Taxonomic analysis and hybrid assembly of environmental DNA from ancient middens in Greenland

Lisa Vader (s6192610)

Abstract
Environmental samples contain a wide variety of DNA from plants, animals, bacteria and fungi, which can be used to gain unique insights into current taxonomic diversity, but also to identify the origin of ancient remains. We sequenced 57 metagenomic samples from middens left behind by the Norse colonisers of Greenland and used taxonomic assignment to predict organisms that were important for their diet. We provide evidence for the exploitation of fish, shellfish, wheat and barley by the Norse population. We also identified the causative agent of plague *Yersinia pestis* from middens, which has not been found before in Greenland. Furthermore, we show that genome assembly of these environmental samples can be improved by using a combination of long- and short-read sequencing data.

Layman's summary
From the early eleventh century until the end of the fourteenth century, the south of Greenland was populated by Norse settlers. In the remnants of their civilisation many middens have been found, which are piles of domestic waste where food remains and faeces were discarded. We sequenced the DNA present in these middens to investigate the diet of the Norse Greenlanders at the time. Interestingly, we found large amounts of DNA from fish and shellfish, but also of the cereals barley and wheat. This suggests that both fishing and agriculture were part of the lifestyle of the Norse. We also examined the pathogens present at these sites and found DNA evidence of *Yersinia pestis*, the bacteria that causes the plague. Finally, we explored different methods to assemble the midden DNA into genomes. Different DNA sequencing methods result in either short or long stretches of DNA. By combining both of these read types, the quality of genome assembly could be considerably improved.

Part 1: Taxonomic analysis of environmental DNA from middens in Greenland

Introduction

A brief history of Greenland

Greenland was one of the last regions in the world to be populated by humans. The first colonisation event was in 2400 BC by the Paleo-Inuit, who originated from Siberia and had reached Greenland via the Americas. For a long time the Paleo-Inuit were the only inhabitants of Greenland, until they were replaced by the whale hunting Neo-Inuit of the Thule culture in 1100 AD, who had migrated from Siberia independently.[1,2] The first time Europeans ever set foot on Greenland was in 1000 AD, when the south was colonised by the Norse as part of a wider expedition into the North Atlantic. There were two Norse settlements, the Eastern settlement that covered the southern tip of Greenland, and the Western settlement that was located near the current capital Nuuk.[3] It has been estimated that at the peak of the Norse occupation, their population had a size of about 2000 people.[4] However, by 1400 AD the Norse settlements had been completely abandoned, either because the population had died or because they had left the island. Many theories have been put forward to explain the disappearance of the Norse in Greenland, including dwindling resources due to climate change[5], the inability to adjust to an arctic climate[6] and unfavourable economic changes[7].

The diet of the Greenland Norse

Gathering of sufficient food was paramount to survival in the harsh circumstances of the Arctic and must have been a major focus for the Norse. Therefore, an important method to get more insight into the lifestyle of the Greenland Norse, is by tracing back their diet. In the past, this has been researched using the few historical texts available, combined with excavations of farm sites and middens. Middens are piles of domestic waste, frequently containing animal bones, faeces, plant remains and shells. From written accounts we know that the Greenland Norse kept domestic animals that they had brought from their homeland, but that they also relied on hunting and fishing wild animals. Animal bone records have identified cattle, sheep and goat as the main domesticates, while pig, horse and dog are occasionally found. Regarding game animals, the most common finds are seal and reindeer, although walrus and polar bear have also been identified.[3] Interestingly, isotope studies of human bones have shown that over the course of their settlement, the Greenland Norse shifted from a predominantly terrestrial to a predominantly marine diet.[8,9] While it is certain the Greenlanders relied heavily on both domestic and game animals, the use of plants has been less widely studied. Findings of quern stones, cereal pollen and barley kernels have given rise to the belief that the Greenland Norse used to grow cereals and bake bread, although it remains uncertain to which extent they utilised this food source.[3,10] We also know that they cultivated flax, used as animal fodder or for producing linen, due to findings of flax pollen in midden deposits of the Western Settlement.[3,11,12]

Environmental DNA analysis

Due to high-throughput sequencing, it is now possible to identify organisms by metagenomic analysis of environmental DNA, as an alternative to more traditional archaeological methods. While excavation techniques by their nature disrupt the sampling environment, the collection of DNA samples is generally much less invasive. In addition, this method is able to gather information from organic material other than bone, such as skin, hair, faeces and plant remains and could thereby provide a more complete picture of local species distribution.[13,14] Studies of ancient environmental DNA have already had a major impact on our understanding of evolutionary history. For example, metagenomic sequencing of ice cores has provided the first evidence for a conifer forest in southern Greenland that existed more than 450 000 years ago.[15]

Previous studies

Two studies so far have used environmental DNA to investigate the remains of Norse civilisation in Greenland, both of them regarding the Western Settlement. The first study was done in 2009 by Arneborg et al. and analysed a soil core extracted from an open field of one of the farms called 'The farm under the sand' (Figure 1).[16] Mitochondrial DNA of mammals was measured using amplicon sequencing and compared over time by carbon dating of the soil layers. Only livestock animals were found (cattle, goat and sheep), except for reindeer which was found in one sample. No livestock animals were found in layers younger than 1450 AD, consistent with the abandonment of the Western Settlement in the late fourteenth century. The absence of wild animals makes sense, as the soil was sampled specifically from a grazing field and not from middens which contain more general food waste.

A more recent study by Seersholm et al. (2016) used shotgun sequencing to identify mammals in a midden located at the Sandnes farm (Figure 1), as well as three middens left behind by Inuit.[14] While the Sandnes midden was dominated by DNA from cattle, goat and sheep, the Inuit middens contained exclusively wild animals such as seal, whale, reindeer and wolf. In smaller quantities, harp seal, reindeer and walrus were also found in the Sandnes midden. However, the relative abundance of wild animals found in this study (10-25 %) was much lower than estimated from the Sandnes bone record, in which seal and walrus take up around 70% of the most abundant mammals.[17]

Study design

We collected samples from three different midden sites of the Western Settlement (Figure 1). Shotgun sequencing was used to determine taxonomic abundance at the different sites. In contrast to the previous studies, we incorporate groups that are of interest in the dietary context besides mammals; namely fish, shellfish and cereals. Furthermore, we compare samples across soil depths ranging from 5 to 72 cm, in order to find temporal patterns.

Identification of pathogens

As a separate objective, we examine the human pathogen content across our samples. Recently, thawing of permafrost due to climate change has raised concerns about the possible (re-)emergence of pathogenic microorganisms.[18–20] This includes both ancient pathogens that have been inactive for a long time, but also extant pathogens that might gain new traits due to genetic exchange with the ancient microbial community. The danger of these emerging pathogens lies in the fact that they are unfamiliar to the human immune system. Especially in combination with the acquisition of antimicrobial resistance, they could form a serious health threat. Cryospheric environments have been found to contain a large variety of antimicrobial resistance genes.[19]

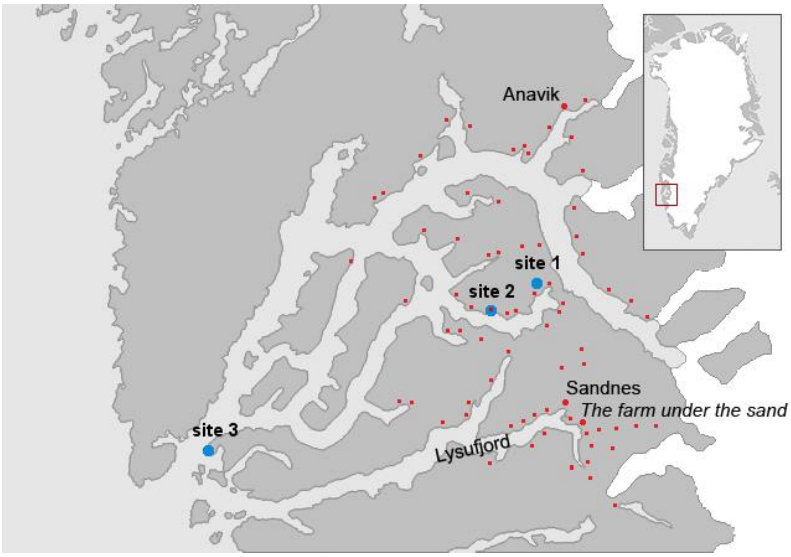

Figure 1: Map of the Western Settlement of the Greenland Norse. Known locations of farms are marked in red. The sampling sites used in this study are marked in blue. Adapted from *Western Settlement*, by Masae, 2007. (https://commons.wikimedia.org/wiki/File:Western-settlement-eng.png) CC BY-SA

The soil of the former Western Settlement is characterised by discontinuous permafrost[21], which means that some regions thaw completely during the summer, while others remain frozen year-round. Here, we report which extant pathogen species are most abundant in this constantly freezing and thawing environment. These pathogens could potentially be involved in the emergence of new diseases from thawing permafrost.

