Metabarcoding of modern soil DNA gives a highly local vegetation signal in Svalbard tundra

Edwards, M.E.[1]*, Alsos, I.G. [2], Yoccoz, N.[3], Coissac E.[4,5], Goslar, T.[6,7], Gielly, L.[4,5], Haile, J. [8], Langdon, C.T.[1], Tribsch, A.[9], Binney, H.A.[1], von Stedingk, H.[1,10], Taberlet, P.[4,5]

[1*] Geography & Environment, University of Southampton, University Road, Southampton, SO17 1BJ, UK

[2] Tromsø Museum, University of Tromsø – The Arctic University of Norway, NO-9037 Tromsø, Norway

[3] Department of Arctic and Marine Biology, University of Tromsø – The Arctic University of Norway, NO-9037 Tromsø, Norway

[4]Laboratoire d'Ecologie Alpine, Université Grenoble Alpes, F-38000 Grenoble, France

[5] Laboratoire d'Ecologie Alpine, CNRS, F-38000 Grenoble, France

[6] Faculty of Physics, Adam Mickiewicz University, Umultowska 85, 61-614 Poznań, Poland

[7] Poznan Radiocarbon Laboratory, Rubież 46, 61-612 101 Poznań, Poland

[8] Centre for Geogenetics, Statens Naturhistoriske Museum, Øster Farimagsgade 2C, 1352 København K, Denmark

[9] Department of Biosciences, University of Salzburg, Hellbrunnerstr. 34, 5020 Salzburg, Austria

[10]Present address: FSC Sweden, S:t Olofsgatan 18, 753 11 Uppsala, Sweden

*Corresponding author.

Abstract

Environmental DNA retrieved from modern soils (eDNA) and late-Quaternary palaeosols and sediments (aDNA and sedaDNA) promises insight into the composition of present and past terrestrial biotic communities, but few studies address the spatial relationship between recovered eDNA and contributing organisms. Svalbard's vascular plant flora is well known, and a cold climate enhances preservation of eDNA in soils. Thus, Svalbard plant communities are excellent systems for addressing the representation of plant eDNA in soil samples. In two valleys in the inner fjord region of Spitsbergen, we carried out detailed vegetation surveys of circular plots up to a 4-m radius. One or three near-surface soil samples from each plot were used for extraction and metabarcoding of soil-derived eDNA. Use of PCR replicates and appropriate filtering, plus a relevant reference metabarcode catalogue, provided taxon lists that reflected the local flora. There was high concordance between taxa recorded in plot vegetation and those in the eDNA, but floristic diversity was under-sampled, even at the scale of a 1-m radius plot. Most detected taxa grew within <0.5-1.0 m of the sampling point. Taxa present in vegetation but not in eDNA tended to occur further from the sampling point, and most had above-ground cover of <5%. Soil-derived eDNA provides a highly local floristic signal, and this spatial constraint should be considered in sampling designs. For palaeoecological or archaeological studies, multiple samples from a given soil horizon that are spatially distributed across the area of interest are likely to provide the most complete picture of species presence.

KEY WORDS: aDNA, eDNA, metabarcoding, soil, vascular plants, Svalbard.

Introduction

The analysis of environmental DNA (eDNA) derived from soil and a range of sediment types is a rapidly growing area of research with applications in contemporary biological monitoring, archaeology and palaeoecology (Brown and Barnes, 2015; Pedersen et al., 2015; Parducci et al., 2017; Zimmerman et al., 2017a, b). Use of eDNA may help reveal information that hitherto has been inaccessible (Giguet-Covex et al., 2014; Rawlence et al., 2014). Furthermore, it is a cost-effective method that has minimal impact on the environment during sampling (Yoccoz et al., 2012; Thomsen and Willerslev, 2015).

Understanding the signal of DNA in soils and other terrestrial sediments

For both modern and ancient DNA studies, the local source area, source materials, likely biases, and possibilities for contamination by material transported over distance into the sampling area must be considered in the interpretation of results (Thomsen and Willerslev, 2015; Barnes and Turner, 2016; Alsos et al., 2018). To date, rather few studies have directly addressed these issues, although several studies on plant DNA derived from late-Quaternary loess or lake sediment (sedaDNA) make a comparison with the traditional proxies: pollen and plant macrofossils (e.g., Jørgensen et al., 2012; Pedersen et al., 2013; Parducci et al., 2013, 2015; Alsos et al., 2016; Sjögren et al., 2017; Zimmermann et al., 2017 a, b). The diversity recorded and the degree of overlap among proxies depend on several factors: the available reference collection/DNA reference library, site-specific characteristics such as floristic diversity and characteristics of sediment deposition, and the achievable taxonomic resolution. Comparisons are also influenced by the sums of pollen or macrofossils counted (Birks and Birks, 2016). This comparative approach provides an incomplete picture of the potential of DNA as a form of proxy data. Direct comparisons of modern DNA content of soil samples against modern vegetation have been carried out by Yoccoz et al. (2012) and Taberlet et al. (2012). Yoccoz et al. (2012) studied tundra sites near the boreal forest limit in north Norway. They reported two key results. First, for major functional groups (woody shrubs, graminoids, and forbs), vegetation biomass in 15x15-m vegetation stands and DNA read numbers retrieved from small (~10-ml) soil samples showed robust quantitative relationships. Second, for two floristically distinct plant communities (heath and meadow), DNA assemblages mirrored modern plant assemblages and distinguished them reliably. The study had limitations: the relatively large size of the sampled stands precluded an accurate assessment of the source area for the DNA, and, as only one polymerase chain reaction (PCR) was

carried out for each sample, PCR bias, particularly in relation to uncommon sequences, was not assessed.

For practical reasons, late-Quaternary soil/sediment samples and those used in contemporary DNA-based scoping surveys of local floras and faunas are small (100-1000 ml; Taberlet et al., 2013; Willerslev et al., 2014; Pansu et al., 2015; but see Taberlet et al., 2012). This could make them particularly susceptible to sampling bias, and a clear understanding of how deposits formed and of subsequent processes that might affect their age or DNA content (e.g., bioturbation) is necessary to avoid misinterpretation of results. The controversy over recent DNA-based archaeological findings (Smith et al., 2015; Bennett, 2015; Weiß et al., 2015) underlines this need. For studies of past biotic communities, the current paucity of direct comparisons between proxy (here eDNA) and vegetation reduces rigour compared with palynology, for example, where a large body of data can be used to demonstrate pollen-vegetation relations in different environments (see Sugita, 2007a, b; Seppä, 2013). Thus, further quantification of the spatial catchment for plant DNA retrieved from a range of late-Quaternary deposits is needed. Here, we address this need via a detailed comparison of soil-derived eDNA with the surrounding vegetation in a tundra ecosystem.

Palaeoecological studies in cold-climate regions

Arctic tundra vegetation often features subtle compositional mosaics that reflect slight differences in site factors, and the slow growth and tendency for vegetative spread of perennial taxa can lead to temporal stability of patches (Bliss 1988). Pollen and plant macrofossil studies have contributed much of what is known of past vegetation and flora. Low pollen production of many entomophilous or autogamous taxa leads to over-representation of anemophilous taxa, particularly shrubs, and pollen derived from long-distance transport may be prominent (e.g., Anderson and Brubaker, 1986; Fréchette et al., 2008). Taxonomic resolution can be limited because pollen of several key families, with some exceptions, is not easily differentiated into genus or species via pollen morphology (e.g., Poaceae, Cyperaceae, Caryophyllaceae, Brassicaceae; Faegri and Iversen, 1989). Pollen samples can nevertheless reveal useful information about tundra vegetation composition at the landscape scale (e.g., Ritchie 1974; Lamb and Edwards, 1988; Hicks 2001). Plant macrofossils, when present, can provide a more floristically resolved record of tundra vegetation (e.g. Birks, 1991, 2003; Bigelow, 2013; Kienast, 2013), but they are not always well preserved or consistently present. Notably, successful sedaDNA records of past flora

(Willerslev et al., 2003, 2014; Zimmerman et al., 2017b) and fauna (Cooper et al., 2015; Graham et al., 2016) have been retrieved from high-latitude sites, which are often permafrost-affected and benefit from the superior preservation of fossil material in frozen settings. Overall plant-taxonomic richness can be as high or even higher than that of detailed pollen and macrofossil studies (Sonstebø et al., 2010; Willerslev et al., 2014; Zimmerman et al., 2017b).

Study design

In this study, we partially address PCR bias by using multiple PCR replicates per sample, and we use what we consider to be more appropriate, fine-scale but highly detailed survey plots for vegetation (4-m radius), compared with Yoccoz et al. (2012). We focus on several key features of the soil-derived DNA records that are important for ensuring plausible interpretation of the data. First, the vegetation source area is critical for understanding the spatial information contained in reconstructions based on modern eDNA or on sedaDNA: for a given species, how does the distance from a sampling point influence the probability of detecting that species? Second, does the probability of detection of a taxon by DNA vary with its abundance in the vegetation? Third, we need to know the accuracy with which the DNA flora reflects the modern floristic composition in the source area: do DNA taxa and vegetation taxa match, or are there many "false positives" (DNA taxa that are not recorded in the vegetation) or "false negatives" (DNA does not record species present in the vegetation)? Fourth, do sampling design and data filtering influence detection success?

Based on analogy with studies of plant macrofossils (e.g., Zazula et al., 2006; Birks, 2007), we developed a taphonomic model to test the spatial representation of soil-derived eDNA (hereafter called DNA) records (Figure S1). We expect taxa in local vegetation (within a few metres of the sampling location) and DNA to be tightly correlated if local sources (fine roots, rootlets, litter from above-ground plant parts) predominate. In this case, the DNA from different plots should reflect variation in the tundra mosaic across distances of tens to hundreds of metres. However, both lateral transport and vertical mixing of soil might introduce DNA of taxa not currently present near the sampling site and possibly generate a reservoir effect, for example, eroded and transported older DNA contaminating recent material (Haile et al. 2007). DNA is likely to be complexed with soil particles that move down-slope (Pedersen et al., 2015), or material may be displaced vertically if soil/sediments are mixed by bioturbation or frost heave. To test the latter possibility, we radiocarbon-dated

a range of sampled materials. At broader scales, plant parts transported by animals or wind from afar (Glaser, 1981) may introduce taxa not found in the study area. Also, concerns have been raised that long-distance transport of pollen, especially gymnosperm pollen, which may contain plastids/chloroplasts (Mogensen, 1996), contributes to the DNA signal (Parducci, 2012a; Birks et al., 2012; Parducci et al., 2012b). Results of Sjögren et al. (2017) and Zimmerman et al. (2017a) suggest such contamination is minimal; nevertheless, we also obtained pollen counts from the samples taken for DNA analysis.

Materials and methods

Study area

The high-arctic Svalbard archipelago largely lies between 77 and 80°N. It has a small, intensively studied arctic flora (Alsos et al., 2017). All common species in the known vascular plant flora are represented in the Ecochange metabarcoding catalogue (Sønstebø et al., 2010; Willerslev et al., 2014); of 52 species observed in this study only *Saxifraga svalbardensis* is not included. This provides a robust system for estimating the effectiveness and scope of the DNA record from soil in relation to modern vegetation. We studied vegetation in two valleys on Spitsbergen, the largest island in the Svalbard Archipelago: Endalen and Colesdalen (Figures 1 and 2).

