UiT The Arctic University of Norway

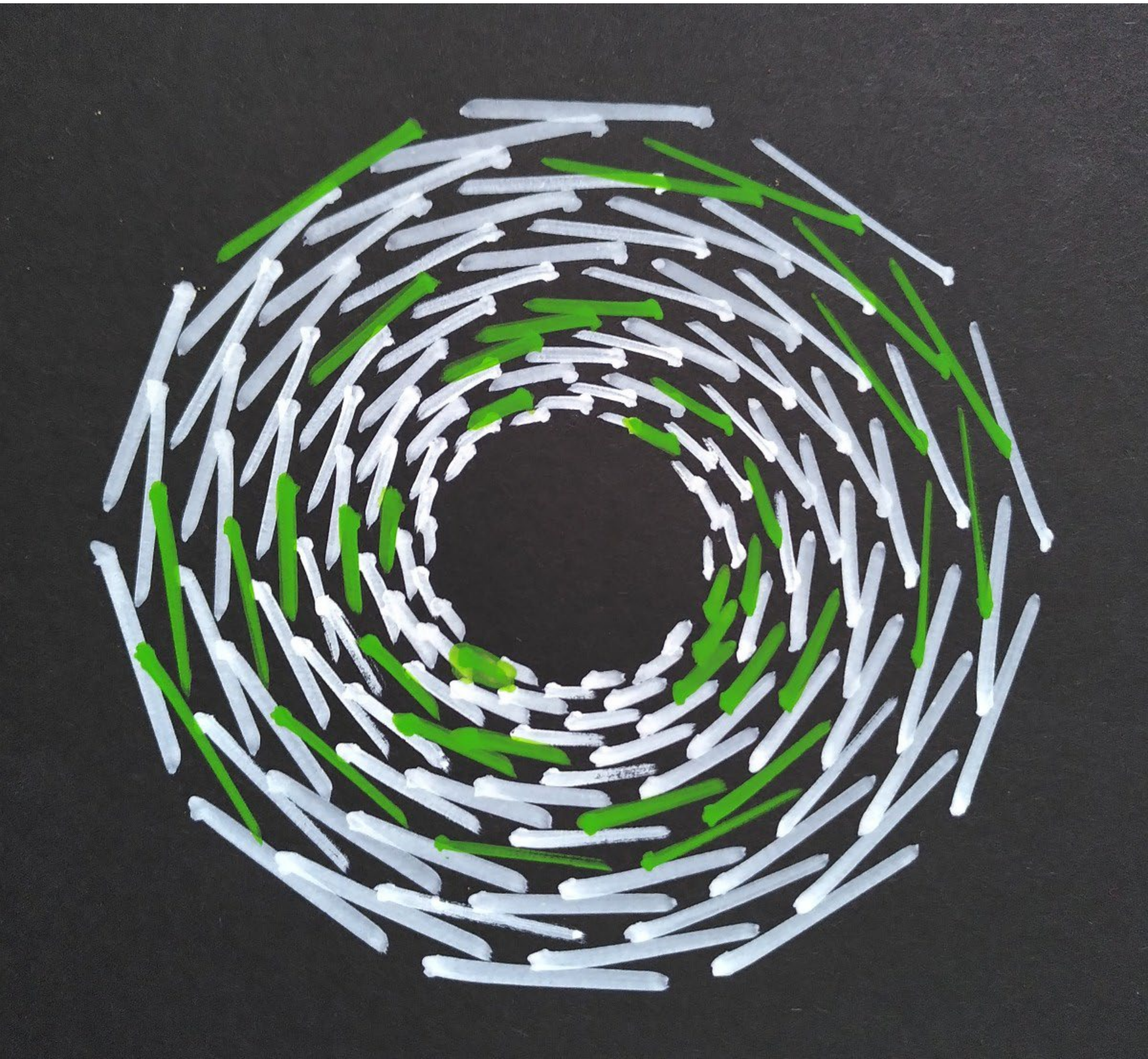
Department of Arctic and Marine Biology

Unlocking the potential of ancient sedimentary DNA

Approaches to reconstructing past vegetation assemblages

Lucas Dane Elliott

A dissertation for the degree of Philosophiae Doctor November 2024



Unlocking the potential of sedimentary ancient DNA

Acknowledgements

I would first like to thank Inger and Tony for giving me the opportunity to work on this project. I could not have dreamed of a better supervisor than Inger who has empowered me as an independent researcher while steering me in the right direction. A special thanks to everyone in the botany building at Tromsø museum for creating the best working environment I have ever experienced. I wouldn't have survived in the lab without the guidance and support of Marie, Iva, Jenny, and Ingebjørg. I feel lucky to have followed in the footsteps of Youri who has made the life of a bioinformatician easy in the ECOGEN lab.

Thank you to my parents who have supported me living an ocean away. Thanks to the wonderful community of people in Tromsø who have kept me sane throughout my PhD, especially Ferdinand, Joséphine, and Mathilde.

I would also like to thank Éric Coissac and Frédéric Boyer of the Laboratoire d'Ecologie Alpine. Spending two summers working with you both in Grenoble has been one of my fondest research memories. Thank you also to Kathleen Stoof-Leichsenring and Janine Klimke of the Alfred Wegener Institute for kindly guiding me through the two week process of shotgun sequencing. The *wholeskim* pipeline wouldn't have been possible without the work and guidance of Téo Lemane. Every technical aspect of this thesis is supported by contributors to stackoverflow and other forums as well as the authors and maintainers of Free Open Source Software.

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement no. 819192 (to Inger Greve Alsos and Antony Brown supporting me) for the IceAGenT project.

Abstract

Past ecosystem dynamics provide valuable information to inform current and future management. Ancient DNA has rapidly become a powerful tool in uncovering these lost environments. We demonstrate the ecological inferences made possible by metabarcoding analysis on sedimentary ancient DNA by investigating the role climate and glacial activity have had on a catchment's vegetation community. We identified four plant assemblages spanning the Holocene at Jøkelvatnet, a lake in northern Norway, that shifted concurrently with glacial activity and mean summer temperatures. The taxonomic resolution of metabarcoding allowed for traits of the assemblages, such as soil disturbance dependence and temperature optimum, to be reconstructed through time.

However, metabarcoding and standard shotgun sequencing analysis methods are targeting only a small fraction of the total DNA preserved in sediment archives. We present two workflows that expand the genomic breadth of information able to be retrieved from these repositories. Multiplex PCR allows for the reliable recovery of intraspecific variable sites for an organism of interest. We demonstrate this ability by identifying four populations of *Vaccinium uliginosum* in five sediment cores spanning the Holocene. In shotgun sequencing analysis, reference databases are currently limited to assembled genomic regions which comprise only a small fraction of total biodiversity. Here we present *wholeskim*, a pipeline that allows for unassembled genome skims to be used as a reference for annotating ancient metagenomes, unlocking previously inaccessible nuclear genomic regions for study. These tools show promise for obtaining larger amounts and more varied information from sedimentary ancient DNA.

Sammendrag

Endringer i fortidens økosystemer gir verdifull innsikt for å veilede nåværende og fremtidige miljøtiltak. Urgammelt DNA (aDNA) har raskt etablert seg som et effektivt verktøy for å utforske disse tapte økosystemene. Vi demonstrerer økologiske innsikter muliggjort av metabarcoding-analyse på urgammelt DNA fra innsjøsedimenter (sedaDNA), ved å utforske hvordan klimaendringer og bre-aktivitet har påvirket vegetasjonsutviklingen. Ved Jøkelvannet, en innsjø i Nord-Norge, identifiserte vi fire vegetasjonsperioder i Holocen som endret seg i takt med bre-aktivitet og gjennomsnittlige sommertemperaturer. Metabarcodingens taksonomiske oppløsningen muliggjorde rekonstruksjon av plantenes egenskaper over tid, inkludert deres tilpasning til forstyrrelser og temperaturforhold.

Metabarcoding og standard shotgun sequencing metoder fokuserer imidlertid kun på en liten brøkdel av det totale DNA som er bevart i sedimentene. Vi presenterer to teknikker som utvider den genomiske bredden av informasjon som kan ekstraheres fra disse kildene. Multiplex PCR muliggjør pålitelig utvinning av intraspesifikke variable DNA koder for fokusorganismer. Vi demonstrerer denne metoden ved å identifisere fire populasjoner av blokkebær i fem sedimentkjerner gjennom Holocen. I shotgun sequencing er referansedatabaser for tiden begrenset til sammensatte genomiske regioner som bare omfatter en liten brøkdel av det totale biologiske mangfoldet. Vi introduserer *wholeskim*, en bioinformatisk prosess som gjør det mulig å bruke primære DNA fragmenter som referanse for å identifisere urgammle metagenomer, og derved åpne opp tidligere utilgjengelige nukleære genomiske områder for forskning. Disse verktøyene viser stort potensiale for å utvide mengder og variasjon av informasjon fra urgammelt DNA.

List of papers

Paper I: Sedimentary ancient DNA reveals local vegetation changes driven by glacial activity and climate. 2023. Elliott, L. D., Rijal, D. P., Brown, A. G., Bakke, J., Topstad, L., Heintzman, P. D., & Alsos, I. G. Quaternary, 6(1), 7. doi.org/10.3390/quat6010007

Paper II: Multiplexing PCR allows the identification of within‑species genetic diversity in ancient eDNA. 2024. Lammers, Y., Taberlet, P., Coissac, E., Elliott, L.D., Merkel, M.F., Pitelkova, I., PhyloAlps Consortium, PhyloNorway Consortium and Alsos, I.G.. Molecular Ecology Resources, p.e13926. doi.org/10.1111/1755-0998.13926

Paper III: Wholeskim: utilizing genome skims for taxonomically annotating ancient DNA metagenomes. In prep. Elliott, L. D., Boyer, F., Lemane, T. PhyloNorway Consortium, Alsos, I.G., and Coissac, E.

Paper IV: Comparison of target enrichment and shotgun sequencing of lake sedaDNA metagenomes. Elliott, L. D., Murchie, T., Stoof-Leichsenring, K., Strandberg, N., Merkel, M. F., Pitelkova, I., Alsos, I. G.

Author contributions

	Paper I	**Paper II**	**Paper III**	**Paper IV**
Concept and idea	LE, IGA, AB	YL, PT, IGA	LE, EC, IGA	LE, IGA
Study design and methods	LE, IGA, AB, DR	YL, PT, EC, IGA	LE, EC, FB, TL	LE, IGA, TM, KS

Data gathering and interpretation	LE, IGA, AB, DR, JB, LT, PH	YL, PT, LE, MM, IP, IGA, EC	LE, EC, FB, IGA	LE, IGA, TM, KS, NS, MM, IP, DR
Manuscript preparation	LE	YL	LE	LE

LE = Lucas Elliott

IGA = Inger Alsos

JB = Jostein Bakke

FB = Frederic Boyer

AB = Antony Brown

EC = Eric Coissac

PH = Peter Heintzman

TL = Teo Lemane

MM = Marie Merkel

TM = Tyler Murchie

IP = Iva Pitelkova

DR = Dilli Rijal

KS = Kathleen Stoof-Leichsenring

NS = Nichola Strandberg

PT = Pierre Taberlet

LT = Lasse Topstad

Introduction

The Arctic region has seen a rise in surface temperature at a rate of more than twice the global average (Cohen et al. 2014; IPCC 2023) with substantial shifts in arctic vegetation in response to warming have already been documented (Bjorkman et al. 2019). Effective conservation efforts must account for these shifts by anticipating populations' adaptive potential and rates of dispersal (Alsos et al. 2012; Barnosky et al. 2017). Investigating ecosystem dynamics during past climate change events can aid in predicting populations' evolutionary adaptive potential, phenotypic plasticity, and rates of dispersal (Nogués-Bravo et al. 2018). Long-term monitoring of ecosystems can provide valuable insights on these phenomena, but only reaches timescales of a century at the extreme (Goldberg and Turner 1986), when these processes can span millennia.

Two main sources of data are used to make inferences of past ecosystem dynamics: 1) indirect evidence of past demography by examining current genetics through the lens of phylogeography or 2) direct evidence of species' presence using macrofossils, pollen, and sedimentary ancient DNA.

Phylogeography can provide information on past colonization routes using contemporary samples, but recent migrations or demographic changes can mask the genetic signals of older events (Alsos et al. 2015). Similarly, species distribution models are often based on current ranges and climate, a limited snapshot in time, while palaeoecological records can uncover dynamics hidden by previously temporally limited information (Alsos et al. 2024). The most powerful tool for reconstructing these histories is DNA that is deposited from organisms in a lake catchment and is subsequently bound and preserved on mineral surfaces (Laura Parducci et al. 2017; Freeman et al. 2023). This resource has primarily been exploited by metabarcoding where conserved genetic regions are used to PCR amplify a wide taxonomic range of organisms' DNA while the interspecifically variable barcode region is used to identify a particular organism (Taberlet et al. 2018; Capo et al. 2021). A single barcode is sufficient to detect the presence of taxa in an environmental sample to species-level (Taberlet et al. 2018), while multiple regions are required to differentiate populations as demonstrated in modern eDNA samples (Andres et al. 2021). This combination of multiple primers in a single PCR has been termed "multiplexing" and has not yet been demonstrated on ancient eDNA samples to recover population-level information.

Once taxa have been identified in a sedaDNA time series, it is possible to track ecosystem changes and correlate these changes to climate or anthropogenic factors (Garcés-Pastor et al. 2022). The species-level resolution of metabarcoding allows for plant assemblages to be characterized using trait databases (Tyler et al. 2021; Inger Greve Alsos et al. 2022). However, single metabarcoding primers are unable to provide species-level resolution for all taxa (Taberlet et al. 2018). Using alternative methods to retrieve greater than species-level resolution allows for the detection of population turnover events and introgression (Schulte et al. 2021).

One of the greatest challenges when working with ancient DNA is that it is characteristically highly fragmented with a reported average length of < 35 base pairs (Pedersen et al. 2016). Metabarcoding is unable to retain the vast majority of genomic information in sedaDNA as even the “gh” primers of the short *trn*L p6-loop barcode are 39 bp themselves while inserts range from 10 - 143 bp (Taberlet et al. 2007). As an alternative, the DNA content of an environmental sample can be directly sequenced without amplification producing genome-wide information from many organisms termed the “metagenome”.

In contrast with metabarcoding, metagenomic datasets are not limited to a single locus, instead, the reads are composed of sequences distributed throughout the entire genome. This necessitates a reference database ideally encompassing the whole genomes of all potential organisms of interest. The most comprehensive set of reference sequences for metagenomic analysis is currently provided by the International Nucleotide Sequence Databases, a collaborative project of GenBank, the DNA Data Bank of Japan (DDBJ), and the European Nucleotide Archive (ENA) (http://www.insdc.org). However, these databases still fall short of offering complete genomic sequences for all species. Several ongoing initiatives aim to sequence and assemble the genomes of all known species on Earth (Gilbert et al. 2014; Lewin et al. 2018), but these efforts are projected to span several decades (Lewin et al. 2022). Until these projects come to fruition, genome skimming—low-coverage, non-targeted sequencing of all DNA extracted from an organism's tissues—can provide genome-wide information for numerous taxa with minimal cost and effort (Straub et al. 2011; Coissac et al. 2016). The scalability of genome skimming is exemplified by the PhyloNorway and PhyloAlps projects, which have sequenced representatives of the entire vascular flora of Norway/Polar Regions and the European Alps/Carpathians, respectively (Alsos et al. 2020). The ~2,100 genome skims of PhyloNorway cover 1,845 species and have a mean value of 4.64 million read pairs (sd = 1.58 million, Alsos et al. 2020). A subset of these genome skims have been used to annotate metagenomic sedaDNA datasets from across the arctic and are able to annotate 23x more reads to Viridiplantae than NCBI's nt database (Wang et al. 2021). However, this study assembles the low coverage genome skims into contigs with an average length of 216 base pairs in order for the mapping program, *bowtie2* (Langdon 2015), to be able index these reference sequences (Langdon 2015; Wang et al. 2021). With PhyloNorway's genome skims averaging 0.5 - 1.0x depth of coverage (Alsos et al. 2020), the raw genome skims are losing a large amount of informative sequences during this assembly.

Metagenomic analysis faces the dual challenge of compiling a comprehensive reference database encompassing the entire genomes of all target organisms and developing specialized algorithms to efficiently compare the vast number of sequence reads generated for each sample to this terabyte-scale database. Published pipelines for taxonomic annotation of metagenomes typically utilize software from two main categories: mapping tools, such as Centrifuge (Kim et al. 2016) and Bowtie2 (Langmead and Salzberg 2012), and diagnostic k-mer based algorithms, such as Kraken2 (Wood et al. 2019). Although these programs can index databases such as NCBI's RefSeq, they struggle with the task of indexing the large number of short reads from genome skims. Mapping-based software requires reference sequences to be substantially longer than the query reads to ensure efficient processing and sensitive alignments. Similarly, many k-mer based approaches are unable to efficiently index the terabytes of short reads produced by genome skims with high k-mer complexity (Lemane et al. 2024). One solution to this problem has been demonstrated by *kmindex* which leverages the probabilistic Bloom filter data structure to index the k-mers of large metagenomic datasets and accurately queries them using the findere algorithm to significantly reduce false positive identifications (Robidou and Peterlongo 2021; Lemane et al. 2024). Bloom filters will never produce a false negative, but have an intrinsic false positive rate based on the number of bits used to construct the index and the number of entries, but this rate can be reduced by using multiple, successive queries of k-mers (Bloom 1970; Robidou and Peterlongo 2021).

There are two main approaches to metagenomic sedaDNA studies; shotgun sequencing in which a prepared DNA library is directly sequenced (Pedersen et al. 2016; Parducci et al. 2019; Wang et al. 2021) and target enrichment (or hybridization capture) in which the sequencing library is first enriched for taxa or genomic regions of interest (Mamanova et al. 2010; Murchie et al. 2021; Schulte et al. 2021).

Shotgun sequencing has the advantage of providing a relatively unbiased image of the total DNA content of a sample since no filter or preselection of molecules is occurring. However, this untargeted approach also leads to a large portion of unidentifiable reads. Typically >95% of shotgun sequenced sedaDNA samples are unidentified molecules with the next largest category composed of Prokaryotes (Heintzman et al. 2023). Metagenomic sedaDNA studies using primarily the NCBI nt or RefSeq databases as references reported very low proportions of reads identified to any taxonomic level of Viridiplantae, with (Slon et al. 2017) identifying a mean of 0.07%, (Courtin et al. 2022) identifying 0.05%, and (Parducci et al. 2019) identifying only 0.0002% of queried reads. To compensate, studies targeting animal or plant taxa drastically increase the depth of sequencing, up to 16 billion reads generated (Kjær et al. 2022), to recover these small proportions of reads. The number of sequenced reads has been used to model plant taxa abundance (Wang et al. 2021), but it is uncertain how well sedaDNA represents biomass/abundance since the taphonomic process of DNA preservation in sediments is not well understood (Giguet-Covex et al. 2023) and since genomic coverage in the DNA reference library strongly affects detection (Wang et al. 2021).

Target enrichment selects a subset of DNA reads corresponding to a specific group of taxa or genomic region. This is accomplished by first designing a “bait” set of complementary DNA molecules of interest that are used to bind a portion of the sequencing library while the remainder is discarded. These baits can be designed and synthesized or used from modern amplified DNA extracts (Maricic et al. 2010). The hybridizing temperature can be adjusted to control how similar the molecules binding to the bait set are, allowing for deaminated fragments, individual variation, and closely related taxa to still be retained (Heintzman et al. 2023). However, this can also lead to off-target sequences being enriched. Compared to shotgun sequencing, hybridization capture significantly reduces the required depth of sequencing as the resulting library is greatly enriched for DNA of interest. Studies report up to a 1600x increase in the number of target sequences obtained with target enrichment compared to shotgun sequencing (Schulte et al. 2021). Both of these approaches have been used for the assembly of organelles and nuclear regions and calling of haplogroups given sufficient coverage (Lammers et al. 2021; Pedersen et al. 2021; Vernot et al. 2021).

Research questions

How can metabarcoding sedaDNA data be used to answer ecological and phylogeographic questions in plants? (Papers I and II)

Are metagenomic annotations of *sed*aDNA using unassembled genome skims able to provide informative assignments? (Paper III)

How do the metagenomic annotations of target enrichment and shotgun sequencing compare? (Paper IV)

What are the specificity and sensitivity errors produced by metagenomic sequence annotation? (Papers III and IV)

Methods

Study sites

The data used in this thesis is primarily sedaDNA from a set of 22 lakes in northern Norway that were selected based meeting most or all of the following three criteria; topography providing small inflow streams to the lake, surrounding vegetation representing a variety of ecosystems from boreal forest to alpine heath, and putatively undisturbed sedimentation from human or natural forces (Supplementary Table 1). One lake from this set, Jøkelvatnet, is examined in detail in Paper I. Surface sediment samples from the full set of lakes are used to compare and contrast target enrichment and shotgun sequencing approaches in Paper IV. In Paper II, sedaDNA from three additional lakes are included; Hopschusee and Krumschnabelsee from the Alps (Garcés-Pastor et al. 2022) and Bolshoye Shchuchye located in the Polar Urals, Russia (Clarke et al. 2018). Three additional sedaDNA samples from archaeological middens in northern Norway (Komatsu et al. in prep) are also included to demonstrate the taxonomic annotation efficacy of *wholeskim* in Paper III.

Vegetation Surveys

Vegetation surveys provide important information that allow for a form of "ground-truthing" taxa identified by sedaDNA as well as providing a description of the current day vegetation community of the catchment (Alsos et al. 2018). We conducted a vegetation survey at Jøkelvannet during July 2021 for Paper I and integrated this information alongside the vegetation surveys for the other 21 lakes for Paper IV. We covered the perimeter of the lake and recorded every species that was growing within 2 meters of the shore as well as any visible macrophytes. Additionally, we made an effort to identify species growing further away from the lake in different habitats such as the birch forest, moraines, talus slopes, and springs present in the catchment. Herbarium vouchers were collected for select taxa and deposited at the Arctic University Museum of Norway Herbarium in Tromsø (TROM).

Modern plant tissue collection

To supplement the genomic information contained in PhyloNorway, we aimed to add genome skim information from multiple individuals of the same species from different populations across Norway to capture intraspecific variation. Candidate species were chosen and given a priority rating based on their abundance in the *seda*DNA record. As this was conducted during corona, six sampling sites across Norway were chosen to capture the diversity of plant populations; Kirkenes, Nordkapp, Tromsø, Andøya, Karmøy, and Kristiansand (Supplementary Table 1). Fresh leaf material was collected from up to ten individuals at each site in May - July 2020 (Supplementary Table 2). Individuals were spaced at least 10 m apart and the elevation of each site ranged from sea level to 852 m.o.h. Recent vegetative buds were targeted as they contain the highest concentration of DNA as well as the fewest possible contaminants. The tissue was immediately stored in silica gel containers to desiccate and herbarium vouchers were collected for all populations.

Genome skims

From the modern plant tissue, we produced 25 genome skims from 16 species (Supplementary Table 3). The DNA extraction and library preparation protocols follow those described in Alsos et al. 2020 and were performed at the Arctic University Museum of Norway in Tromsø. Briefly, the Macherey-Nagel Nucleospin 96 Plant II kit was used to extract DNA from 20g of dried leaf tissue. The resulting DNA extract was then sonicated using the E210 Covaris instrument (Covaris, Inc., USA) and NEBNext DNA Modules Products (New England Biolabs, MA, USA) were used for end-repair, 3'-adenylation and ligation of NextFlex DNA barcodes (Bio Scientific Corporation). Three libraries were sequenced using 151 base-length read chemistry in a paired-end flow cell on an Illumina MiniSeq sequencer (Illumina, USA) from September - November 2020. Nine of these genome skims were used to simulate sedaDNA reads for pipeline benchmarking in Paper III and the *Vaccinium uliginosum* skims were used in the multiplex PCR testing in Paper II.

Coring and subsampling

The surface sediment samples were collected from a subset of the 20 northern Norwegian lakes using a Kajak corer (mini gravity corer) with a diameter of 3 cm and a length of 63 cm (Alsos et al. 2018) and from the remaining lakes using a UWITEC USC 06000 corer with a diameter of 5.9 cm (Rijal et al. 2021). A Nesje piston core (Nesje 1992) totalling 258 cm was collected from the northern basin of Jøkelvatnet. It was split longitudinally and samples were taken from the length of the core with precautions minimizing contamination and open negative controls to monitor for potential ambient DNA (Rijal et al. 2021; Capo et al. 2021) (Figure 1).

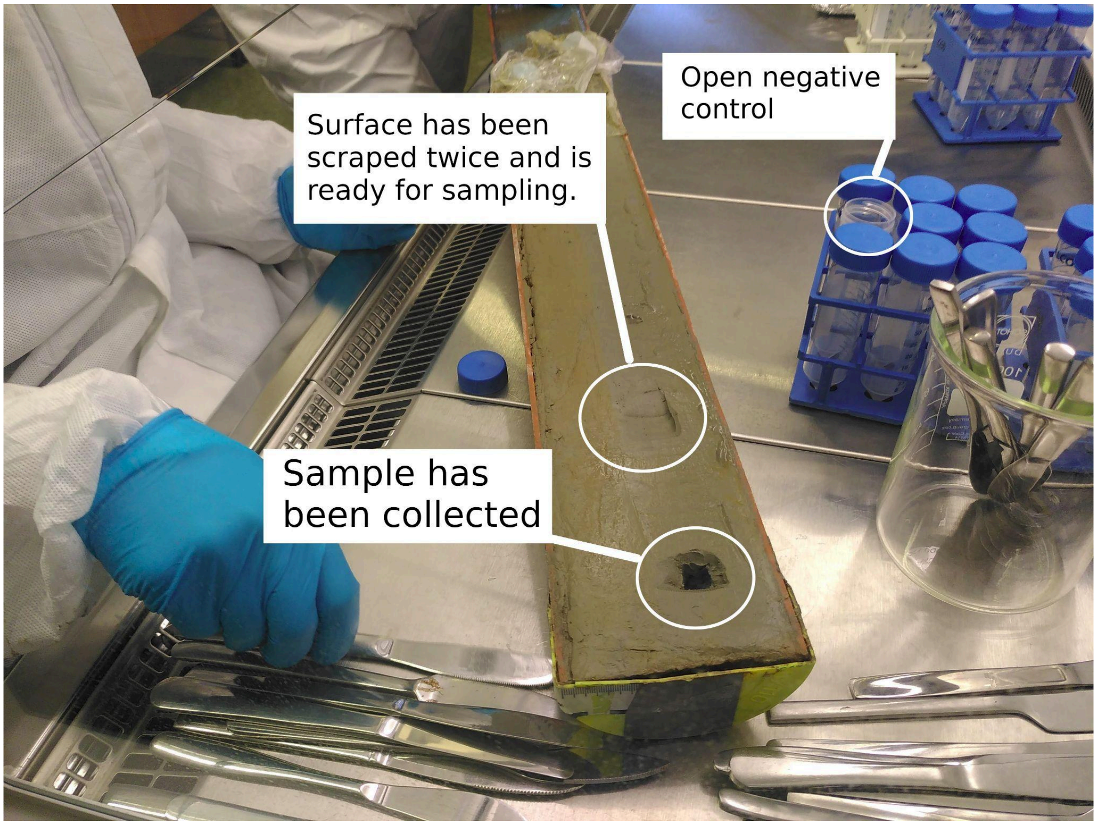

Open negative control
Surface has been scraped twice and is ready for sampling.
Sample has been collected

Figure 1. Alois Revéret demonstrating sediment subsampling techniques under a fumehood minimizing and monitoring for contamination. (Photo: Lucas Elliott)

Metabarcoding

From Jøkelvatnet, 40 samples were amplified with the "gh" metabarcoding primers targeting the *trn*L p6-loop of the chloroplast (Rijal et al. 2021). The resulting two libraries were sequenced on ~10% of 2x 150-cyclemid-output flow cell on the Illumina NextSeq platform at the Genomics Support Centre Tromsø at The Arctic University of Norway. The multiplex amplicons were designed to have an average length of 79 bp and the primer sequences were optimized to have a melting temperature of 55.7 °C. Sequencing runs were performed on ~10% of 2 x 150-cycles on the Illumina MiSeq platform at the Norwegian College of Fishery Science at The Arctic University of Norway.

Shotgun sequencing

For the northern Norwegian surface sediments, DNA was extracted from 250 mg of surface sediment from each sample using the DNeasy PowerSoil Extraction kit (Qiagen, Germany) and following the Murchie et al. (2021) protocol in the ancient DNA laboratory at the Arctic University Museum of Norway in Tromsø in January 2023. This protocol features an additional centrifuge step with Dabney binding buffer to remove inhibitors. An extraction blank was processed alongside each group of 8 samples. The three archaeological midden samples were extracted separately using the method described previously in the metabarcoding section in May 2022 (Komatsu et al. in prep).

Metagenomic shotgun sequencing of these 23 samples was performed using single-stranded library preparation designed specifically for highly degraded ancient DNA (Gansauge et al. 2017; Gansauge and Meyer 2013). From the extraction process described above, 15 ng DNA was used as a template for libraries that were prepared for sequencing at the Alfred Wegener Institute in Potsdam, Germany. Indexing PCR was performed in 10–14 cycles (for samples and blanks) depending on the library concentration with indexed P5 and P7 primers from the Nextera Index kit. Amplificates were purified with the MinElute PCR Purification Kit (Qiagen, Germany) and library size distribution was checked on the Agilent TapeStation using the D1000 ScreenTape (Agilent Technologies, USA). The three pools, composed of in total 23 samples, three extraction blanks, and one library blank, were sequenced in paired-end mode (2 × 100 bp) on a

NextSeq2000 at the Alfred Wegener Institute Helmhotz Centre for Polar and Marine Research, Bremerhaven, Germany. The data from the three archaeological middens (Komatsu et al. in prep) were used in Paper III while all other samples were used as the main dataset in Paper IV.

Target Enrichment

DNA was extracted from 250 mg of surface sediment from 22 surface sediment samples using the DNeasy PowerSoil Extraction kit (Qiagen, Germany) and following the Murchie et al. (2021) protocol in the ancient DNA laboratory at the Arctic University Museum of Norway in Tromsø in October 2019. This protocol is identical to the one used to extract the shotgun sequencing samples discussed previously.

The target enrichment and sequencing of these samples was performed at McMaster Ancient DNA Centre, Hamilton, Canada in 2019. The sequencing library was prepared with 15 ng of each DNA extract using Meyer and Kircher's (2010) double stranded method with modifications from Kircher et al. (2012), and a modified end-repair reaction to account for the lack of uracil excision. The adapter-ligated, dual-indexed libraries were then enriched using the PalaeoChip ArcticPlant-1.0 bait-set, which had been designed in collaboration with Arbor Biosciences (Murchie et al. 2019). The bait-set targets ~2100 circumarctic plant taxa, based on the databases available from Sønstebø et al. (2010), Soininen et al. (2015), and Willerslev et al. (2014). The chloroplast locus *trn*L is the primary target of these reference databases; additional full *trn*L loci from GenBank were added to the bait-set to augment some of the particularly short sequences (<50 bp) available in the original references. The loci *rbc*L and *mat*K were also added where available to further increase the chloroplast targeting scope. Libraries were pooled with the goal of attaining ~1,000,000 sequenced reads per library and sequenced on an Illumina HiSeq 1500 with a 2 x 90 bp paired-end protocol at the Farncombe Metagenomics Facility (McMaster University, ON).

Bioinformatic tools

Metabarcoding

Metabarcoding data was parsed using OBITools (Boyer et al. 2016) and custom Python and R scripts following the procedure in (Alsos et al. 2022). Sequences were matched with 100% identity to the following four databases; 1) PhyloNorway (Inger Greve Alsos et al. 2020), 2) a combination of 815 arctic (Sønstebø et al. 2010) and 835 boreal (Willerslev et al. 2014) vascular plant taxa and 455 arctic-boreal bryophytes (Soininen et al. 2015) from the circumpolar region (ArcBorBryo, n = 2280 sequences of which 1053 are unique), 3) PhyloAlps (n = 4604 specimens of 4437 taxa collected in the Alps and Carpathians, (Alsos et al. 2020) (data.phyloalps.org/browse (accessed on 26 September 2022)), and 4) EMBL (release 143, n = 159,748 sequences of 74,936 taxa). A final manual check of all matches were done based on knowledge of regional flora and cover of the reference library (Alsos et al. 2022). For Paper I, we assigned vascular plant taxa ecological trait values from (Tyler et al. 2021) that are likely to be influenced by climate and glacial activity: moisture, temperature optimum, and soil disturbance. Additionally, we retrieved reconstructed climatic data (“mean temperature of warmest quarter” (bio10) and “annual precipitation” (bio12)) for Jøkelvatnet from the CHELSA-TraCE21k model (Karger et al. 2021).

Wholeskim

We developed *wholeskim*, a tool for indexing unassembled genome skims and accurately annotating sedaDNA metagenomes. This pipeline is able to index the large number of short reads produced by genome skims through the use of *kmindex*, a k-mer based software that leverages probabilistic Bloom filters, and *findere*, an algorithm for reducing the number of false positive produced by this data structure (Bloom 1970; Robidou and Peterlongo 2021; Lemane et al. 2024).

With *wholeskim*, the decision to assign a taxon to a query read, Q, is based on the number of shared k-mers, S_G, it has with each of the genome skims, G. The assignment algorithm is:

- Calculate $N_Q = L_Q - k + 1$, the total number of k-mers present in Q, a read of length L_Q.

- Identify S_{max}, the maximum number of k-mers shared between Q and any of the indexed genome skims G.
- Calculate $t_{max} = S_{max}/N_Q$.
- If $t_{max} \geq t_c$, the cutoff proportion for a positive match, define $t_{min} = t_{max} - \Delta$, a threshold for similar matches.
- Calculate $S_{min} = t_{min} * N_Q$.
- Select all genome skims G with $S_G \in [S_{min}; S_{max}]$.
- Assign to Q the lowest common ancestor (LCA) of all taxa associated with the selected genome skims, using the NCBI taxonomy as a reference.

To reduce the noise of false positive assignments, only assignments to taxa that appeared in greater than a proportion, r, of the total reads were retained. After optimization through testing with simulated datasets, the three parameters of the procedure have been set to $t = 0.7$, $\Delta = 0.1$, $r = 10^{-5}$.

In order to test the robustness of the *wholeskim* pipeline using unassembled genome skims (*wholeskim*-unassembled), we compared the accuracy of this workflow with the *Holi* pipeline using the assembled version of the same genome skims (Pedersen et al. 2016; Wang et al. 2021) on both simulated *sed*aDNA reads and true ancient metagenomes. Additionally, we examined the influence of the genomic and taxonomic completeness of the reference database used with *wholeskim*-assembled on taxonomic assignment accuracy. We examined the effect of the sequencing depth of the genome skims by building databases including 0, 1, 2, 3, …, 19, 20, 24, 28, …, 68 M reads of *Vaccinium uliginosum* and assigning simulated *sed*aDNA reads of the same species. For the effect of taxonomic completeness of the reference database, we simulated and assigned *sed*aDNA reads of *Thesium alpinum*, which has no family members in the reference dataset, and *Salix retusa*, which is not represented in the reference dataset, but 39 other *Salix* species are present.

Simulating sedaDNA datasets

To assess the specificity and sensitivity of the *wholeskim*-unassembled workflow, we simulated sedaDNA datasets from the genome skims of eleven species. Two species are not represented in the PhyloNorway database and genome skims were taken from the PhyloAlps project (Thesium alpinum (PHA009155) and Salix retusa (PHA007876)). The other nine of these species have a representative genome skim in the PhyloNorway database (*Avenella flexuosa, Betula nana, Betula pubescens, Bistorta vivipara, Caltha palustris, Dryas octopetala, Picea abies, Pinus sylvestris,* and *Vaccinium uliginosum*) and genome skims produced for this thesis were used as a basis for the simulations. Sets of simulated reads were obtained from a genome skim using the adrsm software (Borry 2018). Read simulation consisted of reducing the actual length of the sequencing reads of the genome skims to mimic the size distribution observed in ancient DNA metagenomes (mean insert size set to 35 bp based on the fragment length profiles from median age samples in (Pedersen et al. 2016)). A set of bacterial genome assemblies and a *Homo sapiens* genome assembly (GRCh38) were subjected to the same simulation procedure to provide some off-target reads for the query dataset, but these reads were never assigned by any pipeline so they were not included in further analyses. This process resulted in approximately 200,000 reads for each genome skim.

Shotgun sequencing

Taxonomic annotation of the shotgun reads was performed by the *wholeskim* pipeline (Paper III). Merged reads from the shotgun dataset less than 34 bp were discarded as they are unable to be identified using wholeskim with an effective k-mer size of 33. A k-mer similarity cutoff was set to 0.7 and taxa were only retained if they composed at least 1% of the total reads identified to Embryophyta with a minimum read count of 5 (Paper III).

For the set of northern Norwegian lakes, the reference database used for annotation was constructed using the PhyloNorway database comprising unassembled genome skims from 1,823 species (Alsos et al. 2020) as well as the NCBI RefSeq entries with a “complete genome” assembly level for bacteria (2), fungi (4751), and algae which were compiled by collecting the plant entries which belonged to the following groups; Stramenopiles (33634), Rhodophyta (2763), and Chlorophyta (3041). No other plant sequences were added to the database as PhyloNorway provides a relatively even coverage reference for all Norwegian flora including common *sed*aDNA contaminants such as *Triticum aestivum* and *Solanum tuberosum*. Reads from the shotgun dataset were assigned to the LCA using the built-in parsing from *wholeskim* which considers all matches within 10% k-mer similarity of the maximum match (Paper III).

Target enrichment

Merged target enrichment reads less than 30 bp were discarded as taxonomic resolution is poor for these short fragments (Pedersen et al. 2016). Since the target enrichment bait set targeted the *trn*L, *rbc*L, and *mat*K loci all located on the chloroplast, the reads were mapped to a custom database of assembled chloroplasts using bowtie2 with default parameters (Langdon 2015). The custom database was constructed by compiling the 1,845 assembled plastid genomes produced by the PhyloNorway project (Alsos et al. 2020) as well as the NCBI RefSeq entries with a "complete genome" assembly level for bacteria (2), fungi (4751), and algae which were compiled by collecting the plant entries which belonged to the following groups; Stramenopiles (33634), Rhodophyta (2763), and Chlorophyta (3041). Reads from the target enrichment dataset were assigned to the lowest common ancestor (LCA) using ngsLCA with a minimum edit distance proportion of 0.97 (Wang et al. 2022). Taxa were only retained in the final dataset if they had at least 5 reads present in a sample.

Results and Discussion

Ecological inferences from metabarcoding (Paper I)

Jøkelvatnet is a distal glacier-fed lake in northern Norway, directly downstream from the Langfjordjøkelen ice cap. This ice cap currently makes up 18% of the lake's catchment while the rest is largely composed of steep talus slopes, with a smaller area of alpine heath directly surrounding the lake. The catchment of Jøkelvatnet was first deglaciated around 12.9 ka, and the ice cap was absent or very small during the Holocene Thermal Maximum from 10.0 ka to 4.1 ka, where Langfjordjøkelen reformed with frequent centennial-scale fluctuations (Wittmeier et al. 2015). In this paper, we examined the vegetation community's response to these climatic and glacial changes over the Holocene by metabarcoding 38 samples from a sediment core spanning 10.4 ka years.

We identified 193 plant taxa including 133 vascular plants and 60 bryophytes. The samples showed a trend of continually increasing species richness until the present day. We identified four distinct vegetation assemblage zones though CONISS; 1) 10.4-9.8 ka (4 samples) which is characterized by few taxa per sample containing primarily cold tolerant dwarf shrubs and forbs, 2) 9.7-8.7 ka (5 samples) marks the arrival of woody taxa *Betula* and *Sorbus* as well as the arrival of new fern and forb taxa, 3) 8.2 - 4.5 ka (9 samples) starts a gradual increase in total vegetation diversity with *Phyllodoce* and *Vaccinium* arriving, and 4) 4.2 - 0.0 ka (20 samples) begins with a sharp increase in taxa diversity with many graminoids, forbs, and bryophytes arriving in this period.

From the eligible 133 vascular plant taxa, 90 taxa were found to have informative ecological trait values for “soil disturbance”, 81 taxa for “moisture”, and 74 taxa for “temperature optimum” (Tyler et al. 2021). The soil disturbance value started high at the beginning of the record and reached a minimum value at 9.7 ka when the Langfjordjøkelen ice cap was thought to be absent (Figure 2C). This value started to gradually increase in the Late Holocene as the ice cap reformed and went through cycles of growth and retrait. Temperature optimum values showed an inverse trend to soil disturbance. They indicate cold-adapted taxa in Zone 1, but then quickly increase to a maximum at 9.7 ka during the start of the Holocene Thermal Optimum. This is followed by a gradual decrease of the temperature optimum value in the Middle and Late Holocene as regional temperatures decreased. Few taxa disappear from the sediment record after first being detected, implying these trait value shifts are caused by new vegetation communities establishing in the catchment rather than the full turnover of communities.

Significant shifts in the vegetation community at 9.7 and 4.3 ka mirror inflection points in Langfjordjøkelen’s glacial activity (Figure 2F). The vegetation trait values are also correlated to the glacial activity changes ($R^2 > 0.4$) throughout the core. The correlation of vegetation soil disturbance trait values and glacial activity implies that the ice cap had a direct impact on the vegetation community in the catchment. As glacial melting increased soil erosion both in the Early and Late Holocene, more non-competitive taxa established in the catchment. Other lakes in northern Norway without a glacier in their catchment record much more stable soil disturbance and temperature optimum values during the Middle and Late Holocene (Alsos et al. 2022). The taxonomic resolution provided by the *trn*L marker allows for these vegetation trait analyses and the ability to reconstruct the ecosystem dynamics of Jøkelvatnet. Questions remain about where these plants dispersed from and if there were any population turnover events during the climatic and glacial shifts over the Holocene. To answer these questions, it is necessary to expand the genetic target of our *seda*DNA analysis from a single locus to multiple regions in the genome.

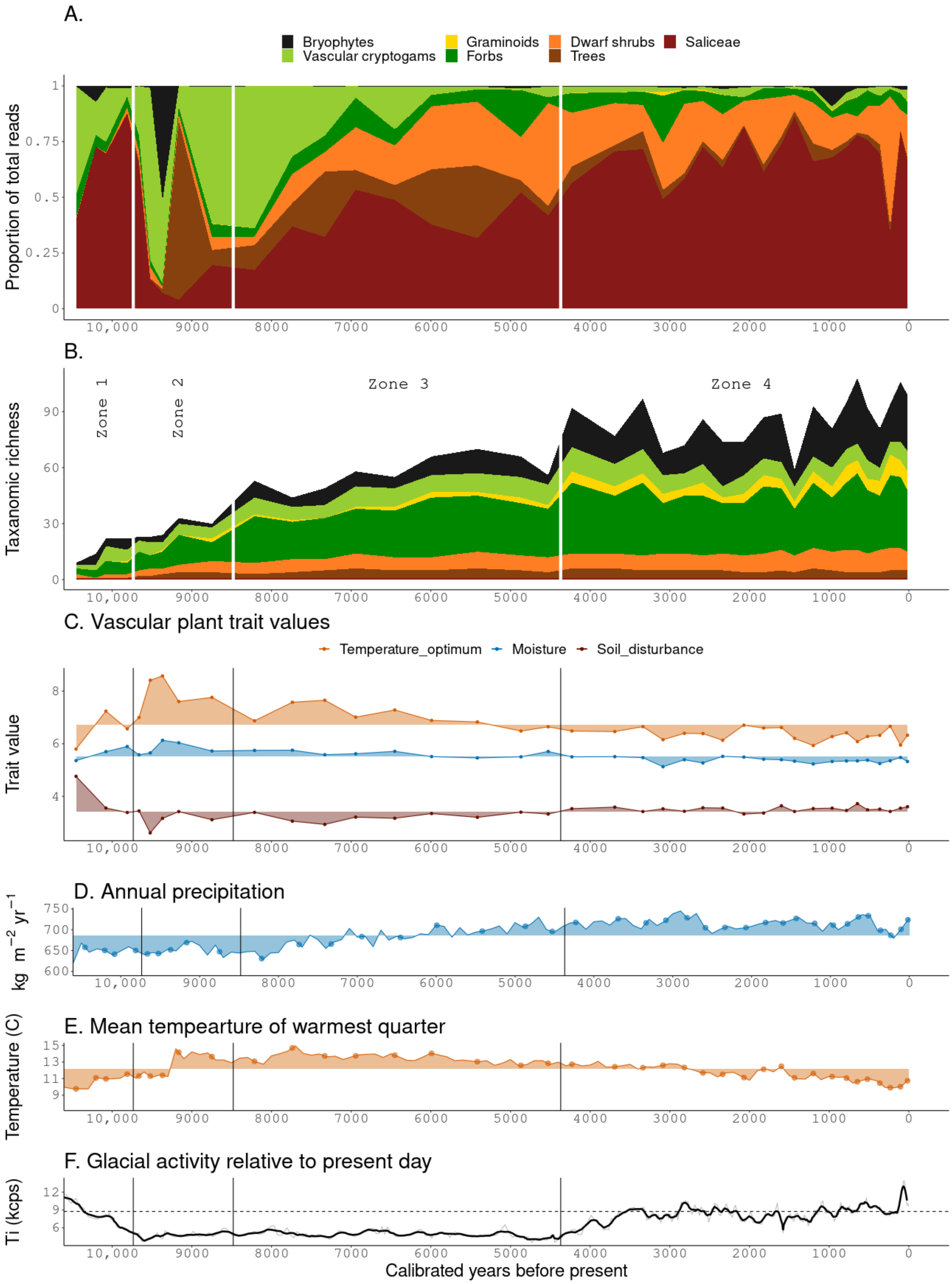

A.
Bryophytes
Vascular cryptogams
Graminoids
Forbs
Dwarf shrubs
Trees
Saliceae
Proportion of total reads
1
0.75
0.5
0.25
0
10,000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
B.
Taxanomic richness
90
60
30
0
Zone 1
Zone 2
Zone 3
Zone 4
C. Vascular plant trait values
Temperature_optimum
Moisture
Soil_disturbance
Trait value
8
6
4
D. Annual precipitation
kg m^{-2} yr^{-1}
750
700
650
600
E. Mean tempearture of warmest quarter
Temperature (C)
15
13
11
9
F. Glacial activity relative to present day
Ti (kcps)
12
9
6
Calibrated years before present

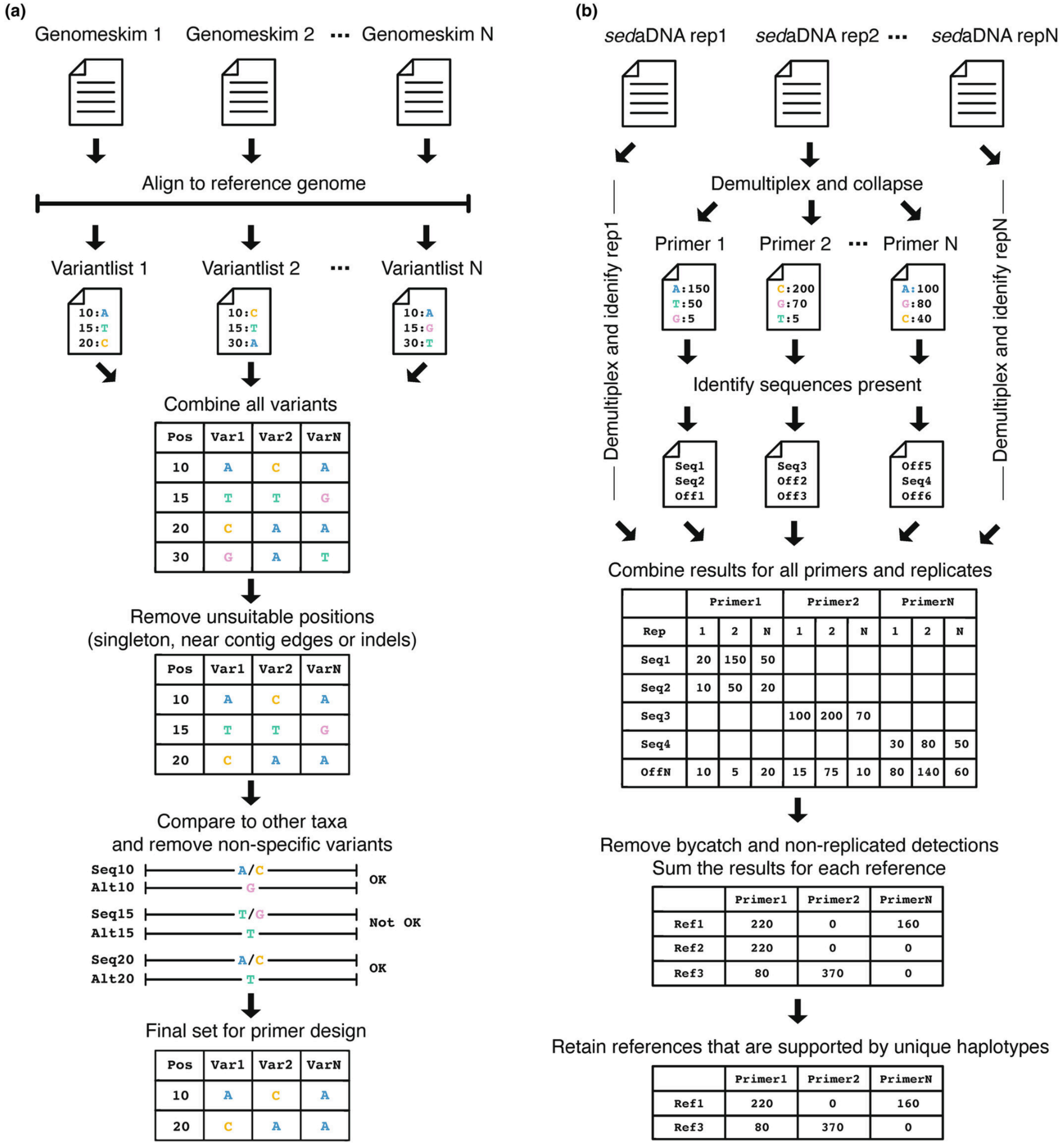
Figure 2. An overview of sedaDNA results, climate reconstructions, and glacial activity. CONISS zone boundaries are demarcated with vertical bars at 9.7, 8.4, and 4.3 ka. (A) Proportion of total identified reads by plant functional group. (B) Stacked taxonomic richness for each functional group. (C) Average weighted vascular plant trait values are based on plants identified in the sedaDNA combined with plant trait values reported in (Tyler et al. 2021), note that temperature optimum index values are inverted. High values indicate high temperature optimum, high moisture requirement and high dependents of soil disturbance. (D) Annual precipitation data (bio12) from [24]. Points represent the age of samples taken from the core. (E) Mean temperature of the warmest quarter (bio10) from (Karger et al. 2021). Points represent the age of samples taken from the core. (F) Glacial activity relative to the present day (dashed horizontal line) adapted from (Wittmeier et al. 2015).

Expanded metabarcoding for population-level information (Paper II)

The application of metabarcoding to sedimentary ancient DNA has previously been limited to taxonomic annotation at the species-level or higher. However, other methodological approaches such as shotgun sequencing (Lammers et al. 2021; Pedersen et al. 2021) or target enrichment (Schulte et al. 2021) have shown the ability to retrieve intraspecific variation from these community samples. Both of these methods are more resource and computationally intensive than metabarcoding and have limited capacity to scale to large numbers of samples. By combining multiple primers in one reaction, it's possible to retrieve population-level information from eDNA samples as demonstrated on contemporary samples (Andres et al. 2021). By targeting multiple intraspecifically variable, but interspecifically diagnostic regions with different primers, it's possible to detect the presence of one or many populations in a given sediment layer.

Here we used *Vaccinium uliginosum* as a model organism for methods development since it has several, well-differentiated populations (Alsos et al. 2005, Eidesen et al. 2007) and it is prevalent in the sedaDNA record (Alsos et al. 2022; Clarke et al. 2019). We used genome skims from 18 individuals collected across the Arctic as well as the Alps to detect candidate regions for primer design. From the aligned genome skims, we determined that the four main lineages previously described for *V. uliginosum* (Alsos et al. 2005) were represented among these 18 individuals. An initial total of 1,059 candidate regions were identified from these genome skims. After culling overlapping regions, positions close to the ends of contigs, indels, homopolymers, and non-species specific regions, a set of 61 multiplex compatible primers were produced for laboratory testing (Figure 3). Initial multiplex PCR testing was performed on two DNA extracts from modern *V. uliginosum* tissue collected in Honningsvåg and Kirkenes, Norway (Supplementary Table 1) as well as two lake sediment extracts from Langfjordvannet and Jøkeltvannet, Norway that were known to contain *V. uliginosum* DNA from previous *trn*L metabarcoding (Rijal et al. 2021). A subset of 38 primers successfully amplified DNA from these extracts and could distinguish between the four main lineages of *V. uliginosum*. This subset was then applied to the full dataset of 20 sediment samples spanning from the present day to 11.1 ka in five different lakes located in Norway, the Alps, and Russia. All 38 primer sets and all four lineages were observed among the 20 samples with 28.1 (SD 11.7) primers amplifying per sample.

These successful tests demonstrate the ability to recover genomic intraspecific variation from sedaDNA through multiplex PCR. In contrast to target enrichment and shotgun sequencing, this approach requires less sequencing depth and bioinformatic processing resources. As a result, larger numbers of samples can be processed providing increased temporal and spatial resolution. As a tradeoff, an upfront effort must be made to select genomic sites and design primers. Additionally, this approach is limited in the taxonomic and genomic breadth of the information produced when compared to shotgun sequencing. However, the information from multiplex PCR can elucidate population turnover and sources of species migration for the emerging field of palaeo-phylogeography.