Methods

All scripts used can be found at https://github.com/lisavader/Greenland_Middens. R version 4.1.0[22] was used for all R scripts.

Sample collection

Soil and midden cores were obtained from different sampling sites at four locations in Greenland (Nuuk, Narsarsuaq, Sermermiut and Qajaa) using a cylindrical soil probe. Triplicate samples were extracted from the soil cores at various depths, ranging from 0 – 72 cm. To avoid cross-contamination, only the centre of the soil core was used for sampling. Samples were stored in nucleic acid preserving buffer (e.g. RNA*later*) and kept frozen at -80 °C until extraction. Only the samples that were collected in the area of Nuuk (n = 57) were used in this analysis.

DNA extraction and sequencing

DNA was extracted using the DNeasy PowerSoil Pro Kit (Qiagen) according to the manufacturer's instructions. Libraries were prepared using a PCR-free protocol and sequenced using the Illumina NovaSeq S4 platform in 150bp paired end mode. Reads were trimmed with BBDuk2 v 36.49[23], using a kmer size of 19 (mink=11), a minimum phred score of 20 and a minimum read size of 50 bp.

Taxonomic assignment

Reads were mapped against the NCBI Refseq database, accessed on 08/02/2022 (https://www.ncbi.nlm.nih.gov/refseq/), and the Nordic vascular plant database PhyloNorway (https://www.phylonorway.no/), using Bowtie2 v 2.4.2[24] in local alignment mode. Perl and python scripts were used for filtering of the hits. Hits were discarded when the forward and reverse reads did not map to the same reference sequence. Only the best reference hit per read pair was kept, based on local alignment score (in case of a tie the first hit was kept). A minimum read coverage of 0.8, minimum identity of 0.9 and a total alignment length of at least 70 bp were required for both the forward and reverse reads. These thresholds were chosen to maximise the true positive/false positive ratio. True/false positives were detected by comparing vascular plant hits to a list of plants previously found in Greenland (Figure S1). After filtering, taxon information was gathered by converting reference ids to NCBI taxids, and hits were counted per taxon. On average, 1.1 million reads per sample could be mapped, which is 1.3% of the total reads. The percentage of reads mapped was positively correlated with bacterial content (Figure S2).

Data analysis

Subsequent data analysis was performed in R. A principal component analysis was done based on vertebrate composition, with as input a matrix of counts per family. Families with < 20 counts in all samples were excluded from the analysis. Primate families were excluded, as it is difficult to distinguish human DNA from contamination. The river dolphin family *Lipotidae* was also excluded, due to unreliable results. Zeroes were imputed using Bayesian multiplicative replacement and count data was transformed onto a Euclidian space by centred log ratio transformation (clr). Furthermore, the relative distribution of food related taxa, and of the ten most common bacterial and fungal pathogens were visualised, based on counts normalised for sequencing depth.

Results

Organisms present in each of the samples were predicted by taxonomic assignment of environmental DNA. To get a general overview of the similarity between samples, their vertebrate content was compared using principal component analysis (Figure 2). Vertebrates represent most of the species we are interested in, and the vertebrate DNA we find is expected to be ancient, unlike the DNA of other taxonomic groups such as bacteria, fungi and nematodes. The PCA plot shows that vertebrate distribution is related to sampling site, but not to sampling depth. Even though they are closely together in distance, site 1 and site 2 show most dissimilarity in the PCA. This is due to lower vertebrate richness in site 2 (Figure S3).

Next, we examined taxonomic groups that are of relevance as a food source or during farming or hunting. Relevant organisms found in our samples include domesticated mammals (pig, sheep, cattle, horse, dog/wolf and cat), wild mammals (reindeer, bear, seal and sperm whale), cereals (wheat, barley, rye and oat), fish and shellfish. In total, these taxonomic groups represent 0.094%, 0.085%, and 0.093 % of all organisms found at site 1, site 2 and site 3 respectively. Cereals are by far the most abundant of all the food groups, representing 45-65% of the reads (Figure 2a). Fish are the second most abundant group (15-30%), followed by shellfish (10-15%). While mammals are comparatively rare at sites 2 and 3 (4-8%), they have a much larger presence at site 1 (20%).

We also examined the abundance of these food groups across different sampling depths (Figure 2b). We expect the layers where we find most of the mammal remains to correspond to midden deposits. Mammal DNA is most abundant in the deeper layers, which is consistent across sampling sites. Conversely, cereal remains are mostly found at shallow depth. This could indicate that cereals mainly grew after farms were abandoned by the Norse. Fish and shellfish appear at all depths, although they seem to correlate somewhat with mammal remains. However, it is hard to interpret patterns across depth without knowing the age of the sediment layers.

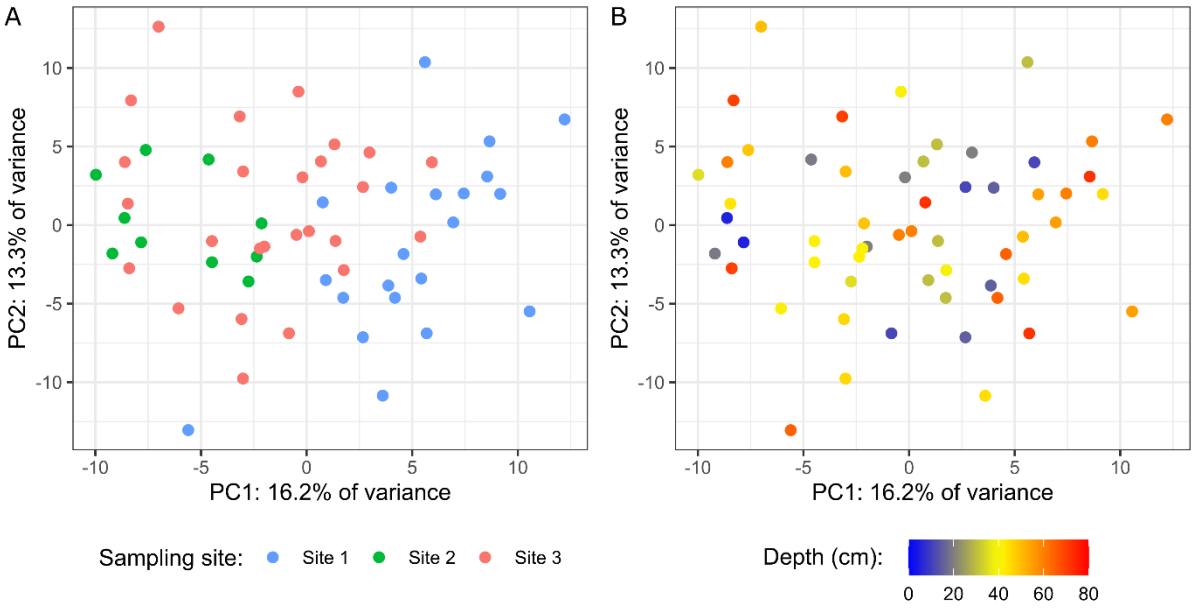

Figure 2: Principal component analysis of the vertebrate family distribution per sample (n=57), based on clr values. Coloured by sampling site (a) and sampling depth (b). Vertebrate families with a count < 20 in all samples were excluded from the analysis.

It is interesting that fish are present in substantial amounts at each site, because even though it is known from historical references that fishing was part of the Norse culture in Greenland, findings of fish bones at archaeological sites have been rare.[3] Many different species of fish are present, but the most common one found is whitefish (*Coregonus* sp., 39%). Findings of shells are usually not documented, although they have previously been found at middens of the Sandnes farm, predominantly consisting of blue mussels (*Mytilus edulis*).[17] However, most of our hits are to either scallops (*Pecten maximus*, 86%) or oysters (*Crassostrea gigas*, 8.6%). It is hard to distinguish whether these represent the true species found in Greenland, or are false hits due to an incomplete database. Either way, it seems likely that shellfish were consumed by the Greenland Norse. It is for instance also known that shells were eaten in Iceland as a dietary supplement or famine food.[25]

The large majority of mammals found in our samples are domestic animals (75-85%, Figure 2c). This pattern is similar to what was found by Seersholm et al. at the Sandnes farm. The most abundant mammal varies by site; over half of the reads at site 1 are from pig, whilst sheep and cat are most abundant in sites 2 and 3 respectively. The amount of reads that map to cattle is relatively low compared to the Sandnes study, where this was the dominant mammal. What is also remarkable is that we don't find any goat remains in our samples. An explanation for this could be that goat reads were incorrectly assigned to sheep, as the two species are closely related genetically. Pigs are characteristic of the early settlement period[3], so seeing them at all sites is indicative that the middens are relatively old. Furthermore, cat bones have rarely been found in the Norse settlements[17], so it is interesting that they are so abundant at site 3. The wild animals found consist almost completely of reindeer. Marine mammals (sperm whale and seal) were only found in very small amounts. Walrus was also present, but not shown due to very low read count in only one sample. The low amount of seal DNA found is remarkable, because seal is known to have been an important part of the Norse diet.

