein D,
385 translational regulator).
386 We leveraged the ecosystem information associated with each reference in the IMG/VR data to
387 gain insights into the environmental origins of the references identified through both approaches
388 (Fig. 6E). Consistent with the source tracking analyses, most of the references originate from
389 aquatic freshwater biomes. Notably, the references with the highest ANI^R and amino acid
390 similarity are derived from environments that resemble permafrost conditions. The biome labelled
391 Aquatic;Marine;Coastal corresponds to samples obtained from arctic subzero sea-ice and cryopeg
392 brines, while the Non-marine Saline and Alkaline;Saline biome represents samples collected from
393 Lost Hammer Spring in Axel Heiberg Island situated in the Arctic Ocean, as well as various lakes
394 in Antarctica.

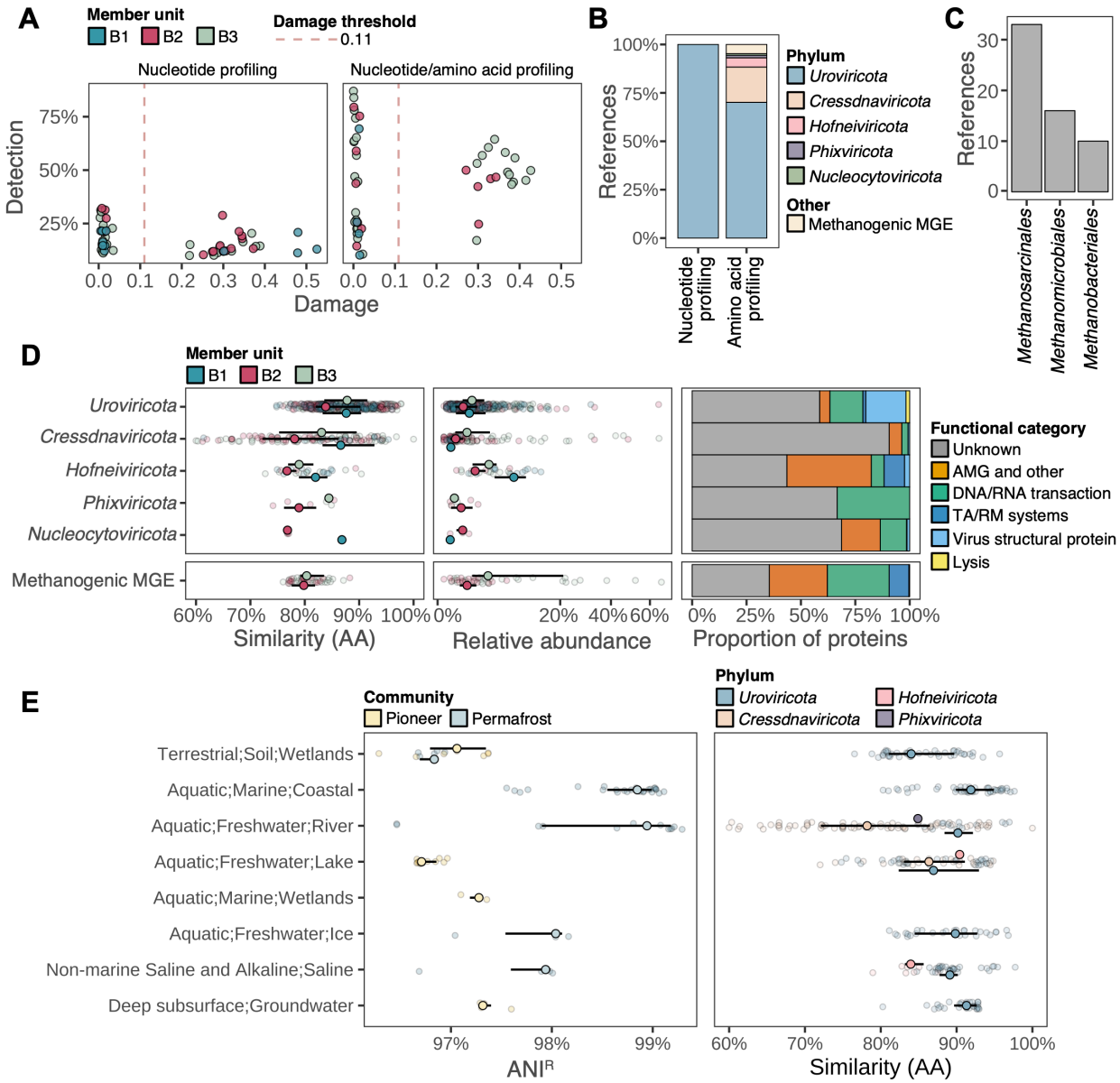


Fig. 6 | The Kap København Formation virome A, Each point depicts a viral reference identified in a sample by nucleotide profiling on the left panel and by nucleotide and amino acid profiling on the right panel. Red dashed line depicts the damage threshold we used to delimit the pioneer from the permafrost communities. B, Proportion of references detected in the whole dataset by phylum and methanogenic mobile genetic elements. C, Number of references associated with each predicted host for the methanogenic MGEs. D, The first two panels show the distribution of 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

A "time-travelling microbe" at the Kap København Formation

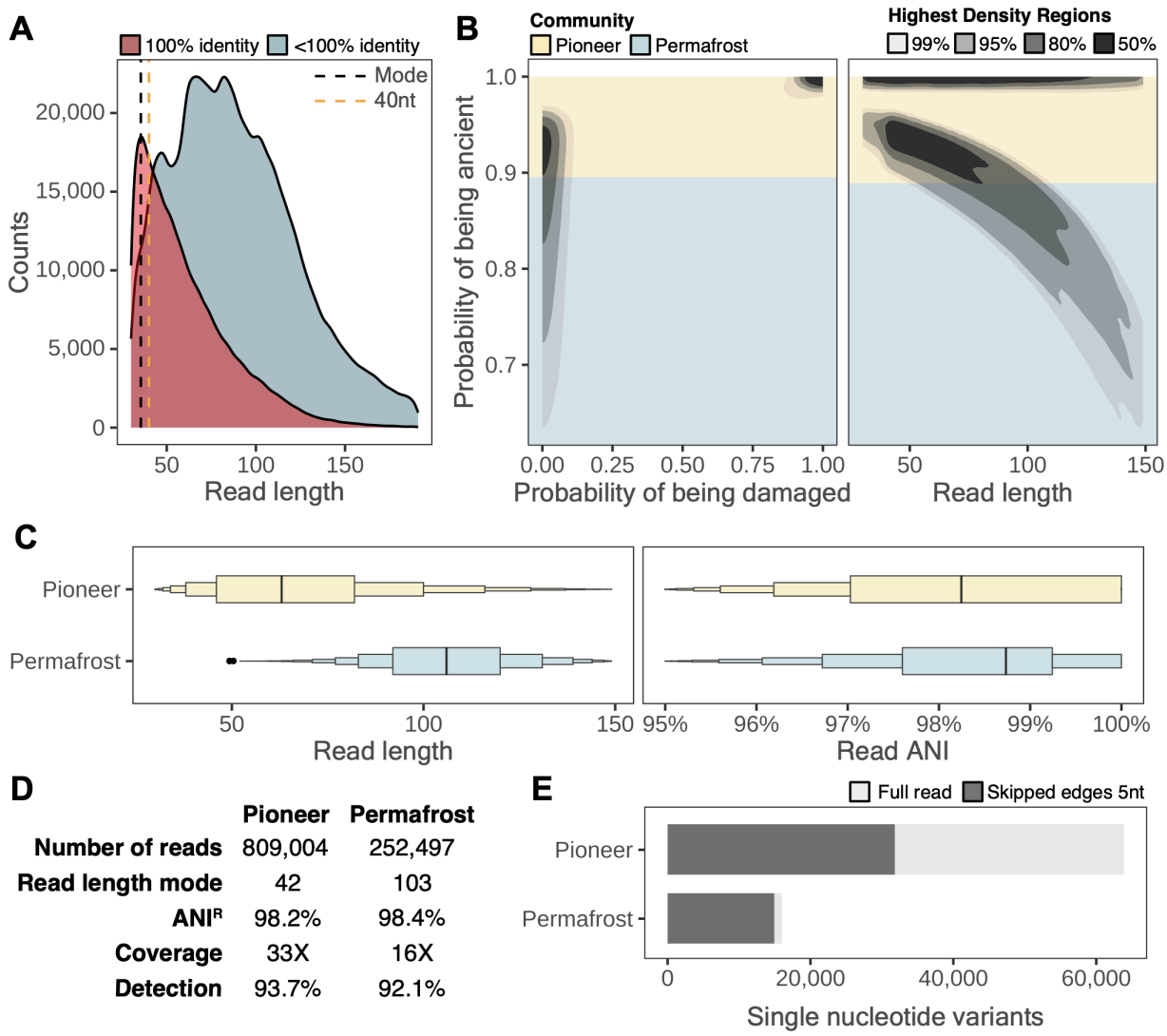
The remarkable sequence similarity between the pioneer microbial communities in Kap København and modern references from permafrost thawing gradients[28] offers an exceptional opportunity to gain ecological and evolutionary insights from our data. Members of the genus *Methanoflorens* (*g__Bog-38*), which had the highest read recruitment level in our study at high identity (>98%) despite the presence of damage (median damage across samples of 0.3), represent a fascinating case. *Methanoflorens* is a CO_2-reducing methanogen which is widely distributed in high methane-flux habitats. In the context of thawing permafrost, *Methanoflorens* shows a high abundance compared to other methanogenic microorganisms and is a significant contributor to methane production[32]. The *Methanoflorens* reference 3300025461_7, which recruited most of the reads in all samples of our study, is a metagenome-assembled genome derived from a drained thaw lake basin in the Arctic coastal plain near Barrow, Alaska[71].

We decided to test whether the reads showing >98% similarity to contemporary *Methanoflorens* from thawing permafrost could originate from "time-travelling" microbes[16]. Each of our samples represents an isolated microcosm that has undergone multiple environmental transitions, beginning as terrestrial sediment accumulating in the delta, passing through shallow marine sediments, and ultimately becoming permafrost for the past million years following the uplift of the formation (Fig. 1C). If these reads indeed derive from "time-travelling microbes", we expect to find a mixture of both damaged and undamaged DNA fragments from the same *Methanoflorens* lineage, spanning different periods. The pioneer communities, those from the deepest past, should have clear post-mortem damage signals, such as shorter fragment lengths and deamination patterns. In contrast, fragments from organisms that have remained dormant should lack these damage patterns and be longer.

To disentangle the signal of these pioneer methanogens (reads with post-mortem damage) and those who survived in the permafrost (reads without post-mortem damage), we analysed the data collected from sample 119_50, where the genome with the accession 3300025461_7 successfully recruited 1,114,032 reads (estimated damage: 0.26, detection: 94% and ANI^R: 98.2%). We used a novel method[72] 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 from our analysis reads that exhibited a 100% identity to the reference and had a read length smaller than 40 nt (modal read length + 5nt) from our analysis.

438 This precautionary step minimised potential interferences caused by a skewed distribution of
439 identical reads (Fig. 7A). An assumption of the method is the requirement for prior knowledge of
440 the modern contamination rate in the sample. Given the inherent difficulty in estimating this rate
441 for our sample type, we adopted a pragmatic approach. We employed modern contamination rates
442 of 5% and 10% and then calculated the average of the resulting probabilities. The results of the
443 inference highlight that we have a group of reads that have a high probability of being ancient
444 (AN) and damaged (PD) and a second group that has a low probability of being damaged but a
445 high probability of being ancient due to their read length (Fig. 7B, left panel). The pioneer fraction
446 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
447 permafrost had $\boldsymbol{\lambda}$=0.23, $\boldsymbol{\delta}_d$=0, $\boldsymbol{\delta}_s$=0, $\boldsymbol{\nu}$=0. The larger values of $\boldsymbol{\delta}_d$, $\boldsymbol{\delta}_s$, and $\boldsymbol{\nu}$ and smaller values of
448 $\boldsymbol{\lambda}$ typically indicate a more severe level of DNA damage. Thus, Briggs parameters suggest that the
449 pioneer fraction shows clear indications of damage, unlike the permafrost fraction, confirming the
450 ancient nature of reads from the sample 119_50 that matched the reference 3300025461_7 with
451 high levels of identity.
452 We also investigated the association between the probability of a given read to be ancient and its
453 length, where a gradual decrease in probability as a function of increasing read lengths is expected.
454 To establish the ANI threshold for distinguishing read fractions associated with the pioneer or
455 permafrost communities, we selected the lowest limit within the 50% highest density region,
456 considering both PD and read length. We added the excluded reads due to their 100% identity to
457 the reference to the pioneer read fraction; it is a common feature in ancient metagenomic studies
458 that many short reads are 100% identical to the reference. The reads assigned to pioneer
459 communities had a significantly smaller read length (mode: 42 nt) than those assigned to
460 permafrost communities (mode: 103 nt). Despite these differences, the ANI^R values remained
461 remarkably consistent (Fig. 7C). In both cases, we recovered more than 92% of the modern
462 reference genome, with coverage of 33X and 16X for the pioneer and permafrost fractions,
463 respectively. However, differences in coverage are unlikely quantitative since our extraction
464 protocols prioritise the recruitment of short DNA fragments, which reduces DNA recovery from
465 living organisms.
466 Lastly, we implemented a new profiling method in anvi'o[73] to exclude the edges of short reads
467 while calculating single-nucleotide variants (SNVs) for more accurate characterisation of genuine
468 genetic variants in ancient DNA by minimising the impact of post-mortem changes. Using this

approach, we re-profiled each fraction by excluding five nucleotides 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 extent of genetic variation we observed 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 genetic variants in the permafrost fraction demonstrates how the community is shaped by selection for survival and growth under relatively stable 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 "time-travelling" microbe 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, The Highest Density Regions plot shows 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 reference 3300025461_7 for both fractions. D, Table summarising the statistics of the mapping of each 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).

