1 # Environmental palaeogenomic reconstruction of an Ice Age algal

2 # population

3 Youri Lammers[1], Peter D. Heintzman[1,*], Inger Greve Alsos[1,*]

4

5 [1]The Arctic University Museum of Norway, UiT - The Arctic University of Norway, Tromsø,

6 Norway

7 [*]Contributed equally to this work

8

9 ## Abstract

10 Palaeogenomics has greatly increased our knowledge of past evolutionary and ecological

11 change, but has been restricted to the study of species that preserve as fossils. Here we show

12 the potential of shotgun metagenomics to reveal population genomic information for a taxon

13 that does not preserve in the body fossil record, the algae *Nannochloropsis*. We shotgun

14 sequenced two lake sediment samples dated to the Last Glacial Maximum and identified *N.*

15 *limnetica* as the dominant taxon. We then reconstructed full chloroplast and mitochondrial

16 genomes to explore within-lake population genomic variation. This revealed at least two

17 major haplogroups for each organellar genome, which could be assigned to known varieties

18 of *N. limnetica*. The approach presented here demonstrates the utility of lake sedimentary

19 ancient DNA (*sed*aDNA) for population genomic analysis, thereby opening the door to

20 environmental palaeogenomics, which will unlock the full potential of *sed*aDNA.

21 ## Keywords

22 Sedimentary ancient DNA, palaeogenomics, shotgun metagenomics, haplotype diversity, ice

23 age, *Nannochloropsis*

24

25 ## Introduction

26 Palaeogenomics, the genomic-scale application of ancient DNA, is revolutionizing our

27 understanding of past evolutionary and ecological processes, including population dynamics,

28 hybridization, extinction, and the effects of drivers of change [1–4]. Despite extensive

29 application to, and innovations using, body fossils [5–7], its use on another major source of

30 ancient DNA - the environment - has been almost entirely limited to inferring the presence or

31 absence of taxa through time [8–14]. However, a nuanced understanding of ecological and

32 evolutionary dynamics requires population genomic information. The direct recovery of this

33 information from cave sediment has recently been shown [12], but - to our knowledge - has not

34 yet been demonstrated for lake sediments.

35   Lake sediments provide an ideal source of sedimentary ancient DNA (*sed*aDNA)

36 that originates from both the catchment and the lake itself, as well as providing a stable

37 environment required for optimal aDNA preservation [15,16]. As a result, lake *sed*aDNA has

38 been used to infer the taxonomic composition of past communities [16,17], regardless of whether

39 those taxa preserve in the body fossil record. The most commonly applied method is DNA

40 metabarcoding, which allows for the targeting of particular groups of organisms [18,19].

41 However, the ability to confidently identify barcodes is constrained by the completeness of

42 appropriate reference databases, and the length and variability of the barcode targeted. Short

43 barcodes are necessarily targeted for fragmented aDNA [20], which can therefore impede

44 species-level identification. An alternative approach is shotgun metagenomics, which is non-

45 targeting and preserves aDNA damage patterns that, in contrast to metabarcoding, allows for

46 authentic aDNA to be distinguished from modern contamination [16,21–23]. For palaeogenomic

47 reconstruction however, either deep shotgun sequencing or target enrichment of *sed*aDNA is

48 required, which allows for robust species-level identification [12,14,24], as well as the potential

49 exploration of population genomic variation.

50   Andøya, an island located off the coast of northwest Norway, was partially

51 unglaciated during the Last Glacial Maximum (LGM, Figure 1) and has therefore been a

52 focus of palaeoecological studies [25], especially for its potential as a cryptic northern refugium

53 [26–28]. Studies focussing on sediment cores from three lakes (Endletvatn, Nedre Æråsvatnet,

54 Øvre Æråsvatnet)[26,29–35] have reported the presence of an Arctic community during the LGM,

55 which includes taxa such as grasses (Poaceae), crucifers (Brassicaceae), and poppy

56 (*Papaver*)*,* along with bones of the Little Auk (*Alle alle*). Furthermore, recent geochemical

57 and DNA metabarcoding analyses indicate the presence, and an inferred high abundance, of

58 the algae *Nannochloropsis* in LGM sediments from Andøya [35].

59   *Nannochloropsis* is a genus of single-celled microalgae of the Eustigmatophyceae. All

60 species have high lipid contents and are therefore of interest as a potential source of biofuels

61 [36,37]. As a result of this economic interest, the organellar and nuclear genomes have been

62 sequenced for six of the eight described species (Supplementary Table S1) [37–40]. All species

63 are known from marine environments, with the exception of *N. limnetica*, which is known

64   from freshwater and brackish habitats, and comprises five varieties [41–43]. The genus has a

65   cosmopolitan distribution, with the marine species being reported from most oceans [44–46],

66   whereas the freshwater/brackish *N. limnetica* is known from lakes in Europe [41], Asia [42], North

67   America [43,47], and Antarctica [48]. Species-level identification of *Nannochloropsis* from water

68   is problematic, due to its small size (2 to 6 μm in diameter [43,46]) and, in contrast to diatoms [49],

69   lack of diagnostic morphological structures [47,50]. Reliable species identification is however

70   possible with short genetic markers [43,47,51]. In sediments, *Nannochloropsis* has not been

71   reported from macrofossil and pollen/spore profiles, and may therefore only be identifiable

72   using *sed*aDNA [10,35,52].

73   In this study, we shotgun sequenced two broadly contemporaneous LGM lake

74   sediment layers from Andøya that a previous metabarcoding study had shown to contain

75   *Nannochloropsis* [35]. The depth of our shotgun metagenomic data, together with the

76   availability of a reference genome panel, allowed us to demonstrate that *N. limnetica*

77   dominates the identifiable taxonomic profile. Through reconstruction of complete chloroplast

78   and mitochondrial genomes, we show that at least two variants of *N. limnetica* are

79   represented. We thus demonstrate, and to the best of our knowledge for the first time, that it

80   is possible to estimate past population genomic diversity both from total *sed*aDNA and from

81   a taxon not preserved in the body fossil record.

82

Results

84   **Metagenomic analysis and species-level determination of *Nannochloropsis***

85   We shotgun sequenced two LGM samples, dated to 17,700 (range: 20,200-16,500) and

86   19,500 (20,040-19,000) calibrated years before present (cal yr BP), to generate 133-224

87   million paired-end reads, of which we retained 53-127 million sequences after filtering

88   (Supplementary Table S2). We first sought to identify the broad metagenomic profiles of the

89   samples and the species-level identification of *Nannochloropsis* from Lake Øvre Æråsvatnet.

90   First, for each sample, we compared two non-overlapping one-million sequence

91   subsets of the filtered data to the NCBI nucleotide database. The taxonomic overlap between

92   the two one million read subsets was 88-93% within each sample, demonstrating that our

93   subsets are internally consistent. We then merged the two subsets from each sample, which

94   resulted in the identification of 29,500-32,700 sequences (Table 1). The majority of the

95   identified sequences were bacterial, with 21-26% identified as *Mycobacterium*, although the

96   majority of these sequences could not be identified to a specific strain. Within the eukaryotes,

97   *Nannochloropsis* constituted ~20% of the assigned sequences in both samples, with ~33% of

98   these identified as *N. limnetica* (Table 1; Supplementary Figure S1; Supplementary Table

99   S3).

100   To further investigate the metagenomic profile of the samples, we aligned all filtered

101   sequences to each nuclear genome within a reference panel derived from four *Mycobacterium*

102   strains and 38 eukaryotes, with the latter either exotic (implausible) or non-exotic (plausible)

103   to LGM Andøya (Supplementary Table S4). We mapped 310,000-680,000 sequences to the

104   *N. limnetica* genome, which translates to 9.3-20.3 thousand sequences per megabase

105   (kseq/Mb), and a nuclear genomic coverage of 0.48-1.13x. We observed a far lower relative

106   mapping frequency to all other *Nannochloropsis* nuclear genomes (up to 2.5-4.8 kseq/Mb). If

107   we consider sequences that are only mappable to a single genome, then the relative mapping

108   frequency falls to 7.4-17 kseq/Mb for *N. limnetica* and up to 0.6-1.3 kseq/Mb for all other

109   *Nannochloropsis* genomes. The most abundant non-*Nannochloropsis* eukaryotic taxon in the

110   sequence data was human, with 2-11 thousand sequences mapped (0.7-3.4 seq/Mb). As

111   expected from the metagenomic analyses, the next most abundant group is *Mycobacterium*.

112   As the relative mapping frequency was consistent across all four strains, based on both all

113   sequences aligned (up to 1.1-1.7 kseq/Mb) and only retaining sequences unique to a strain (up

114   to 0.6-0.9 kseq/Mb), we infer that the *Mycobacterium* strain or strains present in LGM

115   Andøya are not closely related to any that have been sequenced to date (Figure 1;

116   Supplementary Table S4). The relative frequencies of sequences mapping to plausible and

117   implausible eukaryotic genomes are comparable for non-*Nannochloropsis* taxa. Based on

118   both raw counts and those corrected for genome size, these analyses therefore indicated that

119   *N. limnetica* is the best represented taxon in the panel (Figure 1; Supplementary Table S4).

120   We sought to confirm whether sequences identified from the three best represented

121   taxonomic groups (*Nannochloropsis*, *Mycobacterium*, human) were likely to be of ancient

122   origin or to have derived from modern contamination. The sequences aligned to the human

123   genome did not exhibit typical patterns of ancient DNA damage, which include cytosine

124   deamination and depurination-induced strand breaks and are therefore considered to be of

125   modern contaminant origin (Supplementary Figure S2). In contrast, we find that sequences

126   aligned to two *Nannochloropsis limnetica* and the *Mycobacterium avium* genomes exhibit

127   authentic ancient DNA damage (Supplementary Figures S3 and S4), with patterns that are

128   near identical for both taxonomic groups, consistent with their preservation in the same

129   environment of broadly contemporaneous age.