The cereals we find are predominantly wheat and barley, although small amounts of rye and oat are also present (Figure 2d). Flaxseed was present in a few samples, but with very low read count. Wheat and barley are not native to Greenland, so they have likely been introduced by the Norse population. That we find their remains in large amounts is therefore evidence that they were cultivated by the Norse.

We also investigated the bacterial and fungal pathogens present in the midden samples. We searched for soil pathogens that can cause disease in humans by direct interaction with the soil, as reviewed by Steffan et al.[26] These pathogens represent 0.29%, 1.5% and 0.067% of the total amount of bacteria and fungi at site 1, site 2 and site 3 respectively. The ten most common pathogens are shown in Figure 3. Of these, the gut bacterium *Escherichia coli* is by far the most abundant, and it is responsible for the high pathogen content at site 2. *E. coli* has been found in soil at many places in the world, and is known to survive well at low temperatures.[27] Other highly abundant species are the bacteria *Salmonella enterica* and *Legionella* sp. and the fungus *Blastomyces dermatitidis*. The pathogens found aren't correlated to a certain depth, nor to the layers that contain the most mammals (Figure S4). A large number of reads at site 2 were identified as *Yersinia pestis*, the causative agent of the plague. Plague is considered a threatening illness as it is highly contagious, and it is thought that soil plays an important role as a reservoir for the disease.[28,29]

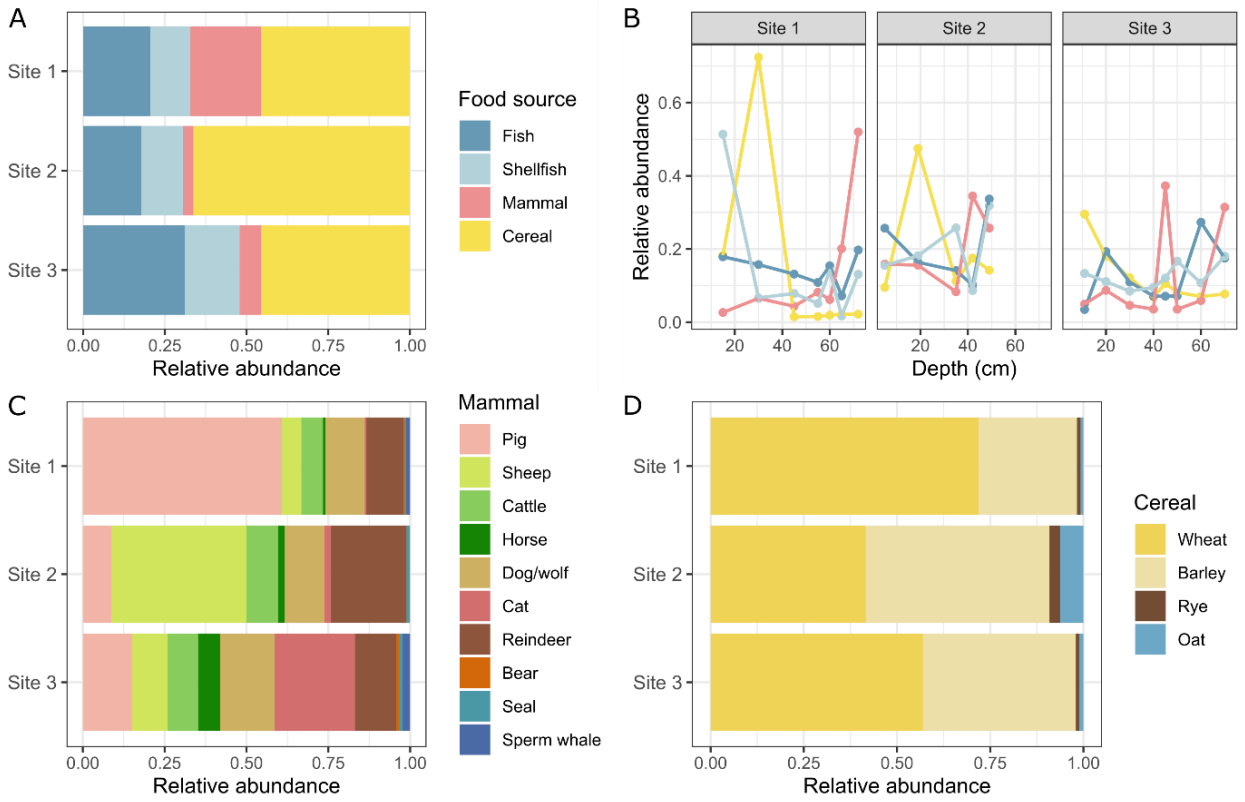

Figure 3: Relative abundance of taxa that are relevant as a food source. (a) Abundance of the four food groups fish, shellfish, mammals and cereals by site. (b) Abundance of the four food groups fish, shellfish, mammals and cereals by depth (same colours as in figure a). (c) Abundance of mammals by site. (d) Abundance of cereals by site. Counts were normalised for sequencing depth. Corresponding taxon per group: Fish = Actinopteri, Shellfish = Bivalvia, Pig = *Sus scrofa*, Sheep = *Ovis*, Cattle = *Bos*, Horse = *Equus*, Reindeer = *Cervidae*, Dog/wolf = *Canis lupus*, Cat = *Felis catus*, Bear = *Ursus*, Seal = *Phocidae*, Sperm whale = *Physeter catodon*, Wheat = *Triticum aestivum*, Barley = *Hordeum vulgare*, Rye = *Secale cereale*, Oat = *Avena sativa*. The mammal and cereal groups consist of the taxa shown in (c) and (d) respectively.

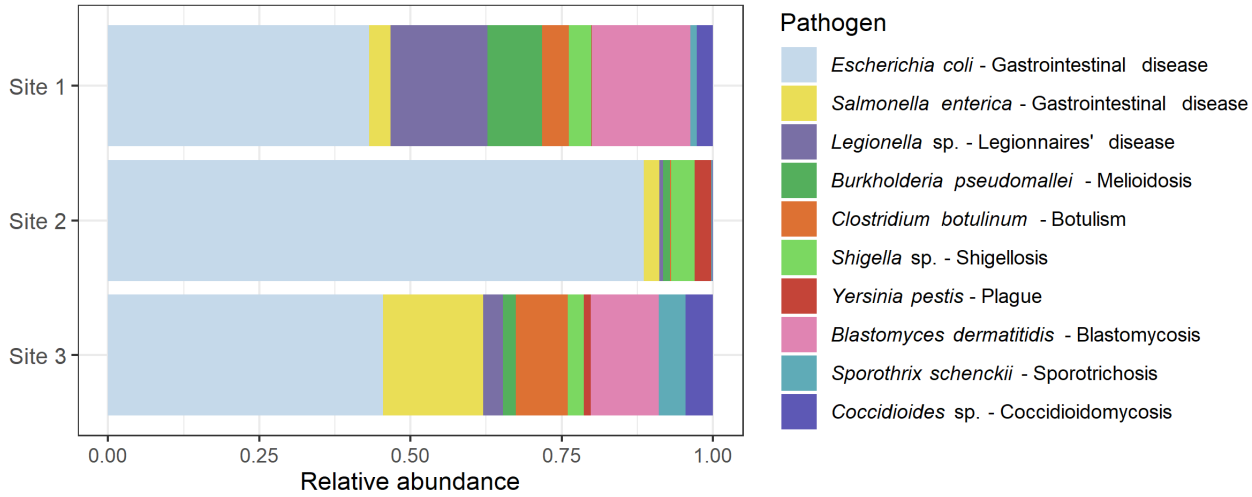

Figure 4: Relative abundance of the ten most common bacterial and fungal pathogens, per sampling site. Counts were normalised for sequencing depth. Pathogens are annotated with their associated diseases, based on Steffan et al.[26].

Discussion

We sequenced a total of 57 environmental samples from middens of the former Western Settlement in Greenland and examined their taxonomic profiles. First of all, we investigated organisms that are relevant in the context of the Norse diet. The main mammal species found in the middens are pig and sheep, which are commonly described domesticates kept by the Norse. Besides mammals, we also explored other taxonomic groups as possible food sources. We found large amounts of fish and shellfish in the midden samples, which indicates that these were part of the diet of the Greenland Norse. Furthermore, we also found substantial presence of wheat and barley, species that are not native to Greenland. Along with previous findings of these cereals at farm sites, this strongly suggests that they were introduced to Greenland by the Norse.

