

Single-cell genomics of a bloom-forming phytoplankton species reveals population genetic structure across continents

Raphael Gollnisch^{1,2,*}, Dag Ahrén³, Karin Rengefors¹

¹Department of Biology, Aquatic Ecology, Lund University, 22362 Lund, Sweden

²Department of Earth Sciences, University of Oxford, Oxford OX1 3AN, United Kingdom

³National Bioinformatics Infrastructure Sweden (NBIS), SciLifeLab, Department of Biology, Lund University, 22362 Lund, Sweden

*Corresponding author: Raphael Gollnisch, Department of Earth Sciences, University of Oxford, South Parks Road, Oxford OX1 3AN, United Kingdom.

Email: raphael.gollnisch@gmail.com

Abstract

The study of microbial diversity over time and space is fundamental to the understanding of their ecology and evolution. The underlying processes driving these patterns are not fully resolved but can be studied using population genetic approaches. Here we investigated the population genetic structure of *Gonyostomum semen*, a bloom-forming phytoplankton species, across two continents. The species appears to be expanding in Europe, whereas similar trends are not observed in the USA. Our aim was to investigate if populations of *Gonyostomum semen* in Europe and in the USA are genetically differentiated, if there is population genetic structure within the continents, and what the potential drivers of differentiation are. To this end, we used a novel method based on single-amplified genomes combined with Restriction-site Associated DNA sequencing that allows de novo genotyping of natural single-cell isolates without the need for culturing. We amplified over 900 single-cell genomes from 25 lake populations across Europe and the USA and identified two distinct population clusters, one in Europe and another in the USA. Low genetic diversity in European populations supports the hypothesized recent expansion of *Gonyostomum semen* on this continent. Geographic population structure within each continent was associated with differences in environmental variables that may have led to ecological divergence of population clusters. Overall, our results show that single-amplified genomes combined with Restriction-site Associated DNA sequencing can be used to analyze microalgal population structure and differentiation based on single-cell isolates from natural, uncultured samples.

Keywords: restriction-site associated DNA (RAD) sequencing, single-cell whole-genome amplification (WGA), *Gonyostomum semen*, dispersal, adaptation, ecological divergence

Introduction

Phytoplankton play vital roles in biogeochemical processes and ecosystem functioning on Earth. They are responsible for a major part of oxygen production, carbon sequestration, nutrient cycling, and form the base of aquatic planktonic food webs [1, 2]. To fully understand the ecology and evolution of phytoplankton, their diversity in time and space, including their biogeographic patterns, must be studied. At the beginning of the 21st century, microbes were proposed to lack biogeographic patterns because their small size and enormous population sizes suggest essentially unlimited dispersal [3]. Instead, microbes were hypothesized to be constrained by environmental filters only [4, 5]. However, later studies showed that spatial structure of microbial diversity was driven by both geographical (physical dispersal limitation) and historical effects, as well as environmental filtering [6–8]. This can in part be attributed to methodological advances in molecular genetics that have increased the resolution to differentiate more closely related populations. Several studies now provide evidence of biogeographical patterns in both marine (e.g. [9, 10]) and freshwater (e.g. [11]) phytoplankton.

To date, the processes underlying biogeographic and population genetic structure in phytoplankton are not fully resolved. Biogeographic patterns are generated by processes at the population level, including selection, mutation, genetic drift, as well as gene flow (dispersal). Consequently, population genetic analyses are used to study these processes. In lake phytoplankton, population genetic structure and genetic differentiation between populations within species indicate that gene flow is limited (e.g. [12, 13, 14]). This limitation could be because of either physical or biological barriers [15]. The latter could, for example, be because of priority effects (advantage of first colonizers that shape the population genetic structure) and/or local adaptation (when the fitness of a local population is tweaked to the local trait optimum through selection against poorly suited phenotypes). To address which processes drive population genetic structure, studies are needed that encompass a range of distances and numerous lakes. However, few studies have been able to combine large scale sampling (across continents) and enough strains to utilize population genetic methods.

The main factor hampering population genetic studies of phytoplankton (and many other microbial eukaryotes) is the

Received: 30 September 2023. Revised: 30 December 2023. Accepted: 12 March 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the International Society for Microbial Ecology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

enormous sampling and culturing effort needed to collect enough strains per population. Culturing is needed to obtain enough DNA for downstream analyses, but only a minor fraction of microorganisms can be grown in the laboratory [16, 17]. Even among species that can be cultured, sampling can usually not be easily combined with immediate isolation and culturing, and the survival rate is often low. To circumvent these problems, we have applied a new method, single-amplified genomes combined with RADseq (SAG-RAD) [18], that we have developed to perform population genomic analyses on single protist cells. This approach combines whole-genome amplification from single cells (single-amplified genomes, SAG) with Restriction-site Associated DNA (RAD) sequencing to produce reduced representation sequencing libraries.

The aim of this study was to investigate the large-scale population genetic structure of the limnic phytoplankton species *Gonyostomum semen* (Raphidophyceae). We used SAG-RAD because *G. semen* is difficult to grow and has a large genome (diploid genome size $2C \approx 32$ Gbp) [14], making whole-genome sequencing an unfeasible approach. Moreover, pooling enough cells from a natural sample to avoid genome amplification is impossible. *Gonyostomum semen* is a species that forms nuisance algal blooms in freshwater sites around the world [19, 20]. An increasing incidence of blooms has been observed in northern Europe during the past five decades [19–22]. It is currently considered invasive with potential southwards range expansion in Europe, colonizing new habitats and forming nuisance blooms in a variety of freshwater systems [19, 20, 23–25]. In contrast to Europe, blooms of *G. semen* appear to be less common across North America [26].

Earlier population genetic studies of *G. semen* using monoclonal cultures and Amplified Fragment Length Polymorphism demonstrated that the lake populations in Finland and Scandinavia are genetically distinct, yet very similar [27, 28] suggesting a recent rapid expansion. Single-nucleotide polymorphisms (SNPs) obtained from RADseq on the same samples [14], suggested an east–west divide as well as a dispersal direction from the north-east/east toward the southwest/west. However, the large-scale biogeographic patterns and population structure across continents is not known, nor what shapes these patterns.

The specific aims of this study were to identify whether populations of *G. semen* in Europe and in the USA form distinct population clusters, if the genetic differentiation of populations is more pronounced between continents than within, and if there are signs of recent expansion in these populations. We further investigated whether there is a biogeographic population structure within each of the two continents, if isolation-by-distance (i.e. physical dispersal limitation) limits dispersal between populations, and if there are differences in environmental variables that are associated with the population structure.

Materials and methods

To investigate the biogeographic pattern of *G. semen* at the population level, single cells were isolated from lakes both in Europe and in the USA to cover a wide geographic distribution. The sampled lakes were chosen to represent both the geographic distribution of *G. semen* as well as lakes of diverse physical, chemical, and biotic characteristics, reflecting the wide environmental range of *G. semen* occurrence. The genomes of single cells were analyzed using the SAG-RAD method [18].

Sampling and single-cell isolation

Gonyostomum semen was sampled from lakes in Estonia, Lithuania, Czech Republic, Poland, Germany, the Netherlands, Denmark,

Sweden, and Portugal in the summer of 2017 (Fig. 1A and Table S1) and in the USA in 2015 (North Carolina) and in 2018 in Washington, Michigan, Maine, and Massachusetts (Fig. 1B and Table S1). Plankton samples were collected using a plankton net (mesh size 20 μm) and filtered through a 150- μm mesh to exclude larger grazers. Single cells of *G. semen* were then isolated (32–82 per sampling site, Table S2), washed, and sorted manually using custom micropipettes as previously described [18]. Single-cell isolates were frozen immediately, transported at -20°C , and stored at -80°C until amplification.