Discussion

Our study provides unprecedented insights into the structure and function of the bacterial, archaea, and viral communities in the Kap København Formation during a 20,000-year period two million years ago. By distinguishing pioneer microbial communities from those that later thrived in the permafrost, we reinforce evidence of an ice-free North Greenland with a warmer, wetland-like ecosystem during the Pliocene/Pleistocene transition. Additionally, our reconstruction of carbon processing pathways suggests that high-latitude ecosystems may have contributed to moderate methane emissions driven by the northward expansion of wetlands. This aligns with early Pleistocene ice core data, which indicate that atmospheric methane concentrations and their glacial and interglacial fluctuations were not significantly higher than those observed after the mid-Pleistocene transition. However, while Arctic wetlands may have served as an additional methane source, other processes, such as methane consumption by high-affinity methanotrophs, could have reduced emissions from these high-latitude ecosystems[66], counterbalancing these effects.

A striking aspect of our findings is the potential re-emergence of dormant methanogens preserved in permafrost over geological timescales. These "time-travelling" microbes[2,16], capable of reviving when conditions become favourable[28,71,74,75], provide a compelling perspective on microbial evolution and resilience. Members of the recovered genus *Methanoflorens* appear to play a central role in methane cycling, potentially reinforcing climate feedback loops in thawing permafrost[32]. Remarkably, despite being buried for at least two million years, these methanogens show >98% nucleotide sequence identity to modern relatives that have persisted across vast temporal and spatial scales, including a 3,000 km geographical separation from their closest known modern reference[71]. This observation supports the idea that not all organisms respond to environmental change in the same way. While some may remain dormant for millennia until favourable conditions return, others evolve or are replaced by new lineages in response to environmental change[2,16].

The recovery of pioneer microbial and viral communities, alongside those of plants and animals prior to permafrost formation, provides a more comprehensive reconstruction of the Arctic ecosystem two million years ago, when temperatures were 10C warmer. Intriguingly, while the plant and animal eDNA[1] reveal communities that vastly differ from what we can observe in the Arctic nowadays, the pioneer microbial community shows striking similarities to those of today's thawing Arctic. This suggests that the initial response to rising temperatures is a shift in the microbial community composition, serving as early indicators of climate-driven ecosystem transformations. It also raises the possibility that initial microbial changes similar to those seen at the Pliocene/Pleistocene transition shape the Arctic landscape for an eventual change of the plant and animal communities into something similar to what is seen at the Kap København Formation two million years ago.

525 ## Material and methods

526 ### Metagenomic data

527 We retrieved the shotgun metagenomic data from Kjær et al. (ENA project accession
528 PRJEB55522) and pre-processed it according to the same protocol described in Kjær et al. We
529 analysed 41 samples (53 metagenomic libraries) distributed across three stratigraphic units, B1,
530 B2 and B3, spanning five different localities (50, 69, 75 and 119). We created a short version of
531 the sample names used by Kjær et al. (Supplementary Table S1)

532 ### Post-depositional uplift history of Kap København Formation

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

541 ### Taxonomic database generation

542 We used a simplified version of the taxdb-integration workflow
543 (https://github.com/chassenr/taxdb-integration) to compile a unified taxonomy with ten levels that
544 combined NCBI and GTDB taxonomies for our genomic references. We sourced data from various
545 genomic datasets, including NCBI, PhyloNorway plastid sequences, complete GTDB v207, and
546 high-quality IMG/VR v4 genomic data. We also incorporated metagenome-assembled genomes
547 from TARA Oceans, Genomes from Earth's Microbiomes, the glacier microbiome catalogue, and
548 those recovered from a permafrost study.
549 To ensure consistency with GTDB v207, we re-annotated the bacterial and archaeal metagenome-
550 assembled genomes using gtdb-tk v2. Subsequently, we used derep-genomes
551 (https://github.com/aMG-tk/derep-genomes) to dereplicate the assemblies at the species or vOTU
552 level for viruses. We then employed the viral, archaeal, bacterial, and organelle dereplicated data

553 to construct a Bowtie2 database. Before the database construction, we concatenated the contigs of
554 each assembly with stretches of 50Ns.

555 Functional database generation

556 To generate a non-redundant version of the KEGG GENES database (downloaded in February
557 2022), we clustered the amino acid sequences of each KEGG orthologous entry using MMseqs2
558 (version b0b8e85f3b8437c10a666e3ea35c78c0ad0d7ec2). We used the following parameters for
559 clustering: *-c 0.8, --min-seq-id 0.9, --cov-mode 0, and --cluster-mode 2*. To ensure reproducibility,
560 we developed a Snakemake workflow, available at https://github.com/aMG-tk/kegg-db-setup.

561 Best search parameters estimation

562 To determine the optimal parameters for different searches, we used a workflow
563 (https://github.com/aMG-tk/aMGSIM-smk) that combined our workflow and *aMGSIM*
564 (https://github.com/aMG-tk/aMGSIM) tools to generate synthetic metagenomes. We used ten
565 samples from the Kap København formation (Supplementary Information, section 4;
566 Supplementary Table S18-S22) to model the community composition (bacteria, archaea, and
567 viruses), the fragment length distribution, and the damage patterns for each reference. Each
568 synthetic metagenome contained a maximum of 1000 damaged references (determined using
569 Bayesian damage estimates), 500 non-damaged references, and 10 million reads. We then
570 annotated these synthetic metagenomes using the taxonomy module in our workflow, exploring
571 different bowtie2 parameters (-N, -k [100, 250, 500, 750, 1000]), read ANI filtering thresholds
572 (92%, 93%, 94%, 95%, 96%), and breadth filtering (Supplementary Table S18-S22). We evaluated
573 the sensitivity and specificity of the taxonomic profiling, the abundance and damage estimations
574 using precision, recall, F1, and F05 for classification problems, and Spearman correlation and
575 median absolute error for quantitative comparisons.

576 Taxonomic profiling

577 To recover the taxonomic profiles from ancient metagenomic data, we employed the taxonomic
578 module of our workflow, an ancient metagenomics analysis workflow. We converted the reads
579 into super-reads using Tadpole, a kmer-based assembler included in the BBTools[76] software suite,
580 with strict parameters. The resulting super-reads were then dereplicated using VSEARCH[77] *–fastx-*
581 *uniques*. We used the dereplicated reads for functional and taxonomic profiling. Then, we mapped

582 the super-read dereplicated reference sequences to the original quality-controlled ancient reads,
583 and using Bowtie2, our workflow mapped the reads against the database we generated for this
584 study. We removed mapping duplicates using the MarkDuplicates program from Picard[78], and
585 filtered the BAM files with the *filterBAM* program (https://github.com/aMG-tk/bam-filter) using
586 the following parameters: *-N -g auto -e auto -n 100 -b 0.75 -c 0.4 -A 94 -a 90 --read-length-freqs*
587 *--read-hits-count --include-low-detection --min-breadth 0.01*. Finally, we estimated the post-
588 mortem damage using Bayesian estimates in metaDMG in local and LCA modes. For the specific
589 versions of the program used, check the files at https://github.com/GeoGenetics/2025-kapk-
590 microbial

591 Abundance table generation

592 We used *filterBAM* to estimate the taxonomic abundances based on the 80% truncated average
593 depth (TAD80). Briefly, we estimated the number of reads in the TAD80 region using the formula
594 N = (G * C) / L, where G stands for the length of the region where we estimated the TAD80, C for
595 the TAD80 coverage value and L for the average read length mapped to the reference. We then
596 estimated the taxonomic abundance by normalising the number of reads by the length of the region
597 where we estimated the TAD80 and scaled by one million.
598 As we used the *--include-low-detection* parameter in *filterBAM*, there will be cases when the
599 TAD80 estimated taxonomic abundance is 0. In this case, we used the taxonomic abundances
600 estimated by the number of reads mapped normalised by the reference length and scaled by 1
601 million.

602 Damage threshold estimation

603 To determine the intervals of post-mortem damage and the minimum number of reads required for
604 reliable damage estimates of non-eukaryotic references, we used metaDMG to estimate damage in
605 local mode and with a significance threshold of at least 2, focusing on the estimates for chloroplasts
606 and mitochondria references (Extended Data Figure 4A). We used letter-value plots[79] to analyse
607 the lower tails of the damage estimate distribution for all samples and references. Subsequently,
608 we considered only those references classified as damaged in all downstream analyses.

609 Contamination identification and removal

610 We followed the same processing procedure for the 13 negative extraction and library controls as
611 for the Kap København Formation samples (Supplementary Table S2). For each reference found
612 in a control, we assessed the number of mapped reads and their damage patterns and excluded
613 them from further analysis using an approach similar to that of Kjær et al.[1].
614

615 Microbial source tracking analyses

616 To create the source dataset for the source tracking analyses, we used *getBiomes*
617 (https://github.com/aMG-tk/get-biomes) to retrieve the biome information and associated raw
618 sequence data from MGnify and ENA. We ran getBiomes with the parameters *--ena-filter*
619 *{library_layout: PAIRED, library_strategy: WGS, library_source: METAGENOMIC,*
620 *library_selection: RANDOM, read_count: 10000000}' --combine --exclude-terms human,16S.*
621 Raw sequences were then processed using the QC module in our workflow with modern
622 metagenomics settings. This involved using fastp to filter low-quality reads and selecting the
623 forward pair for any read pairs that did not merge. The resulting metagenomes were then
624 taxonomically annotated using our workflow with modern metagenomics settings, which turned
625 off the read-extension step and set the read ANI to 95%.
626 We used the same database and procedure to annotate, filter, and estimate the taxonomic
627 abundances of the sinks. We combined the taxonomic annotations of the sources and sinks. We
628 filtered the combined dataset by excluding samples with fewer than 10,000 counts and species
629 observed fewer than three times in at least 20% of the samples, with a coefficient of variation of
630 less than three and a mean proportion across all samples of less than 1e-5. This filtering step was
631 undertaken to ensure that only consistently present and abundant species in the samples were
632 included in the final dataset. The resulting tables were exported as BIOM objects that were used
633 as input for meta-Sourcetracker using the parameters *--sink_rarefaction_depth 0 --*
634 *source_rarefaction_depth 0 --per_sink_feature_assignments --restarts 100 --draws_per_restart 5*
635 *--diagnostics*.
636 We used decOM with default parameters as a complement to meta-Sourcetracker. To create a
637 custom k-mer matrix, we employed the kmtricks pipeline, using the same sources as meta-

638 Sourcetracker. The parameters we used for kmtricks were *--kmer-size 29 --mode kmer:pa:bin --*
639 *nb-partitions 20000 --restrict-to-list 1000 and --recurrence-min 3*.

640 Extraction of reads from damaged references

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

647 Biome-associated metagenomic read-recruitment analyses

648 To investigate the distribution of reads found in damaged references across multiple biomes, we
649 constructed two additional bowtie2 databases. The first database consisted of MAGs from the
650 Genomes from Earth's Microbiomes catalogue[27] and the glacier microbiome catalogue[23]. For the
651 second database, we aimed to obtain a comprehensive overview of a permafrost thaw gradient
652 (palsa, bog, fern) by utilising MAGs recovered from Woodcroft et al.[28]. We employed the same
653 mapping, filtering (expected breadth ratio >= 0.25), and damage estimation strategy used in our
654 previous analyses.

655 Lipid biomarker and isotope analyses

656 Ancient eDNA results are supported by the analysis of microbial lipid biomarkers and their carbon
657 isotopic compositions for a subset of ten samples. For each of these, about 10 g of sediment was
658 extracted following a modified Bligh & Dyer procedure: Samples were ultrasonicated and
659 extracted into a solvent mixture of methanol, dichloromethane (DCM), and aqueous buffer
660 (2:1:0.8). A phosphate buffer (8.7 g L^{-1} KH_2PO_4, pH 7.4) and a trichloroacetic acid buffer (50 g L^{-}
661 ^{1}, pH 2) were each used twice. The sonicated sediment was centrifuged, the supernatants were
662 combined in a separatory funnel and DCM and milliQ water were added for phase separation.
663 After transferring the organic phase, the aqueous phase was extracted three more times with DCM.
664 The obtained total lipid extract (TLE) was dried under a gentle stream of nitrogen and stored at -
665 20°C.

666 Bacteriohopanepolyols and archaeal lipids were directly quantified in the TLE by high-
667 performance liquid chromatography-mass spectrometry (HPLC-MS) following the protocol
668 developed by Hopmans et al.[42]. A maXis quadrupole time-of-flight MS (Bruker Daltonik)
669 equipped with an electrospray ionisation source and coupled to a Dionex Ultimate 3000RS
670 UHPLC was employed. MS detection was carried out in positive ionisation mode while scanning
671 a *m*/*z* range from 150 to 2000. MS^2 scans were obtained in data-dependent mode and with the
672 addition of an inclusion list featuring the exact masses of known BHPs. Compounds were
673 identified according to exact precursor mass to charge (*m*/*z*), retention time and fragmentation
674 patterns.
675 Additionally, for one selected sample (69_34 in unit B2), the carbon isotopic compositions of
676 archaeol and diplopterol were obtained. From 10% of the dried TLE, *n*-hexane-soluble maltenes
677 were recovered and subsequently separated into fractions of increasing polarity using a SUPELCO
678 LC-NH2 glass cartridge. Compounds were analysed as trimethylsilyl ethers after reaction with
679 *N,O*-bis(trimethylsilyl)trifluoroacetamide (BSTFA) in pyridine at 70°C for half an hour and
680 identified on a Thermo Scientific ISQ gas chromatography-MS (GC-MS) system using a 30 m
681 Restek-5ms column (0.25 mm internal diameter, 0.25 µm film thickness). The initial GC oven
682 temperature was held at 100 °C for 1 min, increased to 150 °C at a rate of 10 °C min^{-1}, then to 320
683 °C at a rate of 4 °C min^{-1} and held for 41.5 min. Compound specific isotope analysis was
684 performed on a Trace GC Ultra coupled via a GC Isolink interface to a Delta V plus isotope ratio
685 MS (Thermo Fischer Scientific) using the same temperature program as for GC-MS analysis.
686 Stable carbon isotope values are reported in the delta notation as $\delta^{13}C$ values relative to Vienna
687 Pee Dee Belemnite with a precision of better than 1‰ based on repetitive measurements of an in-
688 house *n*-alkane standard. These values were corrected by mass balancing for the contribution of
689 carbon atoms added during derivatisation.

