Environmental DNA

WILEY

Environmental DNA
Open Access
Dedicated to the study and use of environmental DNA for basic and applied sciences

ORIGINAL ARTICLE OPEN ACCESS

Capture Probe, Metabarcoding, or Shotgun Sequencing: Which Best Reflects Local Vegetation?

Nichola A. Strandberg[1] | Lucas Dane Elliott[1] | Dilli Prasad Rijal[1] | Dorothee Ehrich[2] | Youri Lammers[1] | Aloïs N. Revéret[1] | Nigel G. Yoccoz[2] | Iva Pitelkova[1] | Antony G. Brown[1] | Tyler J. Murchie[3,4] | Kathleen Stoof-Leichsenring[5] | Inger Greve Alsos[1]

[1]The Arctic University Museum of Norway, UiT, The Arctic University of Norway, Tromsø, Norway | [2]Department of Arctic and Marine Biology, UiT, The Arctic University of Norway, Tromsø, Norway | [3]McMaster Ancient DNA Centre, Department of Anthropology, McMaster University, Hamilton, Ontario, Canada | [4]Hakai Institute, Heriot Bay, British Columbia, Canada | [5]Polar Terrestrial Environmental Systems, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Potsdam, Germany

Correspondence: Lucas Dane Elliott (lucas.elliott@uit.no)

Received: 23 October 2025 | **Revised:** 3 March 2026 | **Accepted:** 17 March 2026

Keywords: capture enrichment | capture probe | metabarcoding | palaeoecology | plant DNA | sedaDNA | shotgun sequencing | target capture

ABSTRACT

Metabarcoding is the most widely applied method for studying plant communities using environmental DNA, with shotgun sequencing and capture probes being alternative methods that aim to retrieve multiple markers or genome-wide information. Any method's ability to detect and correctly identify plant taxa varies with DNA preservation, DNA reference library, and the diversity of the local flora, making it difficult to compare results from different environments. Here we compare these three methods using lake surface-sediments from Northern Fennoscandia with the PhyloNorway genome skim reference library (1500 taxa) that includes nearly all species of the regional flora. We also undertook vegetation surveys from around the lakes to estimate the true positive detection rate, identify false positive detections, and provide optimal filtering cut-off thresholds for the three methods. Applying these thresholds, the rate of false positives was too high for reliable identification at the species level based on shotgun (49%) and capture probes (62%), whereas it was low for metabarcoding (5%–12%). All methods were reliable at genus and family levels after applying the optimal filtering thresholds (<4% false positives). Our results show that in these lake sediments, metabarcoding on average detects 2.1 times as many true positive taxa as shotgun sequencing and 6.4 times as many taxa as capture probes. The proportion of a taxon's sequenced reads for the metabarcoding and shotgun methods was significantly related to the taxon's abundance category from the vegetation surveys, but this was not the case for capture probe data. We expect the false positive rate of shotgun sequencing to decrease with increasing genome completeness in the reference libraries and the method to be advantageous for highly degraded DNA with fragments too short for metabarcoding. At present, metabarcoding provides the highest detectability and taxonomic resolution for correct identification and quantification of vascular plants.

1 | Introduction

Sedimentary ancient DNA (*seda*DNA) analysis is a fast-growing approach in palaeoecology and has several benefits compared to traditional approaches, since it can detect more taxa than macrofossil analysis and pollen analysis (Garcés-Pastor et al. 2023), providing a powerful tool for assessing biodiversity dynamics over long, for example, millennial timescales (Capo et al. 2023). Currently, the most widely applied *seda*DNA method is metabarcoding (Alsos et al. 2024; Von Eggers et al. 2024), but

Nichola A. Strandberg, Lucas Dane Elliott share cofirst authorship.

alternatives exist in shotgun sequencing (Liu et al. 2024) and capture probes (Murchie, Kuch, et al. 2021; Murchie, Monteath, et al. 2021).

When metabarcoding, a single locus, for example, a taxonomic diagnostic fragment of the plant chloroplast DNA, is amplified through PCR (polymerase chain reaction) using specific primer pairs (Garcés-Pastor et al. 2023). PCRs can be repeated multiple times for each sample extract, thus improving the chances of detecting rare or degraded ancient DNA sequences (Ficetola et al. 2015). However, the PCR step may skew the relative abundances of DNA fragments (Garcés-Pastor et al. 2023). For example, for the commonly used universal vascular plant *trn*L g/h primers (Taberlet et al. 2007) some plant families like Poaceae and Cyperaceae generally show lower abundances than expected (Alsos et al. 2018), possibly due to GC content, fragment length, complexity, or primer mismatch (Nichols et al. 2018). Despite this, metabarcoding eDNA soil samples has been demonstrated to accurately reconstruct surrounding vegetation richness and relative abundance (Goodell et al. 2025). Although taxonomic resolution is limited in some families, for example, Salicaceae and Cyperaceae (Sønstebø et al. 2010), this is the method with overall highest taxonomic resolution (Revéret et al. 2023), with 40%–50% of taxa identified to species level (Garcés-Pastor et al. 2025; Julián-Posada et al. 2025; Rijal et al. 2021).

Shotgun sequencing, also known as metagenomics, is the non-targeted sequencing of extracted DNA and does not use primers, thus overcoming issues with PCR bias. Shotgun sequencing can take advantage of whole genome references to maximize information and has the added advantage of retaining deamination damage at the ends of DNA sequences, allowing the distinction between ancient and modern DNA (Dabney et al. 2013). At present, a major drawback of shotgun sequencing is that it cannot reliably identify taxa at the species level due to incomplete reference databases, particularly for taxonomic groups lacking complete reference genomes such as plants (Revéret et al. 2023). This can result in more false positives (FPs) since the typically short ancient DNA sequences may be matched to incorrect taxa within these incomplete reference databases (Elliott et al. 2025; Wang et al. 2021), and many false negatives (FNs) since whole genome reference libraries are currently incomplete. In addition, shotgun sequencing typically generates large amounts of data with a small fraction allocated to eukaryotes, a larger proportion assigned to prokaryotes, and many sequences which remain unidentified (Wang et al. 2021). Together, these limitations, high data volume with low target yield, and higher costs currently restrict the effectiveness and efficiency of shotgun sequencing for *seda*DNA studies compared to more targeted approaches.

Capture enrichment (also known as hybridisation capture, capture enrichment and target capture) uses a bait-set to bind targeted DNA sequences. The binding affinity to targets can be adjusted during bait design by adjusting bait tiling density, sequence identity and overlap through bait clustering, soft-masking, and taxonomic binning to improve specificity, and during wet lab processing by varying the hybridisation temperature, allowing closely related taxa, deaminated DNA fragments, and DNA fragments with individual variations to be retained (Capo et al. 2023). To date it is the least applied *seda*DNA method and has mainly been used to target a single genus (Meucci et al. 2021; Schulte et al. 2021). A number of probe sets exist that target species across many vascular plant families; the 353 probe set targeting nuclear Angiosperm genes (Johnson et al. 2019), the OZBaits_CP set, targeting 20 plastid gene regions (Foster et al. 2024) and the PaleoChip, targeting *mat*K, *rbc*L and *trn*L of the chloroplast (Murchie, Kuch, et al. 2021; Murchie, Monteath, et al. 2021). To our knowledge, only the latter probe set has been applied to ancient samples of several sites (Kjær et al. 2022; Murchie, Kuch, et al. 2021; Murchie, Monteath, et al. 2021), so it is difficult to assess the method's efficiency. Also, as most studies do not have independent data available that could be used to distinguish true from false positive identifications, the optimal filtering criteria for reliable identification is largely unknown.

Studying *seda*DNA records from Northern Fennoscandia allows for a unique comparison of the methods as the regional flora is well known and relatively small, with a total of ~600 species (Elven et al. 2022) and 250–550 species per 75 × 75 km grid (Grytnes et al. 1999). Almost all vascular plant species in the region are included in the genome skim reference library PhyloNorway, which on average covers 60% of the genome (Alsos et al. 2020; Wang et al. 2021). Here, we applied all three methods using established protocols to analyze lake surface sediment samples, and we validated these methods by comparing the results to field observations of plant biodiversity around the lakes. We first use the vegetation data to define filtering criteria that balance the loss of true positives versus retaining false positives. We then compare the three methods in taxonomic resolution and ability to detect local vegetation, taking into account the ranked abundance of each taxon.

2 | Materials and Methods

2.1 | Selection of Lakes

Lake catchments were selected covering the major vegetation types and climate gradients in northern Fennoscandia (Figure 1, Table S1). They represent the northernmost pine forest, widespread birch forest, as well as arctic-alpine heath. The lakes have small inflows and are undisturbed by human activities except for reindeer grazing and a conifer plantation within one catchment (Sierravannet).

