J Paleolimnol
DOI 10.1007/s10933-014-9779-1

ORIGINAL PAPER

A combined paleolimnological/genetic analysis of diatoms reveals divergent evolutionary lineages of *Staurosira* and *Staurosirella* (Bacillariophyta) in Siberian lake sediments along a latitudinal transect

Kathleen R. Stoof-Leichsenring · Nadine Bernhardt · Luidmila A. Pestryakova · Laura S. Epp · Ulrike Herzschuh · Ralph Tiedemann

Received: 29 April 2013 / Accepted: 15 May 2014

Abstract Diatom diversity in lakes of northwest Yakutia (Siberia) was investigated by microscopic and genetic analysis of surface and cored lake sediments, to evaluate the use of sedimentary DNA for paleolimnological diatom studies and to identify obscure genetic diversity that cannot be detected by microscopic methods. Two short (76 and 73 bp) and one longer (577 bp) fragments of the ribulose 1,5-bisphosphate carboxylase/oxygenase (rbcL) gene, encoding the large subunit of the rbcL, were used as genetic markers. Diverse morphological assemblages of diatoms, dominated by small benthic fragilarioid taxa, were retrieved from the sediments of each lake. These minute fragilarioid taxa were examined by scanning electron microscopy, revealing diverse morphotypes in *Staurosira* and *Staurosirella* from the different lakes. Genetic analyses indicated a dominance of haplotypes that were assigned to fragilarioid taxa and less genetic diversity in other diatom taxa. The long *rbcL_577* amplicon identified considerable diversification among haplotypes clustering within the *Staurosira*/*Staurosirella* genera, revealing 19 different haplotypes whose spatial distribution appears to be primarily related to the latitude of the lakes, which corresponds to a vegetation and climate gradient. Our rbcL markers are valuable tools

Electronic supplementary material The online version of this article (doi:10.1007/s10933-014-9779-1) contains supplementary material, which is available to authorized users.

K. R. Stoof-Leichsenring · N. Bernhardt · R. Tiedemann
Unit of Evolutionary Biology/Systematic Zoology, Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Strasse 24-25, 14476 Potsdam, Germany

K. R. Stoof-Leichsenring (✉) · L. S. Epp · U. Herzschuh
Periglacial Research, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Telegrafenberg 43A, 14473 Potsdam, Germany
e-mail: Kathleen.Stoof-Leichsenring@awi.de

Present Address:
K. R. Stoof-Leichsenring
Periglacial Research, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Telegrafenberg 43A, 14473 Potsdam, Germany

Present Address:
N. Bernhardt
Taxonomy and Evolutionary Biology, Leibniz Institute of Plant Genetics and Crop Research, 06466 Gatersleben, Germany

L. A. Pestryakova
Department for Geography and Biology, North-Eastern Federal University of Yakutsk, ul. Belinskogo 58, 677000 Yakutsk, Russia

U. Herzschuh
Institute of Earth and Environmental Science, University of Potsdam, Karl-Liebknecht-Strasse 24-25, Potsdam, Germany

Published online: 10 June 2014

for tracking differences between diatom lineages that are not visible in their morphologies. These markers revealed putatively high genetic diversity within the *Staurosira*/*Staurosirella* species complex, at a finer scale than is possible to resolve by microscopic determination. The rbcL markers may provide additional reliable information on the diversity of barely distinguishable minute benthic fragilarioids. Environmental sequencing may thus allow the tracking of spatial and temporal diversification in Siberian lakes, especially in the context of diatom responses to recent environmental changes, which remains a matter of controversy.

Keywords Arctic lake sediments · Diatoms · Environmental DNA · Intraspecific variation · RbcL

Introduction

Diatoms are traditionally identified by the morphology of their frustules, but genetic analyses and mating experiments have demonstrated that the resulting species determinations can be coarse, as similar morphotypes may include genetically and/or reproductively distinct lineages (Amato et al. 2007; Behnke et al. 2004), and there can be genetic diversification without morphological separation (Quijano-Scheggia et al. 2009). In such cases, genetic methods offer an alternative approach on both interspecific and intraspecific levels (Evans et al. 2007, 2008). One relatively new method for determining the genetic diatom diversity of sediment samples is the analysis of sedimentary diatom DNA. Such sequencing studies on sediment material employed genetic markers to recover recent- or paleo-biodiversity from taxa that are difficult to identify from their morphologies alone (Anderson-Carpenter et al. 2011; Boere et al. 2011; Jørgensen et al. 2012), either because they leave no fossil remains or because the taxonomic diversity is concealed within similar or identical morphotypes (Stoof-Leichsenring et al. 2012). Most genetic studies of environmental samples have aimed at species determination, but a few attempts have been made to recover genetic variations within a species or a species-complex (Epp et al. 2010; Härnström et al. 2011).

Ribosomal genes, cytochrome oxidase I (COI), and the large subunit of the ribulose 1,5-bisphosphate carboxylase/oxygenase (rbcL) have been tested as suitable markers (i.e. DNA barcodes) for species discrimination in diatoms (Evans et al. 2007; Hamsher et al. 2011; MacGillivary and Kaczmarska 2011; Zimmermann et al. 2011). Working with environmental DNA, which is a mixture of genomic DNA from various organisms, ribosomal and mitochondrial markers with rather universal primer binding sites likely amplify non-specific DNA from related, or even distinct, taxonomic groups. Because the barcoding genes suggested for diatoms were established and tested on monocultures, the difficulties in amplifying these genes from environmental samples have not been taken into consideration. The rbcL gene used in our study is particularly suitable for environmental DNA applications because the primers are specifically designed to amplify rbcL fragments from diatom DNA and the increasing number of rbcL diatom reference DNA sequences in GenBank facilitates species identification. As rbcL is part of the chloroplast genome, the likelihood of amplifying DNA from non-photosynthetic organisms is greatly reduced, and it occurs in multiple copies within a cell. This increases the amplification probability, especially in older samples. However, in most diatom species, existence of different rbcL variants, their inheritance and copy number, and also the number of chloroplasts per cell, is not well investigated. Chloroplast genome sequencing studies on diatoms have indicated the presence of only one rbcL copy in some centric and pennate diatoms (Kowallik et al. 1995; Oudot-Le Secq et al. 2007; Tanaka et al. 2011), but have not been established in other diatoms.

Diatom valves are established bioindicators in palaeolimnological studies, and are used to infer past ecological changes within a lake and its surrounding catchment. Although diatoms in Arctic freshwater habitats show a wide range of environmental tolerances, individual species often require very specific hydrochemical conditions (Laing et al. 1999). Increases in air temperature can reduce the duration of seasonal ice cover in a lake, affecting light exposure, growth rates, exchanges of gases and nutrients, and thermal stratification and mixing in the lake (Smol et al. 2005), such that diatom species shifts can be attributed to a variety of temperature-related environmental factors.

Diatom species of the Fragilariaceae family, especially the genera *Staurosirella* (Williams and Round),

Staurosira (Ehrenberg), and *Pseudostaurosira* (Williams and Round), are common in Alpine, Arctic, and sub-Arctic lakes (Bouchard et al. 2004; Laing et al. 1999; Perren et al. 2012; Schmidt et al. 2004). Prolonged ice cover and low temperatures have been cited as possible factors responsible for the high abundances of small fragilarioid taxa in these areas (Laing and Smol 2000).

Schmidt et al. (2004) attempted to determine the environmental factors that influence species assemblages in the Fragilariaceae family and suggested that pH, length of ice cover, mean July temperature, and the level of dissolved organic carbon are the main factors that influence fragilarioid species composition in Alpine lakes. Because of this climate sensitivity, the fragilarioids are well suited for investigations into climate-related ecological changes in the Arctic. A reliable reconstruction of environmental variables can only be achieved if the diatom species are correctly identified. Paull et al. (2008) illustrated the difficulties encountered in the identification of small fragilarioid forms: distinguishing between *Pseudostaurosira elliptica* (Schumann) Edlund et al. (synonym: *Fragilaria elliptica* Schumann), *Staurosira venter* (Ehrenberg) Cleve and Möller, and *Staurosirella pinnata* (Ehrenberg) Williams and Round can be problematic, as these species have very similar, small, elliptic forms in cold waters.

Northern Yakutia (northeastern Siberia, Russia) is characterised by permafrost, hence thermokarst ponds and lakes are widely distributed in the lowland areas. The vegetation changes from north to south in accordance with the temperature gradient, from Arctic semi-desert, through typical tundra vegetation, to northern taiga. Although treeline advances into tundra have been documented in the Russian Arctic (Esper and Schweingruber 2004), little is known about climate change in this region and its impacts on ecosystems, as such research has focused mainly on Alaska, Canada, and Scandinavia. A number of studies on high-latitude ponds and lakes in these areas have, however, illustrated how dramatically the biota has been impacted by Arctic climate change, resulting in pronounced changes in the freshwater communities (Prowse et al. 2006; Smol et al. 2005).

We investigated sediment samples from thermokarst lakes in the Russian Arctic using both genetic and microscopic diatom analyses and compared results from the two approaches, addressing the following questions: (1) Are lake sediments a suitable source of diatom DNA, such that DNA constitutes a valuable proxy in palaeolimnological research? (2) Is there intraspecific or obscure genetic diatom diversity that cannot be observed by morphological analysis?

Materials and methods

Localities and sampling

Sediment and water samples were collected from six lakes along a latitudinal transect in the northern lowlands of Yakutia, Russia (Fig. 1; Table 1). The localities range from the edge of the Arctic tundra near the Laptev Sea coastline, to the tundra and taiga forests southeast of Saskylakh (SA) and Tiksi (Tik), crossing the current treeline. The entire Yakutian region is characterised by deep (down to 400 m), ice-rich, continuous permafrost. The lakes investigated were formed by thermokarst processes; they are freshwater, oligotrophic, relatively shallow, and small both in size and catchment area.