The mammals we find are mainly domesticates, which contrasts with what is known from bone records of the Western Settlement, in which wild animals such as seal and reindeer are much more abundant. A similar bias for domestic animals was detected in the study by Seersholm et al.[14] It could be that there is more material present in middens that originate from domestic animals, such as urine, faeces and hair. Marine mammals such as seal and walrus are especially lacking in our data. A possible explanation is that the skinning of these animals took place directly after catching and not at the farm, causing less material to end up in middens. Walrus was hunted for their valuable hide and tusk, which were exchanged for other goods and therefore could be missing from midden samples.[5]

Furthermore, we don't find any apparent differences in vertebrate composition between the different soil depths sampled. A possible explanation could be DNA leaching, i.e. movement of DNA through soil layers over time.[30] However, DNA leaching seems to be less common at low temperatures and has not been found to affect permafrost at all.[31] Even though sampling was done carefully, cross-sample contamination during this stage could also cause samples from different layers to appear more similar to each other. A possible way to check for this kind of contamination is by using trace substances which are smeared on exposed surfaces and equipment during sampling[34], which could be a useful implementation for future sampling.

Although they don't correspond to a specific depth, we generally find mammals to appear in peaks at the deeper layers. The same depth at different sites does not necessarily correspond to the same geological age, so these samples might very well represent the same time period. In order to properly compare samples across sites, we will need to estimate their age by carbon dating. Knowing the age of the layers is also very important for the interpretation of the origin of the mammals, fish and cereals we found.

Furthermore, we also examined the human pathogens present in the midden samples. The main pathogen found is *E. coli*, a species that is not dangerous in and of itself, but contains various highly pathogenic subtypes. Whether any pathogenic *E. coli* is present in the midden soil should be determined by a sequence type specific analysis. The relative amount of *E. coli* is likely somewhat overestimated compared to other bacteria, due to the higher abundance of *E. coli* in the Refseq database. Strikingly, we also found DNA evidence of the causative agent of plague, *Yersinia pestis*, to which a total of 1481 reads aligned across our samples. To our knowledge, *Y. pestis* has not been found before in Greenland soil. Although the Black Death reached all the way into Iceland in the early and late fifteenth century[32], the disease has not been documented in Greenland. However, the human flea which is a host to *Y. pestis* has been found during excavations of the Western Settlement.[33] It would be interesting to investigate the age of the sediments where *Y. pestis* has been found, in order to approximate when this pathogen arrived in Greenland.

We performed rigorous filtering in order to discard false alignments as much as possible. However, the results should be validated, either by examining read coverage over the reference sequences, or by genome assembly and alignment of contigs. Furthermore, there are a few possible improvements for the methodology. A common problem during taxonomic assignment is that highly conserved sequences align equally well to multiple species, but can only be assigned to one taxon. In our current pipeline, reads like this will be randomly assigned to a species. An alternative approach would be to evaluate the lowest common ancestor of all the alignments per read. This can be done for each read by saving all alignments that have the best alignment score, converting the reference identifier to NCBI taxid and moving up the taxonomic hierarchy until the taxids of all hits become the same. Although species level assignment will not be available for each read, this method increases confidence in the validity of the alignments. Another issue is that species that have a larger genome, can be overrepresented in the result just because there is more DNA available per organism. Because genome size is not always exactly known, it can be difficult to adjust for this bias. A feasible way to deal with this in the case of eukaryotes is to use a mitochondrial database instead of a whole genome database. However, this also leads to a loss of sequencing information, which can be a disadvantage.

Overall, we show that metagenomic sequencing of environmental DNA can give valuable insight into the taxonomic diversity of middens that is not found by traditional archaeological methods. We provide evidence for the exploitation of fish, shellfish, wheat and barley by the Norse population in Greenland. We also found presence of the pathogen *Yersina pestis*, which has not been identified before in Greenland.

Part 2 – Hybrid assembly for metagenomics

Introduction

In order to detect genetic changes in microorganisms or to predict previously unknown species, high quality genome reconstructions are indispensable. However, *de novo* assembly from metagenomic data is notoriously difficult, due to the many intra- and intergenomic repeats present, which often can't be fully covered by reads.[35] Still, there exist many useful bioinformatic programs that aim to make assemblies as complete as possible. For assembly of genomes as well as metagenomes, the computational strategy depends on the sequencing data used as input. We will briefly review these strategies.

Short-read assembly
Most genome sequencing nowadays is performed using Illumina technology, which yields reads that have a length of 100 – 300 bp and a median error rate of 0.1 – 0.6% depending on the specific platform used.[36] A common and effective method for assembling short reads is by building a de Bruijn graph. Briefly, from the reads, substrings of a fixed length k (k-mers) are extracted, which represent the nodes of the de Bruijn graph. Nodes that have an overlap of k-1 are then connected to each other by edges, creating a graph. Assemblies are found by computing a Eulerian path through the graph; a path that visits each edge exactly once. High sequencing accuracy is important, because single base sequencing errors introduce false k-mers which can lead to misassembly.[37,38] Assemblers employing de Bruijn graph based methods include SPAdes[39], MEGAHIT[40] and IDBA-UD[41].

Long-read assembly
The main limiting factor of short-read assembly is that repeated regions larger than the (paired-end) read size cannot be resolved. Nanopore technology is able to produce much longer reads of up to 2 million bp[42], although at the cost of a much higher error rate of 6-8%[43]. Long reads are assembled using the overlap-layout-consensus approach. The pairwise overlap between the reads is computed by a dynamic programming algorithm, permitting a certain number of mismatches and gaps. An overlap graph is constructed where nodes correspond to the reads, and edges are drawn between overlapping nodes. During the layout stage, this graph is simplified as much as possible, resulting in contig predictions. Finally, the consensus sequence per contig is found by multiple sequence alignment of the original reads to the predicted contig.[38] Popular assemblers using overlap-layout-consensus are Flye[44], Canu[45] and miniasm[46].

Hybrid assembly
The advantages of short- and long-read data can be combined to create an assembly with both high continuity and little errors. Assembly methods that utilise both types of sequencing data are called hybrid assembly methods. As described in the previous two paragraphs, short- and long-read data require different assembly strategies. Therefore, simply using a combination of both read types as input to a (short- or long-read) assembler does not work. Depending on which type of sequencing data is most abundant, either a short-read first approach or a long-read first approach can be chosen. The short-read first approach means that gaps in a short-read assembly are filled by long reads, so that longer contigs can be produced. Many short-read assemblers include a hybrid option where long reads are used for scaffolding. In contrast, the long-read first approach is based on a long-read assembly where errors are corrected by short reads. This polishing step is usually performed by a secondary program such as POLCA[47], Polypolish[48] or NextPolish[49].

Hybrid assembly is considered a promising approach for completing genomes, but has not been reviewed much in the context of metagenomics. Here, we assemble 18 environmental samples from

Greenland using short-read (SPAdes), long-read (Flye) and hybrid (hybridSPAdes) methods, and compare the results. We also show how the quality of hybrid assembly can be improved in the case of limited long-read data.

Methods

All scripts used can be found at https://github.com/lisavader/Greenland_Middens. R version 4.1.0[22] was used for all R scripts.

DNA extraction and sequencing

Sampling was carried out as described in part 1 of this report. A subset of 18 samples was sequenced using both Illumina and Nanopore technology. This includes 13 samples from Nuuk, 3 samples from Sermermiut and 2 samples from Narsarsuaq. Separate whole genome DNA extractions were carried out prior to short- and long-read sequencing, using the DNeasy PowerSoil Pro Kit (Qiagen) and the Quick-DNA HMW MagBead Kit (Zymo) respectively. Short-read libraries were prepared using a PCR-free protocol and sequenced with Illumina NovaSeq S4 in 150bp paired end mode. Long-read sequences were generated using Oxford Nanopore GridION and R10.3 flow cells, with six samples per flow cell. The output was basecalled using Guppy v 5.0.16 in super accurate mode.

Reads pre-processing and taxonomic alignment

Illumina reads were trimmed with BBDuk2 v 36.49[23], using a kmer size of 19 (mink=11), a minimum phred score of 20 and a minimum read size of 50 bp. Both read sets were aligned to the SILVA ribosomal RNA database[50], accessed on 16/01/2020 (https://www.arb-silva.de/) using KMA v 1.2.10.[51]

Genome assembly

Short-read assembly was performed by SPAdes v 3.13.0[52] in metagenomic mode using custom k-mers 27,47,67,87,107,127. Long-read assembly was done using Flye v 2.9[53] in metagenomic mode. Short-read first hybrid assembly was achieved by providing either long reads, or contigs assembled from long reads by Flye, as extra input to SPAdes (hybridSPAdes).[54] Flye contigs missing from the hybrid assembly were detected using CD-HIT v 4.8.1[55] with a sequence identity threshold of 0.95.