2.2 | Vegetation Surveys

Vegetation surveys recorded every plant taxon observed within 2 m from the lakeshores. For the lakes included in Alsos et al. (2018), aquatic plants were surveyed from a boat using water binoculars and a long-handled rake. For the other lakes, the aquatic vegetation was surveyed by wading to collect and identify taxa. We noted plant taxa observed in the catchment > 2 m away from the lakeshore, although this extended vegetation survey was not comprehensive. All plant vouchers are at the herbarium at The Arctic University Museum of Norway (TROM). Both terrestrial and aquatic plant taxa identified were given a ranked abundance score 1–4 with 1 being rare (only a few ramets), 2 being scarce (ramets occur throughout but at low abundance), 3 being common (common throughout but not the

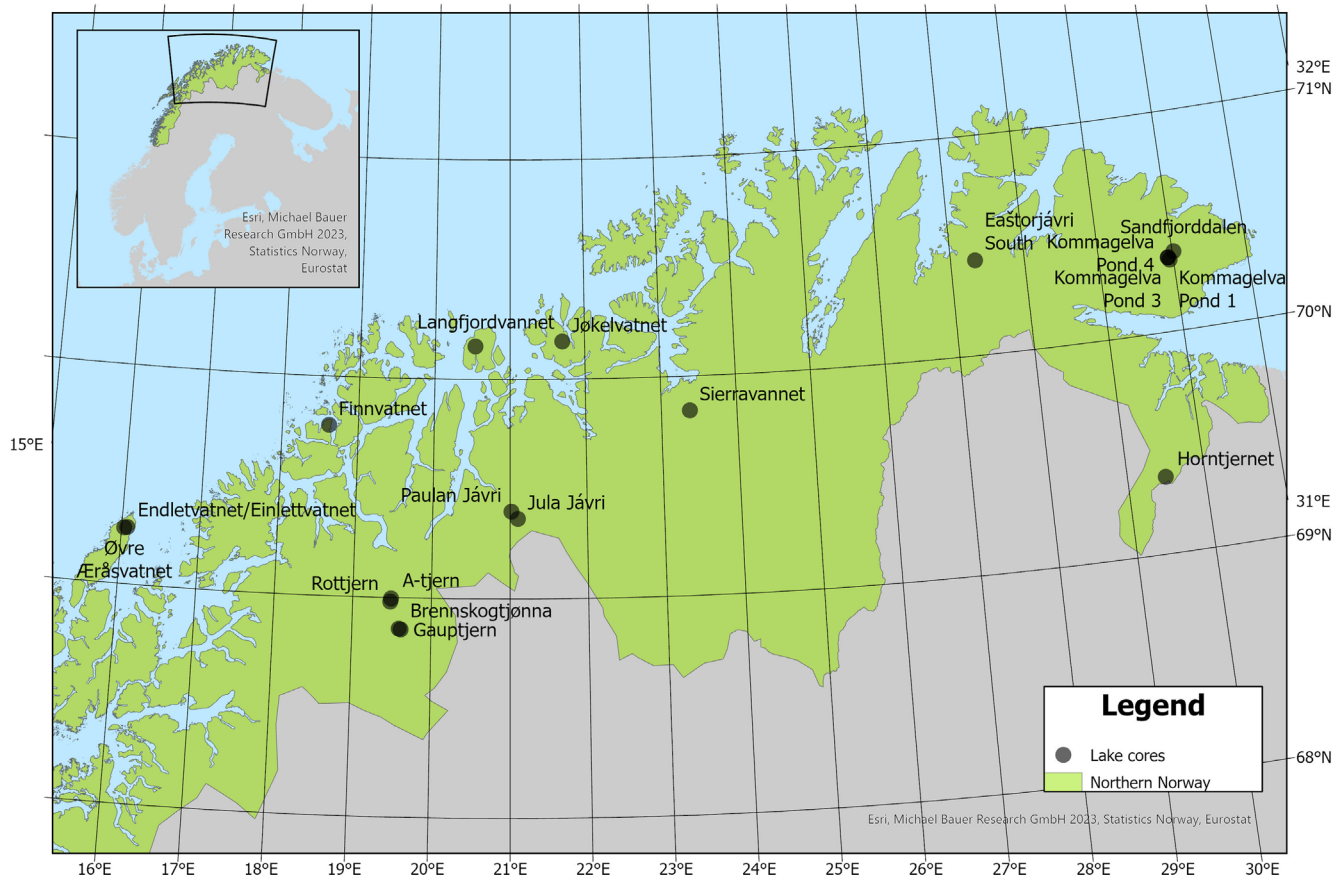

FIGURE 1 | Location of lakes where vegetation surveys and DNA methods are compared.

most abundant ones), and 4 being dominant (making up most of the biomass of the field, shrub or tree layer).

2.3 | Coring

Surface sediments were retrieved using either a Kajak 3 cm diameter mini gravity corer, a 5.9 cm diameter UWITEC USC 06000 corer, or a 4 cm diameter rod-operated Multisampler 12.42.01B (Alsos et al. 2018; Rijal et al. 2021). Eight samples were previously collected by (Alsos et al. 2018), six samples by Rijal et al. (2021), one sample was used by both studies and three samples are new to this study. The cores were opened and subsampled in a dedicated ancient DNA laboratory at The Arctic University Museum of Norway in Tromsø. The top sample from each core was selected for analysis to best reflect the contemporary vegetation recorded in the surveys. These surface samples were then processed through three *sed*aDNA workflows (Figure 2).

2.4 | Capture Probe

The DNA used for the capture probe analysis was extracted following a modified version of the Qiagen DNeasy PowerSoil PowerLyzer protocol as described by Rijal et al. (2021), which had an additional overnight centrifuge step to remove inhibitors (Supporting Information). An extraction negative control was processed alongside each group of 8 samples. The extracts were library prepared, along with a library negative control, following the double-stranded method (Meyer and Kircher 2010) with dual-indexing modifications from Kircher (2012). The libraries were enriched using the PalaeoChip ArcticPlant-1.0 bait-set (Murchie, Kuch, et al. 2021; Murchie, Monteath, et al. 2021). This bait-set primarily targets the plastid *trn*L (UAA) region from ~2100 arctic vascular plant and bryophyte taxa, along with the full *rbc*L and *mat*K loci to increase the capture scope. The enriched libraries were then sequenced on an Illumina HiSeq 1500 at 2 × 90 bp at the Farncombe Metagenomics Facility (McMaster University, ON) (Supporting Information). The sample from lake Nesservatnet (NESS) only produced 3763 read pairs and was excluded from further analysis.

2.5 | Metabarcoding

For metabarcoding, the DNA extraction, PCR, library cleaning and preparation, and sequencing protocol followed Rijal et al. (2021). The 10 samples from Alsos et al. (2018) were re-extracted as these were initially only processed with six replicates. Nine extraction negative controls were included during the separate rounds of extractions. The *trn*L p6-loop region was amplified for the samples with the "g/h" primers (Taberlet et al. 2007), along with seven negative and seven positive PCR controls. Eight PCR replicates were generated for each sample. The library-prepared samples were sequenced on an Illumina NextSeq 550 at 2 × 150 bp at the UiT Genomics Support Centre in Tromsø (Supporting Information).

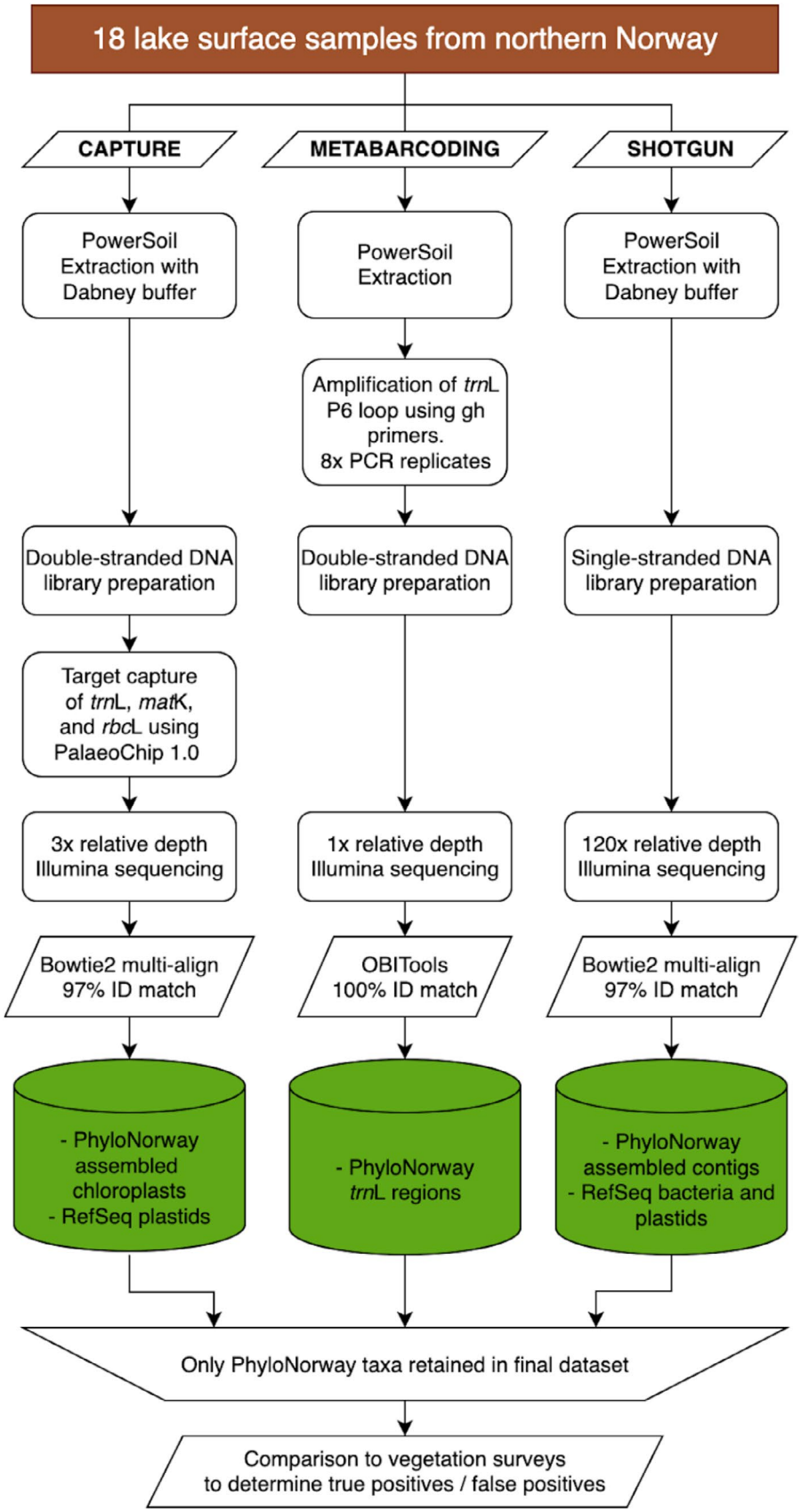

FIGURE 2 | Workflow for the capture probe, metabarcoding and shotgun sequencing approaches.