Six surface sediment samples (07-SA-17, 07-SA-21, 07-SA-26, 07-SA-31, 07-SA-34 and 09-Tik-09) and two additional samples derived from short cores (07-SA-31-26 and 07-SA-34-55) were analysed. Sediments were taken with a grab sampler that took the top 3 cm of surface sediment. The short cores were 8 cm in diameter and had lengths of 28 cm (core 07-SA-31) and 33 cm (core 07-SA-34). Samples were taken at 0.4-cm intervals along each core, in the field, and placed into sterile, plastic Whirl–Pak® bags. For this study, we analysed samples from depths of 17.4 cm (07-SA-31-26) and 31.0 cm (07-SA-34-55). Sampling was carried out in 2007 and 2009, during field trips conducted by the Alfred Wegener Institute, Potsdam, in cooperation with the North-Eastern Federal University of Yakutsk. We minimised cross-contamination between sediments by sampling the lakes on different days and washing the sediment grab several times before sampling. Sediment material was stored in the dark at 7 °C at the Alfred Wegener Institute. For the genetic assessment, small quantities from each sediment sample were transferred with a sterile spatula into sterile Falcon tubes, thoroughly mixed with Queens Tissue Buffer (QTB) (Seutin et al. 1991), and kept in the dark at 10 °C until required.

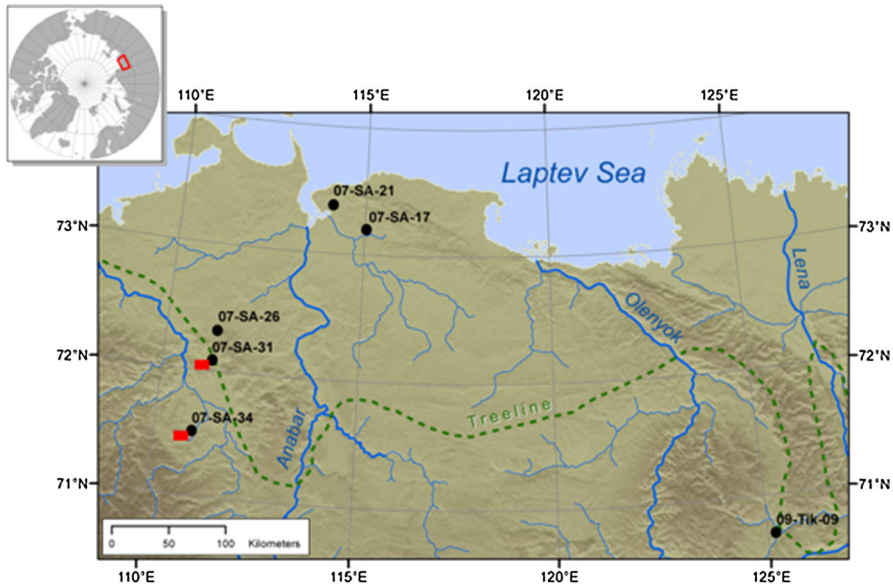

Fig. 1 Locations studied. *Black dots* indicate surface-sediment samples and *red boxes* cored sediment samples

Table 1 Physical and chemical parameters for the investigated lakes

Lake parameters	07-SA-17	07-SA-21	07-SA-26	07-SA-31	07-SA-34	09-Tik-09
Longitude (°)	114.9634	113.9485	111.189	111.1176	110.819	125.0744
Latitude (°)	73.1934	73.3871	72.3200	72.0709	71.5030	70.6984
Elevation (m)	10	4	86	38	80	74
Vegetation zone	Tundra[b]	Tundra[a]	Forest tundra	Forest tundra	Forest[c]	Forest[c]
Lake area (km^2)	0.25	0.02	0.16	0.02	0.06	0.64
Max. depth (m)	7.4	3.3	6.8	6.9	7.0	6.3
Sample depth (m)	7.0	3.3	5.0	5.7	4.8	4.7
Secchi depth (m)	2.3	2.3	4.5	1.8	3.5	1.6
pH	7.40	7.20	7.32	7.15	7.18	6.23
Cond. ($\mu S\ cm^{-1}$)	50	54	30	20	47	14
Alkalinity ($mg\ L^{-1}$)	25.5	20.7	16.2	7.0	24.1	5.0

[a] Typical tundra

[b] Southern tundra

[c] Northern forest

Dating of short cores

Short cores were dated using lead (^{210}Pb) and caesium (^{137}Cs) dating methods, at the University of Liverpool (Appleby 2001). The analyses revealed relatively uniform sedimentation rates through the cores. Linear extrapolations of the sedimentation rates were used to estimate the entire age of the short cores. Samples of the 07-SA-34 core were also radiocarbon dated and detailed results of the chronology are given in Herzschuh et al. (2013). According to these estimation the analysed samples date back to approximately the tenth century (07-SA-34-55) and the early sixteenth century (07-SA-31-26; unpublished data).

Morphological diatom analyses

Morphological diatom analyses were conducted on about 0.02 g of sediment. Calcareous and organic components were removed by heating with HCl

(10 %) and H_2O_2 (30 %), respectively. Cleaned diatom samples were mounted on microscope slides using Naphrax®. Valves were counted along a defined transect using a Zeiss microscope equipped with differential interference contrast, at a magnification of 1000x. The literature used for diatom determination is listed in Pestryakova et al. (2012). Inspections using the scanning electron microscope (SEM, Zeiss Gemini® Ultra plus), with 7.3 mm working distance and 10.00 kV, were performed solely on small fragilarioid diatoms from surface-sediment samples, for which the sediments were again subsampled and prepared in a way similar to that for light-microscope investigations. Samples were coated with a gold/palladium alloy (Au/Pd) and analysed at GFZ Potsdam.

Principal component analysis

Principal component analysis (PCA) was used to study the similarity of lakes with respect to diatom species composition. The PCA was applied to the morphological LM diatom assemblage data. To minimise the effect of abundant species, the data were log (x + 1) transformed. Samples (lakes) and species are presented in a covariance biplot showing the first two principal components. In addition, we calculated a canonical correspondence analysis (CCA), using environmental and species data to get an indication of which environmental variables explain the patterns of diatom distribution. PCA and CCA were calculated and results were visualized with PAST v. 2.14 (Hammer et al. 2001).

DNA isolation from sediment samples

DNA isolation was performed using a PowerSoil™ DNA Isolation Kit (MoBio Laboratories, California). The PowerSoil™ bead solution was transferred to a sterile Eppendorf tube. The samples with QTB were thoroughly mixed immediately prior to isolation and 500 µl of each sample-buffer mixture was transferred to the bead solution tube. Each sample was centrifuged at 10,000*g* for 1 min and the supernatant was removed, leaving 150–200 mg of sediment (Stoof-Leichsenring et al. 2012). Surface-sediment samples were isolated in the general isolation laboratory and core sediments were extracted in a dedicated laboratory for historic-DNA study, separated from the general genetic laboratories. Laboratory procedures for core sediment analysis were carried out following a number of protocols that ensure the authenticity of results when working with ancient or degraded DNA. Those included the separation of work areas for isolation and polymerase chain reaction (PCR) amplification, blank controls for isolations and PCR and reproducibility of results with multiple PCRs. Isolated DNA was stored at −20 °C.

Primer specificity and polymerase chain reaction (PCR) set-up

Polymerase chain reaction of surface sediment isolations was set up in the general PCR set-up laboratory, whereas core sediment PCR was set up in the historic-DNA lab under a special PCR-UV-Hood. For both surface and core sediments, short fragments of the large subunit of the rbcL were amplified using the *Diat_rbcL_705F* and *Diat_rbcL_808R* primers (76 base pairs amplicon, *rbcL_76;* Stoof-Leichsenring et al. 2012) and a slightly modified set of primers (*Diat_rbcL_708F* and *Diat_rbcL_808Rw*; 73 base pairs amplicon, *rbcL_73*). In addition, we amplified a longer (577 bp) rbcL fragment (*rbcL_577*) only from surface sediments, using the *Diat_rbcL_405F* and *Diat_rbcL_1006R* primers [Electronic Supplementary Material (ESM) Table 3]. With all primer pairs, we performed an *in silico* PCR on a database compiled from the entire EMBL Nucleotide Sequence Database, downloaded from ftp://ftp.ebi.ac.uk/pub/databases/embl/release/ (release embl_117, September 2013), thereby also reanalysing the primers for *rbcL_76* (Stoof-Leichsenring et al. 2012) using a newer, much larger sequence database. The output of the *in silico* PCR presents a list of all sequence entries that could potentially be amplified by these primers, including taxonomic information. The following criteria were used for the *in silico* PCR: (1) amplicon lengths of between 20 and 1,000 base pairs, and (2) a maximum of three errors between the query sequence and the primer sequence, but no errors allowed in the last two bases at the 3′ end. As an estimate of the specificity of the primers to diatoms, we calculated the percentage of diatoms within the *in silico* amplicons, excluding sequences with uncertain taxonomic identification. We also analysed the expected taxonomic resolution potential of the amplicons by calculating the ratio of unambiguously identified taxa at a given taxonomic level to the total number of amplified diatom taxa (Ficetola et al. 2010), within the diatoms only and within all potentially amplified sequences.