Assembly quality assessment

The quality of the assemblies was assessed using QUAST v 5.0.2[56] with a contig length cut-off of 1000 bp. An R script was used to compare quality metrics across different assembly methods.

Results

It should be noted that there was substantially more Illumina sequencing data available than Nanopore sequencing data. The average amount of Illumina reads per sample was 96 million, corresponding to 14 gigabasepair (Gbp). For Nanopore, this was 850 000 reads per sample, corresponding to 2.1 Gbp.

Non-hybrid assemblies were obtained using SPAdes and Flye for short and long reads respectively. On average, Flye produced assemblies with a total length of 65 Mbp and an N50 of 87 000 bp. For SPAdes the average assembly size was 330 Mbp, while the average N50 was 2300 bp (Figure 5a and 5b). Because there was far more short-read data than long-read data available, we used a short-read first approach for hybrid assembly, with SPAdes as the main assembler. SPAdes provides two options for hybrid assembly, using either raw long reads (--nanopore flag) or long reads that have been pre-assembled into contigs (--trusted-contigs flag). Both types of data are used to close gaps and resolve

repeated regions in the assembly graph. In the case of the trusted-contigs option, contigs are also used for graph construction.

Hybrid assemblies were produced with SPAdes, using both read and contig based options. In the latter case, the contigs were used that had previously been assembled by Flye. Both hybrid methods produced larger assemblies (reads: 370 Mbp, contigs: 390 Mbp) and better N50 values (reads: 2800 bp, contigs: 3200 bp) than the short-read only assembly (Figure 5a and 5b). Nevertheless, these assemblies still consisted mainly of very small contigs, as can be seen from the N50 values. Due to the lack of long-read data, this is partly unavoidable. However, when we look at the largest contigs produced per assembly (Figure 5c), it becomes clear that not all contigs present in the Flye assembly were used by SPAdes for constructing the hybrid assemblies. Apparently, there are certain differences between the Illumina and Nanopore sequences, which cause at least some long reads to find no overlap with the short-read contigs and to be discarded from the assembly. These differences can be explained by the fact that short and long reads originate from different DNA extraction methods, and therefore different organisms are represented by both read sets. Indeed, read alignment to the SILVA taxonomic database revealed organisms present in the long reads that were not present in the short reads. (Table S1).

We aligned the Flye contigs back to the hybridSPAdes assemblies that used these contigs as input, in order to find contigs that had not been incorporated into the assemblies. These contigs were added manually to the hybrid assemblies, so that all sequencing information would be preserved. This resulted in an average assembly size of 460 Mbp, N50 of 5000 bp and a largest contig of 2.6 Mbp, which is the best result out of all hybrid assembly methods tested (Figure 5). It should be noted that these values vary greatly across samples, especially for the largest contig. To exemplify this, the lower bounds are 120 Mbp for assembly size, 1500 bp for N50 and 0.057 Mbp for the largest contig, while the upper bounds are 1200 Mbp for assembly size, 16000 bp for N50 and 6.6 Mbp for the largest contig. Interestingly, these quality metrics are positively correlated with the amount of long-read data, but not with the amount of short-read data (Figure S5). This indicates that for our dataset, the amount of long-read data is a large bottleneck.

Discussion

Here, we show that using a hybrid approach greatly increases the quality of metagenomic assembly, even when long-read data is limited. We are dealing with long-read data that is both more sparse than and not completely overlapping with the available short-read data. We propose a method where long reads assembled with Flye are used for scaffolding during hybridSPAdes assembly. Long-read contigs that cannot be incorporated in this way are manually added to the assembly. This approach yields improved results compared to short-read only assembly with SPAdes, showing an increase in N50 of more than 200%.

There are a few possibilities to improve the quality of assemblies even more. Sequencing errors in the long-read contigs can be removed by incorporating a polishing step after Flye assembly, which uses short reads for correction. This should have a positive impact on the overall accuracy of the assemblies. Considering the possibility that some long-read contigs are rejected from hybrid assembly because of sequencing errors, this might also have an impact on the contingency of the assemblies. However, contingency can only be substantially improved with the addition of more long-read data. The most obvious way to do this is by using less samples on the same flow cell during sequencing – one Nanopore flow cell can yield 15 – 150 Gbp of data depending on the type.[57] But this of course also increases the sequencing cost per sample. An important consideration is therefore whether to sequence many samples with low depth, or fewer samples with high depth,

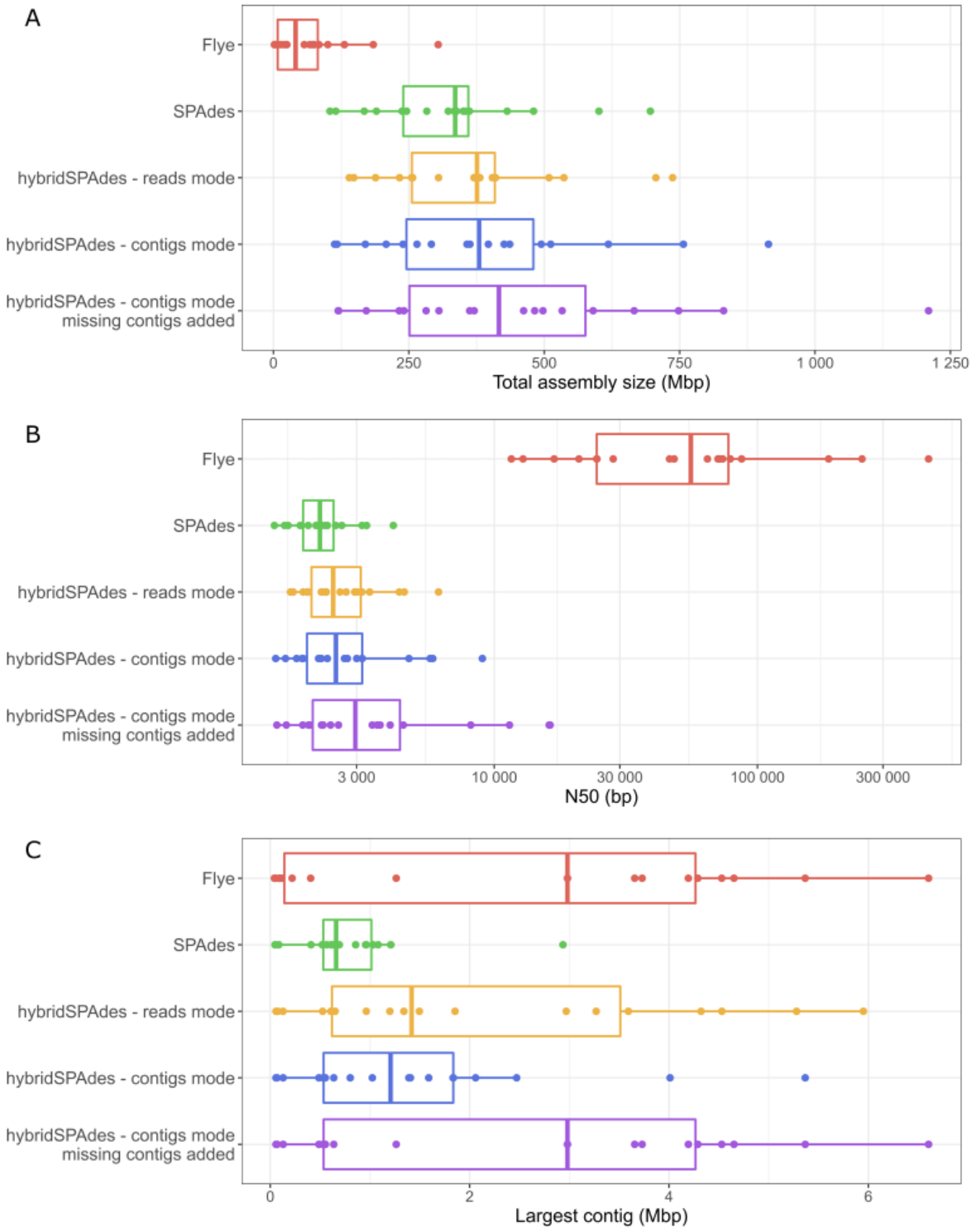

Figure 5: Comparison of assembly quality metrics across long-read (Flye), short-read (SPAdes) and hybrid (hybridSPAdes) assembly methods (n=18). (a) Total assembly size in megabasepairs. (b) N50 of contig lengths. (c) Largest contig in megabasepairs. Contigs with a of size less than 1000 bp were excluded from the calculations.

which is highly dependent on the study objective. Furthermore, the use of different DNA extractions for sequencing of long and short reads seems to hinder hybrid assembly. On the other hand, using specialised extraction methods such as high molecular weight protocols for Nanopore sequencing can improve the quality of the DNA extracted, and thereby theoretically produce longer reads. The chosen DNA extraction method has a big impact on the ability to perform hybrid assembly and should not be overlooked.