2.6 | Shotgun Sequencing

Shotgun extractions followed the protocol described for the capture probe analysis. As no DNA extracts remained, the same homogenized sediment samples were reextracted (Supporting Information) with an extraction negative control included for each group of eight samples. The samples were library prepared at the paleogenetic laboratories in the Alfred Wegener Institute (AWI) Helmholtz Centre for Polar and Marine Research in Potsdam, Germany, using a single-stranded library preparation method designed specifically for highly degraded ancient DNA (Gansauge et al. 2017; Gansauge and Meyer 2013; Schulte et al. 2021). The library prepared samples, along with the three extraction negative controls, and one library blank, were sequenced on an Illumina NextSeq2000 at 2×100 bp at the sequencing facility at the Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany (Supporting Information).

2.7 | Bioinformatics

2.7.1 | Capture Probe Data

The sequenced reads were merged and adapter sequences 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). Reads were then mapped to a custom database using *bowtie2* with default parameters allowing for up to 1000 matches (Langdon 2015). The custom database was constructed by compiling 1541 assembled plastid genomes produced by the PhyloNorway project (Alsos et al. 2020) as well as the NCBI RefSeq plastid entries to allow for competitive mapping of potential cyanobacteria and algae sequences. Reads were assigned to the lowest common ancestor (LCA) using *ngsLCA* with a minimum edit distance proportion of 0.97 (Wang et al. 2022). For example, if a read was mapped to a conserved region of two congeneric species, it would be assigned at the genus level. Only those assignments to taxa at family level or lower represented in PhyloNorway with at least three reads were retained for further analysis.

2.7.2 | Metabarcoding

The paired-end reads were merged and further processed using the OBITools pipeline (Boyer et al. 2016) largely following the protocol detailed in Rijal et al. (2021). Taxonomic annotation was performed using a reference database of 1541 P6-loop reference sequences from PhyloNorway (Alsos et al. 2020). Reads were assigned with 100% identity to a reference sequence. If a reference sequence is shared among members of a genus or family, the read was assigned to the lowest common ancestor (LCA) of these members. Taxa were only retained with a minimum of three reads using a custom R script.

2.7.3 | Shotgun Data

The sequenced reads were merged and adapter sequences removed using *fastp* v0.23.4 (Chen 2023) with a minimum length of 30 bp. Taxonomic annotation was performed by *bowtie2* with default parameters allowing up to 1000 matches to a custom database (Langdon 2015). This database was constructed by compiling the 1541 partially assembled genome skims from PhyloNorway (Wang et al. 2021) as well as RefSeq bacteria and plastid entries to allow for competitive mapping. Reads were assigned to the LCA using *ngsLCA* with a minimum edit distance proportion of 0.97 (Wang et al. 2022). For example, if a read was mapped to a conserved region of a bacterium and plant genome, it was assigned at the level of "cellular organism" and not included in further analysis. Only those assignments to taxa at family level or lower represented in PhyloNorway with at least three reads were retained for further analysis.

2.8 | Comparison of DNA With Vegetation Surveys

The DNA detections that matched to the vegetation surveys at the site level were designated as true positives (TP). We then matched the remaining DNA detections with a list of plant taxa from the regional checklist for Northern Norway (Often and Alm 1996) and designated these matches as regional flora (RF). These were not used further as we could not determine if they represented true or false positives. The detections with no match to either the vegetation surveys or regional flora were deemed false positives (FP). Designations were performed at the species, genus, and family taxonomic levels.

2.9 | Cutoff Thresholds

Two approaches were used to set a minimum read threshold for a taxon to be retained in the molecular data. Both are based on the read proportion of a taxon relative to the total reads queried for that sample and aim to minimize the proportion of FPs relative to the total amount of TPs and FPs in the dataset. The first threshold is determined by finding the first local minimum value under 5% for FP/(FP + TP) while the second takes value at which all FP are discarded. These calculations were performed on the level of family, genus, and species, but thresholds for further analysis were set using genus-level values. Lower-level taxonomic detections and read numbers were collapsed at higher ranks (e.g., the TP detection of both *Vaccinium uliginosum* and *Vaccinium vitis-idaea* in one sample only counted as one TP detection at genus-level of *Vaccinium*, but the read counts were added together).

2.10 | DNA Reads and Vegetation Abundance

We performed an ordinal logistic regression using both the proportion of DNA reads and absolute read numbers (after applying optimal filtering) to predict the abundance categories for each taxon, and the transitions between the four abundance categories. This analysis was carried out at genus and species level. Furthermore, the relationship was examined using log, square root and double square root transformations. Models were fitted using the package *ordinal* (Christensen 2012) in R version 4.4.3 with the clmm function (i.e., using a cumulative link model). "Lake" was included as a random effect to account for possible differences in the vegetation survey ranked

abundance estimations. For each sedaDNA method and transformation, we tested whether a flexible or equidistant model fitted the data better by comparing them with a likelihood ratio test. The fit of the models using different transformations of read proportions or numbers was compared with Akaike's information criterion (AIC).

3 | Results

3.1 | Taxa Detected in Vegetation Surveys

A total of 347 unique taxa were identified across all 18 lakes for both the 2 m and extended vegetation surveys. These taxa included 280 species belonging to 159 genera and 59 families (Table S2). The highest number of taxa was counted at Øvre Æråsvatnet (112), whereas only 33 species were counted at Kommagelva Pond 3 (Figure S4).

3.2 | Sequencing Results

From the capture probe samples, a total of 24,386,721 read pairs were obtained with an average of 1,274,382 ± 634,466 read pairs per sample (Table S3). The two extraction controls and library blank totalled 17,426 read pairs. After adapter trimming and merging, 19,705,695 (77.3%) and 9216 (52.9%) sequences were retained for the samples and controls respectively. Four taxa were identified in the negative extraction controls with > 10 reads: Pinaceae/*Pinus* with 835 reads, Triticeae/*Triticum* with 26 reads, *Myrica* with 18 reads, and *Sparganium* with 15 reads. As Pinaceae/*Pinus* and *Myrica* do not appear in any samples from the same extraction group as these controls (Table S4), we do not believe their presence in the dataset to be a result of contamination, so these taxa are retained in the dataset for analysis. *Triticum* appears in five samples from the same extraction group as the control and *Sparganium* appears in one sample, but as these are both designated regional flora (RF), their presence does not impact the evaluation of capture probe's performance.

From the metabarcoding samples, a total of 7,923,720 read pairs were obtained with an average of 381,629 ± 88,837 per sample across eight PCR replicates. The nine extraction controls totalled 291,138 reads. After merging and processing through the OBITools pipeline, a total of 5,854,199 (76.7%) reads were retained and matched to a barcode. Across the nine extraction controls, 14 taxa were identified, with six of them occurring in the samples from the same extraction group (Table S4). Four of these barcodes are only at family-level resolution and do not impact analyses, while the two others, *Cannabis* and *Pinus*, are both designated regional flora at their respective sites and do not impact performance evaluation.

From the shotgun samples, a total of 904,790,751 read pairs were obtained with an average of 50,266,152 ± 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, a total of 699,592,884 (67.2%) sample sequences and 15,675 (4.4%) control sequences were retained; only two taxa were identified among the negative controls with > 10 read counts: *Avena* (oat) with 32 reads and *Triticum* (wheat) with 12 reads (Table S4). Both appear as only regional flora in the dataset and do not impact performance evaluation. The partially assembled PhyloNorway genome skims ranged in size from 7.7 M—2.1 B base pairs, but there was no correlation between a taxon's reference database size and the number of reads assigned to that taxon across the 18 lakes (Figure S5).

3.3 | Optimal Read Filtering Threshold

At the species level, the number of false positives at relaxed filtering was high for both capture probe and shotgun sequencing (Table 1, Figure 3a, Table S5), and it remained high (> 30%) even with more stringent filtering criteria, making these methods unreliable at the species level (Table 1, Figure 3b). Thus, for capture probe and shotgun sequencing, the larger number of genus-level identifications were used for more robust filtering estimates (Figures S1–S3). The controls were not used for optimizing the filtering threshold due to the relative lack of taxa detected. Metabarcoding detected 224 species at optimal filtering criteria, but the error rate was still relatively high (12%). This was partly caused by 17 fern detections that are known to appear in the gametophyte stage outside the range of the sporophyte (Brock 2025). These gametophytes could contribute eDNA to the sediment archive while being overlooked by vegetation surveys resulting in an erroneous false positive record. Disregarding ferns, the error rate for metabarcoding at species level was 5%.