All PCR reactions were run in the Post-PCR laboratory, which is separated from the general isolation and from the historic-DNA lab to prevent contamination of isolates through DNA molecules amplified via PCR. Amplifications of core samples and surface samples were run on different days. For PCR runs we used a T3000 thermocycler (Biometra, Göttingen, Germany), and the following reagents: fragment-specific primers (each at a final concentration of 0.32 mM), 1 mM Tris–HCl (pH 9.0), 5 mM KCl, 1.5 mM $MgCl_2$, 0.4 mM each of dNTP, 1.5 U *Taq* polymerase (Qbiogene, Carlsbad, CA, USA), and 2–4 µl of DNA template solution. Amplifications of the *rbcL_76* and *rbcL_73* fragments were performed using the following reaction profile: PCR tubes were heated to 94 °C, followed by one cycle at 94 °C for 5 min, and then 50 cycles each at 94 °C for 30 s, at 43.6 °C for 30 s, at 72 °C for 30 s, and a final extension at 72 °C for 10 min. Amplification of the *rbcL_577* fragment in surface samples followed a two-step approach to reduce PCR chimeras (Thompson et al. 2002). For the initial step, PCR tubes were heated to 94 °C followed by one cycle at 94 °C for 30 min, and then 20 cycles each at 94 °C for 30 s, at 50 °C for 30 s, at 72 °C for 2 min, and a final extension at 72 °C for 10 min. The amount of DNA template solution for the initial PCR varied between 2 and 4 µl. For the second PCR (reconditioning step), the tubes were heated to 94 °C, followed by one cycle at 94 °C for 30 min, and then 10 cycles each at 94 °C for 30 s, at 50 °C for 30 s, at 72 °C for 2 min, and a final extension at 72 °C for 10 min. We used fresh PCR reagents and 1 µl of the initial PCR as a template for the PCR-reconditioning step. If a PCR product was too weak, amplifications were concentrated using a SpeedVac centrifuge (Eppendorf, Hamburg, Germany) at 60 °C for 10 min. All PCR products were cloned using the TOPO® TA Cloning® Kit for Sequencing (Invitrogen, Carlsbad, CA, USA). As many positive clones as available (up to a maximum of 32) were sequenced. Some PCR products were cloned twice to yield a sufficient number of positive clones.

Post-sequencing analyses, taxonomic identification and phylogenetic inferences

DNA sequences retrieved from cloning were aligned using Clustal W with default settings as implemented in BioEdit (Hall 1999). Haplotypes were verified following a stringent verification protocol, i.e. to prevent single polymerase errors being considered true variations, only sequences that occurred in two independent PCR reactions were considered authentic. Sequences that derived from only one PCR reaction were, however, considered to be real, if (1) they were present in at least two clones, and (2) they showed nucleotide substitutions in positions known to be polymorphic within the entire haplotype data set. Single sequences that showed unique nucleotide substitutions in otherwise non-polymorphic positions were considered to have resulted from polymerase errors and were excluded. Prior to verification of the *rbcL_577* sequences, we checked all 167 sequences for chimeras using the Bellerophon server (Huber et al. 2004) because, for environmental samples, chimera formation is more likely to occur in longer PCR fragments. The Bellerophon server identified 53 sequences as being of chimeric origin. The remaining 114 sequences were checked according to the verification rules above and 91 haplotypes were eventually considered to be authentic. A Bayesian phylogenetic approach was used to assign a probability of taxonomic origin to haplotypes, using SAP v.1.0.6 software (http://ib.berkeley.edu/labs/slatkin/munch/StatisticalAssignmentPackage.html; Munch et al. 2008; with settings of 0.2 for significance value, 10 for best hits, and 0.8 for minimum similarity). The SAP software provides a statistical measure of phylogenetic relationships between a new query sequence and a set of selected reference sequences and provides statistically meaningful measures of confidence for assignment to different taxonomic levels (class, family, genus, species). As with any method, the certainty of the taxonomic assignment depends on the completeness of the available reference data set.

A Bayesian phylogenetic analysis was conducted for the longer *rbcl_577* haplotypes including only haplotypes assigned to Fragilariaceae or lower taxonomic levels within this family and all available GenBank entries of Fragilariaceae. The phylogenetic tree was calculated with MrBayes (Huelsenbeck and Ronquist 2001), using *Plagiogramma* sp. (JX413563), the sister group to pennates (Theriot et al. 2010), as the outgroup. Two independent runs were performed for 1,000,000 generations. A subset (*Staurosira/Staurosirella*-like only) of *rbcL_577* haplotypes was analysed independently, constructing a 95 % probability parsimony network with TCS software (Clement et al. 2000).

Evolutionary distances and genetic differentiation

We estimated the nucleotide distances between the lakes using the *rbcl_577* data set (*Staurosira/Staurosirella*-like clade only). Therefore, we calculated the between-group mean distances (*p*-distance) and the net between-group mean distances (net D_A) between the six lakes and neglected the haplotype frequencies (=set all haplotype frequencies to one, because the haplotype frequencies are strongly biased by PCR effects and the number of clones investigated). The pairwise group mean distances were calculated as a measure of the mean evolutionary distance, and the net between-group mean distances (net D_A) were calculated to measure the nucleotide distances (corrected for within-group variation) between all pairs of sampled localities.

In order to evaluate the taxonomic level among our haplotypes, we used published diatom sequences from well-investigated diatom species to reanalyse the genetic distances within the *rbcL_577* fragment on the intra- and inter-specific level and compared these results with the *Staurosira/Staurosirella*-like haplotypes identified in our study. Therefore, we calculated the mean nucleotide distances of known rbcL sequences from *Pinnularia* and *Sellaphora*. All genetic distance measurements were calculated using the composite maximum likelihood method implemented in MEGA5.2 (Tamura et al. 2007, 2011).

Results

Diatom assemblages identified by light microscope analysis

Using a light microscope (LM), 6,157 diatom valves (500–1,500 valves/sample) and 194 different diatom taxa were identified. Surface sediments from different lakes harbour different diatom communities (Fig. 2; ESM Table 1): 07-SA-17, 07-SA-26, 07-SA-31, 07-SA-34, and 09-Tik-09 are all dominated by small epiphytic/benthic fragilarioid taxa, mainly from the *Staurosira*, *Staurosirella*, *Pseudostaurosira,* and *Tabellaria* (Ehrenberg) genera, with a single predominant fragilarioid species in each lake (Fig. 2). Only in lake 07-SA-26 is the diatom community co-dominated (in addition to the fragilarioids) by the planktonic alga *Aulacoseira perglabra* (Østrup) E.Y. Haw. The most northerly lake (07-SA-21) is distinct from all others, predominantly harbouring *Nitzschia amphibia* Grunow and *Psammothidium subatomoides* (Hust.) Bukht. and Round.

These differences in recent diatom assemblages are visualised in a covariance biplot of the first and second principal components (ESM Fig. 1a). This plot indicates the distinctiveness of lake 07-SA-21 and the similarity in diatom community between lakes 07-SA-26 and 34, and between lakes 07-SA-17, 31 and 09-Tik-09. Results of the PCA give an eigenvalue for the first axis of 31.2 %, whereas the second axis accounts for 25.0 % of the variance in the data set. The results of the CCA indicate that axis one (eigenvalue 30.9 %) mainly resembles a gradient from shallow to deeper waters, also indicated by the change from bottom-living (e.g. *N. amphibia*) to planktonic species (*A. perglabra*) along the first axis. The second axis (eigenvalue 26.8 %) possibly indicates a gradient from small to larger Secchi depth and might reflect different light availability in the lakes. Chemical variables of the lake may also influence the distribution of diatoms, but are less indicative in this data set (ESM Fig. 1b).

The light microscopic investigations on the 07-SA-31-26 and 07-SA-34-55 core samples reveal a different species composition compared to surface sediments of the same lakes (Fig. 2, ESM Table 1). Interestingly, the 07-SA-31-26 core sample (dated to the early sixteenth century) is enriched in benthic forms such as *Pinnularia hemiptera* (Kütz.) Rabenhorst, *Stauroneis phoenicenteron* (Nitzsch) Ehrenb., and *Amphora libyca* Ehrenb., relative to the surface sediment, and has fewer fragilarioid taxa. The 07-SA-34-55 core sample (dated back to about the tenth century) is completely dominated by *S. pinnata*, in contrast to the surface sediment layer, which exhibits increased diatom diversity and, in particular, a variety of benthic taxa.