(a)
Genomeskim 1
Genomeskim 2
Genomeskim N
Align to reference genome
Variantlist 1
Variantlist 2
Variantlist N
10:A
15:T
20:C
10:C
15:T
30:A
10:A
15:G
30:T
Combine all variants
Remove unsuitable positions
(singleton, near contig edges or indels)
Compare to other taxa
and remove non-specific variants
Seq10 A/C
Alt10 G
OK
Seq15 T/G
Alt15 T
Not OK
Seq20 A/C
Alt20 T
OK
Final set for primer design
(b)
*sed*aDNA rep1
*sed*aDNA rep2
*sed*aDNA repN
Demultiplex and collapse
Demultiplex and idenify rep1
Demultiplex and idenify repN
Primer 1
Primer 2
Primer N
A:150
T:50
G:5
C:200
G:70
T:5
A:100
G:80
C:40
Identify sequences present
Seq1
Seq2
Off1
Seq3
Off2
Off3
Off5
Seq4
Off6
Combine results for all primers and replicates
Remove bycatch and non-replicated detections
Sum the results for each reference
Retain references that are supported by unique haplotypes

Pos	Var1	Var2	VarN
10	A	C	A
15	T	T	G
20	C	A	A
30	G	A	T

Pos	Var1	Var2	VarN
10	A	C	A
15	T	T	G
20	C	A	A

Pos	Var1	Var2	VarN
10	A	C	A
20	C	A	A

	Primer1			Primer2			PrimerN		
Rep	1	2	N	1	2	N	1	2	N
Seq1	20	150	50						
Seq2	10	50	20						
Seq3				100	200	70			
Seq4							30	80	50
OffN	10	5	20	15	75	10	80	140	60

	Primer1	Primer2	PrimerN
Ref1	220	0	160
Ref2	220	0	0
Ref3	80	370	0

	Primer1	Primer2	PrimerN
Ref1	220	0	160
Ref3	80	370	0

Figure 3. (a) Workflow utilized for initial variant discovery and primer design. Starting with aligning N genome skims to a shared reference genome. The resulting variants are combined, and unsuitable positions (rare, near contig edges or indels) are removed. Each variant is compared to other reference taxa, and only those that are distinct are retained for primer design. (b) Workflow for the multiplex data analysis. Each replicate for a sample is demultiplexed separately, and for each primer present the variants are identified. After identification, the

replicate data for a sample are combined. Off-target identifications and non-replicated variants are removed. Variants are retained if they are supported by at least one unique haplotype.

Pipeline for metagenome annotation using unassembled genome skims (Paper III)

The metagenomic annotation tools *kraken2* (Wood et al. 2019), *centrifuge* (Kim et al. 2016), and *bowtie2* (Langdon 2015) were unable to index the 1.9 TB of unassembled genome skims from PhyloNorway due to requesting too much RAM. We benchmarked the two main workflows that could process this data as a reference database; our tool *wholeskim*-unassembled, which used a k-mer similarity of the query sequence to the unassembled PhyloNorway genome skims and *Holi*-assembled which mapped the query reads to the assembled contigs of PhyloNorway Pedersen et al. 2016; Wang et al. 2021).

The initial decontamination cleaning of the PhyloNorway genome skims reduced their size by a median value of 0.015%. Among the rejected reads, the median percentages of 0.007%, 0.004%, 0.002%, and 0.0005% were identified as contamination from algae, bacteria, fungi, and humans, respectively (Figure 4). However, some genome skims exhibited significantly higher contamination levels, with maximum values for these categories reaching 0.08%, 1.7%, 2.2%, and 0.18% respectively. A median value of 3.7% of the genome skims were identified by the filter as Viridiplantae reads, while the remaining ~96% did not match any of the reference database categories and were retained as unsequenced parts of the plant genome.

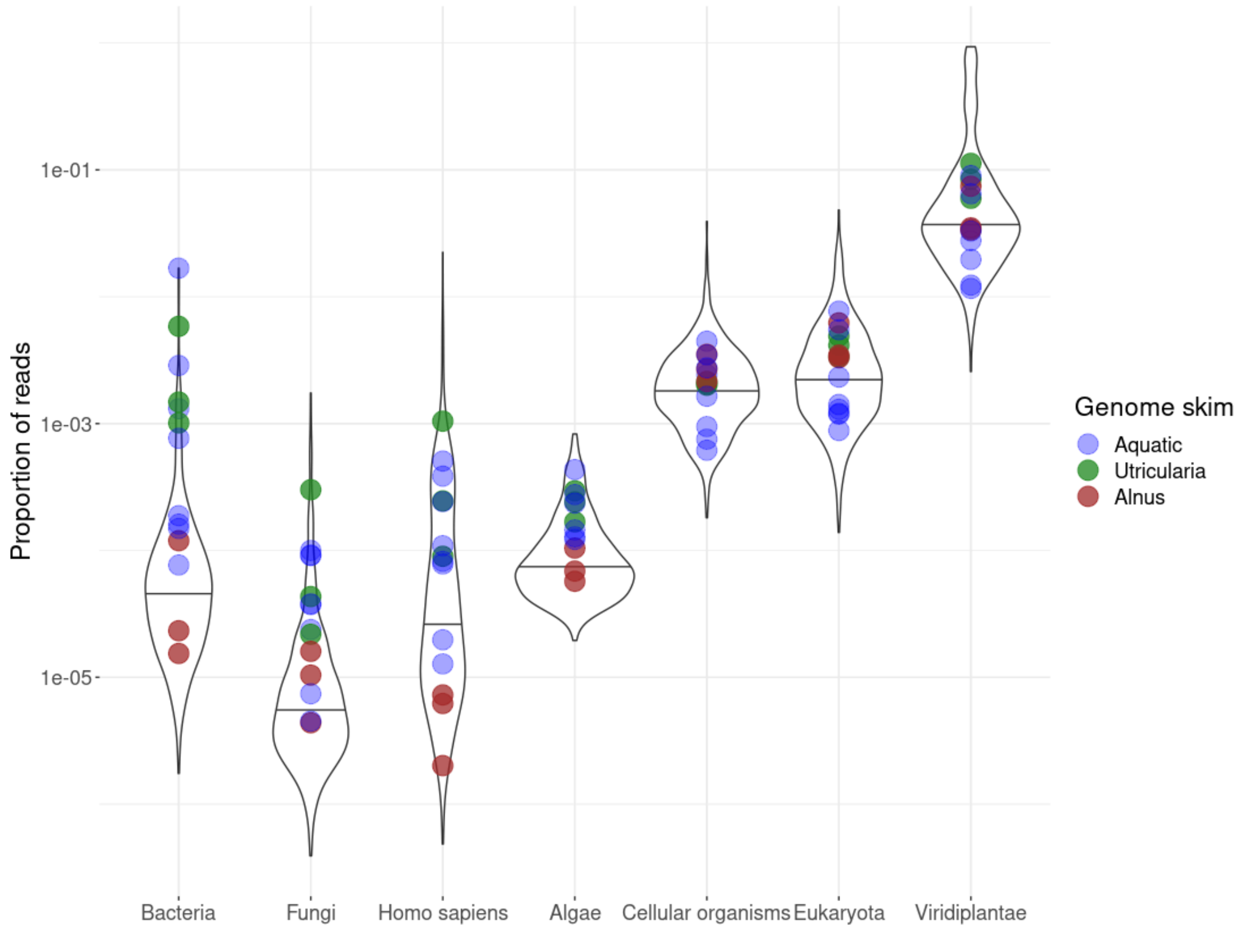


Figure 4. The proportions of contaminants detected in the initial genome skims from PhyloNorway. Identified bacterial, fungal, human, and algal reads were removed from the skims. The colored points indicate skims of *Utricularia*, *Alnus*, and aquatic taxa appearing in the 20 lakes dataset.

To assess the sensitivity and specificity of *wholeskim*-unassembled and *Holi*-assembled, a total of 2.1 million reads were simulated from eleven genome skims that were not included in PhyloNorway. From the nine species that are present in PhyloNorway, *wholeskim*-unassembled correctly identified 20.2% of the reads at a species or genus-level while *Holi*-assembled correctly identified 15.3% of the reads at these levels. *Holi*-assembled and *wholeskim*-unassembled incorrectly assign 3.0% and 3.4% of reads respectively with >90% of these misassignments to a congeneric species of the target. This misassignment could be attributed to the genome skims' incomplete genome coverage. If a conserved genomic region is sequenced by chance for one species, but not for a congeneric species, a query read originating from that region would be incorrectly annotated to species-level.

There is little overlap in the sets of reads that each workflow annotates with only 27.5% of the total correctly assigned reads shared between workflows (Figure 5). This difference in annotation can be explained by two workflow-specific factors. Compared to the contigs assembled by Wang et al. 2021, the unassembled genome skims had a median of 10.1x more k-mers (k=31). While some of this information undoubtedly originates from sequencing errors produced by the Illumina sequencing platform, another portion of the information comes from low depth of coverage regions of the genome skim that were unable to be assembled into contigs. With the PhyloNorway genome skims having an average depth of coverage of 0.5 - 1.0x (Alsos et al. 2020), these discarded regions likely constitute large portions of the nuclear genome that are absent from the *Holi*-assembled database. This factor could explain the reads solely identified by *wholeskim*-unassembled. Conversely, these sequencing errors could be generated on the query reads or intraspecific variation could cause single base pair mismatches when compared to the reference genome skim. Since *wholeskim*-unassembled annotates query reads using k-mer similarity with an effective k-mer size of 34, any single mismatch near the center of a typically short aDNA molecule would result in no shared k-mers with the reference genome. The *bowtie2* mapping approach of *Holi*-assembled is able to accommodate these single mismatches near the center of fragments while employing a sequence-similarity cut-off. This factor likely results in the target reads that are only correctly annotated by *Holi*-assembled. Ancient DNA is typically characterized by C to T transitions (Dalén et al. 2023) which would result in a mismatch with a reference genome, but these damaged base pairs are found at the ends of DNA fragments and would not have a large effect on the k-mer similarity score of *wholeskim*-unassembled given that the fragment is slightly longer than the k-mer size.

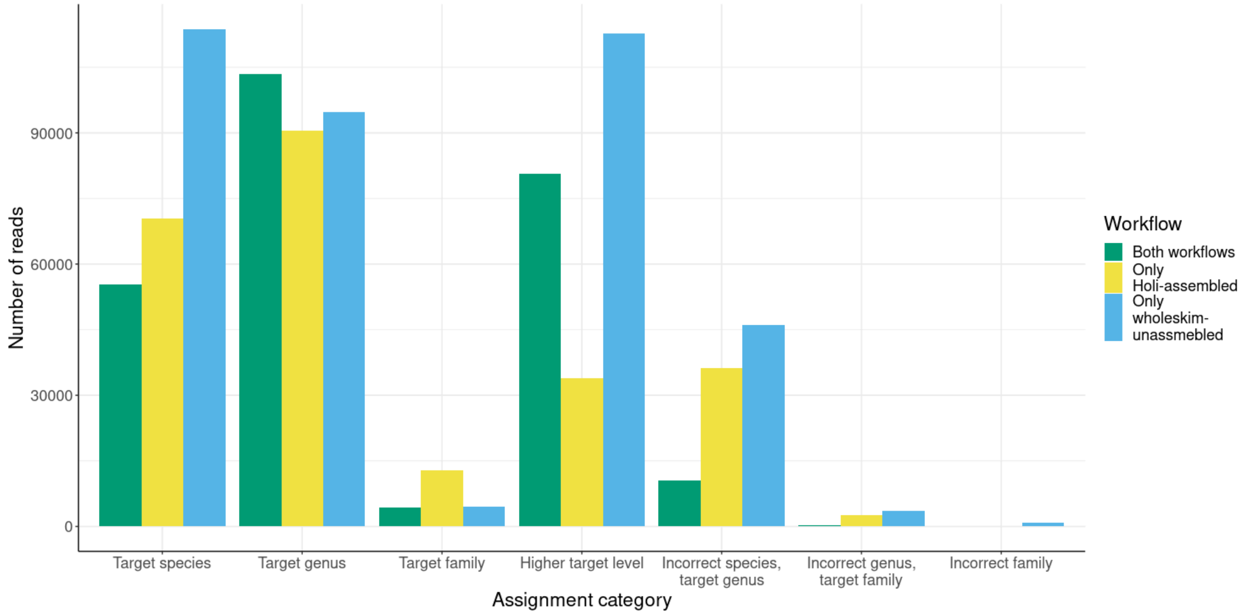


Figure 5. The overlap of simulated reads identified by the Holi-assembled and wholeskim-unassembled workflows. The set of reads is composed of all nine test species present in the PhyloNorway reference database.

Additionally, we investigated the effects of two different forms of reference database completeness on the accuracy of assignments; taxonomic completeness and genomic completeness. By simulating and assigning reads from *Salix retusa*, which species has no representative in PhyloNorway, and *Thesium alpinum,* which genus has no representative in PhyloNorway, we demonstrated that the lack of taxonomic reference database completeness does not create false positives originating from other plant DNA. Less than 0.2% of these simulated reads were assigned to species outside the target genera. In the case of *Salix retusa*, 24.9% of the reads were still correctly identified at genus-level, suggesting that missing species in the reference database are not a cause for false negatives in well-represented genera like *Salix*. However, taxonomic completeness is obviously still critical to prevent false negatives in the case of *Thesium*. Increasing the genomic completeness of a taxon in the reference database results in an increasing accuracy of assignments, until roughly the size of the reference genome where it continues to increase accuracy, but with diminishing returns (Figure 6). The continually increasing accuracy even after 1x coverage could be a result of sequencing errors in the reference genome skim allowing for some "fuzzy" matching to k-mers. This process does not appear to cause large numbers of false positive assignments to the genome skim species in question as *V.*

uliginosum has the highest depth of coverage in our reference dataset (at roughly 11x), but does not appear as a misassignment in the simulated sedaDNA dataset assignments.

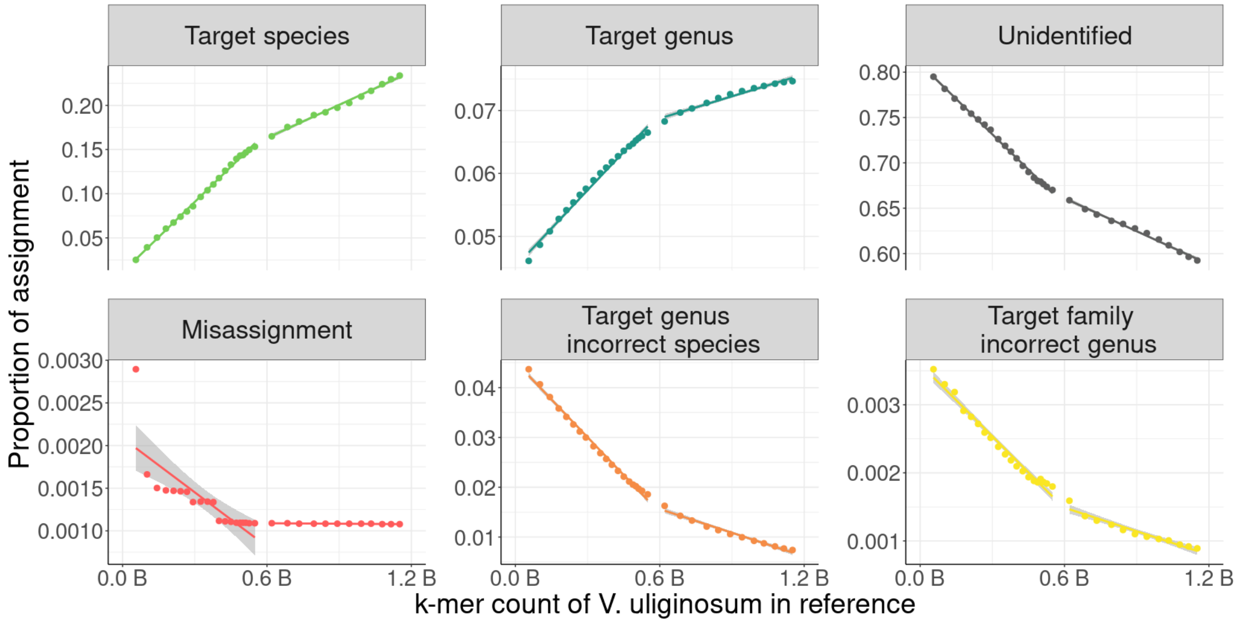


Figure 6. The taxonomic assignment of 200k simulated *Vaccinium uliginosum* reads as increasing amounts of unique k-mers of *V. uliginosum* are added to the reference database. Note the varying y-axis scales. Lines of best fit are added to points before and after estimated 1x coverage of the *V. uliginosum* genome (~0.6 B k-mers).

Finally, we demonstrated the effectiveness of *wholeskim*-unassembled by annotating the metagenomes produced from three ancient sedaDNA samples from northern Norway (Komatsu et al in prep). *Wholeskim*-unassembled was able to annotate between 2.1 - 5.6% of the overall dataset to Viridiplantae, 2.48x more reads than *Holi*-assembled and considerably more than previous studies examining vegetation in lake sedaDNA (Parducci et al. 2019; Courtin et al. 2022). Previous applications of the *Holi*-assembled pipeline annotated 1.7% (Wang et al. 2021) and 2.9% (Kjær et al. 2022) of total reads to Viridiplantae, largely consistent with the 0.8 - 2.3% annotation percentage of this workflow on our simulated data. While the taxa identified by each workflow were largely identical, *wholeskim*-unassembled recovered more reads for each taxon than *Holi*-assembled.

While collapsing all assignments to genus-level minimizes false positives to 0.27% of the data, the resolution of annotation is actually variable among taxa. The *Vaccinium* genus is well-differentiated at genus-level, with the majority of reads being assigned to the target species (Supplementary Figure 1). However, the simulated query reads from *Betula nana* are largely assigned to genus-level with nearly even amounts assigned to each constituent species (Supplementary Figure 1). By handling assignments on a per-taxa basis, species-level resolution could be attained for some taxa. If a study aims to perform genome assembly or to maximize the number of reads annotated, using both *wholeskim*-unassembled and *Holi*-assembled identifies the largest pool of target reads. This information could be used for population-level analyses if that diversity is represented in the reference database (Bohmann et al. 2020). However, *wholeskim*-unassembled run independently is sufficient for detecting which taxa are present in a sample.

Comparison of target enrichment and shotgun sequencing sedaDNA (Paper IV)

A final set of surface samples from 20 lakes was processed with target enrichment and shotgun sequencing. From the shotgun sequenced data analyzed with the *wholeskim*-unassembled pipeline, we identified 33 taxa across the 20 samples with the most abundant in read counts being *Hippuris, Utricularia, and Betula* appearing in 11, 14, and 8 samples respectively. The number of taxa detected per lake ranged from zero to ten. Several of these taxa have little or no known distribution in northern Norway (*Elatine, Lemna*, and *Zannichellia*). Competitively matching the results from non-decontaminated genome skims to bacteria, fungi, and algae removed a variable percentage of reads from different taxa ranging from 4.2% (*Myriophyllum*) to 72.6% (*Cochlearia*) (Supplementary Figure 2). All taxa with >90% of reads conserved in this filtering have known populations in northern Norway suggesting that some genome skims potentially contain algal or bacterial DNA that was not filtered during decontamination. To further investigate these potential false positives, we mapped the identified reads of *Alnus* and *Utricularia* to the three largest assembled chromosomes of species in their respective genera. While 50.3% of the *Alnus* reads mapped fairly evenly across the chromosomes, only 0.15% of the *Utricularia* reads mapped suggesting that these reads are misassigned. Mapping the shotgun reads to the assembled chloroplast database produced for the target enrichment pipeline resulted in aquatics such as *Lemna*, *Utricularia*, and *Zannichellia* disappearing from the dataset (Supplementary Figure 3). High proportions of contaminant reads were initially detected in the *Utricularia* and other aquatic plant species' reference genome skims and subsequently removed (Figure 4), but they likely signal the presence of other bacteria, fungi, and algae that are not represented in RefSeq and are unable to be filtered.

We identified 21 taxa in the target enrichment data across the 20 samples. One sample, Paulan Járvi (PAUL), had two taxa with dominant read counts; *Callitriche* (22,250) and its family Plantaginaceae (23,267) which compose 59.7% of the entire Embryophyta target enrichment dataset. The number of taxa detected in each lake ranged from zero, with five lakes having fewer than 5 reads identified to Embryophyta taxa, to nine taxa. Ten of the 21 taxa identified in the target enrichment dataset do not have a direct counterpart in the shotgun dataset; Betulaceae, Ericaceae, Ericoideae, Cyproideae, Saliceae, Potamogetonaceae, *Dryopteris, Dryas, Equisetum,* and *Stuckenia*, although constituent genera of the first six clades are found in the shotgun dataset. Within each lake sample, there is very little overlap in taxa identified by the two methods and consequently, there is little clustering by lake in the nMDS ordination, except for those lakes where shotgun sequencing did not detect any aquatic taxa; Langfjordvannet and Sierravennet (Figure 7).

Of the total 38 taxa identified across the two workflows, nine taxa were not previously detected by vegetation or metabarcoding in the catchments (Alsos et al. 2018). From these nine taxa, three have distributions in northern Norway (*Braya, Cochlearia*, and *Struthiopteris*), while the other six have not been recorded in northern Norway or are extremely rare (*Elatine, Lemna, Minuartia, Mononeuria, Scilla,* and *Mononeuria*) (Artsdatabanken.no). These six taxa absent from northern Norway were only detected by shotgun sequencing. We hypothesize that the discordance between taxa identified by shotgun sequencing and target enrichment from the same samples could originate from a combination of several factors. The DNA used by each workflow was extracted from the same homogenized subsample using identical protocols, but at two different time points. The library preparation method used with target enrichment targeted only double-stranded molecules while the method used for shotgun sequencing recovered both double- and single-stranded molecules (Gansauge et al. 2017). Additionally, biologically active surface sediment samples likely contain large proportions of bacteria, fungi, and algae that either mask or can be misidentified as plant DNA from the catchment (Capo et al. 2021).

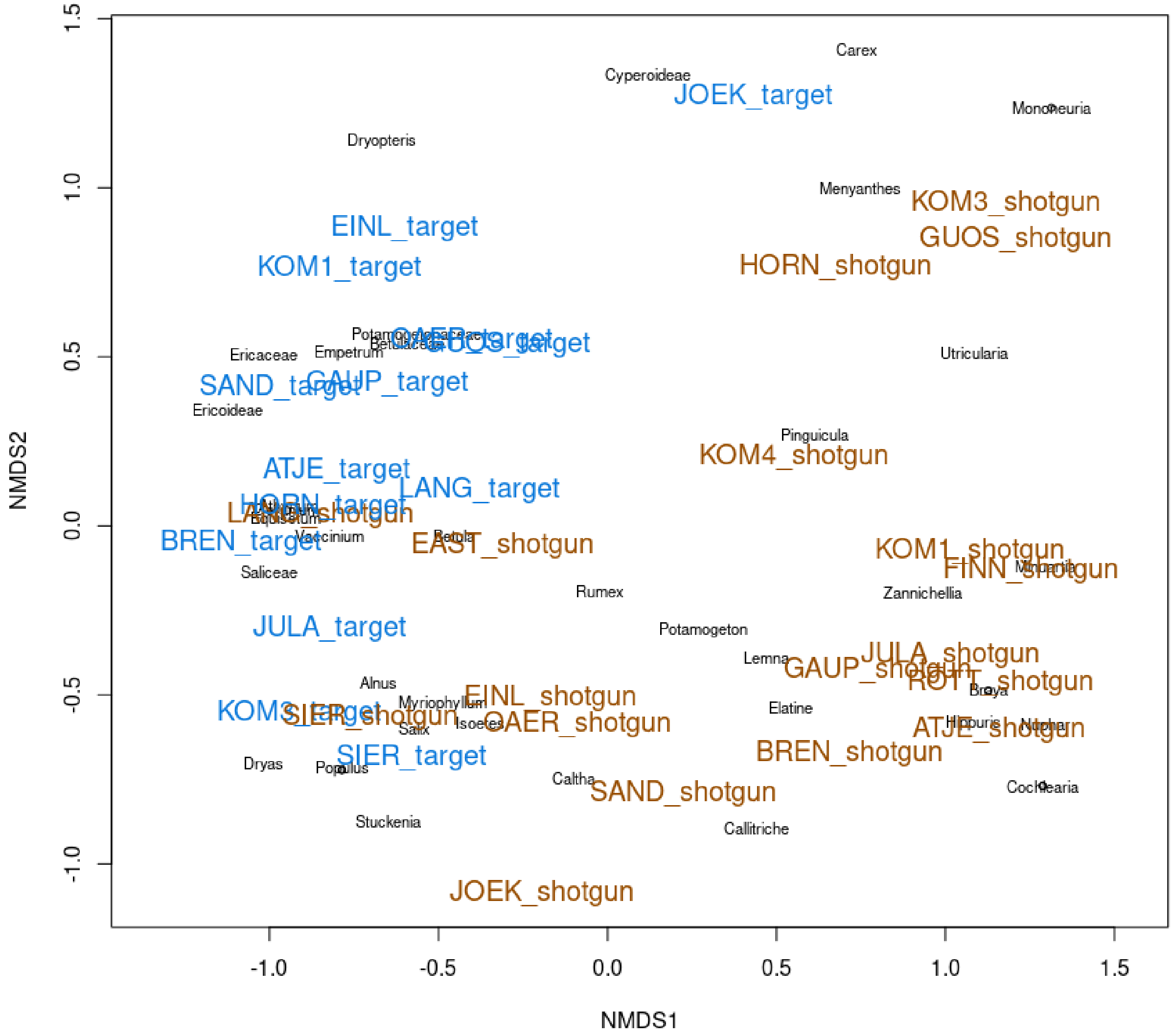


Figure 7. Nonmetric multidimensional scaling (NMDS) representing the vegetation communities detected at each lake by target enrichment and shotgun sequencing. Note that the five lakes without any target enrichment reads are not included, as well as Paulan Javri which is a distant outlier to all other lakes.

In initial tests of *wholeskim* using simulated data, we included some bacterial and human reads alongside the plant reads at equal concentration for annotation. These known “contaminants” were never falsely annotated as Viridiplantae so they were not reported in the tests. However, either including a more realistic proportion of these off-target reads (>0.98, (Wang et al. 2021) or using the specific bacteria and algae found as contaminants in the *Utricularia* genome skim to simulate off-target reads could potentially result in the false positive rates observed in the 20 lakes shotgun sequenced dataset. Mapping the shotgun dataset to the assembled chloroplasts produces a more accurate list of taxa present with few suspected false positives, but at the cost of identifying < 3% of the total reads that the *wholeskim*-unassembled pipeline was able to annotate.

Conclusion

Vegetation reconstructions from lake sediment records provide valuable ecological information as shown by Paper I. However, the metabarcoding data used in this study represents just a small fraction of the total sedaDNA content preserved in these archives. In Papers II and III we demonstrate how this previously inaccessible information can be obtained by two distinct methods. Multiplex PCR allows for intraspecific variation to be reliably detected with minimal resources as shown in Paper II. This information can shed light on routes of plant migrations and pinpoint past population turnover events. The majority of sedaDNA content is nuclear DNA that is undetected by most metabarcoding primers and mapping methods to incomplete databases. In Paper III, we demonstrate how the *wholeskim* pipeline can annotate metagenomes using unassembled genome skims to access this nuclear content, allowing for genome-wide sites to be simultaneously recovered for all taxa in a sample. Paper IV highlights some potential pitfalls of this nuclear metagenomic annotation, but none of them are insurmountable. Small amounts of contamination in genome skims can result in relatively large proportions of false positive annotation even with partial assembly of the genome skims. However, nuclear annotation of shotgun sequenced data remains a potentially powerful tool since it can accurately recover orders of magnitude more sequences than other methods for some taxa. With the goal of detecting plant diversity, metabarcoding remains the optimal method in terms of resource efficiency and being able to recover the largest number of taxa. As reference databases continue to expand in the future with full genome assemblies and sequencing costs continue to decrease, shotgun sequencing will likely surpass this method by accurately identifying the total genomic DNA content of a sample, overcoming the biases of PCR amplification and allowing for a variety of subsequent analyses with this data.

References

Alsos, I. G., T. Engelskjøn, L. Gielly, P. Taberlet, and C. Brochmann. 2005. “Impact of Ice Ages on Circumpolar Molecular Diversity: Insights from an Ecological Key Species.” *Molecular Ecology* 14 (9): 2739–53.

Alsos, Inger Greve, Youri Lammers, Nigel Giles Yoccoz, Tina Jørgensen, Per Sjögren, Ludovic Gielly, and Mary E. Edwards. 2018. “Plant DNA Metabarcoding of Lake Sediments: How Does It Represent the Contemporary Vegetation.” *PloS One* 13 (4): e0195403.

Alsos, Inger Greve, Sebastien Lavergne, Marie Kristine Føreid Merkel, Marti Boleda, Youri Lammers, Adriana Alberti, Charles Pouchon, et al. 2020. “The Treasure Vault Can Be Opened: Large-Scale Genome Skimming Works Well Using Herbarium and Silica Gel Dried Material.” *Plants* 9 (4). https://doi.org/10.3390/plants9040432.

Alsos, Inger Greve, Dilli Prasad Rijal, Dorothee Ehrich, Dirk Nikolaus Karger, Nigel Gilles Yoccoz, Peter D. Heintzman, Antony G. Brown, et al. 2022. “Postglacial Species Arrival and Diversity Buildup of Northern Ecosystems Took Millennia.” *Science Advances* 8 (39): eabo7434.

Andres, Kara J., Suresh A. Sethi, David M. Lodge, and Jose Andrés. 2021. “Nuclear eDNA Estimates Population Allele Frequencies and Abundance in Experimental Mesocosms and Field Samples.” *Molecular Ecology* 30 (3): 685–97.

Bloom, Burton H. 1970. “Space/time Trade-Offs in Hash Coding with Allowable Errors.” *Communications of the ACM* 13 (7): 422–26.

Bohmann, Kristine, Siavash Mirarab, Vineet Bafna, and M. Thomas P. Gilbert. 2020. “Beyond DNA Barcoding: The Unrealized Potential of Genome Skim Data in Sample Identification.” *Molecular Ecology* 29 (14): 2521–34.

Boyer, Frédéric, Céline Mercier, Aurélie Bonin, Yvan Le Bras, Pierre Taberlet, and Eric Coissac. 2016. “Obitools: A Unix-Inspired Software Package for DNA Metabarcoding.” *Molecular Ecology Resources* 16 (1): 176–82.

Capo, Eric, Charline Giguet-Covex, Alexandra Rouillard, Kevin Nota, Peter D. Heintzman, Aurèle Vuillemin, Daniel Ariztegui, et al. 2021. “Lake Sedimentary DNA Research on Past Terrestrial and Aquatic Biodiversity: Overview and Recommendations.” *Quaternary* 4 (1): 6.

Clarke, C. L., M. E. Edwards, L. Gielly, D. Ehrich, P. D. M. Hughes, L. M. Morozova, H. Haflidason, J. Mangerud, J. I. Svendsen, and I. G. Alsos. 2019. “Persistence of Arctic-Alpine Flora during 24,000 Years of Environmental Change in the Polar Urals.” *Scientific Reports* 9 (1): 19613.

Cohen, Judah, James A. Screen, Jason C. Furtado, Mathew Barlow, David Whittleston, Dim Coumou, Jennifer Francis, et al. 9/2014. “Recent Arctic Amplification and Extreme Mid-Latitude Weather.” *Nature Geoscience* 7 (9): 627–37.

Courtin, J., A. Perfumo, A. A. Andreev, and T. Opel. 2022. “Pleistocene Glacial and Interglacial Ecosystems Inferred from Ancient DNA Analyses of Permafrost Sediments from Batagay Megaslump, East Siberia.” *The Environmentalist*. https://onlinelibrary.wiley.com/doi/abs/10.1002/edn3.336.

Dalén, Love, Peter D. Heintzman, Joshua D. Kapp, and Beth Shapiro. 2023. “Deep-Time Paleogenomics and the Limits of DNA Survival.” *Science* 382 (6666): 48–53.

Freeman, C. L., L. Dieudonné, O. B. A. Agbaje, M. Žure, J. Q. Sanz, M. Collins, and K. K. Sand. 2023. “Survival of Environmental DNA in Sediments: Mineralogic Control on DNA Taphonomy.” *Environmental DNA (Hoboken, N.J.)* 5 (6): 1691–1705.

Garcés-Pastor, Sandra, Eric Coissac, Sébastien Lavergne, Christoph Schwörer, Jean-Paul Theurillat, Peter D. Heintzman, Owen S. Wangensteen, et al. 2022. “High Resolution Ancient Sedimentary DNA Shows That Alpine Plant Diversity Is Associated with Human Land Use and Climate Change.” *Nature Communications* 13 (1): 6559.

Giguet-Covex, Charline, Stanislav Jelavić, Anthony Foucher, Marina A. Morlock, Susanna A. Wood, Femke Augustijns, Isabelle Domaizon, Ludovic Gielly, and Eric Capo. 2023. "The Sources and Fates of Lake Sedimentary DNA." In *Tracking Environmental Change Using Lake Sediments: Volume 6: Sedimentary DNA*, edited by Eric Capo, Cécilia Barouillet, and John P. Smol, 9–52. Cham: Springer International Publishing.

Heintzman, Peter D., Kevin Nota, Alexandra Rouillard, Youri Lammers, Tyler J. Murchie, Linda Armbrecht, Sandra Garcés-Pastor, and Benjamin Vernot. 2023. "The Sedimentary Ancient DNA Workflow." In *Tracking Environmental Change Using Lake Sediments: Volume 6: Sedimentary DNA*, edited by Eric Capo, Cécilia Barouillet, and John P. Smol, 53–84. Cham: Springer International Publishing.

Karger, Dirk Nikolaus, Michael P. Nobis, Signe Normand, Catherine H. Graham, and Niklaus E. Zimmermann. 2021. "CHELSA-TraCE21k v1.0. Downscaled Transient Temperature and Precipitation Data since the Last Glacial Maximum." *Clim. Past*. https://doi.org/10.5194/cp-2021-30.

Kim, Daehwan, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. 2016. "Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences." *Genome Research* 26 (12): 1721–29.

Kjær, Kurt H., Mikkel Winther Pedersen, Bianca De Sanctis, Binia De Cahsan, Thorfinn S. Korneliussen, Christian S. Michelsen, Karina K. Sand, et al. 2022. "A 2-Million-Year-Old Ecosystem in Greenland Uncovered by Environmental DNA." *Nature* 612 (7939): 283–91.

Lammers, Youri, Peter D. Heintzman, and Inger Greve Alsos. 2021. "Environmental Palaeogenomic Reconstruction of an Ice Age Algal Population." *Communications Biology* 4 (1): 220.

Langdon, W. B. 2015. "Performance of Genetic Programming Optimised Bowtie2 on Genome Comparison and Analytic Testing (GCAT) Benchmarks." *BioData Mining* 8 (1): 1.

Lemane, Téo, Nolan Lezzoche, Julien Lecubin, Eric Pelletier, Magali Lescot, Rayan Chikhi, and Pierre Peterlongo. 2024. "Indexing and Real-Time User-Friendly Queries in Terabyte-Sized Complex Genomic Datasets with Kmindex and ORA." *Nature Computational Science* 4 (2): 104–9.

Mamanova, Lira, Alison J. Coffey, Carol E. Scott, Iwanka Kozarewa, Emily H. Turner, Akash Kumar, Eleanor Howard, Jay Shendure, and Daniel J. Turner. 2010. "Target-Enrichment Strategies for next-Generation Sequencing." *Nature Methods* 7 (2): 111–18.

Murchie, Tyler J., Melanie Kuch, Ana T. Duggan, Marissa L. Ledger, Kévin Roche, Jennifer Klunk, Emil Karpinski, et al. 2021. "Optimizing Extraction and Targeted Capture of Ancient Environmental DNA for Reconstructing Past Environments Using the PalaeoChip Arctic-1.0 Bait-Set." *Quaternary Research* 99 (January): 305–28.

Nesje, Atle. 1992. "A Piston Corer for Lacustrine and Marine Sediments." *Arctic and Alpine Research* 24 (3): 257–59.

Parducci, L., I. G. Alsos, and P. Unneberg. 2019. "Shotgun Environmental DNA, Pollen, and Macrofossil Analysis of Lateglacial Lake Sediments from Southern Sweden." *Frontiers in Ecology and the Environment*. https://www.frontiersin.org/articles/10.3389/fevo.2019.00189/full.

Parducci, Laura, Keith D. Bennett, Gentile Francesco Ficetola, Inger Greve Alsos, Yoshihisa Suyama, Jamie R. Wood, and Mikkel Winther Pedersen. 2017. "Ancient Plant DNA in Lake Sediments." *The New Phytologist* 214 (3): 924–42.

Pedersen, Mikkel Winther, Bianca De Sanctis, Nedda F. Saremi, Martin Sikora, Emily E. Puckett, Zhenquan Gu, Katherine L. Moon, et al. 2021. "Environmental Genomics of Late Pleistocene Black Bears and Giant Short-Faced Bears." *Current Biology: CB* 31 (12): 2728–36.e8.

Pedersen, Mikkel W., Anthony Ruter, Charles Schweger, Harvey Friebe, Richard A. Staff, Kristian K. Kjeldsen, Marie L. Z. Mendoza, et al. 2016. "Postglacial Viability and Colonization in North America's Ice-Free Corridor." *Nature* 537 (7618): 45–49.

Rijal, Dilli P., Peter D. Heintzman, Youri Lammers, Nigel G. Yoccoz, Kelsey E. Lorberau, Iva Pitelkova, Tomasz Goslar, et al. 2021. "Sedimentary Ancient DNA Shows Terrestrial Plant Richness Continuously Increased over the Holocene in Northern Fennoscandia." *Science Advances* 7 (31). https://doi.org/10.1126/sciadv.abf9557.

Robidou, Lucas, and Pierre Peterlongo. 2021. "Findere: Fast and Precise Approximate Membership Query." *bioRxiv*. https://doi.org/10.1101/2021.05.31.446182.

Schulte, Luise, Nadine Bernhardt, Kathleen Stoof-Leichsenring, Heike H. Zimmermann, Luidmila A. Pestryakova, Laura S. Epp, and Ulrike Herzschuh. 2021. “Hybridization Capture of Larch (Larix Mill.) Chloroplast Genomes from Sedimentary Ancient DNA Reveals Past Changes of Siberian Forest.” *Molecular Ecology Resources* 21 (3): 801–15.

Soininen, Eeva M., Gilles Gauthier, Frédéric Bilodeau, Dominique Berteaux, Ludovic Gielly, Pierre Taberlet, Galina Gussarova, et al. 2015. “Highly Overlapping Winter Diet in Two Sympatric Lemming Species Revealed by DNA Metabarcoding.” *PloS One* 10 (1): e0115335.

Sønstebø, J. H., L. Gielly, A. K. Brysting, R. Elven, M. Edwards, J. Haile, E. Willerslev, et al. 2010. “Using next-Generation Sequencing for Molecular Reconstruction of Past Arctic Vegetation and Climate.” *Molecular Ecology Resources* 10 (6): 1009–18.

Taberlet, P., A. Bonin, L. Zinger, and E. Coissac. 2018. “Environmental DNA: For Biodiversity Research and Monitoring.” https://books.google.ca/books?hl=en&lr=&id=1e9IDwAAQBAJ&oi=fnd&pg=PP1&ots=UY8TqcnfoR&sig=8kUMa4OFtZC1Z2n5fT7MV7cTuSo.

Taberlet, P., E. Coissac, F. Pompanon, L. Gielly, C. Miquel, A. Valentini, T. Vermat, G. Corthier, C. Brochmann, and E. Willerslev. 2007. “Power and Limitations of the Chloroplast trnL (UAA) Intron for Plant DNA Barcoding.” *Nucleic Acids Research* 35 (3): e14–e14.

Tyler, Torbjörn, Lina Herbertsson, Johan Olofsson, and Pål Axel Olsson. 2021. “Ecological Indicator and Traits Values for Swedish Vascular Plants.” *Ecological Indicators* 120 (January): 106923.

Wang, Yucheng, Mikkel Winther Pedersen, Inger Greve Alsos, Bianca De Sanctis, Fernando Racimo, Ana Prohaska, Eric Coissac, et al. 2021. “Late Quaternary Dynamics of Arctic Biota from Ancient Environmental Genomics.” *Nature* 600 (7887): 86–92.

Willerslev, Eske, John Davison, Mari Moora, Martin Zobel, Eric Coissac, Mary E. Edwards, Eline D. Lorenzen, et al. 2014. “Fifty Thousand Years of Arctic Vegetation and Megafaunal Diet.” *Nature* 506 (7486): 47–51.

Wittmeier, Hella E., Jostein Bakke, Kristian Vasskog, and Mathias Trachsel. 2015. “Reconstructing Holocene Glacier Activity at Langfjordjøkelen, Arctic Norway, Using Multi-Proxy Fingerprinting of Distal Glacier-Fed Lake Sediments.” *Quaternary Science Reviews* 114 (April): 78–99.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. “Improved Metagenomic Analysis with Kraken 2.” *Genome Biology* 20 (1): 257.

Appendix

Locality	Sub locality	Plot	Date	Elevation (m)	Latitude	Longitude	Habitat
Kristiansand	Åseral	Sostelifjellet	20200528	700	58.61546	7.37654	Fjellhei
Kristiansand	Åseral	Joneset	20200529	265	58.64898	7.44189	Meadow
Kristiansand	Suleskar	Suleskarvegen	20200530	821	59.02182	6.97571	Fjellhei
Kristiansand	Ljosland	Ljosland	20200529	517	58.78833	7.35426	Elveleie
Karmøy	Øverland	Dalvanuten top	20200531	591	59.371167	5.843984	Fjellhei
Karmøy	Øverland	Dalvanuten	20200531	520	59.369724	5.854202	Fjellhei
Karmøy	Øverland	Leirå	20200531	2	59.344409	5.875522	Strandeng
Karmøy	Vikedal	Horganuten	20200601	850	59.521144	6.044061	Fjellhei
Karmøy	Vikedal	Ilstveitvegen	20200601	400	59.510353	6.022557	Mixed forest/Meadow
Andøya	Bleik	Storvatnet	20200616	5	69.262497	15.960031	Alpine
Andøya	Æråsvatnet	Øvre Æråsvatnet	20200616	40	69.258115	16.04379	Mire
Andøya	Andenes	Andhauet	20200616	250	69.29302	16.03875	Alpine
Tromsø	Tønsvika	Høgmelelva	20200717	120	69.73561	19.2167	River canyon
Tromsø	Fløya	Fløya	20200716	430	69.62767	19.00537	Alpine
Kirkenes	Jarfjorden	Pandurneset	20200619	5	69.667766	30.321468	Mixed salix/betula forest close to meadow and beach
Kirkenes	Lyngberget	Lyngberget	20200619	70	69.748551	30.127171	Mire/Betula pubesence landscape
Kirkenes	Geithøgda	Geithøgda	20200619	280	69.810257	30.241749	Fjellhei
Kirkenes	Øretoppen	Øretoppen	20200619	465	69.819497	30.305994	Tørr fjellhei

Nordkapp	Nordkapp	Plateu	20200620	300	71.164417	25.800533	High alpine
Nordkapp	Magerøya	Honningsvåg	20200621	15	70.994229	25.942215	Alpine

Supplementary Table 1. The sites for modern plant tissue sampling.

Species	Priority	Nordkapp	Kirkenes	Andøya	Karmøy	Kristiansand
Alnus incana	1	0	10	10	7	6
Angelica archangelica	1	8	0	0	10	0
Avenella flexuosa	1	10	1	4	10	10
Bistorta vivipara	1	10	10	10	3	0
Calluna vulgaris	1	10	10	10	10	10
Caltha palustris	1	0	10	10	5	0
Dryas octopetala	1	10	10	10	1	0
Juniperus communis	1	10	10	10	10	10
Picea abies	1	0	0	10	10	10
Pinus sylvestris	1	0	10	10	10	10
Prunus padus	1	0	8	0	10	10
Saxifraga oppositifolia	1	10	3	0	10	0
Sorbus aucuparia	1	0	10	10	10	10
Vaccinium myrtilis	1	10	10	10	10	10
Vaccinium uliginosum	1	10	10	10	10	10
Vaccinium vitis-idaea	1	10	10	10	10	10
Arctous alpina	2	5	5	5	4	3
Betula nana	2	5	5	5	5	0
Betula pubescens	2	0	5	5	5	5
Empetrum nigrum	2	5	5	5	5	5
Kalmia procumbens	2	5	5	5	5	3
Oxyria digyna	2	5	5	0	2	0
Rumex acetosa	2	0	0	0	0	2

Rumex acetosella	2	4	5	5	3	5
Rumex sp	2	0	0	0	3	0
Sibbaldia procumbens	2	0	5	0	2	0
Silene acaulis	2	5	5	5	5	0
Myositis alpestris	3	0	0	1	0	0
Salix herbaceae	3	2	2	2	3	0
SUM		134	169	162	178	129

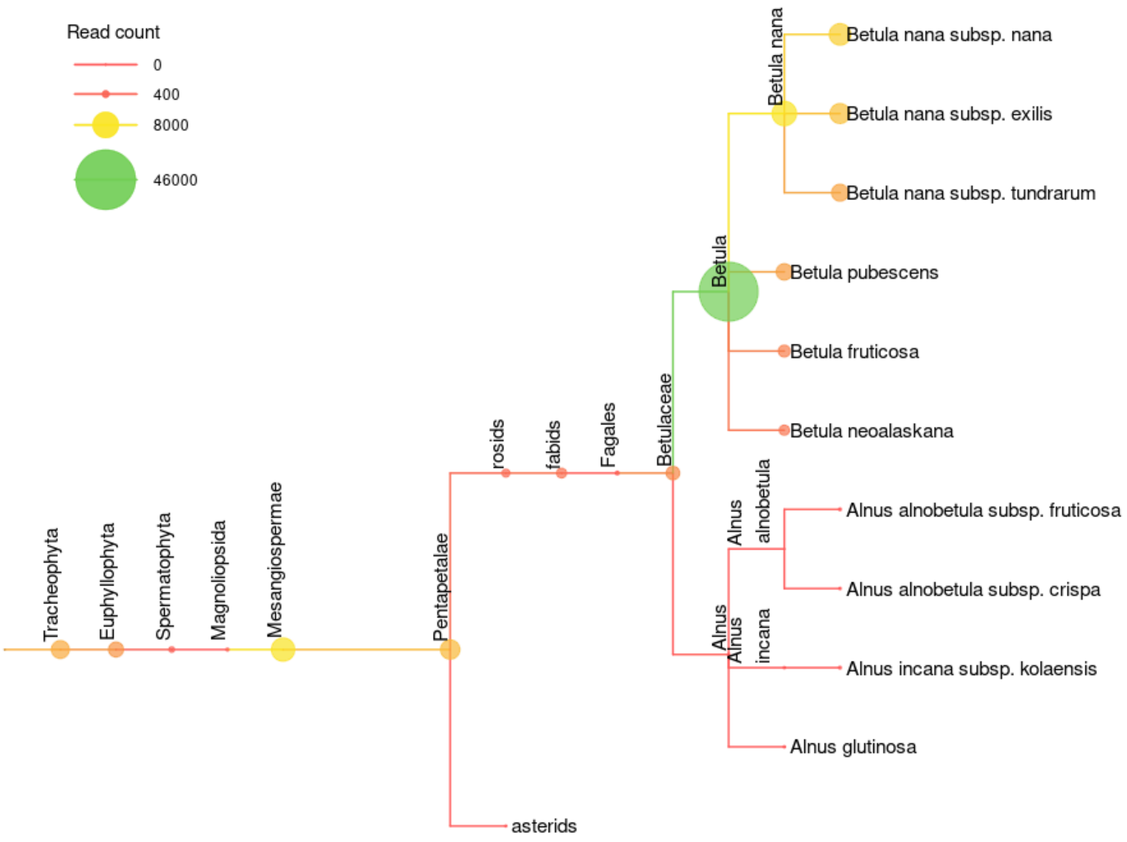
Supplementary Table 2. The number of specimens collected from each species in 2020 at five of the localities in Norway. The coloring represents the proportion of samples collected; green is all specimens collected (ten individuals for priority level 1 species and five individuals for priority level 2 species), yellow is less than the desired amount collected, and red is none collected.

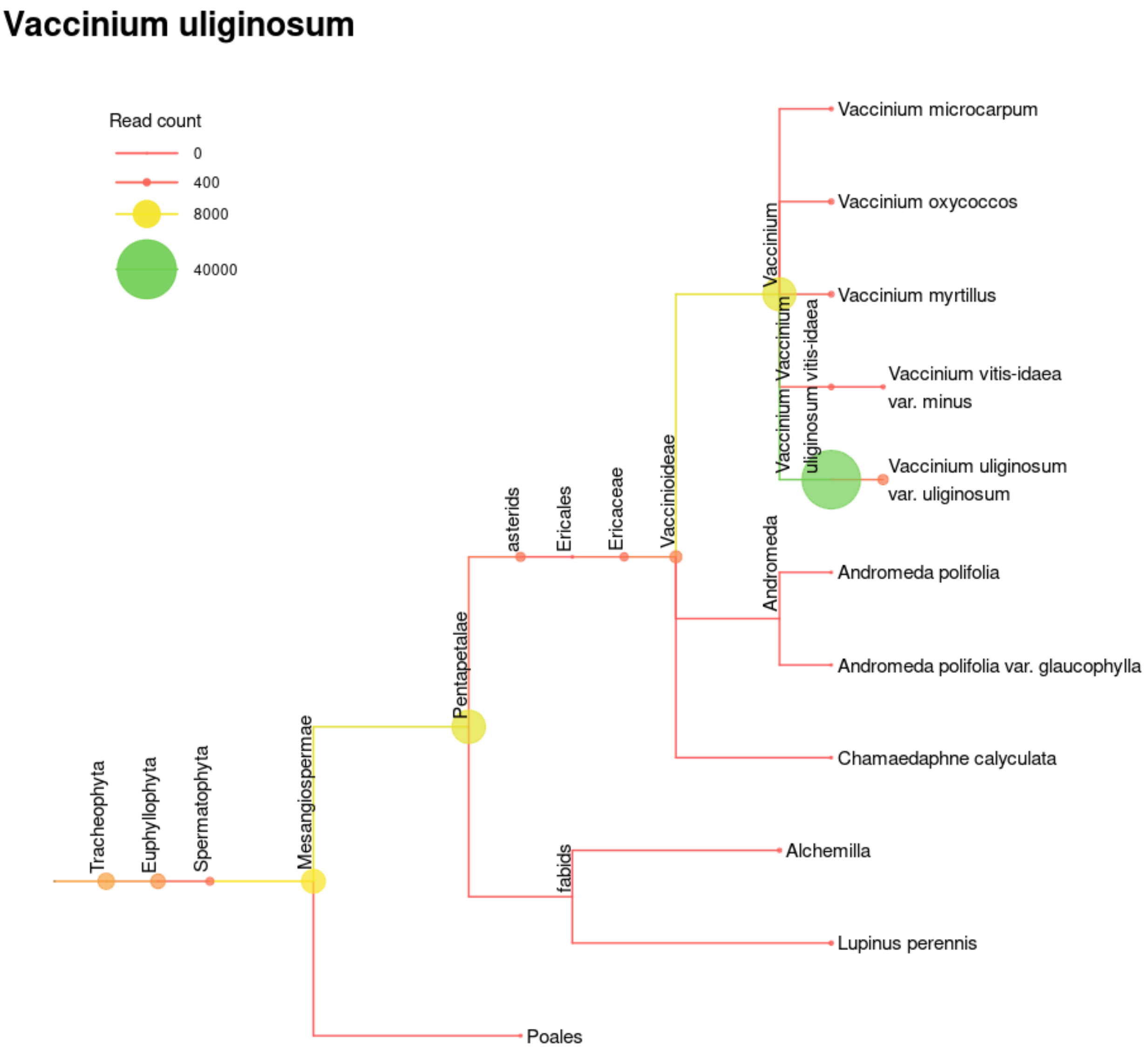
Taxa	**ID**	**Lat**	**Long**	**Fylke**	**Kommune**	**Number of read pairs**
Alnus_incana	P03-12	69.7408	30.11419	Troms og Finnmark	Lyngberget	3519298
Alnus_incana	P06-12	59.10094	7.53567	Agder	Rysstad	2650781
Angelica_archangelica	P01-01	71.164417	25.800533	Troms og Finnmark	Nordkapp	5940265
Avenella_flexuosa	P06-02	58.61546	7.37654	Agder	Sostelifjellet	4240204
Betula_nana	P05-01	59.344409	5.875522	Rogaland	Ieira	4964937
Betula_pubescens	P01-02	70.994229	25.942215	Troms og Finnmark	Honningsvåg	4553653
Bistorta_vivipara	P04-04	69.258115	16.04379	Nordland	Andøya	3178731
Calluna_vulgaris	P03-11	69.810257	30.241749	Troms og Finnmark	Geithøgda	1450132
Calluna_vulgaris	P06-11	58.61546	7.37654	Agder	Sostelifjellet	2079472
Caltha_palustris	P03-05	69.66524	30.32136	Troms og Finnmark	Pandurneset	9061419

Caltha_palustris	P04-05	69.25878	16.0428	Nordland	Andøya	7151936
Caltha_palustris	P05-05	59.344409	5.875522	Rogaland	Ieira	8706303
Dryas_octopetala	P01-06	71.164417	25.800533	Troms og Finnmark	Nordkapp	6212891
Dryas_octopetala	P03-06	69.810257	30.241749	Troms og Finnmark	Geithøgda	4540652
Dryas_octopetala	P04-06	69.25878	16.0428	Nordland	Andøya	3226899
Dryas_octopetala	P05-06	59.344409	5.875522	Rogaland	Ieira	5110327
Picea_abies	P06-13	59.10094	7.53567	Agder	Rysstad	2100610
Pinus_sylvestris	P06-14	59.10094	7.53567	Agder	Rysstad	2909535
Prunus_padus	P03-15	69.810257	30.241749	Troms og Finnmark	Geithøgda	4172184
Prunus_padus	P06-15	59.10094	7.53567	Agder	Rysstad	1585312
Salix_herbacea	P01-02	71.164417	25.800533	Troms og Finnmark	Nordkapp	4525627
Vaccinium_myrtillus	P06-07	59.10094	7.53567	Agder	Rysstad	7424312
Vaccinium_uliginosum	P01-08	71.164417	25.800533	Troms og Finnmark	Nordkapp	6851900
Vaccinium_uliginosum	P06-08	59.10094	7.53567	Agder	Rysstad	3101972
Vaccinium_vitis-idaea	P01-09	71.164417	25.800533	Troms og Finnmark	Nordkapp	4106246

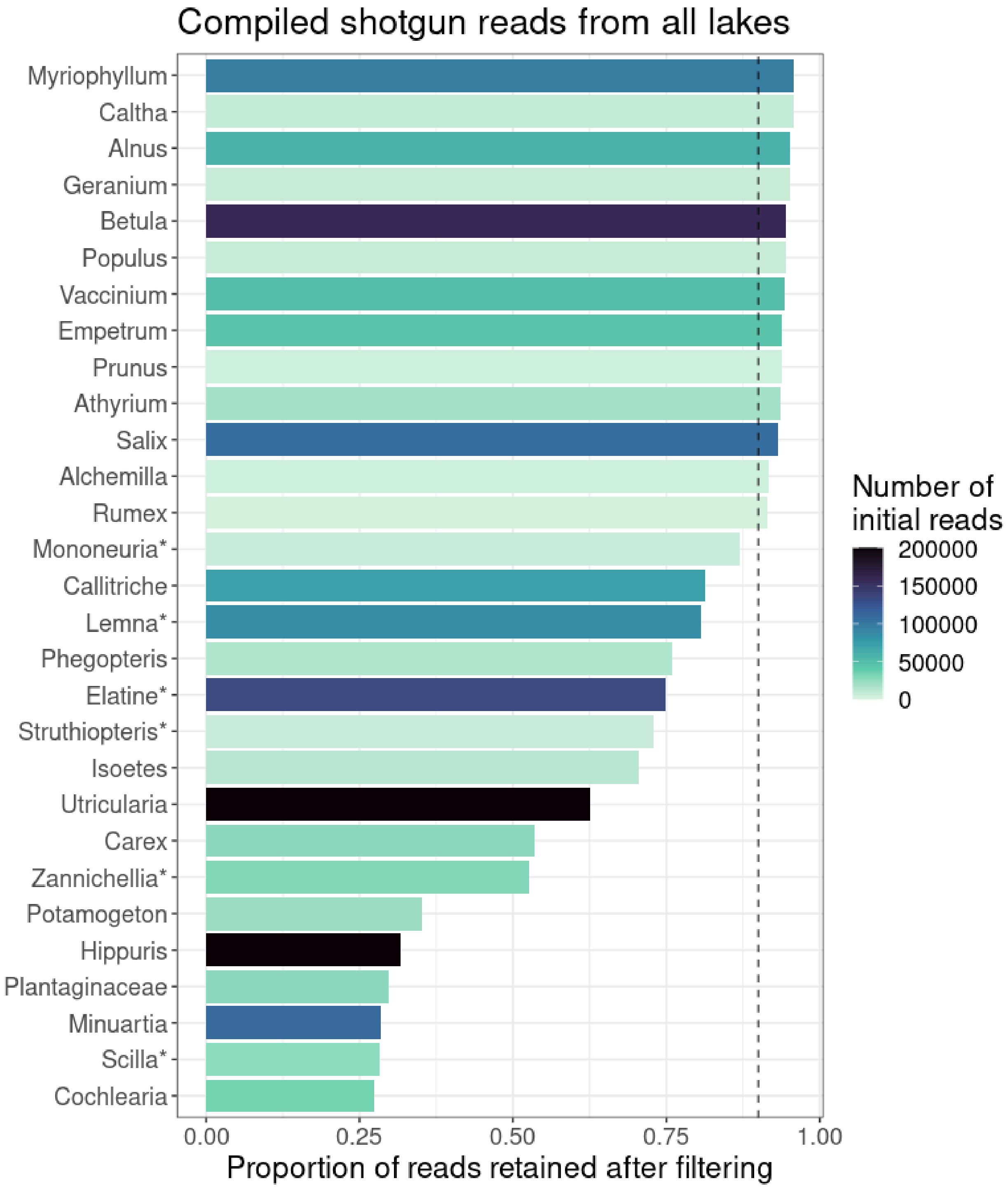
Supplementary Table 3. Genome skims produced for this thesis from modern plant tissue collected in Supplementary Table 2.