At the genus level, the total number of taxa detections (FP and TP) under the base filtering of 3 reads was highest for shotgun, but it had also the highest error rate (Table 1). Increasing the filtering stringency reduces both the total number of detections and the proportion of false positives across all methods, with a genus level error rate of 3% for capture probes, 1% for metabarcoding, and 4% for shotgun (Table 1). When applying the optimal filtering level, the detection of TP genera was 2.1 times higher for metabarcoding than shotgun, and 6.4 times higher for metabarcoding than capture probes. These numbers do not take into account the 41 times deeper sequencing depth for shotgun than capture probes and 131 times deeper for shotgun than metabarcoding. When normalized by sequencing depth, capture probes detected 8 times more vascular plant genera than shotgun sequencing, and metabarcoding detected 26 times more than capture probes and 256 times more than shotgun sequencing. At the family level, the false positive rate was low for all three methods even at relaxed filtering (Figure 3a,b). The mean number of taxa detected at all three taxonomical levels was highest for metabarcoding, followed by shotgun sequencing, and lowest for capture probes (Figure 3c). The low number of taxa detected by capture probes may be due to technical problems encountered when performing this method (see discussion: methods performance).

3.4 | Ability to Detect Local Vegetation

Since shotgun and capture probe had high error rates at the species level even after our stringent filtering approach (Figure S6), we hereafter compare results at the genus level and only include true positives. Metabarcoding identified the greatest proportion

TABLE 1 | Table of false positive (FP) and true positive (TP) taxa detections and unique taxa retained at the genus- and species-level at different cutoff thresholds for each sequencing method. The cutoff thresholds are in order of increasing strictness starting with the initial base cutoff of three reads. The FP proportion is calculated by FP/(FP + TP) disregarding regional flora identifications.

Method	Capture	Capture	Capture	Metabarcoding	Metabarcoding	Metabarcoding	Shotgun	Shotgun	Shotgun
Cutoff	3 reads	5.6×10e-6	1.96×10e-5	3 reads	3.0×10e-5	3.4×10e-4	3 reads	3.2×10e-6	4.32×10e-5
FP genera detection proportion	0.12	0.03	0.00	0.01	0.01	0.00	0.35	0.04	0.00
FP genera detections	11	2	0	5	4	0	371	7	0
TP genera detections	81	62	32	399	365	220	696	189	37
FP genera	10	2	0	1	1	0	108	5	0
TP genera	32	23	11	81	80	63	129	72	25
FP species detection proportion	0.74	0.62	0.30	0.12	0.10	0.07	0.79	0.49	0.36
FP species detections	49	23	3	30	23	10	1859	103	9
TP species detections	17	14	7	224	212	134	482	109	16
FP species	20	7	1	8	7	5	503	64	9
TP species	5	4	1	47	47	36	160	48	11

Note: The yellow columns highlight cutoffs determined to be "optimal" for each method, while the FP taxa rows are shaded from green to red indicating low to high proportions of FPs

of genera within our 2 m vegetation surveys, and the detection rate was greater for the genera with higher ranked abundances (Figure 4). Rare taxa were detected at a rate of 2.5%, 7.8%, and 23% out of all possible observations by capture probe, shotgun, and metabarcoding methods respectively. In contrast, dominant genera were detected at a rate of 22%, 49%, and 77% respectively. The proportion of genera detected by all three methods was low, especially for less dominant taxa (Figure 4). Metabarcoding showed the greatest overlap with shotgun sequencing, while capture probe sequencing detected fewer genera overall. Further, there were more genera uniquely detected in metabarcoding (16%–36%) than shotgun (3%–5%) and capture probe (~1%–2%). At all ranked abundance categories, metabarcoding identified a greater proportion of genera than both metagenomic methods, while shotgun outperformed capture probe.

3.5 | Correlation Between Sequence Reads and the Abundance in the Vegetation

The proportion of reads for TPs at the genus level increased with abundance across taxa detected by metabarcoding (median proportion of reads: 0.00048, 0.00092, 0.00223, and 0.02105 for rare, scattered, common, and dominant genera, respectively). For shotgun sequencing there was a similar overall trend (median proportion of reads: 0.01034, 0.00690, 0.01153, and 0.02837 for rare, scattered, common, and dominant genera, respectively), whereas for capture probe data the relationship was less clear as the highest proportion of reads was recorded for scattered taxa (median proportion of reads: 0.01233, 0.07030, 0.02580, 0.04468; Figure 5).

The ordinal regressions showed a significant relationship of the proportion of reads obtained from metabarcoding to the abundance category of plants observed within 2 m from the lake shore (Table S6). Comparing different transformations of the read proportions with AIC suggested that double square root transformation resulted in the best model at both the genus and species level. At the species level, a log transformation was also supported ($\Delta AIC = 1.43$). For shotgun data, relationships between the proportion of reads and the abundance categories were also significantly positive, but the results from AIC regarding the different transformations were less clear. All transformations seemed equally adequate at the species level, while at the genus level only the model without transformation appeared less supported ($\Delta AIC = 4.17$). For the capture probe data, the proportion of reads was not related to the abundance categories at the genus level, while there were too few TPs at the species level to evaluate the relationship. The poor performance of this method may have been due to technical problems (see discussion). Results were similar when calculated for the number of reads (Table S7).

4 | Discussion

Our findings show that shotgun sequencing and capture probes require stringent filtering, whereas metabarcoding has a lower initial rate of false positives. All methods reliably identified taxa at genus level after optimal filtering, whereas only metabarcoding produced reliable species-level identifications. Metabarcoding identified a larger proportion of genera at each

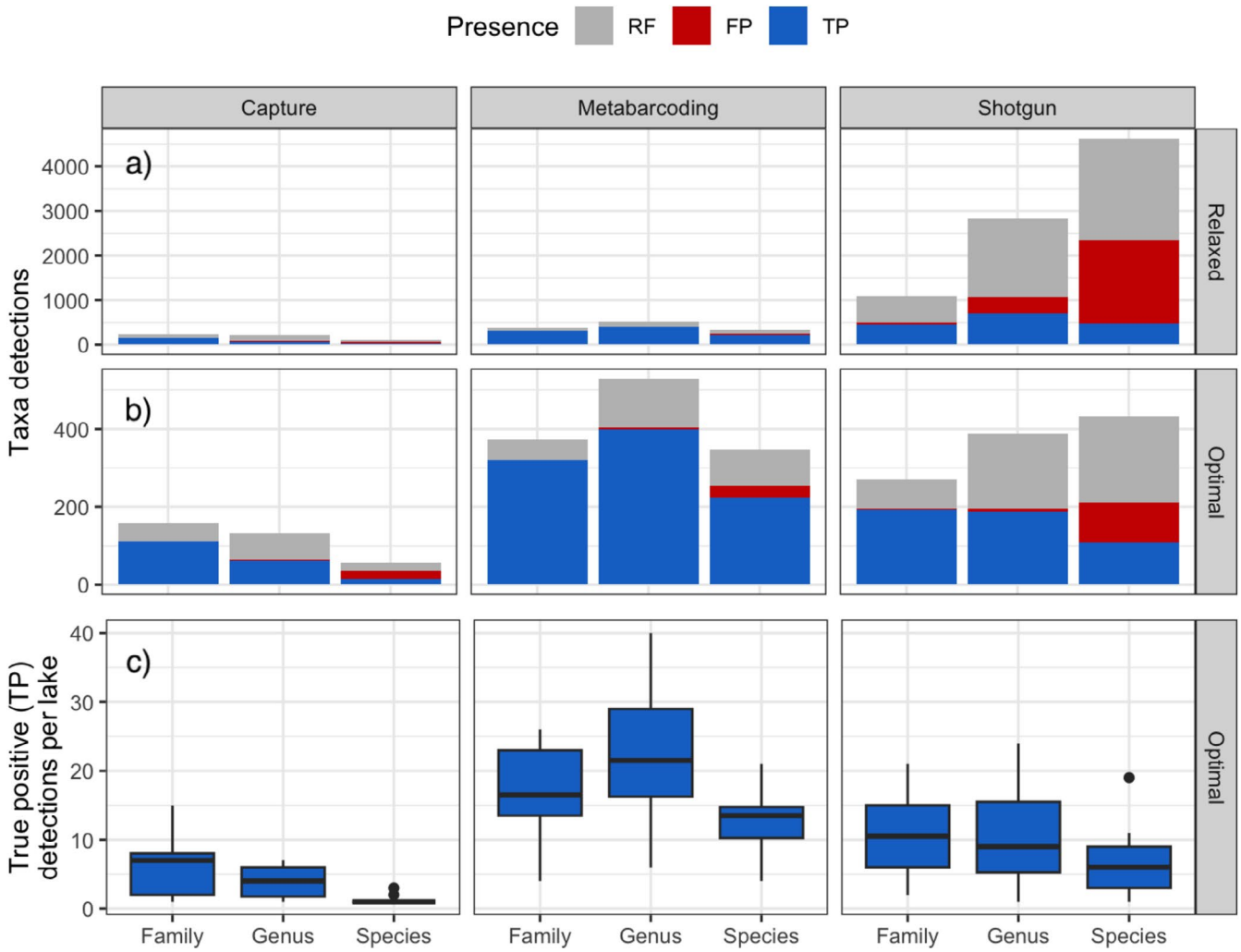

FIGURE 3 | The count of true positive (TP), false positive (FP), and regional flora (RF) taxa identified by capture probe, metabarcoding, and shotgun methods. The values are displayed at the taxonomic levels of family, genus, and species for (a) a relaxed cut-off of 3 reads for all methods; (b) optimal cut-off of 3 reads for metabarcoding and 3.2×10^{-6} and 5.6×10^{-6} of reads for shotgun, and capture respectively, and (c) number of true positive taxa detected at each lake after optimal filtering. A sequence identified at species- or genus-level is also included in higher taxonomic levels.

abundance category compared to other methods. Additionally, DNA reads from metabarcoding and shotgun sequencing reliably predicted vegetation abundance, suggesting a link between read frequency and plant biomass.