Investigation of fragilarioid diatoms using scanning electron microscopy

To examine the morphological differences between the dominant fragilarioids, new preparations of surface sediments from all lakes except lake 07-SA-31 were analysed by SEM. In contrast to the LM investigations, our qualitative assessment with SEM (only a low number of valves were investigated)

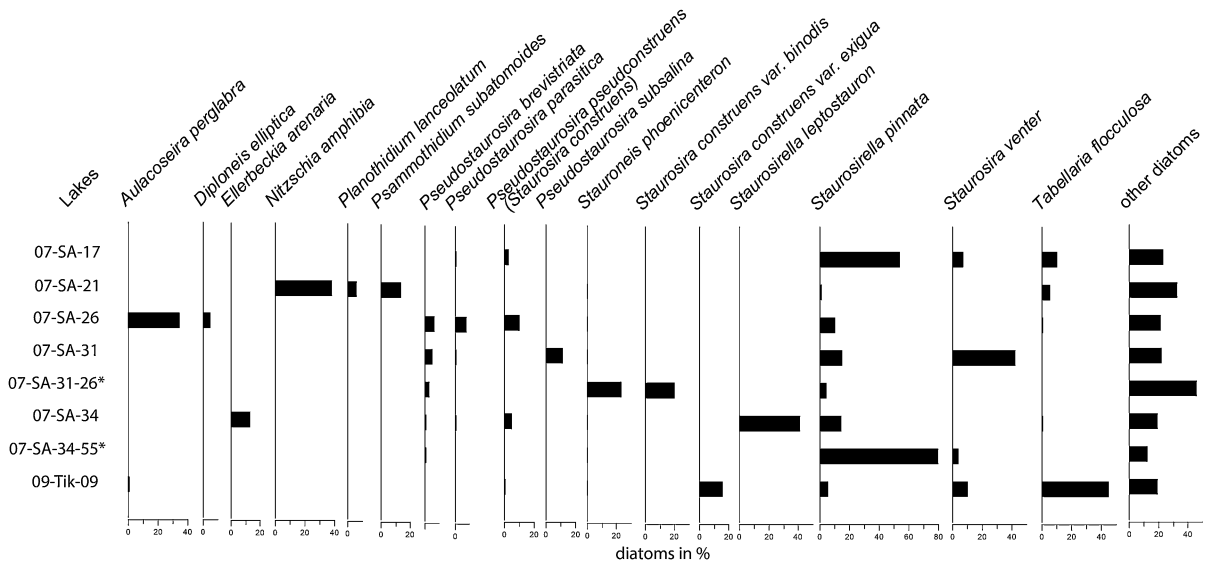

Fig. 2 Relative abundances of most dominant diatom species per sediment sample (*asterisks* indicates the core sediment samples)

determined three slightly diverse morphological variants each of *S. pinnata*, *Pseudostaurosira* spp. (Williams and Round), and *S. venter* (ESM Fig. 2). We detected the presence of *S. pinnata* morphotypes in all lakes, as with the LM counts, but predominantly in the northern lakes 07-SA-17 and 07-SA-21, whereas *S. venter* morphotypes and *Pseudostaurosira* spp. were detected only in the more southerly lakes 07-SA-26, 07-SA-34, and 09-Tik-09 (ESM Table 2).

Genetic assessment of diatom diversity

The *in silico* PCR analysis revealed that the primer pair used to generate the *rbcL_76* amplicon was highly specific, with 88 % of the identified sequences amplified *in silico* belonging to diatoms (ESM Table 3). In the *in silico* PCR, the primers of the shorter (*rbcL_73*) and the longer (*rbcL_577*) amplicons, were less specific to diatoms, with 48 % of the shorter and 29 % of the longer *in silico* amplificates being diatoms. The *rbcL_76* and *rbcL_73* amplicons mainly identified identical haplotypes; the *rbcL_76* amplicon had a tendency to amplify *Staurosira* types, whereas the *rbcL_73* amplicon identified haplotypes from diatoms of the genera *Asterionella*, *Pinnularia*, *Sellaphora,* and *Cymbopleura*. In PCR experiments the established primer pairs successfully amplified the target sequence in several algae cultures (data not shown) and the *rbcL_76* fragment was successfully applied to sediment samples from other areas (Stoof-Leichsenring et al. 2012).

Genetic investigations reveal a total of 41 diatom haplotypes from the *rbcL_76* and *rbcL_73* amplicons (labelled from A to Z and AA to AO), whereas 22 haplotypes are represented by the *rbcL_577* amplicon (numbered from 1 to 22) (Table 2; ESM Table 4; ESM Fig. 3). All of these haplotypes were defined according to a stringent verification protocol (see "Materials and methods").

Taxonomic determination of *rbcL_73* and *rbcL_76* haplotypes using SAP analyses (with a significance level of 95 %) reveals that 24 % of all haplotypes are assigned to species level, 17 % to genus level, and 57 % are assigned to higher taxonomic levels (Table 2). Twenty-three haplotypes are assigned to the Fragilariaceae family, of which 11 haplotypes show highest identities to the genus *Staurosira*. Fourteen (63.8 %) *rbcL_577* haplotypes are assigned to the species *Staurosira elliptica*, whereas all other *rbcL_577* haplotypes are assigned to higher taxonomic levels. Most of the short haploytpes are obtained from both the *rbcL_76* and *rbcL_73* amplicons, and are also represented in the longer *rbcL_577* haplotypes (ESM Table 4). The fact that independent PCR reactions and different primers reveal identical or similar haplotypes supports the authenticity and sensitivity of the results.

An additional Bayesian analysis was performed with the *rbcL_577* haplotypes (Fig. 3) to explore the

Table 2 Assignment of haplotypes from the *rbcL_73/76* (*A_76* to *AO_76*) and the *rbcL_577* (*1_577* to *22_577*) approaches, listing the assignment of taxonomic levels (SAP results) for class/family, genus and species, accession number and sequence identity of the primarily assigned GenBank entry, the final assignment of haplotype and taxonomic rank (95 % significance level)

Haplotype	Class*, Family %	Genus %	Species %	Accession number	Sequence identity	Assigned taxa (95 % significance level)
A_76	Fragilariaceae 100	*Staurosira* 78	*Staurosira elliptica* 58	HQ912451	0.95	Fragilariaceae
B_76	Fragilariaceae 100	*Staurosira* 78	*Staurosira elliptica* 65	HQ912451	0.96	Fragilariaceae
C_76	Fragilariaceae 100	*Staurosira* 94	*Staurosira elliptica* 86	HQ828193	0.97	Fragilariaceae
D_76[a]	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	0.99	*Staurosira elliptica*
E_76	Fragilariaceae 100	*Staurosira* 98	*Staurosira elliptica* 96	HQ828193	0.97	*Staurosira elliptica*
F_76	Fragilariaceae 100	*Staurosira* 90	*Staurosira elliptica* 83	HQ828193	0.96	Fragilariaceae
G_76	Fragilariaceae 100	*Staurosira* 80	*Staurosira elliptica* 69	HQ828193	0.95	Fragilariaceae
H_76	Fragilariaceae 100	*Staurosira* 66	*Staurosira construens* 25	HQ912451	1.00	Fragilariaceae
I_76	Fragilariaceae 100	*Staurosira* 99	*Staurosira elliptica* 98	HQ828193	0.99	*Staurosira elliptica*
J_76	Fragilariaceae 100	*Staurosira* 97	*Staurosira elliptica* 96	HQ828193	0.97	*Staurosira elliptica*
K_76	Fragilariaceae 100	*Staurosira* 88	*Staurosira elliptica* 77	HQ912451	0.96	Fragilariaceae
L_76	Fragilariaceae 100	*Staurosira* 97	*Staurosira elliptica* 94	HQ828193	0.99	*Staurosira*
M_76[a]	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 98	HQ828193	0.96	*Staurosira elliptica*
N_76	Fragilariaceae 100	*Staurosira* 98	*Staurosira elliptica* 97	HQ828193	0.96	*Staurosira elliptica*
O_76	Fragilariaceae 100	*Staurosira* 69	*Staurosira elliptica* 51	HQ828193	0.93	Fragilariaceae
P_76	Fragilariaceae 99	*Staurosira* 65	*Staurosira elliptica* 58	HQ828193	0.95	Fragilariaceae
Q_76	Coscinodiscophyceae* 92	*Chaetoceros* 27	*Chaetoceros socialis* 2	JQ217346	0.91	Bacillariophyta
R_76	Mediophyceae* 61	–	–	–	–	Bacillariophyta
S_73	Sellaphoraceae 99	*Sellaphora* 99	*Sellaphora* cf. *sem.* 63	EF143280	1.00	*Sellaphora*
T_76	Coscinodiscophyceae* 70	*Chaetoceros* 36	*Chaetoceros gracilis* 10	FJ002154	0.89	Bacillariophyta
U_76	Coscinodiscophyceae* 81	*Attheya* 11	*Attheya septentrionalis* 11	FJ002154	0.91	Bacillariophyta
V_76	Bacillariophyceae*100	*Pinnularia* 90	*Pinnularia borealis* 38	HQ912443	0.93	Bacillariophyceae
W_76	Bacillariophyceae*89	*Craticula* 35	*Craticula importuna* 28	JQ354692	0.93	Bacillariophyta
X_73	Bacillariophyceae*100	*Cymbopleura* 24	*Cymbopleura naviculif.* 24	AM710464	0.96	Bacillariophyceae
Y_76	Fragilariaceae 100	*Staurosira* 89	*Staurosira elliptica* 78	HQ828193	0.97	Fragilariaceae
Z_76	Bacillariophyceae*100	*Stauroneis* 33	*Stauroneis gracilor* 9	AM710475	0.96	Bacillariophyceae
AA_73	Fragilariaceae 100	*Staurosira* 30	*Staurosira construens* 9	HQ912451	1.00	Fragilariaceae
AB_76	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	0.97	*Staurosira elliptica*
AC_76	Bacillariophyceae*100	*Stauroneis* 70	*Stauroneis anceps* 21	AM710475	0.97	Bacillariophyceae
AD_76	Coscinodiscophyceae* 100	*Aulacoseira* 98	*Aulacoseira gra.*v. *ang* 94	FJ002130	1.00	*Aulacoseira*
AE_76[a]	Aulacoseiraceae 100	*Aulacoseira* 100	–	–	–	*Aulacoseira*
AF_76	Fragilariaceae 100	*Staurosira* 99	*Staurosira elliptica* 97	HQ828193	0.97	*Staurosira elliptica*
AG_76	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	1.00	*Staurosira elliptica*
AH_76[a]	Fragilariaceae 100	*Staurosira* 96	*Staurosira elliptica* 93	HQ828193	0.96	*Staurosira*
AI_76	Bacillariophyceae* 100	*Sellaphora* 56	*Sellaphora pupula* 6	HQ337608	0.95	Bacillariophyceae
AJ_76	Sellaphoraceae 100	*Fallacia* 100	*Fallacia* sp. 85	HQ912469	0.99	*Fallacia*
AK_73	Fragilariaceae 95	*Asterionella* 97	*Asterionella formosa* 97	AB430671	1.00	*Asterionella formosa*
AL_73	Fragilariophyceae* 100	*Tabellaria* 56	*Tabellaria flocculosa* 56	HQ912448	0.96	Fragilariophyceae
AM_73	Fragilariophyceae* 36	*Fragilaria* 29	*Fragilaria vireseens* 4	HQ912492	0.89	Bacillariophyta
AN_76	Bacillariophyceae* 100	*Craticula* 19	*Craticula importuna* 10	JQ354692	0.95	Bacillariophyceae
AO_73	Pinnulariaceae100	*Pinnularia* 100	*Pinnularia* sp. 9	JN418673	0.99	*Pinnularia*
1_577	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	0.98	*Staurosira elliptica*
2_577	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	0.98	*Staurosira elliptica*