690 Functional profiling

691 We used the functional profiling module of our workflow to annotate the dereplicated super-reads
692 with MMseqs2, using fine-tuned parameters for ancient DNA amino acid searches (*--comp-bias-*
693 *corr 0 --mask 0 -e 1e-5 --exact-kmer-matching 1 --sub-mat PAM30.out -s 3 -k 6 --spaced-kmer-*
694 *mode 1 --spaced-kmer-pattern 11011101 --min-length 15 --format-mode 2 -c 0.8 --cov-mode 2 --*
695 *min-seq-id 0.6*). To perform the annotation, we used the non-redundant KEGG GENES database

696 and dbCAN2 v11 gene sequences (CAZyDB.08062022). We applied *xFilter*, which implements a
697 modified version of FAMLI to identify the KEGG/dbCAN2 genes most likely present in the
698 samples with the parameters *-n 25 -b 20 -e 1e-5 --breadth-expected-ratio 0 -f depth_evenness --*
699 *depth-evenness 1*. For KEGG-related results, xFilter produced output compatible with the
700 enzymes-txt mode of anvi-estimate-metabolism in anvi'o.
701 We expanded the standard KEGG modules collection in anvi'o using anvi-setup-user-modules to
702 include custom modules (Supplementary Table S12; defined as Woodcroft2018 and Borrel2023
703 in the module class) designed to explore carbon metabolism in permafrost. We ran anvi-estimate-
704 metabolism with the parameters *--add-coverage --output-modes*
705 *modules,module_paths,module_steps,hits --include-kos-not-in-kofam --user-modules*, and we
706 parsed the output to identify pathways with 100% completeness.
707 Furthermore, we extracted hits to CAZy families related to cellulose degradation (GH5, GH9,
708 3.2.1.4, GH51, GH6, GH7, GH48, 3.2.1.91), xylan degradation (GH5, GH8, GH10, GH11, GH43,
709 3.2.1.8, GH3, GH30, GH39, GH52, GH54, GH116, GH120, 3.2.1.37, GH67, GH115, 3.2.1.139,
710 CE1, CE2, CE3, CE4, CE5, CE6, CE7, CE12, 3.1.1.72), and xylose degradation (3.2.1.37, GH3,
711 GH30, GH39, GH10, GH43).

712 Reference metabolism estimation

713 We used the program *anvi-estimate-metabolism* from anvi'o with default parameters to estimate
714 the metabolisms of the archaeal and bacterial references that recruited reads in our study. Note that
715 we used a version of KEGG downloaded in January 2023 (for reproducibility, the hash of the
716 KEGG snapshot available via *anvi-setup-kegg-kofams* is d20a0dcd2128)

717 Virome characterisation

718 To increase the sensitivity of the viral taxonomic classification, we envisioned an approach to
719 identify which viral reference is most likely present in a sample. We created a protein database
720 containing 15,948,451 sequences from the viral sequences from IMG/VR v4 present in the
721 database we used for the nucleotide-based taxonomic profiling, and we enriched it with 16,441
722 proteins from a curated collection of mobile genetic elements (MGE) associated with
723 methanogenic archaea. We used MMseqs2 with the fine-tuned parameters for ancient DNA amino
724 acid searches (*--comp-bias-corr 0 --mask 0 -e 1e-5 --exact-kmer-matching 1 --sub-mat PAM30.out*

725 *-s 3 -k 6 --spaced-kmer-mode 1 --spaced-kmer-pattern 11011101 --min-length 15 --format-mode*
726 *2 -c 0.8 --cov-mode 2 --min-seq-id 0.6*) and then filtered the results with *xFilter* to identify the
727 most likely proteins present in the samples with the parameters *-n 25 -b 20 -e 1e-5 --breadth-*
728 *expected-ratio 0 -f depth_evenness --depth-evenness 1*. Then, we estimated the proteome
729 completion ratio for each reference and selected all the ones with a value over 10%. We
730 functionally annotated the selected proteins using HHblits using two iterations in combination with
731 PFAM v35 and PHROG v4. We followed a similar approach to that of Vanni et al.[80] to remove
732 overlapping domains and select those hits with a probability over 90% and a target or query
733 coverage larger than 40%. We manually curated the results and classified them into six functional
734 categories: "Unknown", "Virus structural protein", "DNA/RNA transactions/processing",
735 "Lysis", "TA/RM systems" and "AMG and other".

736 Estimation of probabilities of being ancient and being damaged

737 First, we extracted and aligned reads from damaged references against the taxonomic profiling
738 database. Unlike previous alignments, we did not use the -k option in bowtie2 and reported the
739 best alignment. The filtering and damage estimation steps were conducted as described earlier. We
740 then used the program *getRPercId* (https://github.com/aMG-tk/get-read-percid) to subset the BAM
741 files to the reference 3300025461_7 and calculate for each the read length and average nucleotide
742 identity (ANI). After exploring the read length and ANI, we subset the BAM files to reference
743 3300025461_7 and calculate the read length and average nucleotide identity (ANI) for each read.
744 Then, for a more formal characterisation of the deamination patterns observed in each ancient
745 DNA strand for each fraction of data, we estimated the Briggs parameters, namely $\boldsymbol{\delta}_d$ (deamination
746 rates within the double-stranded regions), $\boldsymbol{\delta}_s$ (deamination rates within the single-stranded
747 regions), $\boldsymbol{\nu}$ (nick occurring rate in the sample), and $\boldsymbol{\lambda}$ (overhang length distribution) using
748 ngsBriggs. The method is based on the counts of nucleotide changes (including but not limited to
749 C to T and G to A) at the first and last 15 cyclic positions of the strands and the position-specific
750 error rates. The tool conducts a multinomial regression to estimate the four parameters ($\boldsymbol{\delta}_d$, $\boldsymbol{\delta}_s$, $\boldsymbol{\nu}$
751 and $\boldsymbol{\lambda}$) that describe the deamination pattern of the ancient DNA samples. Under the assumption
752 that the length distributions of the ancient pioneer strands and modern contaminated strands are
753 distinguishable, the tool will then go through each strand within the length range [30,150) and
754 estimate the best-fit ancient and modern length distributions if the prior knowledge of the modern

755 contamination rate of the sample is provided. The previous four estimates can help to calculate the
756 deaminated probability conditioned on the focal strand being ancient (AN) and the posterior
757 probability of being ancient for each focal strand, given its information on nucleotide changes and
758 fragment length (PD). We used the following parameters for the inference *-isrecal 1 -model b -eps*
759 *0.05 and -eps 0.1*. After identifying the pioneer and permafrost fractions, we estimated the Briggs
760 parameters. Lastly, we used a new profile mode in anvi'o, specially designed for ancient DNA, to
761 infer the SNVs in each fraction where we exclude a certain number of positions from the read
762 edges using the option *--skip-edges 5*.
763
764

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 was supported by a research grant (15467) from the Danish foundation VILLUM FONDEN. YW acknowledges support from the NSFC BSCTPES Project (No. 41988101), the CAS Youth Interdisciplinary Team Fund, National Key Scientific and Technological Infrastructure project "Earth System Numerical Simulation Facility" (EarthLab, 2023-EL-ZD-000111), National Supercomputer Center in Wuxi, and SMU's Center for Research Computing. 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 utilising decOM and Nadine T. Smit for guidance in the analysis of bacteriohopanepolyols.

Data availability

All sequencing data used in this study are accessible at the European Nucleotide Archive under accession PRJEB55522. The re-extracted sequence data, taxonomic database, and other supplementary data can be found at ERDA with the following DOIs: https://doi.org/n59w, DOI: https://doi.org/n6ks, and DOI: XXX respectively.

Code availability

All code used in this study is available at https://github.com/GeoGenetics/2025-kapk-microbial and archived on ZENODO (DOI:XXX). The repository contains a reproducible workflow for generating the basic data, as well as the analytical pipeline that produced all tables and figures. The various tools included in the ancient metagenomic toolkit can be accessed at https://github.com/aMG-tk

References

1. Kjær, K. H. *et al.* A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. *Nature* **612**, 283–291 (2022).
2. Pedrós-Alió, C. Time travel in microorganisms. *Syst. Appl. Microbiol.* **44**, 126227 (2021).
3. Wunderling, N. *et al.* Climate tipping point interactions and cascades: a review. *Earth Syst. Dyn.* **15**, 41–74 (2024).
4. Hillebrand, H., Kuczynski, L., Kunze, C., Rillo, M. C. & Dajka, J.-C. Thresholds and tipping points are tempting but not necessarily suitable concepts to address anthropogenic biodiversity change—an intervention. *Mar. Biodivers.* **53**, 1–8 (2023).
5. Svenning, J.-C. & Sandel, B. Disequilibrium vegetation dynamics under future climate change. *American Journal of Botany* **100**, 1266–1286 (2013).
6. Burke, K. D. *et al.* Pliocene and Eocene provide best analogs for near-future climates. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 13288–13293 (2018).
7. Hansen, J. E. *et al.* Global warming in the pipeline. *arXiv [physics.ao-ph]* (2022).
8. Popescu, S.-M. *et al.* Pliocene and Lower Pleistocene vegetation and climate changes at the European scale: Long pollen records and climatostratigraphy. *Quat. Int.* **219**, 152–167 (2010).
9. Gosse, J. C. *et al.* PoLAR-FIT: Pliocene landscapes and Arctic remains—frozen in time. *Geosci. Can.* **44**, 47 (2017).
10. Bennike, O. & Böcher, J. Forest-tundra neighbouring the North Pole: Plant and insect remains from the Plio-Pleistocene Kap københavn formation, north Greenland. *Arctic* **43**, 331–338 (1990).
11. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).
12. 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).
13. Capo, E. *et al.* Environmental paleomicrobiology: using DNA preserved in aquatic sediments to its full potential. *Environ. Microbiol.* **24**, 2201–2209 (2022).
14. Torti, A., Lever, M. A. & Jørgensen, B. B. Origin, dynamics, and implications of extracellular DNA pools in marine sediments. *Mar. Genomics* **24 Pt 3**, 185–196 (2015).
15. Pedersen, M. W. *et al.* Ancient and modern environmental DNA. *Philos. Trans. R. Soc.*

Lond. B Biol. Sci. **370**, 20130383 (2015).

16. Brennan, G. L. & Logares, R. Tracking contemporary microbial evolution in a changing ocean. *Trends Microbiol.* **31**, 336–345 (2023).
17. Funder, S. V. *et al.* Late Pliocene Greenland - The Kap København Formation in North Greenland. *Bull. Geol. Soc. Den.* (2001).
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. Rodriguez-R, L. M., Tsementzi, D., Luo, C. & Konstantinidis, K. T. Iterative subtractive binning of freshwater chronoseries metagenomes identifies over 400 novel species and their ecologic preferences. *Environ. Microbiol.* **22**, 3394–3412 (2020).
22. 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.
23. Liu, Y. *et al.* A genome and gene catalog of glacier microbiomes. *Nat. Biotechnol.* **40**, 1341–1348 (2022).
24. Margesin, R. & Collins, T. Microbial ecology of the cryosphere (glacial and permafrost habitats): current knowledge. *Appl. Microbiol. Biotechnol.* **103**, 2537–2549 (2019).
25. McGhee, J. J. *et al.* Meta-SourceTracker: application of Bayesian source tracking to shotgun metagenomics. *PeerJ* **8**, e8783 (2020).
26. Duitama González, C. *et al.* decOM: similarity-based microbial source tracking of ancient oral samples using k-mer-based methods. *Microbiome* **11**, 243 (2023).
27. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0718-6.
28. Woodcroft, B. J. *et al.* Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**, 49–54 (2018).
29. 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

863 methane emissions. *Remote Sens. Environ.* **113**, 380–391 (2009).
864 30. Freeman, C. L. *et al.* Survival of environmental DNA in sediments: Mineralogic control on
865 DNA taphonomy. *bioRxiv* 2020.01.28.922997 (2023) doi:10.1101/2020.01.28.922997.
866 31. Wasmund, K. *et al.* Genomic insights into diverse bacterial taxa that degrade extracellular
867 DNA in marine sediments. *Nat Microbiol* **6**, 885–898 (2021).
868 32. Mondav, R. *et al.* Discovery of a novel methanogen prevalent in thawing permafrost. *Nat.*
869 *Commun.* **5**, 3212 (2014).
870 33. Borrel, G. *et al.* Stratification of Archaea in the deep sediments of a freshwater meromictic
871 lake: vertical shift from methanogenic to uncultured archaeal lineages. *PLoS One* **7**, e43346
872 (2012).
873 34. Schwarz, J. I. K., Eckert, W. & Conrad, R. Community structure of Archaea and Bacteria in
874 a profundal lake sediment Lake Kinneret (Israel). *Syst. Appl. Microbiol.* **30**, 239–254 (2007).
875 35. Purdy, K. J., Nedwell, D. B. & Embley, T. M. Analysis of the sulfate-reducing bacterial and
876 methanogenic archaeal populations in contrasting Antarctic sediments. *Appl. Environ.*
877 *Microbiol.* **69**, 3181–3191 (2003).
878 36. Zepp Falz, K. *et al.* Vertical distribution of methanogens in the anoxic sediment of Rotsee
879 (Switzerland). *Appl. Environ. Microbiol.* **65**, 2402–2408 (1999).
880 37. Prasitwuttisak, W., Hoshiko, Y., Maeda, T., Haraguchi, A. & Yanagawa, K. Microbial
881 Community Structures and Methanogenic Functions in Wetland Peat Soils. *Microbes*
882 *Environ.* **37**, (2022).
883 38. Chan, O. C. *et al.* Vertical distribution of structure and function of the methanogenic
884 archaeal community in Lake Dagow sediment. *Environ. Microbiol.* **7**, 1139–1149 (2005).
885 39. Rodrigues-Oliveira, T., Belmok, A., Vasconcellos, D., Schuster, B. & Kyaw, C. M.
886 Archaeal S-Layers: Overview and Current State of the Art. *Front. Microbiol.* **8**, 2597
887 (2017).
888 40. Kish, A. *et al.* Preservation of Archaeal Surface Layer Structure During Mineralization. *Sci.*
889 *Rep.* **6**, 26152 (2016).
890 41. Stevens, K. M. & Warnecke, T. Histone variants in archaea - An undiscovered country.
891 *Semin. Cell Dev. Biol.* **135**, 50–58 (2023).
892 42. Hopmans, E. C., Smit, N. T., Schwartz-Narbonne, R., Sinninghe Damsté, J. S. & Rush, D.
893 Analysis of non-derivatized bacteriohopanepolyols using UHPLC-HRMS reveals great