Although we aimed to optimize each laboratory protocol, this did result in different extraction and library preparation protocols used, which could introduce biases in the results. The sequencing depth for each method also varied, with shotgun sequencing having the highest sequencing depth, followed by capture probe and then finally metabarcoding.

4.1 | Filtering

It is difficult to assess if our optimal filtering threshold of 5.6×10^{-6} of the vascular plant reads for capture probes is more or less strict than that used in previous studies, as they used different approaches to filtering (Kjær et al. 2022; Murchie, Monteath, et al. 2021), but our data show that the cutoff level has a strong impact on detections and error rate for this type of data.

Our results suggest that a threshold of three or ten reads is sufficient for metabarcoding. This is similar to a thresholds used in many studies, for example, five (Stoof-Leichsenring et al. 2022) or ten reads (Garcés-Pastor et al. 2025; S. Liu et al. 2024; Rijal et al. 2021; Wang et al. 2021). Some studies apply additional criteria based on the repeatability across PCR replicates (Rijal et al. 2021) or clustering approaches (Y. Liu et al. 2025). While the absolute number of reads would vary with sample size and sequencing depth, metabarcoding seems to be reliable with a general low filtering threshold, and due to the earlier studies on threshold level, the metabarcoding community in general apply adequate filtering thresholds.

For shotgun sequencing, our estimated optimal cut-off level of 3.2×10^{-6} of total reads assigned is less stringent than that used by others. Wang et al. (2021) applied an initial minimum filter of 1% of *Viridiplantae* reads and further removed

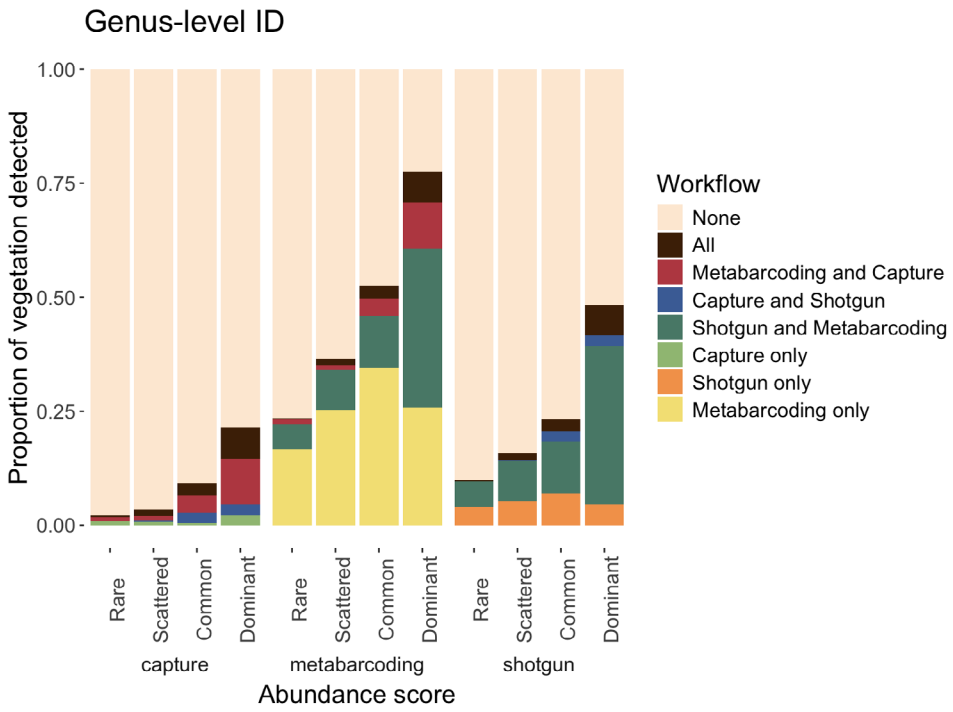

FIGURE 4 | The proportion of 2 m vegetation survey taxa with a DNA match at the genus level grouped by DNA method detected by and ranked abundance. The DNA data shown are those done after the final optimal filtering.

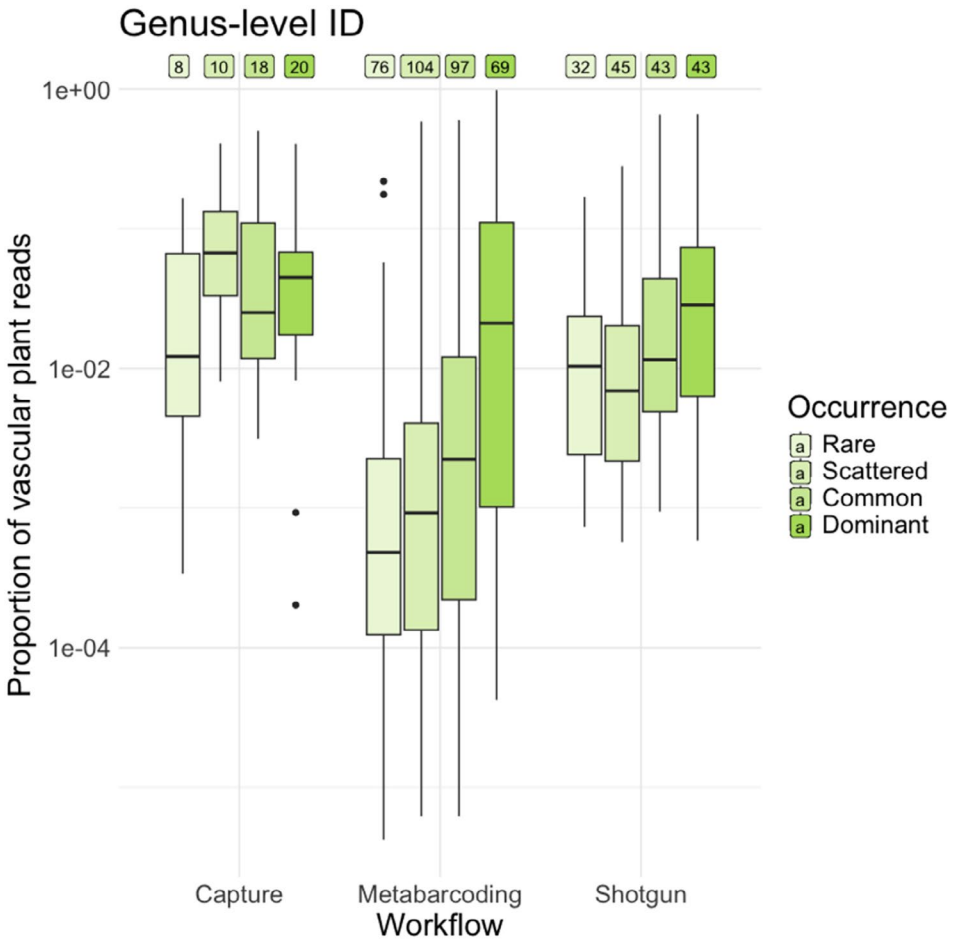

FIGURE 5 | Proportions of vascular plant reads of the true positives at the genus level of the best estimated cutoff approach plotted according to vegetation survey ranked abundance. The horizontal line shows the median of the data and the boxes show the interquartile range. The numbers above the boxplots represent the number of genus observations within each category of plant abundance in vegetation surveys.

the taxa with a combined read percentage lower than 5.39%, which is the median value they observed for Arctic genera. Courtin et al. (2022) eliminated any sequences with an abundance of less than 0.2% of total *Viridiplantae* reads. Liu et al. (2024) retained only those taxa detected in at least two samples with a combined read count of ≥ 5, therefore retaining 94% of their reads, which is less stringent than our filtering. Kjær et al. (2022) were more strict as they filtered out low-abundance taxa at the genus level by setting a threshold at half the median read count and retained taxa found in at least 3 samples. Thus, our use of vegetation data as an estimate of true and false positives may serve as a guideline for future studies.

For all three methods, the number of FPs is underestimated, as we only matched the sequences to PhyloNorway. While using global reference libraries such as EMBL may increase detection, it may also inflate the FPs (Alsos et al. 2022). In cases where the local flora is less well covered in reference libraries or the reference library includes more erroneous annotated sequences, we advocate using a stricter cut-off level. As shotgun uses the whole genome, the effect of incomplete or erroneously annotated sequences will be considerably higher than for capture probes. Another source of potential FPs in the shotgun data is exogenous DNA within the reference databases. Even a well curated database like PhyloNorway (Alsos et al. 2020; Wang et al. 2021) contains endophytes such as bacteria and fungi (Elliott et al. 2025; Oskolkov et al. 2025), though increased reference material for these endophytes may reduce the false positive rate for shotgun data. Given the importance of filtering for biodiversity estimates, transparent reporting of these filtering criteria is crucial for reproducibility and meaningful comparisons.