Table 2 continued

Haplotype	Class*, Family %	Genus %	Species %	Accession number	Sequence identity	Assigned taxa (95 % significance level)
3_577	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	0.98	*Staurosira elliptica*
4_577	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 98	HQ828193	0.98	*Staurosira elliptica*
5_577	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	0.98	*Staurosira elliptica*
6_577	Fragilariaceae 100	*Staurosira* 96	*Staurosira construens* 72	HQ828193	1.00	*Staurosira*
7_577	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 98	HQ828193	0.97	*Staurosira elliptica*
8_577	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	0.98	*Staurosira elliptica*
9_577	Fragilariaceae 100	*Staurosira* 99	*Staurosira elliptica* 99	HQ828193	0.97	*Staurosira elliptica*
10_577	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	0.98	*Staurosira elliptica*
11_577	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	0.98	*Staurosira elliptica*
12_577	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	0.98	*Staurosira elliptica*
13_577	Fragilariaceae 100	*Staurosira* 72	*Staurosira elliptica* 70	HQ912451	0.97	Fragilariaceae
14_577	Fragilariaceae 100	*Staurosira* 38	*Pseusostaurosira* sp. 2	HQ828198	0.98	Fragilariaceae
15_577	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	0.98	*Staurosira elliptica*
16_577	Fragilariaceae 100	*Staurosira* 100	*Staurosira elliptica* 100	HQ828193	0.98	*Staurosira elliptica*
17_577	Fragilariaceae 100	*Staurosira* 96	*Staurosira elliptica* 96	HQ828193	0.98	*Staurosira elliptica*
18_577	Fragilariaceae 100	*Staurosira* 1	*Staurosira elliptica* 1	HQ828193	0.96	Fragilariaceae
19_577	Fragilariaceae 100	*Staurosira* 8	*Staurosira elliptica* 8	HQ828193	0.97	Fragilariaceae
20_577	Fragilariaceae 100	*Fragilariaforma* 11	*Fragilariaforma virescens* 11	HQ828193	0.96	Fragilariaceae
21_577	Bacillariophyceae* 100	–	–	–	–	Bacillariophyceae
22_577	Aulacoseiraceae	*Aulacoseira* 100	*Aulacoseira distans* 94	JQ003564	0.97	*Aulacoseira*

relationships between haplotypes assigned to Fragilariaceae and to allow the phylogenetic position of those haplotypes with less certain taxonomic assignments to be inferred. This phylogenetic tree reveals that except for *20_577*, all haplotypes are retrieved in a single clade, in which two major clades can be recognised (*group I* and *group II*). Besides the sequences retrieved from the sediments, *group I* contains sequences of *Staurosira construens* (HQ912451), *Staurosirella* sp. (HQ828195), *Staurosirella martyi* (HQ828192), *Punctastriata* sp. (HQ828199, HQ828197) and *Pseudostaurosira* sp. (HQ828198). The second group of haplotypes (*group II*) are associated with GenBank sequence of *S. elliptica* (HQ828193). To simplify the description of this entire clade, we refer to it as the *Staurosira/Staurosirella*-like clade (Fig. 3).

Comparison of morphological counts and short rbcL haplotypes in surface and core sediments

In general, morphological assessment yields a higher species diversity than does the genetic approach, whereas the genetic approach is able to reveal deeper insight into the diversity of *Staurosira/Staurosirella*-like haplotypes, but is less powerful in identifying other diatom sequences to species level.

In both the light microscopic and genetic data (short and long amplicons), small fragilarioids (most abundant microscopically were *S. pinnata* and *S. venter*) are dominant in all lakes, with the exception of 07-SA-21. In this most northern lake, both methods indicate a distinct diatom community, predominantly revealed by the microscopic counts of *N. amphibia* and *P. subatomoides* and haplotypes of the Bacillariophyceae family (*W_76*, *AN_76*), with only a few *S. pinnata* counts and *Staurosira/Staurosirella*-like haplotypes detected (ESM Tables 1 and 4). In the other five lakes, the diversity within the Fragilariaceae family, obtained by both morphology and genetics, is most impressive. LM and SEM examinations reveal a number of species within the genera *Staurosira*, *Staurosirella, Pseudostaurosira*, and *Fragilaria*. This diversity is corroborated by the genetic data, as most haplotypes are assigned to the GenBank entries of *S. elliptica*, *Staurosira* sp., and Fragilariaceae (ESM Tables 1

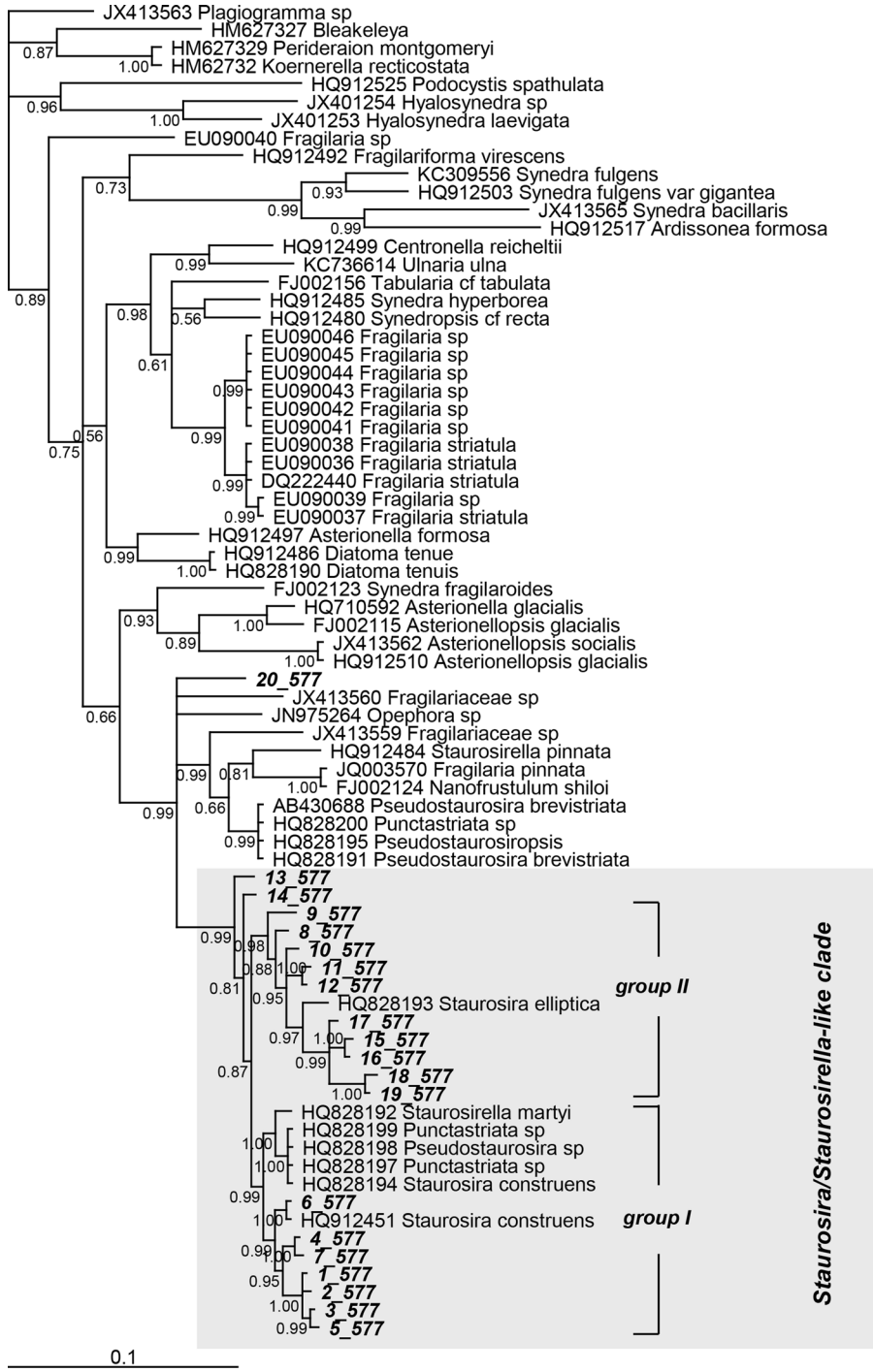

Fig. 3 Phylogenetic tree (Bayesian analyses) containing fragilarioid *rbcL_577* haplotypes (*1_577* to *20_577*) and 54 additional sequences from fragilarioid species obtained from GenBank including *Plagiogramma* sp. as the sister group to pennate diatoms. Spelling mistakes in GenBank (HQ828190 *Diatoma tenuis* = *Diatoma tenue*; FJ002123 *Synedra fragilaroides* = *Synedra fragilarioides*)

and 2). Many of the fragilarioid haplotypes occur in more than one lake, but haplotypes such as *AB_76*, *AF_76* and *AG_76* are restricted to single lakes.