130   We next aligned our sequence data against two organellar reference panels consisting

131   of either 2742 chloroplast or 8486 mitochondrial genomes (Supplementary Table S5). Both

132   analyses also recovered *N. limnetica* as the best represented taxon, with 23,600-37,900 and

133   8,600-14,100 sequences aligning to the chloroplast and mitochondrial genome of this taxon,

134   respectively. After mapping the filtered sequence data to the *Nannochloropsis* chloroplast

135   genomes individually, the number of sequences uniquely aligned to *N. limnetica* fell to 7,700-

136   11,300 (Supplementary Table S4). The most abundant non-*Nannochloropsis* taxon was the

137   algae *Choricystis parasitica* with 255-1,784 and 38-276 of sequences mapped to its

138   chloroplast and mitochondrial genome.

139

140   **Reconstruction of *Nannochloropsis* organellar palaeogenomes and their phylogenetic**

141   **placement**

142   We reconstructed complete composite organellar palaeogenomes for *Nannochloropsis*

143   present in LGM Andøya, using *N. limentica* as a seed sequence. The resulting complete

144   chloroplast sequence was 117.7 kilobases (kb) in length and had a coverage depth of 64.3x.

145   The mitochondrial genome was 38.5 kb in length with a coverage of 62.4x (Figure 2;

146   Supplementary Figure S5; Supplementary Table S6). We observed two major structural

147   changes in our reconstructed chloroplast as compared to the *N. limnetica* seed sequence, in

148   which the reconstructed chloroplast was inferred to share the ancestral structural state with

149   the remaining *Nannochloropsis* taxa. This included a 233 bp region in a non-coding region

150   between the *thiG* and *rpl27* genes, which is absent in the *N. limnetica* seed sequence and of

151   varying length among all other *Nannochloropsis* taxa (Supplementary Figure S6). A 323 bp

152   insertion in a non-coding region between the genes *rbcS* and *psbA*, is present in the *N.*

153   *limnetica* seed sequence, but lacking from our reconstructed chloroplast and all other

154   *Nannochloropsis* taxa (Supplementary Figure S6). We noted that the combined coverage was

155   reduced across these two regions, with 21x and 32x chloroplast genome coverage (the latter

156   calculated from 100 bp upstream and downstream of the deletion), which may be suggestive

157   of within-sample variation. Both the reconstructed composite organellar genomes displayed

158   authentic ancient DNA damage patterns (Supplementary Figures S7 and S8).

159   To account for within-sample variants in our reconstructed organellar palaeogenomes,

160   we created two consensus sequences that included either high or low frequency variants at

161   multiallelic sites. We performed phylogenetic analyses to confirm the placement of the high

162   and low frequency variant consensus genomes relative to other *Nannochloropsis* taxa. For

163   this, we used full organellar genomes and three short loci with high taxonomic representation

5

164 in NCBI Genbank (18S, ITS, *rbc*L; Table S7). Altogether, these analyses from three different

165 markers (chloroplast, mitochondrial, nuclear) were congruent and resolved the high

166 frequency variant consensus sequences as likely deriving from *N. limnetica* var. *globosa* and

167 the low frequency variant consensus sequences as *N. limnetica* var. *limnetica* (Table 2; Figure

168 3; Supplementary Figures S9).

169 We attempted to reconstruct composite chloroplast genomes using alternative

170 *Nannochloropsis* taxa as seed sequences, but these analyses failed to resolve a complete

171 composite sequence (Supplementary Table S8). A phylogenetic analysis of these alternative

172 composite chloroplast genomes displays a topology consistent with the biases associated with

173 mapping to increasingly diverged reference genomes (Supplementary Figure S10). These

174 alternative composite chloroplast genomes were therefore not used further, but provide

175 supporting evidence that *N. limnetica* is the most closely related extant taxon.

176

177 ***Nannochloropsis limnetica* allelic variation and haplotype estimation**

178 In the absence of a catalog of chloroplast and mitochondrial genomes from the *N. limnetica*

179 variants, we sought to explore the frequencies and proportions of allelic variants present in

180 our data set. We combined all sequences aligned to the high and low frequency variant

181 consensus genomes into a single data set for each sample. We restricted our analyses to

182 transversion variants only in order to exclude artifacts derived from ancient DNA damage,

183 and defined the reference allele as that present in the reconstructed composite organellar

184 genomes. We detected 299-376 and 81-112 variants within the *N. limnetica* chloroplast and

185 mitochondrial genomes, respectively (Supplementary Table S9). For each sample and across

186 the entire organellar genome, the average proportion of the transversion-only alternative

187 allele is 0.39-0.42 for chloroplast variants and 0.39-0.43 for mitochondrial variants (Figure

188 4).

189 After pooling data from both samples, we used the phasing of adjacent alleles, which

190 were linked by the same read, to infer the minimum number of haplotypes in each

191 reconstructed composite organellar genome. We identify 70 and 21 transversion-only phased

192 positions in the chloroplast and mitochondrial genomes, respectively. Within each sample,

193 the average number of haplotypes observed, based on the linked alleles in the chloroplast

194 genome, is 1.93-2.09. The equivalent average for the mitochondrial genome is 2.05-2.29

195 (Figure 4; Supplementary Table S10).

196

Discussion

All of our analyses identified *Nannochloropsis* as the most abundant eukaryotic taxon in the LGM lake sediments from Andøya, consistent with a previous study based on plant DNA metabarcoding [35]. We observed ancient DNA deamination patterns for all reference sequence combinations, which supports the authenticity of our data. The phylogenetic placement of our organellar palaeogenomes, as well as other short loci, indicate that the *Nannochloropsis* taxon detected in Andøya is *N. limnetica*, with at least two varieties present: *N. limnetica* var. *globosa* and *N. limnetica* var. *limnetica*.

The low overall proportion of sequences identified by the NCBI-based metagenomic analysis is broadly consistent with other shotgun metagenomic studies from *sed*aDNA [10,13,14] and suggests that the vast majority of taxonomic diversity in the sediment record is currently unidentifiable. We recovered comparable and low relative mapping frequencies for all non-*Nannochloropsis* eukaryotic taxa in our genome panel, regardless of their plausibility of occurring at LGM Andøya. We therefore suggest that these mappings are artifacts resulting from the spurious mapping of short and damaged ancient DNA molecules coupled with the vast diversity of sequences present in *sed*aDNA[53,54]. However, we identified a component of *Mycobacterium* sequences, which display ancient DNA damage patterns, although, unlike *Nannochloropsis*, no dominant strain could be identified. This indicates that our samples contain one or multiple unsequenced strains, some or all of which may be extinct.

We explored whether Andøya *Nannochloropsis* could potentially comprise more than one species. The detection of a low number of uniquely mapped sequences to non-*N. limnetica Nannochloropsis* genomes indicates that we cannot exclude the possibility of other rare *Nannochloropsis* taxa being present in the sequence data. However, we suggest that *N. limnetica* is the sole taxon present and that evolutionary divergence and potential technical artifacts can explain the conflicting results. Our phylogenetic analyses suggest that the Andøya *N. limnetica* variants are not evolutionarily close to any available organellar reference genomes. We therefore might expect some regions of our *N. limnetica* organellar genomes to be evolutionarily closer to non-*N. limnetica* than to *N. limnetica*. This is supported by the trend of decreasing quality of, and increased impact of reference bias upon, the chloroplast genome sequences reconstructed using alternative seed genomes. This could have been compounded by the aforementioned artifacts associated with ancient DNA [53,54].

We detected within-sample allelic variation in both the *N. limnetica* chloroplast and mitochondrial genome reconstructions, which we split into two consensus sequences

7

230　containing either the high or the low frequency variants. In both organellar genomes, the high

231　frequency variants for both samples, assigned as *N. limnetica* var. *globosa*, clustered

232　separately from the low frequency variants, which we identified as *N. limnetica* var.

233　*limnetica*. This demonstrates that the results from each of our broadly contemporaneous

234　samples are replicable, which is further confirmed by comparable results obtained by the

235　estimation of the minimum number of haplotypes. For both samples, our analyses recover at

236　least two haplotypes, with a small proportion of phased positions containing three. We note

237　that our method is conservative, given the strict filtering criteria and limited window size, and

238　almost certainly underestimates true haplotype diversity. To accurately estimate the diversity

239　and proportions of haplotypes, it is likely that an extensive reference database of *N. limnetica*

240　haplotypes will be required. However, this may be particularly problematic for taxa that lack

241　body fossils, which are currently required to reconstruct extinct haplotypes. Future

242　methodological and statistical advances will therefore be required to estimate and quantify

243　haplotype variation for taxa from within a *sed*aDNA population sample.

244　　　　The sheer abundance of *N. limnetica* sequences in our identified *sed*aDNA shotgun

245　sequence data suggests a high biomass of this algae in Lake Øvre Æråsvatnet during the

246　LGM. *Nannochloropsis* is known to undergo blooming events that can reach up to 10^{10} cells

247　per litre of water [55], which have been reported for *N. gaditana* in the Comacchio Lagoons,

248　Italy [56] and *N. granulata* in the Yellow Sea, China [55]. *N. limnetica* itself was first described

249　from spring blooms in Germany, reaching concentrations up to 5.7×10^9 cells per litre [41]. Such

250　blooms could explain the observed high sequence abundance in our data. Independent proxies

251　from the same LGM sediments, including high loss-on-ignition (LOI) values [32] and organic

252　elemental (C/N) proportions [35], are consistent with a blooming scenario resulting from high

253　nutrient input. Stable isotope data suggest that the C/N is of a high trophic origin, most likely

254　bird guano from an adjacent bird cliff [35], which corresponds with the detection of bird bones

255　(little auk, *Alle alle*) in the LGM sediments [30,33,35]. The high inflow of nutrients into the lake

256　could have resulted in eutrophication of the lake ecosystem and thus initiated blooms of *N.*

257　*limnetica*.