Svalbard has an arctic climate that is tempered by the North Atlantic Drift. Average January temperature in the Longyearbyen area (Figure 1) is -11.7°C and that for July is +5.2°C. Average annual precipitation in the Longyearbyen area is 191 mm (Førland et al. 2011). Some inner fjord areas experience atypically warm climates, as is the case with our study sites (Alsos et al., 2004, Engelskjøn et al., 2003), which lie in arctic subzone C (mid-arctic tundra zone; Walker et al., 2005). Endalen and Colesdalen support relatively lush vegetation that includes the thermophilic taxa *Betula nana* ssp. *tundrarum* (both sites), *Vaccinium uliginosum* and *Euphrasia wettsteinii* (Colesdalen). Dominants that have high overall cover on the valley slopes include *Cassiope tetragona* (Endalen only), *Salix polaris, Dryas octopetala, Equisetum arvense,* bryophytes, and grasses such as *Alopecurus borealis* and *Poa arctica*. Lower-lying, waterlogged areas are characterized by *Dupontia fisheri*. Soils are generally shallow, depths varying from a few cm to >10 cm, and comprise an organic upper horizon overlying poorly weathered parent material; in the case where there is a relatively thick organic O-horizon, this tends to be dominated by moss remains. Active-layer depths are

typically shallow but can reach up to 0.5 m on the lower slopes of Colesdalen (pers. obs. 2007). Cryoturbation is widespread.

The bedrock geology of both valleys is dominated by sedimentary formations, including coal beds. Both valleys have been the site of past coal mining, and coal fragments occur on the ground near to mining installations (such as aerial cable lines). The presence of mined coal at the surface poses difficulties for radiocarbon dating (see below), as coal from workings or native coal may be present in the soil or subsoil.

Field sampling

We established vegetation plots along a mid-slope contour to avoid i) debris slides and other mass wasting features near the steep valley wall (Endalen), ii) potentially disturbed floodplain surfaces (Endalen), and iii) flat areas in the valley bottom with mire vegetation (Colesdalen). In both valleys the plots were located on the south-facing valley side (Figure 1). Plots were spaced ca. 100-200 m apart but placed subjectively, as we wished some plots to include less common elements of the tundra mosaic that include, for example, *Betula* or *Vaccinium*. Eight plots in Colesdalen and nine in Endalen were studied. At the centre point of each plot we set a 0.5 x 0.5-m quadrat. Percent cover of vascular plants, bryophytes, lichen, bare ground and rock was assessed visually. We then extended tapes to 4.0 m from the central point in eight equally-spaced directions and visually assessed cover in each of 32 segments defined by the tapes and by 1-m increments from the centre, as measured on the tapes (Figure 3). The central plot overlapped the innermost segments; its data were used separately in some analyses.

The work was carried out over five days in August 2007. As up to five botanists estimated the cover, we cross-checked estimates to minimize differences between observers. We also created rough maps of the main vegetation mosaic by walking outwards ca. 50 m in different directions from the intensively measured plots making observations of the communities present at a set of points (located by GPS and transferred to a hand-drawn map). Inevitably, given the number of sectors surveyed and the limited time available, a few sectors were missed at both sites (three at Endalen and one at Colesdalen), but given the large number of sectors amalgamated to produce presence-absence and cover values, these omissions are unlikely to affect the conclusions.

After the vegetation had been described we sampled soil from the central plot for DNA analysis. In Colesdalen, we cleared away vegetation, then took three soil samples, about 15 cm apart, using either factory-sealed 50-ml plastic tubes driven into the ground (and subsequently capped) or (when the substrate was too hard to push in the tube) a sharp trowel washed with bleach solution prior to taking each sample, with the extracted 5-10 cm column of soil sealed into previously unopened plastic bags (i.e., double bagging). In Endalen, the O horizon at most plots was thick (>5 cm) and spongy. We therefore changed our sampling strategy, using a spade to dig a monolith (ca 0.25 x 0.25 m area) that included the surface vegetation, underlying peaty material, and in some cases the mineral substrate beneath. Monoliths were wrapped securely in clean plastic bags, taped and returned to the laboratory. One sample was collected from each monolith, except for Endalen 3, where the sample was lost.

Radiocarbon dating

Two types of material were isolated for AMS radiocarbon dating at the Poznan Radiocarbon Laboratory, Poland: plant macrofossils and the more decomposed soil matrix. Samples came from the residual material used for DNA extraction (Colesdalen; 10 dates) or a subsample of material taken directly adjacent to the DNA subsample (Endalen monoliths; 31 dates). Protocols followed Brock et al. (2010). Samples were dated using accelerator mass spectrometry. For near-modern samples, calibration was via comparison with post-bomb atmospheric ^{14}C concentrations (Reimer et al., 2004, Hua et al., 2013). We did not calibrate older dates, as the purpose of the dating was to establish whether samples were a few decades or many centuries old (or older). Care was taken to ensure the chances of contamination by coal were minimized by repeated washing of macrofossils and the sieving out of coal fragments from the soil matrix. For the macrofossil samples, attention was focused on excluding rootlets and taking fragments of leaves, stems, and twigs. In some cases, we amalgamated items because of their individual low weights. Some lower samples from Endalen monoliths were also dated to assess whether there were coherent depth-age relationships in the monolith profiles.

DNA extraction and amplification

Each of the three soil samples per plot from Colesdalen were treated as separate units for DNA extraction. The Endalen monoliths were first unwrapped, split vertically down the middle with a knife cleaned with bleach, and then subsampled, the sample for DNA extraction being taken where the material first changed from uncompacted plants and plant remains to more compacted and humified material. Intra- and extracellular DNA was extracted from approximately 10 ml of material using a PowerMax soil kit as described in Willerslev et al. (2014). The short and variable P6 loop region of the chloroplast *trn*L (UAA) intron (Taberlet et al., 2007) was used as the diagnostic marker, amplified with the following universal primers:

"g" (5'-GGGCAATCCTGAGCCAA-3'), and

"h" (5'-CCATTGAGTCTCTGCACCTATC-3'), as described in Willerslev et al. (2014). Four PCR replicates were done for Endalen and up to 12 for Colesdalen, Purified products were sequenced 2 x 108-bp paired-end reads using an Illumina GA IIx platform. One PCR replicate of each sample from Colesdalen was also included in two initial test runs: one on an Illumina GA platform (8 for Colesdalen plot 5) and one on a Roche Genome Sequencer FLX platform. To check for contamination, the final sample set included 19 extraction blanks, plus five PCR negative controls for each Illumina run.

Sequence analyses and filtering

The DNA approach used to date in most studies of past vegetation in northern regions uses selected short sequences, typically 20 - 150 bp, from either the chloroplast or nuclear genomes; longer sequences tend not to be preserved in older sediments (see Taberlet et al., 2007; Valentini et al., 2009). More recently, attempts at shotgun sequencing of whole genomes have identified taxa from sediment samples (*e.g.*, Pedersen et al., 2016). In this study, we used detailed metabarcode catalogues for arctic and boreal taxa developed by the Ecochange consortium (Sønstebø et al., 2010; Willerslev et al., 2014).

The total analysed reads for both sites was ~5.7M. About ~5M were from Colesdalen, from which there were far more samples. Initial filtering using Obitools and the arctic-boreal reference library follows Willerslev et al. (2014). All 256 PCR replicates had more than 1000 reads and were kept initially. We then deleted sequences shorter than 10 nucleotides. We retained 237 replicates, which represent 2-4 and 3-12 PCR and sequencing replicates for Endalen and Colesdalen, respectively (Table S1). Thereafter, we standardized data to 1000 reads per sample using rarefaction to account for higher reads that occurred in one

preliminary run. Multiple PCR replicates can be used as a filter to decide whether a molecular taxonomic unit (MOTU) should be retained in the dataset (i.e., presence in a minimum number of replicates); this has been shown to be effective at removing false positives (Ficetola et al., 2014; Alsos et al., 2016). We used a threshold to define "true" presence: the sequence had to occur with a minimum of 10 reads in more than 50% of the available replicates. We consider this a "strict" threshold, one with a high potential to exclude false positives, but which may also exclude some true positives.

Relating vegetation to MOTUs

The DNA sequences are placed into MOTUs, which vary in taxonomic resolution from sub-family to species; many MOTUs contain several taxa. Because the flora of Svalbard is well documented, we used biogeographical knowledge and parsimony to assign MOTUs to extant species (Table 1). For example, the MOTU *Dryas* contains seven species, but we assume it represents *D. octopetala*, as this is the only *Dryas* species occurring in Svalbard. Because some grass taxa were identified at the 98% level, all grass species were placed the MOTU Pooideae. When at least one grass species was present in a plot, we considered it a match.

To analyse the relationship between observed vegetation (viewed here as "true" presence and the comparator for DNA) and the recovered DNA, we used both abundance and presence-absence data for the vegetation and presence-absence data for the DNA. (Endalen plot 3 has no DNA data, so vegetation plot 3 was omitted from comparison with the DNA.) When using the data from both sites for comparison with vegetation, we combined the three Colesdalen samples, but we also examined the effect of multiple replicates at Colesdalen separately.

Pollen analysis

Sub-samples were taken from residual material not used for DNA/dating (Colesdalen, plots 1-6, 3 replicates), or from material extracted adjacent to a monolith DNA sample (Endalen, plots 1-8). Small volumes (2-5 ml) were processed for pollen using conventional techniques (Berglund and Ralska-Jasiewiczowa, 1986). Counting was done under x400 magnification with x1000 high-power capability. Pollen concentrations were low and multiple slides were counted. Where feasible, pollen sums were at least 100 grains. Pollen diagrams were created using TILIA software (Grimm, 1990).

Results

Vegetation

We identified a total of 52 species of vascular plant, 43 in Colesdalen and 24 in Endalen; these correspond to 32 taxa that potentially could be identified in the DNA analyses, accounting for sequence-sharing among species. Complete data on species abundances for each segment of each circular plot for Colesdalen and Endalen are available from the corresponding author. The surveyed area expands relative to the square of the distance from the plot centre, and the cumulative species-area curves rise quite steeply from the central plot, approaching an asymptote by rings 3 and 4, indicating the plot data sampled the local flora effectively (Endalen, Figure 4). An ordination (correspondence analysis in R package ADE4; not shown) confirms that many plots had similar composition, being dominated by grass species and *Salix polaris*, plus forbs, while several plots containing uncommon dwarf shrubs (*Betula nana*, *Vaccinium uliginosum*, and *Cassiope tetragona*) stood out as compositionally different.