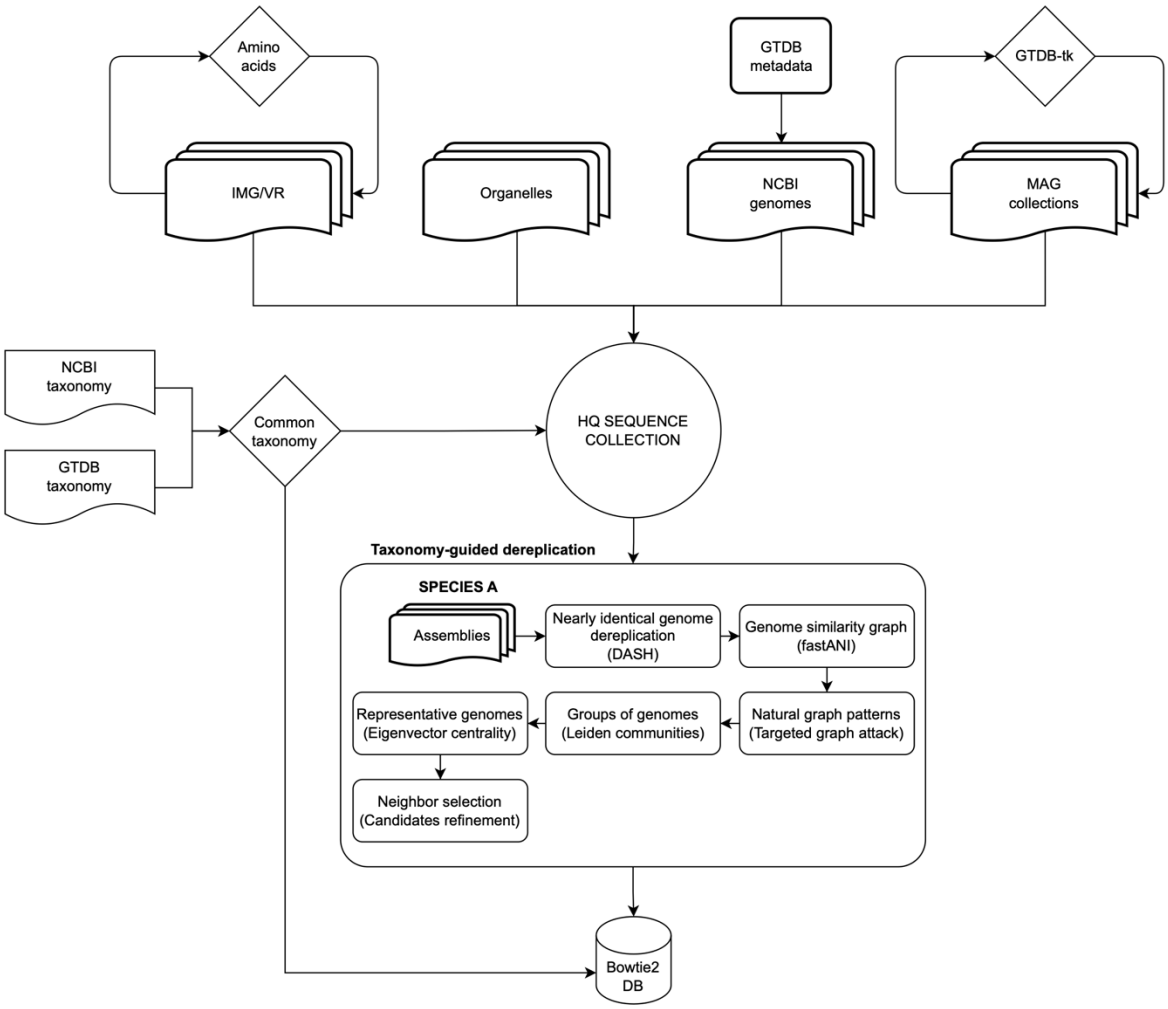
structural diversity in environmental lipid assemblages. *Org. Geochem.* **160**, 104285 (2021).

43. Glass, J. B. *et al.* Microbial metabolism and adaptations in Atribacteria-dominated methane hydrate sediments. *Environ. Microbiol.* **23**, 4646–4660 (2021).
44. 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).
45. 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).
46. 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).
47. 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).
48. D'Imperio, L. *et al.* Spatial controls of methane uptake in upland soils across climatic and geological regions in Greenland. *Commun. Earth Environ.* **4**, (2023).
49. Nguyen, N.-L. *et al.* Genomic Insights Into the Acid Adaptation of Novel Methanotrophs Enriched From Acidic Forest Soils. *Front. Microbiol.* **9**, 1982 (2018).
50. 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).
51. Koga, Y. & Morii, H. Recent advances in structural research on ether lipids from archaea including comparative and physiological aspects. *Biosci. Biotechnol. Biochem.* **69**, 2019–2034 (2005).
52. Cooke, M. P. *et al.* Bacteriohopanepolyol biomarker composition of organic matter exported to the Arctic Ocean by seven of the major Arctic rivers. *Org. Geochem.* **40**, 1151–1159 (2009).
53. 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).
54. Schink, B., Thompson, T. E. & Zeikus, J. G. Characterization of Propionispira arboris gen.

925 nov. sp. nov., a Nitrogen-fixing Anaerobe Common to Wetwoods of Living Trees.
926 *Microbiology* **128**, 2771–2779 (1982).
927 55. Horn, M. A., Matthies, C., Küsel, K., Schramm, A. & Drake, H. L. Hydrogenotrophic
928 methanogenesis by moderately acid-tolerant methanogens of a methane-emitting acidic peat.
929 *Appl. Environ. Microbiol.* **69**, 74–83 (2003).
930 56. Langwig, M. V. *et al.* Large-scale protein level comparison of Deltaproteobacteria reveals
931 cohesive metabolic groups. *ISME J.* **16**, 307–320 (2022).
932 57. Begmatov, S., Beletsky, A. V., Dedysh, S. N., Mardanov, A. V. & Ravin, N. V. Genome
933 analysis of the candidate phylum MBNT15 bacterium from a boreal peatland predicted its
934 respiratory versatility and dissimilatory iron metabolism. *Front. Microbiol.* **13**, 951761
935 (2022).
936 58. Seidel, L., Broman, E., Turner, S., Ståhle, M. & Dopson, M. Interplay between
937 eutrophication and climate warming on bacterial communities in coastal sediments differs
938 depending on water depth and oxygen history. *Sci. Rep.* **11**, 23384 (2021).
939 59. Pessi, I. S., Rutanen, A. & Hultman, J. Candidatus Nitrosopolaris, a genus of putative
940 ammonia-oxidizing archaea with a polar/alpine distribution. *FEMS Microbes* **3**, (2022).
941 60. Zeng, Y. *et al.* Potential Rhodopsin- and Bacteriochlorophyll-Based Dual Phototrophy in a
942 High Arctic Glacier. *MBio* **11**, (2020).
943 61. Overholt, W. A. *et al.* Carbon fixation rates in groundwater similar to those in oligotrophic
944 marine systems. *Nat. Geosci.* **15**, 561–567 (2022).
945 62. Liang, R. *et al.* Genomic reconstruction of fossil and living microorganisms in ancient
946 Siberian permafrost. *Microbiome* **9**, 110 (2021).
947 63. Sun, J. *et al.* Recoding of stop codons expands the metabolic potential of two novel
948 Asgardarchaeota lineages. *ISME Commun* **1**, 30 (2021).
949 64. Veseli, I. *et al.* Microbes with higher metabolic independence are enriched in human gut
950 microbiomes under stress. *bioRxiv* (2023) doi:10.1101/2023.05.10.540289.
951 65. Tveit, A., Schwacke, R., Svenning, M. M. & Urich, T. Organic carbon transformations in
952 high-Arctic peat soils: key functions and microorganisms. *ISME J.* **7**, 299–311 (2013).
953 66. Oh, Y. *et al.* Reduced net methane emissions due to microbial methane oxidation in a
954 warmer Arctic. *Nat. Clim. Chang.* **10**, 317–321 (2020).
955 67. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).

68. Medvedeva, S., Borrel, G., Krupovic, M. & Gribaldo, S. A compendium of viruses from methanogenic archaea reveals their diversity and adaptations to the gut environment. *Nat. Microbiol.* **8**, 2170–2182 (2023).
69. 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).
70. Stahl, L. M. & Olson, J. B. Environmental abiotic and biotic factors affecting the distribution and abundance of Naegleria fowleri. *FEMS Microbiol. Ecol.* **97**, (2020).
71. Lipson, D. A. *et al.* Metagenomic insights into anaerobic metabolism along an Arctic peat soil profile. *PLoS One* **8**, e64659 (2013).
72. Zhao, L., Henriksen, R. A., Ramsøe, A., Nielsen, R. & Korneliussen, T. S. Revisiting the Briggs ancient DNA damage model: A fast maximum likelihood method to estimate post-mortem damage. *Mol. Ecol. Resour.* **25**, e14029 (2025).
73. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* **6**, 3–6 (2021).
74. 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).
75. Ernakovich, J. G. *et al.* Microbiome assembly in thawing permafrost and its feedbacks to climate. *Glob. Chang. Biol.* **28**, 5007–5026 (2022).
76. Bushnell, B., Rood, J. & Singer, E. BBTools software package. Preprint at (2014).
77. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
78. Toolkit, P. Picard toolkit. *Broad Institute, Github Repository* (2019).
79. Hofmann, H., Wickham, H. & Kafadar, K. Letter-Value Plots: Boxplots for Large Data. *J. Comput. Graph. Stat.* **26**, 469–477 (2017).
80. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *Elife* **11**, (2022).