4.2 | Methods Performance

Our capture samples had on average 16 vascular plant taxa, thus lower than the average of 48.9 found by Murchie, Kuch, et al. (2021); Murchie, Monteath, et al. (2021) in their capture probe study from Pleistocene–Holocene transition permafrost samples and the 76.2 taxa per sample recorded by Kjær et al. (2022) in their shotgun study on 2 million years old permafrost samples from Greenland. While capture probe is expected to give ~1000× more 'on-target' DNA proportionally than comparable shotgun sequencing libraries generated using the same methodology (More et al. 2025), our capture identification was lower than the number of taxa found in shotgun analyses of the same samples. Thus, we suspect that the capture hybridization may not have worked optimally, although differences in extract and library preparation complicate interpretations. Capture libraries were generated and sequenced in February–March 2020 during the onset of the COVID-19 pandemic. Subsequent COVID-19 lockdowns and laboratory access restrictions in Canada prevented reprocessing or independent replication; therefore, potential batch effects and other laboratory variations cannot be excluded. Degradation due to long transport time of the samples to Canada, issues such as enzyme/reaction failures during lab preparation or suboptimal bait hybridization could have reduced performance. Additionally, we have noted that lake sediments may yield less DNA captured than permafrost samples using the PalaeoChip Arctic v1.0 Plant bait-set (Murchie pers. obs.). Further optimizations to the taxonomic breadth and genomic depth of an arctic plant bait set beyond *rbc*L, *mat*K, and *trn*L are recommended for future capture research if the focus is specifically on plants. Thus, overall, the method may have higher potential than found in our study. We caution against interpreting these results as representative of capture performance in general given these dataset specific limitations.

The average number of TP and regional flora taxa found per sample using metabarcoding, 37.2, is higher than in the previous study of surface sediments from 11 lakes in the region (mean 19.7, Alsos et al. 2018), probably due to improved laboratory procedures. It is within the range of ancient samples in the region (20.6–65.5, $n = 316$ Rijal et al. 2021) and surface samples in NE Siberia (15–54, $n = 32$, Niemeyer et al. 2017). It is considerably lower than 50–150 taxa per sample ($n = 705$) found in the Alps, where a similar local DNA reference library is available (Garcés-Pastor et al. 2025), but that is probably due to the smaller species pool in the north. Thus, we conclude that the metabarcoding analyses here are representative for the method.

For shotgun analyses, the mean number of taxa per sample (25.5 when counting both TP and regional flora at family and genus level) was higher than the 8.13 and 17.5 taxa (at the genus level) detected by (Wang et al. 2021) and (Courtin et al. 2022) respectively. It was lower than (S. Liu et al. 2024), who found 40 taxa per sample, and considerably lower than (Kjær et al. 2022), who found 76.21 taxa per sample, but had ten times higher sequencing depth. Given the high impact of both sequencing depth and cut-off level used on taxa richness reported, we conclude that our samples are within the range of expectation, and that they are representative of the shotgun method.

4.3 | Detectability and Taxonomic Resolution

Based on our results, metabarcoding appears to be currently the most effective method for detecting local vegetation from lake sediments. The addition of capture probe and shotgun sequencing provided only marginal improvements in taxon detection while also increasing false positive detections and may not justify the increased laboratory time and costs for many studies. Similarly, a review of eDNA-based studies that have detected aquatic macrophytes concluded that the detectability in general is higher using metabarcoding of the P6 loop than shotgun sequencing or capture probes (Revéret et al. 2023). Further, while shotgun and capture probes were not reliable at the species level, metabarcoding was able to identify 47 species at < 10% FP rate.

While the capture probe workflow annotated 3.4× more reads than shotgun sequencing relative to sequencing effort, the overall detection was still low. One limitation of the PalaeoChip bait set used in this study is that it targets the entire genes of *rbc*L, *trn*L, and *mat*K, which include highly conserved regions. These conserved regions are efficiently enriched but are often taxonomically uninformative, leading to plant identifications being assigned to higher taxonomic ranks such as phylum, class, or order (Murchie, Monteath, et al. 2021). However, a key advantage of targeting full genes is that it enables broader genomic

coverage, which can support downstream analyses such as population genetics and the assessment of postmortem DNA damage patterns. Another benefit of capture-based approaches is their compatibility with standard DNA barcoding markers, which, in contrast to the nonstandard minibarcode P6 loop, are available for the flora of most regions (https://boldsystems.org/).

While metabarcoding only explores a minor fraction of the plant genome and therefore has been argued to be less efficient than shotgun sequencing (Wang et al. 2021), our current comparison, as well as the majority of studies published earlier (Revéret et al. 2023), report considerably higher detection and taxonomic resolution for metabarcoding than shotgun data. Reliable species level identifications enable the reconstruction of both abiotic conditions such as temperature, pH, moisture, as well as biotic interactions such as dispersal, pollination, and mycorrhiza based on known species traits (Alsos et al. 2022; Garcés-Pastor et al. 2025). Also, metabarcoding is a more cost-efficient method as a much higher proportion of the sequences belong to target taxa, thus facilitating high time and space resolution analyses (Alsos et al. 2024). Furthermore, it is efficient in terms of bioinformatic requirements and relies on a reference library for only a small part of the genome.

Metabarcoding also has some down-sides compared to shotgun and capture probes. If whole ecosystem reconstruction is the aim, additional analyses are required for other groups of organisms, by using mammal, fish, birds, or fungi specific primers (Capo et al. 2023), where the efficiency of each additional primer depends on the available reference data. Another downside is that ancient DNA metabarcoding relies on relatively long fragments (~90 bases), making it less suitable for very degraded material. The oldest samples successfully analyzed so far are Eemian lake sediments from arctic Canada (Crump et al. 2021), but for older samples or material from warmer regions, shotgun or capture probes may provide better results. Further, metabarcoding does not provide aDNA damage patterns due to the annealing of primers. Exploring damage patterns is desirable when focusing on species that are commonly found as contaminants in environmental and archeological DNA studies (Smith et al. 2015; Weiß et al. 2015). It is also important when investigating caves or other nonwater logged sediments where leaching is likely (Haile et al. 2007). However, no leaching/translocation has been documented for lake sediments (Sjögren et al. 2017), and they are generally regarded to provide reliable timelines for sedaDNA and many other determinants (Garcés-Pastor et al. 2023). Additionally, the genome-wide information provided by shotgun and capture probe approaches allow for the investigation of functional genes, haplotype reconstructions, and other analyses that go beyond the scope of taxonomic classification.

The low richness of shotgun may have resulted from both the high proportion of noninformative conserved genomic regions and the incomplete reference libraries, preventing reliable identification at species level. When comparing sequencing depth of PhyloNorway to Kew plant genome sizes, PhyloNorway may have on average only approximately 60% genome coverage, which can introduce false positives by assigning reads to incorrect but closely related species. Increasing the amount of reference material should increase the true positive match for shotgun sequencing (Elliott et al. 2025), and thus projects that provide deep sequencing and/or fully assembled genomes for plants such as the Darwin Tree of Life (https://www.darwintreeoflife.org/) may considerably increase the potential for correct identifications. However, given the more than 300,000 species of vascular plants (Christenhusz and Byng 2016), with relatively large and complex genomes, it will take time to complete. In regions with smaller floras, for example the Arctic islands, a complete reference database will be easier to obtain.

4.4 | Ability to Record Quantitative Information

Our data showed that both metabarcoding and shotgun data show a clear increase in reads with increase in ranked abundance in the vegetation. For metabarcoding, similar results have also been found for a single lake compared to large sample vegetation surveys (Ataman et al. 2025) and for $200 \times 1\,m^2$ vegetation surveys compared to soil samples (Ariza et al. 2024). Thus, while PCR bias clearly does take place and, for example, grasses and sedges may be underrepresented (Alsos et al. 2018), the overall quantitative pattern may be relatively robust. Similarly, although shotgun data may be biased in terms of unequal representation of the available reference genomes, the overall pattern remains robust. For both methods, one needs to take the taphonomical processes into account, which commonly causes over-representation of taxa growing along the streams and in the lakes. The lack of correlation of capture probe data with plant abundance may have been due to the suboptimal performance of the method here.

5 | Conclusion

Each of the methods has its pros and cons; whilst capture probes and shotgun sequencing both have the advantage of displaying DNA damage patterns and thus may be the best choice if the focus is archeological or nonwaterlogged sediments, their ability to reflect plant diversity is limited by the lack of reliable information at the species level. Metabarcoding is currently the most effective method for detecting vegetation given the current state of comparative reference databases, performing well at both the species and genus levels compared to shotgun sequencing and capture probe techniques. While shotgun and capture probe data can benefit significantly from filtering to reduce noise, metabarcoding tends to produce fewer false positives inherently. Additionally, both metabarcoding and shotgun methods show a positive correlation between ranked taxonomic abundance and read counts, supporting their use in semiquantitative analyses. Capture probes may increase the amount of target DNA sequenced, but further studies are needed to get a better understanding of its ability to capture quantitative patterns. In this dataset, capture recovered fewer on-target plant taxa than expected relative to shotgun sequencing, despite the typical enrichment of on-target DNA achieved by hybridization capture. This outcome is consistent with workflow-specific differences (including different extracts and single- vs. double-stranded library conversion efficiency) and/or batch-related impacts on capture hybridization efficiency. Given differences among workflows

and because capture libraries could not be independently replicated, we interpret the low capture recovery as study-specific and caution against generalizing it to capture performance in general. While metabarcoding detected most true positives in our analysis, the high false positive rate associated with shotgun sequencing is expected to decline as more complete whole genome reference libraries become available. This advancement may make shotgun sequencing particularly advantageous for analyzing highly degraded DNA, especially when fragment lengths are too short for successful metabarcoding.