Multiple displacement amplification of single-cell genomes (SAGs)

Single-cell genomes of *G. semen* cells were amplified (25–58 per sampling site, Table S2) through Multiple Displacement Amplification (MDA) using the REPLI-g Single-Cell Kit (QIAGEN, Hilden, Germany) as previously described [18] to produce SAGs. Each reaction contained 0.5- μM SYTO13 (Invitrogen) fluorescent nucleic acid stain to monitor amplification curves on a real-time PCR system (CFX96 Touch, BIO-RAD). The amplification critical point (Cp) was determined from amplification curves as the x-value of the inflection point (i.e. the time required until the inflection point of the exponential phase is reached in an amplification). Successful MDA reactions (16–45 per sampling site, Table S2) were purified using AMPure XP beads (Beckman Coulter, Brea, CA, USA) according to the QIAGEN Supplementary Protocol RG34 replacing TE with EB buffer (QIAGEN). Fluorometric DNA quantification was performed following instructions in the QIAGEN REPLI-g Single-Cell Kit handbook. To confirm species identity, one SAG per population was amplified in two PCR assays; a *Gonyostomum*-specific 18S assay (primers GS2-F and GS5-R) [29] and a universal *cox1* assay (primers LCO1490 and HCO2198) [30]. The PCR products were sequenced using Sanger sequencing (in house) and the species identity was confirmed using BLASTN in BLAST+ v2.14.1 [31] against the NCBI nucleotide database (Table S3).

Preparation of RADseq libraries and sequencing

Single-digest RADseq library preparation was carried out as previously described [18]. The P1 adapters used here (Table S4) contained unique 8-bp barcodes. Samples with different P1 adapters were randomly pooled into 18 libraries, each containing 34–40 samples, prior to DNA shearing.

Size selection of the sheared libraries was performed using SPRIselect beads (Beckman Coulter). A 0.75 ratio of bead suspension to sample followed by a 0.55 ratio was used to select for fragment sizes of 300–600 bp. P2 adapters (Table S4) were then ligated to each of the 18 libraries of randomly pooled samples. Each P2 adapter ligation reaction was incubated at room temperature for 60 min, followed by reaction cleanup and removal of P2 adapter dimers (left-side size selection as described above for removal of P1 adapter dimers) and eluted in 45- μl EB buffer. Following a final reaction cleanup (AMPure XP beads) after PCR amplification, the pools were sequenced (paired-end, 150 bp) on a Nova Seq 6000 S4 flow cell (Illumina) at the SNP&SEQ Technology Platform of the SciLifeLab facility in Uppsala, Sweden.

Data processing and analysis

Sequences were processed using the software Stacks 2 v2.59 [32]. Stacks process_radtags was used for demultiplexing and to filter reads. Reads that contained adapter sequences, reads with an uncalled base, reads with low quality scores, and reads with no intact restriction enzyme cut site were discarded. This was followed by removal of potential contaminant reads using the taxonomic sequence classifier Kraken 2 v2.1.2 [33]. Additional

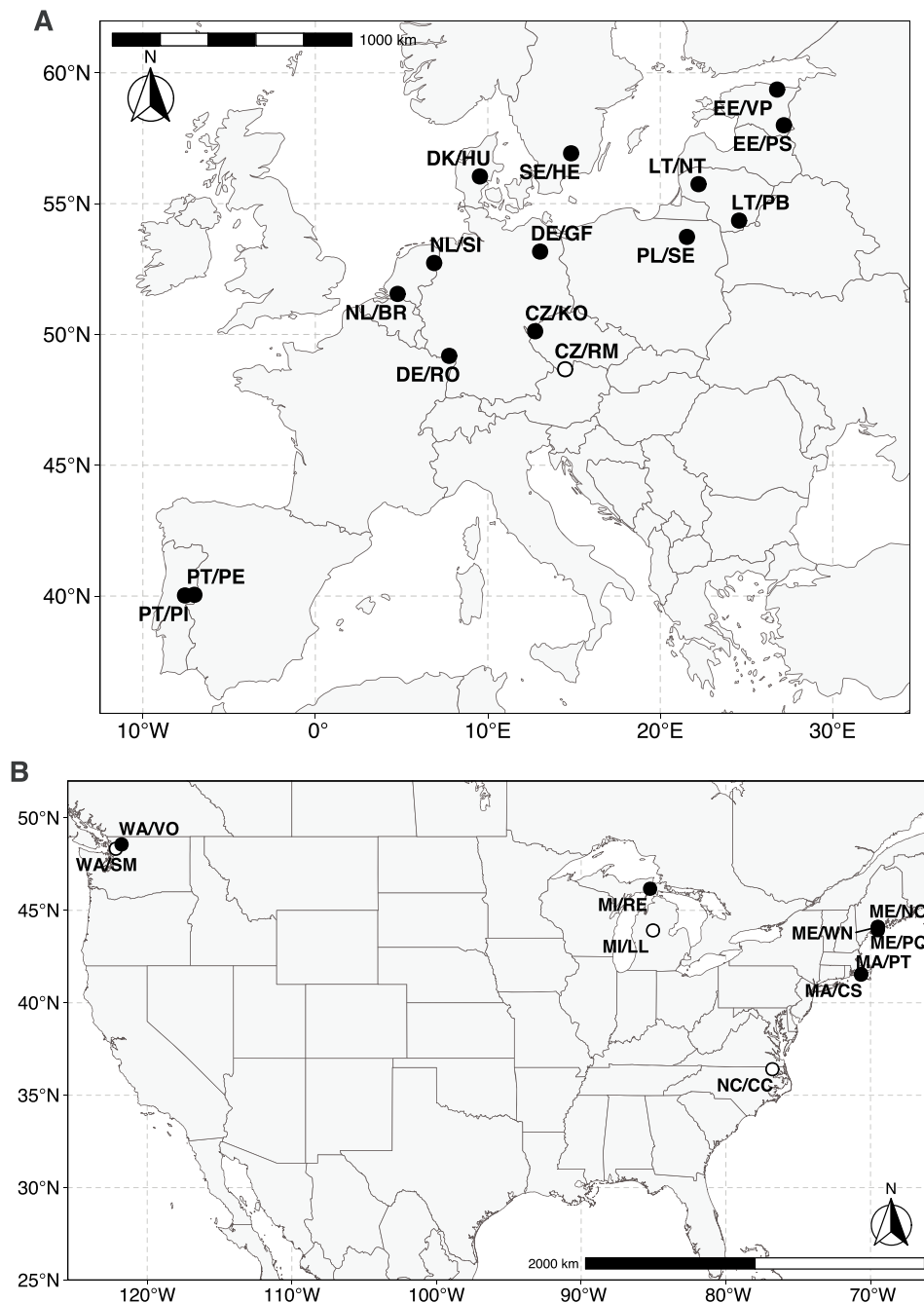


Figure 1. Maps of *G. semen* sampling sites in Europe (A) and in the USA (B). (A) Fifteen lakes in Europe—EE: EE/VP, EE/PS; LT: LT/NT, LT/PB; CZ: CZ/KO, CZ/RM; PL: PL/SE; DE: DE/RO, DE/GF; NL: NL/BR, NL/SI; DK: DK/HU; SE: SE/HE; PT: PT/PE, PT/PI. Population CZ/RM (white circle) was omitted in parts of the analysis because of few shared loci with other populations. See Table S1. For lake coordinates, sampling date and *G. Semen* abundance. (B) Ten lakes in the USA—WA: WA/VO, WA/SM; MI: MI/RE, MI/LL; ME: ME/NO, ME/NO, ME/PQ; MA: MA/PT, MA/CS; NC: NC/CC. Populations WA/SM, MI/CC, and NC/CC (white circles) were omitted in parts of the analysis because of few shared loci with other populations. See Table S2 for lake coordinates, sampling date and *G. semen* abundance.

quality assessment of all sample sequences was performed using FastQC v0.11.9 and MultiQC v1.11 [34]. Loci (RAD loci, i.e. genomic regions up- or downstream of a restriction site) were built and analyzed de novo by running the Stacks pipeline manually.