Radiocarbon dating of soil samples

The radiocarbon dates from both valleys fell into two groups with markedly different ages. In Colesdalen, macrofossils selected from the soil samples all appeared modern (i.e., with carbon largely derived from the post-bomb atmosphere), while only one soil sample (plot 5 #1) gave a modern date (Table S2). Soil samples from plots 4 and 6 (~7700 ^{14}C yr BP), which are closer to the old Colesdalen mine workings, had markedly older ages than samples from plots 7 and 8 (~2300 ^{14}C yr BP). In Endalen, samples were taken from directly underneath the surface vegetation mat to up to 8 cm further down the organic profile of the soil monoliths. Of 31 ages obtained for plant macrofossils all but one (1380 ^{14}C yr BP) were <400 years old, many being modern. Most samples that were single fragments gave a small scatter of ages with many of them in the range 110-115 pMC, which corresponds to atmospheric ^{14}C concentrations in the years 1993-1997. Subsamples consisting of several fragments (mostly leaves), gave a wider scatter of ages. In contrast, ages of the bulk organic material varied between ~1930 and ~14,630 ^{14}C yr BP. For upper samples (taken 0-4 cm from base of plant material) most ages range from ~2000-7000 ^{14}C yr BP. Three deeper samples (4-8 cm) and one at 3 cm had ages >10,000 ^{14}C yr BP.

DNA

The raw sequence data are either already available on DRYAD (see Willerslev et al., 2014) or will be submitted to DRYAD on publication. After filtering, 60 unique sequences with a >98% match to the database were identified, 53 of which had a 100% match. Some sequences were assigned to the same taxon, resulting in 38 different taxa. Of these, we excluded 18 bryophytes and three exotic taxa that were filtered out: *Rumex* (2 samples from Endalen), *Pinus and Trientalis* (one sample each from Colesdalen), leaving 17 different DNA taxa.

We used the vegetation data and the metabarcode databases to align observed species and MOTUs. Creating molecular taxa directly comparable with taxa recorded in the vegetation (Table 1) yielded 30 potentially retrievable MOTU's, some representing multiple species. These correspond to 51 of the 52 observed vegetation species (exception: *Saxifraga svalbardensis*). The 17 MOTUs from DNA soil samples (Figure 5) corresponded to 37 species identified in the vegetation plots (Table 1), giving a maximum identification potential of 71%. The converse is that 29% of vegetation taxa have no matching MOTU or MOTU group in the DNA data, meaning that these taxa are "silent" in the DNA record. It should be noted that because some vegetation species are lumped within a MOTU (e.g., all grass species into Pooideae), but their identity is known in this study, the potential level of identification of taxa in an unknown flora would likely be lower than the level achieved here—at least if our rigorous filtering protocol were used.

Representation of vegetation-plot species by DNA samples

A set of straightforward observations shows that the DNA data are floristically accurate, that the 4-m plots were moderately-to-strongly under-sampled by DNA, and that soil DNA reflects highly local vegetation. The following observations are based on MOTUs that are aligned with both observed plant species in the vegetation and retrieved DNA.

i) DNA reflected the floristic composition of the vegetation sampling plot accurately. With one exception (see below) all taxa observed in the DNA that remained after filtering were also present in the vegetation of that plot (Figure 5).

ii) *Cardamine bellidifolia* was present in the DNA in Colesdalen 3 but was not recorded in the plot vegetation, although it is present in other Colesdalen plots. It is likely that it was overlooked in the vegetation survey, as it is only a few centimetres tall and a short-lived

plant. Thus, this single example of a "false positive" in the DNA most likely reflects a false negative in the vegetation surveys.

iii) The DNA data are variably effective at reflecting species presence. The dominant dwarf shrub *Salix* was always detected, as was group of common and widespread taxa (e.g., *Bistorta viviparum, Equisetum*). The rare dwarf shrubs *Betula nana* and *Vaccinium uliginosum* were also identified by DNA in the single plots where they were present and dominant, but not all dominant species were identified by DNA. *Cassiope tetragona* was detected only once, yet it occurred in nine plots and in several of those within a one-metre radius of the centre. In both Endalen 6 and 7 it occurred at 50% cover in the centre plot, but it was only detected in Endalen 7. Other species that were present in vegetation but not detected by DNA in most or all plots in which they were present tended to be relatively infrequent and/or have low abundance in the plot. Most were small forbs (e.g., *Euphrasia wettsteinii, Draba* spp., *Saxifraga* spp.), but more robust plants such as *Alopecurus borealis* were strongly under-represented, and *Juncus* was not detected.

iv) In all plots, species richness in the 4-m plot was underestimated by DNA. Endalen vegetation plots had 15-20 recorded MOTUs, whereas the DNA MOTU count was 4-7. The central 0.5x0.5m quadrats, however, contained 5-8 taxa and thus had similar richness to the retrieved DNA (Figure 4). In both valleys, with one exception (*Cardamine*, mentioned above), taxa that were recorded by DNA grew within 3 m of the sampling point, and 77% and 97% grew within the central plot and 1-m radius, respectively (see blue records in Figure 5). Among the plant species that were not detected, 52% were >1 m away from the DNA sampling point (Figure 6). These data underline how highly local the detection range for soil DNA appears to be.

v) Taking the central quadrat alone, the higher a taxon's relative abundance in the vegetation, the more likely it was to be detected in the DNA. All taxa with 4% or higher abundance in the vegetation were detected, with the exception of *Cassiope* (Figure 6).

v) The richer flora of Colesdalen (plot richness 20-29) was reflected by more variable MOTU counts per sample than at Endalen. At Colesdalen, the collection of three soil samples per plot tested the effectiveness of closely-spaced repeated DNA samples in increasing the DNA taxon count. A strong predictor of observing a taxon in all three replicates was that it was

present in the centre (50x50 cm) vegetation quadrat and/or in at least 75% of the 41 sectors in a vegetation plot. Species with lower abundances in the vegetation tended to be present in only one or two soil replicates. One-sample MOTU richness was 2-8, and three-sample richness 4-10. Adding samples increased the taxon count from 0% (the same taxa in all three samples) to 150% (increase from four to ten taxa).

Pollen analysis

Pollen concentrations were generally low, and these were reflected in low pollen counts (51-120 grains after several slides counted). After amalgamating taxa that are taxonomically nested, the pollen flora from the two sites contains 16 vascular plant taxa attributable to Svalbard natives, plus *Sphagnum* (Figures S2a and S2b). The most abundant taxa are *Salix*, Poaceae and Caryophyllaceae; these show variation in abundance among plots. Endalen plot 1 is dominated by *Rumex-Oxyria* and plot 8 by *Salix*, but otherwise both sets of samples show relatively little variation. Five non-native taxa are recorded: *Pinus, Juniperus, Sorbus*-type, *Sparganium-Typha* (not shown) and *Lycopodium annotinum,* and while the majority of *Betula* grains can be attributed to *B. nana*, others (i.e., *Betula/Corylus* type) may be tree-*Betula* and thus also non-native.

DISCUSSION

The data provide useful insights into the source area for DNA in a soil sample, the role of distance and abundance in the likelihood of observing a taxon in the DNA, and the accuracy with which the floristic composition of the DNA sample reflects that of the modern plant community. The chosen method of DNA data processing (strict filtering rules) has resulted in an accurate reflection of the local plant communities, remarkably effective at very fine (≤1.0-m) scales but incomplete at larger ones. Our plots were located on gentle slopes which might have facilitated the transport of DNA in snow-melt and rainfall runoff and through-flow, but we did not detect out-of-plot taxa, suggesting the main source of DNA in soil is highly local above- and below-ground biomass (see Figure S1). Similar conclusions were reached by Yoccoz et al. (2012) in Norway and Taberlet et al. (2013) from soil samples in a tropical rain-forest setting.

Chronology: age and possible sources of the DNA

While we avoided sites with active cryoturbation, past frost-heaving may have led to vertical mixing of the tundra soil, and this concern led to our collection of a large set of radiocarbon

dates, dating discrete pieces of plant material ("macrofossils") as well as the soil matrix (Table S2). Dates on the soil matrix range from near-modern to >14,000 ^{14}C yr BP; this variation cannot be fully explained by cryoturbation, as the oldest ages would pre-date the deglaciation of the area. Rather, given the prevalence of coal in the local environment, we conclude that even though the matrix samples were screened for coal pieces, very fine coal particles contaminate the soil in both valleys, meaning soil-matrix radiocarbon dates are unreliable. Circumstantial support comes from the pattern of older ages: soil radiocarbon ages are older nearer to the abandoned coal-processing installation in Colesdalen, and, while ages of Endalen samples tend to increase with depth, some are implausibly old. In contrast, the macrofossils are young, usually only a few decades old. They were presumably incorporated from the surface litter layer into the topmost part of the soil profile from recently living and contemporary plants. As soil components subject to decomposition, they are likely to be a key source of the retrieved DNA.

DNA-vegetation relationships

The floristic composition of the DNA reflects that of local vegetation accurately—and by "local" we are referring to an extremely small effective sampling area. Taxa that grow close to the sampling point (within 0.25 m) and those that have higher cover values, particularly high cover near the plot centre (i.e., within a 1-m radius), are more likely to be detected in the DNA than other taxa (Figure 6). Not unexpectedly, cover dominants (e.g., *Salix polaris, Bistorta vivipara*, *Equisetum*, grasses) dominate the DNA signal. These taxa likely contribute to the strong link between DNA detection and high abundance in vegetation estimated for the central quadrats (Figures 5 and 6). The only exception is *Cassiope*, which is also poorly represented in lake sediments from Svalbard and North Norway (Alsos et al., 2016; 2018). The reason may be a poly-T region of 14-16 bases present in the barcoding DNA sequence, which is likely to cause PCR and sequencing problems.

If the results of this study are representative of soil DNA (at the small, effective sampling scale of ≤1-m radius), the overall detection rate is extremely good. For the 0.5 x 0.5-m quadrat, all taxa with 4% or more cover in the vegetation are detected, except *Cassiope* (see above). We detect more than 60% of taxa that have only 2% cover in the plot, and nearly 50% of taxa with only 0.5-1% cover. The proportions of all observed plant taxa detected in the DNA in our soil samples (73% and 55% for 0.5- and 1-m plots, respectively) are considerably higher than those found in a similar study focused on plant DNA in lake

sediments (31%, Alsos et al. 2018). In the case of lakes, the catchment for DNA is far larger, proximity to the sampling point is lower, and taphonomic pathways are, presumably, more complex and variable.

Other considerations

Other factors affect the outcomes of eDNA and aDNA studies: the effectiveness of the sequence library or database consulted for identification of sequences, the degree of taxonomic resolution possible with a given flora, and the bioinformatic filters applied to the data. While we grouped the grass species into Pooideae (see above), we kept five Saxifragaceae species as separate MOTUs. The grasses were abundant in the vegetation and the DNA, but all species counted for only one MOTU, while the saxifrages, uncommon in the vegetation (only one occurrence of 1% in a central plot), were absent in the DNA. This means that, on the one hand, we may have encouraged a numerical bias towards not finding taxa in the DNA; on the other hand, for Svalbard, a rich library is available, the flora is limited, and thus the level of identification is high, and accurate. With an unknown fossil flora, multi-species MOTUs could not be related to individual species with such certainty. The interpretation of fossil (aDNA) data is thus more challenging, requiring an appreciation of the effects of filtering and the variable resolution in different taxonomic groups, plus biogeographic knowledge that can be brought to bear on the final dataset to identify residual false positives.