Author Contributions

D.P.R. and I.G.A. designed the study. Vegetation surveys were conducted by N.A.S., L.D.E., D.E., A.G.B., and I.G.A. Metabarcoding data was generated by D.P.R., A.N.R., Y.L., and I.P. Shotgun data was generated by L.D.E. and K.S.-L. Capture data was generated by I.P. and T.J.M. Data was analyzed by N.A.S., L.D.E., D.P.R., D.E., N.G.Y., and I.G.A. The manuscript was written by N.A.S. and L.D.E. with feedback from all authors.

Acknowledgments

We thank Marie Føreid Merkel and Janine Klimke for laboratory assistance. The study was funded by Research Council of Norway grant 250963/F20 (to I.G.A., N.G.Y., D.E.; supported D.P.R.) for the ECOGEN project; The European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement No 819192 for the IceAGenT project (to I.G.A., A.G.B.; supported Y.L.); UiT and the ArcEcoGen Centre (supported L.D.E. and N.A.S.). Bioinformatic analyses were performed on resources provided by UNINETT Sigma2—the National Infrastructure for High-Performance Computing and Data Storage in Norway.

Funding

The study was funded by Research Council of Norway grant 250963/F20 (to I.G.A., N.G.Y., D.E.; supported D.P.R.) for the ECOGEN project; The European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement No 819192 for the IceAGenT project (to I.G.A., A.G.B.; supported Y.L.); UiT and the ArcEcoGen Centre (supported L.D.E. and N.A.S.).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The raw capture probe, metabarcoding, and shotgun sequencing data are deposited in the European Nucleotide Archive (ENA) under the project accession code PRJEB100595. The identified taxa alongside the code for analyses and generating figures are available at https://github.com/salanova-elliott/Comparison_manuscript. Scripts for initial processing of the metabarcoding data are at https://github.com/Y-Lammers/MergeAndFilter.

References

Alsos, I. G., V. Boussange, D. P. Rijal, et al. 2024. "Using Ancient Sedimentary DNA to Forecast Ecosystem Trajectories Under Climate Change." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 379, no. 1902: 20230017.

Alsos, I. G., Y. Lammers, N. Yoccoz, et al. 2018. "Plant DNA Metabarcoding of Lake Sediments: How Does It Represent the Contemporary Vegetation." *PLoS One* 13, no. 4: e0195403.

Alsos, I. G., S. Lavergne, M. K. F. Merkel, et al. 2020. "The Treasure Vault Can Be Opened: Large-Scale Genome Skimming Works Well Using Herbarium and Silica Gel Dried Material." *Plants* 9, no. 4: 432.

Alsos, I. G., D. P. Rijal, D. Ehrich, et al. 2022. "Postglacial Species Arrival and Diversity Buildup of Northern Ecosystems Took Millennia." *Science Advances* 8, no. 39: eabo7434.

Ariza, M., M. Engelstad, E. Lieungh, et al. 2024. "Evaluating the Feasibility of Using Plant-Specific Metabarcoding to Assess Forest Types From Soil eDNA." *Applied Vegetation Science* 27, no. 4: e12806. https://doi.org/10.1111/avsc.12806.

Ataman, T. G., Y. Lammers, I. G. Alsos, D. P. Rijal, and A. G. Brown. 2025. "Sedimentary DNA From Lake Depocenters Maximizes Detection of Catchment Vegetation." *Communications Earth & Environment* 6, no. 1: 1–10.

Boyer, F., C. Mercier, A. Bonin, Y. Le Bras, P. Taberlet, and E. Coissac. 2016. "Obitools: a Unix-Inspired Software Package for DNA Metabarcoding." *Molecular Ecology Resources* 16: 176–182.

Brock, J. M. R. 2025. "Effective Dispersal of Fern Spore and the Ecological Relevance of Zoochory." *Biological Reviews of the Cambridge Philosophical Society* 100, no. 5: 2116–2130.

Capo, E., C. Barouillet, and J. P. Smol. 2023. *Tracking Environmental Change Using Lake Sediments: Volume 6: Sedimentary DNA*. Springer International Publishing.

Chen, S. 2023. "Ultrafast One-Pass FASTQ Data Preprocessing, Quality Control, and Deduplication Using Fastp." *IMeta* 2, no. 2: e107.

Christenhusz, M. J. M., and J. W. Byng. 2016. "The Number of Known Plants Species in the World and Its Annual Increase." *Phytotaxa* 261, no. 3: 201.

Christensen, R. H. B. 2012. "Ordinal: Regression Models for Ordinal Data. R Package Version 2011.08-11." http://paperpile.com/b/StohjL/3eVJL.

Courtin, J., A. Perfumo, A. A. Andreev, et al. 2022. "Pleistocene Glacial and Interglacial Ecosystems Inferred From Ancient DNA Analyses of Permafrost Sediments From Batagay Megaslump, East Siberia." *Environmental DNA (Hoboken, N.J.)* 4, no. 6: 1265–1283.

Crump, S. E., B. Fréchette, M. Power, et al. 2021. "Ancient Plant DNA Reveals High Arctic Greening During the Last Interglacial." *Proceedings of the National Academy of Sciences* 118, no. 13: e2019069118.

Dabney, J., M. Meyer, and S. Pääbo. 2013. "Ancient DNA Damage." *Cold Spring Harbor Perspectives in Biology* 5, no. 7: a012567.

Elliott, L., F. Boyer, T. Lemane, PhyloAlps and PhyloNorway consortia, I. G. Alsos, and E. Coissac. 2025. "Wholeskim: Utilising Genome Skims for Taxonomically Annotating Ancient DNA Metagenomes." *Molecular Ecology Resources* 25: e70001.

Elven, R., C. S. Bjorå, E. Fremstad, H. Hegre, and H. Solstad. 2022. *Norsk Flora (Norwegian Flora)*. 8th ed. Samlaget.

Ficetola, G. F., J. Pansu, A. Bonin, et al. 2015. "Replication Levels, False Presences and the Estimation of the Presence/Absence From eDNA Metabarcoding Data." *Molecular Ecology Resources* 15, no. 3: 543–556.

Foster, N. R., A. R. Jones, O. Serrano, et al. 2024. "Environmental DNA Identifies Coastal Plant Community Shift 1,000 Years Ago in Torrens Island, South Australia." *Communications Earth & Environment* 5, no. 1: 1–11.

Gansauge, M.-T., T. Gerber, I. Glocke, et al. 2017. "Single-Stranded DNA Library Preparation From Highly Degraded DNA Using T4 DNA Ligase." *Nucleic Acids Research* 45, no. 10: e79.

Gansauge, M.-T., and M. Meyer. 2013. "Single-Stranded DNA Library Preparation for the Sequencing of Ancient or Damaged DNA." *Nature Protocols* 8, no. 4: 737–748.

Garcés-Pastor, S., P. D. Heintzman, S. Zetter, et al. 2025. "Wild and Domesticated Animal Abundance Is Associated With Greater Late-Holocene Alpine Plant Diversity." *Nature Communications* 16, no. 1: 3924.

Garcés-Pastor, S., K. Nota, D. P. Rijal, et al. 2023. "Terrestrial Plant DNA From Lake Sediments: Volume 6: Sedimentary DNA." In *Tracking Environmental Change Using Lake Sediments*, edited by E. Capo, C. Barouillet, and J. P. Smol, vol. 21, 275–298. Springer International Publishing.

Goodell, T., R. I. Griffiths, H. S. Gweon, L. Norton, S. B. Busi, and D. S. Read. 2025. "Deciphering Landscape-Scale Plant Cover and Biodiversity From Soil eDNA." *Environmental DNA* 7: e70191.

Grytnes, J. A., H. J. B. Birks, and S. M. Peglar. 1999. "Plant Species Richness in Fennoscandia: Evaluating the Relative Importance of Climate and History." *Nordic Journal of Botany* 19, no. 4: 489–503.

Haile, J., R. Holdaway, K. Oliver, et al. 2007. "Ancient DNA Chronology Within Sediment Deposits: Are Paleobiological Reconstructions Possible and Is DNA Leaching a Factor?" *Molecular Biology and Evolution* 24: 982–989.

Johnson, M. G., L. Pokorny, S. Dodsworth, et al. 2019. "A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes From Any Flowering Plant Designed Using k-Medoids Clustering." *Systematic Biology* 68, no. 4: 594–606.

Julián-Posada, I., G. Gil-Romera, S. Garcés-Pastor, et al. 2025. "Neolithic Pastoralism and Plant Community Interactions at High Altitudes of the Pyrenees, Southern Europe." *Communications Earth & Environment* 6, no. 1: 1–10.

Kircher, M. 2012. "Analysis of High-Throughput Ancient DNA Sequencing Data." In *Ancient DNA: Methods and Protocols*, edited by B. Shapiro and M. Hofreiter, 197–228. Humana Press.

Kjær, K. H., M. Winther Pedersen, B. De Sanctis, et al. 2022. "A 2-Million-Year-Old Ecosystem in Greenland Uncovered by Environmental DNA." *Nature* 612, no. 7939: 283–291.

Langdon, W. B. 2015. "Performance of Genetic Programming Optimised Bowtie2 on Genome Comparison and Analytic Testing (GCAT) Benchmarks." *Biodata Mining* 8, no. 1: 1.

Liu, S., K. R. Stoof-Leichsenring, L. Harms, et al. 2024. "Tibetan Terrestrial and Aquatic Ecosystems Collapsed With Cryosphere Loss Inferred From Sedimentary Ancient Metagenomics." *Science Advances* 10, no. 21: eadn8490.