Betula nana subsp. nana

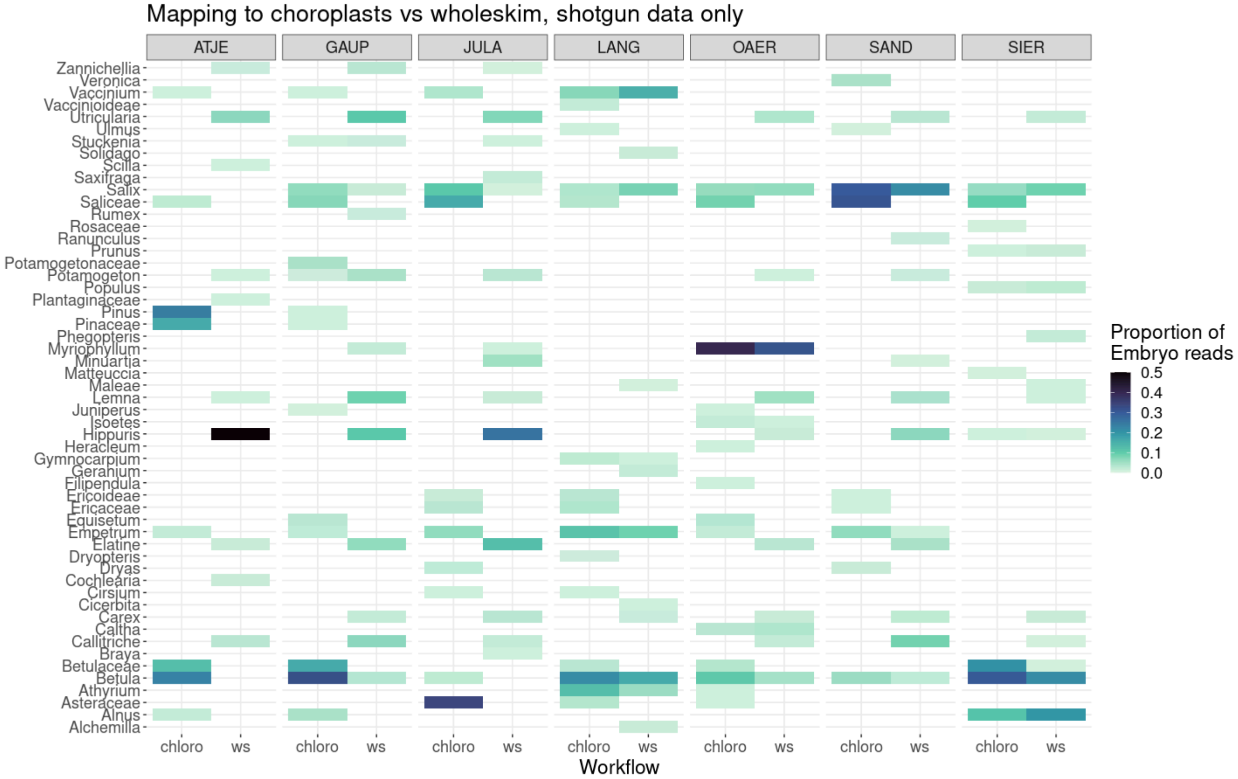




Supplementary Figure 1. Phylogenetic tree of assigned simulated reads for *Betula nana* and *Vaccinium uliginosum* using the *wholeskim*-unassembled workflow. Both the color and size of the node are proportional to the number of reads assigned to that taxon.



Supplementary Figure 2. The proportion of shotgun reads retained for each taxon across the 20 lakes when filtering for bacteria, algae, and fungal reads. The dashed line denotes the 90% retained threshold where all taxa are supported by metabarcoding and local vegetation surveys. Note that *Hippuris* has an initial read count of 1.78 million, an order of magnitude larger than the second most abundant taxon, *Utricularia*, with 269,783 reads. Taxa not recorded (e.g. *Mononeuria*) or rare (e.g. *Elatine*) in N Norway are denoted by an asterisk.



Supplementary Figure 3. A heatmap of the taxa detected through shotgun data from seven lakes when mapping to the chloroplast (chloro) and annotated by *wholeskim* using the unassembled genome skims (ws).

Papers I - IV

quaternary

MDPI

Article

Sedimentary Ancient DNA Reveals Local Vegetation Changes Driven by Glacial Activity and Climate

Lucas D. Elliott [1,*], Dilli P. Rijal [1], Antony G. Brown [1], Jostein Bakke [2], Lasse Topstad [1], Peter D. Heintzman [1,3,4] and Inger G. Alsos [1,*]

[1] The Arctic University Museum of Norway, UiT the Arctic University of Norway, 9006 Tromso, Norway
[2] Department of Earth Science and Bjerknes Centre for Climate Research, University of Bergen, 5007 Bergen, Norway
[3] Centre for Palaeogenetics, Svante Arrhenius väg 20C, 106 91 Stockholm, Sweden
[4] Department of Geological Sciences, Stockholm University, 106 91 Stockholm, Sweden
* Correspondence: lucas.elliott@uit.no (L.D.E.); inger.g.alsos@uit.no (I.G.A.)

Abstract: Disentangling the effects of glaciers and climate on vegetation is complicated by the confounding role that climate plays in both systems. We reconstructed changes in vegetation occurring over the Holocene at Jøkelvatnet, a lake located directly downstream from the Langfjordjøkel glacier in northern Norway. We used a sedimentary ancient DNA (*sed*aDNA) metabarcoding dataset of 38 samples from a lake sediment core spanning 10,400 years using primers targeting the P6 loop of the *trn*L (UAA) intron. A total of 193 plant taxa were identified revealing a pattern of continually increasing richness over the time period. Vegetation surveys conducted around Jøkelvatnet show a high concordance with the taxa identified through *sed*aDNA metabarcoding. We identified four distinct vegetation assemblage zones with transitions at ca. 9.7, 8.4 and 4.3 ka with the first and last mirroring climatic shifts recorded by the Langfjordjøkel glacier. Soil disturbance trait values of the vegetation increased with glacial activity, suggesting that the glacier had a direct impact on plants growing in the catchment. Temperature optimum and moisture trait values correlated with both glacial activity and reconstructed climatic variables showing direct and indirect effects of climate change on the vegetation. In contrast to other catchments without an active glacier, the vegetation at Jøkelvatnet has displayed an increased sensitivity to climate change throughout the Middle and Late Holocene. Beyond the direct impact of climate change on arctic and alpine vegetation, our results suggest the ongoing disappearance of glaciers will have an additional effect on plant communities.

Keywords: *sed*aDNA; glaciers; vegetation reconstruction; climate change; Norway; Holocene

check for updates

Citation: Elliott, L.D.; Rijal, D.P.; Brown, A.G.; Bakke, J.; Topstad, L.; Heintzman, P.D.; Alsos, I.G. Sedimentary Ancient DNA Reveals Local Vegetation Changes Driven by Glacial Activity and Climate. *Quaternary* **2023**, *6*, 7. https://doi.org/10.3390/quat6010007

Academic Editor: James B. Innes

Received: 1 November 2022
Revised: 17 December 2022
Accepted: 20 December 2022
Published: 7 January 2023

1. Introduction

Climate change affects both glaciers and arctic-alpine vegetation through variation in temperature and precipitation. However, glaciers can also have a more direct impact on alpine vegetation as they affect local climate, soil moisture, and soil disturbance [1]. Colonization of post-glacial landscapes is a heterogeneous process with many factors determining the resulting vegetation [2]. Early stages of succession in glacier forefields are largely controlled by geomorphic processes with the unstable, paraglacial landscape limiting vegetation to a few pioneer species [3] whereas later stages of succession are reliant on autogenic processes (e.g., plant colonization, chemical/physical weathering, and soil accumulation) [4].

After the Last Glacial Maximum (~20 thousand years ago, ka), during which northern Fennoscandia was almost completely covered by the Scandinavian Ice Sheet, large areas became ice-free 15 to 14 ka [5]. Deglaciation accelerated at the onset of the Holocene (11.7 ka) when temperatures quickly rose and the continuous ice sheet was broken up into valley/fjord glaciers, which became smaller or entirely absent by the Middle Holocene

(8.3 to 4.2 ka) [6–8]. Several abrupt cold events during this period caused the temporary readvancement or reformation of glaciers, but northern Fennoscandia was almost entirely glacier-free until the Late Holocene (4.2 ka to present) [9,10]. Many valley/fjord glaciers began to reform and underwent rapid fluctuations during the Late Holocene due to predominantly cool but variable temperatures [7,8].

Investigating past vegetation changes in response to a varying climate has been accomplished using pollen and macrofossils preserved in lake sediments [11,12]. However, several properties of pollen and macrofossils complicate the use of these proxies for vegetation reconstruction in the Arctic. The long-distance dispersal of pollen from certain wind-pollinated taxa (e.g., *Alnus alnobetula*, *Pinus sylvestris*) can present false positive signals of local taxa [13]. Many dominant arctic-alpine taxa rely on insect pollination (e.g., *Dryas*, *Saxifraga*) and their pollen is rarely detected in lake sediment records [14,15]. Both pollen and many macrofossils have limited taxonomic resolution and are unable to distinguish between many ecologically important taxa [16].

Studies using sedimentary ancient DNA (*seda*DNA) have demonstrated how this technique overcomes many of the limitations of pollen and macrofossils [17]. *Seda*DNA provides a more localized vegetation signal limited to organisms found within the catchment of a lake [18]. Metabarcoding techniques targeting the trnL P6 loop region of the chloroplast genome, combined with a comprehensive local reference database, are able to identify a majority of vascular plant taxa from *seda*DNA to below the genus-level [19,20]. The high taxonomic resolution and detectability of *sed*aDNA allows for taxa to be matched with trait value databases to provide a more in-depth and precise characterization of past vegetation communities [21,22].

Here, we investigate how the vegetation of Jøkelvatnet changed in relation to the Langfjordjøkelen glacier over the Holocene using *sed*aDNA metabarcoding data together with glacial activity and climatic reconstructions. We examine a subset of the data from [22] and offer a detailed reconstruction of the vegetation and new analyses addressing the causes of observed vegetation changes. vegetation composition is described using ecological indicator trait values from [21] and compared to both reconstructions of Langfjordjøkelen's glacial activity [23] and regional climate reconstructions [24] to enhance our knowledge of how climate affects plants directly (temperature and precipitation) and indirectly through glaciers (soil disturbance).

2. Materials and Methods

2.1. Study Site

Jøkelvatnet (70°10′21″ N 21°42′3″ E at 158 m a.s.l.) is a distal glacier-fed lake located in northern Norway in the valley of Sør-Tverrfjorddalen (Figure 1), immediately downstream from the 7.49 km^2 Langfjordjøkelen ice cap [25]. The surface area of the two-basin lake is ~0.13 km^2 with the deepest part located in the southern basin at ~10 m water depth [23]. The lake's catchment area is ~11 km^2 with 18% currently covered by the ice cap and the remainder primarily composed of talus slopes and steep mountain ridges (Figure 1, [23]). The Sør-Tverrfjorddalen valley was completely deglaciated from 10.0 ka until the glacier reformed at 4.1 ka with large fluctuations in glacial activity over the last two millennia [23]. The valley lies in the Caledonian province of Finnmark and is underlain by allochthonous felsic igneous and metamorphic rocks particularly gabbro and amphibolites. However, the valley floor and gentler slopes are covered by glacial deposits including moraines, till (diamicton) and screes. These deposits have a broader clast lithology which includes sands and gravel with clasts of gneiss, psammite and gabbro. The gabbro and amphibolites are highly fractured and in places highly weathered, producing regoliths and thin soils that have a relatively high nutrient index (high-moderate Mg and Ca) sensu [26].

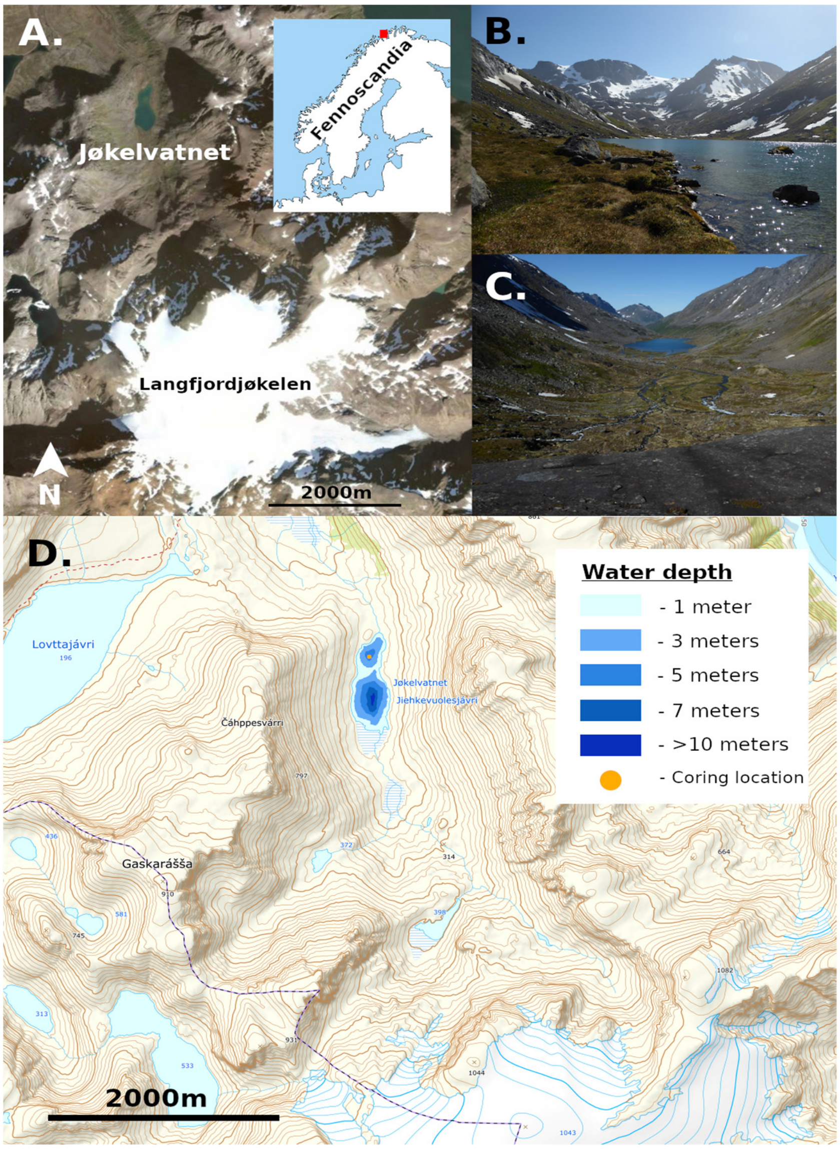


Figure 1. (**A**) Satellite image (norgeskart.no accessed 29 April 2022) of Lake Jøkelvatnet and the Langfjordjøkelen glacier with inset map of Fennoscandia. (**B**) View looking south towards the glacier from the northern outlet of the lake. (**C**) View looking north over the lake from near the glacier. Photographs taken by I. G. Alsos in July 2021. (**D**) Topographic map (norgeskart.no) of the catchment and lake bathymetry showing the coring location of JØP-112.

2.2. Vegetation Surveys

Two vascular plant surveys were conducted in the Sør-Tverrfjorddalen valley with a focus on the catchment of Jøkelvatnet during September 2020 and July 2021. For both surveys, we aimed to record all taxa growing within 2 m of the lake shore and efforts were made to identify taxa growing in the various habitats further upslope. Herbarium vouchers were collected for most taxa and deposited in the herbarium at Tromsø museum. Data from both vegetation surveys was compiled into one dataset consisting of all taxa observed in the catchment.

2.3. Coring, Age-Depth Model, and Stratigraphy

Two piston cores [27] were retrieved from the northern basin of Lake Jøkelvatnet (Figure 1D) in March 2012 and stored at 4 °C until opening [23]. The shielded northern basin was chosen for coring in order to avoid potential disturbances from the main inlet delta in the south. One 258 cm core (JØP-112) was split longitudinally and a total of 40 sediment samples spaced ~7 cm apart, were collected in the clean labs of GeoMicrobiology at the Department of Earth Science, University of Bergen, Norway for sedaDNA analysis as described in [26]. A Bayesian age-depth model was constructed by [22] using Bacon v2.3.4 [28] and the IntCal13 calibration curve [29] using 12 radiocarbon (^{14}C) dates from plant macrofossils found throughout the core [23]. Five samples fell below the basal-most radiocarbon date so extrapolation of the model was explored using different accumulation rate priors and correlating unit transitions with Lake Storvatnet (STP-112) located 5 km downstream in the same valley [22]. No signs of redeposition were detected and full details of the age-depth model construction of JØP-112 can be found in [22]. The core stratigraphy is a basal unit of dark-gray silty clay with some banding but low organic matter content. Above this is strongly laminated dark gray to olive-gray silty gyttja with a moderate organic matter content (6%). The upper 128 cm are variable generally laminated olive-gray and brown-gray silty clays with frequent visible plant remains but low loss on ignition values (2–6%). More details of the core lithostratigraphy can be found in [23].

2.4. sedaDNA Data Generation

*seda*DNA sampling, extraction, amplification, and sequencing steps were described in [22,26] and are summarized here. DNA was extracted from 0.25 to 0.35 g of sediment for each sample using a modified version of the Qiagen DNeasy PowerSoil PowerLyzer (Qiagen Norge, Oslo, Norway) protocol. Six negative controls (composed of water exposed during sediment sampling, extraction, or PCR plating) and one positive PCR control were processed in addition to the 40 samples. DNA was amplified using the "gh" primer set which targets the vascular plant trnL p6-loop locus of the chloroplast genome [19]. The gh primers were uniquely dual-tagged with an 8 or 9 base pair tag, modified from [30]. Eight PCR replicates were amplified for each sediment sample and control DNA extract in a post-PCR laboratory located in a different building from the ancient DNA lab facility at The Arctic University Museum of Norway. Two DNA libraries of these amplicons were prepared using a modified Illumina TruSeq DNA PCR-Free protocol (Illumina Inc., CA, USA) with unique dual indexes, while the magnetic bead cleanup steps were modified to retain short amplicons. The two libraries were sequenced on ~10% of 2x 150-cycle mid-output flow cell on the Illumina NextSeq platform at the Genomics Support Centre Tromsø at The Arctic University of Norway.

2.5. Bioinformatics

The *seda*DNA dataset presented here is a subset of the filtered data presented in [22]. The initial bioinformatic pipeline incorporates OBITools [31] and custom Python and R scripts to filter the data following [26]. Briefly, paired-end reads were merged using SeqPrep (https://github.com/jstjohn/SeqPrep/releases, v1.2 (accessed on 26 September 2022)) and then demultiplexed to individual samples using their 8 base pair tags, followed by the collapsing of identical reads. PCR artifacts were filtered using OBIClean and potential library swaps were corrected using a custom Python script (https://github.com/Y-Lammers/MetabarcodingFilter (accessed on 26 September 2022)). For each PCR replicate, sequences with $\leq$2 reads were discarded. Sequences represented by fewer than 10 reads or three PCR replicates in the entire regional dataset were also removed [22]. Sequences were then matched to the following four databases: (1) PhyloNorway [22], (2) a combination of 815 arctic [32] and 835 boreal [33] vascular plant taxa and 455 arctic-boreal bryophytes [34] from the circumpolar region (ArcBorBryo, n = 2280 sequences of which 1053 are unique), (3) PhyloAlps (n = 4604 specimens of 4437 taxa collected in the Alps and Carpathians, [35] (https://data.phyloalps.org/browse/ (accessed on 26 September 2022)), and (4) EMBL (release 143, n = 159,748 sequences of 74,936 taxa). Sequences with a 100% identification match to at least one taxonomic reference database were retained. Sequences assigned to the same taxon were merged by combining their read counts and taking the maximum number of PCR repeats the sequence was found in at each sample depth. The final taxonomic assignment of the retained species was determined using regional botanical taxonomic expertise by Alsos and following the taxonomy of the Panarctic Flora [36] and Lid's Norsk Flora [37]. All taxa identified in the negative controls were discarded from the overall dataset. Sample quality was assessed using the metabarcoding technical quality (MTQ) and metabarcoding analytical quality (MAQ) scores, which, respectively, measure overall metabarcoding success and the success of retrieving barcodes of interest, following the approach of [26].

2.6. Data Analysis

All further data filtering, analysis, and plotting was performed using custom Python and R scripts (https://github.com/salanova-elliott/jokelvatnet_data (accessed on 26 September 2022)) using the vegan [38], rioja [39], and ggplot2 [40] packages. To compare the compositional changes of the vegetation communities through time, we used a stratigraphically constrained sum of squares (CONISS) cluster analysis (Grimm 1987) with a broken stick model to determine the number of statistically significant clusters. We performed this analysis both using the proportion of PCR replicates that a single taxon appears in per sample and using the proportion of total retained reads that taxon appears in per sample. The results of both analyses are presented here, but we focus on the proportion of PCR replicates since this measure is considered to more clearly reflect community composition changes. Vascular plant taxa were assigned ecological trait values from [21] for the following traits: moisture (12 degree scale), temperature optimum (18 degree scale), and soil disturbance (9 degree scale). These traits were selected as those most likely to be influenced by climate and glacial activity. The temperature optimum index was inverted to provide for a more intuitive interpretation (1 = high alpine/arctic taxa, 18 = subtropical taxa). The assignment of traits to genus- or family-level identifications follows those described in [22]. Briefly, average trait values for all corresponding taxa present in the region today were averaged if within <3 category differences for soil disturbance, <4 for moisture, <5 for temperature optimum. These values were averaged and weighted on PCR replicates for each sample to produce a single value for each trait.

Reconstructed climatic data for Jøkelvatnet was retrieved from the CHELSA-TraCE21k model [24] using a custom python script (https://github.com/salanova-elliott/chelsa_retrieve (accessed on 26 September 2022)) using the coordinates 70.1715 N, 21.7014 E. We specifically examined the variables "mean temperature of warmest quarter" (bio10) and "annual precipitation" (bio12). Summer temperatures were used instead of annual temperatures since arctic flora are more responsive to changes during their growing season [41]. Values for specific years were linearly interpolated from the 100 year resolution climate data. Ti (cps) was used as a proxy for glacial activity in the catchment as presented by [23]. We used simple linear regressions to explore the relationship between the average weighted trait values, glacial activity, and the reconstructed climatic variables.

3. Results

3.1. sedaDNA Dataset Composition

The 8 PCR replicates of 40 samples and 6 controls produced a total of 8,551,653 reads (mean of 198,026 per sample). We retained 38 samples that had a metabarcoding technical quality (MTQ) score > 0.75 (mean of 0.970, σ = 0.055) and a metabarcoding analytical quality (MAQ) score > 0.2 (mean of 0.918, σ = 0.146). All 6 control samples had MTQ and MAQ scores below these thresholds. From the initial 8.5 million reads, ~65% passed quality filtering and were assigned with 100% identity to a taxon. Post-identification filtering resulted in a final dataset of 193 taxa composed of 25% identified at the family-level or above, 26% at genus-level, and 49% at species-level. Of these 193 taxa, 133 are the targeted vascular plant taxa, with 16% at the family-level, 32% at genus-level, and 52% at species-level while the other 60 are bryophytes. We note that some of the identifications considered at genus- and family-level in the aforementioned statistics can be confidently resolved to a subset of lower taxa (e.g., *Ranunculus glacialis/hyperboreus* is considered genus-level despite being narrowed down to two out of 24 potential regional species of *Ranunculus*). Two algal taxa were not included in the final dataset as their presence does not reflect terrestrial vegetation changes and their identification is strongly limited by a lack of representation of algae in the reference library.

3.2. Zonation

A stratigraphically constrained cluster (CONISS) analysis performed on the proportion of PCR replicates assigned to each taxon and compared with a broken stick model suggests the presence of four statistically significant zones in the data. The analysis performed on the read count data provided the same number of zones, with some slightly shifted boundaries (Figures S1 and S2; the oldest zone boundary changes by one sample from 9.7 ka to 9.6 ka, while the middle boundary also changes by one sample from 8.4 ka to 8.0 ka). The nearly identical zone boundaries identified by CONISS using proportion of PCR replicates and proportion of total reads support the interpretation of concurrent increases in taxonomic richness and changes in taxonomic abundance during each zone transition. The full data for each growth form are in Figure S3, whereas the pattern of abundance and richness for each growth form for these zones are displayed in Figure 2A,B.

The vegetation changes throughout the Holocene have largely coincided with changes in glacial activity, as inferred by changes in Ti count rates throughout the sediment core. The major boundary identified by CONISS at ~8.4 ka is at a time with high arrival of new taxa, especially forbs (Figure S3). The other two boundaries identified by CONISS coincide with major turning points in glacial activity identified by [23]. The boundary at ~9.7 ka occurs shortly after the valley becomes entirely deglaciated at 10 ka. The boundary division at ~4.3 ka occurs just prior to the onset of the Late Holocene as the Langfjordjøkelen ice cap begins to reform (Figure 2).

3.2.1. Zone 1

10.4–9.8 ka (4 samples). This zone is characterized by relatively few taxa per sample (mean of 16.8, σ = 6.4) and was dominated in read counts by Saliceae. These Saliceae likely include cold tolerant dwarf shrubs as *Salix herbacea*, *S. reticulata*, and *S. polaris*, which were identified in the vegetation survey close to the present glacier. Bryophytes (*Grimmiaceae*), ferns (*Cystopteris*, *Woodsia*), forbs (*Bistorta vivipara*, *Oxyria digyna*, *Saxifraga oppositifolia*, *Bartsia alpina*), and some dwarf shrubs (*Dryas octopetala*, *Empetrum nigrum*) were also present.

3.2.2. Zone 2

9.7–8.7 ka (5 samples). A dramatic decrease in the read proportion of Saliceae, as well as the appearance of additional woody taxa *Betula* and *Sorbus aucuparia*, occurred at the start of this zone. The arrival of *Athyrium* and *Phegopteris connectilis* caused a large spike in the read proportions of ferns (Figure 2A). Taxonomic richness continued to increase over this zone (mean taxa per sample of 26.6, σ = 4.6) with 18 new forb taxa arriving (Figure 2B).

3.2.3. Zone 3

8.2–4.5 ka (9 samples). Taxonomic richness sharply increased at the start of this zone (mean taxa per sample of 57.4, σ = 8.5). Dwarf shrubs such as *Vaccinium myrtillus*, *Vaccinium vitis-idaea*, *Phyllodoce caerulea*, as well as *Empetrum nigrum*, were responsible for the rise in the read proportions of dwarf shrubs while Saliceae and tree species also increased (Figure 2A).

3.2.4. Zone 4

4.2–0 ka (20 samples). This zone begins with another sharp increase of taxonomic diversity (mean taxa per sample of 86.2, σ = 12.8) with 67 taxa found only in this period. Graminoids in particular became more diverse with 12 new taxa after being composed primarily of *Oreojuncus trifidus* in the previous two zones. Bryophytes and forbs showed an increase in diversity as well, with large fluctuations in total taxonomic richness throughout the zone (Figure 2B). Saliceae makes up >50% of the total read proportion for the majority of this zone (Figure 2A) which likely includes the pioneer species mentioned in Zone 1, but also the shrub-forming *Salix phylicifolia* growing near the lake.

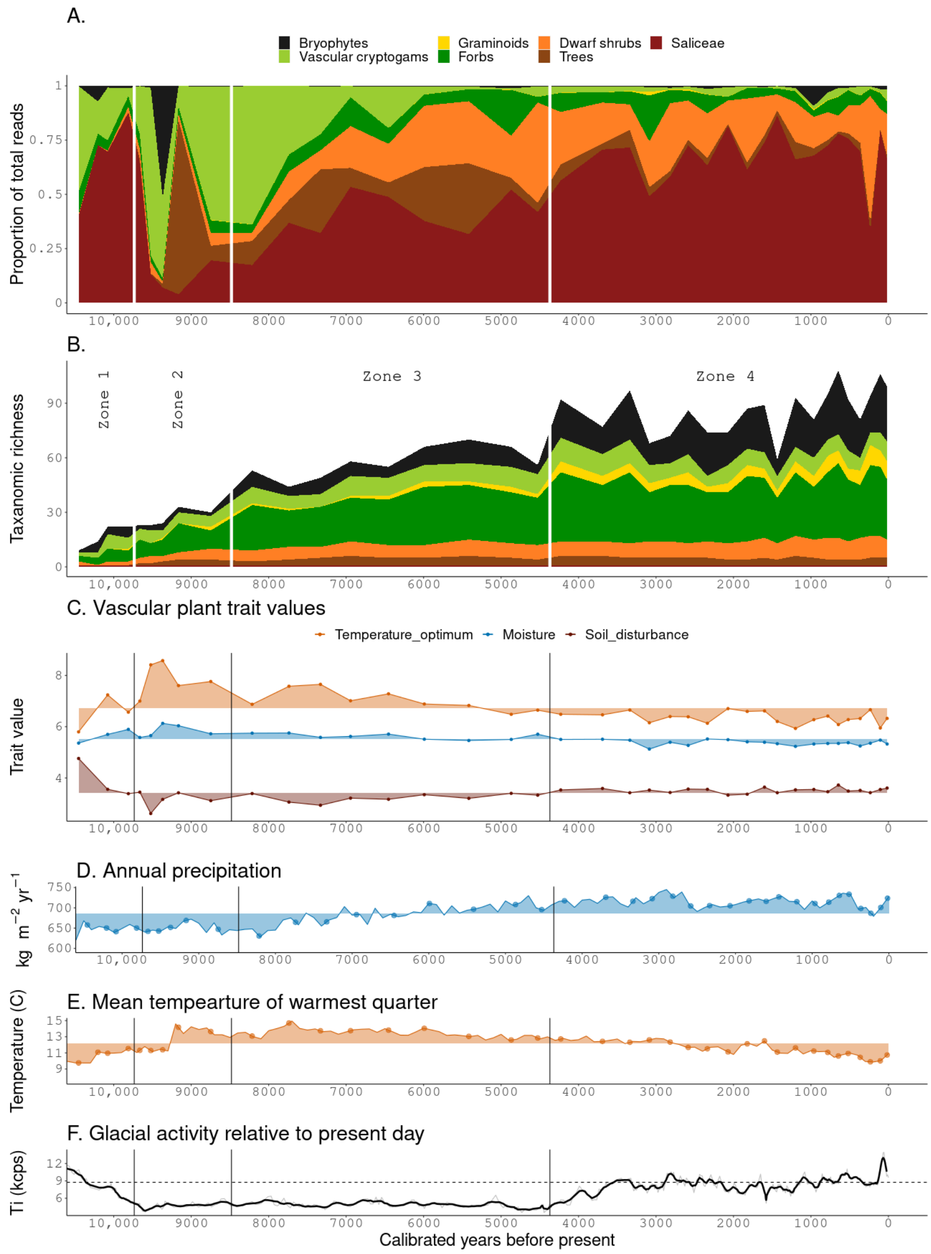


Figure 2. An overview of *sed*aDNA results, climate reconstructions, and glacial activity. CONISS zone boundaries are demarcated with vertical bars at 9.7, 8.4, and 4.3 ka. (**A**) Proportion of total identified reads by plant functional group. (**B**) Stacked taxonomic richness for each functional group. (**C**) Average weighted vascular plant trait values are based on plants identified in the sedaDNA combined with plant trait values reported in [21]. Note that temperature optimum index values are inverted from those reported in [21]. High values indicate high temperature optimum, high moisture requirement and high dependents of soil disturbance. (**D**) Annual precipitation data (bio12) from [24]. Points represent the age of samples taken from the core. (**E**) Mean temperature of the warmest quarter (bio10) from [24]. Points represent the age of samples taken from the core. (**F**) Glacial activity relative to the present day (dashed horizontal line) adapted from [23].

3.3. Ecological Trait Values

From the eligible 133 vascular plant taxa, 90 taxa were found to have informative ecological trait values for "soil disturbance", 81 taxa for "moisture", and 74 taxa for "temperature optimum". The second oldest sample at 10.2 ka had <5 taxa for each trait examined, and was consequently excluded from the trait analyses. The average soil disturbance index of the plant community starts at a high value of 4.7 (5 = requires soil disturbance for reproduction, but established individuals may persist for long (decades–centuries) in undisturbed vegetation), and then decreases to a low value of 2.6 (2 = colonizes already established vegetation, successfully competes with for some time but in the long run outcompeted if there is no soil disturbance) at 9.5 ka (Figure 2C). This value then gradually increases while approaching the present at 3.6 (4 = with some capacity to reproduce also in undisturbed established vegetation, but not sufficient to keep a stable population size). The temperature optimum index follows an inverted trend of this pattern with an initial dominance of cold-adapted plants, increasing temperature optimum values in zone 2 with a peak of 8.6 at 9.3 ka, and then colder taxa gradually become more prevalent towards the current day (Figure 2D). The average moisture trait values show a similar trend on a smaller scale where values peak at 6.1 (6 = moist) at 9.3 ka and then erratically decrease approaching the present day with a minimum value of 5.1 (5 = mesic-moist) at 3.1 ka (Figure 2E). The spike in average moisture trait value at 9.3 ka can be attributed to the temporary disappearance of the "dry-mesic" (moisture trait value = 3) *Cryptogramma crispa* and *Arctous alpina* as well as the appearance of the "moist-wet" *Bartsia alpina* (moisture trait value = 7). The appearance of many "mesic" (moisture trait value = 4) graminoids (*Oreojuncus trifidus*, *Avenella flexuosa*, and *Poa pratensis*/*alpina*/*Anthoxanthum*) and forbs (*Diapensia lapponica* and *Ranunculus acris*/*subborealis*) in the Late Holocene causes the average moisture trait value to decrease through this time period.

3.4. Taxonomic Persistence and Vegetation Surveys

We identified 109 taxa in the combined vascular plant survey, of which 101 are represented in the metabarcoding data (Table S1). From the 133 vascular plant taxa present in the metabarcoding data, 108 were represented across both vegetation surveys (Table S1). The discrepancy between taxa counts of the two measures is due to the differing level of taxonomic identification in the vascular survey compared to metabarcoding (e.g., four *Salix* species were identified in the vascular plant survey whereas metabarcoding can only resolve the tribe Saliceae). The remaining 25 taxa in the metabarcoding dataset were not observed during either vegetation survey despite a directed effort to locate them during our second survey. However, 17 of these taxa were not found in the most recent metabarcoding samples, while 11 appear in $\leq$2 samples suggesting that they may no longer grow in the catchment or are very rare. In contrast, the majority of taxa persisted from one zone to the next, with 94%, 97%, and 96% of taxa detected in zone 1, 2, and 3, respectively, also detected in zone 4 (Figure 3). The few taxa that disappear in the record are bryophytes or forbs that appear in only one sample with <4 replicates.

The vegetation survey revealed that the large floodplain at the inlet in the south end is dominated by *Eriophorum angustifolium* and *Salix herbaceae* with smaller patches of *Nardus stricta* and *Deschampsia cespitosa* and a near 100% bryophyte layer largely composed of *Sphagnum*. The majority of the vegetation 10–20 m from the lake shore is Alpine heath dominated by *Empetrum nigrum*, *Vaccinium myrtillus*, *Phyllodoce caerula*, with the sedge *Carex bigelowii* being co-dominant. Slopes surrounding the lake are dominated by tallus with two small patches of *Betula pubescens* forest located on the north-west slopes of the lake. There were several springs in the hillslopes, one of them very rich with *Angelica archangelica* and *Anthriscus sylvestris*. Moraines towards the current glacier have discontinuous cover of typical arctic-alpine species as *Dryas octopetala*, *Kalmia procumbens* and *Ranunculus glacialis*.

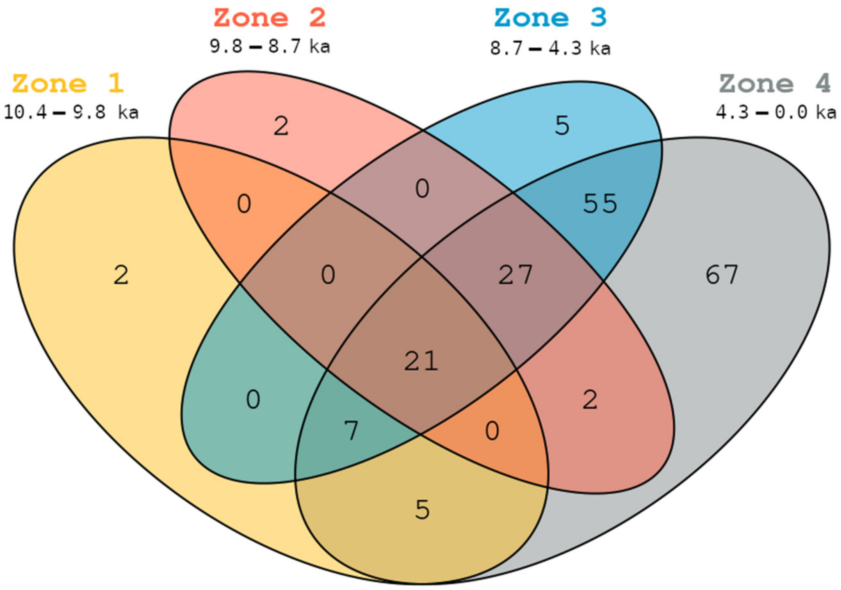


Figure 3. Number of plant taxa identified in the *sed*aDNA analyses that are shared between CONISS demarcated zones.

3.5. Linear Regressions

Soil disturbance, temperature optimum, and moisture values show a significant and intermediate strong correlation with glacial activity (p < 0.001, $R^2 \approx 0.4$; Figure 4). Two samples (9.4 and 9.5 ka) were identified as outliers and removed for linear regressions involving the temperature optimum trait value. These samples occur directly before the spike in summer temperatures reconstructed by CHELSA at 9.3 ka. Plant traits for temperature optimum show a significant correlation with the mean temperature of warmest quarter (bio10) data from CHELSA-TraCE21k (p < 0.001), with more cold tolerant plants present when the mean temperature is lower (Figure 5). Moisture trait values show a similar, but inverted, correlation with annual precipitation (bio12) data (p < 0.001) where dry-adapted plants become more prevalent as precipitation increases (Figure 5). Samples from early in the record have fewer taxa present and display more variability, but still largely follow the described trends (Figures 4 and 5).

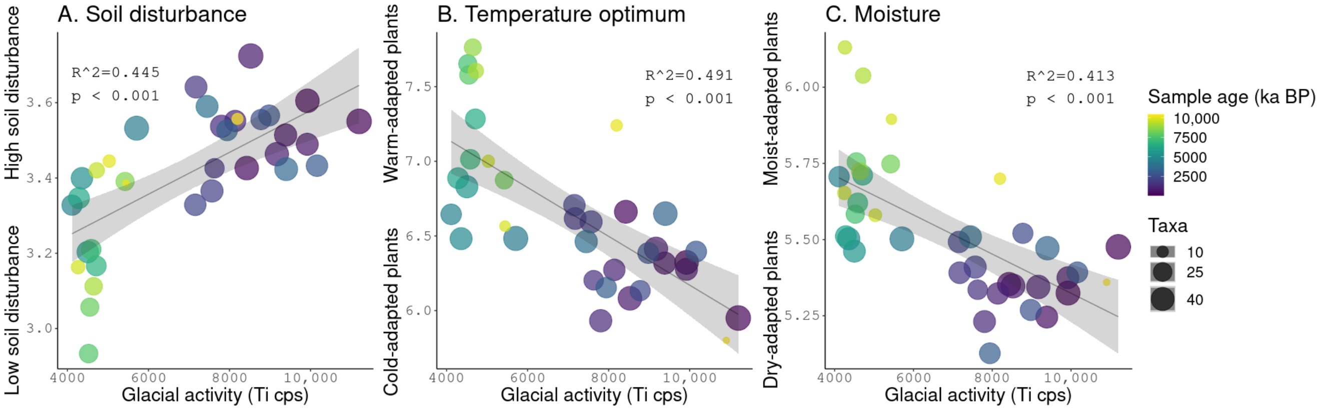


Figure 4. Linear regression of glacial activity and (**A**) soil disturbance, (**B**) temperature optimum, and (**C**) moisture trait values based on plants identified in the *sed*aDNA combined with plant trait values reported in [21]. The number of plant taxa used to calculate the weighted average is shown as point size while sample age is represented as a color scale. Note that temperature optimum index values are inverted from those reported in [21].

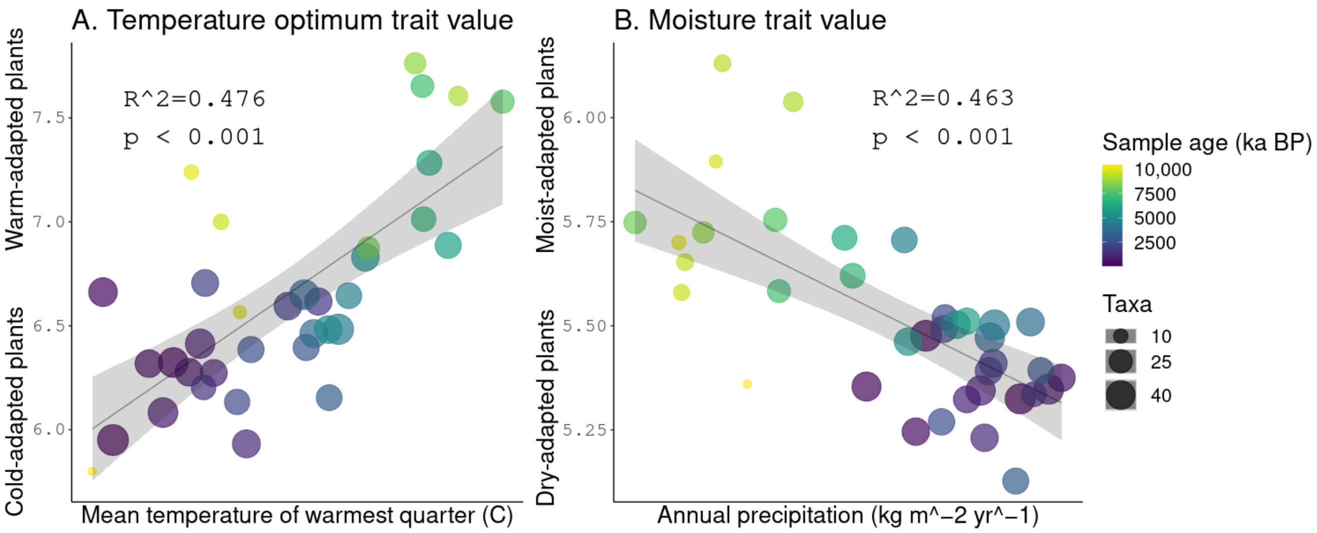


Figure 5. Linear regression of (**A**) the mean temperature of the warmest quarter and the temperature optimum trait value and (**B**) the annual precipitation and moisture trait value. The number of plant taxa used to calculate the weighted average is shown as point size while sample age is represented as a color scale. Note that temperature optimum index values are inverted from those reported in [21].

4. Discussion

Our study shows shifts in both species richness and composition coinciding with changes in glacial activity. A number of other studies have used metabarcoding to examine vegetation changes in a lake catchment containing an active glacier [42–44], but few offer direct analogues to Jøkelvatnet's system of near-complete glacial melting followed by significant reformation during the Late Holocene. With nearly one fifth of the catchment currently occupied by the glacier, vegetation changes as a result of the glacier's influence are readily apparent in the *seda*DNA record. In contrast to pollen's long dispersal distances, *seda*DNA represents a local signal of taxa growing in the catchment [17,18]. Compared to the overall pattern found across ten lakes in northern Fennoscandia, Jøkelvatnet stands out by showing more variation during the Middle to Late Holocene in plant trait values, when the overall pattern of the region is a more stable ecosystem [22]. Here, we posit that the presence of the Langfjordjøkel glacier enhances the effects of Holocene climate change on the vegetation in the catchment.

The overall trend of increasing taxonomic richness throughout the Holocene at Jøkelvatnet is consistent with other lake catchments across northern Fennoscandia [26]. At Jøkelvatnet, the taxonomic richness at each timepoint is increasing at a consistent rate from the beginning of the record until 4.2 ka when it begins to fluctuate dramatically (Figure 2B). This could reflect the rapidly changing landscape of the Late Holocene period where the Langfjordjøkel ice cap is repeatedly reforming and melting. In contrast to expectations of increased soil erosion delaying succession sequences [1], taxonomic richness increased even during periods of high erosion (Figure 2B). In other lake catchments, an influx of glacial flour has been documented as coinciding with a decrease in overall DNA concentration in the sediment, but this decrease is not necessarily reflected in the number of MOTUs retrieved from these periods [43]. Since the vegetation here is recorded from the entire catchment, new taxa are able to colonize the disturbed soils near the glacier as well as the later successional environments at the eastern and western slopes of the lake.

There are only nine taxa that appear in the record and then disappear entirely in subsequent zones (Figure 3). Many of these taxa are bryophytes that are present in only one sample with few PCR replicates (Figure S4) suggesting a low abundance in *seda*DNA and/or difficulty in detection. The detectability of plants is related to their abundance in the catchment [45], and bryophytes in general show less consistent detection than vascular plants [22,46]. Thus, similar to observations for the region of N Fennoscandia [22], we assume that local extirpation has been low in this catchment.

The average temperature optimum trait value's correlation with glacial activity (Figure 5) is in accordance with [23]'s observation of Langfjordjøkelen's glacial activity following regional summer temperatures. Vegetation temperature optimum values at the beginning of the core reflect the relatively rapid warming of the Early Holocene (Figure 2D). Communities of high arctic/alpine taxa are supplemented with warmer taxa at the beginning of the Holocene Thermal Maximum at ~9.5 ka. An increased abundance of cold tolerant forbs and dwarf shrubs (*Oxyria digyna*, *Dryas octopetala*, *Kalmia procumbens*) causes a small dip of cold temperature optimum values at 8.2 ka (Figure 2C) coinciding with the well-documented cold event [10]. This spike is also seen in the detrital parameters such as magnetic susceptibility and Ti count rate which [23] hypothesized could be due to the glacier temporarily reforming or simply cooler temperatures leading to less organic input. The average temperature optimum value of the vegetation gradually shifts colder throughout the Middle Holocene and into the Neoglacial period as Langfjordjøkelen begins to reform. Mean ground surface temperature, which can vary in proximity to a glacier, has been identified as a key explanatory variable for plant community composition in glacial forelands [47]. Similar to ground surface temperature, snow cover is also highly spatially heterogeneous and influential on vegetation composition [48]. Glaciers provide landscape obstacles for snow drift accumulation and can increase snow persistence in their immediate vicinity [49]. Thus, the change towards more cold adapted plants over the Middle and Late Holocene is likely both a direct effect of regional cooling and a local effect of an expanding glacier causing additional local cooling.

The negative correlation between plant moisture trait values and both glacial activity and annual precipitation is the opposite trend as might be intuitively expected. A greater proportion of moist-adapted plants are present during the Early and Middle Holocene when no glacier is present in the catchment (Figure 2E). Similarly, more dry-adapted plants are detected in the Late Holocene which had higher annual precipitation and is often characterized by mire and wetland formation [7,24]. The range of these average moisture trait values is fairly small (1 category: 5.1 (mesic-moist)–6.1 (moist)), but the negative correlation with glacial activity and annual precipitation is significant (Figures 4C and 5B) and could be explained by the locking-up of precipitation in both snow and ice caused by the decreasing temperatures during the Late Holocene.

The catchment of Jøkelvatnet is generally too steep for extensive mire formation, so as the Langfjordjøkel glacier melted and reformed during the Late Holocene, many of the previously mentioned species were growing on either exposed moraine ridges or in the heath directly adjacent to the lake. Additionally, it is important to note that bryophytes were excluded from all trait analyses due to poor taxonomic resolution with the gh primer set and the lack of available trait databases. While some species are highly tolerant of desiccation, bryophytes are highly dependent on locally moist conditions to propagate [50]. At Jøkelvatnet, the abundance and diversity of bryophytes were highest during the Late Holocene and for one spike at 9.3 ka (Figure 2A,B). This pattern is in contrast to that recorded by the vascular plant vegetation and offers a contradictory, and potentially more robust, view of moisture conditions around Jøkelvatnet. In addition, the pattern of moisture tolerance may not reflect overall precipitation as changes in meltwater may alter the inflow of sediments and formation of floodplains. Thus, while patterns of plant temperature optimum traits correlate well with glacial activity and the climate proxies, moisture traits may be highly impacted by the local environment and the analysis of many sites might be required to find regional patterns (e.g., [22]).

The correlation between average soil disturbance trait values and glacial activity (Figure 4A) suggests that the glacier was having a direct effect on the vegetation. Disturbance-dependent forbs such as *Oxyria digyna* and multiple *Saxifraga* species are some of the few taxa recorded in the earliest parts of the record as the valley was first becoming deglaciated. The arrival of more competitive dwarf shrubs (*Vaccinium*) and woody taxa in zones 2 and 3 coincides with the absence of glacial activity. A similar pattern is recorded in the catchment of Lake Bolshoye Shchuchye in the Polar Urals where glaciers had nearly completely melted by 15 ka as vegetation shifted to be shrub-dominated [42,51]. In the Jøkelvatnet catchment, the increasing diversity of disturbance-dependent forbs (additional *Ranunculus* and *Saxifraga* species) during the Neoglacial period (zone 4) could be the result of increased soil erosion from the continually melting and reforming glacier during this period. In contrast, only a few small glaciers reformed in shady cirques around Lake Bolshoye Shchuchye during the Late Holocene while *Ranunculus sulphureus* and various *Saxifraga* species do not return to the sedaDNA record [42,51]. Lake Muzelle in the French Alps also displays this trend as *Saxifraga paniculata*, *Oxyria digyna*, and *Eritrichium* sp. appear towards the end of the Little Ice Age when glaciers in the catchment began to retreat [43]. Glacial activity is a good predictor of habitat availability for pioneer species [52] and the increase and arrival of these early successional stage taxa are detectable through *seda*DNA.

5. Conclusions

We demonstrate that *seda*DNA is a useful tool for both reconstructions of past environments and investigating changes in plant richness and composition. In particular, the more local deposition of sedaDNA compared to pollen enables a finer-scale study of how local environmental conditions may affect vegetation (e.g., the effect of a glacier on a particular catchment's flora). The *seda*DNA record at Jøkelvatnet shows significant changes in the plant community at 9.7 and 4.3 ka corresponding to inflection points in Langfjordjøkel's glacial activity. A correlation of the vegetation's temperature optimum trait value with glacial activity is primarily due to climate but may have been exacerbated by a direct effect of the glacier on the plant community. Additionally the correlation of vegetation soil disturbance trait values and glacial activity implies that the glacial activity has had a direct effect on the vegetation. When compared with catchments lacking a glacier in northern Fennoscandia, Jøkelvatnet's ecosystem displays more variability throughout the Middle and Late Holocene suggesting that the Langfjordjøkel glacier has enhanced the effect of climate change on the plant community. The counterintuitive negative soil moisture trait–glacial activity relationship is likely due to a reduction in soil moisture caused by decreasing temperatures. Beyond the direct impact of climate change on arctic vegetation, our results suggest the disappearance of glaciers will have an additional effect on plant communities. Studying biotic responses to past climate change compliments contemporary monitoring and field experiments when predicting the effects of current climate change on vegetation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/quat6010007/s1. Figure S1. CONISS clustering based on proportions of PCR replicates. Figure S2. CONISS clustering based on proportions of total filtered reads. Figure S3. Number of PCR replicates each species appears in separated by functional group. Figure S4. Proportion of PCR replicates based on functional groups. Figure S5. Diagnostic plots for linear regressions.