258　　　　*Nannochloropsis* has not previously been reported from contemporary northern

259　Norway, based on available Global Biodiversity Information Facility (GBIF) records and the

260　published literature, which could be due to the general difficulty of observing and identifying

261　this algae [47,50]. We note, however, that our reanalysis of modern DNA metabarcoding data

262　from 11 north Norwegian localities [57] shows the presence of *Nannochloropsis* at five sites

263　(Supplementary Table S11), with dominant abundances detected for two sites. In addition to

8

264   LGM Andøya [35], *Nannochloropsis* has either been previously reported, or unreported but

265   present based on our re-analysis, in eight *sed*aDNA-based palaeoecological records from

266   Greenland [58], St. Paul Island, Alaska, USA [9,11], Alberta, Canada [10], Latvia [59], Qinghai, China

267   [60] and Svalbard [52,61] (Supplementary Figure S11). We failed to detect *Nannochloropsis* in the

268   Hässeldala Port, Sweden [13] record. We note that *Nannochloropsis* is particularly well

269   represented in late Pleistocene and early Holocene sediments from these records, at a time

270   when the climate was cooler than present [62]. Assuming these records reflect *N. limnetica*, or

271   an ecological analogue, then these occurrences are consistent with its known climatic

272   tolerances, such as thriving in cold water [43]. Therefore, climate would have been adequate for

273   *N. limnetica* at Andøya during the LGM, whereas the high nutrient input could have

274   stimulated the unusually high concentrations.

275   As *Nannochloropsis* taxa differ in their salinity tolerances, the ability to detect and

276   identify them to the species-level could potentially be used as a palaeoecological proxy to

277   estimate the salinity of coastal marine-lacustrine sedimentary records. The detection of the

278   fresh or brackish water *N. limnetica* in Lake Øvre Æråsvatnet is consistent with earlier

279   studies that indicated a lacustrine LGM sediment record [25,32,34]. However, caution should be

280   applied when assuming ecological preferences of a taxon that is evolutionarily divergent from

281   reference sequences, as is the case here.

282   Our complete *N. limnetica* chloroplast palaeogenome reconstructions represent the

283   first derived from *sed*aDNA, although a near-complete chloroplast sequence has recently

284   been reported for a vascular plant [14]. Although mitochondrial palaeogenomes have previously

285   been reconstructed from cave sediments [12], and archaeological middens and latrines [24,63], ours

286   are the first derived from lake sediments. The high depth of coverage for our sample-

287   combined palaeogenomes (62-64x) allowed us to explore allelic proportions and haplotype

288   diversity using *sed*aDNA, which resulted in us identifying at least two distinct variants.

289   Together with recently published and ongoing studies, our work demonstrates the feasibility

290   of the *sed*aDNA field moving into a new phase of environmental palaeogenomics. This will

291   enable a broad range of ecological and evolutionary questions to be addressed using

292   population genomic approaches, including for communities of taxa that may or may not be

293   preserved in the body fossil record. With further innovations, this approach could also be

294   extended to a suite of broad groups, including plants, invertebrates, and vertebrates, from lake

295   catchments, cave sediments, and archaeological settings, therefore unlocking the full

296   potential of *sed*aDNA.

297

298   Material and methods

299

300   **Site description, chronology, and sampling**

301   A detailed description of the site, coring methods, age-depth model reconstruction, and

302   sampling strategy can be found in [35]. Briefly, Lake Øvre Æråsvatnet is located on Andøya,

303   Northern Norway (69.25579°N, 16.03517°E) (Figure 1). In 2013, two cores were collected

304   from the deepest sediments, AND10 and AND11. Macrofossil remains were dated, with

305   those from AND10 all dating to within the LGM. For the longer core AND11, a Bayesian

306   age-depth model was required to estimate the age of each layer [35]. In this study, we selected

307   one sample of LGM sediments from each of the two cores. According to the Bayesian age-

308   depth model, sample Andøya_LGM_B, from 1102 cm depth in AND11, was dated to a

309   median age of 17,700 (range: 20,200-16,500) cal yr BP. The age of Andøya_LGM_A, from

310   938 cm depth in AND10, was estimated at 19,500 cal yr BP, based on the interpolated

311   median date between two adjacent macrofossils (20 cm above: 19,940-18,980 cal yr BP, 30

312   cm below: 20,040-19,000 cal yr BP). As Andøya_LGM_A falls within the age range of

313   Andøya_LGM_B, we consider the samples to be broadly contemporaneous.

314

315   **Sampling, DNA extraction, library preparation, and sequencing**

316   The two cores were subsampled at the selected layers under clean conditions, in a dedicated

317   ancient DNA laboratory at The Arctic University Museum of Norway in Tromsø. We

318   extracted DNA from 15 g of sediment following the Taberlet phosphate extraction protocol [18]

319   in the same laboratory. We shipped a 210 µL aliquot of each DNA extract to the ancient

320   DNA dedicated laboratories at the Centre for GeoGenetics (University of Copenhagen,

321   Denmark) for double-stranded DNA library construction. After concentrating the DNA

322   extracts to 80 µL, half of each extract (40 µL, totalling between 31.7-36.0 ng of DNA) was

323   converted into Illumina-compatible libraries using established protocols [10]. The number of

324   indexing PCR cycles was determined using qPCR and each sample was dual indexed. The

325   libraries were then purified using the AmpureBead protocol (Beckman Coulter, Indianapolis,

326   IN, USA), adjusting the volume ratio to 1:1.8 library:AmpureBeads, and quantified using a

327   BioAnalyzer (Agilent, Santa Clara, CA, USA). The indexed libraries were pooled

328   equimolarly and sequenced on a lane of the Illumina HiSeq 2500 platform using 2x 80 cycle

329   paired-end chemistry.

330

10

Raw read filtering

For each sample, we merged and adapter-trimmed the paired-end reads with *SeqPrep* (https://github.com/jstjohn/SeqPrep/releases, v1.2) using default parameters. We only retained the resulting merged sequences, which were then filtered with the preprocess function of the *SGA toolkit* v0.10.15 [64] by the removal of those shorter than 35 bp or with a DUST complexity score >1.

Metagenomic analysis of the sequence data

We first sought to obtain an overview of the taxonomic composition of the samples and therefore carried out a BLAST-based metagenomic analysis on the two filtered sequence datasets. To make the datasets more computationally manageable, we subsampled the first and last one million sequences from the filtered dataset of each sample and analysed each separately. The data subsets were each identified against the NCBI nucleotide database (release 223) using the blastn function from the *NCBI-BLAST+* suite v2.2.18+ [65] under default settings. For each sample, the results from the two subsets were checked for internal consistency, merged into one dataset, and loaded into *MEGAN* v6.12.3 [66]. Analysis and visualization of the Last Common Ancestor (LCA) was carried out for the taxonomic profile using the following settings: min score=35, max expected=1.0E-5, min percent identity=95%, top percent=10%, min support percentage=0.01, LCA=naive, min percent sequence to cover=95%. We define sequences as the reads with BLAST hits assigned to taxa post-filtering, thus ignoring "unassigned" and "no hit" categories.

Alignment to reference genome panels

We mapped our filtered data against three different reference panels to help improve taxonomic identifications and provide insight into the sequence abundance of the identified taxa (Supplementary Tables S4 and S5). The first reference panel consisted of 42 nuclear genomes that included taxa expected from Northern Norway, exotic/implausible taxa, human, six *Nannochloropsis* species, and four strains of *Mycobacterium*. The inclusion of exotic taxa was to give an indication of the background spurious mapping rate, which can result from mappings to conserved parts of the genome and/or short and damaged ancient DNA molecules [53,54]. We included *Nannochloropsis*, *Mycobacterium*, and human genomes, due to their overrepresentation in the BLAST-based metagenomic analysis. The other two reference panels were based on either all mitochondrial or chloroplast genomes on NCBI GenBank (as of January 2018). The chloroplast data set was augmented with 247 partial or complete

365 chloroplast genomes generated by the PhyloNorway project [67]. The filtered data were mapped

366 against each reference genome or organellar genome set individually using *bowtie2* v2.3.4.1

367 [68] under default settings. The resulting bam files were processed with *SAMtools* v0.1.19 [69].

368 We removed unmapped sequences with *SAMtools view* and collapsed PCR duplicate

369 sequences with *SAMtools rmdup*.

370   For the nuclear reference panel, we reduced potential spurious or nonspecific

371 sequence mappings by comparing the mapped sequences to both the aligned reference

372 genome and the NCBI nucleotide database using *NCBI-BLAST+*, following the method used

373 by Graham *et al.* [9], as modified by Wang *et al.* [11]. The sequences were aligned using the

374 following *NCBI-BLAST+* settings: num_alignments=100 and perc_identity=90. Sequences

375 were retained if they had better alignments, based on bit score, to reference genomes as

376 compared to the NCBI nucleotide database. If a sequence had a better or equal match against

377 the NCBI nucleotide database, it was removed, unless the LCA of the highest NCBI

378 nucleotide bit score was from the same genus as the reference genome (based on the NCBI

379 taxonID). To standardize the relative mapping frequencies to genomes of different size, we

380 calculated the number of retained mapped sequences per Mb of genome sequence.