Liu, Y., S. Lisovski, J. Courtin, K. R. Stoof-Leichsenring, and U. Herzschuh. 2025. "Plant Interactions Associated With a Directional Shift in the Richness Range Size Relationship During the Glacial-Holocene Transition in the Arctic." *Nature Communications* 16, no. 1: 1128.

Meucci, S., L. Schulte, H. H. Zimmermann, et al. 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, no. 5: 2173–2193.

Meyer, M., and M. Kircher. 2010. "Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing." *Cold Spring Harbor Protocols* 2010, no. 6: pdb.prot5448.

More, K. D., O. Lebrasseur, J. L. Garrido, et al. 2025. "Validating a Target-Enrichment Design for Capturing Uniparental Haplotypes in Ancient Domesticated Animals." *Molecular Ecology Resources* 25, no. 7: e14112.

Murchie, T. J., M. Kuch, A. T. Duggan, 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: 305–328.

Murchie, T. J., A. J. Monteath, M. E. Mahony, et al. 2021. "Collapse of the Mammoth-Steppe in Central Yukon as Revealed by Ancient Environmental DNA." *Nature Communications* 12, no. 1: 7120.

Nichols, R. V., C. Vollmers, L. A. Newsom, et al. 2018. "Minimizing Polymerase Biases in Metabarcoding." *Molecular Ecology Resources* 18: 927–939. https://doi.org/10.1111/1755-0998.12895.

Niemeyer, B., L. S. Epp, K. R. Stoof-Leichsenring, L. A. Pestryakova, and U. Herzschuh. 2017. "A Comparison of Sedimentary DNA and Pollen From Lake Sediments in Recording Vegetation Composition at the Siberian Treeline." *Molecular Ecology Resources* 17, no. 6: e46–e62.

Often, A., and T. Alm. 1996. "Kryssliste for Nord-Norge (Checklist for Northern Norway)." *Polarflokken* 20, no. 2: 165–168.

Oskolkov, N., C. Jin, S. L. Clinton, et al. 2025. "Disinfecting Eukaryotic Reference Genomes to Improve Taxonomic Inference From Ancient Environmental Metagenomic Data." Preprint, bioRxiv, March 19. https://doi.org/10.1101/2025.03.19.644176.

Pedersen, M. W., A. Ruter, C. Schweger, et al. 2016. "Postglacial Viability and Colonization in North America's Ice-Free Corridor." *Nature* 537: 45–49.

Revéret, A., D. P. Rijal, P. D. Heintzman, A. G. Brown, K. R. Stoof-Leichsenring, and I. G. Alsos. 2023. "Environmental DNA of Aquatic Macrophytes: The Potential for Reconstructing Past and Present Vegetation and Environments." *Freshwater Biology* 68: 1929–1950. https://doi.org/10.1111/fwb.14158.

Rijal, D. P., P. D. Heintzman, Y. Lammers, et al. 2021. "Sedimentary Ancient DNA Shows Terrestrial Plant Richness Continuously Increased Over the Holocene in Northern Fennoscandia." *Science Advances* 7, no. 31: eabf9557.

Schulte, L., N. Bernhardt, K. Stoof-Leichsenring, et al. 2021. "Hybridization Capture of Larch (Larix Mill.) Chloroplast Genomes From Sedimentary Ancient DNA Reveals Past Changes of Siberian Forest." *Molecular Ecology Resources* 21, no. 3: 801–815.

Sjögren, P., M. E. Edwards, L. Gielly, et al. 2017. "Lake Sedimentary DNA Accurately Records 20th Century Introductions of Exotic Conifers in Scotland." *New Phytologist* 213, no. 2: 929–941.

Smith, O., G. Momber, R. Bates, et al. 2015. "Sedimentary DNA From a Submerged Site Reveals Wheat in the British Isles 8000 Years Ago." *Science* 347, no. 6225: 998–1001.

Sønstebø, J. H., L. Gielly, A. K. Brysting, et al. 2010. "Using Next-Generation Sequencing for Molecular Reconstruction of Past Arctic Vegetation and Climate." *Molecular Ecology Resources* 10, no. 6: 1009–1018.

Stoof-Leichsenring, K. R., S. Huang, S. Liu, et al. 2022. "Sedimentary DNA Identifies Modern and Past Macrophyte Diversity and Its Environmental Drivers in High-Latitude and High-Elevation Lakes in Siberia and China." *Limnology and Oceanography* 67, no. 5: 1126–1141.

Taberlet, P., E. Coissac, F. Pompanon, et al. 2007. "Power and Limitations of the Chloroplast trnL (UAA) Intron for Plant DNA Barcoding." *Nucleic Acids Research* 35, no. 3: e14.

Von Eggers, J., M.-E. Monchamp, E. Capo, C. Giguet-Covex, T. Spanbauer, and P. Heintzman. 2024. "Inventory of Ancient Environmental DNA From Sedimentary Archives: Locations, Methods, and Target Taxa V2." https://zenodo.org/records/13761348.

Wang, Y., T. S. Korneliussen, L. E. Holman, A. Manica, and M. W. Pedersen. 2022. "*ngs*LCA—A Toolkit for Fast and Flexible Lowest Common Ancestor Inference and Taxonomic Profiling of Metagenomic Data." *Methods in Ecology and Evolution* 13, no. 12: 2699–2708.

Wang, Y., M. W. Pedersen, I. G. Alsos, et al. 2021. "Late Quaternary Dynamics of Arctic Biota From Ancient Environmental Genomics." *Nature* 600: 86–92.

Weiß, C. L., M. Dannemann, K. Prüfer, and H. A. Burbano. 2015. "Contesting the Presence of Wheat in the British Isles 8,000 Years Ago by Assessing Ancient DNA Authenticity From Low-Coverage Data." *ELife* 4: e10005. https://doi.org/10.7554/eLife.10005.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** Top figure shows the number of true positive (TP), false positive (FP), and regional flora (RF) genera retained in the metabarcoding dataset at increasing read cutoffs. The bottom figure shows the proportion of FP genera in the dataset at increasing read cutoffs. The horizontal red line marks 0.05 proportion and the dashed vertical line marks the optimal cutoff threshold. **Figure S2:** Top figure shows the number of true positive (TP), false positive (FP), and regional flora (RF) genera retained in the capture probe dataset at increasing read cutoffs. The bottom figure shows the proportion of FP genera in the dataset at increasing read cutoffs. The horizontal red line marks 0.05 proportion and the dashed vertical line marks the optimal cutoff threshold. **Figure S3:** The top figure shows the number of true positive (TP), false positive (FP), and regional flora (RF) genera retained in the shotgun dataset at increasing read cutoffs. The bottom figure shows the proportion of FP genera in the dataset at increasing read cutoffs. The horizontal red line marks 0.05 proportion and the dashed vertical line marks the optimal cutoff threshold. **Figure S4:** Taxonomic richness in different lakes detected by different methods based on optimal filtering. **Figure S5:** A scatter plot of each taxon included in PhyloNorway comparing the total length of reference sequences present in the bowtie2 database used in the shotgun pipeline (originally produced in Wang et al. 2021) and the number of reads assigned to that taxon across the 18 lakes. Note that both *x*- and *y*-axes are scaled logarithmically. The color of each point represents the average length of all contigs present in the reference dataset. A linear model of Total_reads ~ Total_reference_bp produces an $R^2 = 0.06$. **Figure S6:** The proportion of 2 m vegetation survey taxa with a DNA match at the species-level grouped by DNA method detected by and ranked abundance. The DNA data shown are those done after the final optimal filtering. **Table S1:** Metadata for each site surface samples were collected from. **Table S2:** Vegetation survey data for each site. **Table S3:** Numbers of total reads sequenced for each sample for each workflow. **Table S4:** The taxa identified in negative controls passing the optimal filtering criteria described in the manuscript. **Table S5:** The total dataset of all taxa identified for all three workflows. **Table S6:** Results from ordinal logistic regressions using the proportion of reads to predict plant abundance categories at the species and at the genus level (matches, e.g., species and genus level matches between each DNA method and the vegetation surveys). The data are filtered based on the optimal approach. The ranked abundances are only for the <2 m vegetation surveys. In addition to raw read proportions, three different transformations of the data were evaluated: log, square root and double root. For each model, coefficients are shown with standard error (Est $\pm$ SE), *p*-values and ΔAIC, the difference in AIC to the model with the lowest AIC value. Models with the lowest AIC values and a significant relationship between read proportion and abundance category are highlighted in bold. Models with $\Delta AIC < 2$ that are considered equally suitable are highlighted in italics. "f" indicates that models with a flexible threshold were preferred over models with an equidistant threshold "e" between the categories. **Table S7:** Results from ordinal logistic regressions using the read numbers to predict plant abundance categories at the species and at the genus level (*matches*, e.g., *species and genus level matches between each DNA method and the vegetation surveys*). The data are filtered based on the optimal approach. The ranked abundances are only for the <2 m vegetation surveys. In addition to raw read numbers, three different transformations of the data were evaluated: log, square root and double root. For each model, coefficients are shown with standard error (Est $\pm$ SE), Sp-values and ΔAIC, the difference in AIC to the model with the lowest AIC value. Models with the lowest AIC values and a significant relationship between read proportion and abundance category are highlighted in bold. Models with $\Delta AIC < 2$ that are considered equally suitable are highlighted in italics. "f" indicates that models with a flexible threshold were preferred over models with an equidistant threshold "e" between the categories.