Parameters in Stacks ustacks were set to a minimum stack depth of 5 (parameter *m*) and a distance allowed between stacks of 3 (parameter *M*) based on Rennefors et al. [14] to maximize the number of utilized reads and polymorphic SNPs while maintaining a mean coverage of at least 20x. Loci

with extreme coverage (more than three standard deviations above the mean) were excluded from further analysis. The percentage of repetitive reads (Fig. S2) was determined as the fraction of reads attributed to those loci that were excluded because of extreme coverage. To mitigate effects of missing data caused by individuals with a high amount of missing loci and to ensure retrieval of sufficient loci across samples [35], samples with <50 000 putative loci were excluded from further analysis.

To create the catalog of consensus loci, representative samples with the highest mean coverage within the upper 0.5 quantile of putative loci were selected, restricting the maximum number of individuals per lake to half of its total number of individuals (resulting in five to eight samples per lake and 187 samples in total). The catalog was created in eight steps using *cstacks*, adding one sample per lake in each step (Fig. S4) and allowing for three mismatches between sample loci (parameter *n*).

Informative loci were determined in *Stacks* populations by specifying the minimum number of populations a locus must be present in (*p*) and the minimum proportion of individuals in a population a locus must be present in (*r*). Only the first SNP per RAD locus was utilized for population genetic analyses to avoid using multiple linked neighboring SNPs. Out of 657 sequenced samples in the RAD library, 422 were retained, corresponding to 5–34 samples per sampling location (Table S2). *Stacks* populations runs were carried out on three data sets with the following: *p*=13 and *r*=0.7 for all 15 lakes in Europe, *p*=7 and *r*=0.7 for a data set of seven lakes in the USA, and *p*=20 and *r*=0.6 when combining 22 lakes from both continents in one analysis. Lakes Washington (WA): Summer Lake (WA/SM), Michigan (MI): Lost Lake Fen (MI/LL), and North Carolina (NC): Colly Creek (NC/CC) in the USA were excluded from both data sets because of an extremely low number of loci (< 2%) shared among populations. The software R v3.6.3 [36] with *dplyr* v1.1.2 [37] was used in several analysis steps and the packages *ggplot2* v3.3.0 [37] and *ggpubr* v0.2.1 were used for plotting. Clone correction analysis using *poppr* v2.9.4 [38] identified no duplicated genotypes. Maps of sampling sites were plotted using the additional R packages *rmaple* v0.1.0, *ggspatial* v1.0.3 and *sf* v0.9-6 [39]. The *C_p* (i.e. time at the inflection point of the amplification curve) of MDA reactions was determined from amplification curves using *qpcr* v1.4-1 [40].

Population genetic analyses

Population differentiation (F_{ST}), and nucleotide diversity (π), observed and expected heterozygosity (H_o , H_e) were calculated directly in *Stacks* based on all SNPs utilized by the *Stacks* populations (i.e. the first SNP of each RAD locus). Calculations of nucleotide diversity were based on all called SNPs including low frequency variants. Lake Czechia (CZ): Rychnov nad Malší (CZ/RM) in Europe was excluded in subsequent analyses of coancestry (*fineRADstructure*), discriminant analysis of principal components (DAPC), and significant isolation by distance (IBD), because of a high number of missing loci. The population structure of the *G. semen* populations in Europe and in the USA was analyzed using several independent methods. A clustered coancestry matrix of all populations from Europe and the USA combined was created in *fineRADstructure* [41], using a missing data cutoff of 40%. For MCMC in *fineRADstructure*, 500 000 burn-in iterations were followed by 500 000 sample iterations, and a thin interval of 5000. The coancestry matrix contains counts the number of most similar haplotypes (RAD loci) in pairwise comparisons of individuals. Population structure was further investigated using the Bayesian assignment approach implemented in *Structure* v2.3 [42, 43]. The analysis was performed separately for each of the two continents with the clue for *K* ranging from 1 to 8 (USA) or from 1 to 15 (Europe), assuming an admixture model, correlated allele frequencies, and without location priors. A burn-in period of 50 000 steps followed by 100 000 additional repetitions were performed in each *Structure* run. For each *K*, 20 independent iterations were performed. The most likely *K* to describe population clusters in

the data was estimated with the Evanno method [44, 45] using the R package *pophelper* v2.3.1. *Structure* plots were visualized using *CLUMPAK* [46]. Analysis of molecular variance (AMOVA) was performed in *GenoDive* v3.05 [47]. DAPC was performed on the combined data set, as well as on the Europe and the US data sets separately in R with the packages *ade4* v2.1.5 [48] and *vegan* v2.5-7. The optimal number of principle components (PCs) to retain in DAPC was determined thorough cross-validation and *a-score* optimization. Previously published environmental variables of the different lakes [49] were superimposed on the DAPC ordination plots. No environmental data was available for lakes CZ: Komáří Rybník (CZ/KO) and Sweden (SE): Helgasjön (SE/HE). To test significant IBD, a Mantel test of genetic distances against geographic distances was performed in R with package *ade4* v1.7-18 [50]. This analysis was performed separately for populations in the USA, in Europe, and for the eastern and western population clusters in Europe as identified by the *Structure* analysis.

Results

Library demultiplexing, read filtering, and de novo assembly of RAD loci

The total number of reads in all 18 sequenced RAD libraries of randomly pooled SAGs was 15.3 billion, 98.3% of which were retained after demultiplexing and quality filtering through *Stacks* process_radtags. Around 5.6% of all reads were identified as potential contaminant reads through taxonomic sequence classification using *Kraken* (Fig. S1) and were discarded. Evaluation of amplification curves revealed a strong positive correlation of the amplification *C_p* and the percentage of repetitive reads found in a sample, along with a strong negative correlation of *C_p* and the number of putative loci that are recovered (Fig. S2). This suggests that exclusion of sequenced SAG-RAD samples during the analysis because of missing data could be minimized by evaluating MDA amplification curves and selection of SAGs for sequencing based on the quality of the amplification. Genome recovery (i.e. the number of loci obtained from a sample) was strongly correlated with the timing of amplifications throughout all populations in this study.

After exclusion of samples for which <50 000 loci were recovered, the mean number of putative loci was 92 995 (mean coverage 77.8×) per sample (see Fig. S3 for all samples before filtering). The final catalog created using *Stacks* *cstacks* from 187 samples contained a total of 2 498 256 loci (Fig. S4). Across the 15 populations from lakes in Europe, 664 variant sites were identified, that were present in 70% of the individuals or more in at least 13 of the populations (Table 1). Across the seven populations from lakes in the USA, 708 variant sites were identified that were shared among at least 70% of the individuals in all seven populations (Table 1). In the combined data set with all 22 lake populations from both continents, 2452 variant sites were identified in at least 60% of the individuals in 22 populations or more. To keep the amount of missing data in the analysis to a minimum, most of the analyses of population differentiation and genetic structure were performed on separate data sets of populations in Europe and populations in the USA.

Population genetic metrics

The analysis of population genetic metrics (Table 1) was based on a total number of 664 variant sites from 236 individuals in Europe and 708 variant sites from 124 individuals in the USA. The percentage of polymorphic sites among all sites was overall very low and ranged from 0.005 (Germany (DE): Große Fuchskuhle

Table 1. Population genetic metrics of *G. semen* lake populations in Europe and in the USA. Table contains population ID, number of individuals in each population, number of all sites (variant and fixed), percentage of polymorphic sites, number of private alleles, number of variant sites, nucleotide diversity π , observed heterozygosity H_o , expected heterozygosity H_e , and the inbreeding coefficient F_{is} . Europe: analysis based on a total number of 664 variant sites from 236 individuals. USA: analysis based on a total number of 708 variant sites from 124 individuals.