381   The sequences mapped against the chloroplast and mitochondrial reference panels

382 were filtered and reported in a different manner than the nuclear genomes. First, to exclude

383 any non-eukaryotic sequences, we used *NCBI-BLAST+* to search sequence taxonomies and

384 retained sequences if the LCA was, or was within, Eukaryota. Second, for the sequences that

385 were retained, the LCA was calculated and reported in order to summarize the mapping

386 results across the organelle datasets. LCAs were chosen as the reference sets are composed of

387 multiple genera.

388   Within the *Nannochloropsis* nuclear reference alignments, the relative mapping

389 frequency was highest for *N. limnetica*. In addition, the relative mapping frequency for other

390 *Nannochloropsis* taxa was higher than those observed for the exotic taxa. This could

391 represent the mapping of sequences that are conserved between *Nannochloropsis* genomes or

392 suggest the presence of multiple *Nannochloropsis* taxa in a community sample. We therefore

393 cross-compared mapped sequences to determine the number of uniquely mapped sequences

394 per reference genome. First, we individually remapped the filtered data to six available

395 *Nannochloropsis* nuclear genomes, the accession codes of which are provided in

396 Supplementary Table S4. For each sample, we then calculated the number of sequences that

397 uniquely mapped to, or overlapped, between each *Nannochloropsis* genome. We repeated the

12

398 above analysis with six available chloroplast sequences (Supplementary Table S4), to get a

399 comparable overlap estimation for the chloroplast genome.

400

401 **Reconstruction of the Andøya *Nannochloropsis* community organellar palaeogenomes**

402 To place the Andøya *Nannochloropsis* community taxon into a phylogenetic context, and

403 provide suitable reference sequences for variant calling, we reconstructed environmental

404 palaeogenomes for the *Nannochloropsis* mitochondria and chloroplast. First, the raw read

405 data from both samples were combined into a single dataset and re-filtered with the *SGA*

406 *toolkit* to remove sequences shorter than 35 bp, but retain low complexity sequences to assist

407 in the reconstruction of low complexity regions in the organellar genomes. This re-filtered

408 sequence data set was used throughout the various steps for environmental palaeogenome

409 reconstruction.

410 The re-filtered sequence data were mapped onto the *N. limnetica* reference chloroplast

411 genome (NCBI GenBank accession: NC_022262.1) with *bowtie2* using default settings.

412 *SAMtools* was used to remove unmapped sequences and PCR duplicates, as above. We

413 generated an initial consensus genome from the resulting bam file with *BCFtools* v1.9 [69],

414 using the *mpileup*, *call*, *filter*, and *consensus* functions. For variable sites, we produced a

415 majority-rule consensus using the --variants-only and --multiallelic-caller options, and for

416 uncovered sites the reference genome base was called. The above steps were repeated until

417 the consensus could no longer be improved. The re-filtered sequence data was then re-

418 mapped onto the initial consensus genome sequence with *bowtie2*, using the above settings.

419 The *genomecov* function from *BEDtools* v2.17.0 [70] was used to identify gaps and low

420 coverage regions in the resulting alignment.

421 We attempted to fill the identified gaps, which likely consisted of diverged or

422 difficult-to-assemble regions. For this, we assembled the re-filtered sequence dataset into *de*

423 *novo* contigs with the MEGAHIT pipeline v1.1.4 [71], using a minimum *k*-mer length of 21, a

424 maximum *k*-mer length of 63, and *k*-mer length increments of six. The MEGAHIT contigs

425 were then mapped onto the initial consensus genome sequence with the *blastn* tool from the

426 *NCBI-BLAST+* toolkit. Contigs that covered the gaps identified by *BEDtools* were

427 incorporated into the initial consensus genome sequence, unless a *blast* comparison against

428 the NCBI nucleotide database suggested a closer match to non-*Nannochloropsis* taxa. We

429 repeated the *bowtie2* gap-filling steps iteratively, using the previous consensus sequence as

430 reference, until a gap-free consensus was obtained. The re-filtered sequence data were again

431 mapped, the resulting final assembly was visually inspected, and the consensus was corrected

432 where necessary. This was to ensure the fidelity of the consensus sequence, which

433 incorporated *de novo*-assembled contigs that could potentially be problematic, due to the

434 fragmented nature and deaminated sites of ancient DNA impeding accurate assembly [72].

435 Annotation of the chloroplast genome was carried out with *GeSeq* [73], using the

436 available annotated *Nannochloropsis* chloroplast genomes (accession codes provided in

437 Supplementary Table S12). The resulting annotated chloroplast was visualised with

438 *OGDRAW* [74].

439 The same assembly and annotation methods outlined above were used to reconstruct

440 the mitochondrial palaeogenome sequence, where the initial mapping assembly was based on

441 the *N. limnetica* mitochondrial sequence (NCBI GenBank accession: NC_022256.1). The

442 final annotation was carried out by comparison against all available annotated

443 *Nannochloropsis* mitochondrial genomes (accession codes provided in Supplementary Table

444 S12).

445 If the *Nannochloropsis* sequences derived from more than one taxon, then alignment

446 to the *N. limnetica* chloroplast genome could introduce reference bias, which would

447 underestimate the diversity of the *Nannochloropsis* sequences present. We therefore

448 reconstructed *Nannochloropsis* chloroplast genomes, but using the six available

449 *Nannochloropsis* chloroplast genome sequences, including *N. limnetica*, as seed genomes

450 (accession codes for the reference genomes are provided in Supplementary Table S8). The

451 assembly of the consensus sequences followed the same method outlined above, but with two

452 modifications to account for the mapping rate being too low for complete genome

453 reconstruction based on alignment to the non-*N. limnetica* reference sequences. First,

454 consensus sequences were called with *SAMtools*, which does not incorporate reference bases

455 into the consensus at uncovered sites. Second, neither additional gap filling, nor manual

456 curation was implemented.

457

458 **Assembly of high and low frequency variant consensus sequences**

459 The within-sample variants in each reconstructed organellar palaeogenome was explored by

460 creating two consensus sequences, which included either high or low frequency variants at

461 multiallelic sites. For each sample, the initial filtered sequence data were mapped onto the

462 reconstructed *Nannochloropsis* chloroplast palaeogenome sequence with *bowtie2* using

463 default settings. Unmapped and duplicate sequences were removed with *SAMtools*, as above.

464 We used the *BCFtools mpileup*, *call*, and *normalize* functions to identify the variant sites in

465 the mapped dataset, using the --skip-indels, --variants-only, and --multiallelic-caller options.

14

466   The resulting alleles were divided into two sets, based on either high or low frequency

467   variants. High frequency variants were defined as those present in the reconstructed reference

468   genome sequence. Both sets were further filtered to only include sites with a quality score of

469   30 or higher and a coverage of at least half the average coverage of the mapping assembly

470   (minimum coverage: Andøya_LGM_A=22x, Andøya_LGM_B=14x). We then generated the

471   high and low frequency variant consensus sequences using the consensus function in

472   *BCFTools*. The above method was repeated for the reconstructed *Nannochloropsis*

473   mitochondrial genome sequence in order to generate comparable consensus sequences of

474   high and low frequency variants (minimum coverage: Andøya_LGM_A=16x,

475   Andøya_LGM_B=10x).

476

Analysis of ancient DNA damage patterns

478   We checked for the presence of characteristic ancient DNA damage patterns for sequences

479   aligned to four nuclear genomes: human, *Nannochloropsis limnetica* and *Mycobacterium*

480   *avium*. We further analysed damage patterns for sequences aligned to both the reconstructed

481   *N. limnetica* composite organellar genomes. Damage analysis was conducted with

482   *mapDamage* v2.0.8 [75] using the following settings: --merge-reference-sequences and --

483   length=160.

484

Phylogenetic analysis of the reconstructed organellar palaeogenomes

486   We determined the phylogenetic placement of our high and low frequency variant organellar

487   palaeogenomes within *Nannochloropsis*, using either full mitochondrial and chloroplast

488   genome sequences or three short loci (18S, ITS, *rbc*L). We reconstructed the 18S and ITS1-

489   5.8S-ITS2 complex using DQ977726.1 (full length) and EU165325.1 (positions 147:1006,

490   corresponding to the ITS complex) as seed sequences following the same approach that was

491   used for the organellar palaeogenome reconstructions, except that the first and last 10 bp

492   were trimmed to account for the lower coverage due to sequence tiling. We then called high

493   and low variant consensus sequences as described above.

494   We created six alignments using available sequence data from NCBI Genbank

495   (Supplementary Tables S7) with the addition of: (1+2) the high and low frequency variant

496   chloroplast or mitochondrial genome consensus sequences, (3) a ~1100 bp subset of the

497   chloroplast genome for the *rbc*L alignment, (4+5) ~1800 bp and ~860 bp subsets of the

498   nuclear multicopy complex for the 18S and ITS alignments, respectively, and (6) the

499   reconstructed chloroplast genome consensus sequences derived from the alternative

15

500 *Nannochloropsis* genome starting points. Full details on the coordinates of the subsets are

501 provided in Supplementary Table S7. We generated alignments using *MAFFT* v7.427 [76] with

502 the maxiterate=1000 setting, which was used for the construction of a maximum likelihood

503 tree in *RAxML* v8.1.12 [77] using the GTRGAMMA model and without outgroup specified. We

504 assessed branch support using 1000 replicates of rapid bootstrapping.