The application of the genetic approach to core samples yields positive results and both genetic and morphological results detect distinct diatom

compositions compared to the results of recent sediments. In the core sample (07-SA-31-26) of lake 07-SA-31, we identified *S. phoenicenteron*, *Staurosira binodis,* and the centric diatom *Ellerbeckia arenaria* and detected *Staurosira/Staurosirella*-like and *Aulacoseira* haplotypes as well as some diatom haplotypes of less certain taxonomic assignment. Microscopic and genetic methods indicate that identical species and haplotypes occurred in the past as well as in recent samples in this lake, but both approaches also show an increase in the diversity of *Staurosira* and *Staurosirella* species, corroborated by a reduction of other diatom species from past to present. In the two samples from lake 07-SA-34, we again detected, with both methods, an increase in *Staurosira* and *Staurosirella* numbers and diversity from past to present. Microscopic detections of the remaining species indicate less change in the assemblages, whereas genetic results show diverse haplotypes in the core and recent sediments in this lake.

Genetic diversity, evolutionary distances and spatial diversification in *Staurosira/Staurosirella*-like haplotypes

The longer *rbcL_577* fragment contains additional phylogenetic information compared to the short rbcL haplotypes, and is useful in further deciphering the relationships between the haplotypes. These longer haplotypes were only obtained from surface sediments, as the DNA in core sediments is degraded and not suitable for longer amplificates. In total, we find 19 different *rbcL_577* haplotypes belonging to the *Staurosira/Staurosirella*-like clade (Fig. 3). Haplotypes *1_577* to *5_577* and *7_577*, which occur in all lakes except lake 09-Tik-09, are predominantly found in the northern lakes (07-SA-21 and 07-SA-17) and haplotypes *8_577* to *19_577* do not occur in northernmost lake 07-SA-21, but show a scattered distribution in the southerly lakes. Haplotype *6_577* matched specifically to *S. construens* (HQ912451) and is only found in lake 07-34-SA. The haplotype network gives an overview of the distribution and relatedness of haplotypes across the investigated lakes. This network indicates that 42 % of haplotypes are restricted to one individual lake and 47 % of haplotypes occur in two lakes at the same time (Fig. 4). Only two haplotypes (*13_577* and *17_577*) are found in three lakes. To investigate the nucleotide distances between the groups of haplotypes of each lake independently from their abundances (= set the number of haplotypes to one), we calculated *p*-distances and net D_A between each pair of lakes (ESM Table 5). Lowest nucleotide distances are between the most northern lakes ($p = 0.008$), whereas the highest distances are present between the most northerly 07-SA-21 and most southerly lake 09-Tik-09 ($p = 0.021$). The net D_A was also highest between 07-SA-21 and the most southerly lake 09-Tik-09 (net $D_A = 0.013$) and showed lower numbers for the other comparisons. For 07-SA-26 vs. 07-SA-31 and 07-SA-34, and 07-SA-17 vs. 07-SA-34, there was no genetic distance between the haplotype groups of the lakes, when subtracting within-group variation.

The genetic distance measurement of the *rbcL_577* fragment in other diatom taxa was calculated to evaluate the level of differentiation (intra- and/or inter-specific) in the *rbcL_577* haplotypes identified in this study (ESM Tables 6 and 7). Therefore, we reanalysed the *rbcL_577* fragment for well-known diatom taxa. To measure the nucleotide distances, we calculated *p*-distances between groups of intra- and inter-specific haplotypes. For within-species variation, we selected sequences from *Sellaphora pupula* and *Pinnularia borealis*. Species from the *Sellaphora* species complex, *Pinnularia* species, and fragilarioid species (*Staurosira*, *Staurosirella*, *Pseudostaurosira* and *Punctastriata*) were used to estimate the *p*-distances between species. This reanalysis reveals that *p* for the intraspecific level ranges from 0.023 to 0.032, whereas levels of interspecific variation are all between 0.042 and 0.048 (ESM Table 6). The mean *p*-distance of haplotypes in this study is 0.023 and is even smaller in between-lake comparisons (range from 0.008 to 0.021, ESM Table 5) and thus tends to be a signal of intraspecific variation. Interestingly, the analysis of genetic distance between different fragilarioid species indicates, in some pairs of species, very low levels of differentiation, although this comparison, in contrast to the other, also includes different genera, whereas pairwise distances between two *Staurosirella* species indicate a higher genetic distance of $p = 0.068$ (ESM Table 7).

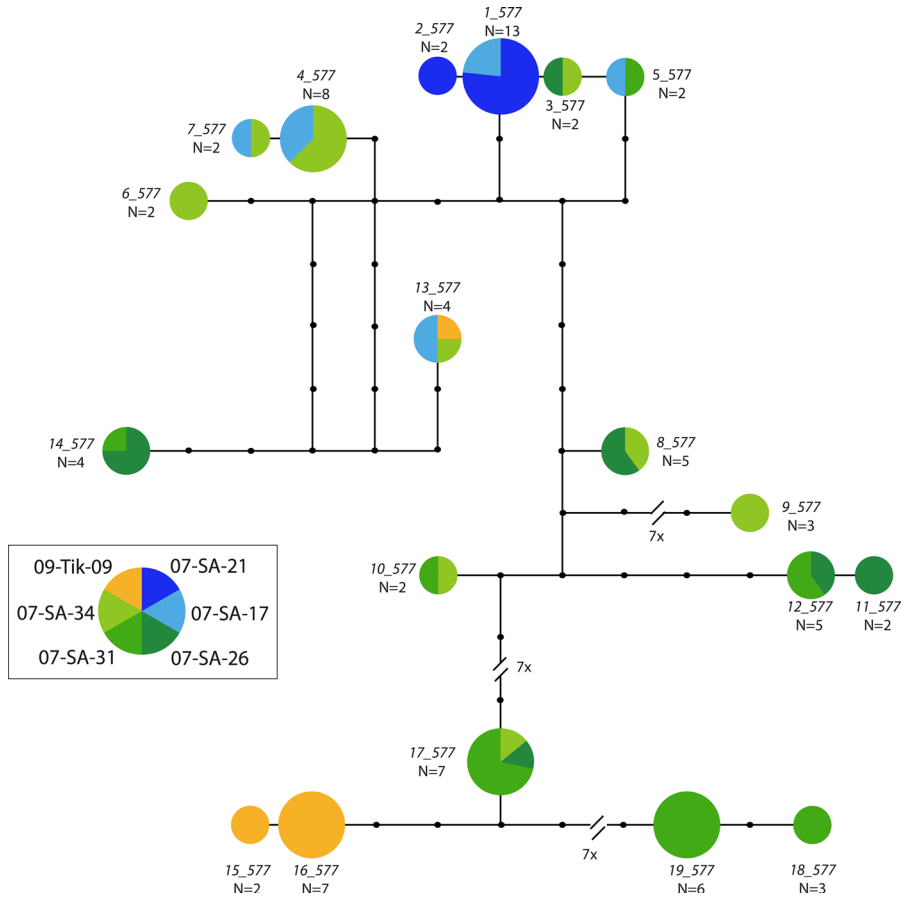

Fig. 4 Haplotype network of *rbcl_577 Staurosira/Staurosirella*-like sequences (haplotypes *1_577* to *19_577*) and their abundance in the lakes. N—number of clones, circle size is proportional to abundance, *dots* indicate missing haplotypes

Discussion

Usability of sedimentary DNA for diatom identification in paleolimnological studies

This study examined the identification of diatoms using microscopic and genetic approaches and, for the first time, identified amplifiable diatom DNA in recent and old sediments from Siberian lakes. Analysis of sedimentary diatom DNA is a useful method for the identification of diatom communities and yields positive results in both surface sediment and core material. This relatively new tool constitutes a promising approach for detecting minute fragilarioid diatoms that are difficult to identify in microscopic investigations. The advantage of genetic applications is that we can achieve finer taxonomic resolution (below species level), which can be important in identifying species that are morphologically hard to distinguish or morphologically very variable and for which the taxonomy is uncertain (Paull et al. 2008). Paull et al. (2008) pointed to the morphological similarity between the small Arctic fragilarioid taxa *S. venter* and *S. pinnata* and claimed this similarity may have caused identification problems and led to false interpretations.