985 # Extended Data figures

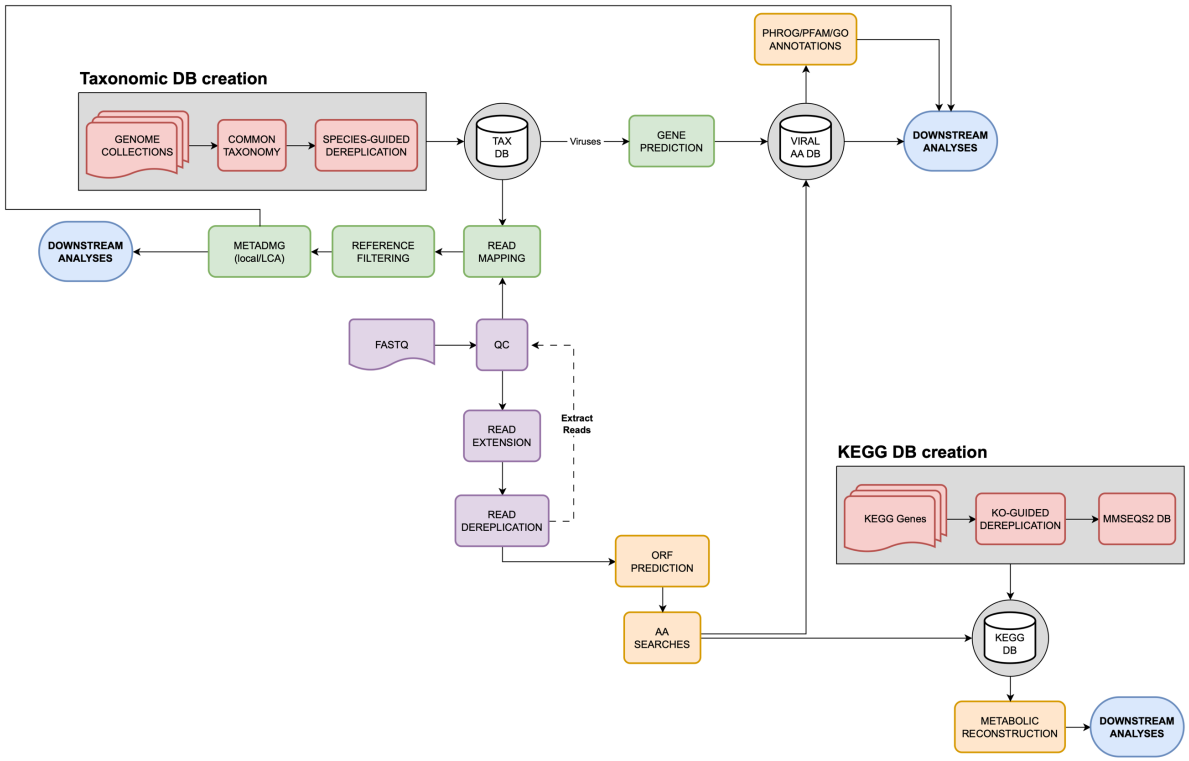


986

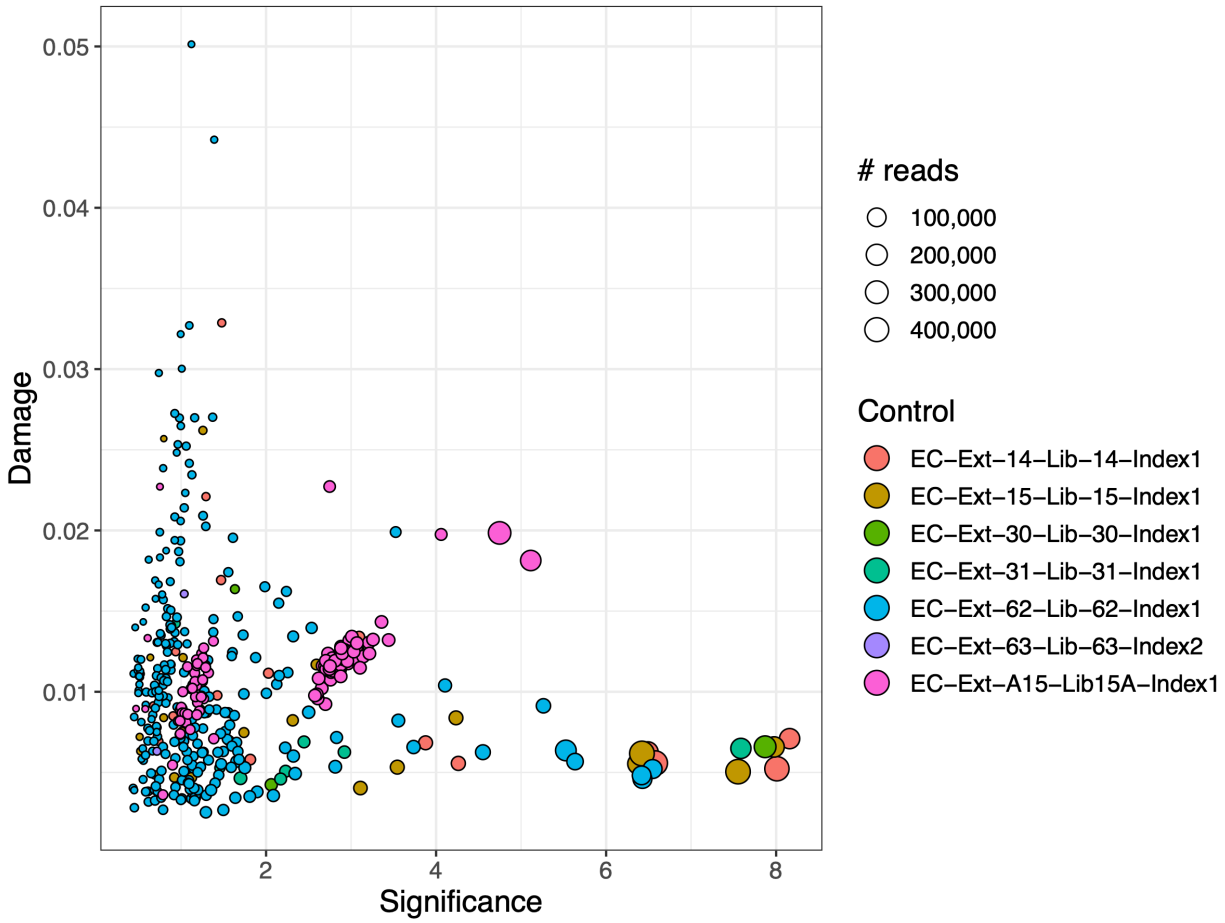
987 **Extended Data Fig. 1 | Taxonomic database creation workflow.** Diagram illustrating the workflow to
988 create the taxonomic database used for profiling the samples in this study.

989

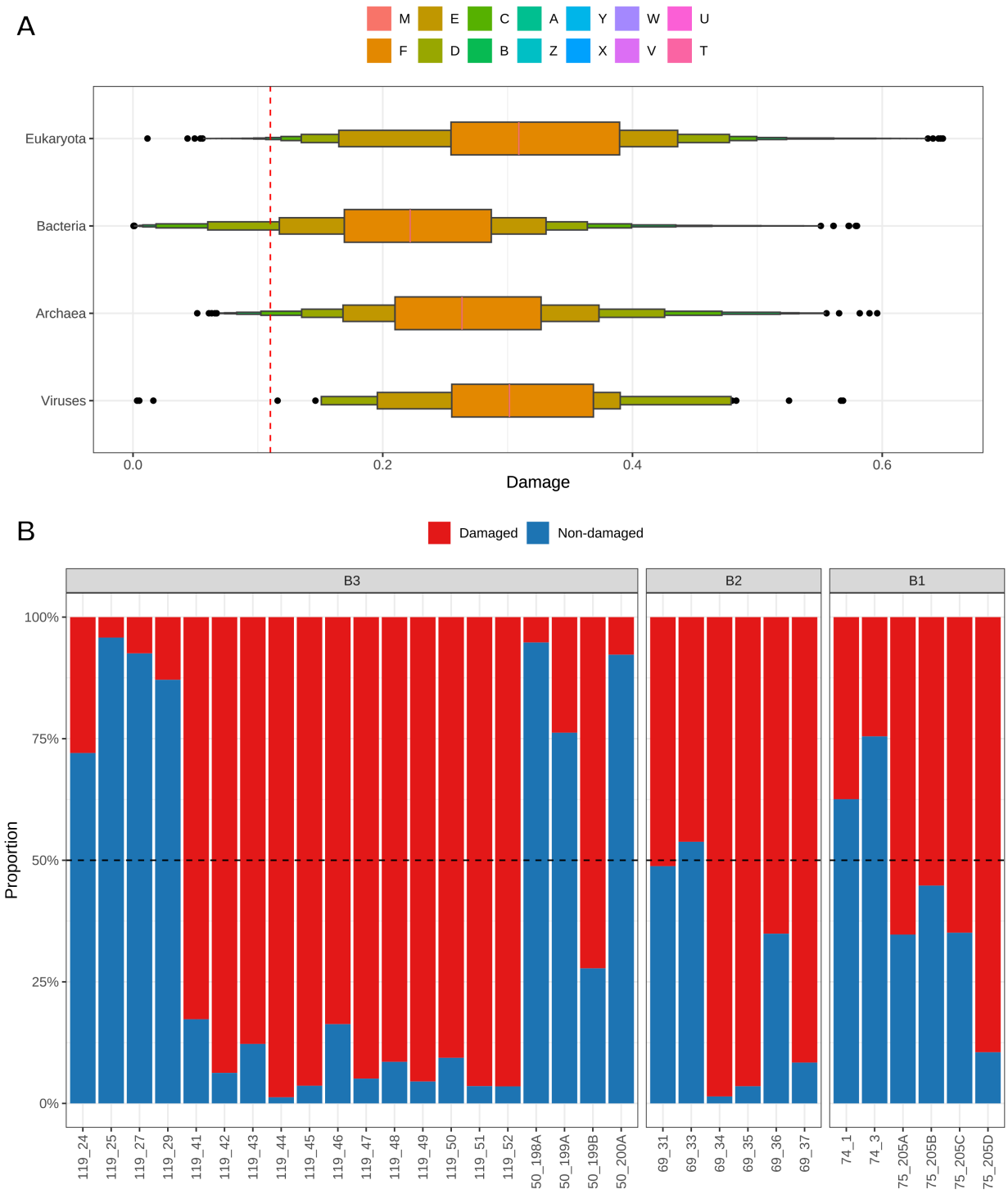
990



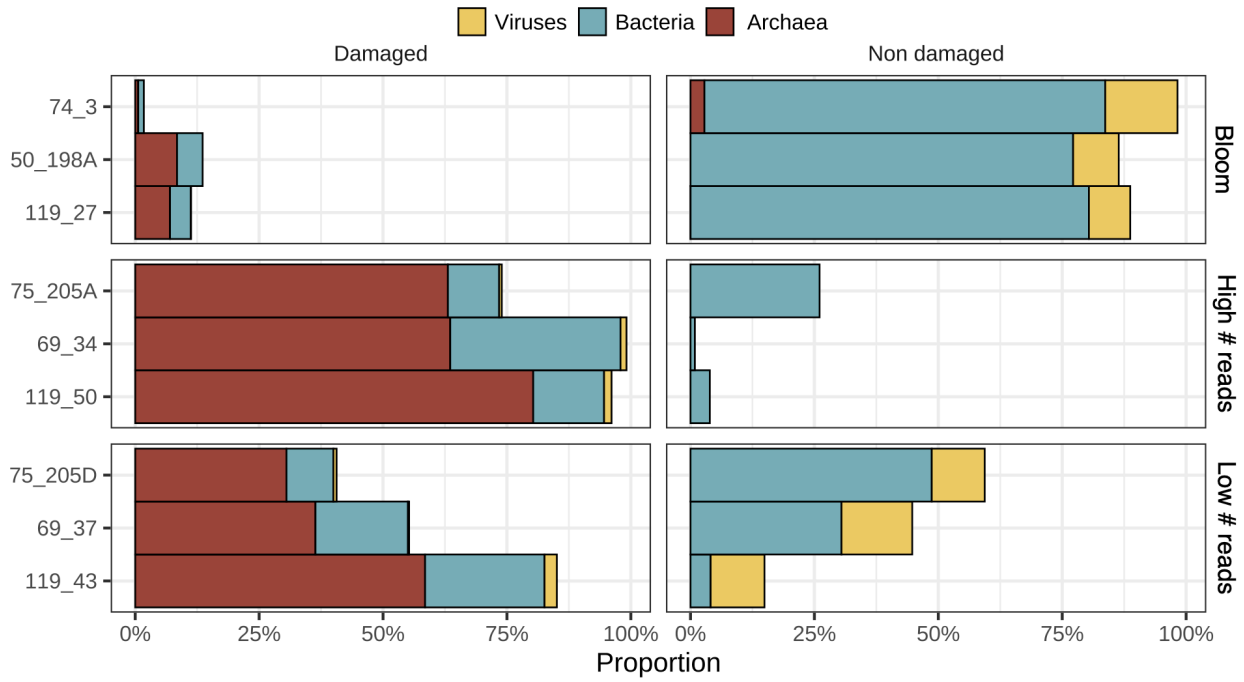
Extended Data Fig. 2 | Taxonomic and functional profiling workflow. Schematic diagram illustrating the main steps implemented in the taxonomic and functional analysis workflow used to process the samples in this study. In red are the steps where databases are created; in purple, the QC post-processing steps involved in creating super-reads and dereplication; in green, the steps for taxonomic profiling; and in orange, the functional annotation steps.



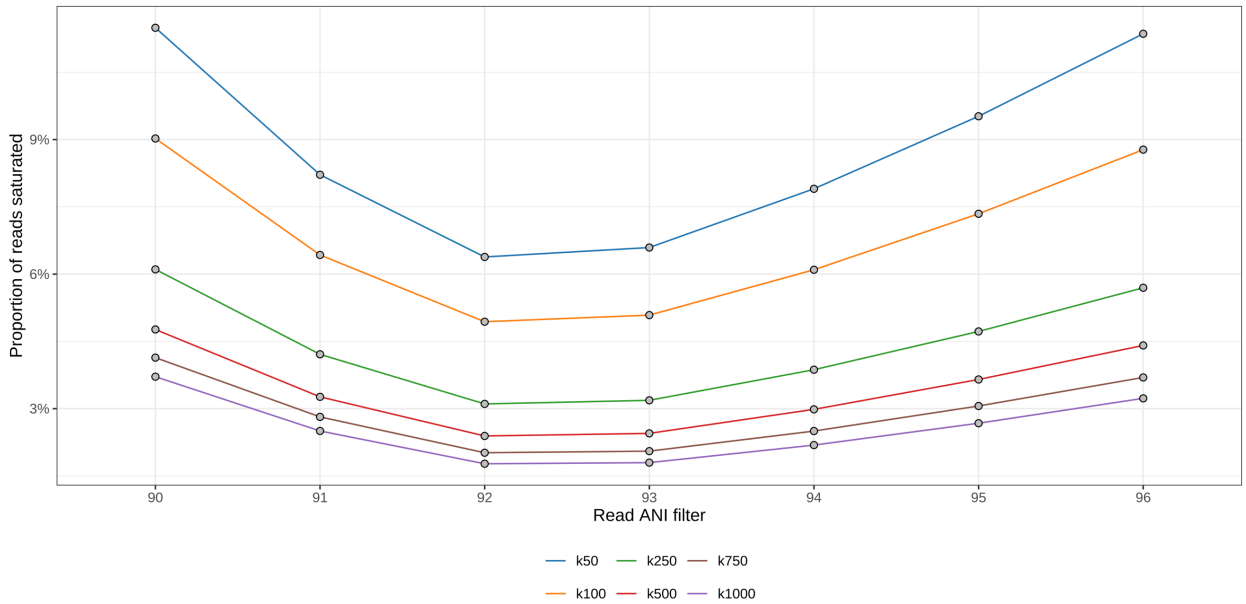
Extended Data Fig. 3 | Extraction controls. Each point in the plot corresponds to a taxon identified in the control samples based on damage and its significance. We did not detect any taxa with damage. The size of the point corresponds to the number of reads mapped, while the colour indicates the control sample (Supplementary Table S2).