505

506 ***Nannochloropsis* variant proportions and haplogroup diversity estimation**

507 To estimate major haplogroup diversity, we calculated the proportions of high and low

508 variants in the sequences aligned to our reconstructed *Nannochloropsis* mitochondrial and

509 chloroplast genomes. For each sample, we first mapped the initial filtered sequence data onto

510 the high and low frequency variant consensus sequences with *bowtie2*. To avoid potential

511 reference biases, and for each organellar genome, the sequence data were mapped separately

512 against both frequency consensus sequences. The resulting bam files were then merged with

513 *SAMtools* merge. We removed exact sequence duplicates, which may have been mapped to

514 different coordinates, from the merged bam file by randomly retaining one copy. This step

515 was replicated five times to examine its impact on the estimated variant proportions. After

516 filtering, remaining duplicate sequences - those with identical mapping coordinates - were

517 removed with *SAMtools* rmdup. We then called variable sites from the duplicate-removed

518 bam files using *BCFTools* under the same settings as used in the assembly of the high and

519 low frequency variant consensus sequences. We restricted our analyses to transversion-only

520 variable positions, to remove the impact of ancient DNA deamination artifacts. For each

521 variable site, the proportion of reference and alternative alleles was calculated, based on

522 comparison to the composite *N. limnetica* reconstructed organellar palaeogenomes. We

523 removed rare alleles occurring at a proportion of <0.1, as these may have resulted from noise.

524   To infer the minimum number of haplogroups in each reconstructed organellar

525 genome sequence, we inspected the phasing of adjacent variable sites that were linked by the

526 same read in the duplicate-removed bam files, akin to the method used by Søe *et al*. [63]. For

527 this, we first identified all positions, from both samples, where two or more transversion-only

528 variable sites occurred within 35 bp windows. We then examined the allelic state in mapped

529 sequences that fully covered each of these linked positions. We recorded the combination of

530 alleles to calculate the observed haplotype diversity at each of the linked positions. We

531 removed low frequency haplotypes, which were defined as those with <3 sequences or <15%

532 of all sequences that covered a linked position, and the remaining haplotypes were scored.

533

Meta-analysis of *Nannochloropsis* in previous *sed*aDNA data sets

We performed a meta-analysis of the global prevalence of *Nannochloropsis* since the last ice age using published and available lake *sed*aDNA data sets. Three published shotgun datasets from Lake Hill, Alaska, USA [9,11], Charlie Lake, BC and Spring Lake, Alberta, Canada [10], and Hässeldala Port, Sweden [13] were reanalysed for the presence of *N. limnetica* using the same nuclear genome method as used in this study (Supplementary Table S13). Furthermore, a metabarcode data set was reanalysed from Skartjørna, Svalbard [61], using the same methods for analysis as the original study, but lowering the minimum barcode length to 10 bp, in order to retain the *Nannochloropsis* barcode (tag-sample lookup is provided in Supplementary Table S14). These data sets were supplemented with *sed*aDNA metabarcoding studies that reported *Nannochloropsis*, including; Bliss Lake, Greenland [58], Qinghai Lake, China [60], Lielais Svĕtiņu, Latvia [59], Lake Øvre Æråsvatnet [35], and Jodavannet, Svalbard [52].

We estimated the occurrence and abundance of *Nannochloropsis* in 5,000-year time windows for the above data sets. Abundance was coarsely divided into four categories for the metabarcode data: (1) dominant, scored when *Nannochloropsis* was the only taxon detected or most abundant of the taxa identified in the sequence data; (2) common, assigned when it was in the top 10 most abundant taxa identified; (3) rare, scored for any other detections, and (4) absent, assigned if *Nannochloropsis* was not detected. The reanalysed shotgun data sets were scored as: (1) dominant, when *Nannochloropsis* made up >=0.1% of the filtered read data; (2) common, 0.09-0.01%; (3) rare, 0.009-0.001%, and (4) absent, with <0.001%.

Data availability

The raw Illumina shotgun sequence datasets are available from EMBL via *ACCESSION CODES*. The reanalysed metabarcoding data from Alsos et al. [61] are available via *ACCESSION CODE*. The reconstructed *Nannochloropsis limnetica* high and low frequency organellar genome sequences are available from NCBI Genbank via *ACCESSION CODES*. The scripts estimating the number of haplotypes across the linked windows are provided in the following GitHub repository at *GITHUB LINK*.

564   References

565   1.   Shapiro, B. & Hofreiter, M. A paleogenomic perspective on evolution and gene function:
566   new insights from ancient DNA. *Science* **343**, 1236573 (2014).

567   2.   Palkopoulou, E. *et al.* Complete Genomes Reveal Signatures of Demographic and
568   Genetic Declines in the Woolly Mammoth. *Curr. Biol.* **25**, 1395–1400 (2015).

569   3.   Slon, V. *et al.* The genome of the offspring of a Neanderthal mother and a Denisovan
570   father. *Nature* **561**, 113–116 (2018).

571   4.   Allaby, R. G., Smith, O. & Kistler, L. Archaeogenomics and Crop Adaptation: Genome-
572   Scale Analysis of Ancient DNA. in *Paleogenomics* (eds. Lindqvist, C. & Rajora, O. P.)
573   vol. 61 189–203 (Springer International Publishing, 2019).

574   5.   Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan
575   individual. *Science* **338**, 222–226 (2012).

576   6.   Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172
577   (2015).

578   7.   Meyer, M. *et al.* Nuclear DNA sequences from the Middle Pleistocene Sima de los
579   Huesos hominins. *Nature* **531**, 504–507 (2016).

580   8.   Smith, O. *et al.* Sedimentary DNA from a submerged site reveals wheat in the British
581   Isles 8000 years ago. *Science* **347**, 998–1001 (2015).

582   9.   Graham, R. W. *et al.* Timing and causes of mid-Holocene mammoth extinction on St.
583   Paul Island, Alaska. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 9310–9314 (2016).

584   10.   Pedersen, M. W. *et al.* Postglacial viability and colonization in North America's ice-free
585   corridor. *Nature* **537**, 45–49 (2016).

586   11.   Wang, Y. *et al.* The southern coastal Beringian land bridge: cryptic refugium or
587   pseudorefugium for woody plants during the Last Glacial Maximum? *J. Biogeogr.* **44**,
588   1559–1571 (2017).

589   12.   Slon, V. *et al.* Neandertal and Denisovan DNA from Pleistocene sediments. *Science* **356**,
590   605–608 (2017).

591   13.   Parducci, L. *et al.* Shotgun Environmental DNA, Pollen, and Macrofossil Analysis of
592   Lateglacial Lake Sediments From Southern Sweden. *Front. Ecol. Evol.* **7**, (2019).

593   14.   Schulte, L. *et al.* Hybridization capture of larch (*Larix* Mill) chloroplast genomes from
594   sedimentary ancient DNA reveals past changes of Siberian forests. *BioRxiv* (2020)
595   doi:10.1101/2020.01.06.896068.

596   15.   Pedersen, M. W. *et al.* Ancient and modern environmental DNA. *Philos. Trans. R. Soc.*

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

598   16. Parducci, L. *et al.* Ancient plant DNA in lake sediments. *New Phytol.* **214**, 924–942
599   (2017).

600   17. Willerslev, E. *et al.* Fifty thousand years of Arctic vegetation and megafaunal diet.
601   *Nature* **506**, 47–51 (2014).

602   18. Taberlet, P. *et al.* Soil sampling and isolation of extracellular DNA from large amount of
603   starting material suitable for metabarcoding studies. *Mol. Ecol.* **21**, 1816–1820 (2012).

604   19. Taberlet, P., Bonin, A., Zinger, L. & Coissac, E. *Environmental DNA for functional*
605   *diversity*. (Oxford University Press, 2018).

606   20. Nichols, R. V., Curd, E., Heintzman, P. D. & Shapiro, B. Targeted Amplification and
607   Sequencing of Ancient Environmental and Sedimentary DNA. in *Ancient DNA* 149–161
608   (Humana Press, 2019).

609   21. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal.
610   *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14616–14621 (2007).

611   22. Skoglund, P. *et al.* Separating endogenous ancient DNA from modern day contamination
612   in a Siberian Neandertal. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2229–2234 (2014).

613   23. Zinger, L. *et al.* DNA metabarcoding—Need for robust experimental designs to draw
614   sound ecological conclusions. *Mol. Ecol.* **28**, 1857–1862 (2019).

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

617   25. Vorren, T. O., Rydningen, T. A., Baeten, N. J. & Laberg, J. S. Chronology and extent of
618   the Lofoten-Vesterålen sector of the Scandinavian Ice Sheet from 26 to 16 cal. ka BP.
619   *Boreas* **44**, 445–458 (2015).

620   26. Parducci, L. *et al.* Glacial survival of boreal trees in northern Scandinavia. *Science* **335**,
621   1083–1086 (2012).

622   27. Birks, H. H. *et al.* Comment on 'Glacial survival of boreal trees in northern
623   Scandinavia'. *Science* **338**, 742; author reply 742 (2012).

624   28. Parducci, L. *et al.* Response to Comment on 'Glacial Survival of Boreal Trees in
625   Northern Scandinavia'. *Science* **338**, 742–742 (2012).

626   29. Vorren, K.-D. Late and Middle Weichselian stratigraphy of Andøya, north Norway.
627   *Boreas* **7**, 19–38 (1978).

628   30. Vorren, T. O., Vorren, K.-D., Torbjørn, A., Gulliksen, S. & Løvlie, R. The last
629   deglaciation (20,000 to 11,000 B. P.) on Andøya, northern Norway. *Boreas* **17**, 41–77
630   (1988).

631 31. Alm, T. & Birks, H. H. Late Weichselian flora and vegetation of Andøya, Northern
632   Norway-macrofossil (seed and fruit) evidence from Nedre Æråsvatn. *Nord. J. Bot.* **11**,
633   465–476 (1991).