Author Contributions: L.D.E., I.G.A. and A.G.B. designed the study. I.G.A., L.D.E. and L.T. did the vegetation surveys. J.B. provided access to the Jøkelvatnet core and data. D.P.R. generated the sedaDNA data. D.P.R., I.G.A. and L.D.E. matched and filtered the sedaDNA data. P.D.H. and A.G.B. provided the age-depth model. L.D.E. analyzed the data and wrote a first draft of the manuscript upon which all co-authors commented. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement no. 819192 (to I.G.A. and A.G.B.; supported L.D.E and L.T.) for the IceAGenT project; Research Council of Norway grant 250963/F20 (to I.G.A) for the ECOGEN project.

Data Availability Statement: All data needed to evaluate the conclusions and generate the figures in the paper are present in the paper and/or the Supplementary Materials. Scripts and data used for generating Figures 2–5 are available at (https://github.com/salanova-elliott/jokelvatnet_data (accessed on 26 September 2022)). All newly generated raw sedaDNA sequence data have been deposited in the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena/browser/home (accessed on 26 September 2022)) with BioProject accession PRJEB39329.

Acknowledgments: We would like to thank M. F. Merkel for participating in the vegetation survey and assistance with technical work; A. Reveret for comments on the manuscript; H. E. Ludwig (Wittmeier) for her previous work and providing data access; R. Paulssen and the GSCT at The Arctic University of Norway for amplicon sequencing. Bioinformatic analyses were performed on resources provided by UNINETT Sigma2–the National Infrastructure for High Performance Computing and Data Storage in Norway.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Matthews, J.A. *The Ecology of Recently-deglaciated Terrain: A Geoecological Approach to Glacier Forelands*; Cambridge University Press: Cambridge, UK, 1992; 386p.
2. Raffl, C.; Mallaun, M.; Mayer, R.; Erschbamer, B. Vegetation Succession Pattern and Diversity Changes in a Glacier Valley, Central Alps, Austria. *Arctic Antarct. Alp. Res.* **2006**, *38*, 421–428. [CrossRef]
3. Eichel, J.; Krautblatter, M.; Schmidtlein, S.; Dikau, R. Biogeomorphic interactions in the Turtmann glacier forefield, Switzerland. *Geomorphology* **2013**, *201*, 98–110. [CrossRef]
4. Wojcik, R.; Eichel, J.; Bradley, J.A.; Benning, L.G. How allogenic factors affect succession in glacier forefields. *Earth-Science Rev.* **2021**, *218*, 103642. [CrossRef]
5. Romundset, A.; Akçar, N.; Fredin, O.; Tikhomirov, D.; Reber, R.; Vockenhuber, C.; Christl, M.; Schlüchter, C. Lateglacial retreat chronology of the Scandinavian Ice Sheet in Finnmark, northern Norway, reconstructed from surface exposure dating of major end moraines. *Quat. Sci. Rev.* **2017**, *177*, 130–144. [CrossRef]
6. Hughes, A.L.C.; Gyllencreutz, R.; Lohne, Ø.S.; Mangerud, J.; Svendsen, J.I. The Last Eurasian Ice Sheets—A Chronological Database and Time-Slice Reconstruction, DATED-1. *Boreas* **2016**, *45*, 1–45. [CrossRef]
7. Sjögren, P.J. An Overview of Holocene Climate Reconstructions in Northernmost Fennoscandia. SapReps. 15 February 2021. Available online: https://septentrio.uit.no/index.php/SapReps/article/view/5747 (accessed on 1 May 2022).
8. Larocca, L.J.; Axford, Y. Arctic glaciers and ice caps through the Holocene:a circumpolar synthesis of lake-based reconstructions. *Clim. Past* **2022**, *18*, 579–606. [CrossRef]
9. Nesje, A.; Bakke, J.; Dahl, S.O.; Lie, Ø.; Matthews, J.A. Norwegian mountain glaciers in the past, present and future. *Glob. Planet. Chang.* **2008**, *60*, 10–27. [CrossRef]
10. Wanner, H.; Solomina, O.; Grosjean, M.; Ritz, S.P.; Jetel, M. Structure and origin of Holocene cold events. *Quat. Sci. Rev.* **2011**, *30*, 3109–3123. [CrossRef]
11. Birks, H.J.B. Contributions of Quaternary botany to modern ecology and biogeography. *Plant Ecol. Divers.* **2019**, *12*, 189–385. [CrossRef]
12. Sjögren, P.; Damm, C. Holocene Vegetation Change in Northernmost Fennoscandia and the Impact on Prehistoric Foragers 12 000–2000 cal. a BP—A Review. *Boreas* **2019**, *48*, 20–35. [CrossRef]
13. Sjögren, P.; van der Knaap, W.; Huusko, A.; van Leeuwen, J.F. Pollen productivity, dispersal, and correction factors for major tree taxa in the Swiss Alps based on pollen-trap results. *Rev. Palaeobot. Palynol.* **2008**, *152*, 200–210. [CrossRef]
14. Ritchie, J.C. Current trends in studies of long-term plant community dynamics. *New Phytol.* **1995**, *130*, 469–494. [CrossRef] [PubMed]
15. Alsos, I.G.; Sjögren, P.J.E.; Edwards, M.E.; Landvik, J.Y.; Gielly, L.; Forwick, M.; Coissac, E.; Brown, A.; Jakobsen, L.V.; Føreid, M.K.; et al. Sedimentary ancient DNA from Lake Skartjørna, Svalbard: Assessing the resilience of arctic flora to Holocene climate change. *Holocene* **2015**, *26*, 627–642. [CrossRef]
16. Birks, H.H.; Birks, H.J.B. Future uses of pollen analysis must include plant macrofossils. *J. Biogeogr.* 2000. Available online: https://www.jstor.org/stable/2655981 (accessed on 14 October 2021).
17. Parducci, L.; Bennett, K.D.; Ficetola, G.F.; Alsos, I.G.; Suyama, Y.; Wood, J.R.; Pedersen, M.W. Ancient plant DNA in lake sediments. *New Phytol.* **2017**, *214*, 924–942. [CrossRef]
18. Sjögren, P.; Edwards, M.E.; Gielly, L.; Langdon, C.T.; Croudace, I.W.; Merkel, M.K.F.; Fonville, T.; Alsos, I.G. Lake sedimentary DNA accurately records 20th Century introductions of exotic conifers in Scotland. *New Phytol.* **2016**, *213*, 929–941. [CrossRef] [PubMed]
19. Taberlet, P.; Coissac, E.; Pompanon, F.; Gielly, L.; Miquel, C.; Valentini, A.; Vermat, T.; Corthier, G.; Brochmann, C.; Willerslev, E. Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.* **2006**, *35*, e14. [CrossRef] [PubMed]
20. Wang, Y.; Pedersen, M.W.; Alsos, I.G.; De Sanctis, B.; Racimo, F.; Prohaska, A.; Coissac, E.; Owens, H.L.; Merkel, M.K.F.; Fernandez-Guerra, A.; et al. Author Correction: Late Quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature* **2021**, *600*, 86–92. [CrossRef]

21. Tyler, T.; Herbertsson, L.; Olofsson, J.; Olsson, P.A. Ecological indicator and traits values for Swedish vascular plants. *Ecol. Indic.* **2020**, *120*, 106923. [CrossRef]
22. Alsos, I.G.; Rijal, D.P.; Ehrich, D.; Karger, D.N.; Yoccoz, N.G.; Heintzman, P.D.; Brown, A.G.; Lammers, Y.; Pellissier, L.; Alm, T.; et al. Postglacial species arrival and diversity buildup of northern ecosystems took millennia. *Sci. Adv.* **2022**, *8*, eabo7434. [CrossRef]
23. Wittmeier, H.E.; Bakke, J.; Vasskog, K.; Trachsel, M. Reconstructing Holocene glacier activity at Langfjordjøkelen, Arctic Norway, using multi-proxy fingerprinting of distal glacier-fed lake sediments. *Quat. Sci. Rev.* **2015**, *114*, 78–99. [CrossRef]
24. Karger, D.N.; Nobis, M.P.; Normand, S.; Graham, C.H.; Zimmermann, N.E. CHELSA-TraCE21k v1.0. Downscaled transient temperature and precipitation data since the last glacial maximum. *Clim. Past.* 2021, pp. 1–27. Available online: https://cp.copernicus.org/preprints/cp-2021-30/ (accessed on 18 May 2022).
25. Andreassen, L.M.; Winsvold, S.H.; Paul, F.; Hausberg, J.E. Inventory of Norwegian glaciers. In *Rapport*; Andreassen, L.M., Winsvold, S.H., Eds.; Norwegian Water Resources and Energy Directorate: Oslo, Norway, 2012; Volume 38, p. 240.
26. Rijal, D.P.; Heintzman, P.D.; Lammers, Y.; Yoccoz, N.G.; Lorberau, K.E.; Pitelkova, I.; Goslar, T.; Murguzur, F.J.A.; Salonen, J.S.; Helmens, K.F.; et al. Sedimentary ancient DNA shows terrestrial plant richness continuously increased over the Holocene in northern Fennoscandia. *Sci. Adv.* **2021**, 7. [CrossRef]
27. Nesje, A. A Piston Corer for Lacustrine and Marine Sediments. *Arct. Alp. Res.* **1992**, *24*, 257. [CrossRef]
28. Blaauw, M.; Christen, J.A. Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Anal.* **2011**, *6*, 457–474. [CrossRef]
29. Reimer, P.J.; Bard, E.; Bayliss, A.; Warren Beck, J.; Blackwell, P.G.; Ramsey, C.B.; Buck, C.E.; Cheng, H.; Edwards, R.L.; Friedrich, M.; et al. IntCal13 and Marine13 Radiocarbon Age Calibration Curves 0–50,000 Years cal BP. *Radiocarbon* **2013**, *55*, 1869–1887. [CrossRef]
30. Taberlet, P.; Bonin, A.; Zinger, L.; Coissac, E. *Environmental DNA: For Biodiversity Research and Monitoring*; Oxford University Press: Oxford, UK, 2018; 253p.
31. Boyer, F.; Mercier, C.; Bonin, A.; Bras, Y.L.; Taberlet, P.; Coissac, E. Obitools: A unix-inspired software package for DNA metabarcoding. *Mol. Ecol. Resour.* **2016**, *16*, 176–182. [CrossRef] [PubMed]
32. Sønstebø, J.H.; Gielly, L.; Brysting, A.K.; Elven, R.; Edwards, M.; Haile, J.; Willerslev, E.; Coissac, E.; Rioux, D.; Sannier, J.; et al. Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Mol. Ecol. Resour.* **2010**, *10*, 1009–1018. [CrossRef] [PubMed]
33. Willerslev, E.; Davison, J.; Moora, M.; Zobel, M.; Coissac, E.; Edwards, M.E.; Lorenzen, E.D.; Vestergård, M.; Gussarova, G.; Haile, J.; et al. Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature* **2014**, *506*, 47–51. [CrossRef] [PubMed]
34. Soininen, E.M.; Gauthier, G.; Bilodeau, F.; Berteaux, D.; Gielly, L.; Taberlet, P.; Gussarova, G.; Bellemain, E.; Hassel, K.; Stenøien, H.K.; et al. Highly overlapping winter diet in two sympatric lemming species revealed by DNA metabarcoding. *PLoS ONE* **2015**, *10*, e0115335. [CrossRef] [PubMed]
35. Alsos, I.G.; Lavergne, S.; Merkel, M.K.F.; Boleda, M.; Lammers, Y.; Alberti, A.; Pouchon, C.; Denoeud, F.; Pitelkova, I.; Pușcaș, M.; et al. The Treasure Vault Can be Opened: Large-Scale Genome Skimming Works Well Using Herbarium and Silica Gel Dried Material. *Plants* **2020**, *9*, 432. [CrossRef] [PubMed]
36. Elven, R.; Murray, D.F.; Razzhivin, V.Y.; Yurtsev, B.A. *Annotated Checklist of the Panarctic Flora (PAF) Vascular Plants*; Natural History Museum, University of Oslo: Oslo, Norway, 2011.
37. Elven, R.; Alm, T.; Berg, T.; Båtvik, J.I.I.; Fremstad, E.; Pedersen, O. *Johannes Lid & Dagny Tande Lid: Norsk Flora*; Det Norske Samlaget: Oslo, Norway, 2005.
38. Oksanen, J.; Blanchet, F.G.; Friendly, M.; Kindt, R.; Legendre, P.; McGlinn, D.; Minchin, P.; O'Hara, R.B.; Simpson, G.; Solymos, P.; et al. vegan: Community Ecology Package. R package version 2.5-6. Available online: https://CRAN.R-project.org/package=vegan (accessed on 10 March 2021).
39. Juggins, S. Rioja: Analysis of Quaternary Science Data. 2020. Available online: https://cran.r-project.org/package=rioja (accessed on 10 March 2021).
40. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016; Available online: https://ggplot2.tidyverse.org (accessed on 10 March 2021).
41. Walker, M.D.; Walker, D.A.; Welker, J.M.; Arft, A.M.; Bardsley, T.; Brooks, P.D.; Fahnestock, J.T.; Jones, M.H.; Losleben, M.; Parsons, A.N.; et al. Long-Term Experimental Manipulation of Winter Snow Regime and Summer Temperature in Arctic and Alpine Tundra. *Hydrol. Process.* **1999**, *13*, 2315–2330. [CrossRef]
42. Clarke, C.L.; Edwards, M.E.; Gielly, L.; Ehrich, D.; Hughes, P.D.M.; Morozova, L.M.; Haflidason, H.; Mangerud, J.; Svendsen, J.I.; Alsos, I.G. Persistence of arctic-alpine flora during 24,000 years of environmental change in the Polar Urals. *Sci. Rep.* **2019**, *9*, 1–11. [CrossRef]
43. Giguet-Covex, C.; Ficetola, G.F.; Walsh, K.; Poulenard, J.; Bajard, M.; Fouinat, L.; Sabatier, P.; Gielly, L.; Messager, E.; Develle, A.L.; et al. New insights on lake sediment DNA from the catchment: Importance of taphonomic and analytical issues on the record quality. *Sci. Rep.* **2019**, *9*, 1–21. [CrossRef]
44. Heinecke, L.; Epp, L.S.; Reschke, M.; Stoof-Leichsenring, K.R.; Mischke, S.; Plessen, B.; Herzschuh, U. Aquatic macrophyte dynamics in Lake Karakul (Eastern Pamir) over the last 29 cal ka revealed by sedimentary ancient DNA and geochemical analyses of macrofossil remains. *J. Paleolimnol.* **2017**, *58*, 403–417. [CrossRef]

45. Alsos, I.G.; Lammers, Y.; Yoccoz, N.G.; Jørgensen, T.; Sjögren, P.; Gielly, L.; Edwards, M.E. Plant DNA metabarcoding of lake sediments: How does it represent the contemporary vegetation. *PLoS ONE* **2018**, *13*, e0195403. [CrossRef] [PubMed]
46. Ariza, M.; Fouks, B.; Mauvisseau, Q.; Halvorsen, R.; Alsos, I.G.; de Boer, H.J. Plant Biodiversity Assessment through Soil eDNA Reflects Temporal and Local Diversity. *Methods Ecol. Evol.* 2022. [CrossRef]
47. Giaccone, E.; Luoto, M.; Vittoz, P.; Guisan, A.; Mariéthoz, G.; Lambiel, C. Influence of microclimate and geomorphological factors on alpine vegetation in the Western Swiss Alps. *Earth Surf. Process. Landforms* **2019**, *44*, 3093–3107. [CrossRef]
48. Niittynen, P.; Luoto, M. The importance of snow in species distribution models of arctic vegetation. *Ecography* **2017**, *41*, 1024–1037. [CrossRef]
49. Schuler, T.V.; Crochet, P.; Hock, R.; Jackson, M.; Barstad, I.; Johannesson, T. Distribution of snow accumulation on the Svartisen ice cap, Norway, assessed by a model of orographic precipitation. *Hydrol. Process.* **2008**, *22*, 3998–4008. [CrossRef]
50. Bates, J.W. Is "life-form" a useful concept in bryophyte ecology? *Oikos* **1998**, *82*, 223. [CrossRef]
51. Svendsen, J.I.; Færseth, L.M.B.; Gyllencreutz, R.; Haflidason, H.; Henriksen, M.; Hovland, M.N.; Lohne, S.; Mangerud, J.; Nazarov, D.; Regnéll, C.; et al. Glacial and environmental changes over the last 60 000 years in the Polar Ural Mountains, Arctic Russia, inferred from a high-resolution lake record and other observations from adjacent areas. *Boreas* **2018**, *48*, 407–431. [CrossRef]
52. Jones, G.A.; Henry, G.H.R. Primary plant succession on recently deglaciated terrain in the Canadian High Arctic. *J. Biogeogr.* **2003**, *30*, 277–296. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Received: 22 August 2023 | Revised: 6 December 2023 | Accepted: 21 December 2023

DOI: 10.1111/1755-0998.13926

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Multiplexing PCR allows the identification of within-species genetic diversity in ancient eDNA

Y. Lammers[1] | P. Taberlet[1,2] | E. Coissac[2] | L. D. Elliott[1] | M. F. Merkel[1] | I. Pitelkova[1] | PhyloAlps Consortium | PhyloNorway Consortium | I. G. Alsos[1]

[1]The Arctic University Museum of Norway, UiT—The Arctic University of Norway, Tromsø, Norway

[2]Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LECA, Grenoble, France

Correspondence
Y. Lammers, The Arctic University Museum of Norway, UiT—The Arctic University of Norway, Tromsø, Norway.
Email: youri.lammers@uit.no

Funding information
Norges Forskningsråd, Grant/Award Number: 226134/F50, 250963/F20 and 255415; Norwegian Biodiversity Information Centre, Grant/Award Number: 14-14 and 70184209; H2020 European Research Council, Grant/Award Number: 819192; France Génomique, Grant/Award Number: ANR-10-INBS-09-08

Handling Editor: Paula Arribas

Abstract

Sedimentary ancient DNA (*seda*DNA) has rarely been used to obtain population-level data due to either a lack of taxonomic resolution for the molecular method used, limitations in the reference material or inefficient methods. Here, we present the potential of multiplexing different PCR primers to retrieve population-level genetic data from *seda*DNA samples. *Vaccinium uliginosum* (Ericaceae) is a widespread species with a circumpolar distribution and three lineages in present-day populations. We searched 18 plastid genomes for intraspecific variable regions and developed 61 primer sets to target these. Initial multiplex PCR testing resulted in a final set of 38 primer sets. These primer sets were used to analyse 20 lake *seda*DNA samples (11,200 cal. yr BP to present) from five different localities in northern Norway, the Alps and the Polar Urals. All known *V. uliginosum* lineages in these regions and all primer sets could be recovered from the *seda*DNA data. For each sample on average 28.1 primer sets, representing 34.15 sequence variants, were recovered. All sediment samples were dominated by a single lineage, except three Alpine samples which had co-occurrence of two different lineages. Furthermore, lineage turnover was observed in the Alps and northern Norway, suggesting that present-day phylogeographical studies may overlook past genetic patterns. Multiplexing primer is a promising tool for generating population-level genetic information from *seda*DNA. The relatively simple method, combined with high sensitivity, provides a scalable method which will allow researchers to track populations through time and space using environmental DNA.

KEYWORDS
ancient DNA, environmental DNA, multiplexing PCR, palaeo-phylogeography, population-level genomics

1 | INTRODUCTION

Ancient DNA (aDNA) has the potential to provide invaluable phylogeographic and population-level genomic information. The ability to reconstruct past population genomic histories can lead to better phylogeographical interpretations of present-day populations, resolve issues with hidden population replacement and reconstruct histories for extinct taxa (Edwards et al., 2022; McGaughran et al., 2022). The use of aDNA for population reconstructions has been demonstrated for a number of taxa such as mammoth (Palkopoulou et al., 2015; van der Valk et al., 2021), silver fir (Schmid et al., 2017), steppe bison (Heintzman et al., 2016), oak (Wagner et al., 2023) and most notably human (Allentoft et al., 2015; Posth et al., 2023; Skoglund et al., 2012). These studies, however, rely on fossil material for aDNA extraction, which are either rare or not available for the majority of species (Pedersen et al., 2015).

An alternative source of aDNA is sedimentary ancient DNA (*seda*DNA), DNA that can persist in cave, permafrost or lake sediments (Capo et al., 2021; Parducci et al., 2017; Pedersen et al., 2015) for extended periods of time (Kjær et al., 2022). *Seda*DNA has primarily been used for the identification of taxa or even reconstructing palaeoecologies (Epp et al., 2012; Parducci et al., 2019; Rijal et al., 2021; Willerslev et al., 2014). However, a number of studies have explored the potential of *seda*DNA for the retrieval of intraspecific variation through the usage of either shotgun metagenomic (Lammers et al., 2021; Meucci et al., 2021; Pedersen et al., 2021; Seersholm et al., 2016), qPCR melting curve assays (Nota et al., 2022) or hybridization capture techniques (Schulte et al., 2021; Zavala et al., 2021; Zhang et al., 2020). These methods, however, can be time intensive in terms of data generation and analysis, and in case of shotgun sequencing require a high concentration of the species of interest. More scalable methods are thus required for large-scale reconstructions of past populations.

Amplicon-based methods such as DNA metabarcoding have proven to be scalable when working with *seda*DNA, but as a single-barcode approach, have been limited to species-level or higher identifications. Multiplexing PCR methods combine multiple primer sets into a single reaction to simultaneously obtain information from different groups of organisms, for example, mammals and plants (De Barba et al., 2014; Taberlet et al., 2018). However, different primer sets can also be included in a reaction that is designed to amplify intraspecific regions, thus obtaining population-level genomic information. So far, the usage of multiplexing PCRs to obtain population-level data is limited to contemporary DNA (Andres et al., 2021; De Barba et al., 2017; Skrbinšek et al., 2010), whereas multiplexing PCRs on ancient sediments have been limited to species-level identifications (Côté et al., 2016). Applying the multiplexing PCR method to *seda*DNA comes with its own challenges given the damaged and highly fragmented nature of ancient DNA (Dabney et al., 2013; Orlando et al., 2021).

In this study, we investigate the potential of multiplexing PCRs for the retrieval of population-level genomic information from ancient sediments. We have selected the dwarf shrub *Vaccinium uliginosum* as our test species. *V. uliginosum* has well-known present-day intraspecific lineages (Alsos et al., 2005; Eidesen et al., 2007) and is frequently detected in *seda*DNA studies (Alsos et al., 2022; Clarke et al., 2019), making it an ideal candidate for method development. We designed multiple intraspecific primers for *V. uliginosum* and tested these on a range of *seda*DNA samples to explore the palaeophylogeography of the species and evaluate the applicability of the method.

2 | METHODS

2.1 | Study species

The dwarf shrub *V. uliginosum* (Ericaceae) has a present-day circumpolar and circumboreal distribution but can also be found in more southern mountain ranges such as the Alps and Pyrenees (Hultén, 1986). The species is long-lived, bird-dispersed and insect-pollinated, though self-pollination has been observed (Jacquemart, 1996). Furthermore, the species has been reported in *seda*DNA studies from Scandinavia (Alsos et al., 2022; Rijal et al., 2021), the Ural mountains (Clarke et al., 2019) and the Alps (Garcés-Pastor et al., 2022; van Vugt et al., 2022). A number of different subspecies or synonym species have been described, but these can be categorized into the following three lineages based on Amplified Fragment Length Polymorphism (AFLPs), ploidy and chloroplastic variation: Amphi-Atlantic, Arctic-Alpine and Beringian (Figure 1; Alsos et al., 2005; Eidesen et al., 2007).

The Amphi-Atlantic lineage is tetraploid and is widespread throughout Europe at lower elevations. Its distribution reaches from the northern Alps to Scandinavia, Iceland, southern Greenland and the northern Atlantic coast of North America (Alsos et al., 2005; Eidesen et al., 2007).

The Arctic-Alpine lineage is primarily diploid and can be split into two sublineages, namely the Arctic and Alpine sublineage. The Arctic sublineage has a circumpolar distribution and is found in northern Asia, Greenland, Svalbard and North America, with single populations in the Carpathians and northern Norway (Alsos et al., 2005). The Alpine sublineage on the other hand has a more limited distribution and is found in the Alps and Pyrenees (Figure 1b). The split of the Arctic-Alpine lineage is most likely the result of isolation during the Quaternary cold periods, where the Alpine sublineage became isolated in southern Europe in relation to the Arctic sublineage (Alsos et al., 2005; Eidesen et al., 2007).

Finally, there is the Beringian lineage, which contains a mixture of di-, tetra- and hexaploids. This lineage has a geographical range from northern Japan and the Kamchatka Peninsula to Alaska, with smaller populations extending down the North American west coast (Alsos et al., 2005; Eidesen et al., 2007; Hirao et al., 2011).

2.2 | *Vaccinium uliginosum* variant discovery

The *V. uliginosum* variant discovery was based on the genome skim data for 18 modern individuals that were either part of the PhyloNorway project ($n=16$; Alsos et al., 2020; Wang et al., 2021) or the PhyloAlps project ($n=2$; Table S1). A reference chloroplast genome was assembled based on individual TROM_V_355013, an Alpine individual from the French Alps, following the methods described in Alsos et al., 2020.

Intraspecific variable positions were identified by mapping each of the 18 individuals separately onto the assembled reference genome using *bowtie2* v2.3.4.1 (Langmead & Salzberg, 2012)

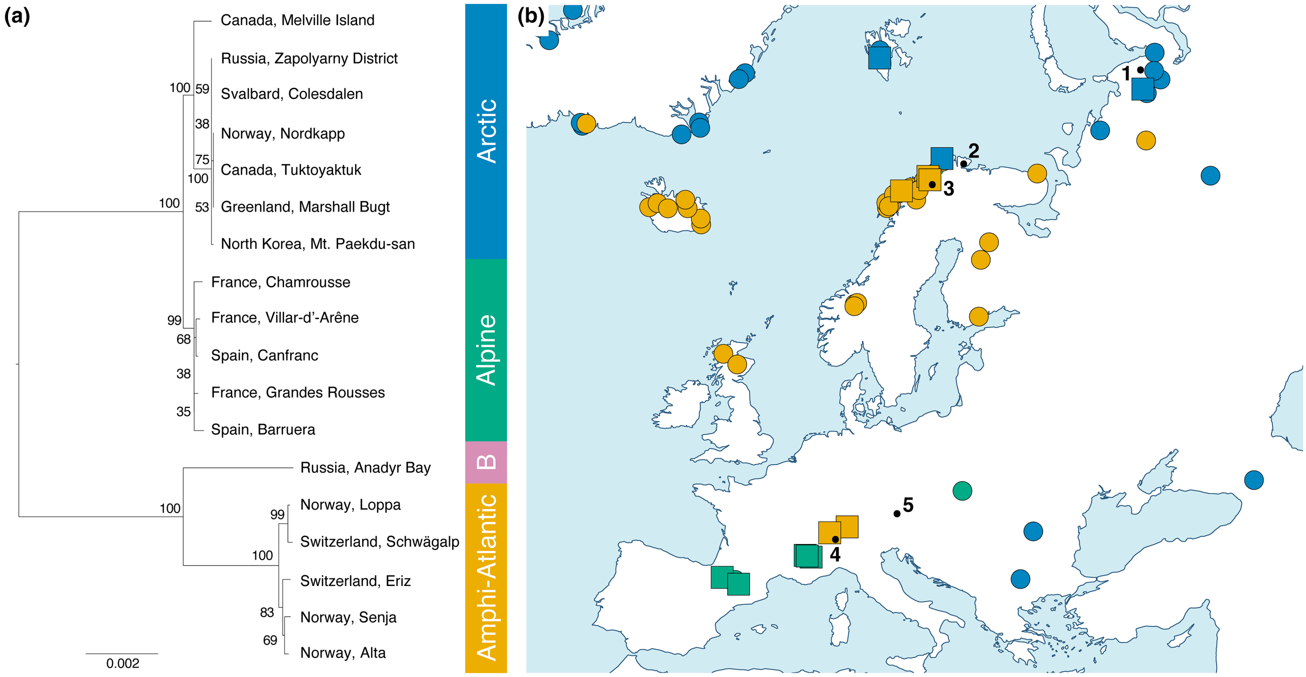


FIGURE 1 (a) Chloroplast maximum likelihood tree for the 18 *V. uliginosum* reference individuals. Node values indicate bootstrap support. (b) Map with the location of the European Arctic, Alpine and Amphi-Atlantic reference individuals indicated by the coloured boxes. Coloured circles indicate the location of previously genotyped individuals by Alsos et al. (2005). The black dots indicate the location of the five lake *sed*aDNA cores, (1) Bolshoye Shchuchye, (2) Nordvivatnet, (3) Jøkelvatnet, (4) Hopschusee and (5) Krumschnabelsee.

with default settings (Figure 2a). The resulting alignments were processed with *SAMtools* v1.7 (Li et al., 2009), in order to remove unaligned reads and duplicate reads with the *view* and *markdup* functions, respectively. Variants were called and filtered with *BCFtools* v1.9 (Li et al., 2009), using the *mpileup*, *call* and *index* functions, with the variants-only and multiallelic-caller options. The resulting variable positions were further filtered to only include those that had a mapping quality ≥30, coverage ≥20, and an alternative allele that was more abundant than the reference allele in case of haplotypic variation (Figure 2a).

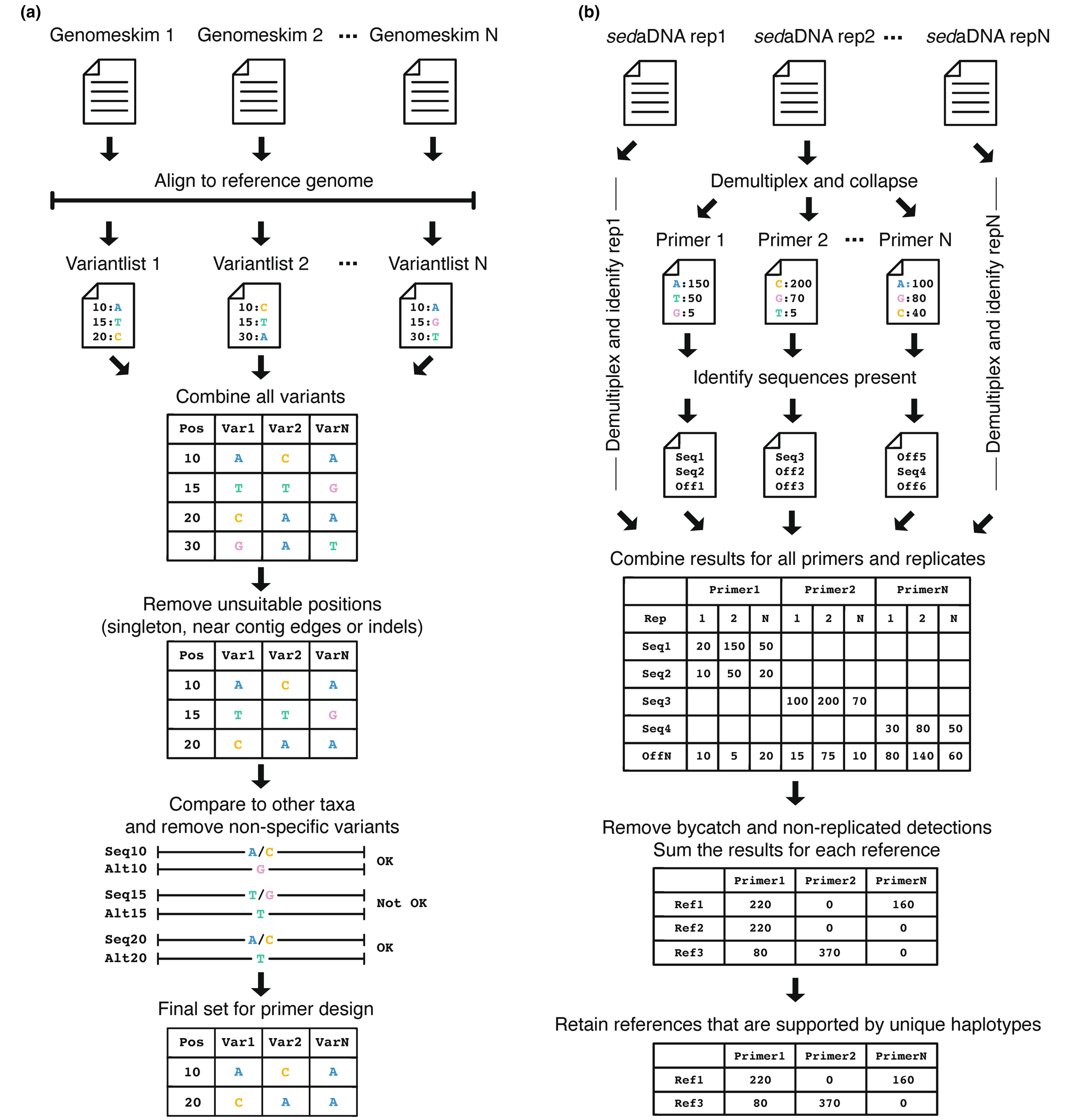
The variable positions for the 18 individuals were combined into one set and processed to find candidate regions for primer design (Figure 2a). First, variants that were located within 100 bp of a contig end were discarded, as these variants could both be the result of incorrect mapping and limit the space available for primer design. Second, variants located within 5 bp of an indel or homopolymer stretch of at least five bases were excluded to minimize the chance of misaligned reads. For the remaining variable positions, the haplotypes were called for each of the 18 individuals using the filtered *bowtie2* alignments. Each haplotype was constructed using the bases with a base quality ≥13 for assemblies with a coverage ≥20, and alleles with a frequency ≥0.1 in case of haplotypic variation (Figure 2a).

Individual variable positions were grouped into larger windows if two adjacent variants were within 50 bp of each other and treated as a single variable position onwards. To remove singleton haplotypes that carry limited population-level information, only positions that had at least two distinct haplotypes across the 18 individuals, with a minimum of two individuals per haplotype, were retained.

To ensure that each haplotype was specific to *V. uliginosum*, the variable positions were compared against all assembled chloroplast genomes in the PhyloNorway ($n=1969$) and PhyloAlps ($n=3923$) genome skim reference databases (Alsos et al., 2020; Wang et al., 2021). For each variable position, the haplotypes present, extended 20 bp up and downstream of the variant, were aligned against the reference databases with the *blastn* function from *NCBI-BLAST+* v2.2.18 (Camacho et al., 2009). Any variable position that contained a haplotype that was not specific to *V. uliginosum* was removed (Figure 2a).

2.3 | Ancient sediment samples

Twenty ancient lake sediment samples were selected for *V. uliginosum* multiplex amplification. The *sed*aDNA samples originated from five different lake sediment cores that are known to contain *V. uliginosum* DNA based on previous *trn*L P6-loop metabarcoding of the sediments, and they cover the present-day extent of the different European *V. uliginosum* lineages (Data S1). Two lakes, Jøkelvatnet and Nordvivatnet, are located in northern Norway (Rijal et al., 2021) where the Amphi-Atlantic lineage is dominating today; two, Hopschusee and Krumschnabelsee, are from the Alps (Garcés-Pastor et al., n.d.) where the Alpine sublineage is most prevalent today, and one, Bolshoye Shchuchye, from the Polar Urals, Russia (Clarke et al., 2019), where the Arctic sublineage is dominant today.



(a)
Genomeskim 1
Genomeskim 2
···
Genomeskim N
Align to reference genome
Variantlist 1
10:A
15:T
20:C
Variantlist 2
10:C
15:T
30:A
···
Variantlist N
10:A
15:G
30:T
Combine all variants

Pos	Var1	Var2	VarN
10	A	C	A
15	T	T	G
20	C	A	A
30	G	A	T

Remove unsuitable positions
(singleton, near contig edges or indels)

Pos	Var1	Var2	VarN
10	A	C	A
15	T	T	G
20	C	A	A

Compare to other taxa
and remove non-specific variants
Seq10 A/C
Alt10 G
OK
Seq15 T/G
Alt15 T
Not OK
Seq20 A/C
Alt20 T
OK
Final set for primer design

Pos	Var1	Var2	VarN
10	A	C	A
20	C	A	A

(b)
sedaDNA rep1
sedaDNA rep2
···
sedaDNA repN
Demultiplex and idenify rep1
Demultiplex and idenify repN
Demultiplex and collapse
Primer 1
A:150
T:50
G:5
Primer 2
C:200
G:70
T:5
···
Primer N
A:100
G:80
C:40
Identify sequences present
Seq1
Seq2
Off1
Seq3
Off2
Off3
Off5
Seq4
Off6
Combine results for all primers and replicates

	Primer1			Primer2			PrimerN		
Rep	1	2	N	1	2	N	1	2	N
Seq1	20	150	50						
Seq2	10	50	20						
Seq3				100	200	70			
Seq4							30	80	50
OffN	10	5	20	15	75	10	80	140	60

Remove bycatch and non-replicated detections
Sum the results for each reference

	Primer1	Primer2	PrimerN
Ref1	220	0	160
Ref2	220	0	0
Ref3	80	370	0

Retain references that are supported by unique haplotypes

	Primer1	Primer2	PrimerN
Ref1	220	0	160
Ref3	80	370	0

FIGURE 2 (a) Workflow utilized for the variant discovery. Starting with aligning N genomes or genome skims to a shared reference genome. The resulting variants are combined, and unsuitable positions (rare, near contig edges or indels) are removed. Each variant is compared to other reference taxa, and only those that are distinct are retained for primer design. (b) Workflow for the multiplex data analysis. Each replicate for a sample is demultiplexed separately, and for each primer present the variants are identified. After identification, the replicate data for a sample are combined. Off-target identifications and non-replicated variants are removed. Variants are retained if they are supported by at least one unique haplotype.

All sites are close to present-day meeting points between the lineages (Figure 1b).

Each site was represented by four *seda*DNA samples, one sample for the oldest known *V. uliginosum* detection, one for the most recent detection, and the remaining two samples spaced in between. The lake sediment cores were previously dated by radiocarbon dating of plant macrofossils using Accelerator Mass Spectrometry (AMS) at the Poznań Radiocarbon Laboratory. Bayesian age-depth models were constructed with “Bacon” v2.3.4 (Blaauw & Christen 2011) using the calibrated radiocarbon dates based on the IntCal13 curves (Reimer et al., 2013) for the Bolshoye Shchuchye, Jøkelvatnet and Nordvivatnet samples (Clarke et al., 2019; Rijal et al., 2021) and IntCal20 curves (Reimer et al., 2020). The age-depth models for Hopschusee and Krumschnabelsee were based on 25 and 12 radiocarbon dates, respectively, and followed the methods above (Garcés-Pastor et al., n.d.).

The material for Bolshoye Shchuchye, Jøkelvatnet and Hopschusee was previously extracted, along with three negative extraction controls (Garcés-Pastor et al., n.d.; Clarke et al., 2019; Rijal et al., 2021). The eight extracts used for Nordvivatnet and Krumschnabelsee were re-extracted for this study. All sampling and extractions were performed in a dedicated ancient DNA laboratory at the Arctic University Museum of Norway, UiT—The Arctic University of Norway, Tromsø, Norway. DNA was extracted from ~0.3 g sediment subsamples using a modified DNeasy PowerSoil kit (Qiagen, Germany). Modifications include a bead beating step and an initial lysis step with proteinase K, as described in (Alsos et al., 2020). A negative extraction control was included for the re-extraction of the Nordvivatnet and Krumschnabelsee sediments.

2.4 | Multiplex PCR

Multiplex primers were developed for all variable positions that passed the filtering criteria. Primer sets were developed using a 100 bp flanking region around the variable positions and were designed such that they included both the *V. uliginosum* intraspecific variable positions and any positions required for *V. uliginosum* specificity. The primer sets were optimized to have a short amplicon length of an average of 79 bp (min 46 bp, max 116 bp; Table S2), to ensure amplification of highly fragmented ancient material (Orlando et al., 2021). Furthermore, the melting temperature was kept similar across all primer sets at an average of 55.7°C (min 55°C, max 57.5°C) to promote even primer annealing and amplification. In total, 61 primer sets were developed and tested (Data S1; Table S2), which resulted in a final set of 38 primers (Table S2). Each primer was modified by adding one of two 3 bp tags on the 5′ end to allow for pooling of up to four PCR products (Binladen et al., 2007). These tags had an edit distance of three to avoid misassignment based on sequencing errors (Table S3). The 38 primer sets were pooled together, with different concentrations per primer set to account for the different amplification efficiencies (Data S1; Table S2).

All PCRs were prepared in a dedicated ancient DNA lab at the Arctic University Museum of Norway, UiT—The Arctic University of Norway, Tromsø, Norway. Each amplification was carried out in a 50 μL final volume, containing 25 μL of Platinum Multiplex PCR master mix (Life Technologies, Carlsbad, CA, USA), 15 μL of molecular grade water, 5 μL of the multiplex primers with a final concentration of 1.735 μM (Table S2) and 5 μL of DNA extract. Two negative PCR controls were included in the sample preparation. The following amplification protocol was used: enzyme activation step (2 min at 95°C), 45 PCR cycles of 30 s at 95°C, 90 s at 53°C and 30 s at 72°C, followed by a final elongation step of 10 min at 72°C. A annealing temperature of 2°C lower than the calculated 55°C temperature was used to increase the efficiency of the reaction (Taberlet et al., 2018). The amplification of the samples and controls was carried out using four multiplex PCR replicates. Positive amplification for a subset of the sample replicates was verified by electrophoresis on a 2% agarose gel. For each sample, an equal volume of PCR product from the four uniquely tagged replicates was pooled together and cleaned following (Voldstad et al., 2020). The pools were converted into DNA libraries with the Illumina TruSeq DNA PCR-Free protocol (Illumina Inc., CA, USA), with unique dual indexes and sequenced at 2 × 150 cycles on an Illumina NextSeq platform at the Genomics Support Centre Tromsø, UiT—The Arctic University of Norway.

2.5 | Analysis

The paired-end sequence data was merged and adapter-trimmed with SeqPrep (https://github.com/jstjohn/SeqPrep/releases, v1.2). For each library, the data were demultiplexed with the *ngsfilter* tool from OBITools (v1.2.12; Boyer et al., 2016) using default settings (Figure 2b). Identical reads were collapsed with *obiuniq*, and singleton sequences were removed with *obigrep*. PCR artefacts were identified with *obiclean* using a head: internal ratio of 0.05. Finally, the sequences were identified using *ecotag* and a curated reference database (Figure 2b). The reference database was created with *ecoPCR* v1.0.1 (Ficetola et al., 2010), using the multiplex primer sets and the reference genomes from the PhyloAlps (4604 specimens of 4437 taxa) and PhyloNorway (2051 specimens of 1899 taxa) projects (Alsos et al., 2020; Wang et al., 2021), allowing for a maximum of two mismatches between the primer sets and references. The identified data were exported with *obitab* for downstream analysis.

The data were further analysed with a custom Python script. Rare sequence detections, defined as those with less than three reads per PCR repeat, were removed from the dataset, as well as any sequences identified as “internal” by *obiclean*. For each sample, the four replicates were merged together, calculating the replication for each sequence in the process. Any sequence that was not present in at least two replicates was removed. The sequences that were identified as *V. uliginosum* were compared to the known haplotypes of the 18 reference individuals. For each sample, a reference individual, or individuals in case of identical haplotypes, was considered present if it was supported by at least one unique haplotype. The reads for each haplotype were split among the present reference individuals. Furthermore, for each sample, the *V. uliginosum* lineages present were determined as well. A lineage was considered present if there was at least one haplotype specific to the lineage detected in a sample (Figure 2b).

3 | RESULTS

3.1 | *Vaccinium uliginosum* variant discovery

Phylogenetic analysis of the 18 *Vaccinium uliginosum* reference chloroplast genomes indicated that all four lineages and sublineages were covered. Seven individuals belonged to the Arctic sublineage, five to the Alpine sublineage, five to the Amphi-Atlantic lineage and finally one individual from the Bergingian lineage (Figure 1a).

A total of 1059 variable positions were found across the 18 *Vaccinium uliginosum* reference individuals. This was reduced to 704 variable positions after those close to contig edges, indels or homopolymers were removed. Grouping of adjacent variable positions into windows and removal of non-descriptive positions reduced the number to 292 variable positions. Out of these, 88 were fully specific to *V. uliginosum* and 81 were selected for multiplex primer design (Table S2). The remaining seven variable regions were excluded due to a high (>0.2) proportion of haplotypic variation among the individuals.

3.2 | Multiplex PCR performance

Multiplex-compatible primer sets were developed for 61 of the variable regions. Initial multiplex PCR tests indicated that 52 of these had both successful amplification and informative results (Data S1). A final set of 38 primers that had the highest read counts in testing were used. Phylogenetic analysis of just the amplifiable regions revealed that they are capable of identifying the four *Vaccinium uliginosum* lineages, though five out of the seven Arctic, two out of the five Alpine and two out of the five Amphi-Atlantic reference individuals had identical haplotypes within their lineage and thus cannot be identified down to the exact reference individual by the primer sets (Figure S1).

We generated 98,947,000 paired-end reads for the sediment samples and controls. After merging and demultiplexing, we obtained 22,376,000 reads, out of which 19,050,000 reads remained after filtering and identification. Of the identified reads, 4,177,000 were identified as *V. uliginosum* and the remainder as off-target bycatch (Table S4). The most common off-target taxon was identified as *Vaccinium*, containing sequences that could either be identified as *V. myrtillus* or *V. vitis-idaea*. *Vaccinium uliginosum* was identified in all 20 sediment samples and was absent from the controls. The controls instead contained sequences identified to higher taxonomic levels (Spermatophyta, asterids, Pentapetalae) and common food contaminants (*Solanum*) (Table S4). Each sediment sample contained on average 28.1 (SD 11.07) and 34.15 (SD 15.4) on-target primer sets and variant sequences, respectively (Table S4). Seven sediment samples had on-target amplification for all 38 primer sets, while the lowest number of on-target primer sets was six, for a sediment sample from Nordvivatnet (Table S4).

All 38 primer sets could be observed in the sediment samples, where on average each primer could successfully amplify *V. uliginosum* in 14.79 (SD 3.11) sediment samples (Table S5). Furthermore, three primer sets had successful on-target amplification in all sediment samples (Table S5).

After filtering and identifying the reads, on average, each primer had 501,000 reads assigned to it, of which 110,000 reads were identified as *V. uliginosum*. The most abundant primer set was Vu_62, with 7,228,000 reads, though only 132,000 reads were on-target. The most abundant on-target primer set was Vu_16, with 699,000 *V. uliginosum* reads across all samples (Tables S5 and S6).

In total, 89 different on-target variant sequences were obtained for the samples, with an average of 2.34 (SD 0.58) sequences per primer (Table S5). The average on-target replication for these sequences was 3.51 (SD 0.75), with 86 out of the 89 variant sequences being fully replicated for at least one sample (Table S7).

All four *V. uliginosum* lineage and sublineages were represented in our data. The Amphi-Atlantic lineage was the most common lineage with 15 detections, while the Beringian lineage was the rarest, with one detection. The Arctic and Alpine sublineages had 5 and 10 detections, respectively. On average, we detected 1.55 lineages per sample, with a minimum of one and a maximum of three lineages per sample (Tables S8–S10). All reference individuals within the lineages were present for at least one sediment sample and were represented by at least one fully replicated primer (Table S8).

When subsampling the data, the main lineage, as defined by containing more than >1% of the unique reads, is present in most samples from 500 reads per repeat onwards. Most additional lineages are also present from 500 reads per repeat onwards, though some lineages, which are only represented by one repeat, are never detected at stricter filtering regimes (Figure S2). The number of identified references on the other hand shows a declining trend for all filtering methods as the number of reads per subsample increases (Figure S3).

3.3 | *Vaccinium uliginosum* sedaDNA results per region

The Bolshoye Shchuchye sediment samples primarily contained reads assigned to the Arctic *V. uliginosum* sublineage. The two oldest samples exclusively contained sequences and repeats assigned to the Arctic sublineage, while the two youngest samples contained material assigned to both the Arctic sublineage and the Amphi-Atlantic lineage, where the Arctic sublineage was the more abundant lineage, with 99.9% and 92.2% of the unique reads and repeats, respectively (Table 1; Figure 3; Tables S8 and S9).

Both northern Norwegian lakes mainly contained material assigned to the Amphi-Atlantic lineage, the most dominant lineage in the region today. For Nordvivatnet, the Amphi-Atlantic lineage was the only lineage in the oldest three samples, but in the most recent sample 10% and 16% of the reads and repeats, respectively, were assigned to the Arctic sublineage. Jøkelvatnet, again mainly contained material identified to the Amphi-Atlantic lineage, but in addition contained material identified as the Alpine sublineage in all samples. The Alpine sublineage, however, was present, on average, in only 0.6% and 7% of the reads and repeats, respectively, across the four samples (Table 1; Figure 3; Tables S8 and S9).

Similar to the Norwegian sites, the Alpine sites contained a mixture of *V. uliginosum* material. The Alpine sublineage was the main lineage in Hopschusee. It was the only lineage present from

TABLE 1 The proportion of unique reads assigned to each *V. uliginosum* lineage for the 20 lake *seda* DNA samples.

Lake	Age (cal. Yr BP)	Proportion of unique Arctic reads	Proportion of unique Alpine reads	Proportion of unique Beringian reads	Proportion of unique Amphi-Atlantic reads
Bolshoye Shchuchye	11,186	1	0	0	0
Bolshoye Shchuchye	8134	1	0	0	0
Bolshoye Shchuchye	4207	0.999	0	0	0.001
Bolshoye Shchuchye	1322	0.999	0	0	0.001
Nordvivatnet	9415	0	0	0	1
Nordvivatnet	6729	0	0	0	1
Nordvivatnet	3204	0	0	0	1
Nordvivatnet	477	0.104	0	0	0.896
Jøkelvatnet	9697	0	0.004	0	0.996
Jøkelvatnet	5954	0	0.003	0	0.997
Jøkelvatnet	2060	0	0.009	0	0.991
Jøkelvatnet	17	0	0.007	0	0.993
Hopschusee	10,561	0	1	0	0
Hopschusee	5636	0	1	0	0
Hopschusee	3151	0	1	0	0
Hopschusee	1240	0	0.664	0.007	0.329
Krumschnabelsee	2919	0	0	0	1
Krumschnabelsee	1801	0	0.636	0	0.364
Krumschnabelsee	622	0	0.502	0	0.498
Krumschnabelsee	244	0	0	0	1

10,500–3100 cal. yr BP and the dominant lineage in the youngest sample. The youngest sample also contained material assigned to the Amphi-Atlantic and Beringian lineages, at 32.9% and 0.7% of the reads, respectively. The sediment samples from Krumschnabelsee only cover the period 2900 cal. yr BP to present, and it contained primarily Amphi-Atlantic material, which was the only lineage in the oldest and youngest samples from this site. The two samples from 1800 to 600 cal. yr BP had an even split between the Alpine sub-lineage and Amphi-Atlantic lineage for both the reads and repeats (Table 1; Figure 3; Tables S8 and S9).

4 | DISCUSSION

4.1 | Multiplex PCR results strengths and weaknesses

Our *V. uliginosum* results indicate that we can reliably amplify multiple informative target regions per *seda*DNA sample. Despite the low coverage of some of our samples, we were still able to determine the main *V. uliginosum* lineages present. Narrowing down the exact reference genomes present is more challenging due to haplotype sharing between a number of reference individuals (Figure S1). As a result, we conclude that the method presented here is most robust for identification of the *V. uliginosum* lineages and that identifications down to an individual reference genome should be interpreted with caution.

The samples with the lowest number of on-target primer sets tended to have a lower on-target replication and relatively few reads assigned to *V. uliginosum* (Tables S4 and S7). The low *V. uliginosum* read coverage observed is mainly the result of off-target amplification. A number of primer sets, most notably Vu_62, are able to amplify taxa outside of *V. uliginosum* and Ericaceae. The amplification of off-target material will both reduce the relative amount of *V. uliginosum* material in these primer sets and reduce the overall amount of *V. uliginosum* within the multiplex reaction. That combined with the already low proportion of *V. uliginosum* template material for some samples, as estimated from previous *trn*L P6-loop metabarcoding results, can result in a limited amount of usable data (Table S4). Removal of less specific primer sets should increase the amount of assigned *V. uliginosum* sequences and result in improved identifications.

This study utilized 38 primer sets to identify the different *V. uliginosum* lineages. Given the experimental nature of this study, redundancy was built into this set of primers, to ensure that the main lineages could be identified even in the event of non-amplification of the target regions. For follow-up studies, this set of primers can be reduced to improve the read coverage per primer at equal sequencing depths or alternatively open up the possibility to include different primer sets targeting different regions or species.