	Pop.	Indiv.	Sites	% Polym. sites	Private alleles	Variant sites	π	H_o	H_e	F_{is}
Europe	CZ/KO	14	713 473	0.010	15	634	0.024	0.024	0.023	0.002
	CZ/RM	11	283 227	0.032	53	250	0.119	0.115	0.112	0.014
	DE/GF	11	718 891	0.005	6	639	0.012	0.014	0.011	−0.005
	DE/RO	21	713 959	0.012	20	645	0.020	0.019	0.019	0.003
	DK/HU	15	696 211	0.010	14	619	0.018	0.016	0.018	0.011
	EE/PS	7	722 635	0.006	7	650	0.018	0.018	0.017	0.002
	EE/VP	9	683 719	0.007	6	609	0.020	0.016	0.018	0.010
	LT/NT	16	705 947	0.007	13	630	0.012	0.011	0.012	0.006
	LT/PB	25	670 825	0.035	111	611	0.037	0.027	0.036	0.070
	NL/BR	34	729 975	0.023	35	656	0.028	0.024	0.027	0.020
	NL/SI	23	730 964	0.024	39	656	0.034	0.028	0.033	0.028
	PL/SE	13	579 642	0.011	21	515	0.020	0.018	0.019	0.003
	PT/PE	14	706 333	0.008	11	636	0.015	0.016	0.014	−0.003
	PT/PI	15	660 157	0.008	17	601	0.015	0.016	0.015	0.001
	SE/HE	8	645 303	0.008	11	583	0.020	0.014	0.018	0.012
USA	MA/CS	20	691 287	0.022	52	708	0.039	0.025	0.038	0.049
	MA/PT	20	691 275	0.021	50	708	0.046	0.031	0.044	0.044
	ME/NO	5	691 148	0.014	58	708	0.044	0.032	0.039	0.031
	ME/PQ	15	691 229	0.027	83	708	0.050	0.032	0.048	0.062
	ME/WN	13	691 174	0.022	54	708	0.044	0.030	0.042	0.044
	MI/RE	23	691 228	0.023	68	708	0.046	0.047	0.045	0.005
	WA/VO	28	691 282	0.030	108	708	0.057	0.061	0.056	−0.005

(DE/GF) to 0.035 (Lithuania (LT): Pabzninkai (LT/PB)) in Europe and 0.014 (Maine (ME): Nobleboro (ME/NO)) to 0.03 (WA: Vogler Lake (WA/VO)) in the USA. The mean nucleotide diversity π of variant sites ranged from 0.012 to 0.119 across all populations with values in the USA generally exceeding those in Europe (Wilcoxon signed-rank test: $p = 0.001$). In Europe, π varied mostly between 0.012 (DE/GF and LT: Natalka (LT/NT)) and 0.037 (LT/PB), but was as high as 0.119 in CZ/RM. In the USA, values for π ranged from 0.039 (Massachusetts (MA): Cedar Swamp (MA/CS)) to 0.057 (WA/VO). Compared with the other populations, CZ/RM only shared a relatively low number of 250 variant sites (283 227 variant and fixed sites) with other populations in Europe and had a relatively high number of 53 private alleles (i.e. an allele that is found in only one of several populations). The number of private alleles varied strongly between populations and ranged from 6 (Estonia (EE): Viitna Pikkjärv (EE/VP) and DE/GF) up to 111 (LT/PB). In the USA, the number of private alleles ranged from 50 (MA: Peterson Pond (MA/PT)) to 108 (WA/VO). Heterozygosity (a measure of genetic variation in a population) was overall very low, especially in European populations, where mean values of observed heterozygosity varied between 0.011 (LT/NT) and 0.028 (the Netherlands (NL): Siepeldijk (NL/SI)) for all populations except CZ/RM with 0.115. In the USA, observed heterozygosity was slightly higher and ranged from 0.025 (MA/CS) to 0.061 (WA/VO). Mean values of expected heterozygosity were similar but tended to be slightly higher than those of observed heterozygosity as is also indicated by inbreeding coefficients (F_{is}) ranging from −0.005 (DE/GF and WA/VO) to a maximum of 0.07 (LT/PB).

A combined analysis of 2452 loci in 360 individuals from lake populations across Europe and the USA (Table S6) showed similar patterns in terms of a generally higher number of private alleles, higher nucleotide diversity, and higher heterozygosity in the USA compared with Europe.

Population differentiation

Analysis of the SNP data from *G. semen* lake populations in Europe and in the USA showed genetic differentiation between all lake populations. Overall, values of pairwise genetic distance (F_{ST}) between lake populations were significant and ranged from moderate to high differentiation. In Europe (Fig. 2A), lowest F_{ST} values around 0.019–0.026 were generally found between lakes that were in relative proximity. Population CZ/RM appeared highly differentiated from all other lakes in Europe and very high F_{ST} values, ranging from 0.309 to 0.447, were observed between CZ/RM and all other lakes. F_{ST} values were generally higher in the USA (Fig. 2B), even between populations within a few kilometers of distance, and ranged from 0.052 (ME: West Neck Pond (ME/WN) and Pemaquid River (ME/PQ)) up to around 0.14 for pairwise comparisons between the Massachusetts lakes and ME/NO. With all F_{ST} values above 0.1, the two lakes in Massachusetts appeared to be highly differentiated from all other lakes in the USA.

In a combined analysis of lakes across Europe and the USA, F_{ST} values between lakes within each of the two continents were generally lower than between lakes on different continents (Fig. S5). The lowest F_{ST} values were observed between lakes within Europe and values between lakes within the USA were overall higher. However, the ranges of F_{ST} values between and within continents were not clearly distinct and largely overlapped.

Population genetic structure between continents

Population structure of the *G. semen* populations in Europe and in the USA was analyzed using several independent methods. Clustering of populations generally correlated with geography, both across and within Europe and the USA.

A combined DAPC of populations in Europe and in the USA revealed a distinct separation of European and US population