634 32. Alm, T. Øvre Æråsvatn - palynostratigraphy of a 22,000 to 10,000 BP lacustrine record
635   on Andøya, northern Norway. *Boreas* **22**, 171–188 (1993).

636 33. Elverland, E. & Alm, T. PhD chapter: A Late Weichselian Alle alle colony on Andøya,
637   northern Norway - a contribution to the history of an important Arctic environment. (UiT
638   - The Arctic University of Norway, 2012).

639 34. Vorren, T. O. *et al.* Palaeoenvironment in northern Norway between 22.2 and 14.5 cal.
640   ka BP. *Boreas* **42**, 876–895 (2013).

641 35. Alsos, I. G. *et al.* Late Glacial Maximum environmental condition of Andøya, a northern
642   ecological 'hotspot'. *Quaternary Science Reviews* (2020).

643 36. Su, C.-H. *et al.* Factors affecting lipid accumulation by *Nannochloropsis oculata* in a
644   two-stage cultivation process. *J. Appl. Phycol.* **23**, 903–908 (2011).

645 37. Radakovits, R. *et al.* Draft genome sequence and genetic transformation of the
646   oleaginous alga *Nannochloropsis gaditana*. *Nat. Commun.* **3**, (2012).

647 38. Vieler, A. *et al.* Genome, functional gene annotation, and nuclear transformation of the
648   heterokont oleaginous alga *Nannochloropsis oceanica* CCMP1779. *PLoS Genet.* **8**,
649   e1003064 (2012).

650 39. Wei, L. *et al. Nannochloropsis* plastid and mitochondrial phylogenomes reveal organelle
651   diversification mechanism and intragenus phylotyping strategy in microalgae. *BMC*
652   *Genom.* **14**, 534 (2013).

653 40. Schwartz, A. S. *et al.* Complete genome sequence of the model oleaginous alga
654   *Nannochloropsis gaditana* CCMP1894. *Genome Announc.* **6**, (2018).

655 41. Krienitz, L., Hepperle, D., Stich, H.-B. & Weiler, W. *Nannochloropsis limnetica*
656   (Eustigmatophyceae), a new species of picoplankton from freshwater. *Phycologia* **39**,
657   219–227 (2000).

658 42. Fietz, S. *et al.* First record of *Nannochloropsis limnetica* (Eustigmatophyceae) in the
659   autotrophic picoplankton from Lake Baikal. *J. Phycol.* **41**, 780–790 (2005).

660 43. Fawley, K. P. & Fawley, M. W. Observations on the diversity and ecology of freshwater
661   *Nannochloropsis* (Eustigmatophyceae), with descriptions of new taxa. *Protist* **158**, 325–
662   336 (2007).

663 44. Hibberd, D. J. Notes on the taxonomy and nomenclature of the algal classes
664   Eustigmatophyceae and Tribophyceae (synonym Xanthophyceae). *Bot. J. Linn. Soc.* **82**,

665   93–119 (1981).

666   45. Karlson, B., Potter, D., Kuylenstierna, M. & Andersen, R. A. Ultrastructure, pigment
667   composition, and 18S rRNA gene sequence for *Nannochloropsis granulata sp. nov.*
668   (Monodopsidaceae, Eustigmatophyceae), a marine ultraplankter isolated from the
669   Skagerrak, northeast Atlantic Ocean. *Phycologia* **35**, 253–260 (1996).

670   46. Suda, S., Atsumi, M. & Miyashita, H. Taxonomic characterization of a marine
671   *Nannochloropsis* species, *N. oceanica sp. nov.* (Eustigmatophyceae). *Phycologia* **41**,
672   273–279 (2002).

673   47. Andersen, R. A., Brett, R. W., Potter, D. & Sexton, J. P. Phylogeny of the
674   Eustigmatophyceae Based upon 18S rDNA, with Emphasis on Nannochloropsis. *Protist*
675   **149**, 61–74 (1998).

676   48. Karlov, D. S. *et al.* Microbial communities within the water column of freshwater Lake
677   Radok, East Antarctica: predominant 16S rDNA phylotypes and bacterial cultures. *Polar*
678   *Biology* **40**, 823–836 (2017).

679   49. Smol, J. P., Birks, H. J. & Last, W. M. *Tracking Environmental Change Using Lake*
680   *Sediments: Volume 4: Zoological Indicators.* (Springer Netherlands, 2001).

681   50. Gladu, P. K., Patterson, G. W., Wikfors, G. H. & Smith, B. C. Sterol fatty acid and
682   pigment characteristics of UTEX 2341 a marine eustigmatophyte identified previously as
683   *Chlorella minutissima* (Chlorophyceae). *J. Phycol.* **31**, 774–777 (1995).

684   51. Kryvenda, A., Rybalka, N., Wolf, M. & Friedl, T. Species distinctions among closely
685   related strains of Eustigmatophyceae (Stramenopiles) emphasizing ITS2 sequence-
686   structure data: *Eustigmatos* and *Vischeria. Eur. J. Phycol.* **53**, 471–491 (2018).

687   52. Voldstad, L. H. *et al.* A complete Holocene lake sediment ancient DNA record reveals
688   long-standing high Arctic plant diversity hotspot in northern Svalbard. *Quat. Sci. Rev.*
689   **234**, (2020).

690   53. Prüfer, K. *et al.* Computational challenges in the analysis of ancient DNA. *Genome Biol.*
691   **11**, R47 (2010).

692   54. Orlando, L., Gilbert, M. T. P. & Willerslev, E. Reconstructing ancient genomes and
693   epigenomes. *Nat. Rev. Genet.* **16**, 395–408 (2015).

694   55. Zhang, X. *et al.* First record of a large-scale bloom-causing species *Nannochloropsis*
695   *granulata* (Monodopsidaceae, Eustigmatophyceae) in China Sea waters. *Ecotoxicology*
696   **24**, 1430–1441 (2015).

697   56. Andreoli, C. *et al.* A Survey on a Persistent Greenish Bloom in the Comacchio Lagoons
698   (Ferrara, Italy). *Bot. Mar.* **42**, (1999).

699   57. Alsos, I. G. *et al.* Plant DNA metabarcoding of lake sediments: How does it represent the
700   contemporary vegetation. *PLoS One* **13**, e0195403 (2018).

701   58. Epp, L. S. *et al.* Lake sediment multi-taxon DNA from North Greenland records early
702   post-glacial appearance of vascular plants and accurately tracks environmental changes.
703   *Quat. Sci. Rev.* **117**, 152–163 (2015).

704   59. Stivrins, N. *et al.* Towards understanding the abundance of non-pollen palynomorphs: A
705   comparison of fossil algae, algal pigments and seda DNA from temperate lake
706   sediments. *Rev. Palaeobot. Palynol.* **249**, 9–15 (2018).

707   60. Li, G. *et al.* Temporal Succession of Ancient Phytoplankton Community in Qinghai
708   Lake and Implication for Paleo-environmental Change. *Sci. Rep.* **6**, 19769 (2016).

709   61. Alsos, I. G. *et al.* Sedimentary ancient DNA from Lake Skartjørna, Svalbard: Assessing
710   the resilience of arctic flora to Holocene climate change. *Holocene* **26**, 627–642 (2016).

711   62. Rasmussen, S. O. *et al.* A stratigraphic framework for abrupt climatic changes during the
712   Last Glacial period based on three synchronized Greenland ice-core records: refining and
713   extending the INTIMATE event stratigraphy. *Quat. Sci. Rev.* **106**, 14–28 (2014).

714   63. Søe, M. J. *et al.* Ancient DNA from latrines in Northern Europe and the Middle East
715   (500 BC-1700 AD) reveals past parasites and diet. *PLoS One* **13**, e0195481 (2018).

716   64. Simpson, J. T. & Durbin, R. Efficient de novo assembly of large genomes using
717   compressed data structures. *Genome Res.* **22**, 549–556 (2012).

718   65. Camacho, C. *et al.* BLAST : architecture and applications. *BMC Bioinform.* **10**, 421
719   (2009).

720   66. Huson, D. H. *et al.* MEGAN Community Edition - Interactive Exploration and Analysis
721   of Large-Scale Microbiome Sequencing Data. *PLoS Comput. Biol.* **12**, e1004957 (2016).

722   67. Alsos, I. G. *et al.* The treasure vault can be opened: large scale genome skimming works
723   equally well for herbarium as silica gel dried material. *Plants* (2020).

724   68. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat.*
725   *Methods* **9**, 357–359 (2012).

726   69. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
727   2078–2079 (2009).

728   70. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing
729   genomic features. *Bioinformatics* **26**, 841–842 (2010).

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

733   72. Seitz, A. & Nieselt, K. Improving ancient DNA genome assembly. *PeerJ* **5**, e3126
734   (2017).
735   73. Tillich, M. *et al.* GeSeq - versatile and accurate annotation of organelle genomes.
736   *Nucleic Acids Res.* **45**, W6–W11 (2017).
737   74. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version
738   1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic*
739   *Acids Res.* **47**, W59–W64 (2019).
740   75. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L.
741   mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage
742   parameters. *Bioinformatics* **29**, 1682–1684 (2013).
743   76. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
744   improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
745   77. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
746   large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
747   78. Hughes, A. L. C., Gyllencreutz, R., Lohne, Ø. S., Mangerud, J. & Svendsen, J. I. The last
748   Eurasian ice sheets - a chronological database and time-slice reconstruction, DATED-1.
749   *Boreas* **45**, 1–45 (2016).

750   ## Acknowledgements

762

763 ## Author contribution

764 YL, PDH, and IGA conceptualised and designed the research, and contributed to the final

765 version of the manuscript; YL analysed the data and wrote the first draft; PDH provided

766 analytical guidance and refined the drafted manuscript; IGA performed fieldwork, DNA

767 extractions, provided resources, acquired funding, and supervised YL.