Overall, we show that long reads are very valuable in the context of metagenomic assembly and that using a hybrid approach improves assembly quality compared to a short-read only approach. Accurate reconstruction of genomes from metagenomic data has many applications, such as for predicting antimicrobial resistance genes and plasmids present in the human gut microbiome[58,59] but also in environmental samples[60]. Therefore, it is expected that hybrid methods such as the one presented here will become increasingly relevant in the field of metagenomics. The presented approach is especially useful in the common scenario where long-read data is available but limited.

Supplementary data

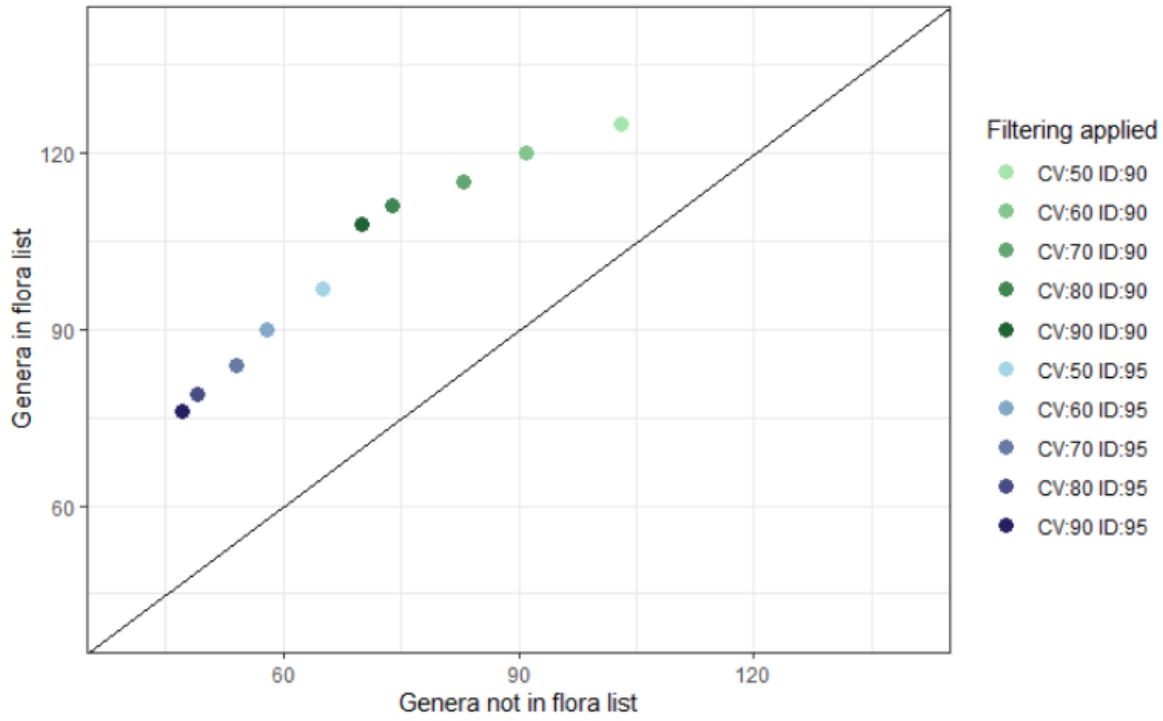

Figure S1: The number of false positive vs. true positive PhyloNorway hits of sample DTU_2021_1010055_1_MG_Nuuk_ID69_S1_StV23C_0_5_inf1, for different coverage and identity thresholds. Hits were assessed on genus level and considered false positives if they were not included in a list of vascular plants found in Greenland. This list is based on *Grønlands Flora* by Böcher et al.[61] and supplemented with more recent findings.

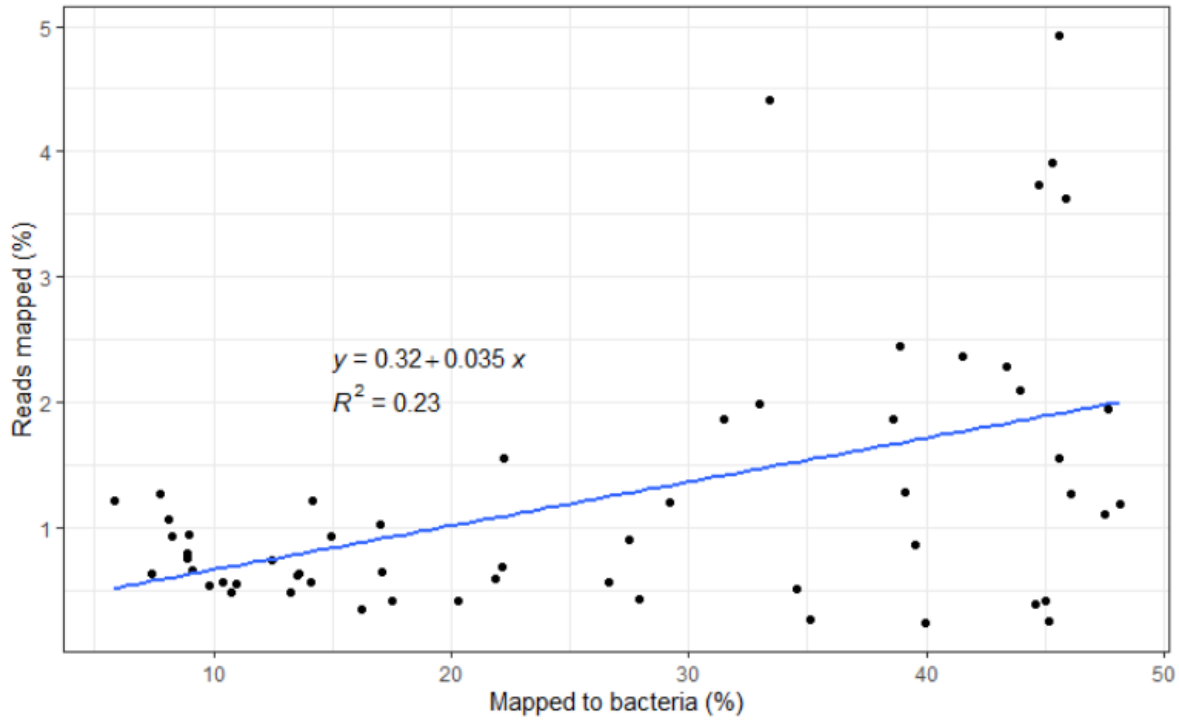

Figure S2: The percentage of reads mapped versus the percentage of mapped reads that are mapped to bacteria, per sample (n=57). Regression line shown in blue, with formula and R squared.

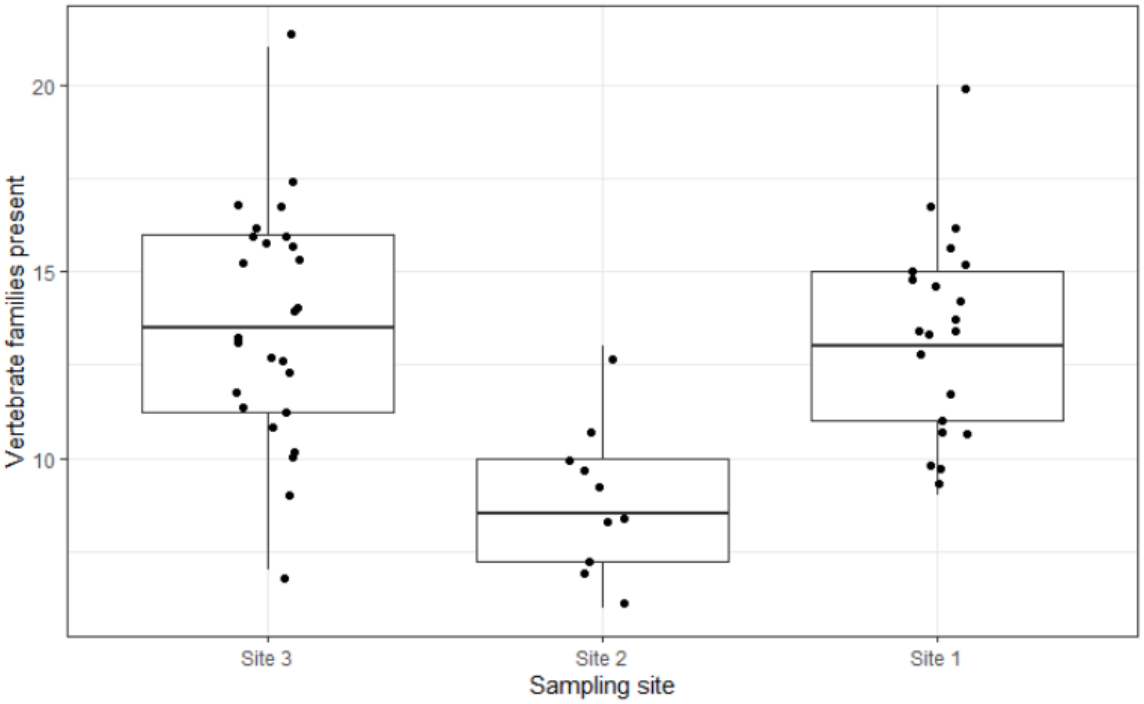

Figure S3: Total amount of vertebrate families present per sample, shown for each sampling site. Families with abundance < 20 in all samples are not shown.