Our study has shown that with regard to estimations of entire species assemblages in a community, stochastic error in a genetic biodiversity assessment may be higher than in microscopic counting, as the number of clones (sequences) obtained is much smaller than the number of valves counted. In future studies, this effect can be compensated for by next-generation amplicon sequencing methods (Kermarrec et al. 2013; Medinger et al. 2010). Despite the fact that more sequence data can be obtained easily, a currently limiting factor in the analysis of this data is the restriction to available diatom reference sequences in GenBank. A more secure identification could be gained by establishing reference libraries, which specifically contain species collected from the sampling localities (Sønstebø et al. 2010). Our study indicates a very high diversity within the *Staurosira/*

Staurosirella haplotypes, even though many of the short haplotypes could not be identified to species level, because of the lack of suitable reference data. Furthermore, such a genetic survey depends on the resolution of the genetic marker. The *rbcL_76* amplicon is very specific to diatoms (88 % specificity to diatoms, of which 53 % of sequences within the embl117 database are assigned to species level). In our study, only 24 % of all short haplotypes are assigned to species, which is a reflection of missing reference material rather than insufficient resolution. The second, short rbcL marker (*rbcL_73*), a modification of the *rbcL_76* amplicon, has less specificity to diatoms, but is used to increase the diversity as it was adjusted to preferentially amplify DNA from *Asterionella* and *Tabellaria*. This modification reduced the specificity towards amplification of diatoms (ESM Table 3), but led to the successful amplification of the preferred taxa in some samples. This demonstrates that PCR primers constitute filters that can bias the results of genetic barcoding in DNA composites, e.g. in environmental samples, and thus can have a strong impact on the number (quantity) and diversity of genetic lineages obtained (Jahn et al. 2007). In contrast, our study indicates that different fragments and independent PCRs result in many identical and/or similar haplotypes, supporting the authenticity and sensitivity of the results. Beyond the amplification of surface samples, we tested the use of the genetic approach to older material. As with dissolution of valves or total absence of diatom fossils in morphological studies, genetic analyses of environmental samples are also generally limited by the quality and quantity of DNA in the sedimentary record (Coolen et al. 2004; Pääbo et al. 2004). Working with older sedimentary material, DNA degradation and fragmentation can bias the amplification success. In surface sediments that integrate both intracellular DNA of living cells and extracellular DNA (Pietramellara et al. 2008), the presence of intact DNA from the living benthic community can lead to overrepresentation of these taxa in the genetic data when analysing entire communities and might lead to quantitative and qualitative differences between microscopic and genetic results. The *rbcL_76* amplicon was previously shown to yield positive results in core samples and indicate turnover of species assemblages through time (Stoof-Leichsenring et al. 2012). However, in this study, we mostly identified variation within the *Staurosira/Staurosirella*-like haplotypes and were able to detect haplotype changes comparing old and recent samples in this group. We also detected changes in the entire diatom community, but missing reference data limited reliable assignments to the species level.

In summary, sedimentary DNA can supplement morphological surveys because it can be used to refine taxonomic determination by detecting differentiation within similar and/or minute morphotypes over spatial and temporal scales. It might also help unravel obscure genetic diversity that cannot be detected solely with microscopic methods. Generally, the genetic approach is limited by: (1) the compromise between length, number of informative sites, and primer specificity of the genetic marker, (2) the quality and quantity of the DNA present, and (3) the availability of genetic reference data. These factors create challenges for efforts to combine environmental genetic research and morphological surveys and to link the genetic data with morphologically identified species. This problem will only be solved when we can ensure the accuracy of morphological identification and increase reliable genetic reference data and studies. These reference data ideally should come from sampling localities themselves, and in the case of diatoms from recent sediments, could be obtained from single-cell experiments or by cultivation and subsequent sequencing of mono-specific cultures.

Detecting obscured genetic diversity in *Staurosira/Staurosirella*-like haplotypes

Our study recorded the existence of distinct genetic lineages of *Staurosira/Staurosirella*-like taxa in lakes located along a latitudinal transect from Arctic tundra to northern larch forests. We identified one clade of 19 highly related haplotypes that split into two groups. The phylogenetic tree primarily assigned haplotypes in *group I* to sequences of *S. martyi*, *Punctastriata* sp., *Pseudostaurosira* sp., and *S. construens*, and the haplotypes in *group II* to *S. elliptica*. Within the first group, the rbcL marker did not distinguish between the different species/genera because they are too similar with respect to their sequences. These reference sequences originate from rivers in and around Lake Constance, Germany (Medlin et al. 2012). Medlin et al. (2012) claimed that the cluster of *Pseudostaurosira*, *Staurosirella,* and *Staurosira* should be considered as one genus because none of them is monophyletic. This

set of sequences is similar to our haplotypes and both results indicate that generic classification of fragilarioids is still uncertain. This argument is supported by the fact that our haplotypes tend to mirror intraspecific rather than interspecific diversity (ESM Tables 6 and 7). The only other reference from the first group in the *Staurosira*/*Staurosirella*-like clade originates from Canadian freshwaters (*S. construens* HQ912451; Theriot et al. 2010). This GenBank entry, however, is questionable, as it was obtained from an alga culture (UTEX, FD 323), which refers, according to the UTEX webpage, to *Pseudostaurosira construens*, but was named *S. construens* in Theriot et al. (2010). If we consider the HQ912451 entry to be *Pseudostaurosira*, the haplotype *6_577* that primarily assigns to this genus could be a representative for *Pseudostaurosira pseudoconstruens*.

Interestingly, *S. pinnata* and *Pseudostaurosira* spp. are predominant in the morphological assessment (LM and SEM), but none of the obtained haplotypes match the GenBank entries for *Pseudostaurosira brevistriata* (AB430688) or *S. pinnata* (HQ912484, JQ003570). Because the sequences are of marine origin, they may constitute divergent lineages.

The obtained *Staurosira/Staurosirella*-like haplotypes are not unambiguously assignable to reference sequences of single species because of the absence of appropriate reference data from limnic *S. venter*, *S. pinnata,* and *Pseudostaurosira* species. We are currently unable to fully resolve these taxonomic issues within the *Staurosira/Staurosirella*-like species complex. Nonetheless, our data clearly indicate the presence of distinct evolutionary lineages in the rbcL gene from different Siberian lakes, which is also suggested by the detection of slightly different morphotypes by electron microscopy (ESM Fig. 2). The light and scanning electron microscopy enumerations identified three dominant taxa: *S. pinnata*, *S. venter,* and *Pseudostaurosira* spp. Because both morphological approaches (LM and SEM) identified *S. pinnata* as dominant in the northern lakes, we assume that the *1-3* and *5_577* haplotypes in the genetic data belong to *S. pinnata*. In contrast, *S. venter* and *Pseudostaurosira* valves are mostly detected in the southerly lakes, coinciding with the detection of distinct haplotypes (*8_577* to *12_577* and *15_577* to *19_577*) in these lakes. Given the uncertainty in parts of the diatom taxonomy, the incompleteness of the available genetic references in GenBank (both in terms of taxonomic and geographic coverage), and the resolution limits of the genetic marker, it is difficult to assess the taxonomic rank of the divergent evolutionary lineages revealed by our analysis. One indication might be the analysis of nucleotide distances in comparison to other species (ESM Tables 6 and 7), which largely refer to intraspecific diversity rather than variation between different species or genera.

From an evolutionary perspective, it is, on the one hand, possible that these lineages evolved rather recently and part of this evolution may have occurred in situ, as we identified eight locally restricted haplotypes (Fig. 4). On the other hand, the genetic divergence could be, in part, a reflection of geographic isolation, as we detect phylogenetically distant haplotypes in the furthermost lakes. As latitude represents geographic distance linked to a vegetation and climate gradient, the position of lakes across the treeline ecotone might be influenced by many different factors at the same time. A recent study suggests that lakewater geochemistry, mostly alkalinity, in Yakutia is highly correlated with catchment forest density and thus affects diatom composition in these lakes (Herzschuh et al. 2013). However, the morphological investigations of diatom assemblages in our study suggest that the distribution is better linked to the maximum depth and the Secchi depth of the lakes, than to chemical variables like alkalinity, which is possibly a consequence of the small number of lakes investigated and their geographic position. The close proximity of the most northern lakes (07-SA-21 and 17) to the seashore might be responsible for the higher ion concentration in these lakes, which blurs the alkalinity gradient across the treeline ecotone highlighted by Herzschuh et al. (2013). Furthermore, latitude is related to a temperature gradient, which in turn affects the duration of ice cover, light penetration, and the growth period. Temperature was previously found to affect Arctic diatom communities (Hobbs et al. 2010; Laing et al. 1999). A few studies report slight shifts in the distribution of diatoms with changes in climate over different time scales, but also reveal the continuous presence of *Staurosira* and *Staurosirella* over long periods (Biskaborn et al. 2012; Westover et al. 2006). The distribution of *Staurosira* and *Staurosirella* haplotypes in our study might be driven by some of these environmental factors, but further investigations into these factors are needed, as is broader geographic sampling, to address this question.

In summary, our study reveals distinct genetic lineages of *Staurosira* and *Staurosirella* in the lakes of the Russian Arctic and detects subtle changes in biodiversity across a tundra–northern taiga transect. The genetic approach has proven to be useful for recent and old sedimentary material and might offer an opportunity to track genetic diversity across spatial and temporal scales to refine biodiversity studies and separate haplotype lineages concealed within similar or identical morphotypes.

Acknowledgments We thank Helga Kemnitz and Ilona Schäpan for help with the scanning electron microscope at GFZ Potsdam, the University of Potsdam for financial support, and Ed Manning and Cathy Jenks for editing the manuscript.