Finally, the primer sets designed for this study primarily differentiate between the Amphi-Atlantic, Arctic and Alpine lineages, given the focus on European *seda*DNA samples. Although a Beringian *V. uliginosum* sample was included in the set used for

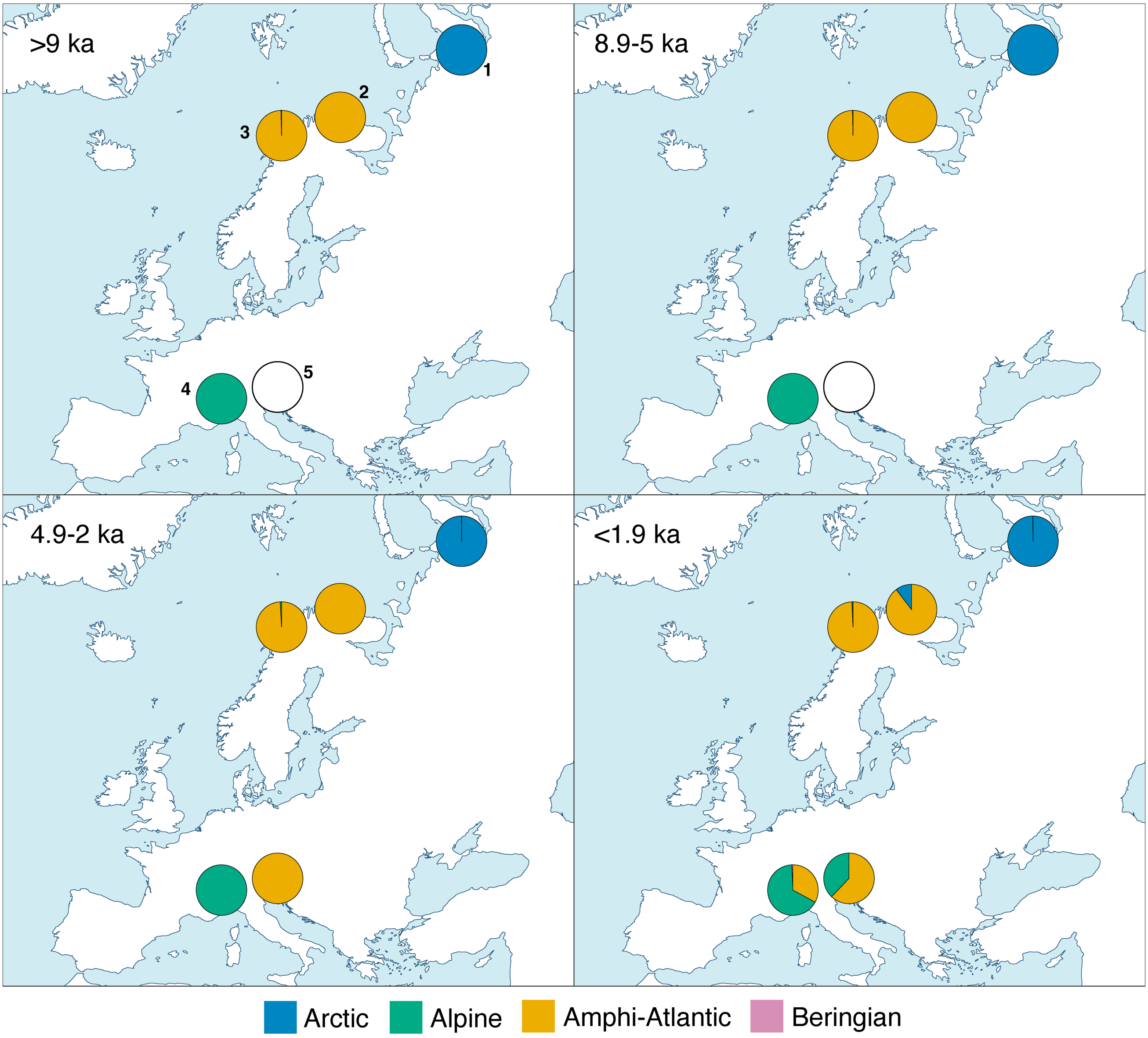


FIGURE 3 Multiplex results for the five lake *seda*DNA cores: (1) Bolshoye Shchuchye, (2) Nordvivatnet, (3) Jøkelvatnet, (4) Hopschusee and (5) Krumschnabelsee. Each lake is represented by a pie chart indicating the proportion of unique reads supporting each lineage. Raw values are provided in Table 1 and Table S9. The oldest sample from Krumschnabelsee (5) is from 2900 cal. yr BP and is thus empty for the periods >9 ka and 8.9–5 ka. The chart in period <1.9 ka for Krumschnabelsee is an average of three samples.

primer design and as a reference, the Beringian region contains substantial variation that likely is not fully captured by a single individual (Alsos et al., 2005; Alsos, 2003; Brochmann et al., 2004; Eidesen et al., 2007). Applying the multiplex method and the primer sets developed here to *seda*DNA samples from Beringia would require sequencing of additional reference individuals and inspecting the amplifiable variation present prior to analysis of *seda*DNA samples.

4.2 | *Vaccinium uliginosum* phylogeography

Due to the proximity of the potential *V. uliginosum* Arctic sublineage refugia east of the Scandinavian ice sheet (Hughes et al., 2016) to northern Norway, we assumed that the first *V. uliginosum* colonization of this region would originate from the Arctic sublineage, with later replacement of the boreal Amphi-Atlantic lineage observed in the region today. Our results, however, indicate that the Amphi-Atlantic lineage was first to arrive in northern Norway, as was detected in Jøkelvatnet and Nordvivatnet. In the case of the more western Jøkelvatnet, this was paired with a background presence of the Alpine sublineage, which could be the result of unknown Arctic variation not present in our reference set, suggesting potential long-term presence of the Arctic-Alpine lineage. On the other hand, at Nordvivatnet only the most recent sample contains a small component of Arctic sublineage reads and repeats. Thus, to determine if the presence of a modern Arctic *V. uliginosum* population at North

Cape (Alsos et al., 2005; Eidesen et al., 2007) is an ancient or more recent introduction in Norway, more sites and samples need to be analysed.

The Alps provide a more dynamic story. Hopschusee was first colonized by the Alpine sublineage, which remained the only lineage present until the arrival of the Amphi-Atlantic lineage in the youngest sample, where both lineages are now present in roughly even amounts. The introduction of the Amphi-Atlantic lineage could represent the migration of the more boreal lineage to higher altitudes. The Beringian lineage is also present in the youngest sample, though at a low abundance. This detection is based on just two haplotype sequences and could be the result of unknown Alpine or Amphi-Atlantic variation that is overlapping with the variation found within the Beringian lineage. *Vaccinium uliginosum* is a relatively recent detection in Krumschnabelsee, where it was first detected at 3000 cal. yr BP although the core goes back to 8600 cal. yr BP. The Amphi-Atlantic lineage was the first to appear but was joined by the Alpine sublineage between 1800 and 600 cal. yr BP. The most recent sediment sample, representing the present, however, indicates that the Amphi-Atlantic lineage is currently the only lineage remaining.

The Bolshoye Shchuchye samples, representing the Polar Urals, are the most stable in terms of *V. uliginosum* lineages present. The Arctic sublineage is the main lineage present, where only the two most recent samples have a minor representation of the Amphi-Atlantic lineages. The present-day extent of the Amphi-Atlantic lineages reaches the Urals (Eidesen et al., 2007), so the more recent detection of this lineage in the Polar Urals samples could indicate a recent background presence in the region.

Although this dataset already provided some novel insights regarding *V. uliginosum* lineage migration, more work needs to be done to provide a complete overview of the species post-glacial migration. This study is limited in its spatial and temporal resolution. Samples were selected to cover a geographically wide area to increase the chance of observing the three European *V. uliginosum* lineages and allow for the evaluation of the multiplex PCR's performance. For a full reconstruction of the species migration, more *sed*aDNA sampling locations are required along the migration paths in Scandinavia, and potential refuge locations such as Western Europe, the Urals and Pyrenees. Furthermore, the temporal resolution needs to be improved so that both the timing of migration events can be narrowed down and increase the ability to observe short-term population turnover.

4.3 | Future applications of multiplexing PCRs

The success of the multiplexing method for *sed*aDNA has implications for other taxa and studies. First, a set of multiplex PCR primers can be used to improve the taxonomic resolution of a group of interest (Côté et al., 2016) or resolve cryptic species identifications (Brosseau et al., 2019), compared to single-barcode methods. Second, combining multiple metabarcoding primer sets into a single multiplex PCR can greatly increase the taxonomic information obtained from a single amplification run (De Barba et al., 2014; Schuette et al., 2022; Weber et al., 2023). A traditional metabarcoding approach can achieve similar results by running multiple distinct metabarcoding runs using different primer sets (Epp et al., 2012; Garcés-Pastor et al., 2022; Willerslev et al., 2014); however, this comes at the cost of additional laboratory expenses and, especially in the case of *sed*aDNA, usage of available DNA extracts.

The multiplex PCR method on the other hand requires some careful planning. First, the method, similar to shotgun metagenomics and hybridization capture methods, requires the presence of an extensive reference database. This database is needed to both identify the intraspecific regions of interest and avoid false positives caused by regions shared among closely related species. Furthermore, the identification of extinct lineages, regardless of *sed*aDNA method, can be problematic if no appropriate reference material is available in the reference databases. Second, to ensure sufficient template material for amplification, multicopy regions are preferred, such as plastid or ribosomal DNA. In addition, taxa with a low biomass might not produce the template DNA required for reliable amplification across all primer sets. Third primer sets must be carefully designed to ensure that they have comparable annealing temperatures and do not form dimers. Fourth, primer sets need to be specific to the target taxa, which requires access to complete reference libraries for both design and identification of the amplicons. Finally, different primer sets and taxa will differ in available template DNA, which can lead to uneven amplicon concentrations and read counts. This can be solved by adjusting the primer concentrations to match the available template, but this can take time to fully optimize (Taberlet et al., 2018). Furthermore, since multiplexing does not retain deamination profiles that are commonly used to authenticate ancient DNA (Dabney et al., 2013; Orlando et al., 2021), sufficient negative controls need to be incorporated to avoid detections based on contamination (Pedersen et al., 2015; Taberlet et al., 2018).

The multiplex PCR method presented here for generating population-level genomic data from *sed*aDNA samples has some advantages compared to the more commonly used shotgun metagenomic and hybridization capture methods. The main advantage is the scalability and time effectiveness of the method. A multiplex reaction, similar to a metabarcode reaction, is relatively simple to set up compared to more complex methods such as hybridization capture (Schulte et al., 2021). Furthermore, the multiplex method requires a lower sequencing depth compared to shotgun metagenomics (Kjær et al., 2022; Wang et al., 2021), which results in lower sequencing costs. The bioinformatic analysis of the multiplex data is also relatively simple compared to shotgun metagenomics and hybridization capture, as the data generated can be processed to some extent by well-established metabarcoding pipelines, which reduces the analysis time considerably. All these factors combined allow the multiplex PCR method to generate population-level genomic data for a large number of samples and thus enable the potential to track populations through time and space, opening up the field of palaeo-phylogeography.

AUTHOR CONTRIBUTIONS

Y.L., P.T. and I.G.A. conceptualized the study; Y.L., P.T., L.D.E., M.F.M. and I.P. were involved in generating the data; Y.L., P.T. and E.C. performed the analysis of sequence data; Y.L. wrote the manuscript with feedback from all co-authors.

ACKNOWLEDGEMENTS

We thank Ingebjørg Fagerheim Tokle for her assistance with the multiplex PCR tests, Sandra Garcés-Pastor for providing help with the sediment extractions, Scarlett Zetter for aiding with the library preparation, Lasse Topstad for collecting fresh *V. uliginosum* samples and Owen S. Wangensteen for preparing the MiSeq sequencing run for the multiplex PCR test libraries. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement no. 819192 for the IceAGenT project. The PhyloNorway project was funded by the Research Council of Norway (226134/F50) and the Norwegian Biodiversity Information Centre (14-14, 70184209). The PhyloAlps project was funded by France Génomique (ANR-10-INBS-09-08). Sediment samples were obtained through the ECOGEN and CHASE projects, both funded by the Research Council of Norway (grant number: 250963/F20 and 255415 respectively). Bioinformatic analyses were performed on resources provided by UNINETT Sigma2—the National Infrastructure for High-Performance Computing and Data Storage in Norway.

DATA AVAILABILITY STATEMENT

The raw DNA sequence data generated for both the test multiplex analysis and main experiment have been deposited in the European Nucleotide Archive (ENA) under BioProject accession code PRJEB65305. The identified OBITools output for the multiplex data and the analysis Python scripts are available on Dryad: https://doi.org/10.5061/dryad.4f4qrfjjc.

COLLABORATORS

Members of the PhyloAlps consortium: Lavergne S., Pouchon C., Coissac E., Roquet C., Smyčka J., Boleda M., Thuiller W., Gielly L., Taberlet P., Rioux D., Boyer F., Hombiat A., Bzeznick B. (Laboratoire d'Ecologie Alpine, CNRS, UGA, Grenoble, France); Alberti A., Denoeud F., Wincker P., Orvain C. (Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Univ. Paris-Saclay, France); Perrier C., Douzet R., Rome M., Valay J.G., Aubert S. (Jardin Alpin du Lautaret, CNRS, UGA, Grenoble, France); Zimmermann N.E., Wüest R.O., Latzin S., Wipf S. (Swiss Federal Research Institute WSL, Birmensdorf, Switzerland); Van Es J., Garraud L., Villaret J.C., Abdulhak S., Bonnet V., Huc S., Fort N., Legland T., Sanz T., Pache G., Mikolajczak A. (Conservatoire Botanique National Alpin, Gap, France); Noble V., Michaud H., Offerhaus B., Pires M., Morvant Y. (Conservatoire Botanique National Méditerranéen, Hyères, France); Dentant C., Salomez P., Bonet R. (Parc National des Ecrins, Gap, France); Delahaye T. (Parc National de la Vanoise, Chambery, France); Leccia M.F., Perfus M. (Parc National du Mercantour, Nice, France); Eggenberg S., Möhl A. (Info-Flora, Bern, Switzerland); Hurdu B., Puşcaş M. (Babeş Bolyai University, Institute of Biological Research, Cluj Napoca, Romania), Slovák M. (Institute of Botany, Bratislava, Slovakia). Members of the PhyloNorway consortium: Alsos I.G., Merkel M.F., Lammers Y. (The Arctic University Museum of Norway, UiT—The Arctic University of Norway, Tromsø, Norway), Coissac E., Pouchon C. (Laboratoire d'Ecologie Alpine, CNRS, UGA, Grenoble, France); Alberti A., Denoeud F., Wincker P. (Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Univ. Paris-Saclay, France).

ORCID

Y. Lammers https://orcid.org/0000-0003-0952-2668
P. Taberlet https://orcid.org/0000-0002-3554-5954
E. Coissac https://orcid.org/0000-0001-7507-6729
L. D. Elliott https://orcid.org/0000-0002-0099-8005
M. F. Merkel https://orcid.org/0000-0002-5072-1071
I. G. Alsos https://orcid.org/0000-0002-8610-1085

REFERENCES

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P. B., Schroeder, H., Ahlström, T., Vinner, L., Malaspinas, A.-S., Margaryan, A., Higham, T., Chivall, D., Lynnerup, N., Harvig, L., Baron, J., Casa, P. D., Dąbrowski, P., … Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, *522*(7555), 167–172. https://doi.org/10.1038/nature14507

Alsos, I. G. (2003). *Conservation biology of the most thermophilous plant species in the Arctic: Genetic variation, recruitment and phylogeography in a changing climate. (Doctoral dissertation).*

Alsos, I. G., Engelskjøn, T., Gielly, L., Taberlet, P., & Brochmann, C. (2005). Impact of ice ages on circumpolar molecular diversity: Insights from an ecological key species. *Molecular Ecology*, *14*(9), 2739–2753. https://doi.org/10.1111/j.1365-294X.2005.02621.x

Alsos, I. G., Lavergne, S., Merkel, M. K. F., Boleda, M., Lammers, Y., Alberti, A., Pouchon, C., Denoeud, F., Pitelkova, I., Puşcaş, M., Roquet, C., Hurdu, B. I., Thuiller, W., Zimmermann, N. E., Hollingsworth, P. M., & Coissac, E. (2020). The treasure vault can be opened: Large-scale genome skimming works well using herbarium and silica gel dried material. *Plants*, *9*(4), 432. https://doi.org/10.3390/plants9040432

Alsos, I. G., Rijal, D. P., Ehrich, D., Karger, D. N., Yoccoz, N. G., Heintzman, P. D., Brown, A. G., Lammers, Y., Pellissier, L., Alm, T., Bråthen, K. A., Coissac, E., Merkel, M. K. F., Alberti, A., Denoeud, F., Bakke, J., & PhyloNorway Consortium. (2022). Postglacial species arrival and diversity buildup of northern ecosystems took millennia. *Science. Advances*, *8*(39), eabo7434. https://doi.org/10.1126/sciadv.abo7434

Andres, K. J., Sethi, S. A., Lodge, D. M., & Andrés, J. (2021). Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples. *Molecular Ecology*, *30*(3), 685–697. https://doi.org/10.1111/mec.15765

Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One*, *2*(2), e197. https://doi.org/10.1371/journal.pone.0000197

Blaauw, M., & Christen, J. A. (2011). Flexible paleoclimate age-depth models using an autoregressive gamma process. Bayesian Analysis, 6(3). https://doi.org/10.1214/11-ba618

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182. https://doi.org/10.1111/1755-0998.12428

Brochmann, C., Brysting, A. K., Alsos, I. G., Borgen, L., Grundt, H. H., Scheen, A. C., & Elven, R. (2004). Polyploidy in arctic plants. *Biological Journal of the Linnean Society*, *82*(4), 521–536. https://doi.org/10.1111/j.1095-8312.2004.00337.x

Brosseau, L., Udom, C., Sukkanon, C., Chareonviriyaphap, T., Bangs, M. J., Saeung, A., & Manguin, S. (2019). A multiplex PCR assay for the identification of five species of the *Anopheles barbirostris* complex in Thailand. *Parasites & Vectors*, *12*(1), 223. https://doi.org/10.1186/s13071-019-3494-8

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Capo, E., Giguet-Covex, C., Rouillard, A., Nota, K., Heintzman, P. D., Vuillemin, A., Ariztegui, D., Arnaud, F., Belle, S., Bertilsson, S., Bigler, C., Bindler, R., Brown, T., Clarke, C. L., Crump, S. E., Debroas, D., Englund, G., Ficetola, G. F., Garner, R., … Parducci, L. (2021). Lake sedimentary DNA research on past terrestrial and aquatic biodiversity: Overview and recommendations. *Quaternary*, *4*(1), 6. https://doi.org/10.3390/quat4010006

Clarke, C. L., Edwards, M. E., Gielly, L., Ehrich, D., Hughes, P. D. M., Morozova, L. M., Haflidason, H., Mangerud, J., Svendsen, J. I., & Alsos, I. G. (2019). Persistence of arctic-alpine flora during 24,000 years of environmental change in the Polar Urals. *Scientific Reports*, *9*(1), 19613. https://doi.org/10.1038/s41598-019-55989-9

Côté, N. M. L., Daligault, J., Pruvost, M., Bennett, E. A., Gorgé, O., Guimaraes, S., Capelli, N., Le Bailly, M., Geigl, E. M., & Grange, T. (2016). A new high-throughput approach to genotype ancient human gastrointestinal parasites. *PLoS One*, *11*(1), e0146230. https://doi.org/10.1371/journal.pone.0146230

Dabney, J., Meyer, M., & Pääbo, S. (2013). Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, *5*(7), a012567. https://doi.org/10.1101/cshperspect.a012567

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. *Molecular Ecology Resources*, *14*(2), 306–323. https://doi.org/10.1111/1755-0998.12188

De Barba, M., Miquel, C., Lobréaux, S., Quenette, P. Y., Swenson, J. E., & Taberlet, P. (2017). High-throughput microsatellite genotyping in ecology: Improved accuracy, efficiency, standardization and success with low-quantity and degraded DNA. *Molecular Ecology Resources*, *17*(3), 492–507. https://doi.org/10.1111/1755-0998.12594

Edwards, S. V., Robin, V. V., Ferrand, N., & Moritz, C. (2022). The evolution of comparative phylogeography: Putting the geography (and more) into comparative population genomics. *Genome Biology and Evolution*, *14*(1), evab176. https://doi.org/10.1093/gbe/evab176

Eidesen, P. B., Alsos, I. G., Popp, M., Stensrud, Ø., Suda, J., & Brochmann, C. (2007). Nuclear vs. plastid data: Complex Pleistocene history of a circumpolar key species. *Molecular Ecology*, *16*(18), 3902–3925. https://doi.org/10.1111/j.1365-294X.2007.03425.x

Epp, L. S., Boessenkool, S., Bellemain, E. P., Haile, J., Esposito, A., Riaz, T., Erséus, C., Gusarov, V. I., Edwards, M. E., Johnsen, A., Stenøien, H. K., Hassel, K., Kauserud, H., Yoccoz, N. G., Bråthen, K. A., Willerslev, E., Taberlet, P., Coissac, E., & Brochmann, C. (2012). New environmental metabarcodes for analysing soil DNA: Potential for studying past and present ecosystems. *Molecular Ecology*, *21*(8), 1821–1833. https://doi.org/10.1111/j.1365-294X.2012.05537.x

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., Taberlet, P., & Pompanon, F. (2010). An in silico approach for the evaluation of DNA barcodes. *BMC Genomics*, *11*, 434. https://doi.org/10.1186/1471-2164-11-434

Garcés-Pastor, S., Coissac, E., Lavergne, S., Schwörer, C., Theurillat, J.-P., Heintzman, P. D., Wangensteen, O. S., Tinner, W., Rey, F., Heer, M., Rutzer, A., Walsh, K., Lammers, Y., Brown, A. G., Goslar, T., Rijal, D. P., Karger, D. N., Pellissier, L., The PhyloAlps Consortium, … Alsos, I. G. (2022). High resolution ancient sedimentary DNA shows that alpine plant diversity is associated with human land use and climate change. *Nature Communications*, *13*(1), 6559. https://doi.org/10.1038/s41467-022-34010-4

Garcés-Pastor, S., Heintzman, P. D., Zetter, S., Yoccoz, N., Brown, A. G., Lammers, Y., Vannière, B., Tribsch, A., Wangensteen, O. S., Schwörer, C., van Vugt, L., Rey, F., Heiri, O., Tinner, W., Coissac, E., Lavergne, S., Giguet-Covex, C., Walsh, K., Ficetola, F., … Alsos, I. G. *Impact of climate and domesticated mammals on Holocene plant richness in the European Alps revealed by sedaDNA* (Unpublished).

Heintzman, P. D., Froese, D., Ives, J. W., Soares, A. E. R., Zazula, G. D., Letts, B., Andrews, T. D., Driver, J. C., Hall, E., Hare, P. G., Jass, C. N., MacKay, G., Southon, J. R., Stiller, M., Woywitka, R., Suchard, M. A., & Shapiro, B. (2016). Bison phylogeography constrains dispersal and viability of the Ice Free Corridor in western Canada. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(29), 8057–8063. https://doi.org/10.1073/pnas.1601077113

Hirao, A. S., Sato, T., & Kudo, G. (2011). *Beringia, the phylogeographic origin of a circumpolar plant, Vaccinium uliginosum, in the Japanese archipelago*. Acta Phytotaxonomica et Geobotanica.

Hughes, A. L. C., Gyllencreutz, R., Lohne, Ø. S., Mangerud, J., & Svendsen, J. I. (2016). The last Eurasian ice sheets—a chronological database and time-slice reconstruction, DATED-1. *Boreas*, *45*(1), 1–45. https://doi.org/10.1111/bor.12142

Hultén, E. (1986). *Atlas of north European vascular plants north of the tropic of cancer*. (M. Fries, Trans.). Koeltz Scientific Books.

Jacquemart, A.-L. (1996). Vaccinium Uliginosum L. *The Journal of Ecology*, *84*(5), 771. https://doi.org/10.2307/2261339

Kjær, K. H., Winther Pedersen, M., De Sanctis, B., De Cahsan, B., Korneliussen, T. S., Michelsen, C. S., Sand, K. K., Jelavić, S., Ruter, A. H., Schmidt, A. M. A., Kjeldsen, K. K., Tesakov, A. S., Snowball, I., Gosse, J. C., Alsos, I. G., Wang, Y., Dockter, C., Rasmussen, M., Jørgensen, M. E., … Willerslev, E. (2022). A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. *Nature*, *612*(7939), 283–291. https://doi.org/10.1038/s41586-022-05453-y

Lammers, Y., Heintzman, P. D., & Alsos, I. G. (2021). Environmental palaeogenomic reconstruction of an ice age algal population. *Communications Biology*, *4*(1), 220. https://doi.org/10.1038/s42003-021-01710-4

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

McGaughran, A., Liggins, L., Marske, K. A., Dawson, M. N., Schiebelhut, L. M., Lavery, S. D., Moritz, C., & Riginos, C. (2022). Comparative phylogeography in the genomic age: Opportunities and challenges. *Journal of Biogeography*, *49*, 2130–2144. https://doi.org/10.1111/jbi.14481

Meucci, S., Schulte, L., Zimmermann, H. H., Stoof-Leichsenring, K. R., Epp, L., Bronken Eidesen, P., & Herzschuh, U. (2021). Holocene chloroplast genetic variation of shrubs (*Alnus alnobetula*, *Betula nana*, *Salix* sp.) at the siberian tundra-taiga ecotone inferred from modern chloroplast genome assembly and sedimentary ancient DNA analyses. *Ecology and Evolution*, *11*(5), 2173–2193. https://doi.org/10.1002/ece3.7183

Nota, K., Klaminder, J., Milesi, P., Bindler, R., Nobile, A., van Steijn, T., Bertilsson, S., Svensson, B., Hirota, S. K., Matsuo, A., Gunnarsson, U., Seppä, H., Väliranta, M. M., Wohlfarth, B., Suyama, Y., & Parducci, L. (2022). Norway spruce postglacial recolonization of Fennoscandia. *Nature Communications*, *13*(1), 1333. https://doi.org/10.1038/s41467-022-28976-4

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., & Warinner, C. (2021). Ancient DNA analysis. *Nature Reviews Methods Primers*, *1*(1), 14. https://doi.org/10.1038/s43586-020-00011-0

Palkopoulou, E., Mallick, S., Skoglund, P., Enk, J., Rohland, N., Li, H., Omrak, A., Vartanyan, S., Poinar, H., Götherström, A., Reich, D., & Dalén, L. (2015). Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Current Biology*, *25*(10), 1395–1400. https://doi.org/10.1016/j.cub.2015.04.007

Parducci, L., Alsos, I. G., Unneberg, P., Pedersen, M. W., Han, L., Lammers, Y., Salonen, J. S., Väliranta, M. M., Slotte, T., & Wohlfarth, B. (2019). Shotgun environmental DNA, pollen, and macrofossil analysis of lateglacial lake sediments from southern Sweden. *Frontiers in Ecology and Evolution*, *7*,189. https://doi.org/10.3389/fevo.2019.00189

Parducci, L., Bennett, K. D., Ficetola, G. F., Alsos, I. G., Suyama, Y., Wood, J. R., & Pedersen, M. W. (2017). Ancient plant DNA in lake sediments. *The New Phytologist*, *214*(3), 924–942. https://doi.org/10.1111/nph.14470

Pedersen, M. W., De Sanctis, B., Saremi, N. F., Sikora, M., Puckett, E. E., Gu, Z., Moon, K. L., Kapp, J. D., Vinner, L., Vardanyan, Z., Ardelean, C. F., Arroyo-Cabrales, J., Cahill, J. A., Heintzman, P. D., Zazula, G., MacPhee, R. D. E., Shapiro, B., Durbin, R., & Willerslev, E. (2021). Environmental genomics of Late Pleistocene black bears and giant short-faced bears. *Current Biology*, *31*(12), 2728–2736.e8. https://doi.org/10.1016/j.cub.2021.04.027

Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., Spens, J., Thomsen, P. F., Bohmann, K., Cappellini, E., Schnell, I. B., Wales, N. A., Carøe, C., Campos, P. F., Schmidt, A. M., Gilbert, M. T., Hansen, A. J., Orlando, L., & Willerslev, E. (2015). Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1660), 20130383. https://doi.org/10.1098/rstb.2013.0383

Posth, C., Yu, H., Ghalichi, A., Rougier, H., Crevecoeur, I., Huang, Y., … Krause, J. (2023). Palaeogenomics of Upper Palaeolithic to Neolithic European hunter-gatherers. *Nature*, *615*(7950), 117–126. https://doi.org/10.1038/s41586-023-05726-0

Reimer, P. J., Bard, E., Bayliss, A., Beck, J. W., Blackwell, P. G., Ramsey, C. B., Buck, C. E., Cheng, H., Edwards, R. L., Friedrich, M., Grootes, P. M., Guilderson, T. P., Haflidason, H., Hajdas, I., Hatté, C., Heaton, T. J., Hoffmann, D. L., Hogg, A. G., Hughen, K. A., … van der Plicht, J. (2013). IntCal13 and Marine13 Radiocarbon Age Calibration Curves 0–50,000 Years cal BP. *Radiocarbon*, 55(4), 1869–1887. https://doi.org/10.2458/azu_js_rc.55.16947

Reimer, P. J., Austin, W. E. N., Bard, E., Bayliss, A., Blackwell, P. G., Bronk Ramsey, C., Butzin, M., Cheng, H., Edwards, R. L., Friedrich, M., Grootes, P. M., Guilderson, T. P., Hajdas, I., Heaton, T. J., Hogg, A. G., Hughen, K. A., Kromer, B., Manning, S. W., Muscheler, R., … Talamo, S. (2020). The IntCal20 Northern Hemisphere Radiocarbon Age Calibration Curve (0–55 cal kBP). *Radiocarbon*, 62(4), 725–757. https://doi.org/10.1017/rdc.2020.41

Rijal, D. P., Heintzman, P. D., Lammers, Y., Yoccoz, N. G., Lorberau, K. E., Pitelkova, I., … Alsos, I. G. (2021). Sedimentary ancient DNA shows terrestrial plant richness continuously increased over the Holocene in northern Fennoscandia. *Science. Advances*, *7*(31), eabf9557. https://doi.org/10.1126/sciadv.abf9557

Schmid, S., Genevest, R., Gobet, E., Suchan, T., Sperisen, C., Tinner, W., & Alvarez, N. (2017). HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract genomic information from ancient DNA. *Methods in Ecology and Evolution*, *8*(10), 1374–1388. https://doi.org/10.1111/2041-210X.12785

Schuette, P., Ebbert, S., Droghini, A., & Nawrocki, T. (2022). Small mammal diet indicates plant diversity, vegetation structure, and ecological integrity in a remote ecosystem. *Biodiversity and Conservation*, *31*(3), 909–924. https://doi.org/10.1007/s10531-022-02370-4

Schulte, L., Bernhardt, N., Stoof-Leichsenring, K., Zimmermann, H. H., Pestryakova, L. A., Epp, L. S., & Herzschuh, U. (2021). Hybridization capture of larch (*Larix* Mill.) chloroplast genomes from sedimentary ancient DNA reveals past changes of Siberian forest. *Molecular Ecology Resources*, *21*(3), 801–815. https://doi.org/10.1111/1755-0998.13311

Seersholm, F. V., Pedersen, M. W., Søe, M. J., Shokry, H., Mak, S. S. T., Ruter, A., Raghavan, M., Fitzhugh, W., Kjær, K. H., Willerslev, E., Meldgaard, M., Kapel, C. M., & Hansen, A. J. (2016). DNA evidence of bowhead whale exploitation by Greenlandic Paleo-Inuit 4,000 years ago. *Nature Communications*, *7*, 13389. https://doi.org/10.1038/ncomms13389

Skoglund, P., Malmström, H., Raghavan, M., Storå, J., Hall, P., Willerslev, E., Gilbert, M. T. P., Götherström, A., & Jakobsson, M. (2012). Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science*, *336*(6080), 466–469. https://doi.org/10.1126/science.1216304

Skrbinšek, T., Jelenčič, M., Waits, L., Kos, I., & Trontelj, P. (2010). Highly efficient multiplex PCR of noninvasive DNA does not require preamplification. *Molecular Ecology Resources*, *10*(3), 495–501. https://doi.org/10.1111/j.1755-0998.2009.02780.x

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). In P. Taberlet, A. Bonin, L. Zinger, & E. Coissac (Eds.), *Environmental DNA: For biodiversity research and monitoring*. Oxford University Press. https://doi.org/10.1093/oso/9780198767220.001.0001

van der Valk, T., Pečnerová, P., Díez-Del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., Sağlıcan, E., Fidan, F. R., Barnes, I., Liu, S., Somel, M., Heintzman, P. D., Nikolskiy, P., Shapiro, B., Skoglund, P., Hofreiter, M., … Dalén, L. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, *591*(7849), 265–269. https://doi.org/10.1038/s41586-021-03224-9

van Vugt, L., Garcés-Pastor, S., Gobet, E., Brechbühl, S., Knetge, A., Lammers, Y., Stengele, K., Alsos, I. G., Tinner, W., & Schwörer, C. (2022). Pollen, macrofossils and sedaDNA reveal climate and land use impacts on Holocene mountain vegetation of the Lepontine Alps, Italy. *Quaternary Science Reviews*, *296*, 107749. https://doi.org/10.1016/j.quascirev.2022.107749

Voldstad, L. H., Alsos, I. G., Farnsworth, W. R., Heintzman, P. D., Håkansson, L., Kjellman, S. E., Rouillard, A., Schomacker, A., & Eidesen, P. B. (2020). A complete Holocene lake sediment ancient DNA record reveals long-standing high Arctic plant diversity hotspot in northern Svalbard. *Quaternary Science Reviews*, *234*, 106207. https://doi.org/10.1016/j.quascirev.2020.106207

Wagner, S., Seguin-Orlando, A., Leplé, J.-C., Leroy, T., Lalanne, C., Labadie, K., Aury, J.-M., Poirier, S., Wincker, P., Plomion, C., Kremer, A., & Orlando, L. (2023). Tracking population structure and phenology through time using ancient genomes from waterlogged white oak wood. *Molecular Ecology*, 00, 1–17. https://doi.org/10.1111/mec.16859

Wang, Y., Pedersen, M. W., Alsos, I. G., De Sanctis, B., Racimo, F., Prohaska, A., Coissac, E., Owens, H. L., Merkel, M. K. F., Fernandez-Guerra, A., Rouillard, A., Lammers, Y., Alberti, A., Denoeud, F., Money, D., Ruter, A. H., McColl, H., Larsen, N. K., Cherezova, A. A., … Willerslev, E. (2021). Late Quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature*, *600*(7887), 86–92. https://doi.org/10.1038/s41586-021-04016-x

Weber, S., Junk, I., Brink, L., Wörner, M., Künzel, S., Veith, M., Teubner, D., Klein, R., Paulus, M., & Krehenwinkel, H. (2023). Molecular diet analysis in mussels and other metazoan filter feeders and an assessment of their utility as natural eDNA samplers. *Molecular Ecology Resources*, *23*(2), 471–485. https://doi.org/10.1111/1755-0998.13710

Willerslev, E., Davison, J., Moora, M., Zobel, M., Coissac, E., Edwards, M. E., Lorenzen, E. D., Vestergård, M., Gussarova, G., Haile, J., Craine, J., Gielly, L., Boessenkool, S., Epp, L. S., Pearman, P. B., Cheddadi, R., Murray, D., Bråthen, K. A., Yoccoz, N., … Taberlet, P. (2014). Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature*, *506*(7486), 47–51. https://doi.org/10.1038/nature12921

Zavala, E. I., Jacobs, Z., Vernot, B., Shunkov, M. V., Kozlikin, M. B., Derevianko, A. P., Essel, E., de Fillipo, C., Nagel, S., Richter, J., Romagné, F., Schmidt, A., Li, B., O'Gorman, K., Slon, V., Kelso, J., Pääbo, S., Roberts, R. G., & Meyer, M. (2021). Pleistocene sediment DNA reveals hominin and faunal turnovers at Denisova Cave. *Nature*, *595*(7867), 399–403. https://doi.org/10.1038/s41586-021-03675-0

Zhang, D., Xia, H., Chen, F., Li, B., Slon, V., Cheng, T., Yang, R., Jacobs, Z., Dai, Q., Massilani, D., Shen, X., Wang, J., Feng, X., Cao, P., Yang, M. A., Yao, J., Yang, J., Madsen, D. B., Han, Y., … Fu, Q. (2020). Denisovan DNA in Late Pleistocene sediments from Baishiya Karst Cave on the Tibetan Plateau. *Science*, *370*(6516), 584–587. https://doi.org/10.1126/science.abb6320

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lammers, Y., Taberlet, P., Coissac, E., Elliott, L. D., Merkel, M. F., Pitelkova, I., PhyloAlps Consortium, PhyloNorway Consortium., & Alsos, I. G. (2024). Multiplexing PCR allows the identification of within-species genetic diversity in ancient eDNA. *Molecular Ecology Resources*, *24*, e13926. https://doi.org/10.1111/1755-0998.13926

`wholeskim`: Utilizing genome skims for taxonomically annotating ancient DNA metagenomes

Lucas Elliott, Frédéric Boyer, Teo Lemane, PhyloAlps and PhyloNorway consortia, Inger Greve Alsos, Eric Coissac

Abstract

Inferring community composition from shotgun sequencing of environmental DNA is highly dependent on the completeness of reference databases used to assign taxonomic information as well as the pipeline used. While the number of complete, fully assembled reference genomes is increasing rapidly, their taxonomic coverage is generally too sparse to use them to build complete reference databases that span all or most of the target taxa. Low-coverage, whole genome sequencing, or skimming, provides a cost-effective and scalable alternative source of genome-wide information in the interim. Without enough coverage to assemble large contigs of nuclear DNA, much of the utility of a genome skim in the context of taxonomic annotation is found in its short read form. However, previous methods have not been able to fully leverage the data in this format. We demonstrate the utility of *wholeskim*, a pipeline for the indexing of k-mers present in genome skims and subsequent querying of these indices with short DNA reads. *Wholeskim* expands on the functionality of *kmindex*, a software which utilizes Bloom filters to efficiently index and query billions of k-mers. Using a collection of thousands of plant genome skims, *wholeskim* is the only software that is able to index and query the skims in their unassembled form. It is able to correctly annotate 1.16x more simulated reads and 2.48x more true *seda*DNA reads in 0.32x of the time required by *Holi*, another metagenomic pipeline that uses genome skims in their assembled form as its reference database input. We also explore the effects of taxonomic and genomic completeness of the reference database on the accuracy and sensitivity of read assignment. Increasing the genomic coverage of the genome skims used as reference increases the number of correctly annotated reads, but with diminishing returns after ~1x depth of coverage. Increasing taxonomic coverage clearly reduces the number of false negative taxa in the dataset, but we also demonstrate that it does not greatly impact false positive annotations. The open-source *wholeskim* pipeline is available at https://github.com/ArcEcoGen/wholeskim with a docker image of the pipeline.

Introduction

Environmental DNA (eDNA) is a powerful tool for both palaeoecological reconstructions and contemporary biomonitoring (Laura Parducci et al. 2017; Taberlet et al. 2018). Targeted PCR amplification of conserved regions of DNA (metabarcoding) has been the most commonly used approach to identifying taxa in environmental samples (Von Eggers, Monchamp, and Capo 2022; Revéret, Rijal, and Heintzman 2023). As the cost of DNA sequencing has decreased and the access to high-performance computing clusters has increased, non-targeted sequencing of the entire DNA content of a sample (metagenomic shotgun sequencing) has become a viable alternative (L. Parducci, Alsos, and Unneberg 2019; Wang et al. 2021). This approach allows for the study of the complete taxonomic community present in a sample and the retrieval of genome-wide loci (Graham et al. 2016; Slon et al. 2017; Zimmermann et al. 2023). However, currently both the available DNA reference libraries and bioinformatic tools may limit our ability to take full advantage of shotgun sequenced eDNA.

Compared to metabarcoding, taxonomic annotation is significantly more challenging with a shotgun approach, since the sequenced reads are not limited to a few loci. Instead, the reads are composed of sequences distributed throughout the entire genome, which necessitates the use of a reference database that ideally covers the whole genome of any potential organisms of interest. To date, the most comprehensive set of reference sequences usable for this metagenomic analysis is provided by the International Nucleotide Sequence Databases which is the collaborative project of GenBank, DNA Data Bank of Japan (DDBJ), and European Nucleotide Archive (ENA) (http://www.insdc.org). However, these are far from providing complete genomic sequences for all species. There are a number of ongoing initiatives to sequence and assemble the genomes of all known species on Earth (Gilbert, Jansson, and Knight 2014; Lewin et al. 2018), but the timeline of these efforts extends decades into the future (Lewin et al. 2022). Until these projects are fully realized, low-coverage, non-targeted sequencing of all DNA extracted from an organism's tissues (genome skimming) can generate genome-wide information for numerous taxa at low cost and effort (Straub et al. 2011; Coissac et al. 2016). The scalability of this approach is demonstrated by the PhyloNorway and PhyloAlps projects, which have sequenced representatives of the entire vascular flora of Norway/Polar Regions and the European Alps/Carpathians, respectively (Alsos et al. 2020).

In addition to the challenge of assembling a complete reference database composed of the whole genomes for all organisms of interest, metagenomic analysis also requires adapted algorithms to efficiently compare the large number of sequence reads produced for each sample to this terabyte-sized database. The pipelines used in published studies to achieve taxonomic annotation of metagenomes rely on software belonging to two main categories: mapping software, such as *centrifuge* (Kim et al. 2016) or *bowtie2* (Langmead and Salzberg 2012), and diagnostic k-mer based algorithms, such as *kraken2* (Wood, Lu, and Langmead 2019). While these programs can index the complete genbank database, both approaches have difficulty achieving the same task on numerous genome skims. Mapping-based software requires reference sequences to be significantly larger than the queried reads for efficient processing and to produce sensitive alignments, while current k-mer approaches are unable to index the high k-mer complexity of terabytes of short reads produced by genome skims.

Recent studies have demonstrated the feasibility of incorporating the genome skims produced by the PhyloNorway project into a reference database used to analyze shotgun-sequenced ancient eDNA (Wang et al. 2021; Kjær et al. 2022). These studies indicated that the addition dramatically improved the proportion of metagenome reads that could be taxonomically annotated to genus level (Wang et al. 2021). These studies relied on the bowtie2, mapping-based *holi* pipeline (Pedersen et al. 2016) as well as the assembly of PhyloNorway's genome skims using BBMap's tadpole (Bushnell 2014) and MEGAHIT (Li et al. 2015) to produce contigs with an average length of 496 bp (Wang et al. 2021). However, this preprocessing is computationally expensive and results in discarding information from low-coverage regions of the genome. With PhyloNorway's genome skims averaging 0.5 - 1.0x depth of coverage (Alsos et al. 2020), the raw genome skims are losing a large amount of information during assembly which we hypothesize is useful for taxonomic annotation.

Here we present, *wholeskim*, a pipeline for annotating ancient DNA metagenomes consisting of very short reads by exploiting all the information contained in a set of unassembled genome skims. This is accomplished through two main steps: first, by building a reference database of these genome skims, each of which is associated with the set of k-mers it contains; second, by annotating each read of the metagenome by comparing its k-mer composition with the k-mer reference database built in the previous step. Expanding on the functionality of *kmindex* (Lemane et al. 2023), wholeskim optimizes storage and accuracy for a large number of input genome skims and performs least common ancestor assignment on each identified read. *Kmindex* leverages the Bloom filter data structure to efficiently index large metagenomic datasets and accurately queries them using the findere algorithm to significantly reduce false positive identifications (Robidou and Peterlongo 2021; Lemane et al. 2023).

In this paper, we demonstrate the functionality and benchmark the performance of *wholeskim* using 1541 genome skims from PhyloNorway and compare these measurements to the *holi* pipeline (Wang et al. 2021). We also examine the impact of reference database completeness on taxonomic assignment by considering two aspects: overall taxonomic coverage of the reference database and genomic completeness on a per species level. Finally, we demonstrate *wholeskim*'s ability to taxonomically annotate three sedimentary ancient DNA metagenomic datasets.

Methods

wholeskim implementation

wholeskim consists of two bash scripts: *prep_indices*, which cleans the genome skims, groups them by k-mer complexity, and finally indexes them and *query_indices* which queries the indices with a DNA metagenome and processes the resulting assignments. The input for *prep_indices* is a collection of genome skim data files in FASTA/FASTQ format. *query_indices* requires the previously built indices and a shotgun-sequenced eDNA sample in FASTA/FASTQ format. The final output is a table with the following information for every read: the maximum proportion of shared kmers with a single database entry, the number of database entries considered for least common ancestor (LCA) assignment, and the taxonomic ID of the LCA. All code and docker image is available at https://github.com/ArcEcoGen/wholeskim.

Building of the k-mer reference database

Genome skims produced by the PhyloNorway project were used for reference database construction (ENA project number: PRJEB43865 (Alsos et al. 2020; Wang et al. 2021)). Prior to indexing, the 1 541 genome skims were concatenated on a species-level to produce 1 323 separate entries for indexing. The programs *centrifuge*, *bwa*, and *kraken2* were all unable to index these 1.9 TB of unassembled genome skims due to requesting > 2 TB of memory during the course of construction. We were able to build the reference database using *kmindex* (Lemane et al. 2023), a software utilizing Bloom filters to space-efficiently index the k-mers of large metagenomic datasets. A Bloom filter is a probabilistic data structure used to store numerous objects, here k-mers, and to test whether an object is a member of that set. The reason for its effectiveness is that a Bloom filter is not based on an exact algorithm, but on heuristics. As a consequence of this characteristic, Bloom filters can produce false positives, answering that the k-mer is a member of the set when it is not, but they cannot produce false negatives. To reduce the number of false postives, *kmindex* relies on the corrective *findere* algorithm (Robidou and Peterlongo 2021; Lemane et al. 2023) which queries *z* neighboring *k*-mers of size k - z + 1, instead of a k-mer of size k. It only reports a positive match if all *z* *k*-mers are indexed in the reference database. Thus, if the Bloom filter has a false positive rate *fp*, the resulting false positive rate after using the *findere* algorithm is fp^z.

wholeskim builds the k-mer reference database in a two-steps process. First, it filters out from the genome skims all the sequence reads that match to commonly suspected DNA contaminants in a genome skim (bacteria, fungi, and human DNA) and are not found in a representative set of other related reference genomes. Second, it builds a Bloom filter for each genome skims with the cleaned reads.

Genome skim cleaning: Prior to indexing, *wholeskim* filters out from the genome skims all reads that match to a commonly suspected DNA contaminant in a plant tissue extract. The following large taxonomic groups are indexed using *kmindex*: Bacteria (NCBI taxonomy ID: 2), fungi (4751), and Homo sapiens (9606). A last bloom filter corresponding to our group of interest Viridiplantae (33090) is also built. Using the taxonomic annotation procedure described below, only reads unidentified or annotated as Viridiplantae are conserved. Here, these cleaning bloom filters are built from genbank's 259 release.

k-mer reference database indexing: Bloom filters encode k-mers as a bit field (an ordered set of 0s or 1s). The false positive rate of a Bloom filter depends on the number of bits (m) used to encode the k-mers and the number of different k-mers stored in the filter. To obtain a given false positive (fp) rate when storing n k-mers, m is defined according to formula 1.

$$m \quad = \quad -\frac{n}{ln\,(1-fp)} \qquad\qquad (1)$$

We aim for a *fp=0.05* for each bloom filter, corresponding to an overall false positive rate at query-time of $0.05^3 = 1.25 \times 10^{-4}$. *wholeskim* works with k-mer of size k=34 as a tradeoff between specificity, efficiency and the ability to annotate the majority of reads present in ancient DNA metagenomes (Supplementary Figure 1). To optimize its computation and storage footprint, kmindex allows for simultaneous indexing of multiple bloom filters with the same bit size (m). To minimize storage and maintain a consistent *fp* rate, wholeskim groups genome skims for indexing based on their k-mer diversity, the count of distinct k-mers present in each skim estimated by *ntcard (Mohamadi, Khan, and Birol 2017)*.

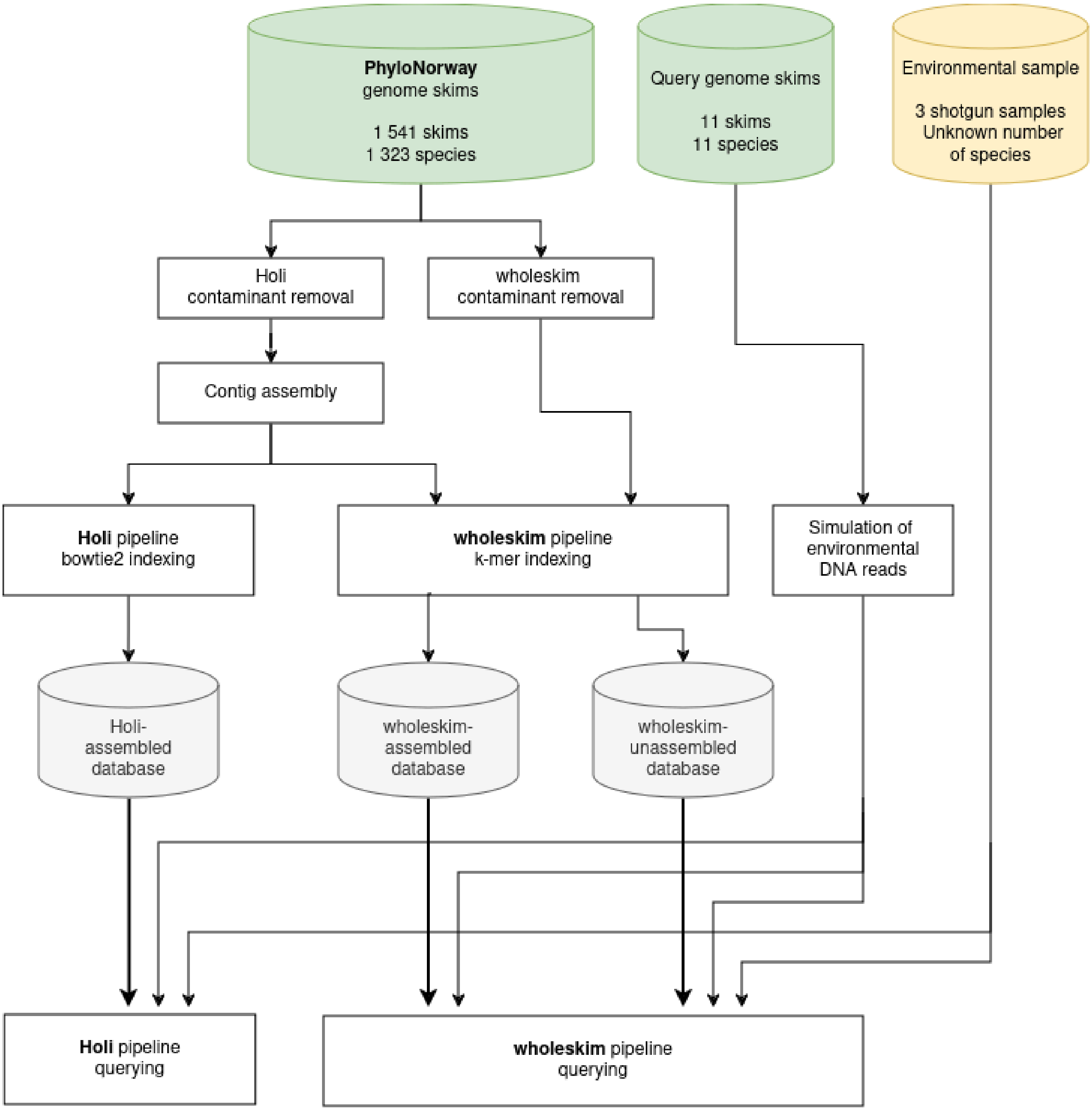
Assigning a taxon to a read from a metagenome

The decision to assign a taxon to a query read, Q, is based on the number of shared k-mers, S_G, it has with each of the genome skims, G. The assignment algorithm is:

- Calculate $N_Q = L_Q - k + 1$, the total number of k-mers present in Q, a read of length L_Q.
- Identify S_{max}, the maximum number of k-mers shared between Q and any of the indexed genome skims G.
- Calculate $t_{max} = S_{max}/N_Q$.
- If $t_{max} \geq t_c$, the cutoff proportion for a positive match, define $t_{min} = t_{max} - \Delta$, a threshold for similar matches.
- Calculate $S_{min} = t_{min} * N_Q$.
- Select all genome skims G with $S_G \in [S_{min}; S_{max}]$.
- Assign to Q the lowest common ancestor (LCA) of all taxa associated with the selected genome skims, using the NCBI taxonomy as a reference.

To reduce the noise of assignments, only assignments to taxa that appeared in greater than a proportion, r, of the total reads were retained. After optimization through testing with simulated datasets, the three parameters of the procedure have been set to t = 0.7, Δ = 0.1, r = 10^{-5} (Supplementary Figure 2).

Evaluating wholeskim performances



PhyloNorway genome skims
1 541 skims
1 323 species
Query genome skims
11 skims
11 species
Environmental sample
3 shotgun samples
Unknown number of species
Holi contaminant removal
wholeskim contaminant removal
Contig assembly
Holi pipeline bowtie2 indexing
wholeskim pipeline k-mer indexing
Simulation of environmental DNA reads
Holi-assembled database
wholeskim-assembled database
wholeskim-unassembled database
Holi pipeline querying
wholeskim pipeline querying

Figure 1. Diagram of the three different workflows that were evaluated. From left to right: the Holi pipeline tested with the assembled contigs, wholeskim tested with the assembled contigs, and finally wholeskim tested with the unassembled reads.

Wholeskim was evaluated on two aspects. 1) Its efficiency in indexing and querying the subset of 1541 PhyloNorway genome skims released in (Wang et al. 2021). This efficiency was measured in terms of indexing speed, querying speed, and memory requirements. 2) The sensitivity and specificity of wholeskim in annotating sequences were compared with those of the Holi pipeline, (Wang et al. 2021; Kjær et al. 2022). These estimates of the sensitivity and specificity of wholeskim were made considering the impact on the completeness of the reference database in terms of taxonomic coverage and sequencing depth for a given species. Computations were performed on an HPC cluster node with 2 x 16-core Intel Xeon Gold 6130 processor 2.10 GHz with 192 GB of memory. Databases were stored on an SSD to facilitate faster read access since *kmindex* does not load the full database into memory.

Sensitivity and specificity tests: Two of the skims used for querying the pipeline, *Thesium alpinum* (PHA009155) and *Salix retusa* (PHA007876), are taken from PhyloAlps while the other nine (*Avenella flexuosa*, *Betula nana*, *Betula pubescens*, *Bistorta vivipara*, *Caltha palustris*, *Dryas octopetala*, *Picea abies*, *Pinus sylvestris*, *Vaccinium uliginosum*) (XXX) were extracted/sequenced in Tromsø following the PhyloNorway protocol (Alsos et al. 2020). None of these skims were included in the assembled or unassembled reference databases. Sets of simulated reads were obtained from a genome skim using the *adrsm* software (Borry 2018). Read simulation consisted of reducing the actual length of the sequencing reads of the genome skims to mimic the size distribution observed in ancient DNA metagenomes (mean insert size set to 35 based on median age samples from (Pedersen et al. 2016)).