The adoption of multiple PCR replicates facilitates a further level of filtering, and in recent studies we have shown that our current filtering standards exclude almost all contaminants (e.g., Alsos et al. 2016). Filtering may, however, remove taxa that are almost certainly valid. Notably, although low levels of *Betula* pollen were frequently found in plots without shrub birch, *Betula* DNA does not reflect this. *Pinus* pollen was also present in several samples, but *Pinus* pollen was not observed in the sample that contained *Pinus* DNA (Colesdalen 3). Furthermore, in this study, *Pinus* was identified as a contaminant and filtered out of the DNA dataset. Thus, we have no reason to think that the presence of pollen of non-local taxa gives rise to a DNA signal. This finding is supported by the results of Sjögren et al. (2017), who studied Scottish lake sediments with higher pollen concentrations than those from Svalbard. It is probable that for most deposits investigated by DNA analysis, pollen contamination is not a source of error.

Multiple soil samples can be more effective at accounting for the total diversity of plant taxa than a single sample (Zinger et al., 2016). In Colesdalen, we took three samples, all from the central 0.5 x 0.5 m plot. The additional samples increased the final MOTU count from most plots. Abundant species were usually detected with one soil sample, whereas less abundant ones had a higher chance of being detected with multiple samples. This pattern is reflected in the fact that Colesdalen DNA samples record most of the detected low-abundance taxa in the total DNA dataset (Figure 5, left side). On the other hand, no DNA samples, including lumped Colesdalen samples, reflected the total species richness of a 4-m vegetation plot, presumably because 4 m is too large a radius for effective detection. The small scale of successful detection has both advantages and disadvantages. For archeological studies, the localized taphonomy of DNA may help distinguish species used for specific purposes (e.g., food, bait, tools, fibres) from those merely present in the surroundings. Thus, sampling should ideally cover features such as middens and the peripheries of hearths but also points at increasing distance from the site itself. Similarly, in modern biodiversity and palaeoecological studies, to register the full diversity present in local plant communities or in target areas the optimal sampling scheme would feature many samples and cover (where practicable) the whole area of interest.

Conclusion

The environmental DNA samples in this study function rather like 1x1-m quadrats in a vegetation survey, though each soil sample has some temporal depth because the slow speed of soil formation and decomposition leads to the incorporation of biological material of different ages into surface layers (Table S3). (Fossil samples from accreting surfaces such as yedoma—frozen, ice-rich silt—would represent an even greater age span.) Information is locally precise, recording species within less than a meter of the sample point. We conclude that this is useful for understanding the fine grain of key portions of a landscape or a human occupation site. The likelihood of under-sampling diversity and the small spatial scale of the signal make this type of sample less appropriate for reconstructing regional vegetation or inferring climate parameters. The general under-sampling of floristic diversity likely to occur with single, widely spaced samples can be partly addressed by more intensive sampling.

The field of metabarcoding and environmental DNA studies is rapidly changing. The development of larger reference databases will provide more scope for a range of studies in different geographic regions. At the same time, current results suggest that while soil and sediments are promising sources of DNA, both ancient and modern, a careful and extensive sampling strategy is required to obtain the best results. Such a strategy will require more resources and effort but should ensure increasing quality and reliability of results. For DNA derived from terrestrial sediments, such as yedoma, paleosols, and archaeological sites, there is nevertheless exciting potential for detailed records of floras and of other organismal groups, which in turn should lead to a greater understanding of trophic dynamics, past ecosystem function, and the dynamics of human-ecological systems.

Acknowledgements

This study was funded by the European Commission, under the Sixth Framework Programme (EcoChange project, contract No. FP6-036866) and the Norwegian Research Council (grant No. 230617/E10). We thank Leanne Franklin-Smith for assistance in the field. The University Centre in Svalbard provided essential logistic support.

References

Alsos IG, Westergaard K, Lund L, Sandbakk BE. 2004. Floraen i Colesdalen, Svalbard. *(The flora of Colesdalen, Svalbard. In Norwegian). Blyttia* 62, 142-150.

Alsos, I.G., Arnesen, G., Sandbakk, B.E. and Elven, R. 2017. The flora of Svalbard. Available at: http://svalbardflora.no .

Alsos I.G., Sjögren P., Edwards M.E., et al. 2016. Sedimentary ancient DNA from Lake Skartjørna, Svalbard: assessing the resilience of arctic flora to Holocene climate change. *The Holocene* 26, 627–642.

Alsos, I. G., Lammers, Y., Yoccoz, N.G., et al. 2018. Plant DNA metabarcoding of lake sediments: How does it represent the contemporary vegetation? *PLoS ONE* 13(4): e0195403. doi.org/10.1371/journal.pone.0195403.

Anderson, P.M., and Brubaker, L.B. 1986. Modern Pollen Assemblages from Northern Alaska. *Review of Palaeobotany and Palynology* 46, 273-291.

Barnes, M. A. and C. R. Turner. 2016. The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics* 17, 1–17.

Bennett, K. D. 2015. Comment on "Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago". *Science* 349, 247.

Berglund, B.E. and Ralksa-Jasiewiczowa, M. 1986: Pollen analysis and pollen diagrams. In Berglund, B.E., editor, *Handbook of Holocene Palaeoecology and Palaeohydrology*. John Wiley, 455-484.

Bigelow, N.H. 2013: Plant macrofossil records: Arctic North America. In Elias S.A and Mock, C.J., editors, *Encyclopedia of Quaternary Science*, 2nd Edition, Elsevier, 746–759.

Birks, H.H. 1991. Holocene vegetational history and climatic changes in west Spitsbergen – Plant macrofossils from Skardtjørna, an Arctic lake. *The Holocene* 1, 209–218.

Birks, H.H. 2003. The importance of plant macrofossils in the reconstruction of Lateglacial vegetation and climate: Examples from Scotland, western Norway, and Minnesota, USA. *Quaternary Science Reviews* 22, 453–473.

Birks, H.H. 2007: Plant macrofossil introduction. In S.A. Elias, editor, *Encyclopedia of Quaternary Science*, Elsevier, 2266–2288.

Birks, H.H., Giesecke, T., Hewitt, G.M., Tzedakis, P.C., Bakke, J. and Birks, H.J.B. 2012. Comment on “Glacial Survival of Boreal Trees in Northern Scandinavia”. *Science* 338, 742.

Birks, H.J.B. and Birks, H.H. 2016. How have studies of ancient DNA from sediments contributed to the reconstruction of Quaternary floras? *New Phytologist* 209, 499–506

Bliss, L.C. 1988: Arctic tundra and polar desert biome. In Barbour, M. G. and W. D. Billings, editors, *North American terrestrial vegetation*. Cambridge, 1-32.

Brock, F., Higham, T., Ditchfield, P. and Ramsey, C. B. 2010. Current pretreatment methods for AMS radiocarbon dating at the Oxford Radiocarbon Accelerator Unit (ORAU). *Radiocarbon* 52, 103–112.

Brown, T.A. and Barnes, I.M. 2015. The current and future applications of ancient DNA in Quaternary science. *Journal of Quaternary Science*, 30, 144–153.

Cooper, A., Turney, C., Hughen, K.A., et al. 2015. Abrupt warming events drove late Pleistocene Holarctic megafaunal turnover. *Science* 349, 602–606.

Engelskjøn T, Lund L, Alsos IG. 2003. Twenty of the most thermophilous vascular plant species in Svalbard and their conservation state. *Polar Research* 22, 317-339.

Fægri, K. and Iversen, J. 1989: *Textbook of pollen analysis*, 4th Edition (by K. Fægri, P.E. Kaland and K. Krzywinski), John Wiley, 328 pp.

Ficetola, G.F., Pansu, J., Bonin, A., et al. 2014. Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources* 15, 543–556.

Fréchette, B., de Vernal, A., Guiot, J. et al. 2008. Methodological basis for quantitative reconstruction of air temperature and sunshine from pollen assemblages in Arctic Canada and Greenland. *Quaternary Science Reviews* 27, 1197-1216.

Førland, E. J., Benestad, R, Hanssen-Bauer, I. et al. 2011. Temperature and precipitation development at Svalbard 1900-2100. *Advances in Meteorology*. Doi 10.1155/2011/893790.

Giguet-Covex, C., Pansu, J., Arnaud, F., et al. 2014. Long livestock farming history and human landscape shaping revealed by lake sediment DNA. Nature Communications 5, Article No. 3211

Glaser, P.H. 1981. Transport and deposition of leaves and seeds on tundra – A late-glacial analog. *Arctic and Alpine Research* 13, 173–182.

Graham, R., Belmecheri, S., Choy, et al. 2016. Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 113, 9310-4.

Grimm, E.C., 1990. TILIA and TILIA*GRAPH. PC spreadsheet and graphics software for pollen data. INQUA Working Group on Data-Handling Methods. Newsletter 4, 5e7.

Haile, J., Holdaway, R., Oliver, K. 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. doi.org/10.1093/molbev/msm016.

Hicks, S. 2001. The use of annual arboreal pollen deposition values for delimiting tree-lines in the landscape and exploring models of pollen dispersal. *Review of Palaeobotany and Palynology* (Special Issue) 117, 1-29.

Hua, Q., Barbetti, M., & Rakowski, A. J. 2013. Atmospheric radiocarbon for the period 1950-2010. *Radiocarbon* 55, 2059-2072.

Jørgensen, T., Haile, J., Möller, P., et al. 2012. A comparative study of ancient sedimentary DNA, pollen and macrofossils from permafrost sediments of northern Siberia reveals long-term vegetational stability. *Molecular Ecology* 21, 1989–2003.

Kienast, F. 2013. Plant macrofossil records: Arctic Eurasia. In Elias, S.A., editor, *Encyclopedia of Quaternary Science*. 2nd Edition, Elsevier, 733–745.

Lamb, H.F. and Edwards, M.E. 1988: The Arctic. In Huntley B and Webb T, III, editors, *Vegetation History: Handbook of Vegetation Science 7*.: Kluwer Academic Publishers, 19–555.

Mogensen, H. L.1996. The hows and whys of cytoplasmic inheritance in seed plants. *American Journal of Botany* 83, 383–404.

Pansu, J., De Danieli, S., Puissant, J. 2015. Landscape-scale distribution patterns of earthworms inferred from soil DNA. *Soil Biology and Biochemistry* 83, 100-105.

Parducci, L., Jørgensen, T., Tollefsrud, et al. 2012a. Glacial survival of boreal trees in northern Scandinavia. *Science* 335, 1083–1086.

Parducci, L., Edwards, M.E., Bennett, K.D., et. al. 2012. Response to Comment on "Glacial Survival of Boreal Trees in Northern Scandinavia". *Science* 338, 742.

Parducci, L., Matetovici, I., Fontana et al. 2013. Molecular- and pollen-based vegetation analysis in lake sediments from central Scandinavia. *Molecular Ecology* 22, 3511–3524.

Parducci, L., Väliranta, M., Salonen, J.S. et al. 2015. Proxy comparison in ancient peat sediments: pollen, macrofossil and plant DNA. *Philosophical Transactions of the Royal Society B* 370, 20130382.