768

769 ## Ethics declarations

770 **Competing interests**

771 The author(s) declare no competing interests.

772

773 ## Tables and figures

774 **Table 1:** Summary of the best represented taxa (>200 identified sequences) detected in the

775 metagenomic analysis. N=Number of identified sequences, I=Percentage of identified

776 sequences, A=Percentage of all sequences included in the metagenomic analysis.

	Andøya_LGM_A			Andøya_LGM_B		
	N	**I**	**A**	**N**	**I**	**A**
Bacteria	18,852	63.93	0.94	21,873	66.85	1.09
Mycobacterium	6268	21.26	0.31	8535	26.08	0.43
M. avium complex	444	1.51	0.02	635	1.94	0.03
M. dioxanotrophicus	0	0	0	229	0.7	0.01
M. sp. EPa45	0	0	0	306	0.94	0.02
M. sp. YC-RL4	0	0	0	207	0.63	0.01
Pseudomonas	920	3.12	0.05	904	2.76	0.05
Eukaryota	9333	31.65	0.47	9563	29.23	0.48
Nannochloropsis	5913	20.05	0.3	6179	18.88	0.31
N. limnetica	2223	7.54	0.11	2272	6.94	0.11
Other	1303	4.42	0.07	1284	3.92	0.06
Total	29,488	100	1.47	32,720	100	1.64

777

24

778 **Table 2:** Placement of the reconstructed high and low frequency organellar genomes and markers in each phylogenetic analysis. BSS=Bootstrap

779 support values, HFV=High frequency variant, LFV=Low frequency variant.

| Sample | Chloroplast | | | | Mitochondrion | | Nuclear | | | | Consensus |
| | whole | | rbcL | | whole | | 18S | | ITS | | |
	placement	BSS (%)	placement	BSS (%)	placement	BSS (%)	placement	BSS (%)	placement	BSS (%)	
Andøya_LGM_A HFV	sister to *N. limnetica*	100	*N. limnetica* var. *globosa*	57	sister to *N. limnetica*	83	*N. limnetica*	76	sister to *N. limnetica*	98	*N. limnetica* var. *globosa*
Andøya_LGM_B HFV	sister to *N. limnetica*	100	*N. limnetica* var. *globosa*	57	sister to *N. limnetica*	83	*N. limnetica*	76	sister to *N. limnetica*	98	*N. limnetica* var. *globosa*
Andøya_LGM_A LFV	sister to *N. limnetica*	100	sister to *N. limnetica* var. *limnetica*	48	sister to *N. limnetica*	83	*N. limnetica* var. *limnetica*	72	sister to *N. limnetica*	49	*N. limnetica* var. *limnetica*
Andøya_LGM_B LFV	sister to *N. limnetica*	100	sister to *N. limnetica* var. *limnetica*	48	sister to *N. limnetica*	83	*N. limnetica* var. *limnetica*	72	sister to *N. limnetica*	49	*N. limnetica* var. *limnetica*

780
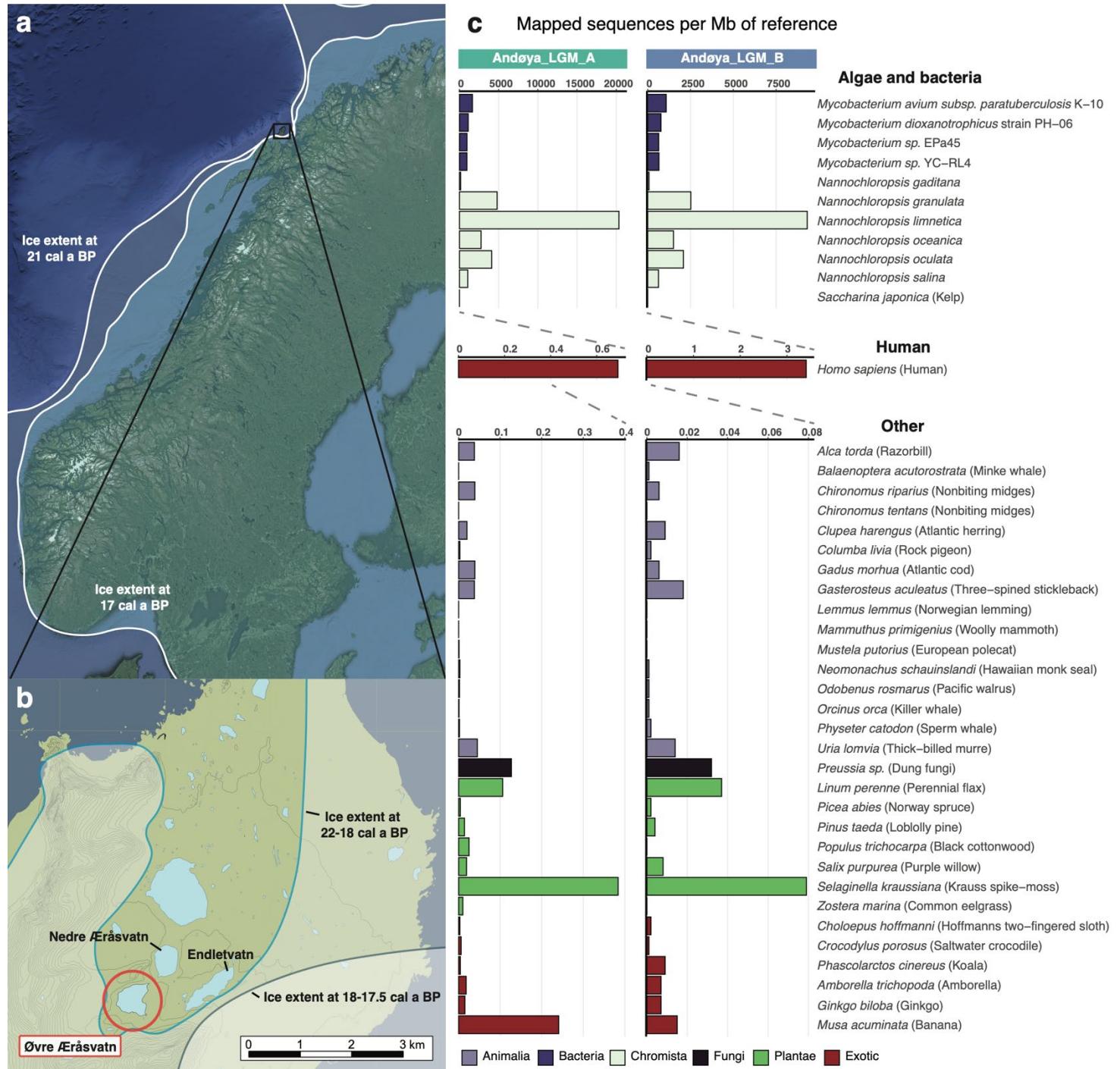
Figure 1: (a, b) Location of Lake Øvre Æråsvatnet (circled in red, panel b) in the ice-free refugium of Andøya in northwest Norway. The regional ice extent for Scandinavia in panel a has been plotted for 22 (outer) and 17 (inner) kcal yr BP and is based on Hughes *et al.* [78]. The local ice extent in panel B is plotted for 22-18 and 18-17.5 kcal yr BP and is based on Vorren *et al.* [25]. (c) Taxonomic composition of the LGM Andøya sediment samples, based on alignment to a reference panel of 42 eukaryotic or bacterial nuclear genomes. For readability, the algal, bacterial, and human results are plotted separately with differing y-axis scales.