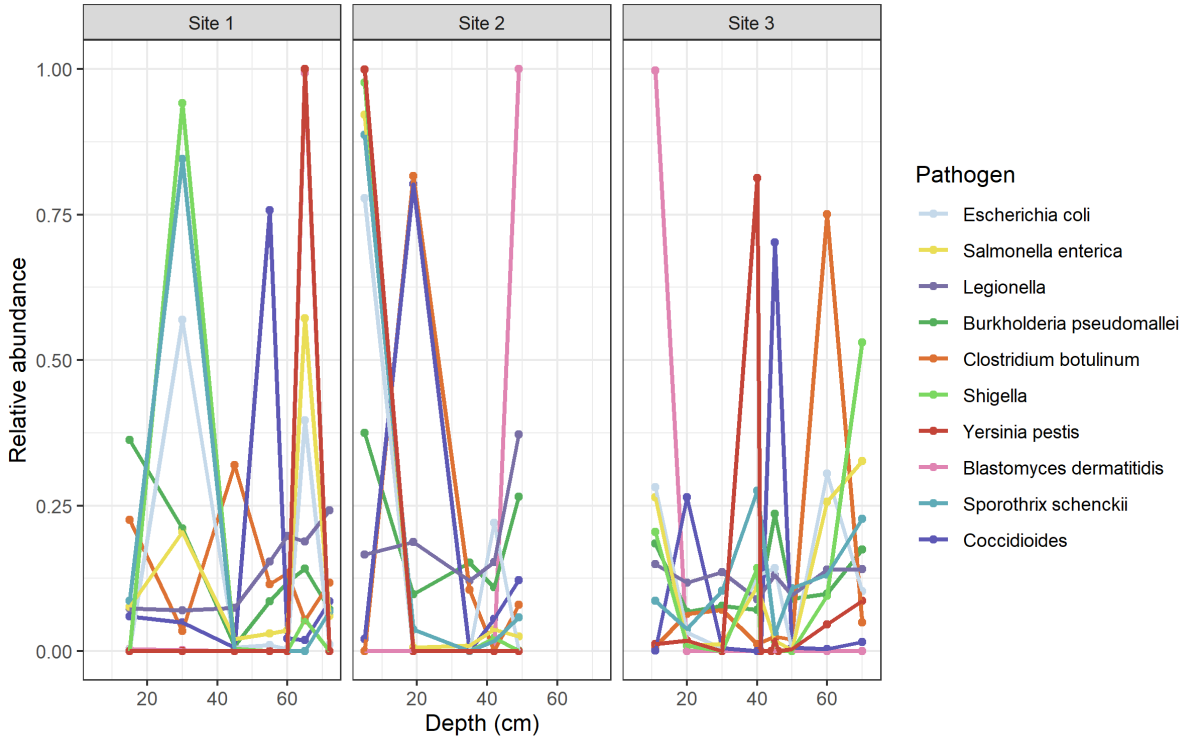

Figure S4: Relative abundance of each pathogen according to depth, split by sampling site. Counts were normalised for sequencing depth.

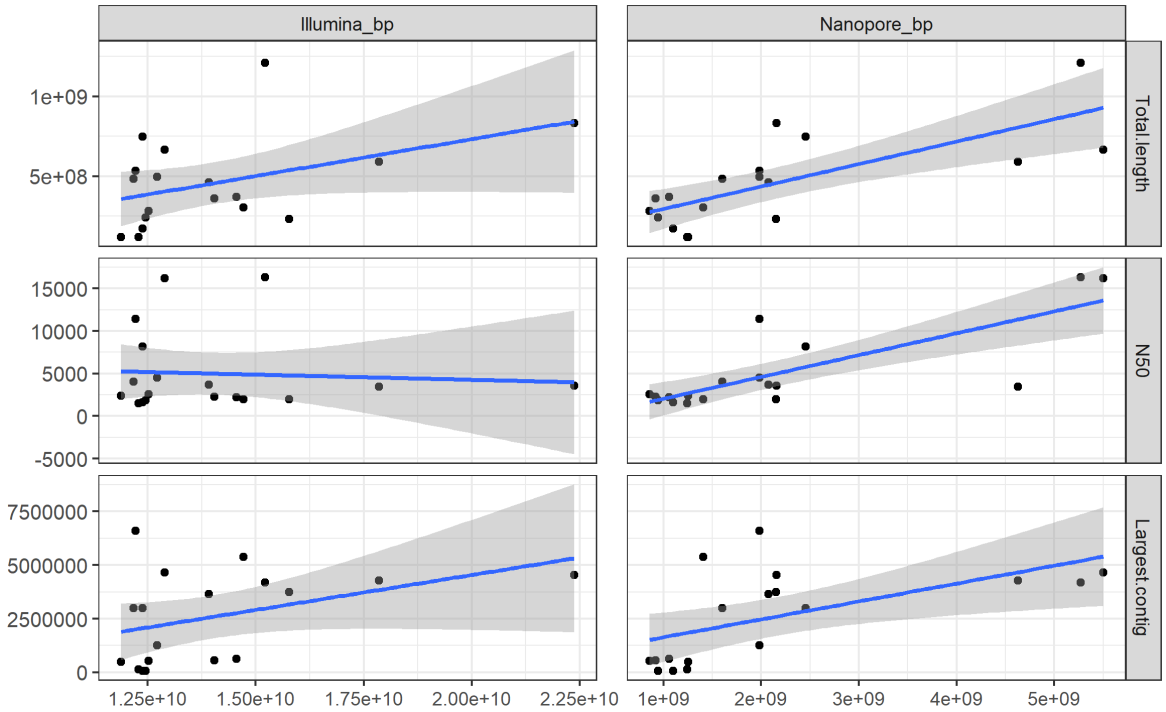

Figure S5: Values of the metrics assembly length, N50 and largest contigs compared to the amount of Illumina data (left) and Nanopore data (right) in basepairs. Linear regression is shown in blue including 95% confidence interval in grey.

Table S1: Overview of taxa found in Nanopore reads but not in Illumina reads, according to alignment to the SILVA database using KMA.

Sample	Nr. of taxa	Total counts
DTU_2021_1010001_1_MG_Nar_ID2_SFA_P1_5_10	27	27
DTU_2021_1010012_1_MG_Nar_ID18_SFB_P5_5_10	43	43
DTU_2021_1010063_1_MG_Nuuk_ID77_S1_StV23C_0_5_inf9	13	13
DTU_2021_1010067_1_MG_Nuuk_ID81_S1_StV23C_5_10_out2	20	21
DTU_2021_1010073_1_MG_Nuuk_ID87_S1_StV23C_5_10_out8	17	17
DTU_2021_1010095_1_MG_Nuuk_ID111_S2_StV23B_5_10_inf8	39	40
DTU_2021_1010100_1_MG_Nuuk_ID116_S2_StV23B_5_10_out2	21	22
DTU_2021_1010119_1_MG_Nuuk_ID137_S3_StV24A_0_5_inf9	29	29
DTU_2021_1010135_1_MG_Nuuk_ID153_S3_StV24A_19_4130_mid4	31	31
DTU_2021_1010143_1_MG_Nuuk_ID161_S3_StV24A_51_5955_mid12	88	90
DTU_2021_1010144_1_MG_Nuuk_ID162_S3_StV24A_59_6360_mid13	110	114
DTU_2021_1010148_1_MG_Nuuk_ID166_S3_StV24A_63_7165_mid17	69	69
DTU_2021_1010173_1_MG_Nuuk_ID191_S5_StNuuk_70_mid21	26	26
DTU_2021_1010184_1_MG_Nuuk_ID207_S4_StV26A_0_5_inf3	10	10
DTU_2021_1010197_1_MG_Nuuk_ID220_S4_StV26A_5_10_inf16	174	185
DTU_2021_1010209_1_MG_Ser_ID446_0_5_Sed1	312	338
DTU_2021_1010216_1_MG_Ser_ID454_0_Cli1	42	43
DTU_2021_1010219_1_MG_Ser_ID457_0_Cli4	74	78