Parducci, L., Bennett, K.D., Ficetola, G.F. et al. 2017. Tansley Reviews: Ancient plant DNA from lake sediments. *New Phytologist* 214, 924-942.

Pedersen, M.W., Ginolhac, A., Orlando et al. 2013. A comparative study of ancient environmental DNA to pollen and macrofossils from lake sediments reveals taxonomic overlap and additional plant taxa. *Quaternary Science Reviews* 75: 161–168.

Pedersen, M.W., Overballe-Petersen,S., Ermini, L. et al. 2015. Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B* 370: 20130383.

Pedersen, M.W., Ruter, A., Schweger, C. et al. 2016. Postglacial viability and colonization in North America's ice-free corridor. *Nature* 537: 45-49.

Rawlence, N.J., Lowe, D.J., Wood, J.R. et al. 2014. Using palaeoenvironmental DNA to reconstruct past environments: progress and prospects. Journal of Quaternary Science 29: 610-626.

Reimer, P.J., Brown, T.A. and Reimer, R.W. 2004. Discussion: reporting and calibration of post-bomb ^{14}C data. *Radiocarbon*, 46, 1299–1304.

Ritchie, J.C. 1974. Modern pollen assemblages near the Arctic treeline, MacKenzie Delta region, Northwest Territories. *Canadian Journal of Botany* 52, 381-396.

Seppä, H. 2013: Pollen analysis, principles. In S.A. Elias and C. J. Mock, editors, *Encyclopedia of Quaternary Science* (Second Edition). Elsevier, 794-804.

Sjögren P., Edwards M.E., Gielly L. et al. 2017. Lake sedimentary DNA accurately records 20th Century introductions of exotic conifers in Scotland. New Phytologist. 213, 929–941.

Smith, O., Momber, G., Bates, R. et al. 2015. Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago. *Science* 347, 998–1001.

Sugita, S. 2007a. Theory of quantitative reconstruction of vegetation I: Pollen from large sites REVEALS regional vegetation composition. *The Holocene* 17, 229–241.

Sugita, S. 2007b. Theory of quantitative reconstruction of vegetation II: All you need is LOVE. *The Holocene* 17, 243–257.

Sønstebø, J.H., Gielly, L., Brysting, A.K. et al. 2010. Using next generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Molecular Ecology Resources* 10, 1009–1018.

Taberlet, P., Coissac, E., Pompanon, F. et al. 2007. Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Research* 35: e14.

Taberlet, P., Prud'Homme, S.M., Campione, E. et al. 2012. Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Molecular Ecology* 21, 1816–1820.

Taberlet, P., Coissac, E. Bonin, A. et al. 2013. Mapping biodiversity in a tropical forest using a DNA metabarcoding approach. Abstract: Association for Tropical Biology and Conservation/Organization for Tropical Studies 50th Anniversary Meeting, June, 2013, San Jose (https://atbc.confex.com/atbc/2013/webprogram/Paper1299.html).

Thomsen, P.F. and Willerslev, E. 2015. Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, 183, 4–18.

Valentini, A., Pompanon, F. and Taberlet, P. 2009. DNA barcoding for ecologists. *Trends in Ecology & Evolution* 24, 110–117.

Walker, D.A., Raynolds, M.K., Daniëls, F.J.A. et al. 2005. The Circumpolar Arctic vegetation map. *Journal of Vegetation Science* 16, 267-282.

Weiß, C.L., Dannemann, M., Prüfer, K. et al. 2015. Contesting the presence of wheat in the British Isles 8,000 years ago by assessing ancient DNA authenticity from low-coverage data. *eLife*. 2015; 4: e10005.

Willerslev, E., Hansen, A.J., Binladen, J. et al. 2003. Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* 300, 791–795

Willerslev, E., Davison, J., Moora, M. et al. 2014. Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature* 506, 47–51.

Yoccoz, N.G., Bråthen, K.A., Gielly, L. et al. 2012. DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology* 21, 3647–3655.

Zazula, G.D., Froese, D.G., Elias, S.A. et al. 2006. Vegetation buried under Dawson tephra (25,300 14C years BP) and locally diverse late Pleistocene paleoenvironments of Goldbottom Creek, Yukon, Canada. *Palaeogeography, Palaeoclimatology, Palaeoecology* 242, 253–286.

Zimmermann, H. H., Raschke1, E, Epp, L.S. 2017a. Sedimentary ancient DNA and pollen reveal the composition of plant organic matter in Late Quaternary permafrost sediments of the Buor Khaya Peninsula (north-eastern Siberia). *Biogeosciences* 14, 575-596.

Zimmermann, H.H., Raschke1, E, Epp, 2017b. The history of tree and shrub taxa on Bol'shoy Lyakhovsky Island (New Siberian Archipelago) since the Last Interglacial uncovered by sedimentary ancient DNA and pollen data. *Genes*, 8, 273.

Zinger, L., Chave, J., Coissac, E., 2016. Extracellular DNA extraction is a fast, cheap and reliable alternative for multi-taxa surveys based on soil DNA. *Soil Biology and Biochemistry* 96, 16-19.

Figure Captions

Figure 1. Spitsbergen, and the Colesdalen (A) and Endalen (B) sampling areas.

Figure 2. Photos of the study sites: a) Colesdalen, b) Endalen.

Figure 3. Field sampling design. Red circles represent the 1-3 soil samples; rings show the vegetation survey layout. The circular plot was divided into eight segments, and cover was recorded for each segment (32 segments in total; the lower half of the circle shows segment layout). A central 0.5x0.5m was surveyed separately.

Figure 4. Cumulative curves for the number of taxa (as MOTUs) observed in vegetation (y axis) and plot area in m^2 (x-axis) at Endalen. Red triangles mark the number of MOTUs in the DNA sample.

Figure 5. A plot-by-species matrix showing representation of taxa in vegetation by DNA. Top - plot name; Side - taxon name (see also Table 2). The left-hand columns are Colesdalen; note the more frequent occurrence of DNA of small-stature, non-dominant forb taxa, which partly reflects the pooling of three repeat samples, compared with Endalen (right-hand columns). Colour key shown at right.

Figure 6. Effect of plant abundance (percent cover) on the DNA detection rate of taxa present in the central (0.5 x 0.5 m) quadrat (all sites). X-axis: categories of percent cover for individual taxa. The number above each column indicates the number of taxa falling within this category (a given taxon may fall into more than one category). Y-axis: detected vs. not detected (percent). Blue = detected, orange = not detected.

Table 1: Summary of DNA and vegetation matches. Column 1 shows the species in the vegetation, column 2 the equivalent DNA MOTU.

Vegetation_taxon_name	**DNA: potential taxonomic resolution**
Alopecurus borealis Trin.	Pooideae
Betula nana L. ssp. *tundrarum* (Perfil.) Á.Löve & D.Löve	*Betula nana*
Bistorta vivipara (L.) S.F. Gray	*Bistorta vivipara*
Calamagrostis neglecta (Ehrh.) P.Gaertn., B.Mey. & Scherb. ssp. *Groenlandica* (Schrank) Matuszk	Pooideae
Cardamine bellidifolia L., ssp. bellidifolia	*Cardamine bellidifolia*
Cardamine pratensis L. ssp. *angustifolia* (Hook.) O.E.Schulz	*Cardamine pratensis*
Carex fuliginosa Schkuhr ssp. *misandra* (R.Br.) Nyman	*Carex fuliginosa*
Carex rupestris All.	*Carex rupestris*
Cassiope tetragona L.D.Don. ssp. *tetragona*	*Cassiope tetragona*
Cerastium arcticum Lange coll.	*Cerastium*
Cerastium arcticum x regelii	*Cerastium*
Cerastium regelii Ostenf.	*Cerastium*
Coptidium lapponicum (L.) Tzvelev	Ranunculaceae
Draba lactea Adams	*Draba*
Draba norvegica Gunn.	*Draba*
Draba sp.	*Draba*
Dryas octopetala L.	*Dryas octopetala*
Dupontia fisheri R. Br.	Pooideae

Equisetum arvense L. ssp. *alpestre* (Wahlenb.) Schönswetter & Elven	*Equisetum*
Equisetum scirpoides Michx.	*Equisetum*
Eriophorum scheuchzeri Hoppe ssp. *arcticum* Novoselova	*Eriophorum scheuchzeri*
Euphrasia wettsteinii G.Gussarova	*Euphrasia wettstenii*
Festuca cf. *edlundiae* S. Aiken, Consaul & Lefkovitch	Pooideae
Festuca rubra L. ssp. *richardsonii* (Hook.) Hultén	Pooideae
Hierochloe alpina (Sw.) Roem. & Schult. ssp. *alpina*	Pooideae
Huperzia arctica (Grossh. Ex Tolm.) Sipliv.	*Huperzia arctica*
Juncus biglumis L.	*Juncus biglumis*
Koenigia islandica L.	*Koenigia islandica*
Luzula confusa Lindeb.	*Luzula*
Luzula nivalis (Laest.) Spreng.	*Luzula*
Micranthes foliolosa (R. Br.) Gornall	*Micranthes foliolosa*
*Micranthes hieracif*olia (Waldst. & Kit. ex Willd.) Haw. ssp. *hieracifolia*	*Micranthes hieracifolia*
Oxyria digyna (L.) Hill	*Oxyria digyna*
Pedicularis dasyantha (Trautv.) Hadac	*Pedicularis*
Pedicularis hirsuta L.	*Pedicularis*
Poa alpina L. var. *vivipara*	Pooideae
Poa arctica R.Br. ssp. *arctica*	Pooideae
Poa pratensis L. ssp. *alpigena* (Fr.) Hiit.	Pooideae
Potentilla hyparctica Malte ssp. *hyparctica*	*Potentilla hyparctica*
Ranunculus hyperboreus Rottb. ssp. *arnellii* Scheutz	Ranuculaceae
Ranunculus nivalis L.	Ranuculaceae
Ranunculus pygmaeus Wahlenb.	Ranuculaceae

Ranunculus sulphureus Sol.	Ranuculaceae
Sagina nivalis (Lindbl.) Fr.	*Sagina nivalis*
Salix polaris Wahlenb.	Saliceae
Saxifraga cespitosa L. ssp. *cespitosa*	*Saxifraga cespitosa*
Saxifraga oppositifolia L. ssp. *oppositifolia*	*Saxifraga oppositifolia*
Saxifraga svalbardensis Øvstedal	*Saxifraga svalbardensis*
Silene acaulis (L.) Jacq.	*Silene acaulis*
Stellaria longipes Goldie coll.	*Stellaria longipes*
Trisetum spicatum (L.) K.Richt. ssp. *spicatum*	Pooideae
Vaccinium uliginosum L. ssp. *microphyllum* Lange	*Vaccinium uliginosum*