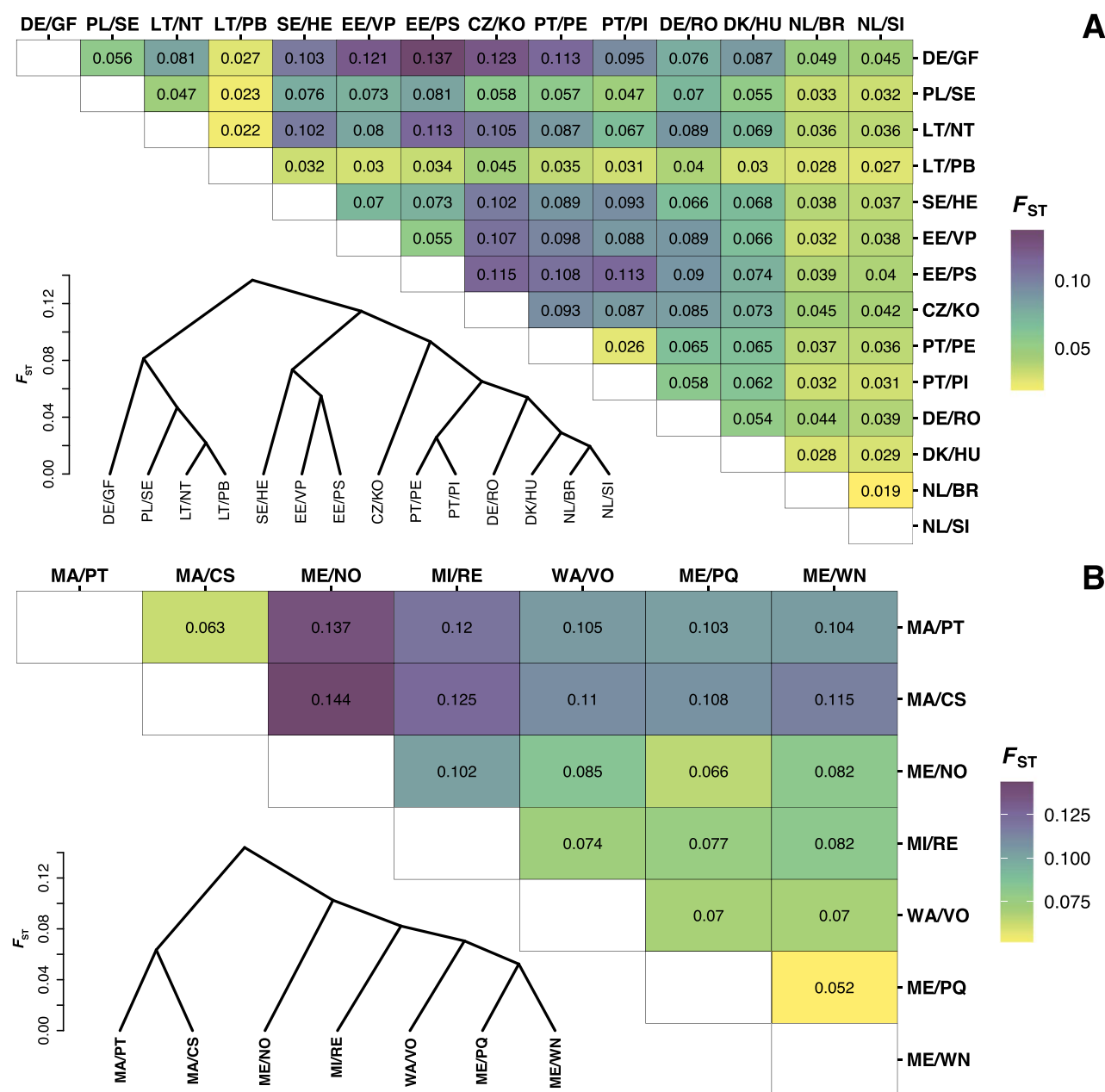


Figure 2. Pairwise genetic distance (F_{ST} values) between lake populations of *G. semen* in Europe* (A) and in the USA (B). Upper right: Heatmap table of F_{ST} values. Lower left: hierarchical clustering tree that was used to arrange the order of population in the heatmap table. * F_{ST} values between population CZ/RM and all other populations ranging from 0.309 to 0.447 were omitted from the heatmap table.

clusters (Fig. S6). An AMOVA of *G. semen* populations across Europe and the USA showed significant differentiation of populations both within (35.3% of the variation; $F = 0.38$; $p = 0.001$) and between continents (37.9% of the variation; $F = 0.57$; $p = 0.001$).

Analysis of coancestry (nearest neighbor haplotype relationships) between all populations in the USA and in Europe using fineRADstructure (Fig. 3) also showed that European and US populations form distinct clusters. Estimated coancestry within population clusters was generally higher in the USA than in Europe. Ten clusters on five main branches were identified within the European population cluster. These included an eastern branch with the populations LT/NT, LT/PB, Poland (PL): Sęczek (PL/SE), and DE/GF. The population CZ/KO formed a separate central branch. A western central branch consisted of Denmark (DK): Hundsø

(DK/HU), NL/SI, and NL: Breda Zuid Oost (NL/BR). A northeast branch included the populations SE/HE, EE/VP, and EE: Partsi Saarijärvi (EE/PS). And a southwest branch contained the populations DE: Rohrwoog (DE/RO), Portugal (PT): Penha Garcia (PT/PE), and PT: Pisco (PT/PI). In the USA, the coancestry analysis identified seven clusters on four main branches, separated by states.

Population genetic structure within continents

Fine scale population structure was analyzed separately for populations in Europe and in the USA using the program Structure. Based on the Evanno method [45], the best clustering of populations in Europe was with two inferred population clusters ($K = 2$, Fig. S7), separating population CZ/RM from all other populations (Fig. 4A). However, more clusters were identified and

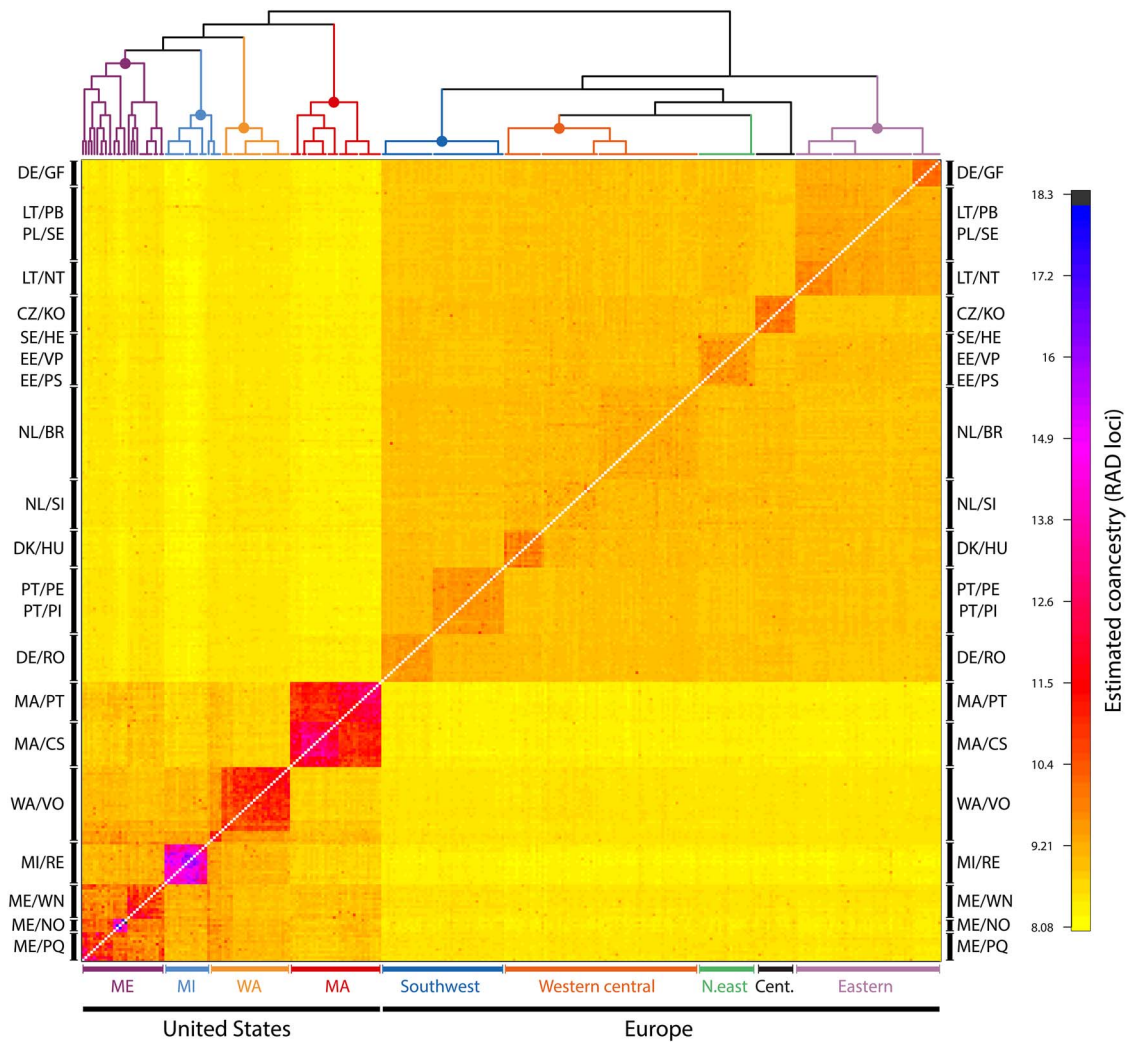


Figure 3. Clustered fineRADstructure coancestry matrix of *G. semen* populations in Europe and in the USA. The heat map depicts pairwise coancestry (i.e. the number of most similar haplotypes (RAD loci) in pairwise comparisons of individuals) based on RAD loci among individuals according to the color scale shown on the right. Labels on the left and right denote clusters of individuals in lake populations. Bottom labels denote major geographic clusters of populations on the two different continents as indicated by the hierarchical clustering tree (color coded) on top of the coancestry matrix.

the log-likelihood for higher values of K increased until $K=5$, showing distinct western and eastern clusters with $K=3$. For $K=4$, the western cluster was further divided into a central European cluster and the three populations DE/RO, PT/PE, and PT/PI in the southwest formed a separate cluster. By including a fifth group ($K=5$) no additional cluster was formed. In the Structure analysis of populations in the USA (Fig. 4B), the best clustering was with three inferred population clusters ($K=3$, Fig. S8). These included one cluster of both lakes in Massachusetts, a second cluster formed by the lakes in Maine, with Michigan and Washington as a third cluster. The Michigan lake population was separated from the Maine populations for higher values of K .