References

Amato A, Kooistra WHCF, Ghiron JHL, Mann DG, Pröschold T, Montresor M (2007) Reproductive isolation among sympatric cryptic species in marine diatoms. Protist 158:193–207

Anderson-Carpenter LL, McLachlan JS, Jackson ST, Kuch M, Lumibao CY, Poinar HN (2011) Ancient DNA from lake sediments: bridging the gap between paleoecology and genetics. BMC Evol Biol 11:30

Appleby PG (2001) Chronostratigraphic techniques in recent sediments. In: Last W, Smol J (eds) Tracking environmental change using lake sediments. Springer, Netherlands, pp 171–203

Behnke A, Friedl T, Chepurnov VA, Mann DG (2004) Reproductive compatibility and Rdna sequence analyses in the *Sellaphora Pupula* species complex (Bacillariophyta)1. J Phycol 40:193–208

Biskaborn BK, Herzschuh U, Bolshiyanov D, Savelieva L, Diekmann B (2012) Environmental variability in northeastern Siberia during the last 13,300 yr inferred from lake diatoms and sediment–geochemical parameters. Palaeogeogr Palaeoclimatol Palaeoecol 329:22–36

Boere AC, Rijpstra WIC, De Lange GJ, Damsté JSS, Coolen MJL (2011) Preservation potential of ancient plankton DNA in Pleistocene marine sediments. Geobiology 9:377–393

Bouchard G, Gajewski K, Hamilton PB (2004) Freshwater diatom biogeography in the Canadian Arctic Archipelago. J Biogeogr 31:1955–1973

Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. Mol Ecol 9: 1657–1659

Coolen MJL, Muyzer G, Rijpstra WIC, Schouten S, Volkman JK, Damsté JSS (2004) Combined DNA and lipid analyses of sediments reveal changes in Holocene haptophyte and diatom populations in an Antarctic lake. Earth Planet Sci Lett 223:225–239

Epp LS, Stoof KR, Trauth MH, Tiedemann R (2010) Historical genetics on a sediment core from a Kenyan lake: intraspecific genotype turnover in a tropical rotifer is related to past environmental changes. J Paleolimnol 43:939–954

Esper J, Schweingruber FH (2004) Large-scale treeline changes recorded in Siberia. Geophys Res Lett 3:L06202. doi:10.1029/2003GL019178

Evans KM, Wortley AH, Mann DG (2007) An assessment of potential diatom "barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). Protist 158:349–364

Evans KM, Wortley AH, Simpson GE, Chepurnov VA, Mann DG (2008) A molecular systematic approach to explore diversity within the *Sellaphora pupula* species complex (Bacillariophyta). J Phycol 44:215–231

Ficetola GF, Coissac E, Zundel S, Riaz T, Shehzad W, Bessière J, Taberlet P, Pompanon F (2010) An in silico approach for the evaluation of DNA barcodes. BMC Genomics 11:434

Hall TA (1999) BioEdit: a user friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98

Hammer O, Harper DAT, Ryan PD (2001) PAST: paleontological statistics software package for education and data analysis. Palaeontol Electron 4:1–9

Hamsher SE, Evans KM, Mann DG, Poulickova A, Saunders GW (2011) Barcoding diatoms: exploring alternatives to COI-5P. Protist 162:405–422

Härnström K, Ellegaard M, Andersen TJ, Godhe A (2011) Hundred years of genetic structure in a sediment revived diatom population. Proc Natl Acad Sci USA 108:4252–4257

Herzschuh U, Pestryakova LA, Savelieva LA, Heinecke L, Bohmer T, Biskaborn BK, Andreev A, Ramisch A, Shinneman AL, Birks HJ (2013) Siberian larch forests and the ion content of thaw lakes form a geochemically functional entity. Nat Commun 4:2408

Hobbs WO, Telford RJ, Birks HJB, Saros JE, Hazewinkel RRO et al (2010) Quantifying recent ecological changes in remote lakes of North America and Greenland Using sediment diatom assemblages. PLoS ONE 5(4):e10026. doi:10.1371/journal.pone.0010026

Huber T, Faulkner G, Hugenholtz P (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics 20:2317–2319

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Jahn R, Zetzsche H, Reinhardt R, Gemeinholzer B (2007) Diatoms and DNA barcoding: a pilot study on an environmental sample. In: Kusber WH, Jahn R (eds), Proceedings of the 1st Central European diatom meeting 2007. Botanic Garden and Botanical Museum Berlin-Dahlem, FU Berlin, Berlin, pp 63–68

Jørgensen T, Kjær KH, Haile J, Rasmussen M, Boessenkool S, Andersen K, Coissac E, Taberlet P, Brochmann C, Orlando L, Gilbert MTP, Willerslev E (2012) Islands in the ice: detecting past vegetation on Greenlandic nunataks using historical records and sedimentary ancient DNA Metabarcoding. Mol Ecol 21:1980–1988

Kermarrec L, Franc A, Rimet F, Chaumeil P, Humbert JF, Bouchez A (2013) Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. Mol Ecol Resour 13:607–619

Kowallik KV, Stoebe B, Schaffran I, KrothPancic P, Freier U (1995) The chloroplast genome of a chlorophyll a+c-containing alga, *Odontella sinensis*. Plant Mol Biol Rep 13:336–342

Laing TE, Smol JP (2000) Factors influencing diatom distributions in circumpolar treeline lakes of northern Russia. J Phycol 36:1035–1048

Laing TE, Rühland KM, Smol JP (1999) Past environmental and climatic changes related to tree-line shifts inferred from fossil diatoms from a lake near the Lena River Delta, Siberia. Holocene 9:547–557

MacGillivary ML, Kaczmarska I (2011) Survey of the efficacy of a short fragment of the rbcL gene as a supplemental DNA barcode for diatoms. J Eukaryot Microbiol 58:529–536

Medinger R, Nolte V, Pandey RV, Jost S, Ottenwälder B, Schlötterer C, Boenigk J (2010) Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. Mol Ecol 19:32–40

Medlin L, Yang I, Sato S (2012) Evolution of the diatoms. VII. Four gene Phylogeny assesses the validity of selected araphid genera. Nova Hedwig 141:505–513

Munch K, Boomsma W, Huelsenbeck JP, Willerslev E, Nielsen R (2008) Statistical assignment of DNA sequences using Bayesian phylogenetics. Syst Biol 57:750–757

Oudot-Le Secq MP, Grimwood J, Shapiro H, Armbrust EV, Bowler C, Green BR (2007) Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: comparison with other plastid genomes of the red lineage. Mol Genet Genomics 277:427–439

Pääbo S, Poinar H, Serre D, Jaenicke-Després V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M (2004) Genetic analyses from ancient DNA. Annu Rev Genet 38:645–679

Paull TM, Hamilton PB, Gajewski K, LeBlanc M (2008) Numerical analysis of small Arctic diatoms (Bacillariophyceae) representing the *Staurosira* and *Staurosirella* species complexes. Phycologia 47:213–224

Perren BB, Wolfe AP, Cooke CA, Kjær KH, Mazzucchi D, Steig EJ (2012) Twentieth-century warming revives the world's northernmost lake. Geology 40:1003–1006

Pestryakova LA, Herzschuh U, Wetterich S, Ulrich M (2012) Present-day variability and Holocene dynamics of permafrost-affected lakes in central Yakutia (Eastern Siberia) inferred from diatom records. Quat Sci Rev 51:56–70

Pietramellara G, Ascher J, Borgogni F, Ceccherini MT, Guerri G, Nannipieri P (2008) Extracellular DNA in soil and sediment: fate and ecological relevance. Biol Fertil Soils 45:219–235

Prowse TD, Wrona FJ, Reist JD, Gibson JJ, Hobbie JE, Lévesque LMJ, Vincent WF (2006) Climate change effects on hydroecology of Arctic freshwater ecosystems. Ambio 35:347–358

Quijano-Scheggia SI, Garces E, Lundholm N, Moestrup Ø, Andree K, Campi J (2009) Morphology, physiology, molecular phylogeny and sexual compatibility of the cryptic *Pseudo-nitzschia delicatissima* complex (Bacillariophyta), including the description of *P. arenysensis* sp nov. Phycologia 48:492–509

Schmidt R, Kamenik C, Lange-Bertalot H, Klee R (2004) Fragilaria and Staurosira (Bacillariophyceae) from sediment surfaces of 40 lakes in the Austrian Alps in relation to environmental variables, and their potential for palaeoclimatology. J Limnol 63:171–189

Seutin G, White BN, Boag PT (1991) Preservation of avian blood and tissue samples for DNA analyses. Can J Zool 69:82–90

Smol JP, Wolfe AP, Birks HJB, Douglas MSV, Jones VJ, Korhola A, Pienitz R, Rühland K, Sorvari S, Antoniades D, Brooks SJ, Fallu MA, Hughes M, Keatley BE, Laing TE, Michelutti N, Nazarova L, Nyman M, Paterson AM, Perren B, Quinlan R, Rautio M, Saulnier-Talbot E, Siitonen S, Solovieva N, Weckström J (2005) Climate-driven regime shifts in the biological communities of arctic lakes. Proc Natl Acad Sci USA 102:4397–4402

Sønstebø JH, Gielly L, Brysting AK, Elven R, Edwards M, Haile J, Willerslev E, Coissac E, Rioux D, Sannier J, Taberlet P, Brochmann C (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. Mol Ecol Resour 10:1009–1018

Stoof-Leichsenring KR, Epp LS, Trauth MH, Tiedemann R (2012) Hidden diversity in diatoms of Kenyan Lake Naivasha: a genetic approach detects temporal variation. Mol Ecol 21:1918–1930

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739

Tanaka T, Fukuda Y, Yoshino T, Maeda Y, Muto M, Matsumoto M, Mayama S, Matsunaga T (2011) High-throughput pyrosequencing of the chloroplast genome of a highly neutral-lipid-producing marine pennate diatom, *Fistulifera* sp. strain JPCC DA0580. Photosynth Res 109:223–229

Theriot EC, Ashworth M, Ruck E, Nakov T, Jansen RK (2010) A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. Plant Ecol Evol 143:278–296

Thompson JR, Marcelino LA, Polz MF (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. Nucleic Acids Res 30:2083–2088

Westover KS, Fritz SC, Blyakharchuk TA, Wright HE (2006) Diatom paleolimnological record of Holocene climatic and environmental change in the Altai Mountains, Siberia. J Paleolimnol 35:519–541

Zimmermann J, Jahn R, Gemeinholzer B (2011) Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. Org Divers Evol 11:173–192