Figure 1

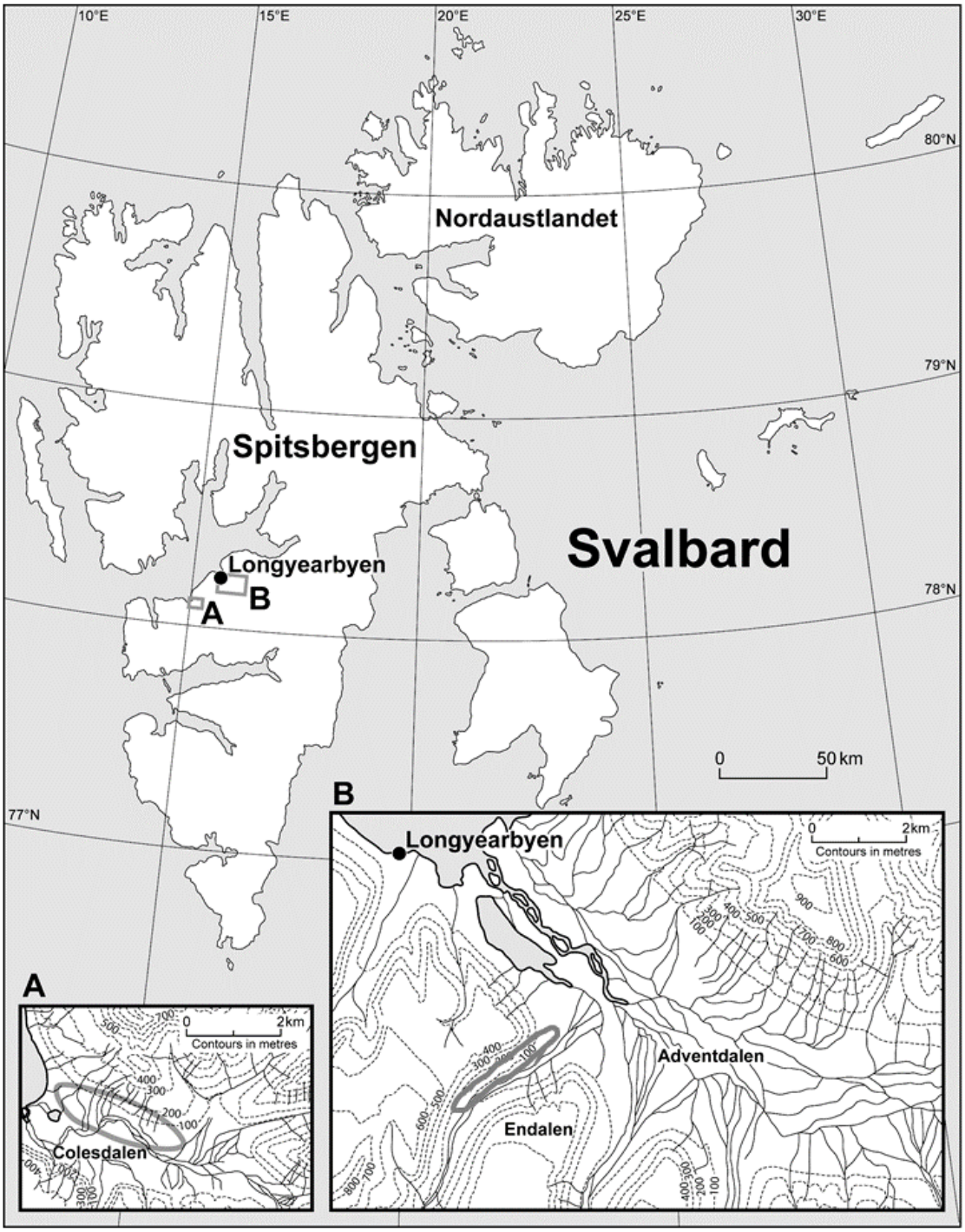


Figure 2

Fig 2a

Fig 2b

Figure 3

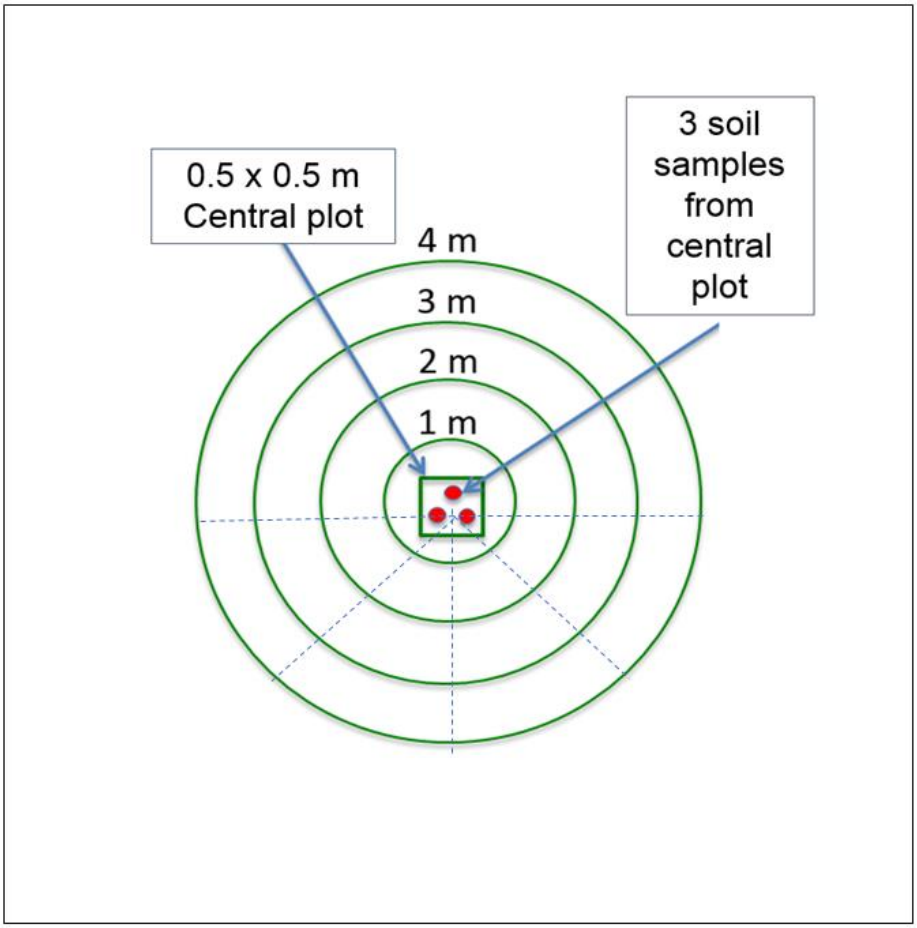

0.5 x 0.5 m
Central plot
3 soil
samples
from
central
plot
4 m
3 m
2 m
1 m

Figure 4

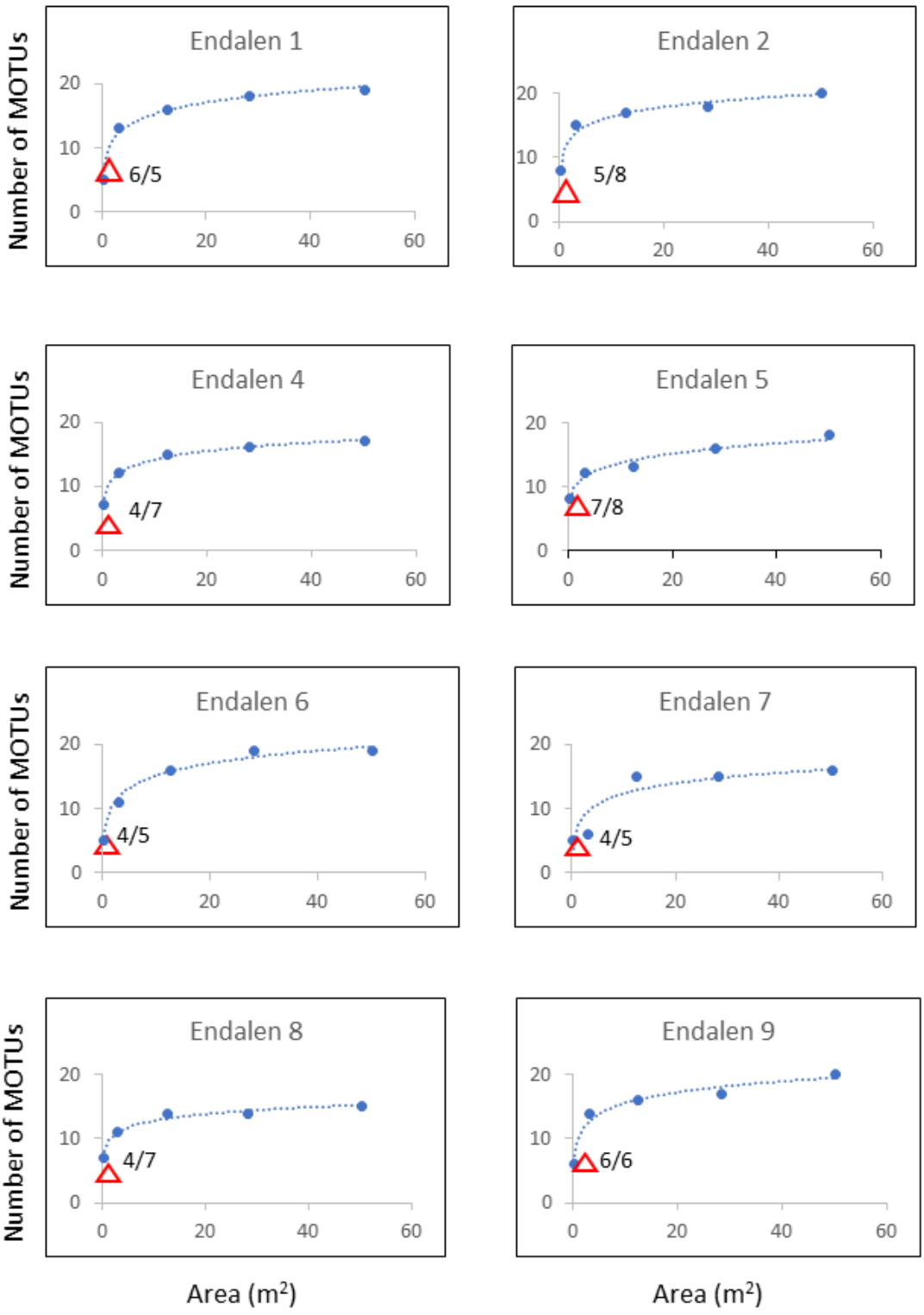
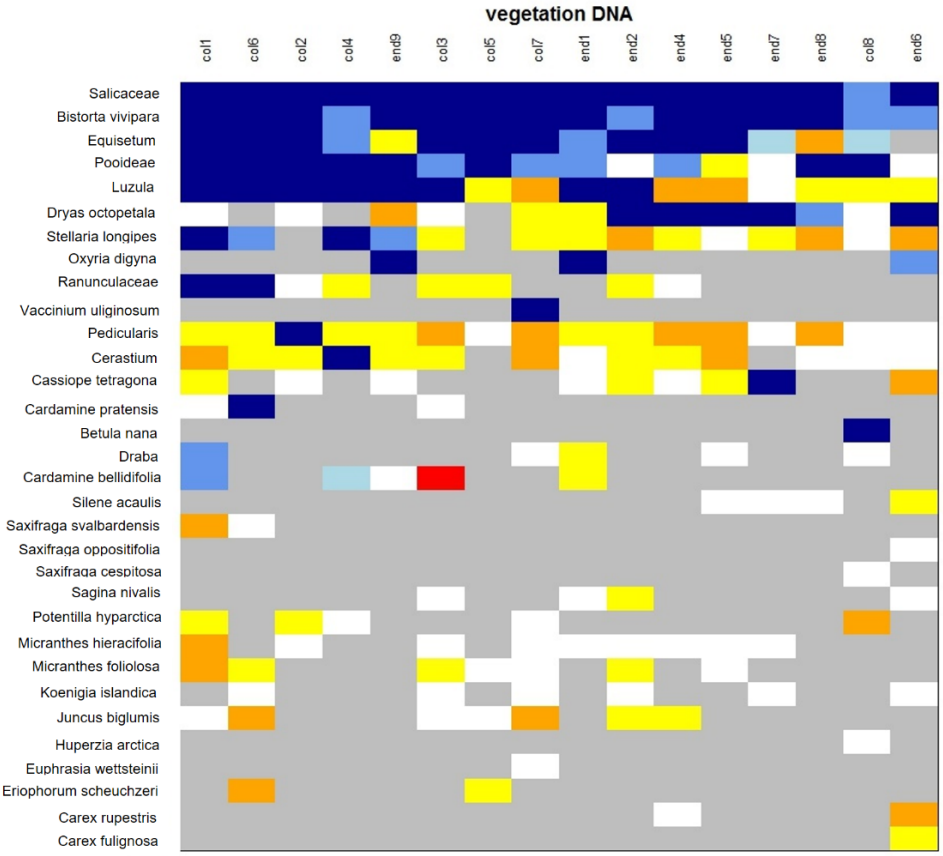