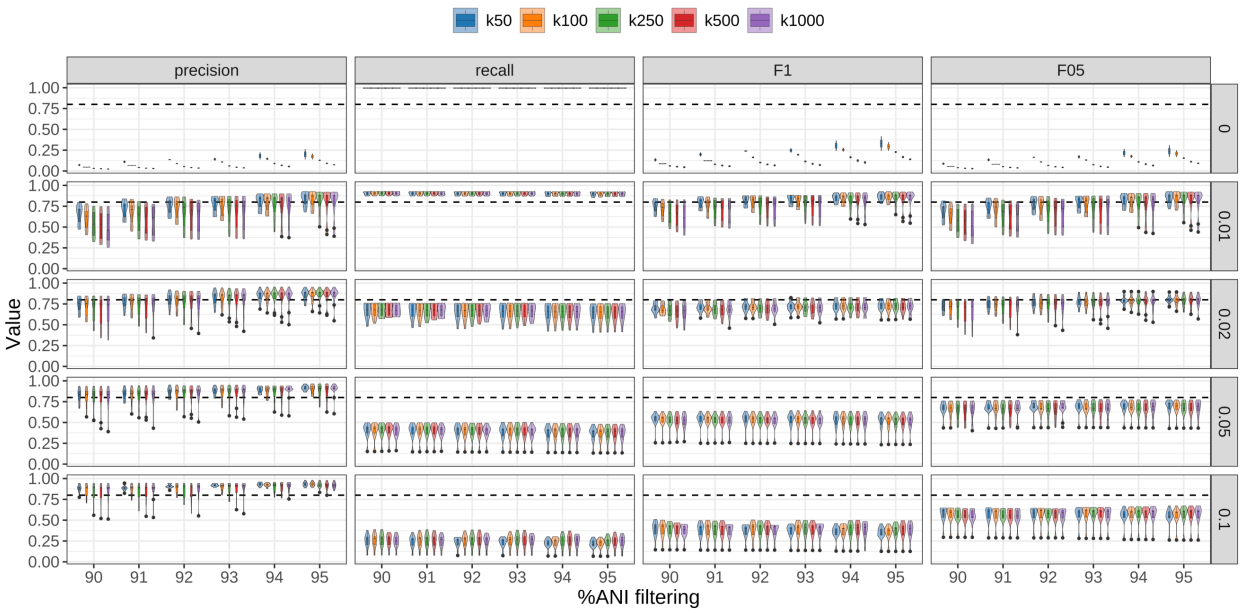
Extended Data Fig. 4 | Damage thresholds and sample selection. A) Letter-value plots illustrating the damage distribution for Archaea, Bacteria, Viruses, and Eukaryotes. The red line indicates the threshold to determine whether a taxon is potentially damaged. B) We discarded those samples that had less than 50% of the estimated TAD80 taxonomic abundances.



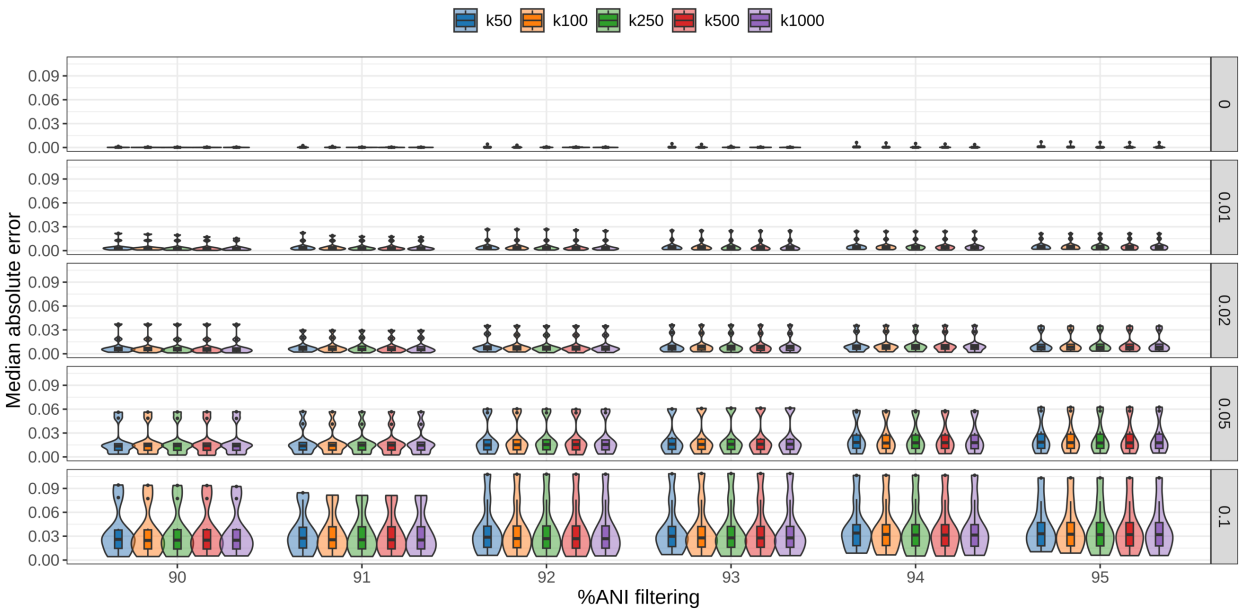
Extended Data Fig. 5 | Extraction bias and bloom assessment. Proportion of different taxa after re-extracting selected samples to assess potential extraction bias and blooming of taxa. We selected three distinct sets of samples: three samples enriched in taxa that do not show damage and six samples where Archaea are abundant and produced either a high or low number of reads in the original dataset from Kjaer et al., 2022.



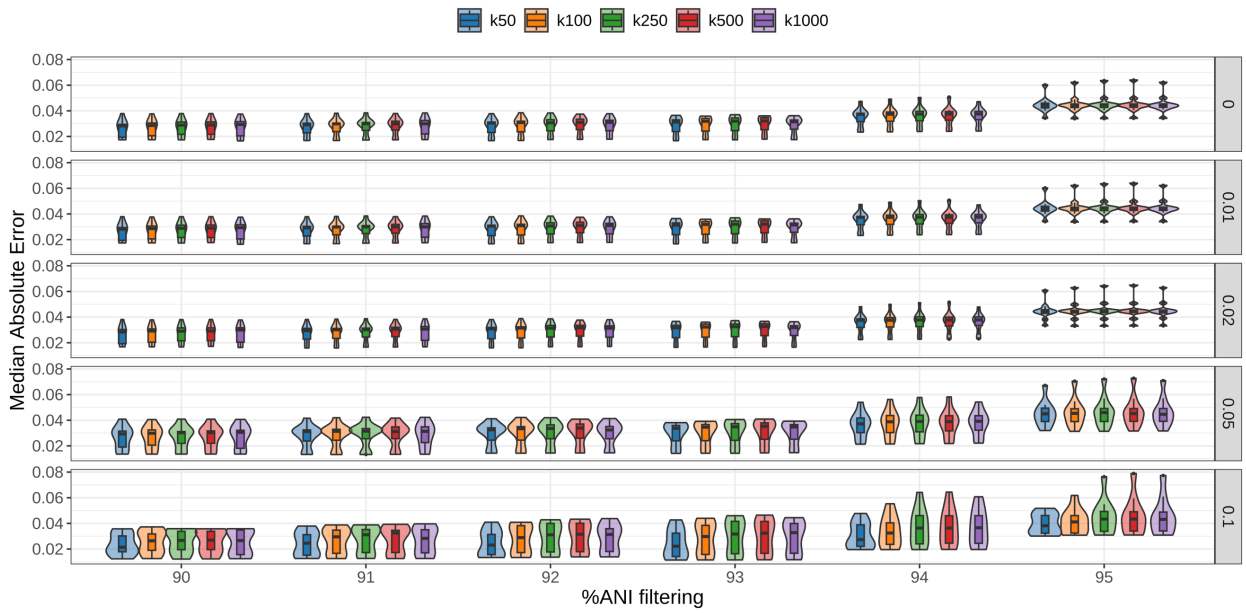
Extended Data Fig. 6 | Effect of read alignment saturation. We explored various values (50, 100, 250, 500, 750, and 1000) for the Bowtie2 parameter -k, determining the number of alignments reported for each read. Our evaluation spanned multiple Average Nucleotide Identity values (90-95%). The lower the values, the better.



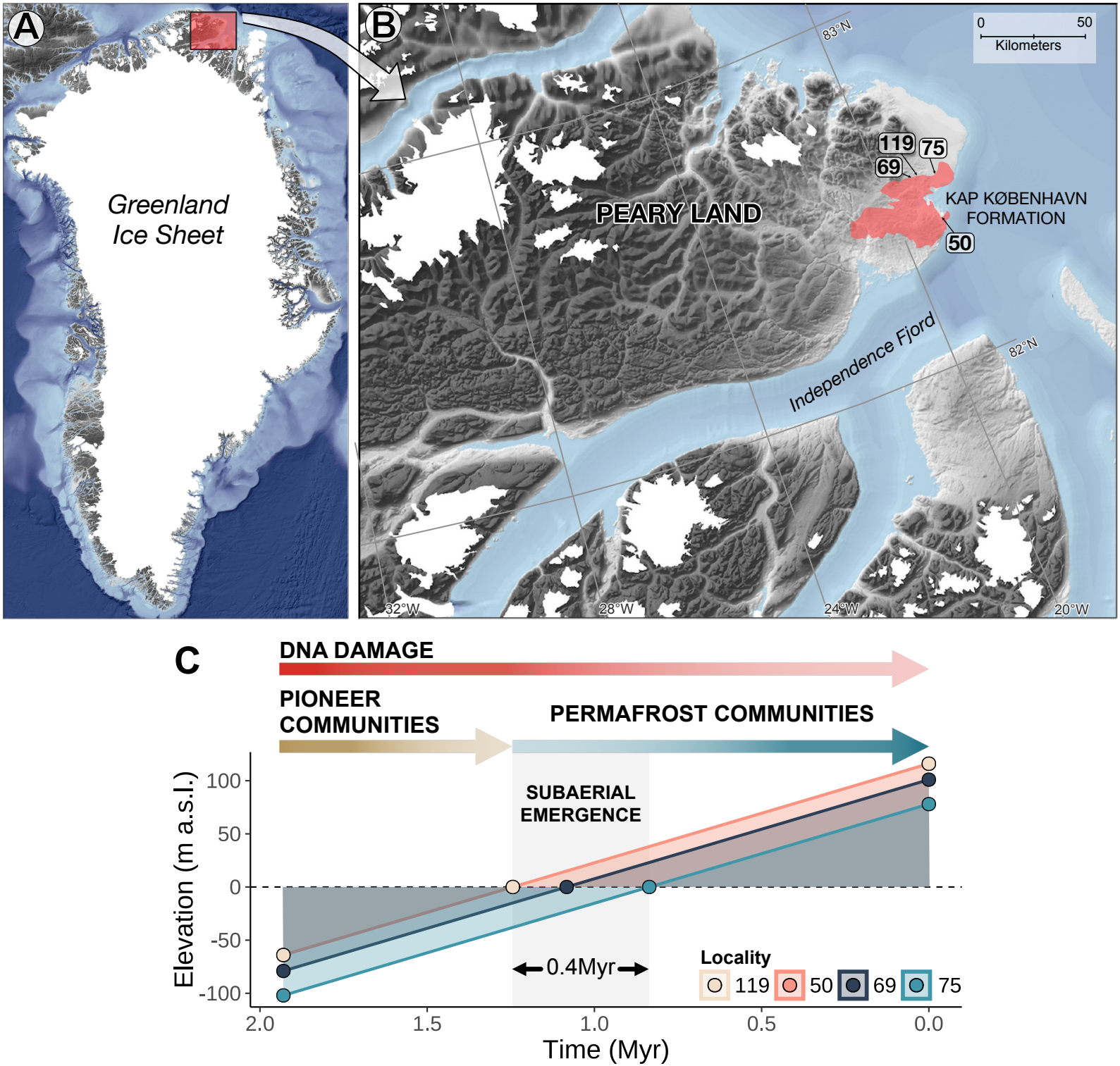
Extended Data Fig. 7 | Assessment of the taxonomic classification. To determine the optimal parameter combination that accurately recovers the correct taxonomic composition of the samples in this study, we examined several values (50, 100, 250, 500, 750, and 1000) for the Bowtie2 parameter -k, alongside filtering thresholds for ANI (90-94%) and breadth of coverage (0-10%) across 10 synthetic ancient metagenomes derived from the samples used in this study. We computed Precision, Recall, F1, and F0.5 for each parameter combination. The higher the values, the better.