Comparison of sensitivity and specificity between the *Holi* and *wholeskim* pipelines: The annotations from both pipelines were sorted into eight categories. Four categories describe different levels of precision in an accurate assignment: i) "target species" when the read is annotated as the correct species, ii) "target genus" when the read is annotated at the genus level only, but the correct one, iii) "target family" when the read is annotated at the family level only, but the correct one, and iv) "higher target level" when the read is annotated at a taxonomic level higher than family, but the correct one. Four categories correspond to different levels of misclassification: i) "incorrect species / target genus", when the read is annotated to a different species in the target genus, ii) "incorrect genus / target family" when the read is annotated to a different species or genus in the target family, iii) "incorrect family", when the read is annotated to a *Viridiplantae* clade outside the previous two categories, and iv) “unidentified” when a read is not assigned.

Three workflows were used to taxonomically assign reads: i) the *Holi* pipeline using the assembled contigs from 1 541 PhyloNorway genome skims downloaded from Dataverse.no (doi.org/10.18710/3CVQAG), ii) the *wholeskim* pipeline using the unassembled 1 541 genome skims, and iii) the *wholeskim* pipeline using the assembled contigs from workflow (i).

Workflow (i) was executed by following the scripts present in https://github.com/miwipe/KapCopenhagen (Kjær et al. 2022) and using only the 1 541 PhyloNorway assembled genome skims as a reference database. An overview of this workflow is presented in Supplementary Figure 3. To harmonize comparison with the *wholeskim* pipeline, reads assigned by *Holi* to taxa comprising less than a cutoff proportion ($r = 10^{-5}$) of the entire dataset were set to unidentified, a generally more conservative threshold than the cutoff proportion of ($r = 10^{-5}$) for only annotated reads as recommended by (Pedersen et al. 2016).

We break down the assignments of four species in detail that represent varying levels of representation in the reference database. They were selected as follows: *Betula nana* was selected as a well-represented taxon because six skims of the *Betula* genus, including four from *Betula nana*, are present in the reference database; *Avenella flexuosa* was selected because, it is the only member of the *Avenella* genus present in the reference database; *Salix retusa* was chosen as the species is not represented in the reference database, but 39 other *Salix* skims representing 30 species are present; and finally for *Thesium alpinum*, no members of the Santalaceae family were present in the reference database. Workflows (i) and (ii) are further contrasted by summing the eight categories of taxonomic assignments across all nine test species that are represented in PhyloNorway and constructing a confusion matrix. This allows us to assess the overall performance of the two pipelines and investigate how reads classified by one workflow are assigned by the other.

Impact of the genome coverage on assignment quality: Seven genome skims of *V. uliginosum* totalling 68 M reads are included in the 1 541 skims of PhyloNorway. With an estimated genome size of 600 MB (Sultana et al. 2020), this represents an expected maximum genome depth of coverage of 11x. To assess the impact of genome coverage and database completeness on *wholeskim*'s accuracy, the simulated *V. uliginosum* eDNA reads were assigned using the following subsets of the reference database: a database excluding all Ericaceae genome skims, a database excluding all *Vaccinium* genome skims, and databases including 0, 1, 2, 3, …, 19, 20, 24, 28, …, 68 M reads of *V. uliginosum*.

Comparative efficiency of wholeskim and holi on true ancient metagenomes

Three shotgun sequenced *sed*aDNA samples from northern Norwegian archaeological midden complexes were annotated with both the *Holi*-assembled and *wholeskim*-unassembled workflows. The samples (GB-6, GB-5, and IG-4) were dated to 4.2, 4.0, and 0.7 ka respectively and extracted using the DNeasy PowerLyzer PowerSoil kit (Qiagen: 12855-100; Komatsu et al. 2024). Libraries were prepared for paired-end sequencing using single-stranded library preparation designed specifically for highly degraded ancient DNA (Komatsu et al. 2024). The paired-end sequenced reads were merged, adapter sequences were removed, and reads less than 34 bp were discarded using *fastp* v0.23.4 (Chen et al. 2018). The resulting 13.9 M, 16.7 M, and 17.5 M reads respectively were annotated using the *Holi*-assembled and *wholeskim*-unassembled workflows. Taxa with less reads than 0.001% of the total were discarded from both workflow's final assignments. Assignments at species-level were collapsed to genus-level as this is a more reliable level of identification for shotgun data (see results), while assignments to taxa above family-level were discarded.

Results

Genome skim cleaning

The genome skims produced by PhyloNorway had a mean value of 4.64 million read pairs (sd = 1.58 million, Alsos et al, 2020). The cleaning step of wholeskim reduced the size of these raw genome skims by a median value of 0.015%. Among the rejected reads, median percentages of 0.007%, 0.004%, 0.002%, and 0.0005% were identified as algae, bacteria, fungi, and human contamination (Supplementary Figure 4). However, a few genome skims had considerably more contamination detected with maximum values for the previous categories of 0.08%, 1.7%, 2.2%, and 0.18%. Among the selected reads, a median value of 3.7% were identified by the filter as Viridiplantae reads, while the remaining ~96% matched none of the tested categories and considered as not yet sequenced part of the plant genome.

Genome skim information content

The information content of skims and contigs was measured by grouping unassembled genome skims or contigs by species and then counting the number of distinct k-mers (k = 31) present in each set. The ratio of the number of k-mers in skims to the number of k-mers in the corresponding contigs for a species varies from 1 to 100 with a mode around 10 (Figure 2A). While a portion of the distinct k-mers present in the unassembled genome skims are due to sequencing error, there is still a drastic reduction in information content caused by the skim assembly step required by mapping-based approaches such as the *Holi* pipeline. Unassembled genome skims with a higher depth of coverage produce increasing numbers of distinct k-mers as expected, but this number does not begin to plateau after reaching > 1x coverage (Figure 2B). The observed number of unique k-mers found in the *V. uliginosum* skims matches the expected number of unique k-mers from given a sequencing error rate of 4.4x10^{-3} (Figure 2B, equation given in Supplementary).

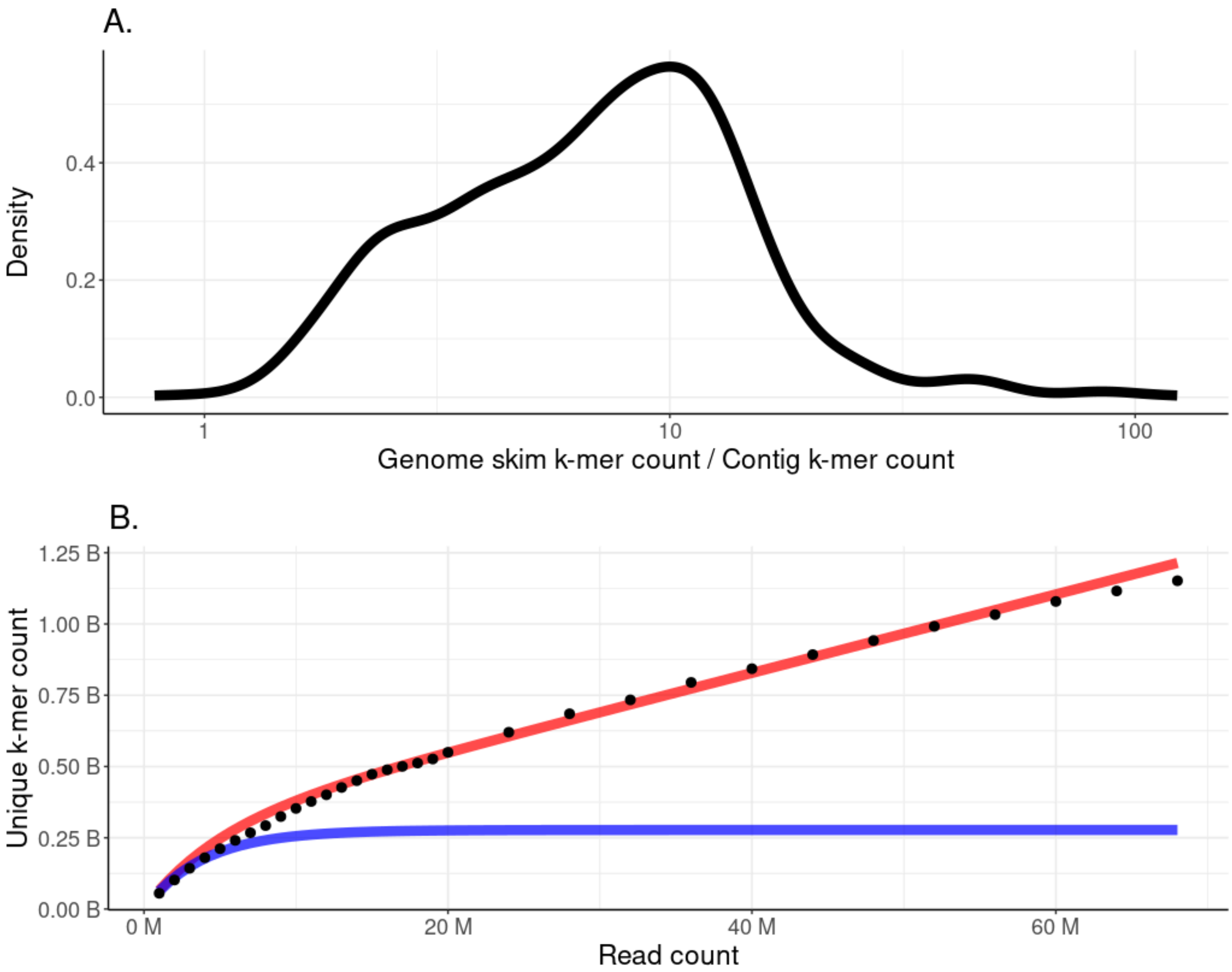


Figure 2: A) Distribution of the ratio between the number of distinct k-mer (k=31) associated with a species in the genome skims dataset and the equivalent number in the corresponding contig dataset. B) The points represent the number of unique k-mers (k=31) for a given number of reads for *Vaccinium uliginosum* from the PhyloNorway genome skims. The blue line represents the expected number of unique k-mers found in an unassembled genome skim for a genome of *V. uliginosum*'s size (600MB) and read size of 101 bp. The red line shows the same relationship, but with an introduced sequencing error rate of 4.4 x 10^{-3}.

Performance of *wholeskim* compared to *Holi*

Indexing efficiency:

The indexing of the 1 541 genome skims or 1.9 TB of sequences by *wholeskim* required 11.8 hours of computation time with a maximum of 54.0 GB of random access memory (RAM) used. To this, 7.6 hours per billion reads of the reference genome skims for the cleaning step have to be added. The *Holi* pipeline required 36.4 hours of computation time to index (through bowtie2-build) the 152 GB of contigs requesting a maximum of 240.0 GB of RAM. An unknown amount of time was used to assemble and preprocess the contigs used as input. The resulting index created by *wholeskim* required 472 GB of storage, larger than the 311 GB occupied by *Holi*'s index.

Querying efficiency:

A total of 2,187,986 simulated environmental DNA sequences from eleven species were taxonomically annotated using both the *wholeskim* and *Holi* pipelines. The querying of these reads by *wholeskim* with the unassembled genome skims required 3.1 hours and a maximum of 4.8 GB of RAM. By comparison, *Holi* required 3.7 hours to index the same set of reads while loading significant portions of the reference database in memory and requesting a maximum of 115 GB of RAM.

Overall sensitivity and specificity of workflows

The *wholeskim*-assembled workflow performed the worst, misassigning more reads and correctly identifying less reads for every taxon than both other workflows (Supplementary Figure 4). Due to this poor performance, we exclude this workflow from future comparisons. The *wholeskim*-unassembled and *Holi*-assembled workflows both correctly annotated a plurality of reads for each species with few misassignments. Averaged over the simulated reads from the nine species that are present in PhyloNorway, *wholeskim*-unassembled correctly identified 20.2% of the reads at a species or genus-level while *Holi*-assembled correctly identified 15.3% of the reads at these levels. At the species level, *wholeskim*-unassembled assigns nearly 1.34x more reads than *Holi*-assembled. At the genus-level, *wholeskim*-unassembled assigns 1.16x more reads. This difference is even larger for correct assignments at the taxonomic level of family or above, but these identifications are rarely useful for metagenomic annotation purposes. *Holi*-assembled and *wholeskim*-unassembled incorrectly assign 3.0% and 3.4% of reads respectively with >90% of these misassignments to a congeneric species of the target. There is little overlap in the sets of reads that each workflow annotates with only 21.6% of the total reads annotated to target species shared between workflows (Figure 3).

Taxonomic completeness of reference database

Both the *Holi*-assembled and *wholeskim*-unassembled workflows accurately annotated reads from species with varying levels of representation in the reference database. *Betula nana*, a well-covered species and genus, had 28.1% and 20.6% of reads assigned correctly to genus or species-level by *wholeskim*-unassembled and *Holi*-assembled (Figure 4A). Only 2.4% and 3.0% of reads were incorrectly assigned by each workflow respectively, and all of them were to another species in the *Betula* genus. *Avenella flexuosa* is the only member of the *Avenella* genus present in the database and consequently the only misassignment is 1.1% and 0.8% of reads to taxa in the same family (Figure 4B). Both workflows correctly assigned 16.1% of this species' simulated reads to the species-level. *Salix retusa* had no skims present in the reference dataset, so no reads were assigned correctly to species-level, however 24.2% and 19.7% were assigned to the *Salix* genus. The workflows assigned 3.2% and 5.2% to other *Salix* species and <0.1% to taxa outside the Salicaceae family. Without any representatives of the family Santalaceae in the reference database, 95% of *Thesium alpinum* reads were left unidentified while <0.2% were incorrectly annotated by both workflows.

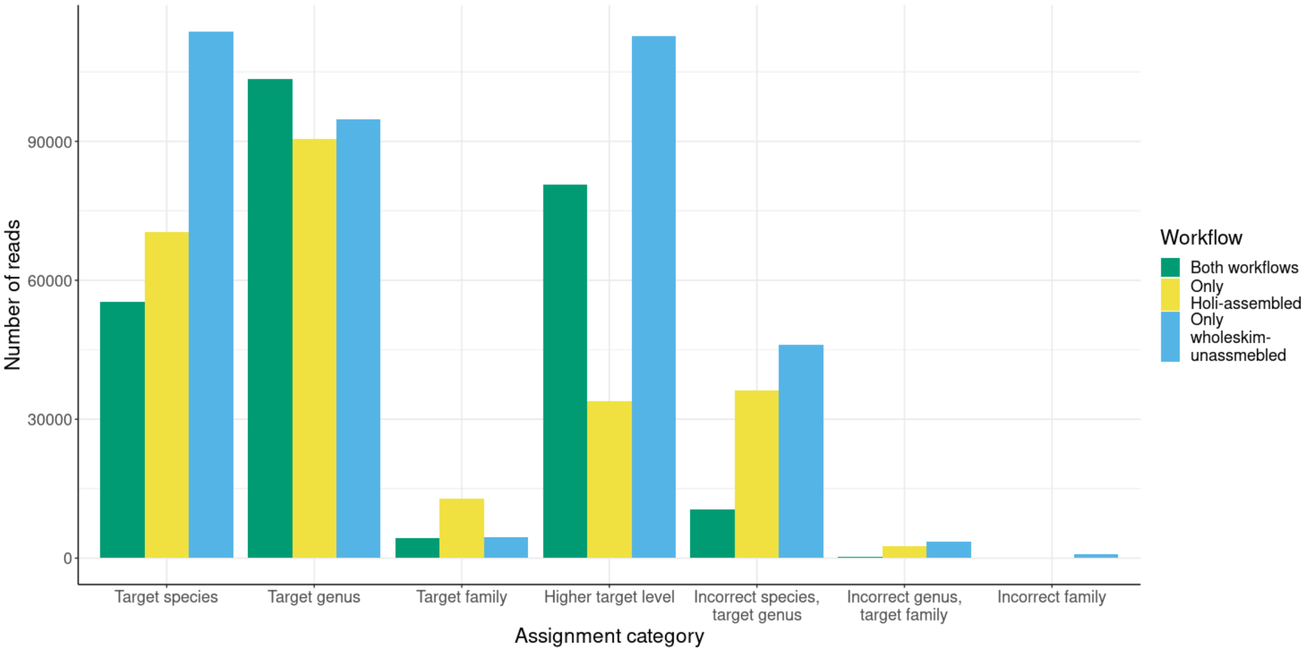


Figure 3. The overlap of simulated reads identified by the *Holi*-assembled and *wholeskim*-unassembled workflows. The set of reads is composed of all nine test species present in the PhyloNorway reference database.

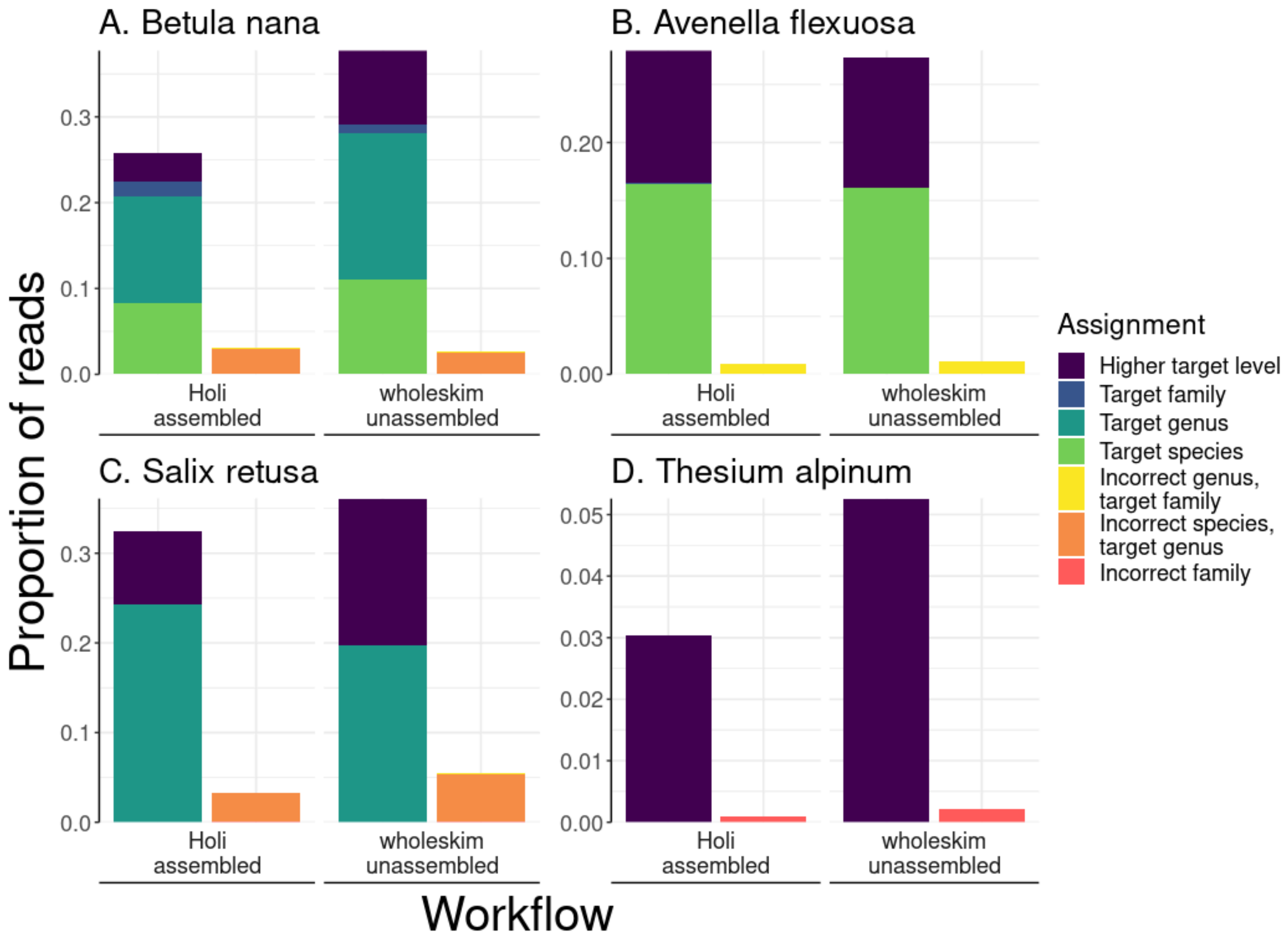


Figure 4: The effect of taxonomic completeness of the reference database on assignment for each workflow. Each workflow's assignments are divided into two stacked bars representing correct and incorrect assignments which are further broken down by color according to the legend. The represented species are A) *Betula nana* which has 4 genome skims in the database along with 2 more skims from the *Betula* genus, B) *Avenella flexuosa* which is present with one species in the database, and it is the only species within this genus in the database, C) *Salix retusa* which is not present in the reference database, but skims of 38 other *Salix* species are included, and D) *Thesium alpinum* which is not present in the database, and no members of the Santalaceae family are in the reference database. Note the differences in scale of the y-axis.

Genomic completeness of reference database

The effect of genome skim sequencing effort on taxonomic annotation accuracy was tested with the *wholeskim*-unassembled workflow by constructing the reference database using increasing numbers of *V. uliginosum* reads. As the number of genome skim reads, and consequently the number of unique k-mers, increased in the reference database, the proportion of correctly annotated simulated *V. uliginosum* reads also increased (Figure 5). This trend is linear until 600 million unique k-mers, roughly corresponding to genome size of *V. uliginosum* (Sultana et al. 2020), where it continues to increase, but with a smaller slope. The proportion of unidentified and misassigned reads follows a similar, but inverted trend as the number of unique k-mers in the reference database increases.

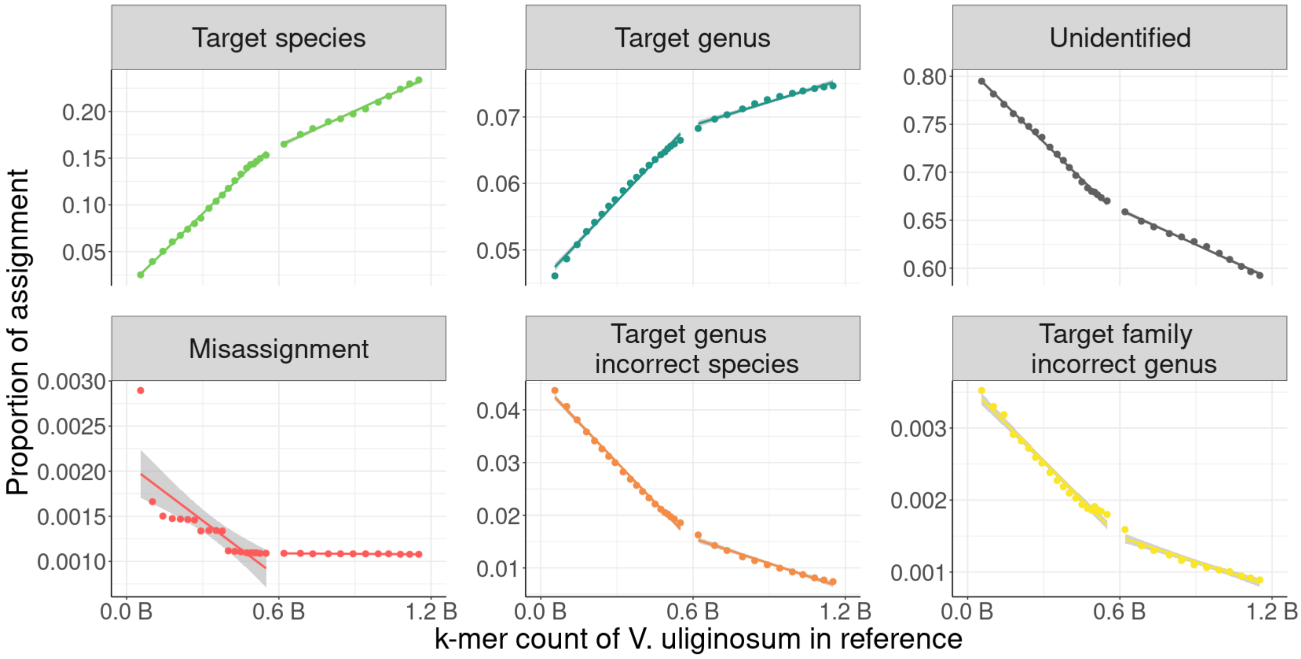


Figure 5. The taxonomic assignment of 200k simulated *Vaccinium uliginosum* reads as increasing amounts of unique k-mers of *V. uliginosum* are added to the reference database. Note the varying y-axis scales. Lines of best fit are added to points before and after estimated 1x coverage of the *V. uliginosum* genome (~0.6 B k-mers).

Performance of analyzing true ancient DNA metagenomes

Between 0.8 - 5.6% of the reads from the three sedaDNA datasets were annotated by either workflow (Figure 6). On average, *wholeskim*-unassembled annotated 2.48x more reads than *Holi*-assembled for all three datasets. Each taxon present in the final annotations had more reads identified by *wholeskim*-unassembled than by *Holi*-assembled. The list of identified taxa is largely consistent between the two workflows with all taxa identified by *Holi*-assembled also being identified by *wholeskim*-unassembled. However, there are seven instances of a taxon being identified as present by *wholeskim*-unassembled, while having too few reads (r < 0.01% of query reads) to be retained by *Holi*-assembled in the final annotations.

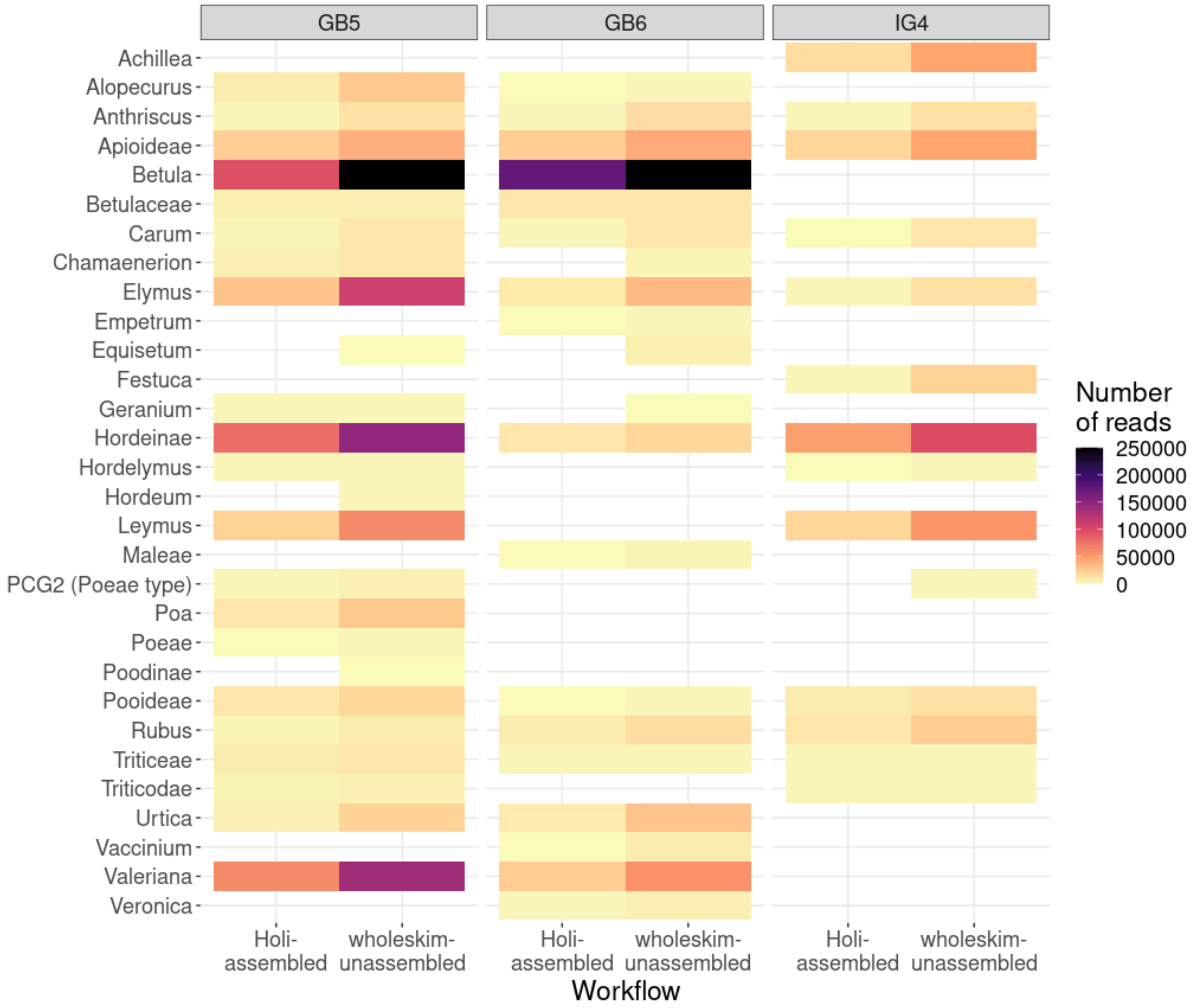


Figure 6. The number of reads annotated by the *Holi*-assembled and *wholeskim*-unassembled workflows of three shotgun-sequenced *seda*DNA samples.

Discussion

The *wholeskim*-unassembled workflow is currently the only approach to accurately annotating DNA metagenomes that allows for the effective indexing and querying of large-scale unassembled genome skim datasets. This workflow correctly annotated 1.16x more simulated reads and generally annotated 2.48x more *seda*DNA reads than the existing metagenomic DNA annotation workflow used on large-scale genome skim datasets, *Holi* (Pedersen et al. 2016; Wang et al. 2021). The *wholeskim*-unassembled workflow is however slightly more prone to erroneous annotations with 1.58x more assignments to taxa outside the genus of interest, however these misassignments only totalled 0.27% of the total simulated read dataset and are distributed between many taxa. This misassignment rate is close to the reported rate of 0.24% from the simulated data in *Holi*'s original publication (Pedersen et al. 2016).

The *Holi*-assembled and *wholeskim*-unassembled workflows operate differently and only jointly identify 27.5% of the total correctly assigned simulated reads. The discrepancy in assignment performance and subset of reads annotated between the workflows could be attributed to two different factors: the method of reference database construction or the matching algorithm itself.

When *wholeskim* was run using the assembled contigs database used with the *Holi* pipeline, it misassigned considerably more reads (Supplementary Figure 5), as expected as each pipeline is tailored to its corresponding reference database composition.

There was a mode 10x loss of information when assembling the genome skims into contigs (Figure 2A). With the Illumina HiSeq 2500 having a sequencing error rate of 1.12e^{-3}, σ = 5.44e^{-3} (Stoler and Nekrutenko 2021), some of these discarded k-mers are a product of sequencing error, however many are also likely from low-coverage regions that were unable to be assembled. Query reads spanning these low coverage areas would not be able to be identified by the *Holi*-assembled workflow, but could be annotated by *wholeskim*-unassembled. Conversely, query reads with a single sequencing error near the center of the read would have a very low number of shared k-mers with the reference index in the *wholeskim*-unassembled workflow. These reads would however be able to be identified by bowtie2's mapping-based approach in *Holi*-assembled.

As a genome skim's depth of coverage increases past 1x, the number of unique k-mers continues to increase, albeit at a smaller slope (Figure 2B). These are likely "erroneous" k-mers being included in the reference dataset, but their inclusion does allow for more accurate assignment of reads (Figure 5). The mechanism behind this improvement could be that the spurious k-mers allow for some "fuzzy" matching to compensate for sequencing errors in the query reads as well as individual genomic variation. Other k-mer based metagenomic tools that have incorporated "fuzzy"-matching reported higher accuracy assignments when compared to exact-matching k-mer tools (Firtina et al. 2023). It does not appear that incorporating these spurious k-mers in the reference database has significantly increased incorrect assignment to these taxa since *V. uliginosum* has the highest genome coverage and unique k-mer count, but is not detected as one of the false positives by *wholeskim*-unassembled in the simulated read datasets. The slightly higher misassignment rate of *wholeskim*-unassembled could be attributed to the hash collisions inherent in the probabilistic Bloom filter data structure (Bloom 1970). The false positives produced by this mechanism are expected to be distributed among all taxa in the database at a low frequency so they are largely filtered out by employing a cutoff of the minimum proportion of reads assigned to a taxon if that taxon is to be retained in the final dataset ($r = 10^{-5}$). Through *in silico* testing with simulated datasets, *Holi* arrived at a similar, but less stringent, threshold ($r = 10^{-5}$ for all assigned reads, rather than queried reads) for discarding false positive taxa attributed to sequencing errors, amplification errors, or DNA damage (Pedersen et al. 2016). With both workflows, misassignments generally cluster around closely-related taxonomic groups with the majority being congeneric species and confamilial taxa (Supplementary Figure 6). These misassignments could be the result of conserved genomic regions that have not been sequenced in one of the low coverage genome skims. An example can be seen in GB5, one of the ancient genomic samples, where a small proportion of the reads were assigned to the cultivated plant *Hordeum* although they probably represent the local native species *Leymus* (Koamtsu et al. 2024). *Hordeum* was assigned a proportion of reads marginally above the cutoff (0.012%), while the family containing *Hordeum*, *Hordeinae*, was assigned a proportion of reads an order of magnitude larger (0.85%). This highlights the shortcomings of using a fixed cutoff (r) for retaining taxa when the DNA content of a metagenome is not expected to be equally distributed among all taxa present in a sample. Instead, the cutoff could be a proportion of the most dominant taxon and take into consideration taxonomic distance.

A major challenge for metagenomic annotation is the taxonomic completeness of reference databases (L. Parducci, Alsos, and Unneberg 2019; Wang et al. 2021). When queried with simulated reads from species absent in the reference database, both workflows produced few misassignments suggesting that while reference database incompleteness produces false negatives, it is not a major contributor to false positives. While some taxa have a majority of reads assigned to species-level in the *wholeskim*-unassembled workflow (e.g. *V. uliginosum*), other taxa like *Betula nana* have a majority assigned to the genus-level with significant portions assigned to congeneric species making a species-level identification unreliable especially in an environmental sample with many taxa present (Supplementary Figure 7). The authors of the *Holi* pipeline have also recognized this limitation and generally identify flora to the genus- or family-level except when combined with species distribution information (Pedersen et al. 2016; Wang et al. 2021). This taxonomic resolution is one of the major limitations of shotgun sequenced metagenome annotation when compared to other methods like metabarcoding (Revéret et al. 2023).

Earlier *seda*DNA metagenomic studies using primarily the NCBI nt or RefSeq databases as references reported very low proportions of reads identified to any taxonomic level of *Viridiplantae,* with (Slon et al. 2017) identifying a mean of 0.07%, (Courtin et al. 2022) identifying 0.05%, and (L. Parducci, Alsos, and Unneberg 2019) identifying only 0.0002% of queried reads. The first application of the *Holi* pipeline using NCBI nt reports 0.05% of total reads assigned to some level of *Viridiplantae* (Pedersen et al. 2016). The addition of the PhyloNorway genome skim contigs to the *Holi* pipeline gave a significant increase to the number of reads annotated to *Viridiplantae*, 1.7% (Wang et al. 2021). This percentage of identified reads is consistent with how many reads were annotated by *Holi*-assembled for the true *seda*DNA datasets in this paper, 0.8-2.3%, while *wholeskim*-unassembled assigned 2.1-5.6% for the same datasets. Other comparisons between metagenomic annotation tools, including *Holi*, show that they largely agree on the taxa present in datasets (Harbert 2018). Here, we report the same as 51/58 taxa observations in the *seda*DNA datasets are shared between workflows (Figure 6). While *wholeskim*-unassembled is not able to identify reads with a mismatch near the center of the fragment, errors near the ends of the fragment do not adversely affect the k-mer similarity score as strongly. Since ancient DNA deaminations typically occur at the ends of fragments (Dabney, Meyer, and Pääbo 2013), *wholeskim*-unassembled still annotated a large proportion of the *seda*DNA reads.

Conclusion

By incorporating information from the entire low-coverage genome skims, the *wholeskim*-unassembled workflow is able to accurately annotate more reads than other metagenomic pipelines. It is clear that increasing the taxonomic coverage of the reference database reduces the number of false negatives, but we also demonstrate that it does not greatly impact the number of false positive annotations in *wholeskim*-unassembled. Similarly, increasing the genomic coverage of the genome skims used as reference increases the number of annotated reads, but with diminishing returns after ~1x depth of coverage. Since *wholeskim*-unassembled and *Holi*-assembled are correctly annotating different sets of reads, their combined use results in the largest number of reads for applications such as metagenomic assembly. However, if the intent of the study is to infer the community composition of the sample, using only the more computationally efficient *wholeskim*-unassembled is sufficient.

Author contributions

L.D.E., F.B., I.G.A., and E.C. conceptualized the study. T.L. designed and coded the software kmindex. L.D.E. wrote the code for wholeskim with feedback from F.B. and E.C. Collection and genome skimming of the nine additional species was performed by L.D.E. Benchmarking of the pipeline was performed by L.D.E. Interpretation of data was done by L.D.E. with input from I.G.A. and E.C. The manuscript was written by. L.D.E. with feedback from all co-authors.

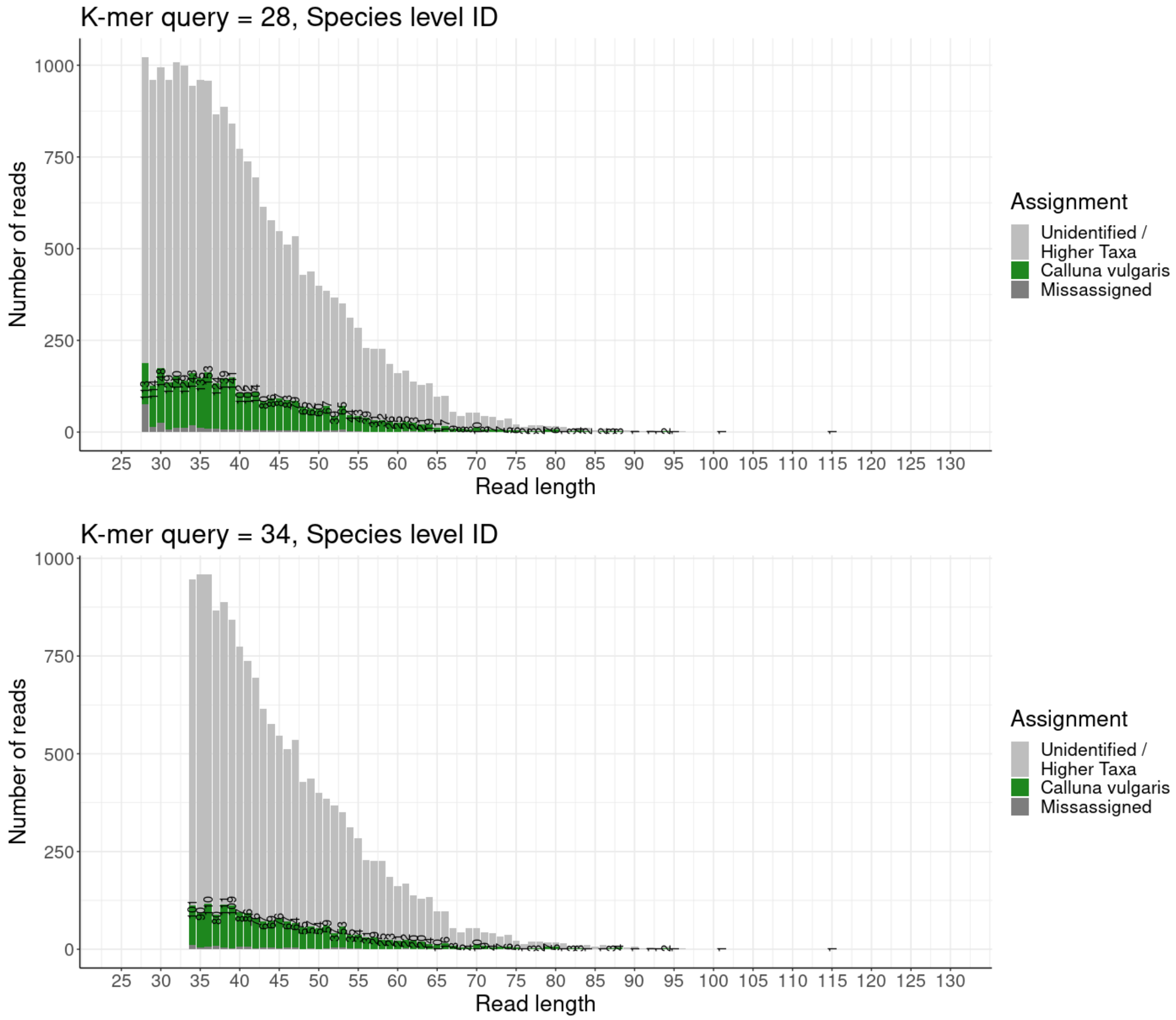
Acknowledgements

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement no. 819192 for the IceAGenT project. PhyloNorway is part of the Norwegian Barcode of Life Network (NorBOL.org) funded by the Research Council of Norway (226134/F50) and the Norwegian Biodiversity Information Centre (14-14, 70184209). All the computations presented in this paper were performed using the GRICAD infrastructure (https://gricad.univ-grenoble-alpes.fr), which is supported by Grenoble research communities. We would like to thank Aya Komatsu for collecting and extracting the *sed*aDNA data, Clement Lionnet for help with parallelizing wholeskim, Janine Klimke and Kathleen Stoof-Leichsenring for collaboration in shotgun sequencing the sedaDNA samples, Lasse Topstad for collaboration in collecting plant samples, and Iva Pitelkova for instruction in plant tissue DNA extractions and genome skim sequencing.

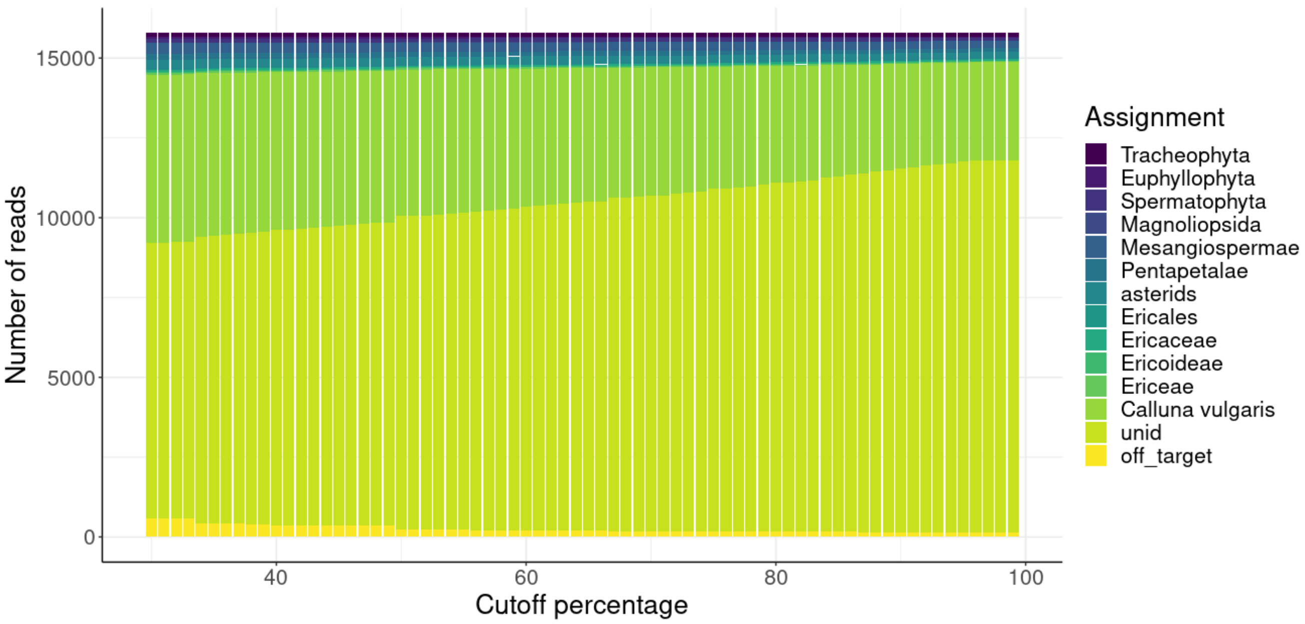
Data availability

Github and ENA accession numbers

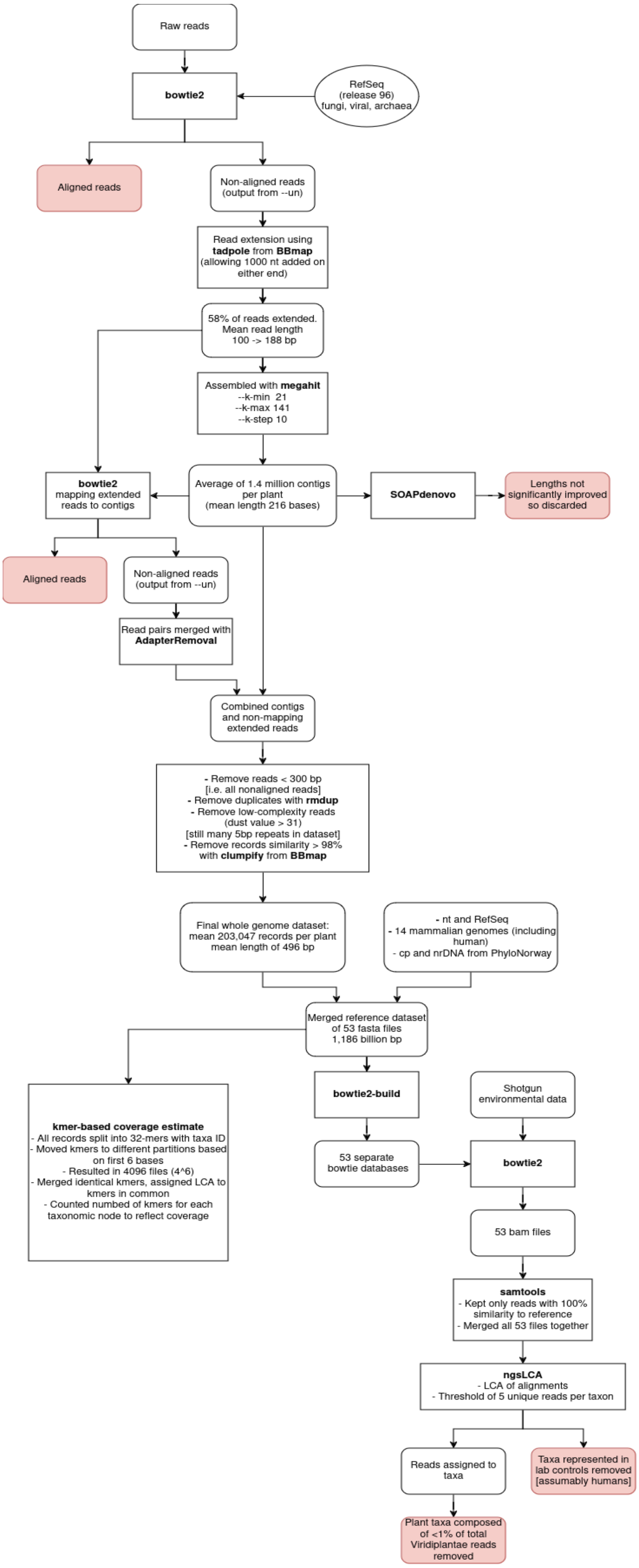
Supplementary information



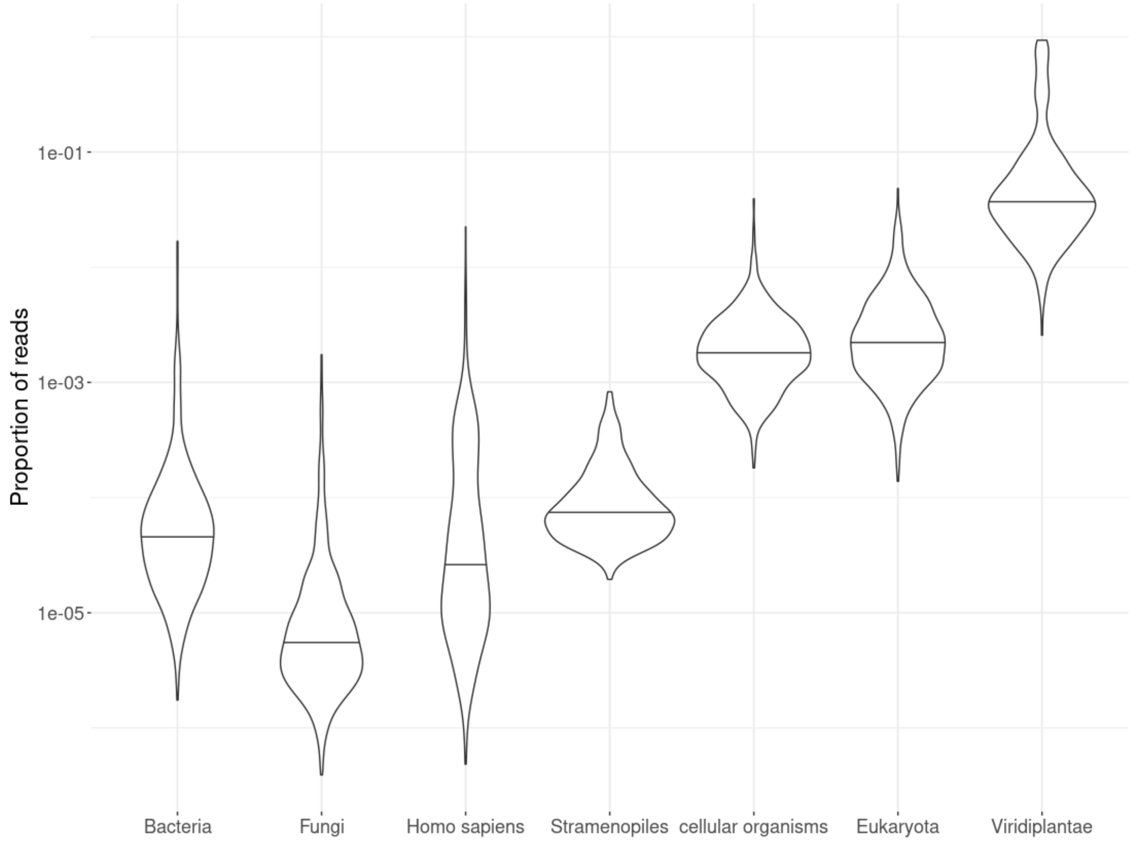
Supplementary Figure 1. The assignment accuracy of k-mer = 28 and k-mer = 34 of simulated *Calluna vulgaris* reads by the *wholeskim*-unassembled workflow. Using an effective k-mer size of 28 is able to assign more reads, but has a significantly larger misassignment rate than a k-mer size of 34.



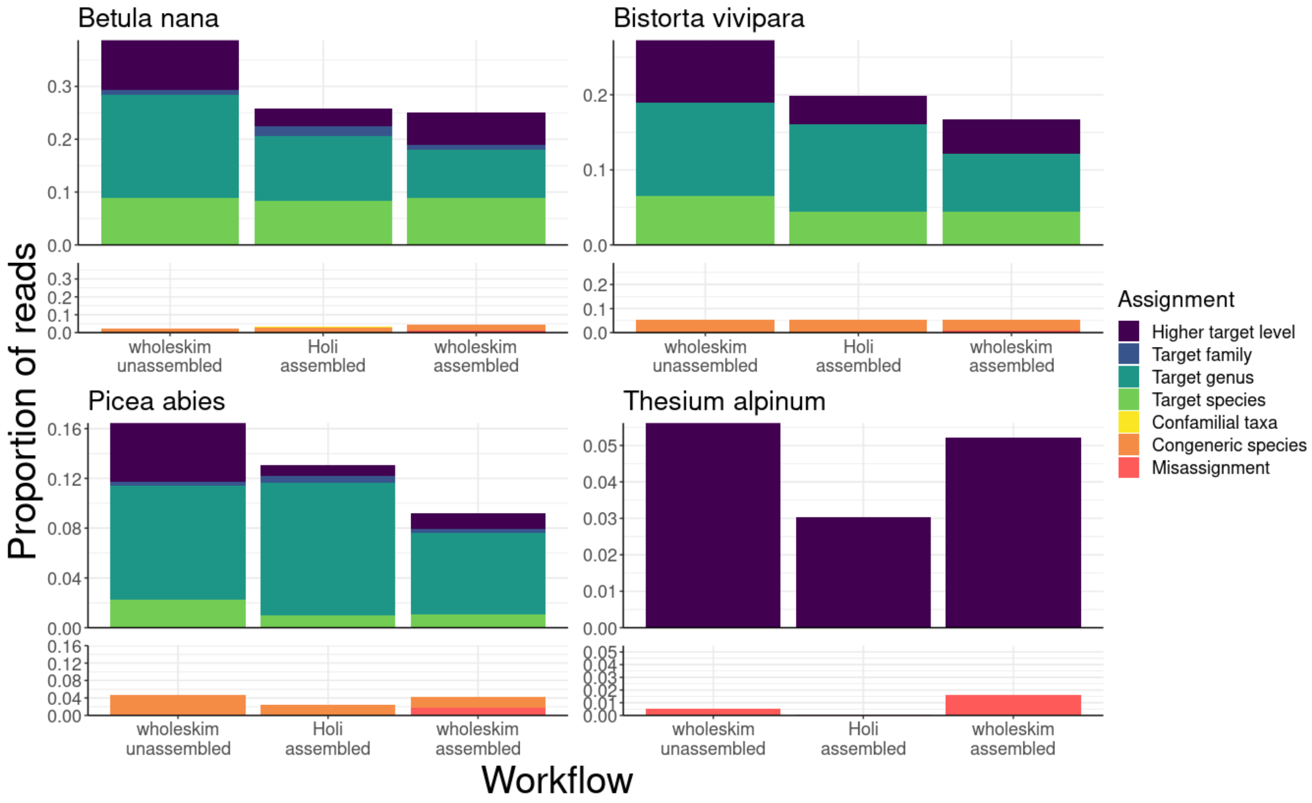
Supplementary Figure 2. The assignment of simulated *Calluna vulgaris* reads by the *wholeskim*-unassembled workflow with varying cutoff levels (t). A cutoff value of 70% provides the most favorable ratio of *Calluna vulgaris* assigned reads to off-target assigned reads (22.9). Note that the taxonomic assignment cutoff (r = 0.0001) is not applied here to the misassigned reads.