26
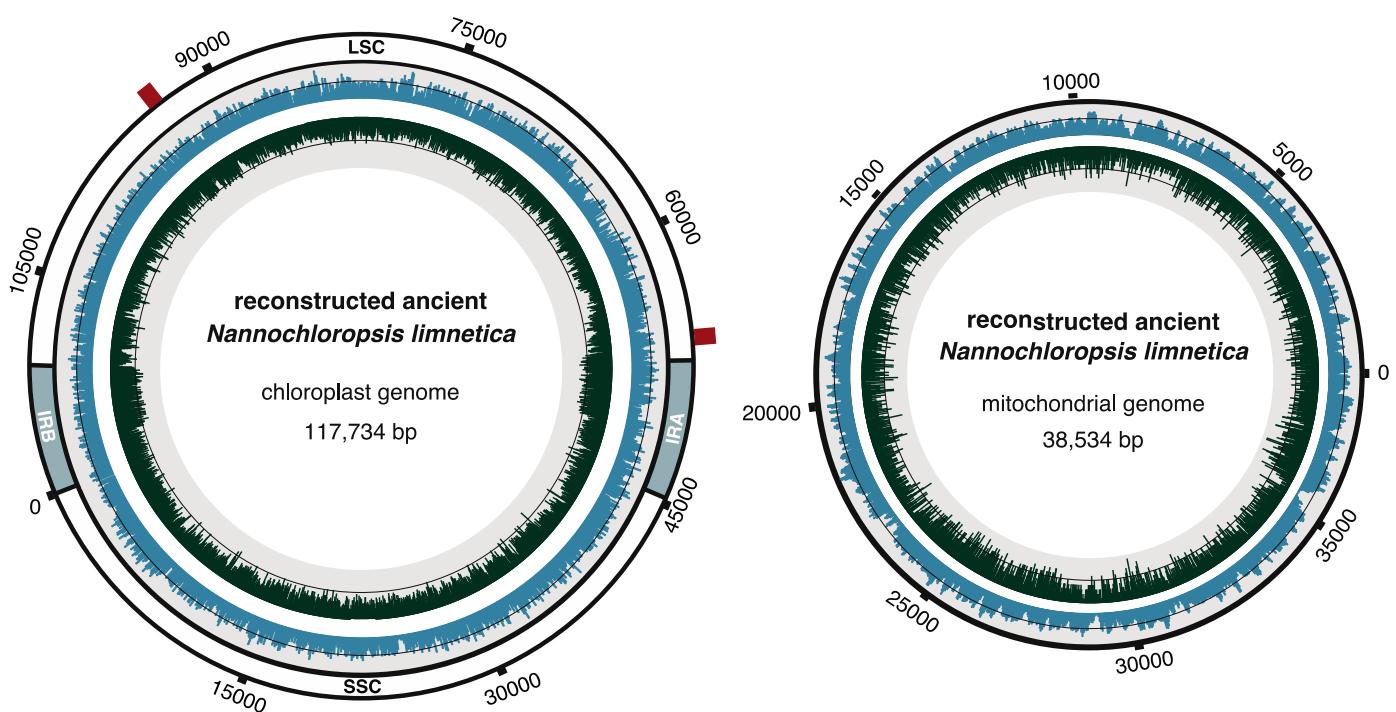
Figure 2: *N. limnetica* chloroplast and mitochondrial palaeogenomes reconstructed directly from *sed*aDNA. The innermost circle contains a distribution of the GC content in dark green, with the black line representing the 50% mark. The outer blue distribution contains the genomic coverage for the assembly, with the black line representing the average coverage of 64.3x for the chloroplast and 64.9x for the mitochondria. For the chloroplast the inverted repeats (IRA and IRB), large single copy (LSC) and small single copy (SSC) regions are annotated. The red bars on the chloroplast indicate the location of the two regions with structural change compared to the *N. limnetica* reference genome.
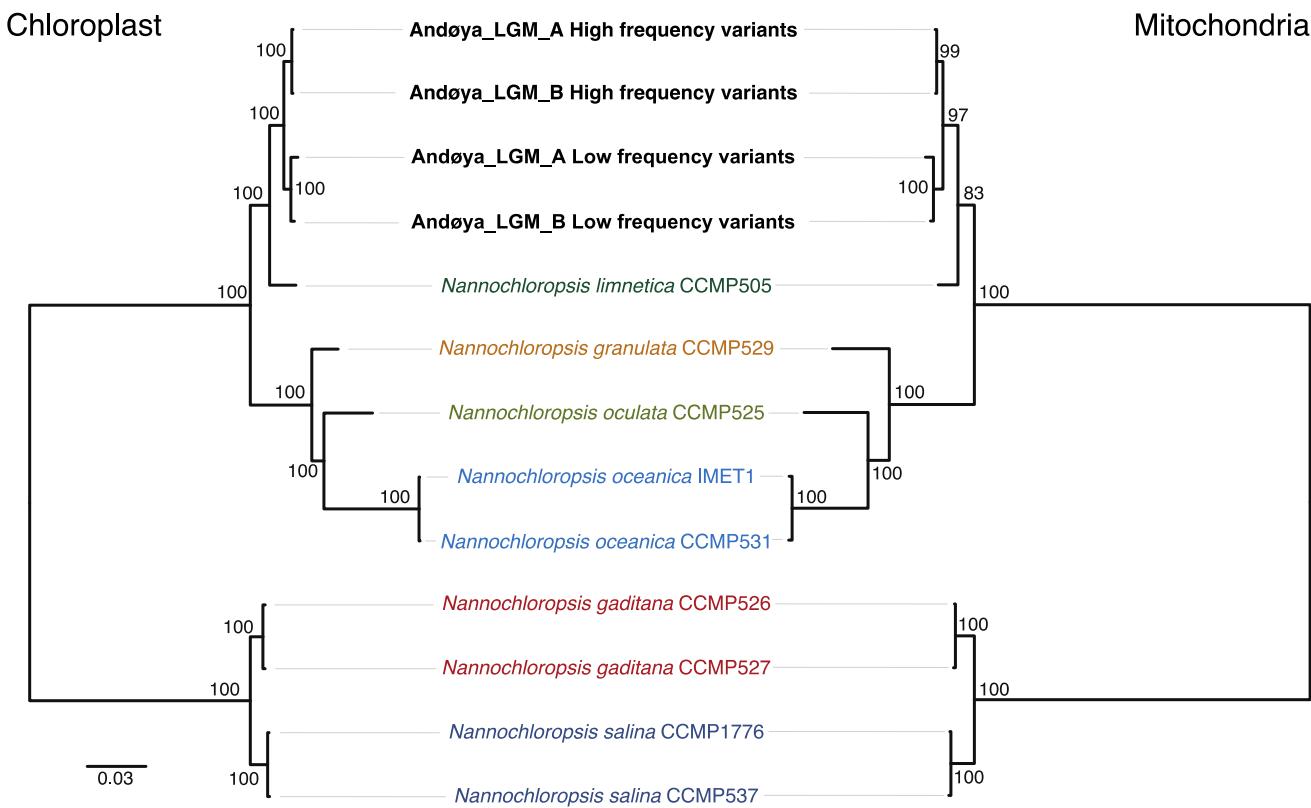
Figure 3: Maximum likelihood phylogeny of *Nannochloropsis* chloroplast (left) and mitochondrial (right) genome sequences, including the reconstructed *N. limnetica* consensus sequences based on either high or low frequency variants.
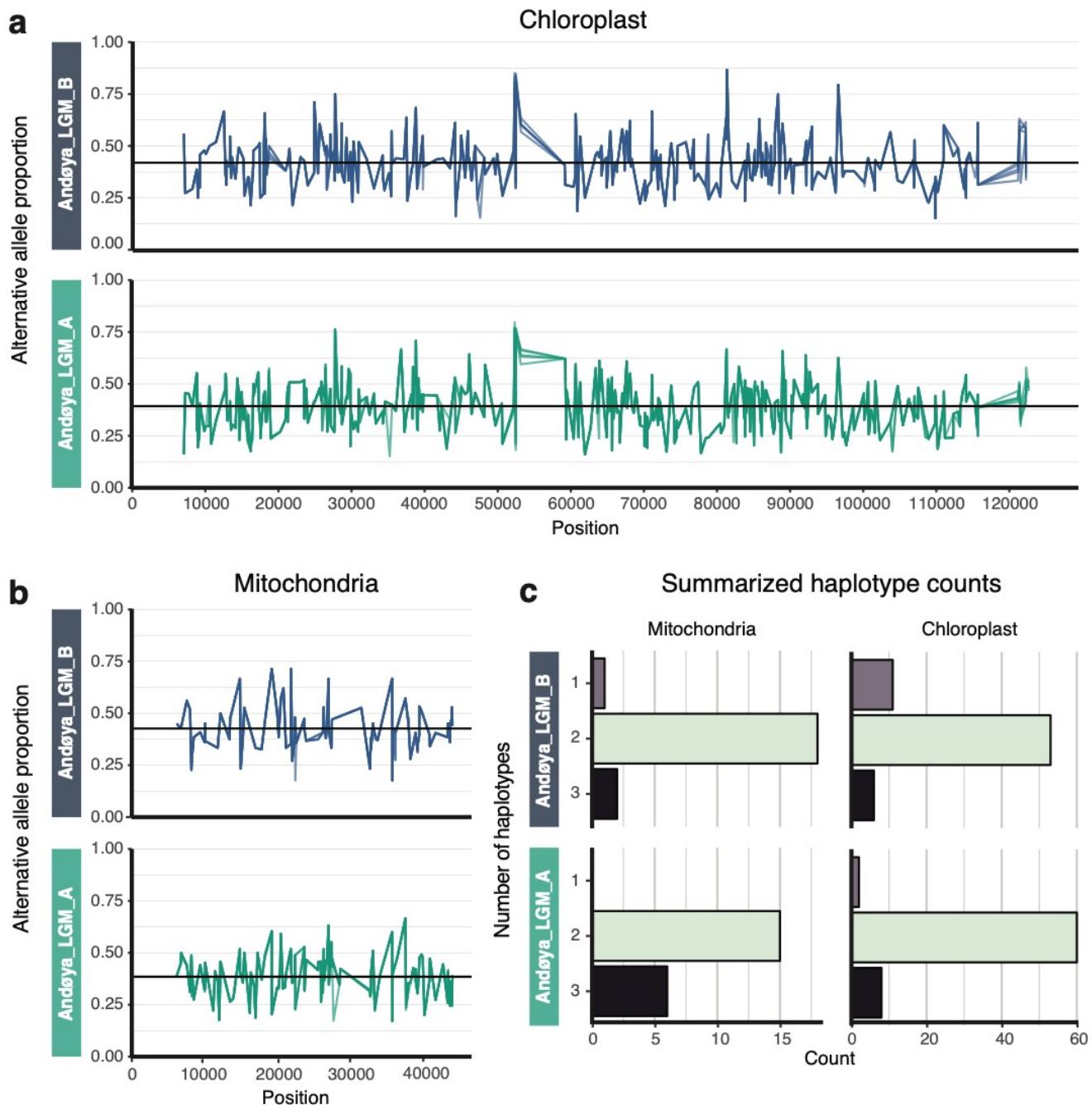
Figure 4: The proportions of alternative alleles across the chloroplast (a) and mitochondrial (b) genomes based on transversions-only. Each proportion plot consists of five independent variant calling runs to account for sampling biases (see Methods). The horizontal black lines represent averages: 0.39 and 0.42 for the chloroplast and 0.39 and 0.43 for the mitochondria, for samples Andøya_LGM_A and Andøya_LGM_B respectively. In (a) and (b), colour denotes sample. (c) Observed minimum haplotype counts based on the linked alleles for the chloroplast and mitochondrial genomes for both samples.