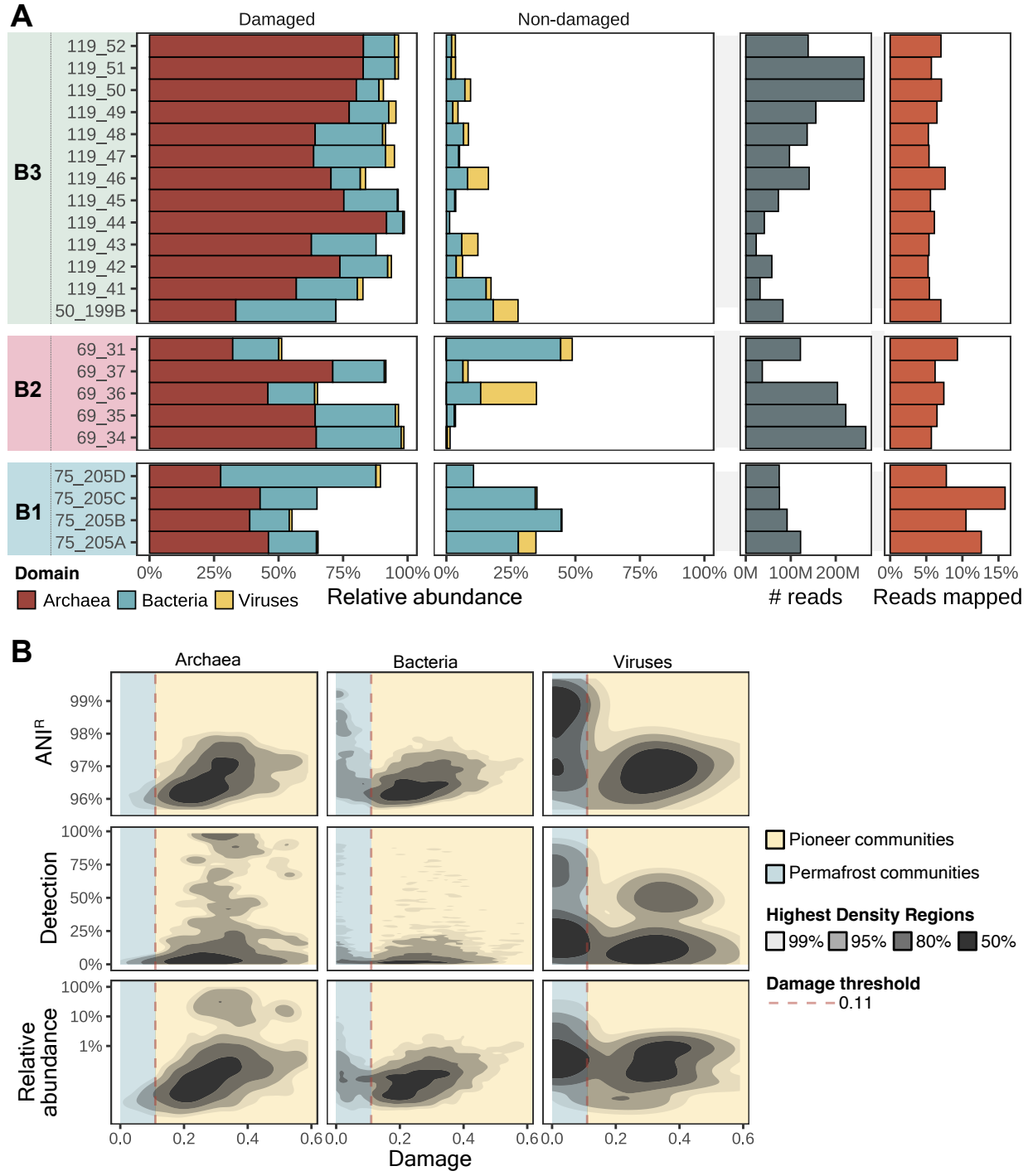


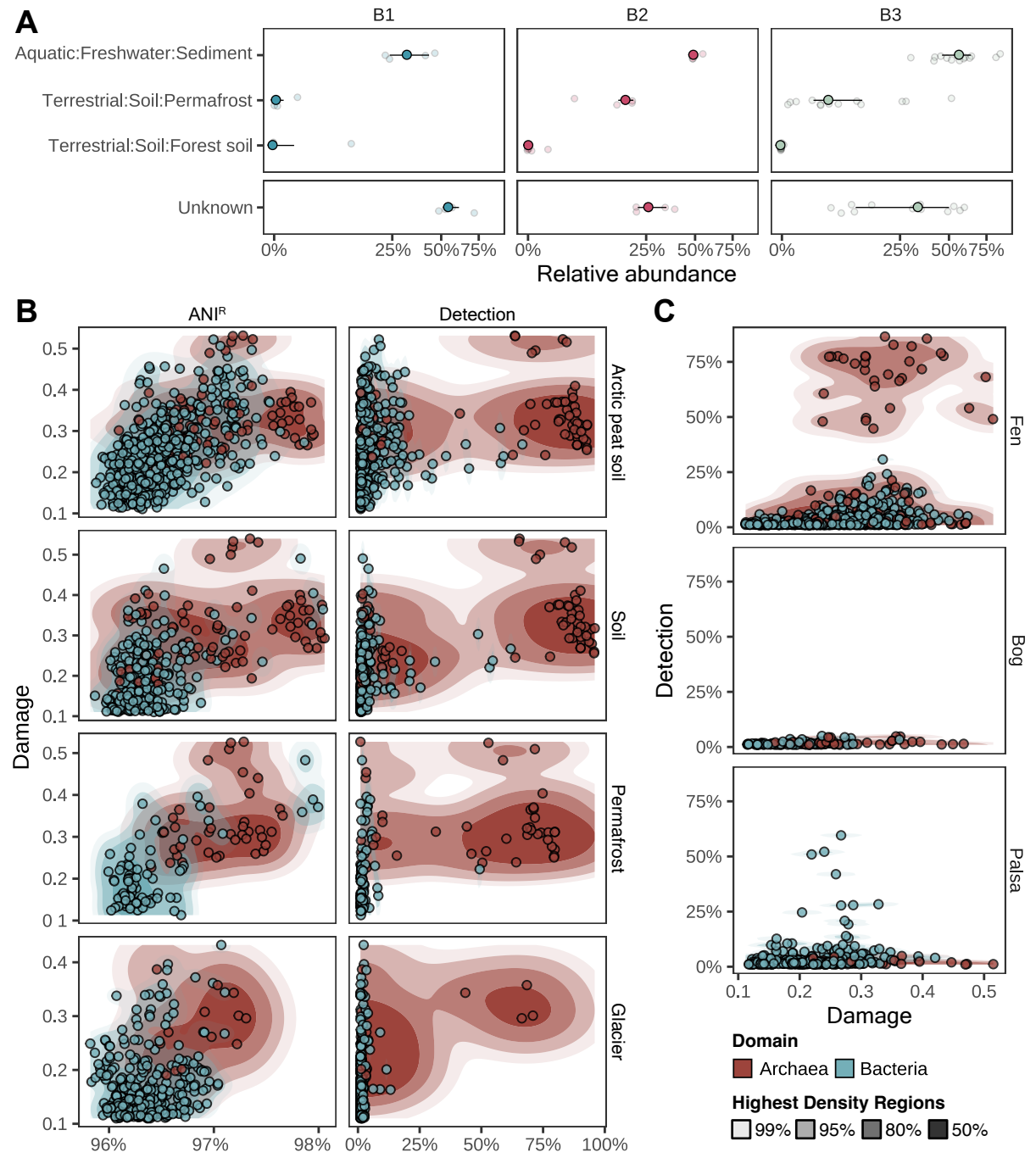
Extended Data Fig. 8 | Assessment of the taxonomic abundances. To determine the optimal parameter combination that accurately recovers the correct taxonomic abundances of the samples in this study, we examined several values (50, 100, 250, 500, 750, and 1000) for the Bowtie2 parameter -k, alongside filtering thresholds for ANI (90-94%) and breadth of coverage (0-10%) across 10 synthetic ancient metagenomes derived from the samples used in this study. For each parameter combination, we computed the Median Absolute Error. The lower the values, the better.

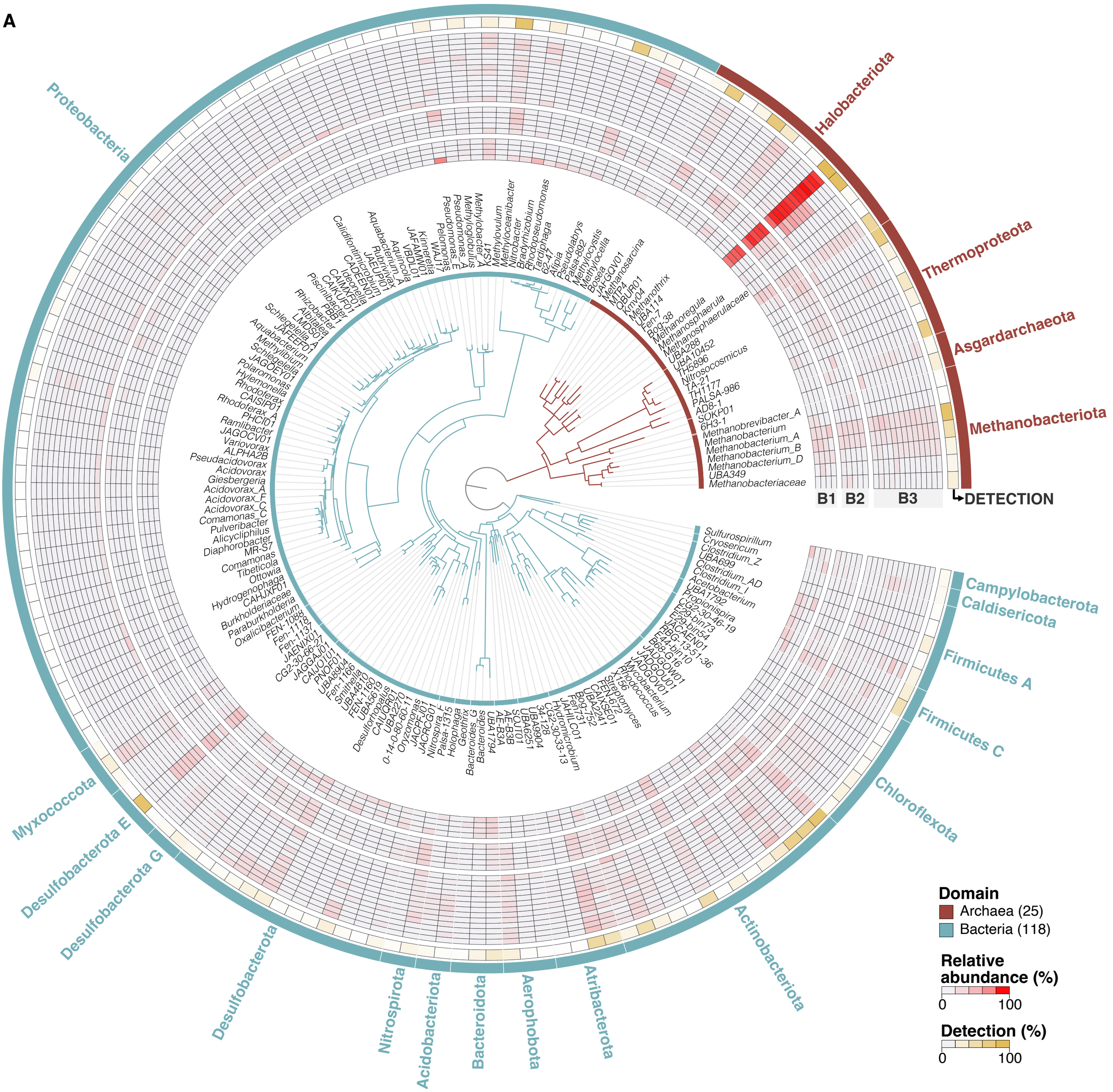


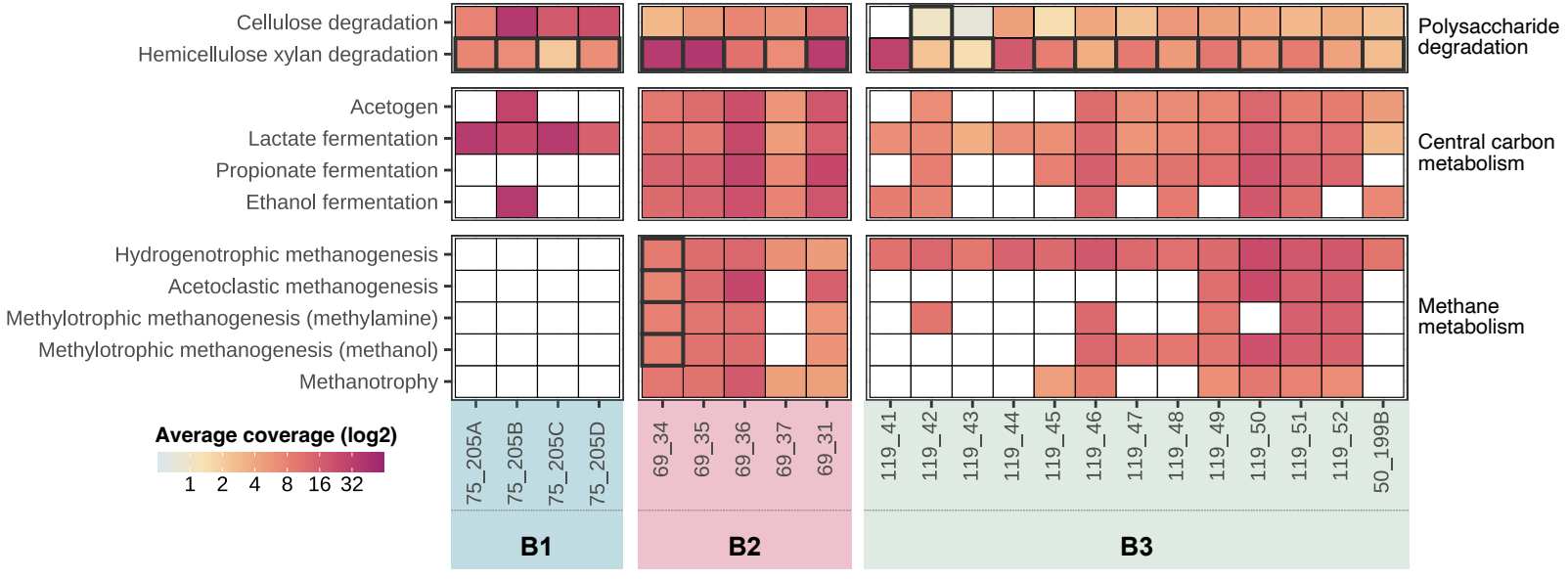
Extended Data Fig. 9 | Assessment of the damage estimates. To determine the optimal parameter combination that accurately recovers the correct taxa damage estimates of the samples in this study, we examined several values (50, 100, 250, 500, 750, and 1000) for the Bowtie2 parameter -k, alongside filtering thresholds for ANI (90-94%) and breadth of coverage (0-10%) across 10 synthetic ancient metagenomes derived from the samples used in this study. For each parameter combination, we computed the Median Absolute Error. The lower the values, the better.