Potential drivers of population differentiation

Overall, population differentiation was correlated with geographic distance between lake populations within a certain range of geographic distance, but not across the full geographic scale. IBD across all populations in Europe (Fig. 5A) was not significant (Mantel statistic $r=0.22$, $p=0.118$). However, when population genetic clusters according to Structure were used, another pattern emerged. The eastern cluster (containing lake populations in

Estonia, Sweden, Lithuania, and northeastern Germany) showed significant IBD (Mantel statistic $r=0.49$, $p=0.028$). Significant IBD (Mantel statistic $r=0.41$, $p=0.038$) was also found in the western cluster (containing lake populations in Denmark, the Netherlands, southwestern Germany, and Portugal). In contrast, the relationship between geographic distance and genetic differentiation was slightly negative in the USA (Fig. 5B) and there was no sign of IBD between populations (Mantel statistic $r=-0.14$, $p=0.569$).

Separate DAPC analyses of Europe (Fig. 6A) and the USA (Fig. 6B) generally showed a clustering pattern that reflected their geography. In Europe, a northeastern, a central and a southwestern cluster were observed. Population CZ/KO in Czech Republic formed a separate cluster. The two populations in Germany were separated with DE/RO being part of the southwestern and DE/GF being part of the northeastern cluster. In the USA, three separate population clusters were identified, similar to the Structure analysis. The populations in Maine and Massachusetts formed two separate clusters adjacent to a cluster including WA/VO and MI: Rexton Bog (MI/RE).

Environmental vectors plotted on the DAPC ordination showed that the separation of the western cluster in Europe correlated

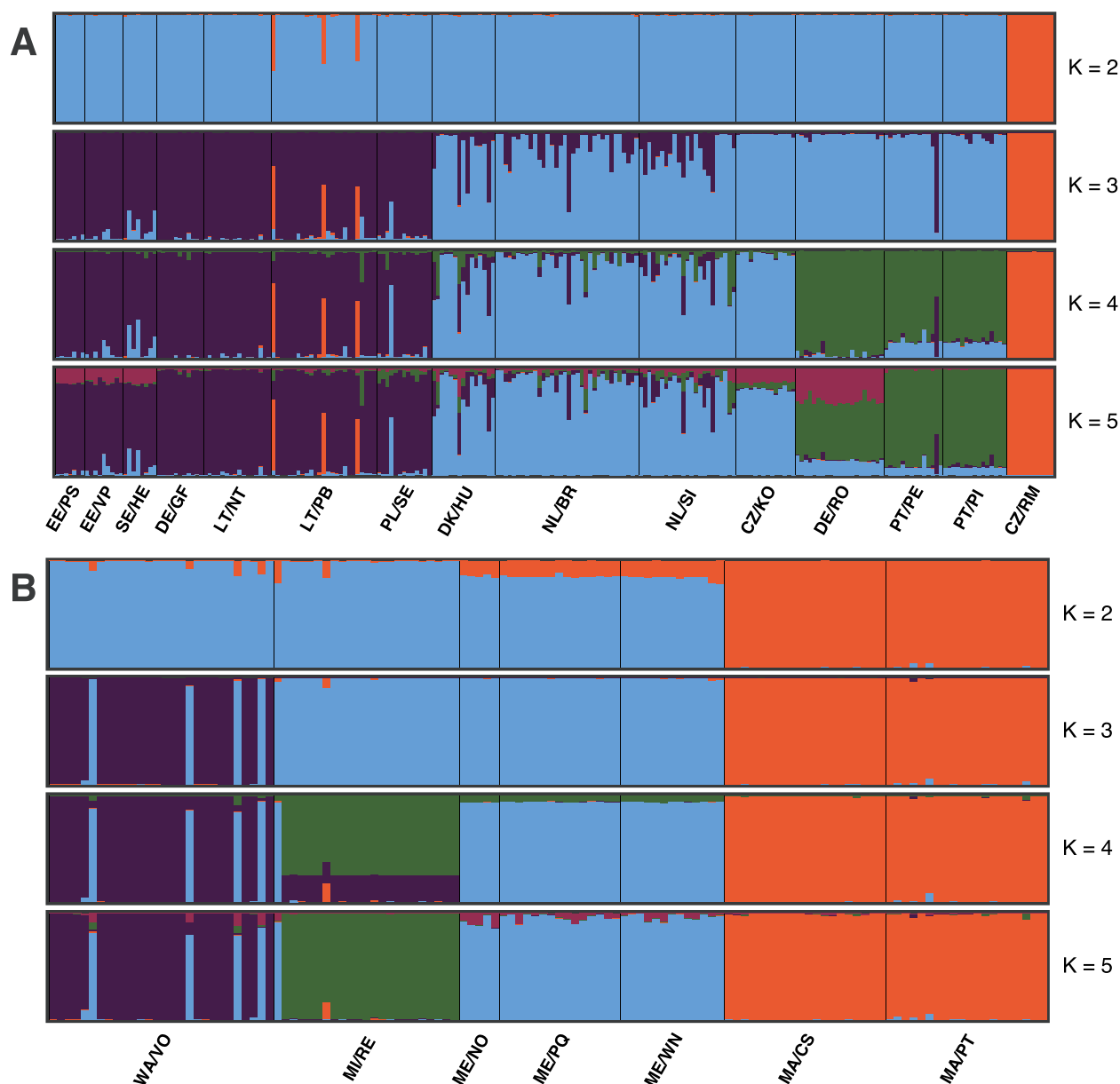


Figure 4. Structure plots of *G. semen* lake populations in Europe (A) and in the USA (B) for clustering values $K=2$ – 5 . Each individual strain is represented by a vertical bar, which is partitioned into K -colored segments that represent estimated assignment fractions of strains in K population clusters (along the y-axis). Bottom labels indicate country/abbreviation of lake names (see Fig. 1).

with higher pH and calcium (Ca) concentration. The eastern cluster, in contrast, correlated with higher concentrations of iron (Fe), total phosphorus (TP), total nitrogen (TN), dissolved organic carbon (DOC), and higher water color (wc). In the USA, separation of the two lake populations in Massachusetts was mainly correlated with lower pH and higher concentrations of Fe, TN, TP, as well as higher wc, compared with the other lakes. The separation of the three lake populations in Maine was correlated with higher concentrations of Ca.

Discussion

Here we present a phytoplankton population genomic study based on single cells isolated directly from lake samples without culturing. We accomplished this by using the new method SAG-RAD, which is based on SAGs and reduced representation

sequencing, followed by SNP analyses. Our main finding was that *G. semen* forms two distinct geographic clusters, one in Europe and one in the USA, demonstrating a substantial dispersal barrier between the continents. In addition, we detected geographic population structure of *G. semen* within each of the two continents connected to environmental factors and isolation-by-distance. Below we discuss these results in relation to dispersal and distribution of phytoplankton, as well as considerations when using SAG-RAD.