Raw reads
bowtie2
RefSeq (release 96) fungi, viral, archaea
Aligned reads
Non-aligned reads (output from --un)
Read extension using tadpole from BBmap (allowing 1000 nt added on either end)
58% of reads extended. Mean read length 100 -> 188 bp
Assembled with megahit --k-min 21 --k-max 141 --k-step 10
bowtie2 mapping extended reads to contigs
Average of 1.4 million contigs per plant (mean length 216 bases)
SOAPdenovo
Lengths not significantly improved so discarded
Aligned reads
Non-aligned reads (output from --un)
Read pairs merged with AdapterRemoval
Combined contigs and non-mapping extended reads
- Remove reads < 300 bp [i.e. all nonaligned reads]
- Remove duplicates with rmdup
- Remove low-complexity reads (dust value > 31) [still many 5bp repeats in dataset]
- Remove records similarity > 98% with clumpify from BBmap
Final whole genome dataset: mean 203,047 records per plant mean length of 496 bp
- nt and RefSeq
- 14 mammalian genomes (including human)
- cp and nrDNA from PhyloNorway
Merged reference dataset of 53 fasta files 1,186 billion bp
kmer-based coverage estimate
- All records split into 32-mers with taxa ID
- Moved kmers to different partitions based on first 6 bases
- Resulted in 4096 files (4^6)
- Merged identical kmers, assigned LCA to kmers in common
- Counted numbed of kmers for each taxonomic node to reflect coverage
bowtie2-build
Shotgun environmental data
53 separate bowtie databases
bowtie2
53 bam files
samtools
- Kept only reads with 100% similarity to reference
- Merged all 53 files together
ngsLCA
- LCA of alignments
- Threshold of 5 unique reads per taxon
Reads assigned to taxa
Taxa represented in lab controls removed [assumably humans]
Plant taxa composed of <1% of total Viridiplantae reads removed

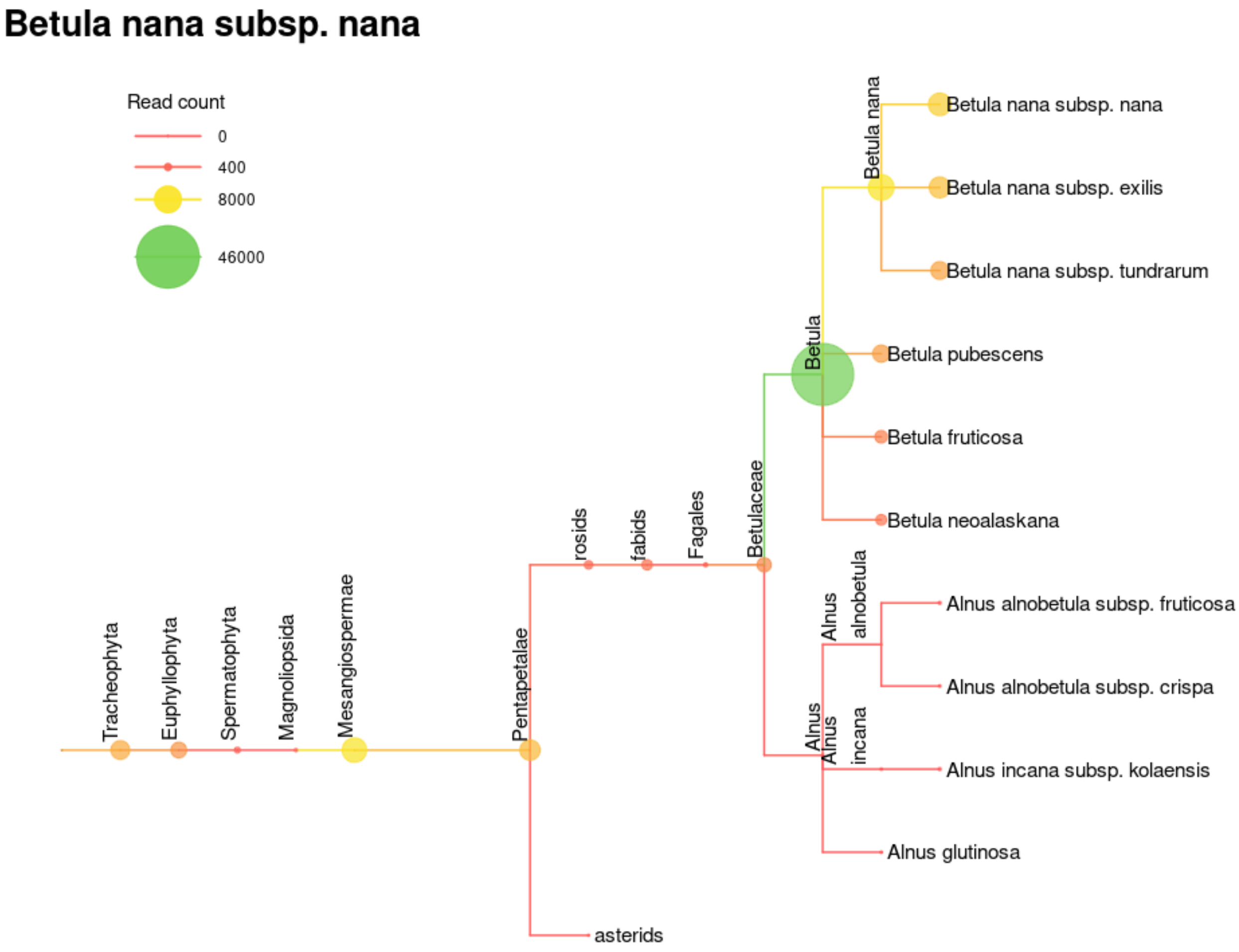
Supplementary Figure 3. A diagrammed workflow of the PhyloNorway genome skims processed by the *Holi* pipeline as applied in Wang et al. 2021.

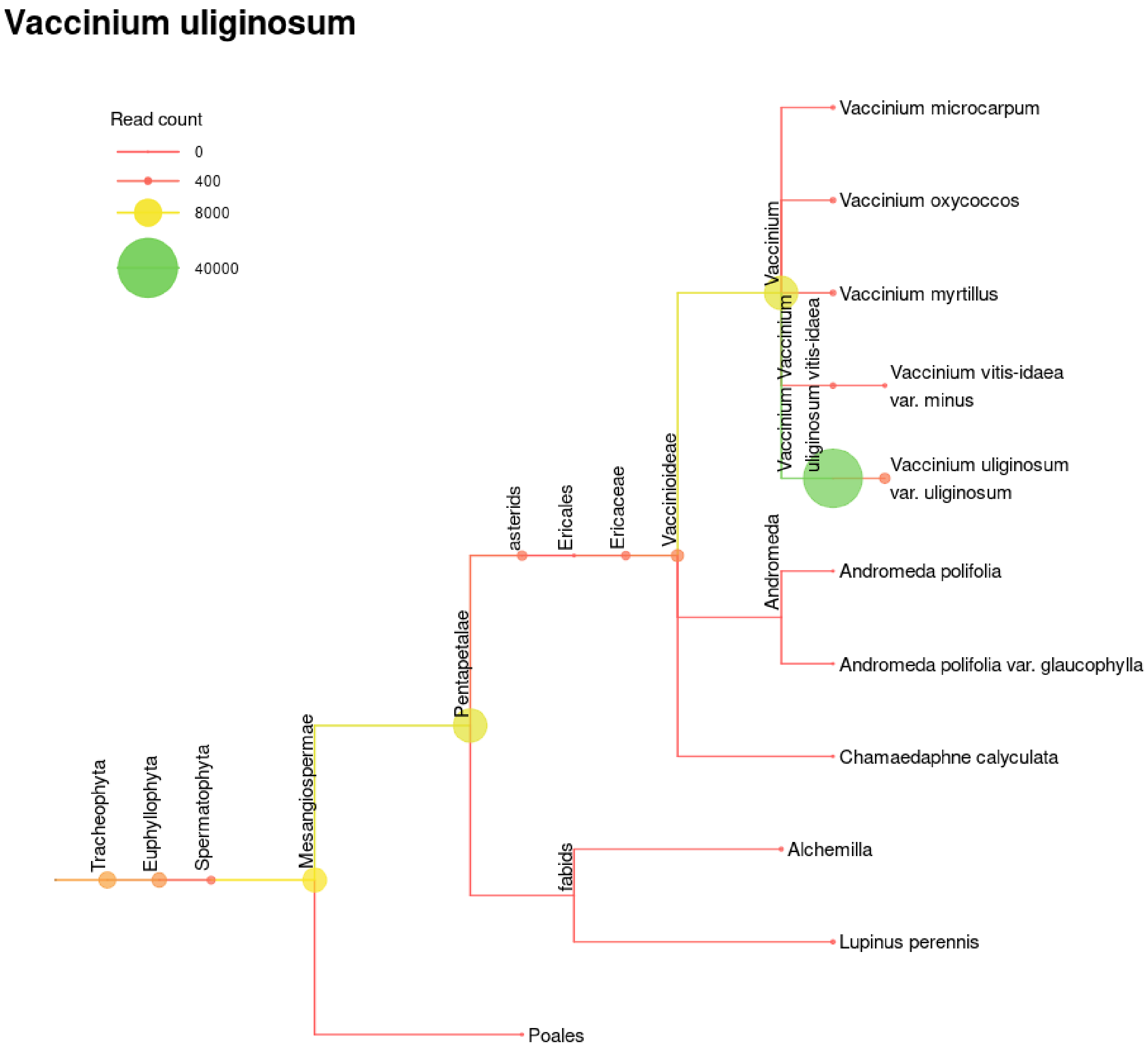


Supplementary Figure 4. Violin plot of the proportion of identified reads for each PhyloNorway genome skim used to construct the reference database. Reads identified to the groups Algae, Bacteria, Fungi, Homo sapiens, or Opisthokonts were discarded as putative contamination. The genome skims of a subset of aquatic plants are highlighted in blue as possessing more than average amounts of contaminated reads.

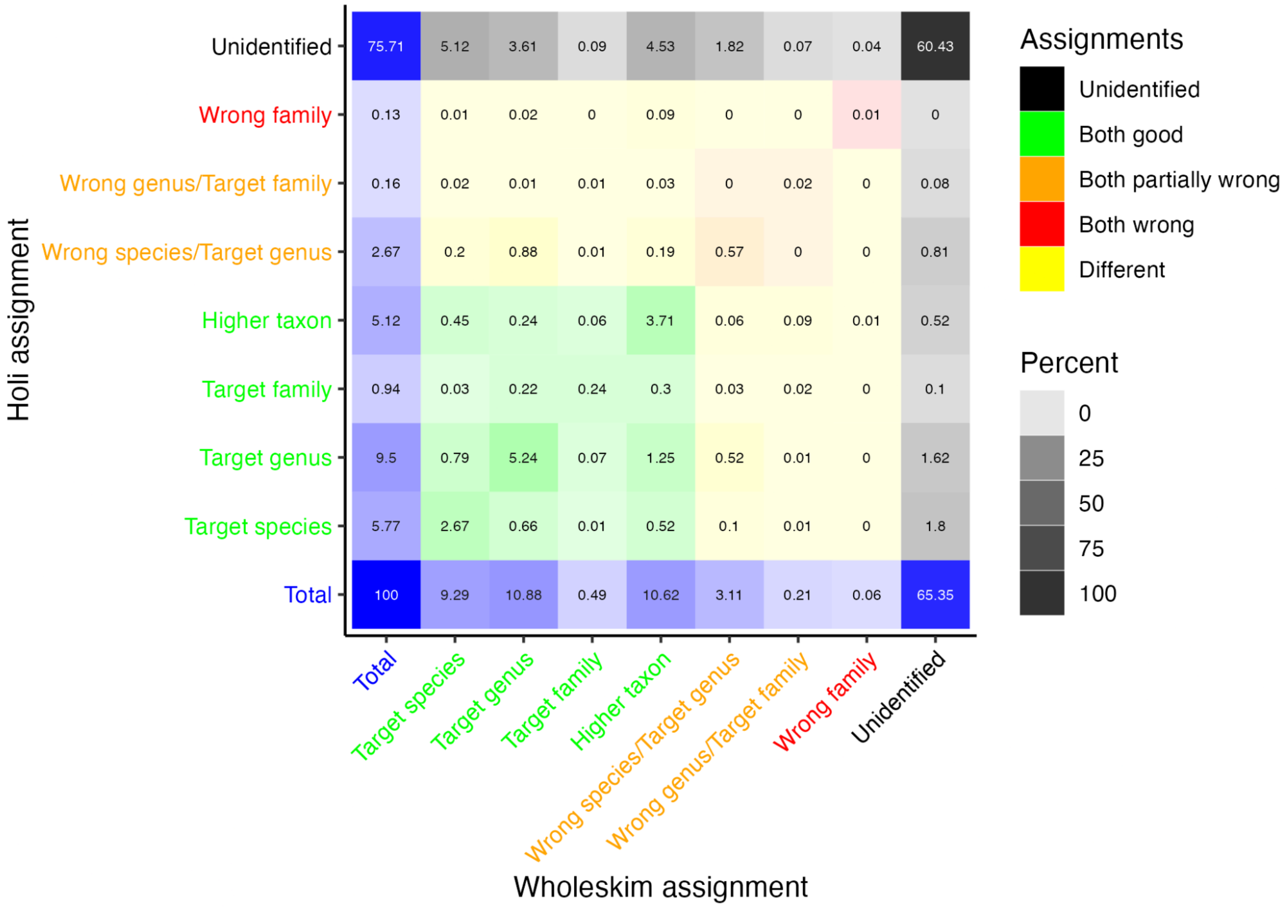


Supplementary Figure 5. Assignment accuracy of simulated reads including the *wholeskim*-assembled workflow. Note the variable y-axis scales between species.





Supplementary Figure 6. Phylogenetic tree of assigned simulated reads for *Betula nana* (A) and *Vaccinium uliginosum* (B). Both the color and size of the node are proportional to the number of reads assigned to that taxon.



Supplementary Figure 7: A confusion matrix of the cumulative assignments for all nine test species present in PhyloNorway. The confusion matrix presented here illustrates how the same read can have different accuracy of assignment from one pipeline to another.
Identifications made on the taxonomic assignment by both the pipelines can be categorized according to their quality. The total columns and row present the intrinsic results respectively of Wholeskim and Holi. All values are expressed in percent of the annotated reads.

References

Alsos, Inger Greve, Sebastien Lavergne, Marie Kristine Føreid Merkel, Marti Boleda, Youri Lammers, Adriana Alberti, Charles Pouchon, et al. 2020. “The Treasure Vault Can Be Opened: Large-Scale Genome Skimming Works Well Using Herbarium and Silica Gel Dried Material.” *Plants* 9 (4). https://doi.org/10.3390/plants9040432.

Bloom, Burton H. 1970. “Space/time Trade-Offs in Hash Coding with Allowable Errors.” *Communications of the ACM* 13 (7): 422–26.

Borry, Maxime. 2018. *ADRSM: Ancient DNA Read Simulator for Metagenomics* (version v0.9.5). https://doi.org/10.5281/zenodo.1462743.

Bushnell, Brian. 2014. “BBMap: A Fast, Accurate, Splice-Aware Aligner,” March. https://escholarship.org/uc/item/1h3515gn.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. “Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor.” *Bioinformatics* 34 (17): i884–90.

Coissac, Eric, Peter M. Hollingsworth, Sébastien Lavergne, and Pierre Taberlet. 2016. “From Barcodes to Genomes: Extending the Concept of DNA Barcoding.” *Molecular Ecology* 25 (7): 1423–28.

Courtin, J., A. Perfumo, A. A. Andreev, and T. Opel. 2022. “Pleistocene Glacial and Interglacial Ecosystems Inferred from Ancient DNA Analyses of Permafrost Sediments from Batagay Megaslump, East Siberia.” *The Environmentalist*. https://onlinelibrary.wiley.com/doi/abs/10.1002/edn3.336.

Dabney, Jesse, Matthias Meyer, and Svante Pääbo. 2013. “Ancient DNA Damage.” *Cold Spring Harbor Perspectives in Biology* 5 (7). https://doi.org/10.1101/cshperspect.a012567.

Firtina, Can, Jisung Park, Mohammed Alser, Jeremie S. Kim, Damla Senol Cali, Taha Shahroodi, Nika Mansouri Ghiasi, et al. 2023. “BLEND: A Fast, Memory-Efficient and Accurate Mechanism to Find Fuzzy Seed Matches in Genome Analysis.” *NAR Genomics and Bioinformatics* 5 (1): lqad004.

Gilbert, Jack A., Janet K. Jansson, and Rob Knight. 2014. “The Earth Microbiome Project: Successes and Aspirations.” *BMC Biology* 12 (August): 69.

Graham, Russell W., Soumaya Belmecheri, Kyungcheol Choy, Brendan J. Culleton, Lauren J. Davies, Duane Froese, Peter D. Heintzman, et al. 2016. “Timing and Causes of Mid-Holocene Mammoth Extinction on St. Paul Island, Alaska.” *Proceedings of the National Academy of Sciences* 113 (33): 9310–14.

Harbert, R. S. 2018. “Algorithms and Strategies in Short‐read Shotgun Metagenomic Reconstruction of Plant Communities.” *Applications in Plant Sciences*. https://bsapubs.onlinelibrary.wiley.com/doi/abs/10.1002/aps3.1034.

Kim, Daehwan, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. 2016. “Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences.” *Genome Research* 26 (12): 1721–29.

Kjær, Kurt H., Mikkel Winther Pedersen, Bianca De Sanctis, Binia De Cahsan, Thorfinn S. Korneliussen, Christian S. Michelsen, Karina K. Sand, et al. 2022. “A 2-Million-Year-Old Ecosystem in Greenland Uncovered by Environmental DNA.” *Nature* 612 (7939): 283–91.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Lemane, Téo, Nolan Lezzoche, Julien Lecubin, Eric Pelletier, Magali Lescot, Rayan Chikhi, and Pierre Peterlongo. 2023. “Kmindex and ORA: Indexing and Real-Time User-Friendly Queries in Terabytes-Sized Complex Genomic Datasets.” *bioRxiv*. https://doi.org/10.1101/2023.05.31.543043.

Lewin, Harris A., Stephen Richards, Erez Lieberman Aiden, Miguel L. Allende, John M. Archibald, Miklós Bálint, Katharine B. Barker, et al. 2022. “The Earth BioGenome Project 2020: Starting the Clock.” *Proceedings of the National Academy of Sciences of the United States of America* 119 (4). https://doi.org/10.1073/pnas.2115635118.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, et al. 2018. “Earth BioGenome Project: Sequencing Life for the Future of Life.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (17): 4325–33.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. “MEGAHIT:

An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph.” *Bioinformatics* 31 (10): 1674–76.

Mohamadi, Hamid, Hamza Khan, and Inanc Birol. 2017. “ntCard: A Streaming Algorithm for Cardinality Estimation in Genomics Data.” *Bioinformatics* 33 (9): 1324–30.

Parducci, L., I. G. Alsos, and P. Unneberg. 2019. “Shotgun Environmental DNA, Pollen, and Macrofossil Analysis of Lateglacial Lake Sediments from Southern Sweden.” *Frontiers in Ecology and the Environment*. https://www.frontiersin.org/articles/10.3389/fevo.2019.00189/full.

Parducci, Laura, Keith D. Bennett, Gentile Francesco Ficetola, Inger Greve Alsos, Yoshihisa Suyama, Jamie R. Wood, and Mikkel Winther Pedersen. 2017. “Ancient Plant DNA in Lake Sediments.” *The New Phytologist* 214 (3): 924–42.

Pedersen, Mikkel W., Anthony Ruter, Charles Schweger, Harvey Friebe, Richard A. Staff, Kristian K. Kjeldsen, Marie L. Z. Mendoza, et al. 2016. “Postglacial Viability and Colonization in North America’s Ice-Free Corridor.” *Nature* 537 (7618): 45–49.

Revéret, A., D. P. Rijal, and P. D. Heintzman. 2023. “Environmental DNA of Aquatic Macrophytes: The Potential for Reconstructing Past and Present Vegetation and Environments.” *Freshwater*. https://onlinelibrary.wiley.com/doi/abs/10.1111/fwb.14158.

Robidou, Lucas, and Pierre Peterlongo. 2021. “Findere: Fast and Precise Approximate Membership Query.” *bioRxiv*. https://doi.org/10.1101/2021.05.31.446182.

Slon, Viviane, Charlotte Hopfe, Clemens L. Weiß, Fabrizio Mafessoni, Marco de la Rasilla, Carles Lalueza-Fox, Antonio Rosas, et al. 2017. “Neandertal and Denisovan DNA from Pleistocene Sediments.” *Science* 356 (6338): 605–8.

Stoler, Nicholas, and Anton Nekrutenko. 2021. “Sequencing Error Profiles of Illumina Sequencing Instruments.” *NAR Genomics and Bioinformatics* 3 (1): lqab019.

Straub, Shannon C. K., Mark Fishbein, Tatyana Livshultz, Zachary Foster, Matthew Parks, Kevin Weitemier, Richard C. Cronn, and Aaron Liston. 2011. “Building a Model: Developing Genomic Resources for Common Milkweed (Asclepias Syriaca) with Low Coverage Genome Sequencing.” *BMC Genomics* 12 (May): 211.

Sultana, Nusrat, Joan Pere Pascual-Díaz, Ahsen Gers, Kübra Ilga, Sedat Serçe, Daniel Vitales, and Sònia Garcia. 2020. “Contribution to the Knowledge of Genome Size Evolution in Edible Blueberries (genus Vaccinium).” *Journal of Berry Research* 10 (2): 243–57.

Taberlet, P., A. Bonin, L. Zinger, and E. Coissac. 2018. “Environmental DNA: For Biodiversity Research and Monitoring.” https://books.google.ca/books?hl=en&lr=&id=1e9IDwAAQBAJ&oi=fnd&pg=PP1&ots=UY8TqcnfoR&sig=8kUMa4OFtZC1Z2n5fT7MV7cTuSo.

Von Eggers, J., M. E. Monchamp, and E. Capo. 2022. “Inventory of Ancient Environmental DNA from Sedimentary Archives: Locations, Methods, and Target Taxa. Zenodo.” https://scholar.google.ca/scholar?cluster=15840593690555241839&hl=en&as_sdt=0,5&sciodt=0,5.

Wang, Yucheng, Mikkel Winther Pedersen, Inger Greve Alsos, Bianca De Sanctis, Fernando Racimo, Ana Prohaska, Eric Coissac, et al. 2021. “Late Quaternary Dynamics of Arctic Biota from Ancient Environmental Genomics.” *Nature* 600 (7887): 86–92.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. “Improved Metagenomic Analysis with Kraken 2.” *Genome Biology* 20 (1): 257.

Zimmermann, Heike H., Kathleen R. Stoof-Leichsenring, Viktor Dinkel, Lars Harms, Luise Schulte, Marc-Thorsten Hütt, Dirk Nürnberg, Ralf Tiedemann, and Ulrike Herzschuh. 2023. “Marine Ecosystem Shifts with Deglacial Sea-Ice Loss Inferred from Ancient DNA Shotgun Sequencing.” *Nature Communications* 14 (1): 1650.

Comparison of target enrichment and shotgun sequencing of lake sedaDNA metagenomes

Lucas Elliott, Tyler Murchie, Kathleen Stoof-Leichsenring, Nichola Strandberg, Marie Føreid Merkel, Iva Pitelkova, Inger Greve Alsos

Introduction

Lake sediment cores provide a valuable archive for characterizing current ecosystems as well as reconstructing past biodiversity changes. DNA from aquatic organisms in the lake itself in addition to terrestrial taxa from the surrounding catchment can be deposited and preserved by adsorbing to minerals (Freeman et al. 2023; Giguet-Covex et al. 2023). A subset of these molecules can be amplified through PCR at taxonomically diagnostic loci to describe what organisms are found in the catchment in a process called metabarcoding (Taberlet et al. 2018). However, metabarcoding's amplification is biased towards certain fragments based on their composition (Nichols et al. 2018) and limited to taxonomically informative genomic regions that are flanked by conserved areas (Taberlet et al. 2007). Additionally, the amplification of DNA fragments erases deamination-based damage patterns that could be used to authenticate DNA as ancient (Dabney et al. 2013; Dalén et al. 2023). By contrast, the metagenomic approach sequences the entire *seda*DNA content without PCR amplifying specified regions (Heintzman et al. 2023; Liu et al. 2024). This can be accomplished by directly sequencing the prepared library in a process termed "shotgun sequencing", (Parducci et al. 2019; Pedersen et al. 2016), or by first enriching the sequencing library for taxa or genomic regions of interest through "target enrichment" or "hybridization capture" (Mamanova et al. 2010; Murchie et al. 2021; Schulte et al. 2021; Schulte et al. 2022).

While sequencing the total *seda*DNA content, the shotgun sequencing approach often leads to a large percentage of unidentifiable reads with the next largest categories being Bacteria and Fungi (Schulte et al. 2021; Parducci et al. 2019). To compensate, studies targeting animal or plant taxa drastically increase the depth of sequencing, up to 16 billion reads generated (Kjær et al. 2022), to recover these small proportions of reads. One of the largest advantages of shotgun sequencing is that it can provide a relatively unbiased image of total DNA content of a sample since a limited preselection of molecules is occurring. The number of sequenced reads has been used to model plant taxa abundance (Wang et al. 2021), but it is uncertain how well *seda*DNA represents biomass/abundance since the taphonomic process of DNA preservation in sediments is not well understood (Heintzman et al. 2023). Additionally, taxonomic and genomic representation in the DNA reference library strongly affects detection (Elliott et al. 2024) while the short read lengths limit detection and the taxonomic resolution of annotations.

In contrast, target enrichment selects a subset of DNA reads corresponding to a specific group of taxa or genomic region. This is accomplished by first designing a “bait” set of complementary DNA molecules of interest that are used to bind a portion of the sequencing library while the remainder is discarded. These baits can be designed and synthesized or used from modern amplified DNA extracts (Maricic et al. 2010). The hybridizing temperature can be adjusted to control how similar the molecules binding to the bait set are, allowing for deaminated fragments, individual variation, and closely related taxa to still be retained (Heintzman et al. 2023). However, this can also lead to off-target sequences being enriched. Compared to shotgun sequencing, hybridization capture significantly reduces the required depth of sequencing as the resulting library is greatly enriched for DNA of interest. Both target enrichment and shotgun sequencing have been used for the assembly of organelles and nuclear regions and calling of haplogroups given sufficient coverage (Lammers et al. 2021; Pedersen et al. 2021; Vernot et al. 2021; Schulte et al. 2021; Schulte et al. 2022).

We aim to compare and contrast the taxonomic annotations of these two different methods of processing metagenomic data from lake surface samples; target enrichment and shotgun sequencing. We processed surface samples from 20 lakes across northern Norway and performed a single-stranded library preparation and direct shotgun sequencing (Gansauge et al. 2017; Schulte et al. 2021) as well as a target enrichment on arctic vascular plant taxa with a double-stranded protocol using the PalaeoChip ArcticPlant-1.0 (Murchie et al. 2021) on the same sediment subsamples. Surface samples were chosen to reflect the contemporary vegetation growing in the catchment (Alsos et al. 2018), which can be used to “ground truth” the observations made by both molecular techniques.

Methods

Study area and coring

A set of 22 lakes across northern Norway was selected based meeting most or all of the following three criteria; topography providing small inflow streams to the lake, surrounding vegetation representing a variety of ecosystems in the region from boreal forest to alpine heath, and putatively undisturbed sedimentation from human or natural forces (Supplementary Table 1). Most of these lakes and catchments were studied in Alsos et al. 2018 and Rijal et al. 2021 using metabarcoding. Surface samples were collected from the lakes in 2012 (Alsos et al. 2018) using a Kajak corer (mini gravity corer) with a diameter of 3 cm and a length of 63 cm and in 2017 using a UWITEC USC 06000 corer with a diameter of 5.9 cm (Rijal et al. 2021).

DNA extraction and purification

DNA was extracted from 250 mg of surface sediment from each lake using the DNeasy PowerSoil Extraction kit (Qiagen, Germany) and following the Murchie et al. (2021) protocol in the ancient DNA laboratory at the Arctic University Museum of Norway in Tromsø. This protocol features an additional centrifuge step where the PowerBead supernatant was added to 17.5 ml of Dabney binding buffer and spun at 4200 rpm for 20 hours overnight to remove inhibitors. An extraction blank was processed alongside each group of 8 samples. This procedure was performed once in 2019 and the extract was then processed for target enrichment. As there were no DNA extracts remaining, another round of extraction on the same homogenized sediment subsamples was performed in 2023 and subsequently processed for shotgun sequencing. In 2023, there was no sediment left from Langfjordvannet (LANG) and Nordvivannet (NORD) so we attempted to recover DNA from the storage tubes of the previous extracts. This process was successful for Langfjordvannet, but resulted in a DNA concentration <0.5 ng/μl for Nordvivannet so this sample was excluded from further analysis.

Shotgun library preparation and sequencing

Library preparation of of 21 samples was performed in the paleogenetic laboratories at Alfred Wegener Institute (AWI) Helmholtz Centre for Polar and Marine Research in Potsdam, Germany, using single-stranded library preparation designed specifically for highly degraded ancient DNA (Gansauge et al. 2017; Gansauge and Meyer 2013; Schulte et al. 2021). From the previously described extraction process, 15 ng DNA was used as a DNA input. In total three library batches containing 24 samples and three library blank controls were included. One library included three other *sed*aDNA extracts from northern Norway not included in the study (Komatsu et al. in prep).

Libraries were quantified using qPCR (Gansauge and Meyer 2013): The qPCR setup contained 1x Maxima™ SYBR™ Green (Thermo Scientific, Germany), 0.2 μM IS7 and IS8, and 1 μL of the libraries diluted 1:20 with TET buffer in a final volume of 25 μL. The qPCR was run in Quant Studio 3 (Thermo Fisher Scientific) with the following settings: 95 °C for 10 min, followed by 40 cycles of 30 s at 95 °C, 30 s at 60 °C and 30 s at 72 °C. The fluorescence was acquired after each cycle and the amplification curve was used to estimate the needed number of amplification cycles during indexing PCR.

Indexing PCR was performed in 10–14 cycles (for samples and blanks) depending on the library concentration with indexed P5 (5′−3′: AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTT; IDT, Germany) and P7 (5′−3′: CAAGCAGAAGACGGCATACGAGATNNNNNNNGTGACTGGAGTTCAGACGTGT, IDT, Germany) primers using AccuPrime Pfx polymerase (Life Technologies, Germany) and 24 μL of each library. The blank was amplified with the same number of cycles, although the initial DNA concentration was 0 ng/μL. Amplificates were purified with the MinElute PCR Purification Kit (Qiagen, Germany) and library size distribution was checked on the Agilent TapeStation using the D1000 ScreenTape (Agilent Technologies, USA). The samples were pooled equimolarly, but with the negative control in a 10:1 ratio to achieve a final molarity of 20 mM. The 21 samples, three extraction blanks, and one library blank, were sequenced in paired-end mode (2 × 100 bp) on an Illumina NextSeq2000 device at the sequencing facility at Alfred Wegener Institute

Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany. One lake, NESS, failed to produce any sequences and was excluded from further analysis.

Target enrichment, library preparation, and sequencing

DNA extracts of 15 ng from the 22 lakes were enriched and sequenced at McMaster Ancient DNA Centre, Hamilton, Canada in 2019. The sequencing library was prepared using Meyer and Kircher's (2010) double-stranded method with modifications from Kircher et al. (2012), and a modified end-repair reaction to account for the lack of uracil excision. Samples were purified after blunt-end repair with a QIAquick PCR Purification Kit (QIAGEN) (to maximally retain small fragments) and after adapter ligation with a MinElute PCR Purification Kit (QIAGEN). Libraries were quantified using the short amplification primer set with PCR duplicates to assess library preparation success, and thereafter were uniquely dual indexed. Libraries with higher amplification RFU values (relative to other libraries in the same indexing reaction set) were removed after 11 PCR cycles, while samples with lower RFU values were left for an additional 3 cycles. All were re-added for the final round of extension.

The adapter-ligated, dual-indexed double-stranded DNA libraries were then enriched using the PalaeoChip ArcticPlant-1.0 bait-set, which had been designed in collaboration with Arbor Biosciences (Murchie et al. 2019). The bait-set targets ~2100 arctic vascular plant and bryophyte taxa, based on the databases available from Sønstebø et al. (2010), Soininen et al. (2015), and Willerslev et al. (2014). The chloroplast locus *trn*L (UAA) is the primary target of these reference databases; additional full *trn*L loci from GenBank were added to the bait-set to augment some of the particularly short sequences (<50 bp) available in the original references. The loci *rbc*L and *mat*K were also added where available to further increase the chloroplast targeting scope.

Hybridization and bait mixes were prepared to a desired concentration of 100 ng of baits per reaction. An indexed library input of 5 µL was combined with Bloligos (blocking oligos which prevent the hybridization between library adapter sequences). The hybridization and bait mixes were pre-warmed to 60°C before being combined with the library-Bloligo mixture. The final reaction was incubated for 45 hours at 60°C for bait-library hybridization. After the two-day hybridization, beads were dispensed (20 µL per reaction), washed three times with an equivalent volume of binding buffer per library, then resuspended in binding buffer and aliquoted into PCR strips. Baits were captured using 20 µL of the bead binding buffer suspension per library, incubated at 60°C for 2.5 minutes, finger vortexed and spun down, then incubated for another 2.5 minutes. Beads were pelleted and the supernatant (the non-captured library fraction) was removed and stored at −20°C as per Klunk et al. (2019). The beads were resuspended in 180 µL of 60°C Wash Buffer X per tube and washed four times following the MYbaits V4 protocol. Beads were eluted in 15 µL EBT, PCR reamplified for 12 cycles, then purified with MinElute columns following manufacturer's protocols, and finally eluted in 15 µL EBT.

Total enriched DNA was quantified using the long amplification primer set with PCR duplicates. These values were used to calculate the dilution ratio for equimolar pooling. The goal of pooling being to attain ~1,000,000 sequenced reads per library, but while maintaining a pool molarity of ≳200 pM that can be detected following a post-pooling purification as well as size-selection procedure. The pools were size-selected with gel excision following electrophoresis for molecules ranging between 150–600 bp. Gel plugs were purified using the QIAquick Gel Extraction Kit (QIAGEN), according to manufacturer's protocol, then sequenced on an Illumina HiSeq 1500 with a 2 x 90 bp paired-end protocol at the Farncombe Metagenomics Facility (McMaster University, ON). The sample from lake Nesservatnet (NESS) only produced 3,763 read pairs and was excluded from further analysis.

Bioinformatic processing

Shotgun data

The quality of the raw sequencing data was checked using FastQC v0.12.0 (Andrews 2010) and reads were then merged and adapter sequences were removed using *fastp* v0.23.4 (Chen 2023). Taxonomic annotation of the shotgun reads was performed by the *wholeskim* pipeline (Elliott et al. 2024). Merged reads from the shotgun dataset less than 34 bp were discarded as they are unable to be identified using *wholeskim* with an effective k-mer size of 33. A k-mer similarity cutoff was set to 0.7 and taxa were only retained if they composed at least 1% of the total reads identified to Embryophyta with a minimum read count of 5 (Elliott et al. in prep). The reference database used for annotation was constructed using the PhyloNorway database comprising genome skims from 1,823 species (Alsos et al. 2020) as well as the NCBI RefSeq entries with a "complete genome" assembly level for bacteria (2), fungi (4751), and algae which were compiled by collecting the plant entries which belonged to the following groups; Stramenopiles (33634), Rhodophyta (2763), and Chlorophyta (3041). No other plant sequences were added to the database as PhyloNorway provides a relatively even coverage reference for all Norwegian flora including common *seda*DNA contaminants such as *Triticum aestivum* and *Solanum tuberosum*. Reads from the shotgun dataset were assigned to the LCA using the built-in parsing from *wholeskim* which considers all matches within 10% k-mer similarity of the maximum match (Elliott et al. in prep). To evaluate how many reads were mapped to both plant and off-target taxa, we also ran the *wholeskim* pipeline using only the PhyloNorway portion of the database.

Target enrichment data

The quality of the raw sequencing data was checked using FastQC v0.12.0 (Andrews 2010) and reads were then merged and adapter sequences were removed using *fastp* v0.23.4 (Chen 2023). Merged reads less than 30 bp were discarded as taxonomic resolution is poor for these short fragments (Pedersen et al. 2016). The target enrichment reads were mapped to a custom database using *bowtie2* with default parameters (Langdon 2015). The custom database was constructed by compiling the 1,845 assembled plastid genomes produced by the PhyloNorway project (Alsos et al. 2020) as well as the NCBI RefSeq entries with a "complete genome" assembly level for bacteria (NCBI taxid: 2), fungi (4751), and algae which were compiled by collecting the plant entries which belonged to the following groups; Stramenopiles (33634), Rhodophyta (2763), and Chlorophyta (3041). Reads from the target enrichment dataset were assigned to the lowest common ancestor (LCA) using *ngsLCA* with a minimum edit distance proportion of 0.97 (Wang et al. 2022). Taxa were only retained in the final dataset if they had at least 5 reads present in a sample.

Workflow comparison

Identifications from both workflows were collapsed to genus-level and any identifications to above family-level were not considered for comparing plant assemblies (although these taxa were used to compute the total number of Embryophyta reads per sample). To characterize the similarity of plant assemblies detected by target enrichment and shotgun sequencing, we used nonmetric multidimensional scaling (nMDS) with Bray-Curtis dissimilarity as implemented in the *vegan* function *metaMDS* (Oksanen et al. 2020; Chen and Ficetola 2020). For this ordination, we used the proportion of each taxon's reads to the total Embryophyta reads identified in each sample. To obtain a stress-value < 0.5, the nMDS was run with 5 dimensions (k = 5 with *metaMDS*).

Mapping

To validate taxonomic assignments, we choose two taxa, *Alnus* and *Utricularia*, with available assembled genomes to map the metagenomic reads to. Reads annotated to these two taxa from the shotgun dataset were compiled from all 20 lakes. As comparison, *Alnus* and *Utricularia* metagenomic reads from the arctic shotgun sequenced dataset Wang et al. 2021 (ENA project code: PRJEB43822) were also compiled. Since the taxonomic annotations of the individual reads were not published by Wang et al. (2021) (uploaded *bam* files have all read FLAGs set to unmapped), the ten sites with the most abundant *Alnus* and *Utricularia* presence were selected from Supplementary Data 5. The truncated and merged reads were downloaded from these sites (Supplementary Table 2) and processed through the *Holi* pipeline as detailed in (https://github.com/miwipe/KapCopenhagen) using the PhyloNorway contig database. This resulted in a total dataset of 11,692 *Alnus* reads and 2,367 *Utricularia* reads. Using *bwa mem* with default parameters, we then mapped these reads to the three largest chromosomes of a representative from each genus; *Alnus glutinosa* (OY340898.1, OY340899.1, and OY340900.1,) and *Utricularia gibba* (CM007989.1, CM007990.1, and CM007991.1). We calculated genomic coverage estimates through *samtools coverage (Li et al. 2009)*.

Results

Shotgun sequencing results

A total of 1,040,366,383 read pairs were obtained from the 20 lakes with an average of 52,018,319 ± 13,462,696 read pairs per sample. The three extraction negative controls and two library blanks totalled 357,831 read pairs. After adapter trimming, merging, and filtering sequences < 34 bp a total of 699,592,884 (67.2%) sample sequences were retained, while 15,675 (4.4%) negative control sequences were retained. A total of 4,256,710 sequences (0.62%) were annotated to Embryophyta while 56,795,531 (8.1%) were annotated to Bacteria (Figure 1). Two taxa were identified in the negative controls with >10 read counts; *Avena* with 32 (reads and *Triticum* with 12 reads, both in one of the extraction controls. However, neither of these appear in the sample dataset. We identified 33 taxa across the 20 samples with the most abundant in read counts being *Hippuris*, *Utricularia*, and *Betula* appearing in 11, 14, and 8 samples respectively. The number of taxa detected per lake ranged from zero (BEAR) to ten (JULA and EINL). Competitively matching the results from non-decontaminated genome skims to bacteria, fungi, and algae removed a variable percentage of reads from different taxa ranging from 4.2% (*Myriophyllum*) to 72.6% (*Cochlearia*) (Figure 2). Taxa with greater than 90% of reads retained all have a known distribution in northern Norway.

Target enrichment sequencing results

A total of 25,487,638 read pairs were obtained from the 20 lakes with an average of 1,274,382 ± 634,466 read pairs per sample. The two extraction controls and library blank totalled 17,426 read pairs. After adapter trimming and merging, a total of 19,705,695 (77.3%) sample sequences were retained, while 9,216 (52.9%) control sequences were retained, all from the two extraction controls. Four taxa were identified in the negative controls with >10 reads; Pinaceae/*Pinus* with 835 reads, Triticeae/*Triticum* with 26 reads, *Myrica* with 18 reads, and *Sparganium* with 15 reads. Pinaceae/*Pinus* has been documented as a low-frequency contaminant in *seda*DNA studies (Alsos et al. 2020) and is the only contaminant taxon that appears in samples with 230 reads annotated from Gauptjern (GAUP) and 77 annotated from Guossajávri (GUOS). As *Pinus sylvestris* has been recorded in the catchment of both sites (Alsos, pers. obs.) they may represent a true positive, but we chose to be conservative and removed them for further analysis. A total of 76,284 (0.39%) were annotated to Embryophyta while 194,170 (0.98%) were annotated to Bacteria.

We identified 21 taxa in the target enrichment data across the 20 samples. One sample, Paulan Järvi (PAUL), had two taxa with dominant read counts; *Callitriche* (22,250) and its family Plantaginaceae (23,267) which compose 59.7% of the entire Embryophyta target enrichment dataset (Figure 1). The number of taxa detected in each lake ranged from zero with five lakes having fewer than 5 reads identified to Embryophyta taxa (BEAR, EAST, FINN, KOM4, and ROTT) to nine taxa (EINL).

Comparing assignments

Ten of the 21 taxa identified in the target enrichment dataset do not have a direct counterpart in the shotgun dataset; Betulaceae, Ericaceae, Ericoideae, Cyproideae, Saliceae, Potamogetonaceae, *Dryopteris*, *Dryas*, *Equisetum*, and *Stuckenia* although constituent genera of the first six clades are found in the shotgun dataset (Supplementary Table 2). Three of the remaining four genera are only found in one sample each while *Equisetum* is detected in four samples. Within each lake sample, there is very little overlap in taxa identified by the two methods (Figure 3). Consequently, there is little clustering by lake in the nMDS ordination, except for those lakes where shotgun sequencing did not detect any aquatic taxa; LANG and SIER (Figure 4). However, there is clear separation between methods in the ordination with shotgun sequenced and target enrichment samples largely clustering together (also with the exception of LANG and SIER). The shotgun sequenced samples are characterized by mostly aquatic taxa that do not appear in the target enrichment samples (e.g. *Zannichelia, Elatine, Lemna,* and *Hippuris*). Conversely, many of the more abundant terrestrial taxa are associated mostly with the target enrichment samples or both methods (e.g. *Empetrum*, *Vaccinium*, and *Alnus*). Performing the nMDS ordination on only the 11 taxa identified by both workflows reduced the separation by method, but did not result in lake samples pairing together (Supplementary Figure 2).

Mapping to reference genome

We compiled reads from all 20 lakes in the shotgun dataset identified to *Alnus* (n = 72,458) and *Utricularia* (n = 177,967) to map their corresponding chromosomes and compare with those reads from *Alnus* (n = 11,692) and *Utricularia* (n = 2,637) found in the Wang et al. 2021 study. From our dataset, 50.3% (36,571) of the *Alnus* reads mapped to the first three chromosomes of *Alnus glutinosa* while a comparable proportion (49.6%, or 5,801 reads) of the Wang et al. 2021 dataset mapped to the same chromosomes (Figure 5A, 5B). Both datasets produced relatively even coverage as expected for a metagenome. In comparison, only 269 reads (0.15%) of the *Utricularia* reads sporadically mapped to the first three chromosomes of *Utricularia gibba* while none of the reads from Wang et al. 2021 mapped (Figure 5C).

Discussion

Using both target enrichment and shotgun sequencing, we detected a limited number of taxa from the metagenomes of surface sediment samples from 20 lakes in northern Norway. The 37 taxa we detect across the 20 samples is roughly equivalent with previous metagenomic sedaDNA studies that report plant results. Before the completion of the PhyloNorway reference database, Pedersen et al. 2016 reported 57 plant taxa over 18 samples in one lake while Parducci et al. 2019 recorded 51 plant taxa over 14 samples in a single lake using NCBI's nt database. In addition, both of these studies recorded significantly less reads identified to Viridiplantae, 0.05% and 0.0002% respectively. Adding PhyloNorway as a reference database, Wang et al. 2021 reported 1,020 plant taxa in 499 circum-arctic samples (average 8.1 taxa per sample) with a greatly improved 1.7% of the total reads assigned to Viridiplantae. Using an expanded version of the PhyloNorway database (Elliott et al. in prep), we only recovered 0.6% Viridiplantae reads with shotgun sequencing and 0.2% with target enrichment. A key determining factor in taxonomic richness of sedaDNA samples is the cutoff criteria used by each workflow determined by optimizing the false negative to false positive ratio. Our conservative approach of requiring a taxon to have >1% of total Embryophyta reads could be allowing an overly abundant taxon to drown out the signal of other true positives. However, surface samples have been documented as problematic for DNA recovery due to their high organic and inhibitor content as well their biologically active community of both prokaryotes and eukaryotes (Capo et al. 2021).

The proportion of Embryophyta reads recovered from the methods is inverted compared to expectations (Schulte et al. 2021; Murchie et al. 2021) with shotgun sequencing recovering proportionally more target reads than every target enrichment sample other than PAUL. We posit that the large number of *Callitriche* and Plantaginaceae reads detected in PAUL either originate directly from plant tissue from a tributary stream as potentially both of these taxa are aquatic or possible contamination since no aquatics were recorded as growing in the lake itself during the vegetation survey (Alsos et al. 2018) and there are no records of *Callitriche* from the valley (Artsdatabanken.no).

Two confounding variables could be influencing the variable taxa detections of the target enrichment and shotgun approaches. With the exception of Langfjordvannet, DNA was extracted at two different points for each workflow. These two rounds of extraction were performed on the same homogenized sediment subsamples using the same protocol and previous studies with amplicons on multiple DNA extractions from the same sample yielded consistent results (Ficetola et al. 2015). Additionally, the same DNA extract from Langfjordvannet produced two very distinct plant assemblies from either method (Figure 2), but this small difference could compound with the different library preparation methods we used. Our target enrichment method targeted double-stranded DNA while the shotgun library preparation targeted both single- and double-stranded DNA molecules (Gansauge et al. 2017). The latter approach leads to more complex libraries and is more effective at retaining damaged DNA (Kapp et al. 2021; Dalén et al. 2023). This difference between workflows could lead to more plant DNA being converted to library molecules in the shotgun sequenced dataset compared to target enrichment. However, despite the high copy number of organellar DNA, the large size of the nuclear genome makes it the main source of plant DNA preserved in sediments (Wang et al. 2021). In previous target enrichment and shotgun sequencing comparisons (Schulte et al. 2021; Murchie et al. 2021), the nuclear component of plant *seda*DNA was underrepresented as the plant entries in NCBI's nt database are largely plastid and mitochondrial sequences. In this study, PhyloNorway (Alsos et al. 2020) is used as the main reference database for both workflows as it contains both nuclear and plastid information for the complete Norwegian flora. The addition of nuclear information to the reference database has been demonstrated to increase identified reads from shotgun sequenced datasets 23 fold (Wang et al. 2021).

However, the difference in proportion of reads identified does not fully explain why the two workflows detect different vegetation communities at nearly all lakes (Figure 3, Supplementary Table 2). In other direct comparisons of the two workflows, the plant reads identified by shotgun sequencing were generally a subset of the target enrichment dataset (Schulte et al. 2021; Murchie et al. 2021). Samples generally clustered by workflow in the nMDS ordination, partly driven by method-specific taxon detections (Figure 4). When only considering the eleven taxa that were detected by both methods, there is less of a separation by method, but the two methods still do not characterize the same vegetation communities for each lake (Supplementary Figure 2).

Of the taxa detected only by shotgun sequencing, many are aquatic plants with relatively high read counts (i.e. *Hippuris*, *Elatine*, *Lemna*, *Minuartia*, *Utricularia,* and *Zannichellia*). Due to their proximity, aquatic plants are generally well-represented in lake *seda*DNA (Alsos et al. 2018; Wang et al. 2021; Revéret et al. 2023), however many of the aquatic plants mentioned previously are likely spurious assignments for both geographic as well as technical reasons. All of these taxa fall below the 90% cutoff reads retained when competitively matching to bacteria, fungi, and algae sources (Figure 2). Three of the genera listed have only one species recorded in northern Norway with very limited geographic distributions and all are included Norway's Red List for species at risk (*Lemna triscula*, *Elatine hydropiper*, and *Zannichellia palustris*) (Artsdatabanken 2021). It is unlikely that these rare taxa would be detected in many *seda*DNA samples while abundant taxa in the catchment are missing from the dataset (Alsos et al. 2018). *Utricularia minor,* has a wide distribution in northern Norway, but was only detected by metabarcoding DNA and/or vegetation survey in 2/11 lakes in Alsos et al. 2018, while being absent from six lakes where it is detected in the shotgun data (FINN, GAUP, JULA, OAER, PAUL, and ROTT). Additionally, the poor mapping of *Utricularia* reads from this dataset as well as those from Wang et al. 2021 to the *Utricularia gibba* reference genome indicate that these reads were likely spuriously annotated (Figure 5C). High proportions of contaminant reads were initially detected in the *Utricularia* and other aquatic plant species' reference genome skims and subsequently removed (Supplementary Figure 3), but they likely signal the presence of other bacteria, fungi, and algae that are not represented in RefSeq and are unable to be filtered (Elliott et al. 2024). By comparison, 50.3% of the *Alnus* reads from the shotgun dataset mapped to a reference genome (Figure 5A), < 5% of the initially identified *Alnus* reads competitively matched to a contaminant category (Figure 2), and the initial reference genome skims for *Alnus* species had low to median amounts of contamination detected (Supplementary Figure 3). All of these factors give us confidence in the assignment of *Alnus* reads in the shotgun dataset, while casting some doubt on the validity of the *Utricularia* assignments.

During the initial processing of the PhyloNorway genome skims, 42 samples were flagged as considerably contaminated with bacterial and/or fungal reads (Alsos et al. 2020), but there was likely some additional contamination that was not detected through this initial screening. *Utricularia* species have exceptionally small genomes with *Utricular gibba*'s consisting of 82 Mb (Ibarra-Laclette et al. 2013) and *Utricularia australis*' consisting of 200 Mb (Veleba et al. 2014), yet only 3.8% of the *Utricularia australis* genome skim of 6.2 million 2 x 101 bp read pairs (TROM_V_165506) is able to be mapped to a full *Utricularia gibba* genome assembly (CoGe Genome ID: 29027) resulting in an average depth of coverage of 0.2x. While a portion of these unmapped reads are certainly due to interspecific variation, there is likely a portion originating from off-target sources. In contrast to the other putatively spurious aquatic taxa, *Myriophyllum* is an aquatic plant with high read counts in the shotgun dataset, but the identified reads have very low overlap with suspected contaminants (Figure 2) and it is also detected by target enrichment in two samples (Supplementary Table 2). Additionally, *Myriophyllum alterniflorum* was recorded as present in EINL and OAER through metabarcoding and vegetation surveys (Alsos et al.

2018) where it is detected with shotgun sequencing (Figure 3). For surface sediments, one precaution to avoid this flooding of reads could be to avoid cell lysis during extraction to target only the extracellular DNA content, although this approach has been shown to recover fewer taxa (Capo et al. 2021).

Overall, while shotgun sequencing resulted in a higher proportion of annotated Embryophyta reads, the target enrichment of barcode regions and direct mapping to a curated reference database produced more robust taxonomic assignments. This study highlights two potential pitfalls of metagenomic analyses on lake sediments; 1) biologically active surface sediment samples likely contain large proportions of bacteria, fungi, and algae that either mask or can be misidentified as plant DNA from the catchment and 2) unassembled genome skims that contain some off-target reads which are unable to be filtered out with incomplete reference databases increase true positives, but can also lead to the increase of false positives.

Acknowledgements

Funding was provided by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program grant agreement no. 819192 (to Inger Greve Alsos and Antony Brown supporting Lucas Elliott) for the IceAGenT project. We would like to thank Dilli Rijal, Youri Lammers, and Peter Heintzmann among others at the Arctic University Museum of Norway, Tromsø and ECOGEN Northern Norway for their contributions in coring the lakes. We thank Janine Klimke for shotgun sequencing laboratory guidance. Bioinformatic analyses were performed on resources provided by UNINETT Sigma2–the National Infrastructure for High Performance Computing and Data Storage in Norway.

Author contributions

The study was conceptualized by L.E. and I.A.. I.P. performed the target enrichment extractions and T.M. performed the target enrichment, library preparation, sequencing, and initial analysis of this data. L.E. and M.F.M. extracted the shotgun samples and performed library preparation and sequencing with the guidance of K.S. Data was analyzed by L.E. with feedback from I.A., T.M., K.S., and N.S.. Manuscript was drafted by L.E. and commented on by all authors.