References

1. Raghavan, M. *et al.* The genetic prehistory of the New World Arctic. *Science* **345**, 1255832 (2014).
2. Harris, A. J. T. *et al.* Dorset Pre-Inuit and Beothuk foodways in Newfoundland, ca. AD 500-1829. *PloS One* **14**, e0210187 (2019).
3. Arneborg, J. *et al.* Norse Greenland Dietary Economy ca. AD 980-ca. AD 1450: Introduction. *J. N. Atl.* **2012**, 1–39 (2012).
4. Lynnerup, N. *The Greenland Norse. A Biological-anthropological Study*. vol. 24 (The Commission for Scientific Research in Greenland, 1998).
5. Dugmore, A. J. *et al.* Cultural adaptation, compounding vulnerabilities and conjunctures in Norse Greenland. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3658–3663 (2012).
6. Jackson, R. *et al.* Disequilibrium, Adaptation, and the Norse Settlement of Greenland. *Hum. Ecol. Interdiscip. J.* **46**, 665–684 (2018).
7. Dugmore, A. J., Keller, C. & McGovern, T. H. Norse Greenland settlement: reflections on climate change, trade, and the contrasting fates of human settlements in the North Atlantic Islands. *Arct. Anthropol.* **44**, 12–36 (2007).
8. Arneborg, J. *et al.* Change of Diet of the Greenland Vikings Determined from Stable Carbon Isotope Analysis and 14C Dating of Their Bones. *Radiocarbon* **41**, 157–168 (1999).
9. Arneborg, J., Lynnerup, N. & Heinemeier, J. Human Diet and Subsistence Patterns in Norse Greenland AD C.980—AD c. 1450: Archaeological interpretations. *J. N. Atl.* **2012**, 119–133 (2012).
10. Henriksen, P. S. Norse agriculture in Greenland: farming at the northern frontier. In *Northern Worlds - landscapes, interactions and dynamics* (ed. Gulløv, H. C.) 423–431 (Syddansk Universitetsforlag, 2014).
11. Sørensen, I. Pollenundersøgelser i møddingen i Niaqussat. *Tidsskr. Grønl.* 296–302 (1982).
12. Fredskild, B. & Humle, L. Plant remains from the Norse farm Sandnes in the western settlement, Greenland. (1991) doi:10.1080/08003839108580400.

13. Pedersen, M. W. *et al.* Ancient and modern environmental DNA. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20130383 (2015).

14. Seersholm, F. V. *et al.* DNA evidence of bowhead whale exploitation by Greenlandic Paleo-Inuit 4,000 years ago. *Nat. Commun.* **7**, 13389 (2016).

15. Willerslev, E. *et al.* Ancient Biomolecules from Deep Ice Cores Reveal a Forested Southern Greenland. *Science* **317**, 111–114 (2007).

16. Hebsgaard, M. B. *et al.* 'The Farm Beneath the Sand' – an archaeological case study on ancient 'dirt' DNA. *Antiquity* **83**, 430–444 (2009).

17. McGovern, T., Amorosi, T., Perdikaris, S. & Woollett, J. Vertebrate zooarchaeology of Sandnes V51: Economic change at a Chieftain's farm in West Greenland. *Arct. Anthropol.* **33**, 94–121 (1996).

18. Wu, R., Trubl, G., Taş, N. & Jansson, J. K. Permafrost as a potential pathogen reservoir. *One Earth* **5**, 351–360 (2022).

19. Sajjad, W. *et al.* Resurrection of inactive microbes and resistome present in the natural frozen world: Reality or myth? *Sci. Total Environ.* **735**, 139275 (2020).

20. Edwards, A. Coming in from the cold: potential microbial threats from the terrestrial cryosphere. *Front. Earth Sci.* **3**, (2015).

21. Westergaard-Nielsen, A., Karami, M., Hansen, B. U., Westermann, S. & Elberling, B. Contrasting temperature trends across the ice-free part of Greenland. *Sci. Rep.* **8**, 1586 (2018).

22. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2021).

23. Bushnell, B. *BBMap (sourceforge.net/projects/bbmap/)*.

24. Fast gapped-read alignment with Bowtie 2 | Nature Methods. https://www-nature-com.proxy.library.uu.nl/articles/nmeth.1923.

25. Amorosi, T., Buckland, P., Ólafsson, G. & Sadler, J. *Site status and the palaeoecological record: a discussion of the results from Bessastadir, Iceland*. (1992).

26. Steffan, J. J., Derby, J. A. & Brevik, E. C. Soil pathogens that may potentially cause pandemics, including severe acute respiratory syndrome (SARS) coronaviruses. *Curr. Opin. Environ. Sci. Health* **17**, 35–40 (2020).

27. Jang, J. *et al.* Environmental Escherichia coli: ecology and public health implications-a review. *J. Appl. Microbiol.* **123**, 570–581 (2017).

28. Ayyadurai, S. *et al.* Long-term persistence of virulent Yersinia pestis in soil. *Microbiology* **154**, 2865–2871.

29. Barbieri, R. *et al.* Yersinia pestis: the Natural History of Plague. *Clin. Microbiol. Rev.* **34**, e00044-19 (2020).

30. Haile, J. *et al.* Ancient DNA Chronology within Sediment Deposits: Are Paleobiological Reconstructions Possible and Is DNA Leaching a Factor? *Mol. Biol. Evol.* **24**, 982–989 (2007).

31. Hansen, A. J. *et al.* Crosslinks rather than strand breaks determine access to ancient DNA sequences from frozen sediments. *Genetics* **173**, 1175–1179 (2006).

32. Callow, C. & Evans, C. The mystery of plague in medieval Iceland. *J. Mediev. Hist.* **42**, 254–284 (2016).

33. Buckland, P. C. & Sadler, J. P. A Biogeography of the Human Flea, Pulex irritans L. (Siphonaptera: Pulicidae). *J. Biogeogr.* **16**, 115–120 (1989).

34. Juck, D. F. *et al.* Utilization of Fluorescent Microspheres and a Green Fluorescent Protein-Marked Strain for Assessment of Microbiological Contamination of Permafrost and Ground Ice Core Samples from the Canadian High Arctic. *Appl. Environ. Microbiol.* **71**, 1035–1041 (2005).

35. Olson, N. D. *et al.* Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief. Bioinform.* **20**, 1140–1150 (2019).

36. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinforma.* **3**, lqab019 (2021).

37. Rice, E. S. & Green, R. E. New Approaches for Genome Assembly and Scaffolding. *Annu. Rev. Anim. Biosci.* **7**, 17–40 (2019).

38. Ghurye, J. S., Cepeda-Espinoza, V. & Pop, M. Metagenomic Assembly: Overview, Challenges and Applications. *Yale J. Biol. Med.* **89**, 353–362 (2016).

39. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

40. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinforma. Oxf. Engl.* **31**, 1674–1676 (2015).

41. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinforma. Oxf. Engl.* **28**, 1420–1428 (2012).

42. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).

43. Delahaye, C. & Nicolas, J. Sequencing DNA with nanopores: Troubles and biases. *PLOS ONE* **16**, e0257521 (2021).

44. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

45. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

46. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinforma. Oxf. Engl.* **32**, 2103–2110 (2016).

47. Zimin, A. V. & Salzberg, S. L. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput. Biol.* **16**, e1007981 (2020).

48. Wick, R. R. & Holt, K. E. Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLOS Comput. Biol.* **18**, e1009802 (2022).

49. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinforma. Oxf. Engl.* **36**, 2253–2255 (2020).

50. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-596 (2013).

51. Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* **19**, 307 (2018).

52. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

53. Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).

54. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015 (2016).

55. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinforma. Oxf. Engl.* **22**, 1658–1659 (2006).

56. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinforma. Oxf. Engl.* **29**, 1072–1075 (2013).

57. Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience* **8**, giz043 (2019).

58. Ye, L. *et al.* High-Resolution Metagenomics of Human Gut Microbiota Generated by Nanopore and Illumina Hybrid Metagenome Assembly. *Front. Microbiol.* **13**, 801587 (2022).

59. Bertrand, D. *et al.* Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* **37**, 937–944 (2019).

60. Brown, C. L. *et al.* Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes. *Sci. Rep.* **11**, 3753 (2021).

61. Böcher, T. W., Fredskild, B., Holmen, K. & Jakobsen, K. *Grønlands Flora*. (P. Haase & Søns Forlag, 1978).