Although *G. semen* is widely distributed, it forms two distinct genetic clusters that separate the populations in Europe and in the USA. This was clear both in the DAPC and fineRADstructure coancestry analyses, which are analyses that are based on very different assumptions, where DAPC does not rely on a particular population genetic model [51]. This distinct pattern implies that gene flow between the continents is unlikely and that the

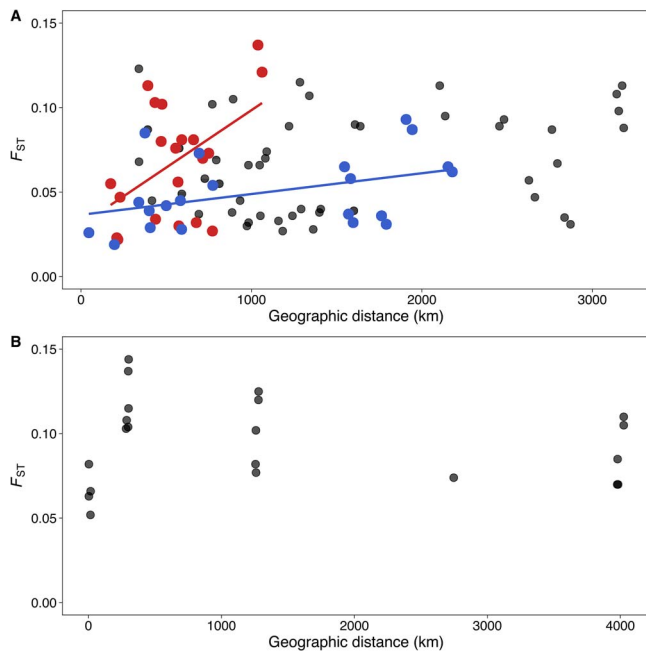


Figure 5. IBD of *G. semen* populations in Europe (A) and in the USA (B). Genetic distance (F_{ST}) is plotted against geographic distance. (A) Europe—no IBD overall (black), IBD significant in the eastern cluster (red with linear fit; $R^2 = 0.24$), IBD significant in the western cluster (blue with linear fit; $R^2 = 0.17$); (B) USA—no IBD overall.

dispersal barriers are high. Given that the two populations are separated by an ocean, this finding was not unexpected, but demonstrates that although *G. semen* is a cosmopolitan species, the populations are highly differentiated.

Within both Europe and the USA, we observed clear geographic clustering based on coancestry (genetic similarity) analyses. There were also differences in intra-population coancestry, with a more pronounced isolation of populations in the USA compared with Europe. Possibly, this difference is related to the direction of the isolation. Within the USA, most of the isolation among lakes is on an east–west scale, which means a greater disconnection in terms of dispersal vectors such as birds, whose flyways tend to be in a north–south direction [52]. Pairwise genetic differentiation (F_{ST}) was generally higher in lake populations between Europe and the USA, compared with F_{ST} of populations within each continent. However, this was not always the case, despite that the USA and Europe are separated by around 5000 km across the Atlantic Ocean. This suggests the relevance of other factors than physical isolation through geographic distance driving differentiation of *G. semen* populations, such as environmental factors leading to local adaptation [53].

In Europe, we found a split between a northeastern and a southwestern cluster of lake populations. The Structure analysis indicated an eastern cluster of lakes in Estonia, Lithuania, Sweden, Poland, and north-east Germany, that was separated from a western cluster spanning lakes in Denmark, the Netherlands, Czech Republic, south-west Germany and Portugal. A divide into an eastern and western population cluster, along a similar longitudinal axis, was observed in Northern Europe between Sweden and Finland in the east and Norway in the west [14]. Drivers behind this separation remain unclear but could potentially be connected to bird migratory paths [14, 54]. In the USA, we similarly observed a clear separation of three distinct population clusters that coincided with the geography of the sampled lakes. These clusters were separated by latitude, further strengthening the

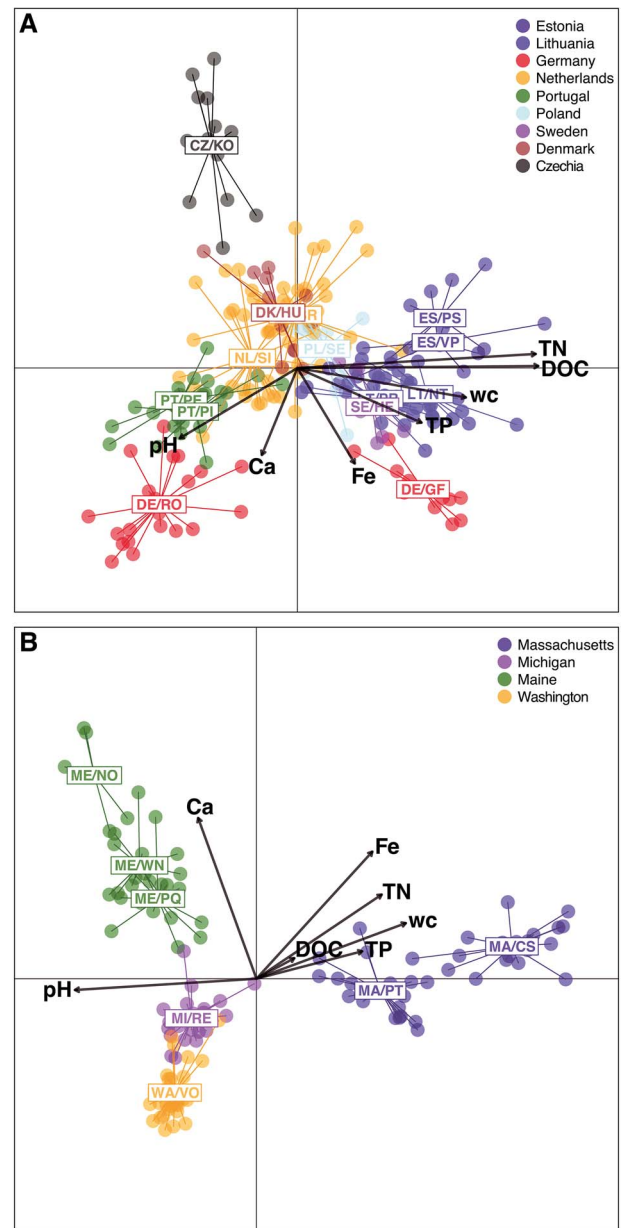


Figure 6. DAPC of *G. semen* lake populations within continents. Points represent individuals and lines connect individuals within populations. The number of retained PCs was determined using cross-validation and a-score optimization. Color code denotes country or state (USA). (A) In Europe, retaining 13 PCs. (B) In the USA, retaining 12 PCs. Vectors of environmental variables are superimposed on the DAPC ordination scores. Superimposed environmental variables (from [49]) are pH, wc, and concentrations of Ca, Fe, TP, TN, and DOC. No environmental data is available for lakes CZ/KO and SE/HE.

bird migratory pathway hypothesis. Coleman [55] found a similar pattern in the green alga *Pandorina morum* indicating that the east coast, mid-west, and west coast did not share lineages. Both Coleman's and our study suggest that east/west transport by westerlies (wind) was less effective than north–south transport (by migrating waterfowl), because their flyways correspond to these divisions [56]. Although bird transport of *G. semen* has not been studied, transport likely occurs through resting cysts, which are more tolerant than motile cells and easily germinate in non-local water [57]. Moreover, Tesson and Santl-Temkiv [58] showed that *G. semen* cells and cysts did not tolerate freezing and desiccation, indicating poor airborne dispersal.

Three populations in the USA (WA/SM, MI/LL, and NC/CC) and to a lesser degree CZ/RM in Europe appeared to be highly differentiated from other populations and had to be omitted in the analysis, or parts thereof. Both microscopic inspection and the partial 18S rRNA gene sequences clearly suggested *G. semen* for all isolates. However, although the partial sequence of *coxI* matched *G. semen* in both NC/CC and MI/LL albeit with a lower percent identity, WA/SM had a weak match to *Chattonella marina* var. *antiqua*. Because it is unlikely that the lineage belongs to *C. marina* (reported from brackish but not freshwater habitats) and the WA/SM lineage had high similarity to *G. semen* in the 18S rRNA gene, the species identification remains unclear. This raises the question whether these lineages represent cryptic species.