Figures

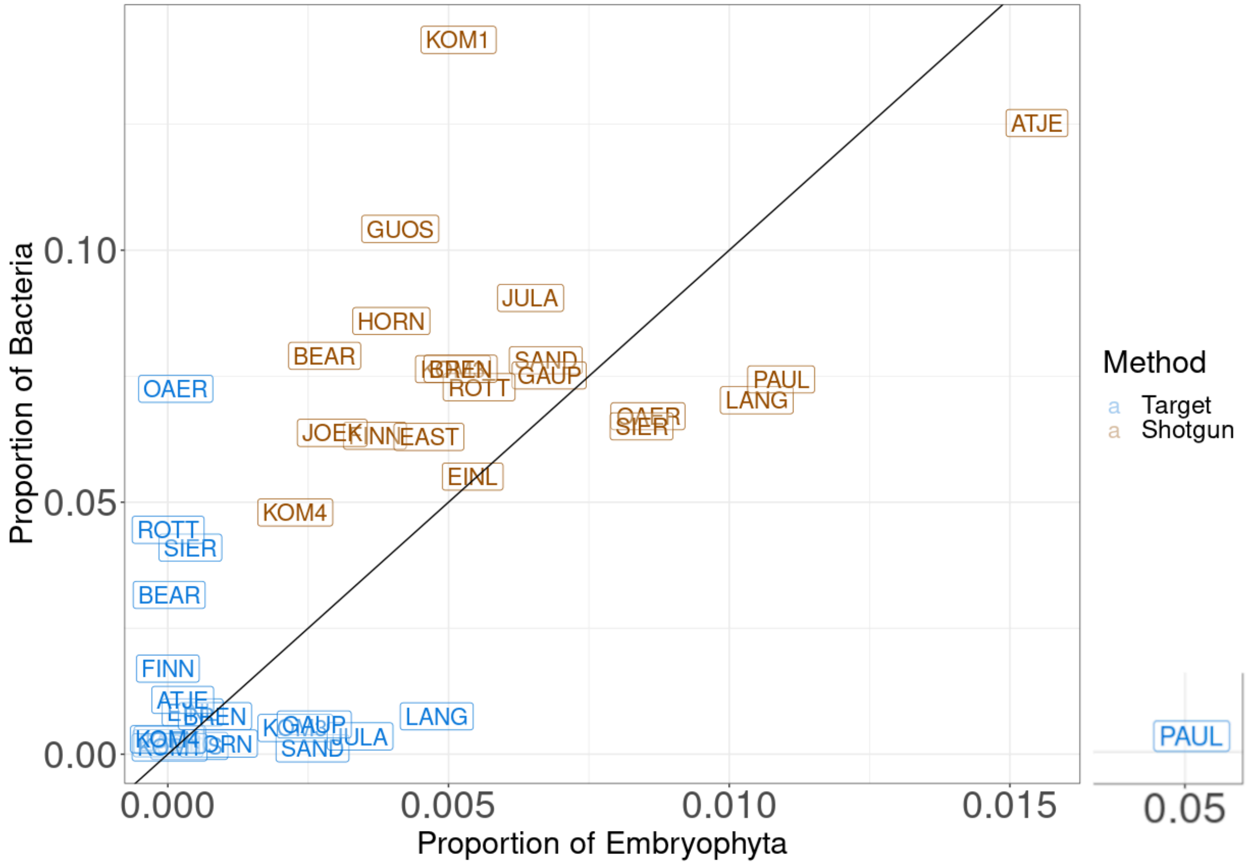


Figure 1. The proportion of annotated Bacteria and Embryophyta reads from each lake for both the capture and shotgun datasets. Note the difference in y- and x-axis scales and that the diagonal line is y = 10x. Also note that target capture from lake Paulanjävri (PAUL) is a distant outlier with a proportion of 0.05 Embryophyta reads and 0.003 Bacteria reads. For lake names, see Supplementary Table 1.

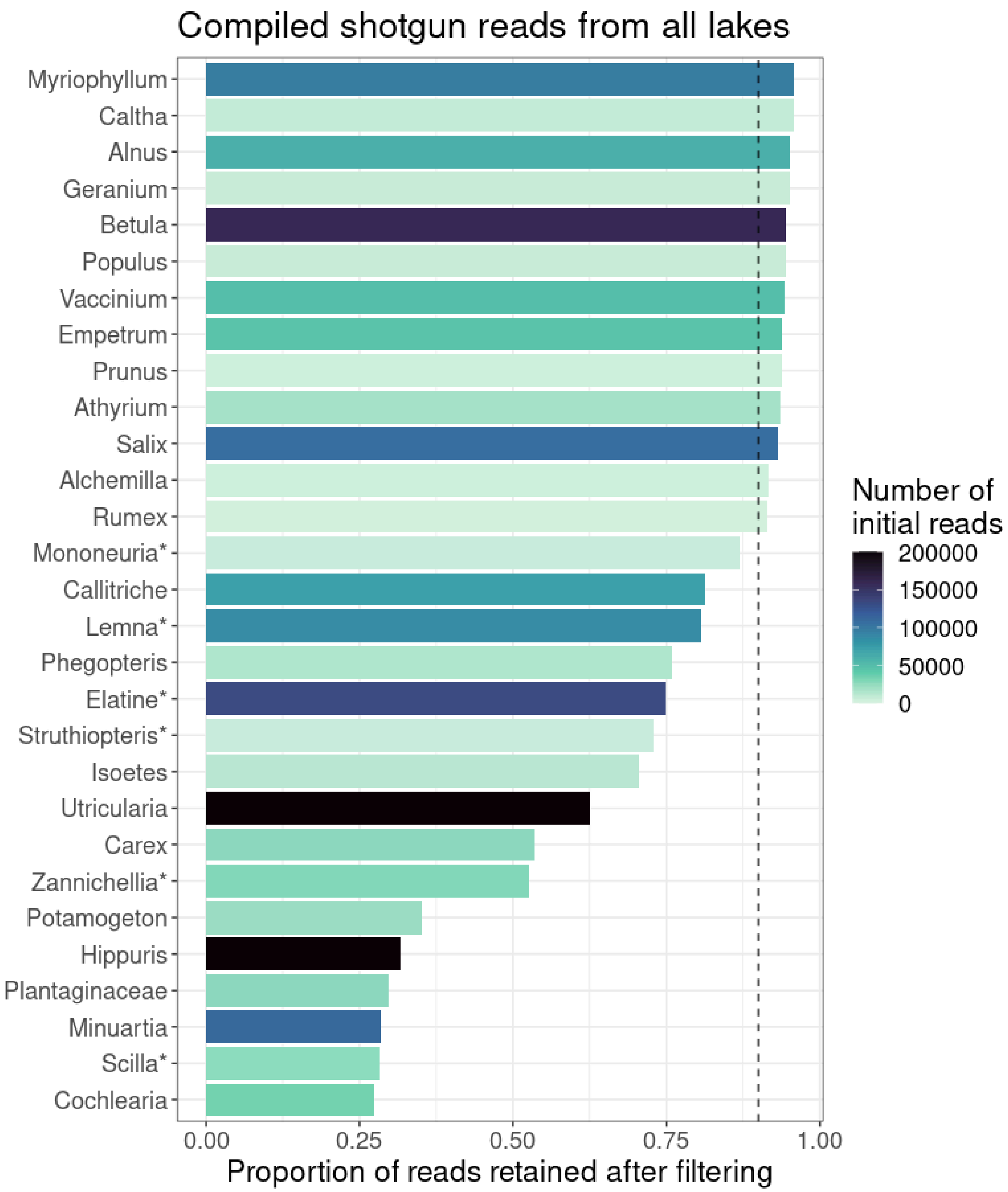


Figure 2. The proportion of shotgun reads retained for each taxon across the 20 lakes when filtering for bacteria, algae, and fungal reads. The dashed line denotes the 90% retained threshold where all taxa are supported by metabarcoding and local vegetation surveys. Note that *Hippuris* has an initial read count of 1.78 million, an order of magnitude larger than the second most abundant taxon, *Utricularia*, with 269,783 reads. Taxa not recorded (e.g. *Mononeuria*) or rare (e.g. *Elatine*) in N Norway are denoted by an asterisk.

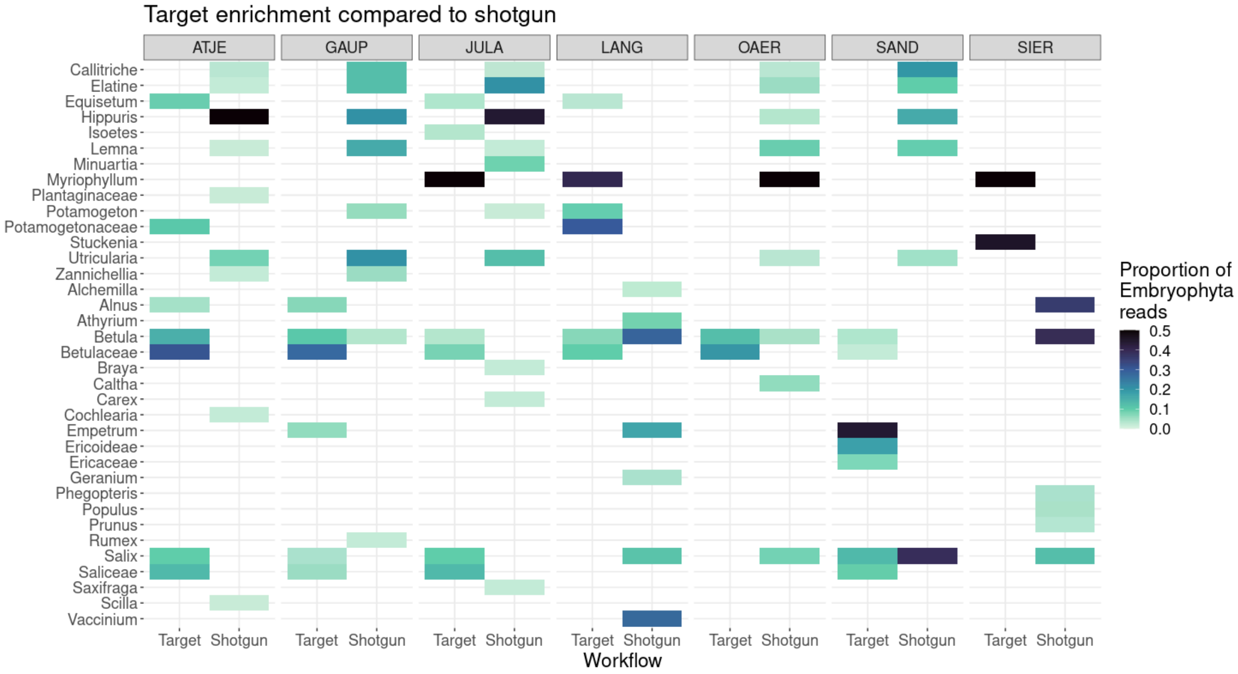


Figure 3. A heatmap of the taxa detected by target enrichment and shotgun sequencing of seven lakes. The color of the cells corresponds to the proportion of that taxon's read count to the sample’s overall Embryophyta read count. Taxa are sorted alphabetically by aquatic taxa and then alphabetically by terrestrial taxa. For abbreviations of lake name, see Supplementary Table 1.

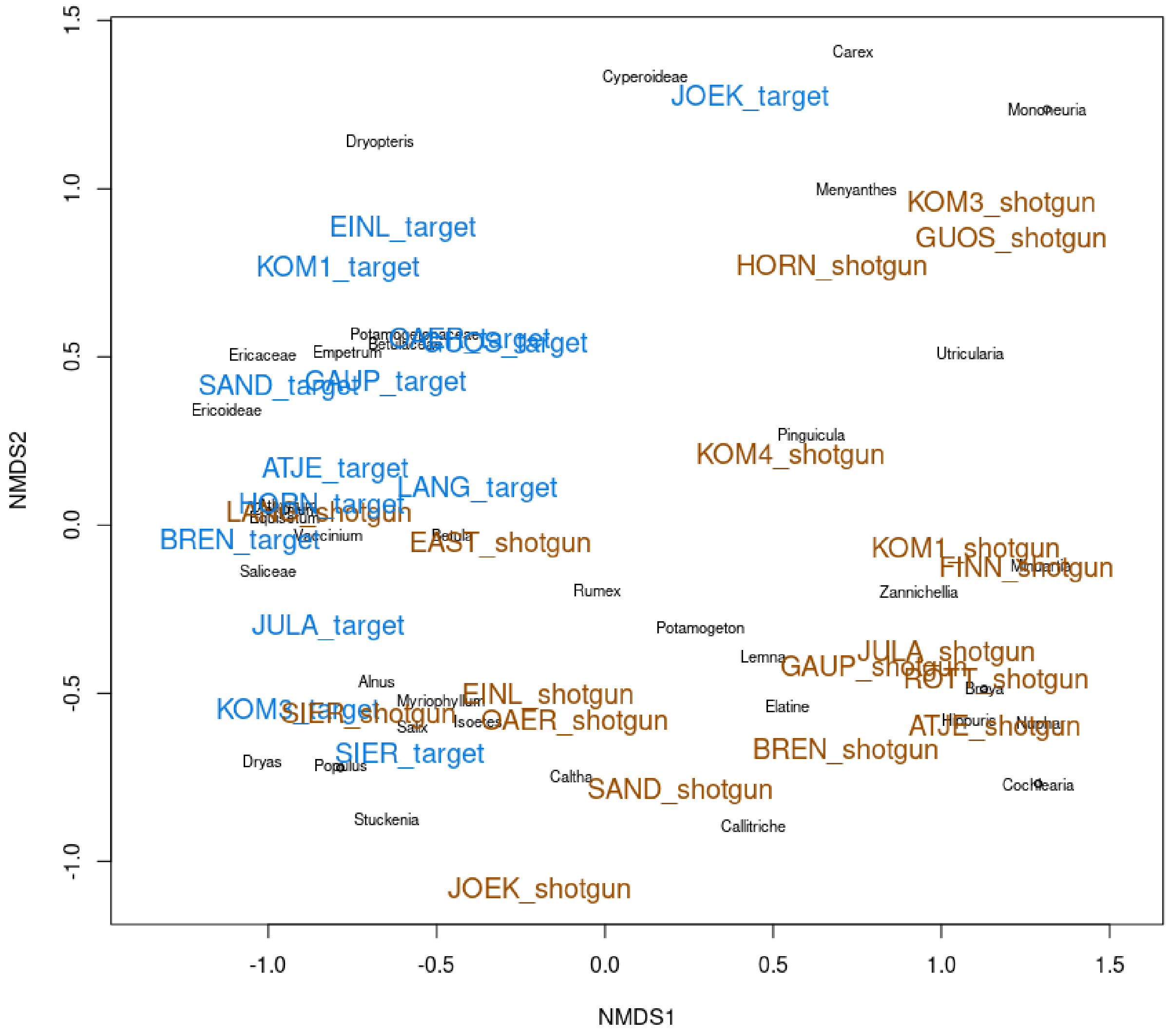
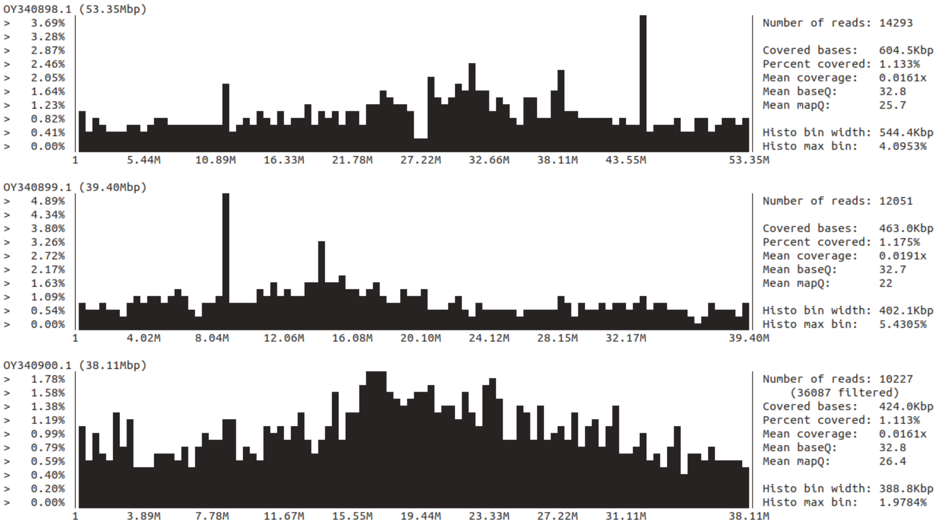


Figure 4. Nonmetric multidimensional scaling (NMDS) representing the vegetation communities detected at each lake by target enrichment and shotgun sequencing. Note that the five lakes without any target enrichment reads are not included, as well as PAUL which is a distant outlier to all other lakes.

A. *Alnus* reads from 20 lakes study

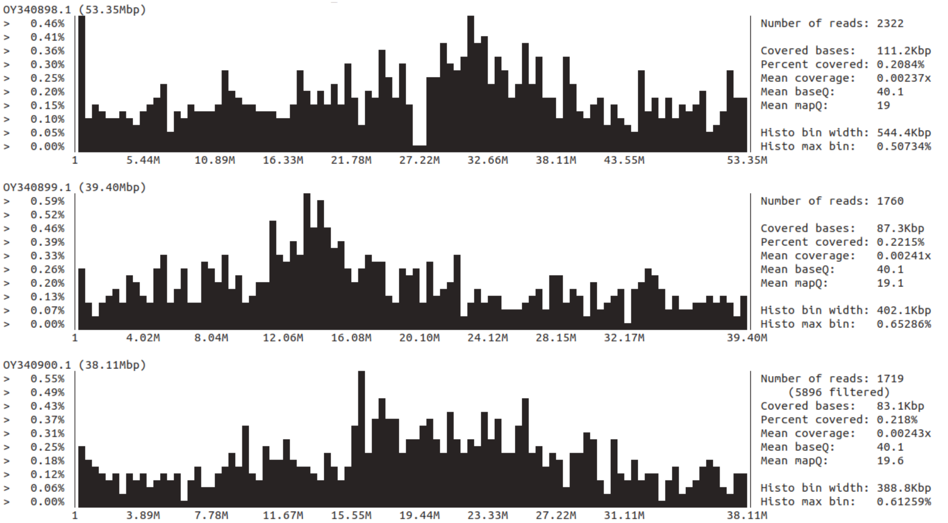


```
OY340898.1 (53.35Mbp)
>   3.69%          Number of reads: 14293
>   3.28%
>   2.87%          Covered bases:   604.5Kbp
>   2.46%          Percent covered: 1.133%
>   2.05%          Mean coverage:   0.0161x
>   1.64%          Mean baseQ:      32.8
>   1.23%          Mean mapQ:       25.7
>   0.82%
>   0.41%          Histo bin width: 544.4Kbp
>   0.00%          Histo max bin:   4.0953%
          1     5.44M   10.89M   16.33M   21.78M   27.22M   32.66M   38.11M   43.55M         53.35M

OY340899.1 (39.40Mbp)
>   4.89%          Number of reads: 12051
>   4.34%
>   3.80%          Covered bases:   463.0Kbp
>   3.26%          Percent covered: 1.175%
>   2.72%          Mean coverage:   0.0191x
>   2.17%          Mean baseQ:      32.7
>   1.63%          Mean mapQ:       22
>   1.09%
>   0.54%          Histo bin width: 402.1Kbp
>   0.00%          Histo max bin:   5.4305%
          1     4.02M    8.04M   12.06M   16.08M   20.10M   24.12M   28.15M   32.17M         39.40M

OY340900.1 (38.11Mbp)
>   1.78%          Number of reads: 10227
>   1.58%              (36087 filtered)
>   1.38%          Covered bases:   424.0Kbp
>   1.19%          Percent covered: 1.113%
>   0.99%          Mean coverage:   0.0161x
>   0.79%          Mean baseQ:      32.8
>   0.59%          Mean mapQ:       26.4
>   0.40%
>   0.20%          Histo bin width: 388.8Kbp
>   0.00%          Histo max bin:   1.9784%
          1     3.89M    7.78M   11.67M   15.55M   19.44M   23.33M   27.22M   31.11M         38.11M
```

B. *Alnus* reads from Wang et al. 2021

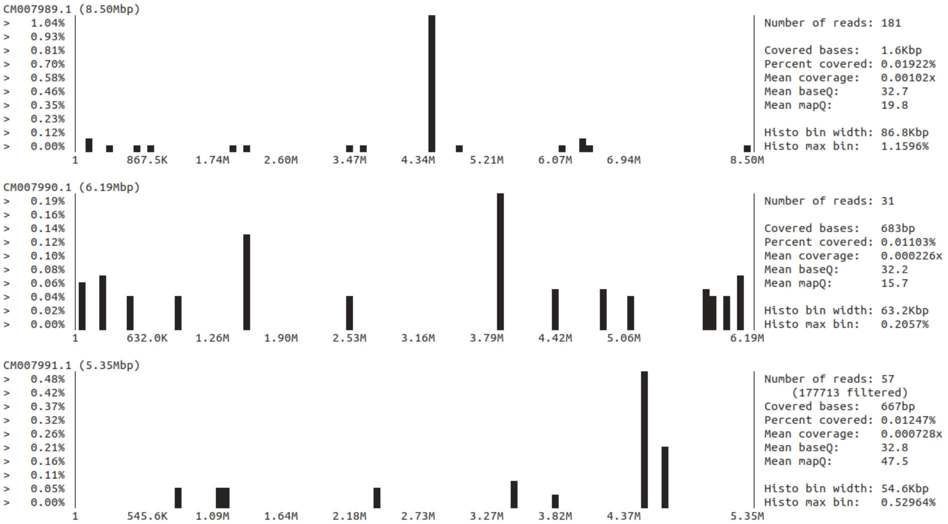


```
OY340898.1 (53.35Mbp)
>   0.46%          Number of reads: 2322
>   0.41%
>   0.36%          Covered bases:   111.2Kbp
>   0.30%          Percent covered: 0.2084%
>   0.25%          Mean coverage:   0.00237x
>   0.20%          Mean baseQ:      40.1
>   0.15%          Mean mapQ:       19
>   0.10%
>   0.05%          Histo bin width: 544.4Kbp
>   0.00%          Histo max bin:   0.50734%
          1     5.44M   10.89M   16.33M   21.78M   27.22M   32.66M   38.11M   43.55M         53.35M

OY340899.1 (39.40Mbp)
>   0.59%          Number of reads: 1760
>   0.52%
>   0.46%          Covered bases:   87.3Kbp
>   0.39%          Percent covered: 0.2215%
>   0.33%          Mean coverage:   0.00241x
>   0.26%          Mean baseQ:      40.1
>   0.20%          Mean mapQ:       19.1
>   0.13%
>   0.07%          Histo bin width: 402.1Kbp
>   0.00%          Histo max bin:   0.65286%
          1     4.02M    8.04M   12.06M   16.08M   20.10M   24.12M   28.15M   32.17M         39.40M

OY340900.1 (38.11Mbp)
>   0.55%          Number of reads: 1719
>   0.49%              (5896 filtered)
>   0.43%          Covered bases:   83.1Kbp
>   0.37%          Percent covered: 0.218%
>   0.31%          Mean coverage:   0.00243x
>   0.25%          Mean baseQ:      40.1
>   0.18%          Mean mapQ:       19.6
>   0.12%
>   0.06%          Histo bin width: 388.8Kbp
>   0.00%          Histo max bin:   0.61259%
          1     3.89M    7.78M   11.67M   15.55M   19.44M   23.33M   27.22M   31.11M         38.11M
```

C. *Utricularia* reads from 20 lakes study



```
CM007989.1 (8.50Mbp)
>   1.04%          Number of reads: 181
>   0.93%
>   0.81%          Covered bases:   1.6Kbp
>   0.70%          Percent covered: 0.01922%
>   0.58%          Mean coverage:   0.00102x
>   0.46%          Mean baseQ:      32.7
>   0.35%          Mean mapQ:       19.8
>   0.23%
>   0.12%          Histo bin width: 86.8Kbp
>   0.00%          Histo max bin:   1.1596%
          1    867.5K    1.74M    2.60M    3.47M    4.34M    5.21M    6.07M    6.94M          8.50M

CM007990.1 (6.19Mbp)
>   0.19%          Number of reads: 31
>   0.16%
>   0.14%          Covered bases:   683bp
>   0.12%          Percent covered: 0.01103%
>   0.10%          Mean coverage:   0.000226x
>   0.08%          Mean baseQ:      32.2
>   0.06%          Mean mapQ:       15.7
>   0.04%
>   0.02%          Histo bin width: 63.2Kbp
>   0.00%          Histo max bin:   0.2057%
          1    632.0K    1.26M    1.90M    2.53M    3.16M    3.79M    4.42M    5.06M          6.19M

CM007991.1 (5.35Mbp)
>   0.48%          Number of reads: 57
>   0.42%              (177713 filtered)
>   0.37%          Covered bases:   667bp
>   0.32%          Percent covered: 0.01247%
>   0.26%          Mean coverage:   0.000728x
>   0.21%          Mean baseQ:      32.8
>   0.16%          Mean mapQ:       47.5
>   0.11%
>   0.05%          Histo bin width: 54.6Kbp
>   0.00%          Histo max bin:   0.52964%
          1    545.6K    1.09M    1.64M    2.18M    2.73M    3.27M    3.82M    4.37M          5.35M
```

Figure 5. A genomic coverage histogram of *Alnus* reads from **A.** the 20 lakes study (36,571 reads mapping, 50.3%) and **B.** the Wang et al. 2021 study (5,801 reads mapping, 49.6%) to the three largest chromosomes of *Alnus glutinosa*. **C.** Genomic coverage histogram of *Utricualria* reads from the 20 lakes study (269 reads mapping, 0.15%) mapped to the three largest chromosomes of *Utricularia gibba*.

Supplementary

Supplementary Table 1. Overview of sites and sequencing results.

unique_name	lake_name	shotgun_name	capture_name	shotgun_reads_postfilter	capture_reads_postfilter	shotgun_reads_embryophyta	capture_reads_embryophyta
ATJE	A-tjern	JK139L-3	TrE-5	36690851	983540	568026	261
BEAR	Bearalveaijohka	LDE001L-4	TrE-18	46263497	41325	129263	1
BREN	Brennskogtjørna	JK139L-5	TrE-4	37864625	887832	197452	746
EAST	Eastorjavri	LDE001L-6	TrE-13	25638198	717392	119254	3
EINL	Einletvatnet	LDE002L-1	TrE-1	30032301	918254	163105	406
FINN	Finnvatnet	LDE002L-5	TrE-10	22566594	686860	83260	0
GAUP	Gauptjern	JK139L-6	TrE-6	33499268	469257	227553	1224
GUOS	Guossajavri	LDE002L-6	TrE-11	35675555	740850	147685	284
HORN	Horntjernet	LDE001L-3	TrE-17	26603757	592222	105918	538
JOEK	Jøkelvatnet	LDE001L-2	TrE-19	35512972	1120458	104045	25
JULA	Jula Javri	LDE002L-2	TrE-3	51426283	571526	332326	1957
KOM1	Pond 1 (Kommagdalen)	LDE001L-7	TrE-20	49615902	1203007	256997	38
KOM3	Pond 3 (Kommagdalen)	LDE001L-1	TrE-21	37905044	1415546	192572	3220
KOM4	Pond 4 (Kommagdalen)	JK139L-2	TrE-22	37666350	618689	85513	2
LANG	Langfjordvannet	LDE002L-7	TrE-9	30893740	2187357	324133	10479
NESS	Nesservatnet	LDE001L-5	TrE-16	0	1627	0	0
NORD	Nordvivatnet	NA	TrE-12	NA	881254	NA	476
OAER	Øvre æråsvatnet	LDE002L-3	TrE-7	26222103	1088028	224253	154
PAUL	Paulan Javri	JK139L-7	TrE-2	36006915	1073262	393375	54023
ROTT	Rottjern	LDE002L-4	TrE-8	24978954	658720	138508	4
SAND	Sandfjorddalen	JK139L-4	TrE-15	38404600	1115874	258585	2876
SIER	Sierravannet	JK139L-1	TrE-14	36125375	106957	304887	43

capture_extraction_control_1	NA	NA	TrE-23	NA	5130	NA	891
capture_extraction_control_2	NA	NA	TrE-24	NA	4086	NA	31
capture_library blank_control_1	NA	NA	TrE-25	NA	0	NA	0
shotgun_extraction_control_1	NA	JK139L-8	NA	10778	NA	73	NA
shotgun_extraction_control_2	NA	LDE001L-8	NA	348	NA	5	NA
shotgun_extraction_control_3	NA	LDE002L-11	NA	3977	NA	142	NA
shotgun_library blank_control_1	NA	LDE002L-12	NA	572	NA	7	NA

Supplementary Table 2. Overview of the taxa detected by target enrichment and shotgun sequencing and how many lakes they were detected in. Workflow specific taxa with an asterisk have an equivalent taxon identified by the other workflow (e.g. *Populus* is only detected by shotgun sequencing, but Salicaceae is detected by target enrichment).

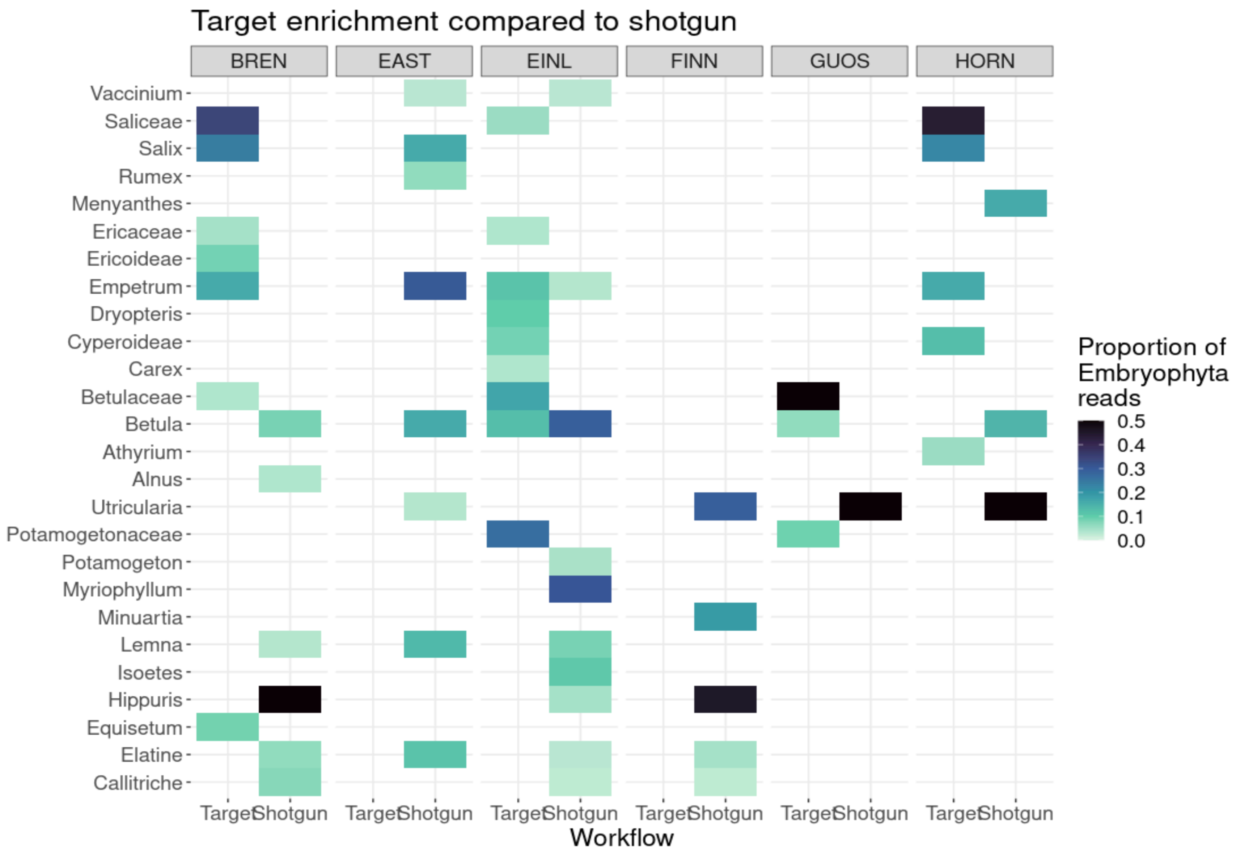
Family	**Shotgun taxa**	**Target capture**	**Number of shotgun lake detections**	**Number of target lake detections**	**Number of common lake detections**
Athyriaceae	Athyrium	Athyrium	1	1	0
Betulaceae	Alnus	Alnus, Betulceae	2	10	1
Betulaceae	Betula	Betula, Betulaceae	8	10	5
Blechnaceae	Struthiopteris		1	0	0
Brassicaceae	Braya		1	0	0
Brassicaceae	Cochlearia		1	0	0
Caryophyllaceae	Minuartia		1	0	0
Caryophyllaceae	Mononeuria		1	0	0
Cyperaceae	Carex	Carex, Cyperoideae	2	3	0
Dryopteridaceae		Dryopteris	0	1	0
Elatinaceae	Elatine		14	0	0
Equisetaceae		Equisetum	0	4	0
Ericaceae	Empetrum	Empetrum, Ericoideae, Ericaceae	4	7	1
Ericaceae	Vaccinium	Ericaceae	3	3	1
Geraniaceae	Geranium		1	0	0
Haloragaceae	Myriophyllum	Myriophyllum	5	0	0
Hyacinthaceae	Scilla		1	0	0
Isoetaceae	Isoetes	Isoetes	1	1	0
Lemnoideae	Lemna		12	0	0
Lentibulariaceae	Pinguicula		1	0	0
Lentibulariaceae	Utricularia		14	0	0
Menyanthaceae	Menyanthes		1	0	0
Nymphaeaceae	Nuphar		1	0	0
Salicaceae	Salix	Salix, Salicaceae	7	9	1
Salicaceae	Populus	Salicaceae	1	9	0
Potamogetonaceae	Potamogeton	Potamogeton, Potamogetonaceae	5	4	0

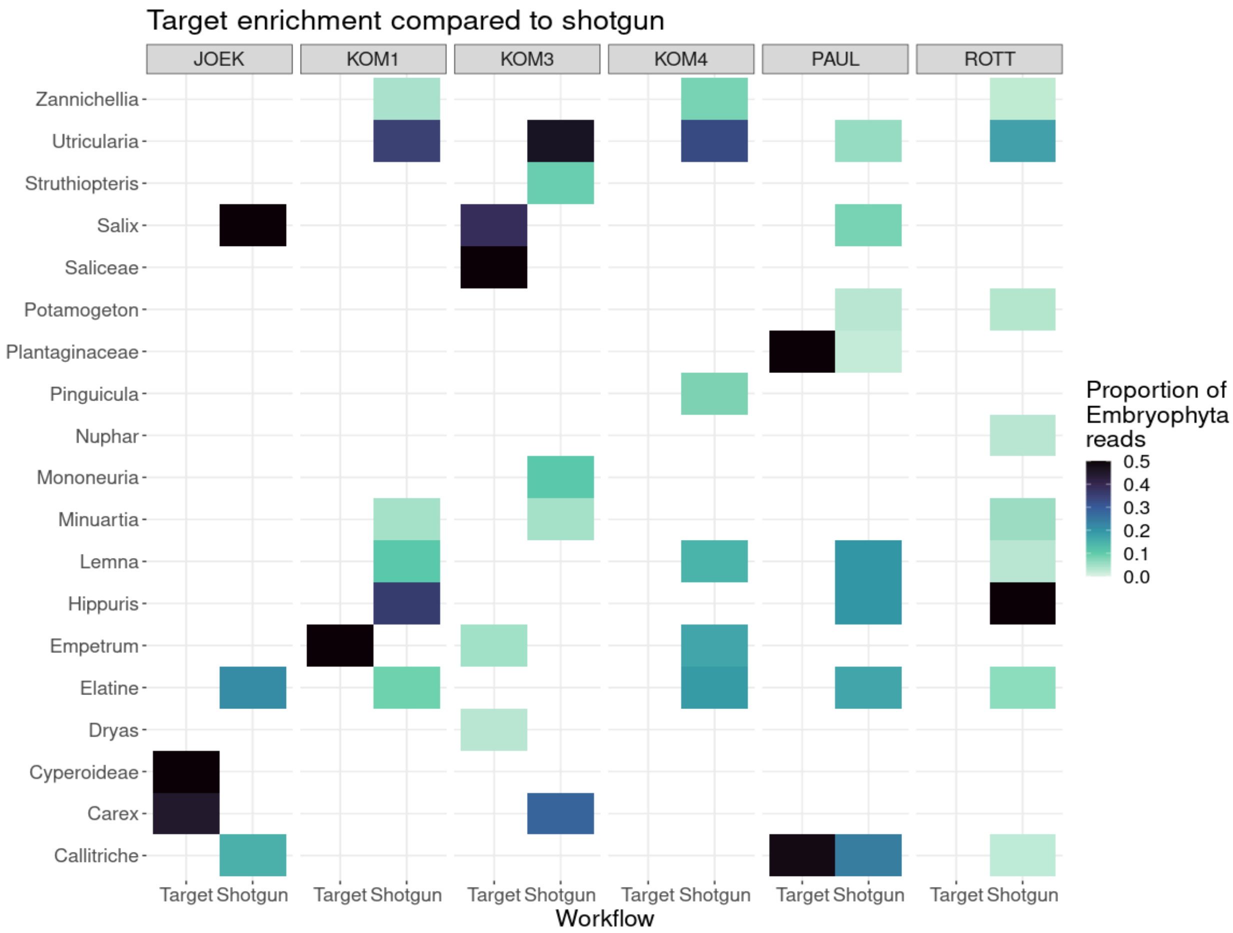
Potamogetonaceae		Stuckenia, Potamogetonaceae	0	5	0
Potamogetonaceae	Zannichellia	Potamogetonaceae	5	4	1
Plantaginaceae	Hippuris, Plantaginaceae	Plantaginaceae	11	1	1
Plantaginaceae	Callitriche, Plantaginaceae	Callitriche, Plantaginaceae	11	1	1
Polygonaceae	Rumex		2	0	0
Ranunculaceae	Caltha		1	0	0
Rosaceae	Alchemilla		1	0	0
Rosaceae		Dryas	0	1	0
Rosaceae	Prunus		1	0	0
Saxifragaceae	Saxifraga		1	0	0
Thelypteridaceae	Phegopteris		1	0	0

Supplementary Table 2. Name of samples from Wang et al. 2021 where reads of the listed genus were subset from.

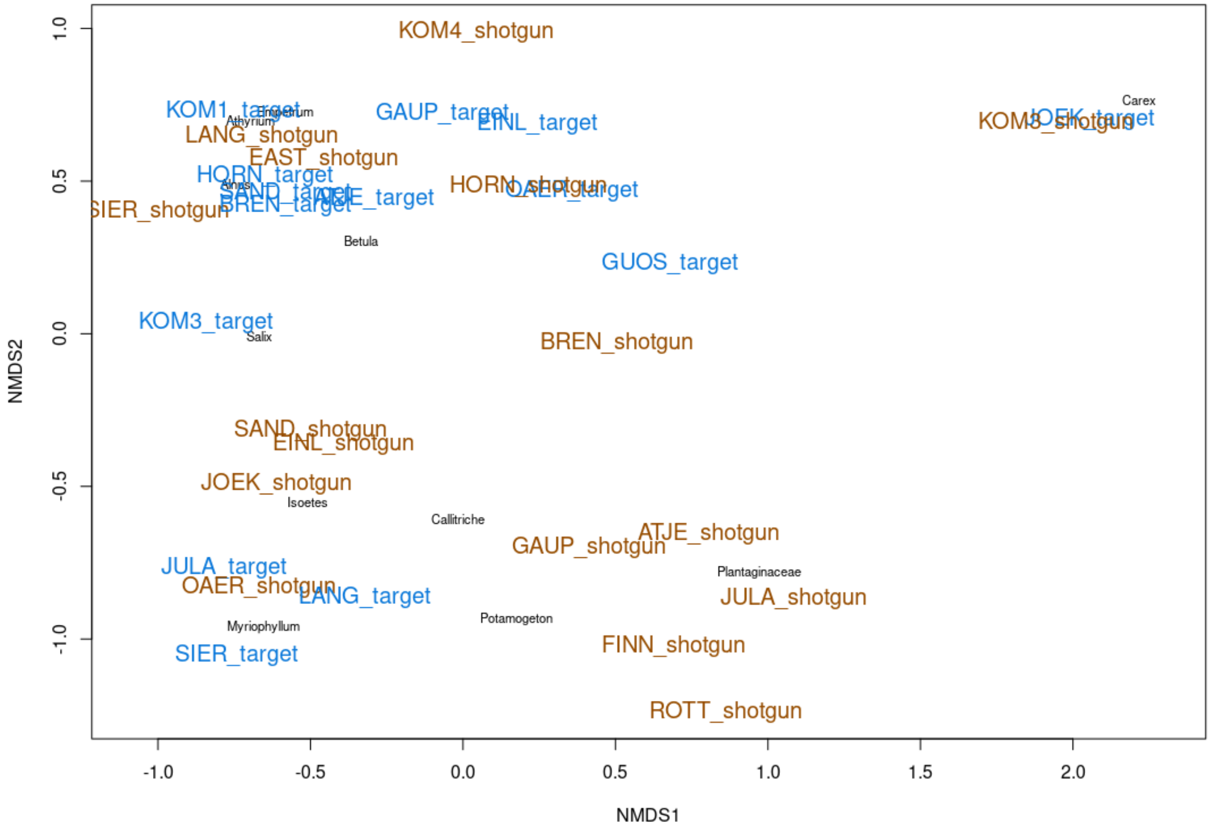
sample_id	**genus**
ar6_30	*Utricularia*
cr1_1	*Utricularia*
cr1_29	*Utricularia*
cr8_27	*Utricularia*
cr9_16	*Utricularia*
cr9_1	*Utricularia*
cr9_3	*Utricularia*
cr9_4	*Utricularia*
cr9_5	*Utricularia*
cr9_6	*Utricularia*
ar6_18	*Alnus*
cr2_23	*Alnus*

cr4_2	*Alnus*
cr4_4	*Alnus*
cr8_1	*Alnus*
cr8_2	*Alnus*
cr8_3	*Alnus*
tm4_1	*Alnus*
tm4_3	*Alnus*
tm6_8	*Alnus*

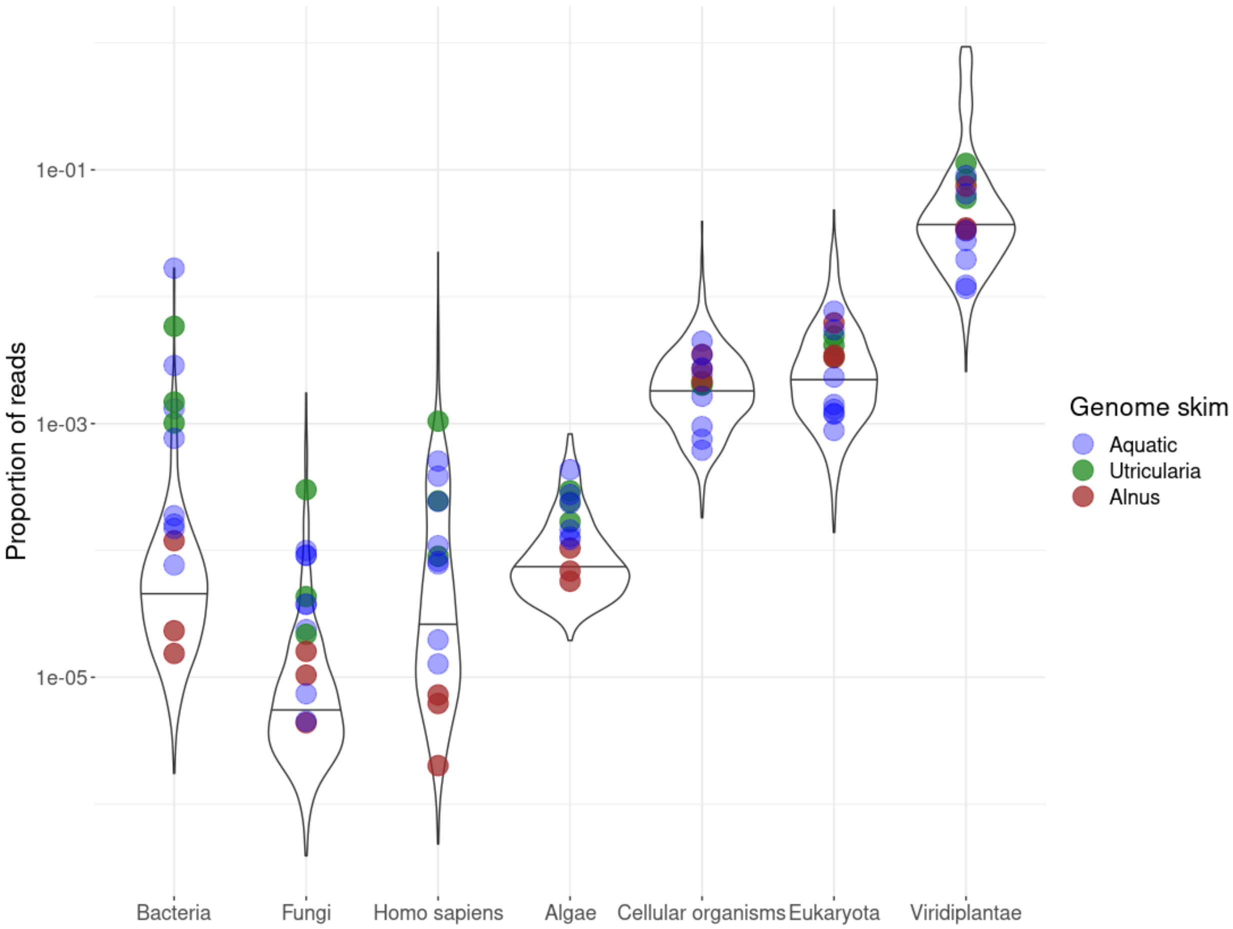

Target enrichment compared to shotgun
BREN
EAST
EINL
FINN
GUOS
HORN
Vaccinium
Saliceae
Salix
Rumex
Menyanthes
Ericaceae
Ericoideae
Empetrum
Dryopteris
Cyperoideae
Carex
Betulaceae
Betula
Athyrium
Alnus
Utricularia
Potamogetonaceae
Potamogeton
Myriophyllum
Minuartia
Lemna
Isoetes
Hippuris
Equisetum
Elatine
Callitriche
Target
Shotgun
Workflow
Proportion of Embryophyta reads
0.5
0.4
0.3
0.2
0.1
0.0



Supplementary Figure 1. Heatmaps for the remaining lakes included in the study. The color of the cells corresponds to the proportion of that taxon's read count to the sample's overall Embryophyta read count. Taxa are sorted alphabetically by terrestrial taxa and then alphabetically by aquatic taxa.



Supplementary Figure 2. Nonmetric multidimensional scaling (NMDS) representing the vegetation communities detected at each lake by target enrichment and shotgun sequencing. Only taxa detected in at least one lake by both methods are included in this ordination.



Supplementary Figure 3. The proportions of contaminants detected in the initial genome skims from PhyloNorway. Identified bacteria, fungi, human, and algae reads were removed from the skims. The colored points indicate skims of *Utricularia*, *Alnus*, and aquatic taxa appearing in the 20 lakes dataset. Figure reproduced and edited from (Elliott et al. 2024).

References

Alsos, Inger Greve, Youri Lammers, Nigel Giles Yoccoz, Tina Jørgensen, Per Sjögren, Ludovic Gielly, and Mary E. Edwards. 2018. "Plant DNA Metabarcoding of Lake Sediments: How Does It Represent the Contemporary Vegetation." *PloS One* 13 (4): e0195403.

Alsos, Inger Greve, Sebastien Lavergne, Marie Kristine Føreid Merkel, Marti Boleda, Youri Lammers, Adriana Alberti, Charles Pouchon, et al. 2020. "The Treasure Vault Can Be Opened: Large-Scale Genome Skimming Works Well Using Herbarium and Silica Gel Dried Material." *Plants* 9 (4). https://doi.org/10.3390/plants9040432.

Alsos, Inger G., Per Sjögren, Antony G. Brown, Ludovic Gielly, Marie Kristine Føreid Merkel, Aage Paus, Youri Lammers, et al. 2020. "Last Glacial Maximum Environmental Conditions at Andøya, Northern Norway; Evidence for a Northern Ice-Edge Ecological 'hotspot.'" *Quaternary Science Reviews* 239 (July): 106364.

Andrews, Simon. 2010. *FastQC: A Quality Control Analysis Tool for High Throughput Sequencing Data*. Github. https://github.com/s-andrews/FastQC.

Capo, Eric, Charline Giguet-Covex, Alexandra Rouillard, Kevin Nota, Peter D. Heintzman, Aurèle Vuillemin, Daniel Ariztegui, et al. 2021. "Lake Sedimentary DNA Research on Past Terrestrial and Aquatic Biodiversity: Overview and Recommendations." *Quaternary* 4 (1): 6.

Chen, Wentao, and Gentile Francesco Ficetola. 2020. "Numerical Methods for sedimentary‐ancient‐DNA‐based Study on Past Biodiversity and Ecosystem Functioning." *Environmental DNA (Hoboken, N.J.)* 2 (2): 115–29.

Dabney, Jesse, Matthias Meyer, and Svante Pääbo. 2013. "Ancient DNA Damage." *Cold Spring Harbor Perspectives in Biology* 5 (7). https://doi.org/10.1101/cshperspect.a012567.

Dalén, Love, Peter D. Heintzman, Joshua D. Kapp, and Beth Shapiro. 2023. "Deep-Time Paleogenomics and the Limits of DNA Survival." *Science* 382 (6666): 48–53.

Freeman, C. L., L. Dieudonné, O. B. A. Agbaje, M. Žure, J. Q. Sanz, M. Collins, and K. K. Sand. 2023. "Survival of Environmental DNA in Sediments: Mineralogic Control on DNA Taphonomy." *Environmental DNA (Hoboken, N.J.)* 5 (6): 1691–1705.

Giguet-Covex, Charline, Stanislav Jelavić, Anthony Foucher, Marina A. Morlock, Susanna A. Wood, Femke Augustijns, Isabelle Domaizon, Ludovic Gielly, and Eric Capo. 2023. "The Sources and Fates of Lake Sedimentary DNA." In *Tracking Environmental Change Using Lake Sediments: Volume 6: Sedimentary DNA*, edited by Eric Capo, Cécilia Barouillet, and John P. Smol, 9–52. Cham: Springer International Publishing.

Heintzman, Peter D., Kevin Nota, Alexandra Rouillard, Youri Lammers, Tyler J. Murchie, Linda Armbrecht, Sandra Garcés-Pastor, and Benjamin Vernot. 2023. "The Sedimentary Ancient DNA Workflow." In *Tracking Environmental Change Using Lake Sediments: Volume 6: Sedimentary DNA*, edited by Eric Capo, Cécilia Barouillet, and John P. Smol, 53–84. Cham: Springer International Publishing.

Kjær, Kurt H., Mikkel Winther Pedersen, Bianca De Sanctis, Binia De Cahsan, Thorfinn S. Korneliussen, Christian S. Michelsen, Karina K. Sand, et al. 2022. "A 2-Million-Year-Old Ecosystem in Greenland Uncovered by Environmental DNA." *Nature* 612 (7939): 283–91.

Lammers, Youri, Peter D. Heintzman, and Inger Greve Alsos. 2021. "Environmental Palaeogenomic Reconstruction of an Ice Age Algal Population." *Communications Biology* 4 (1): 220.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Mamanova, Lira, Alison J. Coffey, Carol E. Scott, Iwanka Kozarewa, Emily H. Turner, Akash Kumar, Eleanor Howard, Jay Shendure, and Daniel J. Turner. 2010. "Target-Enrichment Strategies for next-Generation Sequencing." *Nature Methods* 7 (2): 111–18.

Maricic, Tomislav, Mark Whitten, and Svante Pääbo. 2010. "Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products." *PloS One* 5 (11): e14004.

Murchie, Tyler J., Melanie Kuch, Ana T. Duggan, Marissa L. Ledger, Kévin Roche, Jennifer Klunk, Emil Karpinski, et al. 2021. "Optimizing Extraction and Targeted Capture of Ancient Environmental DNA for Reconstructing Past Environments Using the PalaeoChip Arctic-1.0 Bait-Set." *Quaternary Research* 99 (January): 305–28.

Nichols, Ruth V., Christopher Vollmers, Lee A. Newsom, Yue Wang, Peter D. Heintzman, Mckenna Leighton, Richard E. Green, and Beth Shapiro. 2018. "Minimizing Polymerase Biases in Metabarcoding." *Molecular Ecology Resources*, May. https://doi.org/10.1111/1755-0998.12895.

Oksanen, J., F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan

McGlinn, Peter R. Minchin, et al. 2020. “Vegan: Community Ecology Package. R Package Version 2.5-6. 2019.”

Parducci, L., I. G. Alsos, and P. Unneberg. 2019. “Shotgun Environmental DNA, Pollen, and Macrofossil Analysis of Lateglacial Lake Sediments from Southern Sweden.” *Frontiers in Ecology and the Environment*. https://www.frontiersin.org/articles/10.3389/fevo.2019.00189/full.

Pedersen, Mikkel Winther, Bianca De Sanctis, Nedda F. Saremi, Martin Sikora, Emily E. Puckett, Zhenquan Gu, Katherine L. Moon, et al. 2021. “Environmental Genomics of Late Pleistocene Black Bears and Giant Short-Faced Bears.” *Current Biology: CB* 31 (12): 2728–36.e8.

Pedersen, Mikkel W., Anthony Ruter, Charles Schweger, Harvey Friebe, Richard A. Staff, Kristian K. Kjeldsen, Marie L. Z. Mendoza, et al. 2016. “Postglacial Viability and Colonization in North America’s Ice-Free Corridor.” *Nature* 537 (7618): 45–49.

Schulte, Luise, Nadine Bernhardt, Kathleen Stoof-Leichsenring, Heike H. Zimmermann, Luidmila A. Pestryakova, Laura S. Epp, and Ulrike Herzschuh. 2021. “Hybridization Capture of Larch (Larix Mill.) Chloroplast Genomes from Sedimentary Ancient DNA Reveals Past Changes of Siberian Forest.” *Molecular Ecology Resources* 21 (3): 801–15.

Taberlet, P., A. Bonin, L. Zinger, and E. Coissac. 2018. “Environmental DNA: For Biodiversity Research and Monitoring.” https://books.google.ca/books?hl=en&lr=&id=1e9IDwAAQBAJ&oi=fnd&pg=PP1&ots=UY8TqcnfoR&sig=8kUMa4OFtZC1Z2n5fT7MV7cTuSo.

Taberlet, P., E. Coissac, F. Pompanon, L. Gielly, C. Miquel, A. Valentini, T. Vermat, G. Corthier, C. Brochmann, and E. Willerslev. 2007. “Power and Limitations of the Chloroplast trnL (UAA) Intron for Plant DNA Barcoding.” *Nucleic Acids Research* 35 (3): e14–e14.

Vernot, Benjamin, Elena I. Zavala, Asier Gómez-Olivencia, Zenobia Jacobs, Viviane Slon, Fabrizio Mafessoni, Frédéric Romagné, et al. 2021. “Unearthing Neanderthal Population History Using Nuclear and Mitochondrial DNA from Cave Sediments.” *Science* 372 (6542). https://doi.org/10.1126/science.abf1667.

Wang, Yucheng, Thorfinn Sand Korneliussen, Luke E. Holman, Andrea Manica, and Mikkel Winther Pedersen. 2022. “ngsLCA—A Toolkit for Fast and Flexible Lowest Common Ancestor Inference and Taxonomic Profiling of Metagenomic Data.” *Methods in Ecology and Evolution / British Ecological Society* 13 (12): 2699–2708.