Endalen 1
6/5
Endalen 2
5/8
Endalen 4
4/7
Endalen 5
7/8
Endalen 6
4/5
Endalen 7
4/5
Endalen 8
4/7
Endalen 9
6/6
Number of MOTUs
Area (m²)
0
10
20
40
60

Figure 5



Key to Fig. 5

	present in DNA	absent from DNA
present in vegetation at 0.5 m		
present in vegetation at 1 m		
present in vegetation at 2-4 m		
absent from vegetation		

Figure 6

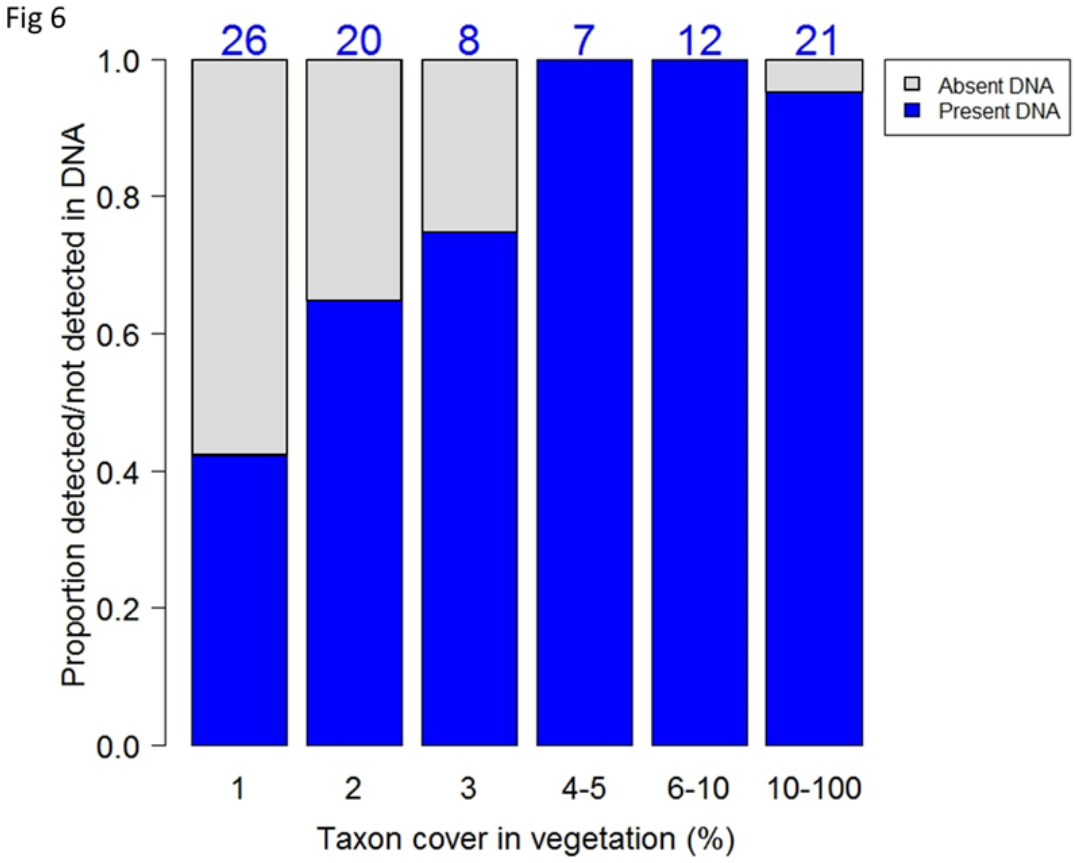


Fig 6
26
20
8
7
12
21
1.0
0.8
0.6
0.4
0.2
0.0
Proportion detected/not detected in DNA
Absent DNA
Present DNA
1
2
3
4-5
6-10
10-100
Taxon cover in vegetation (%)

Supplementary Material

Table S1

Numbers of PCR repeats per plot kept after filtering. Samples 1-3 refer to the replicate soil samples taken at Colesdalen only**.**

plot	**sample 1**	**sample 2**	**sample 3**
col1	9	9	9
col2	8	9	9
col3	9	9	9
col4	9	9	8
col5	11	12	11
col6	8	9	9
col7	6	9	3
col8	6	9	9
end1	4	0	0
end2	3	0	0
end4	4	0	0
end5	4	0	0
end6	2	0	0
end7	4	0	0
end8	4	0	0
end9	4	0	0

Table S2. Radiocarbon dates. Calendar years are estimated from atmospheric post-bomb carbon-14 levels. For each site, the samples are grouped by whether they are derived from plant macrofossils or the sediment matrix. For Colesdalen, where three samples were taken per plot, the sample name indicates plot and replicate; M indicates a macrofossil sample. For Endalen, a single monolith was taken per plot but sampled at varying depths, with replicates taken at some depths. The depth below surface (in cm) is given after the plot ID and before the sample ID, e.g., Endalen Plot 1 **6-8**/1; M indicates a macrofossil sample. Samples marked * have a ^{14}C age corresponding to a brief period in 1955 and are interpreted as most likely representing the average of several plant fragments of different ages in one sample.

Sample name	Material	Poz-ID	^{14}C Age	Error	Estimated Calendar year for post-bomb dates	Comment
COLESDALEN						
Colesdalen Plot 4/2M	plant remains	45289	-912	26	1995	
Colesdalen Plot 5/1M	plant remains	45290	-860	27	1996	
Colesdalen Plot 6/2M	plant remains	45291	-564	41	2004	
Colesdalen Plot 7/2M	plant remains	45292	-37	28	*	
Colesdalen Plot 8/1M	plant remains	45296	-2450	25	1976	
Colesdalen Plot 4/2	organic sediment	45264	7610	50		
Colesdalen Plot 5/1	organic sediment	45265	-983	27	1993	
Colesdalen Plot 6/2	organic sediment	45267	7700	50		
Colesdalen Plot 7/2	organic sediment	45268	2310	35		
Colesdalen Plot 8/1	organic sediment	45269	2425	35		
ENDALEN						
Endalen Plot 1 6-8/1M	plant remains	45850	-2018	25	1980	stem 1fragment (fr)
Endalen Plot 2 0-2/1M	plant remains	45852	-1454	25	1986	twig 1fr

Endalen Plot 2 2-4/1M	plant remains	45853	195	35	-	twig 1fr
Endalen Plot 2 4-6/1M	plant remains	45854	-1482	29	1985	twig 2fr
Endalen Plot 2 6-8/1M	plant remains	45855	-1073	118	1991	twigs 5fr
Endalen Plot 3 2-4/1M	plant remains	45856	-976	27	1993	twig 1fr
Endalen Plot 3 4-6/1M	plant remains	45857	-1438	27	1986	stem 1fr (Equisetum?)
Endalen Plot 4 2-4/1M	plant remains	45858	-883	26	1995	twig 1fr
Endalen Plot 4 4-6/1M	plant remains	45859	-1001	24	1993	stem 1fr (Equisetum?)
Endalen Plot 4 6-8/1M	plant remains	45860	-2291	25	1978	stem 2fr (Equisetum?)
Endalen Plot 5 3-5/1M	plant remains	45962	-63	28	*	twig 1fr
Endalen Plot 5 5-7/1M	plant remains	45863	110	30		twig 1fr
Endalen Plot 6 2-4/1M	plant remains	45864	-1054	26	1992	stem (grass)
Endalen Plot 7 5-7/1M	plant remains	45865	1380	50		stems >5fr moss
Endalen Plot 7 7-9/1M	plant remains	45866	335	30		twig 1fr
Endalen Plot 8 8-10/1M	plant remains	45867	320	35		ears >5fr
Endalen Plot 1 6-8/2M	plant remains	45869	220	30		leaves >5fr
Endalen Plot 2 0-2/2M	plant remains	45870	-626	28	2002	leaves >5fr
Endalen Plot 2 2-4/2M	plant remains	45872	-1093	31	1991	moss twig leaves
Endalen Plot 3 2-4/2M	plant remains	45773	-1100	41	1993	leaves >5fr
Endalen Plot 3 4-6/2M	plant remains	45874	-1283	26	1988	stems >5fr (sedge?)
Endalen Plot 4 2-4/2M	plant remains	45875	-563	27	2004	auricles>5fr
Endalen Plot 4 4-6/2M	plant remains	45876	-3295	22	1971	stems >5fr (moss)
Endalen Plot 4 6-8/2M	plant remains	45877	-1031	31	1992	stems >5fr(sedge?)
Endalen Plot 5 3-5/2M	plant remains	45878	-380	27	2009	leaves >5fr
Endalen Plot 5 5-7/2M	plant remains	45879	-59	28	*	stems, leaves (moss)
Endalen Plot 6 2-4/2M	plant remains	45881	-2587	24	1975	leaves >5fr
Endalen Plot 7 5-7/2M	plant remains	45882	-1455	44	1986	leaves >5fr
Endalen Plot 7 7-9/2M	plant remains	45883	-488	31	2006	stems, leaves (moss) >5fr
Endalen Plot 8 8-10/2M	plant remains	45884	158	29		leaves >5fr

Endalen Plot 9 7-9/2M	plant remains	45885	95	35		leaves >5fr
Endalen Plot 1 6-8	organic sediment	45270	3595	35		
Endalen Plot 2 0-2	organic sediment	45271	7350	50		
Endalen Plot 2 2-4	organic sediment	45272	2800	40		
Endalen Plot 2 4-6	organic sediment	45273	13810	70		
Endalen Plot 2 6-8	organic sediment	45274	14630	80		
Endalen Plot 3 2-4	organic sediment	45275	9780	50		
Endalen Plot 3 4-6	organic sediment	45277	11200	60		
Endalen Plot 4 2-4	organic sediment	45278	1930	30		
Endalen Plot 4 4-6	organic sediment	45279	2575	35		
Endalen Plot 4 6-8	organic sediment	45280	10030	50		
Endalen Plot 5 3-5	organic sediment	45281	6810	50		
Endalen Plot 5 5-7	organic sediment	45282	2335	35		
Endalen Plot 6 2-4	organic sediment	45283	6580	40		
Endalen Plot 7 5-7	organic sediment	45284	6260	40		
Endalen Plot 7 7-9	organic sediment	45285	6180	50		
Endalen Plot 8 8-10	organic sediment	45287	3570	35		
Endalen Plot 9 7-9	organic sediment	45288	6100	40		

Figure captions:

Figure S1. Schematic of probable sources of DNA to a small soil sample in tundra. Material may be derived from above- and below-ground plant matter (litter, larger roots, fine roots) growing on or close to the sampled soil, and it may be transported downslope complexed on soil particles (clays, humic material) in over-ground flow or throughflow in the active layer.