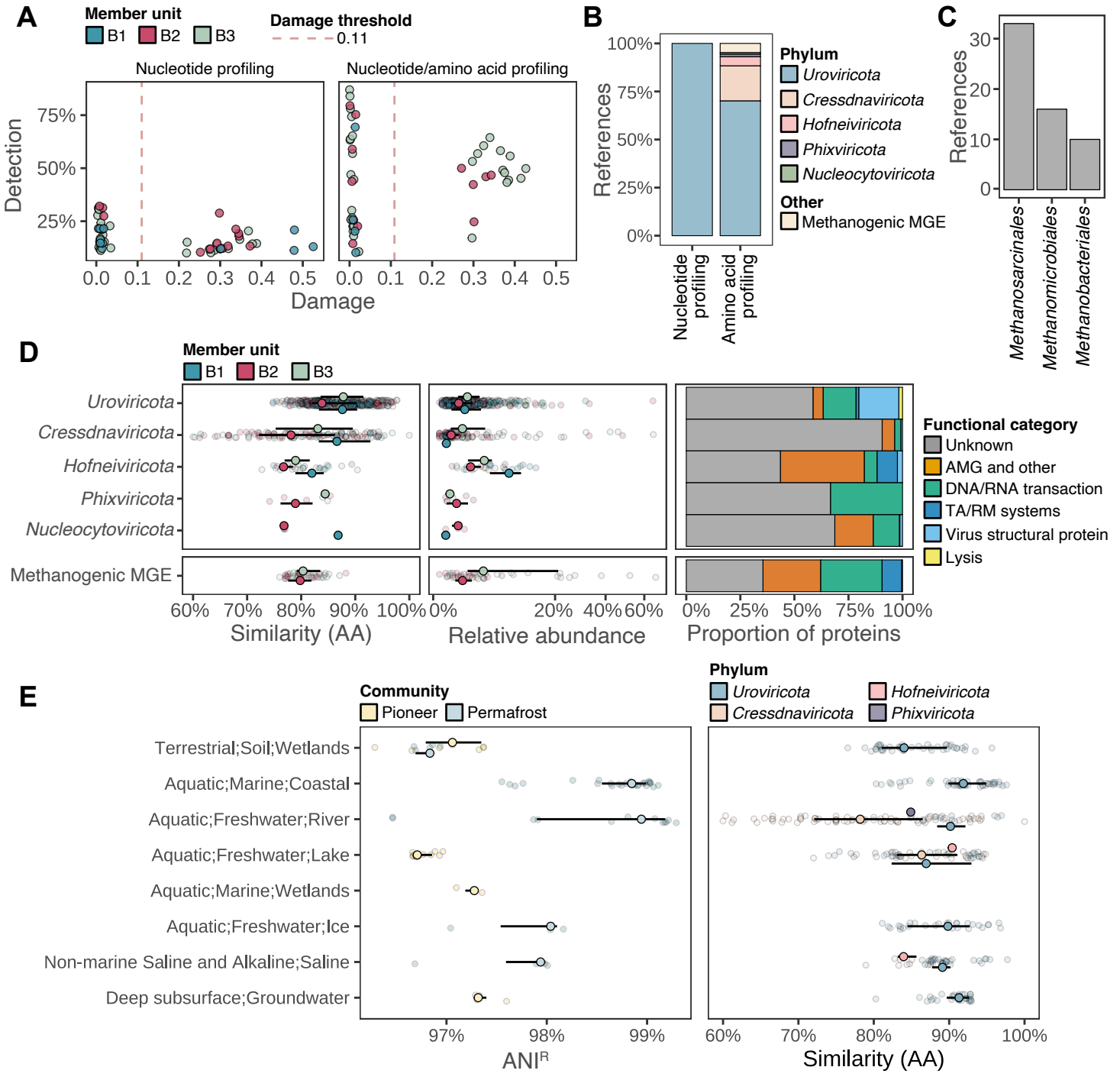

A
Greenland
Ice Sheet
B
0
50
Kilometers
83°N
82°N
119
75
69
50
KAP KØBENHAVN
FORMATION
PEARY LAND
Independence Fjord
32°W
28°W
24°W
20°W
C
DNA DAMAGE
PIONEER
COMMUNITIES
PERMAFROST COMMUNITIES
SUBAERIAL
EMERGENCE
0.4Myr
Locality
119
50
69
75
Elevation (m a.s.l.)
100
50
0
-50
-100
2.0
1.5
1.0
0.5
0.0
Time (Myr)

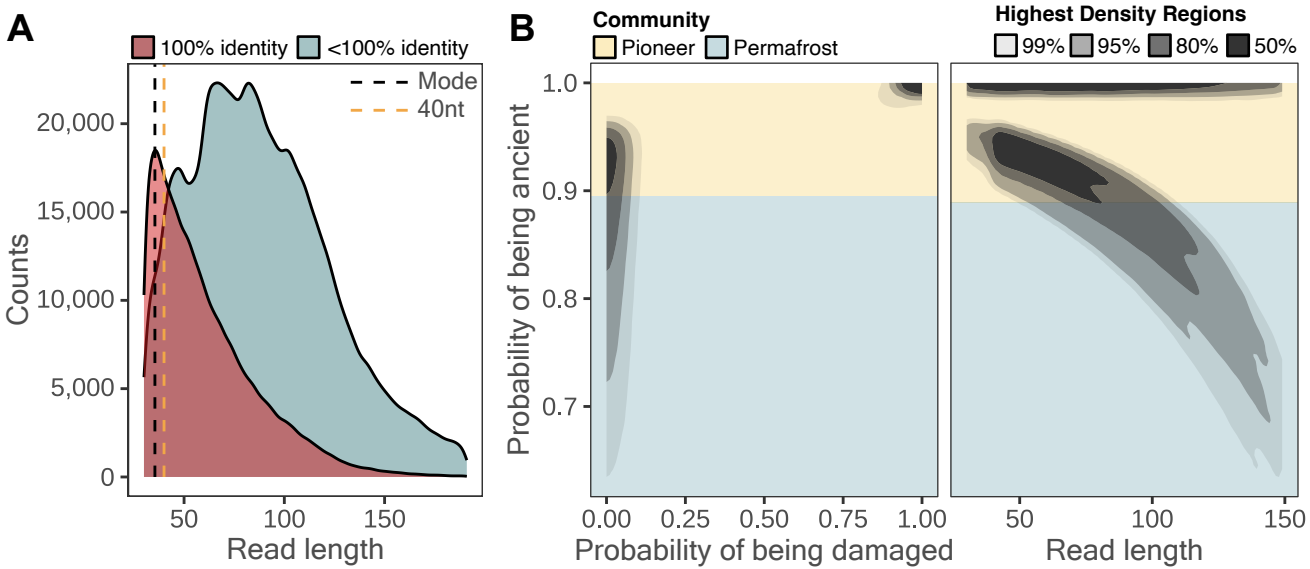

A
Damaged
Non-damaged
B3
119_52
119_51
119_50
119_49
119_48
119_47
119_46
119_45
119_44
119_43
119_42
119_41
50_199B
B2
69_31
69_37
69_36
69_35
69_34
B1
75_205D
75_205C
75_205B
75_205A
Domain
Archaea
Bacteria
Viruses
0%
25%
50%
75%
100%
Relative abundance
0M
100M
200M
reads
0%
5%
10%
15%
Reads mapped
B
Archaea
Bacteria
Viruses
ANI^R
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

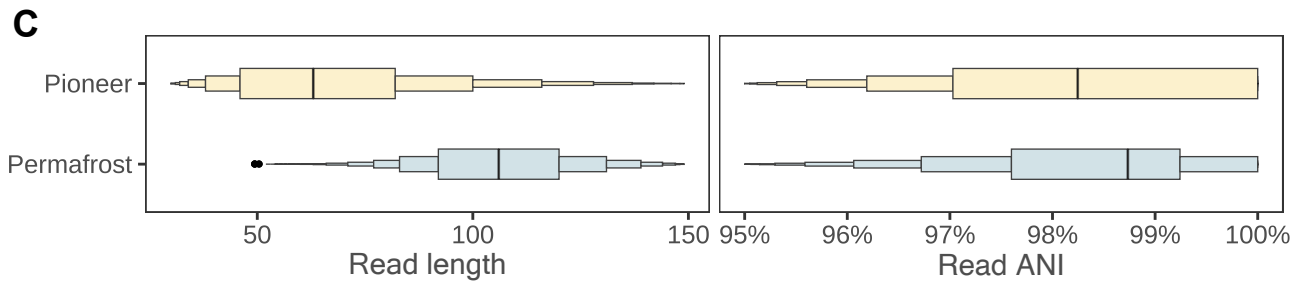

A
B1
B2
B3
Aquatic:Freshwater:Sediment
Terrestrial:Soil:Permafrost
Terrestrial:Soil:Forest soil
Unknown
0%
25%
50%
75%
Relative abundance
B
ANI^R
Detection
Arctic peat soil
Soil
Permafrost
Glacier
Damage
96%
97%
98%
0%
25%
50%
75%
100%
C
Fen
Bog
Palsa
Detection
Damage
Domain
Archaea
Bacteria
Highest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

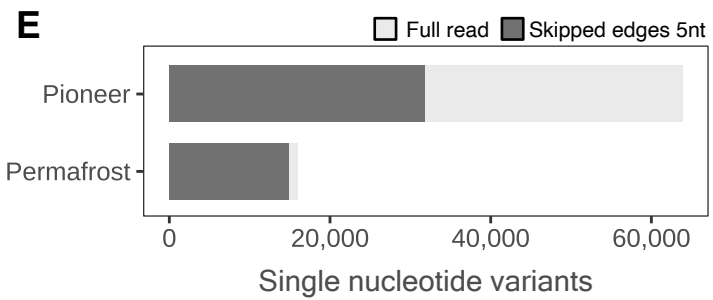

A
Proteobacteria
Halobacteriota
Thermoproteota
Asgardarchaeota
Methanobacteriota
B1
B2
B3
DETECTION
Campylobacterota
Caldisericota
Firmicutes A
Firmicutes C
Chloroflexota
Actinobacteriota
Atribacterota
Aerophobota
Bacteroidota
Acidobacteriota
Nitrospirota
Desulfobacterota
Desulfobacterota G
Desulfobacterota E
Myxococcota
Methylobacter_A
Methylovulum
Methyloceanibacter
Nitrobacter
Bradyrhizobium
Rhodopseudomonas
Tardiphaga
62-47
Afipia
Pseudolabrys
Palsa-892
Methylocystis
Methylocella
Bosea
JAFGQV01
Methanosarcina
MTP4
QBUR01
Kmv04
Methanothrix
UBA114
Fen-7
Bog-38
Methanoregula
Methanosphaerula
Methanosphaerulaceae
Methanosphaerula
UBA288
UBA10452
TH5896
Nitrosocosmicus
TA-21
TH1177
PALSA-986
AD8-1
SOKP01
6H3-1
Methanobrevibacter_A
Methanobacterium
Methanobacterium_A
Methanobacterium_B
Methanobacterium_D
UBA349
Methanobacteriaceae
Sulfurospirillum
Cryosericum
Clostridium_Z
UBA699
Clostridium_AD
Clostridium_I
Acetobacterium
UBA1792
Propionispira
CG2-30-46-19
TF29-bin73
TF29-bin54
JACAEN01
RBG-13-51-36
TF44-bin10
B68-G16
JADGOW01
JADGOU01
JADGOV01
Mycobacterium
Rhodococcus
AY156
Streptomyces
FEN-612
CAIXSE01
UBA2241
Bog-752
Fen-731
JAHILC01
Hydromicrobium
CG2-30-33-13
34-128
UBA9904
UBA6251
SOJT01
AE-B3B
AE-B3A
UBA1794
Bacteroides
Bacteroides_G
Geothrix
Holophaga
Palsa-1315
Nitrospira_F
JACRCG01
JACPCJ01
Oryzomonas
0-14-0-80-60-11
UBA2270
CAIUQR01
Desulfohopalus
UBA5619
FEN-1160
UBA4810
Smithella
Fen-1166
UBA8904
PNOF01
CAIJOT01
JAGGAJ01
CG2-30-66-27
JAENIX01
Fen-1137
Fen-1118
FEN-1088
Oxalicibacterium
Paraburkholderia
Burkholderiaceae
CAHJXF01
Hydrogenophaga
Ottowia
Tibeticola
Comamonas
MR-S7
Diaphorobacter
Alicycliphilus
Pulveribacter
Comamonas_C
Acidovorax_C
Acidovorax_F
Acidovorax_A
Giesbergeria
Acidovorax
Pseudacidovorax
ALPHA2B
Variovorax
JAGOCV01
Ramlibacter
PHCI01
Rhodoferax_A
CAISIP01
Rhodoferax
Hylemonella
Schlegelella
JAGOEY01
Polaromonas
Methylibium
Aquabacterium
Schlegelella_A
JAFEEF01
LMDS01
Albitalea
Rhizobacter
PBB1
Piscinibacter
CAIKUF01
CAIMXF01
Ideonella
CADEEN01
Calditrichotimicrobium
JAEUPI01
Rubrivivax
Aquabacterium_A
Aquincola
VBDL01
JAFAMW01
Kinneretia
WAJ17
Pelomonas
Pseudomonas_E
Pseudomonas_A
Methylogloburus
Methylobacter_A
S41
Methylovulum
Methyloceanibacter
Nitrobacter
Bradyrhizobium
Rhodopseudomonas
Tardiphaga
Domain
Archaea (25)
Bacteria (118)
Relative abundance (%)
0
100
Detection (%)
0
100
B
Soil
Methanotrophs
Various sources
B3
119_51
119_50
119_43
119_42
50_199B
B2
69_37
69_36
69_34
B1
75_205D
75_205C
0%
10%
20%
30%
40%
0%
3%
6%
9%
0%
20%
40%
60%
Proportion
Soil
adenosylhopane
adenosylhopane HG-Me
adenosylhopane HG-diMe
Me-adenosylhopane HG-diMe
diMe-adenosylhopane HG-diMe
Me-adenosylhopane
Me-adenosylhopane HG-Me
adenosylhopane type-2
adenosylhopane type-2 deMe
Methanotrophs
aminopentol
unsaturated aminopentol
aminotetrol
Various sources
aminotriol
BHhexol
BHpentol
unsaturated BHpentol
BHT
BHT-CE
anhydro BHT
unsaturated BHT
Me-BHT


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
100%
75%
50%
25%
0%
References
Nucleotide profiling
Amino acid profiling
Phylum
Uroviricota
Cressdnaviricota
Hofneiviricota
Phixviricota
Nucleocytoviricota
Other
Methanogenic MGE
C
30
20
10
0
References
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%
ANI^R
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%