Figure S2. S2a, pollen diagram for Colesdalen; S2b pollen diagram for Endalen. Y axis – sample number; X axis – pollen frequency (percent terrestrial pollen sum). Aquatic taxa not shown.

Figure S1

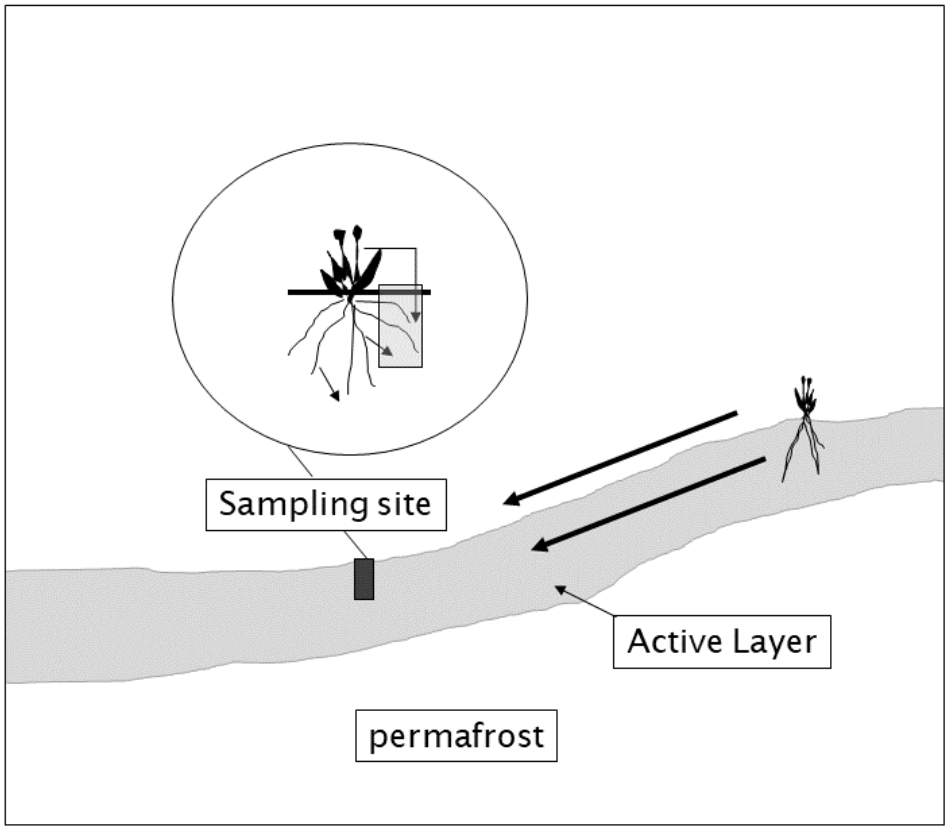


Figure S2a

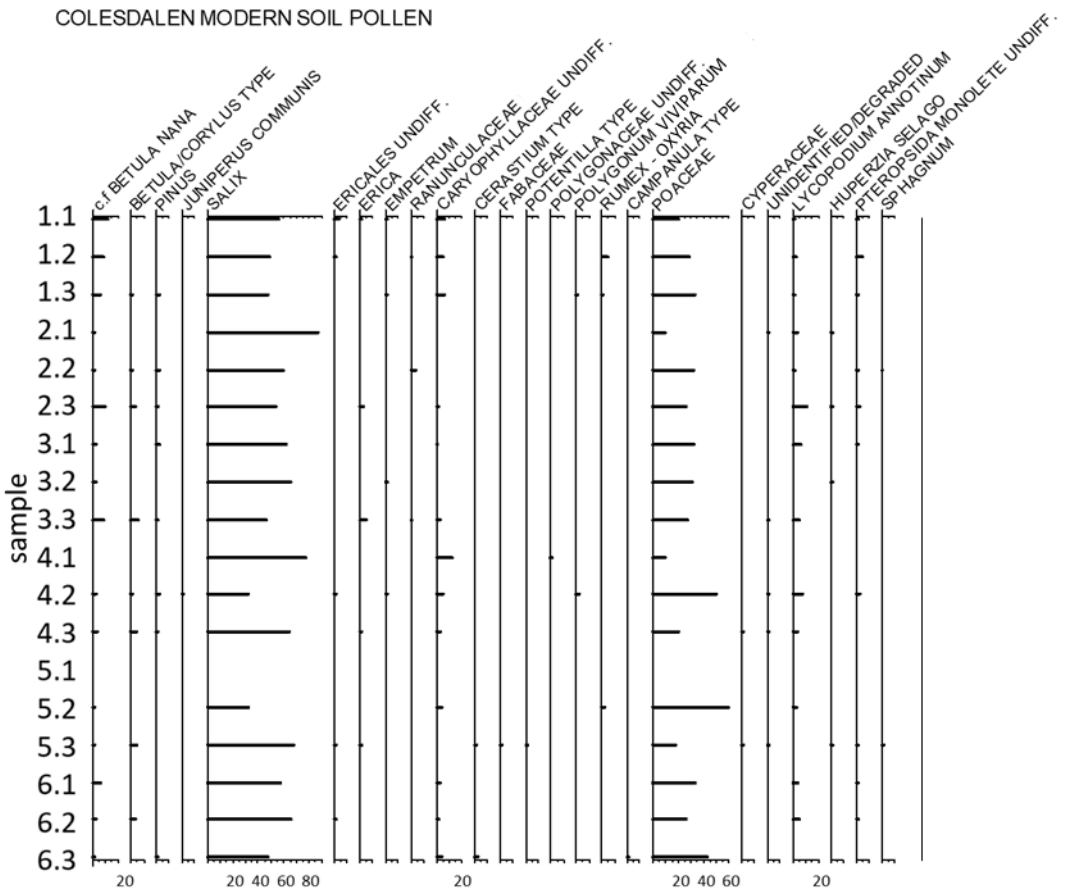
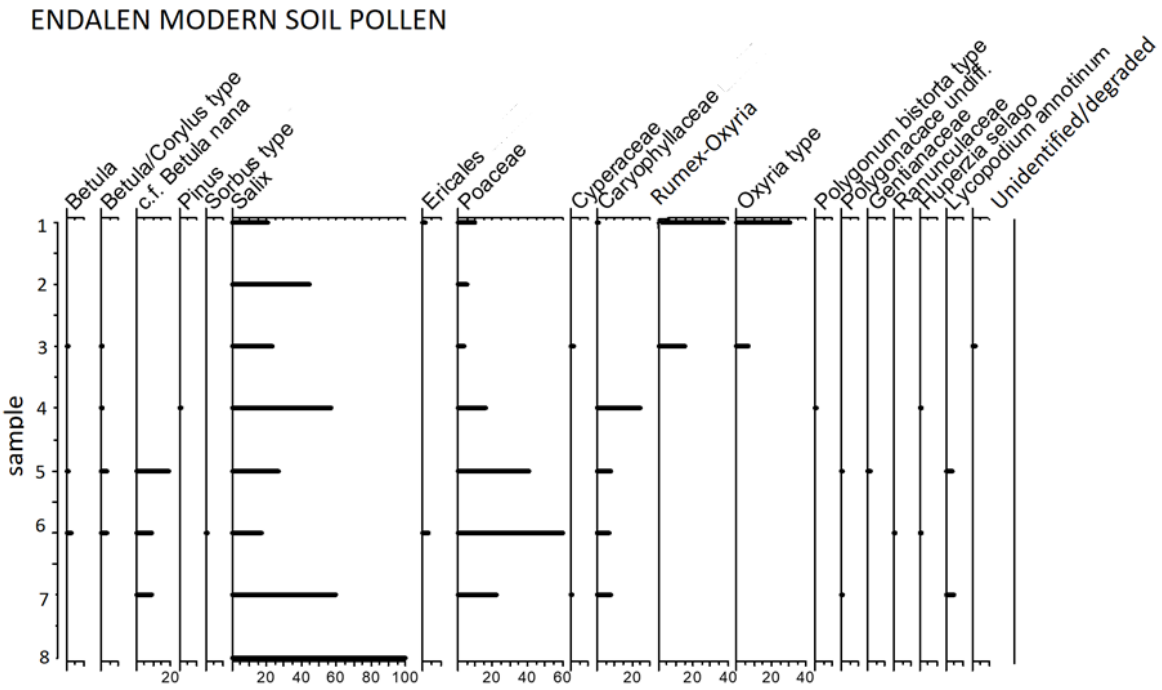

COLESDALEN MODERN SOIL POLLEN
c.f BETULA NANA
BETULA/CORYLUS TYPE
PINUS
JUNIPERUS COMMUNIS
SALIX
ERICALES UNDIFF.
ERICA
EMPETRUM
RANUNCULACEAE
CARYOPHYLLACEAE UNDIFF.
CERASTIUM TYPE
FABACEAE
POTENTILLA TYPE
POLYGONACEAE UNDIFF.
POLYGONUM VIVIPARUM
RUMEX - OXYRIA
CAMPANULA TYPE
POACEAE
CYPERACEAE
UNIDENTIFIED/DEGRADED
LYCOPODIUM ANNOTINUM
HUPERZIA SELAGO
PTEROPSIDA MONOLETE UNDIFF.
SP HAGNUM
sample
1.1
1.2
1.3
2.1
2.2
2.3
3.1
3.2
3.3
4.1
4.2
4.3
5.1
5.2
5.3
6.1
6.2
6.3
20
20 40 60 80
20
20 40 60
20

Figure S2b


ENDALEN MODERN SOIL POLLEN
Betula
Betula/Corylus type
c.f. Betula nana
Pinus
Sorbus type
Salix
Ericales
Poaceae
Cyperaceae
Caryophyllaceae
Rumex-Oxyria
Oxyria type
Polygonum bistorta type
Polygonacace undiff.
Gentianaceae
Ranunculaceae
Huperzia selago
Lycopodium annotinum
Unidentified/degraded
sample
1
2
3
4
5
6
7
8
20
20 40 60 80 100
20 40 60
20
20 40
20 40