Overall, we observed higher nucleotide diversity, higher heterozygosity, and more polymorphic loci in lake populations of *G. semen* in the USA compared with populations in Europe. This suggests that populations in the USA could have been separated for a longer time than those in Europe, supporting our hypothesis of a more recent expansion of *G. semen* in Europe. Although we cannot directly infer directionality from our data, a possible scenario underlying the observed differences in genetic diversity between populations on the two continents is that the European populations of *G. semen* represent a lineage that diverged from an ancestral lineage in the USA followed by a recent expansion in Europe.

Physical dispersal limitation, or geographic distance, does play a role in generating the population genetic patterns we observed in *G. semen*, but not at all distance ranges or in all directions. Lakes in close geographic proximity usually had a low but significant genetic distance (F_{ST} value) both in the USA and in Europe. We further observed a trend of higher genetic distance with higher geographic distance both within the eastern and the western cluster in Europe, but not when combined. We conclude that the correlation between geographic distance and genetic differentiation seems to break down at larger distances. In the USA, we found no signs of isolation-by-distance across the long distances between populations included in this study, which as indicated above could be because of main dispersal by migratory birds. Although water, wind, and animal vectors [59, 60] are potential means of microalgal dispersal, the scales, and the extent at which these drivers act are largely unknown. Sassenhagen et al. [28] suggested the relevance of different dispersal mechanisms at different geographic scales and emphasized the potential role of animal vectors for dispersal between adjacent lakes. At larger geographic scales, patterns of population differentiation might be influenced by different source populations, migratory birds, and more stochastic dispersal events [28]. Finally, priority effects of first colonizers or genetic anchoring effects could explain limited gene flow between populations independent of dispersal vectors or geographic isolation [14].

Environmental selection likely also played a role in population differentiation of *G. semen*. We observed differences in environmental variables between population clusters within Europe and the USA. These environmental differences were mainly high pH and high Ca concentration on the one hand, and higher wc and high concentrations of TN, TP, DOC, and Fe on the other hand. Among the latter are conditions that are typically associated with a dominance of *G. semen* [61] and represent the characteristics of waterbodies where high density blooms of the species were first described (e.g. [62]). In the laboratory, we have shown that *G. semen* growth requires high Fe [63] and field observations connect Fe and presence of *G. semen* blooms [64]. Moreover, in Gollnisch et al. [49] we showed that *G. semen* distribution may be limited

by a combination of high pH and high Ca concentration. The observed environmental heterogeneity could cause local adaptation within population clusters [65], limiting gene flow between different environmental clusters and thus facilitating genetic differentiation [66].

We observed relatively low genetic diversity, i.e. expected and observed heterozygosity in this data set compared with Rengefors et al. [14]. This is likely because of allelic dropout (ADO) during single-cell MDA, when the amplification is biased toward one allele in the presence of two alleles in a heterozygous individual resulting in an underestimation of heterozygosity. Rates of ADO are generally higher in SAG-RAD compared with RAD from extracted DNA of clonal cultures [18]. A critical step in the analysis of SAG-RAD libraries that can cause ADO is likely the construction of the RAD loci catalog. For large data sets, the number of samples used to build the catalog in the Stacks analysis should generally be limited to around 40–200 samples to minimize noise in the catalog [67]. This requires selecting a subset of samples with high coverage that is representative of the genetic diversity in the data set [67]. Therefore, in this study we selected 187 individuals (out of 422), based on number and coverage of loci, to construct the RAD catalog. Construction of the catalog from a subset of samples can, however, cause loss of low minor allele frequency variants in the data set [67]. If an allele is not included in the catalog at this step, it would be discarded when sample loci are later mapped against the catalog and cause ADO during RAD analysis. Heterozygosity could therefore be another criterion, besides number and coverage of loci, to select representative samples used to build the catalog, especially in the analysis of SAG-RAD libraries.

In this study we demonstrated that SAG-RAD provides an important alternative to traditional culture-based studies of population genetics in microalgae by circumventing major limitations associated with algal culturing. However, when using SAG-RAD, the extent of genome recovery and allelic dropout in samples, i.e. true heterozygous sites may appear as homozygous, need to be considered. This effect is because of amplification bias in MDA [68], the method used for amplification of single-cell genomes in SAG-RAD. We show that Cp can be used as a proxy for amplification quality prior to sequencing (Fig. S2).

To conclude, single-cell population genomics using SAG-RAD enabled us to analyze the genetic structure and differentiation of *G. semen* across a large geographic distance including over 20 lakes and hundreds of individuals. This allowed us to identify two distinct population clusters separating *G. semen* in Europe and in the USA indicating a divergence of the species that predates the present distribution. We found that genetic differentiation of the populations is influenced by both physical dispersal limitation and different environmental conditions. Finally, the relatively low genetic diversity in European populations supports the hypothesized recent expansion of *G. semen*. In the future, when a genome will be available for this important bloom-forming species, we anticipate that our population genomic data can be used to further investigate genes under selection, local adaptation, and links between genotypes and phenotypes.

Acknowledgements

Many thanks to Catarina Churro, Dedmer van de Waal, Donald M. Anderson, Dora Certnerova, Hans-Peter Grossart, Ingmar Ott, Judita Koreviene, Paul A. Moore, Pavel Škaloud, Carmelo Thomas, Ramunas Stepanauskas, Robin Matthews and Virginia Edgcomb for their advice and help during field sampling. Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The

facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. The computations and data handling were enabled by resources in project SNIC 2017/7-349 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Supplementary material

Supplementary material is available at *The ISME Journal* online.

Conflicts of interest

None declared.

Data availability

Analysis code is available on GitHub at https://github.com/RGollnisch/Gsemen_SAG-RAD and on Zenodo. Sequence data are available through the NCBI SRA database (BioProject ID PRJNA988296) at <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA988296>.

Funding

This study was supported by the EU ITN program SINGEK (Marie Skłodowska-Curie grant 675752) and the Crafoord Foundation (20170722) to K.R. Additional funding was provided by the Botanical Society Lund, Jörgen Lindströms foundation, Ove Almborn Donation Fund, the Royal Physiographic Society of Lund, and the University of Michigan Biological Station Fellowship.

References

- Falkowski PG, Knoll AH. An introduction to primary producers in the sea: who they are, what they do, and when they evolved. In: Falkowski P.G., Knoll A.H. (eds.), *Evolution of Primary Producers in the Sea*. Academic Press, Cambridge MA USA, 2007, 1–6.
- Simon N, Cras A-L, Foulon E et al. Diversity and evolution of marine phytoplankton. *Comptes Rendus Biologies* 2009;**332**: 159–70. <https://doi.org/10.1016/j.crv.2008.09.009>
- Finlay BJ. Global dispersal of free-living microbial eukaryote species. *Science* 2002;**296**:1061–3. <https://doi.org/10.1126/science.1070710>
- de Wit R, Bouvier T. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environ Microbiol* 2006;**8**:755–8. <https://doi.org/10.1111/j.1462-2920.2006.01017.x>
- Fenchel T, Finlay BJ. The ubiquity of small species: patterns of local and global diversity. *Bioscience* 2004;**54**:777–84. [https://doi.org/10.1641/0006-3568\(2004\)054\[0777:TUOSSP\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2004)054[0777:TUOSSP]2.0.CO;2)
- O'Malley MA. 'Everything is everywhere: but the environment selects': ubiquitous distribution and ecological determinism in microbial biogeography. *Stud Hist Phil Biol Biomed Sci* 2008;**39**: 314–25. <https://doi.org/10.1016/j.shpsc.2008.06.005>
- Martiny JBH, Bohannan BJM, Brown JH et al. Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* 2006;**4**:102–12. <https://doi.org/10.1038/nrmicro1341>
- Škaloud P, Jadrná I, Dvořák P et al. Rapid diversification of a free-living protist is driven by adaptation to climate and habitat. *Curr Biol* 2024;**34**:92–105.e6. <https://doi.org/10.1016/j.cub.2023.11.046>
- Nagai S, Lian C, Yamaguchi S et al. Microsatellite markers reveal population genetic structure of the toxic dinoflagellate *Alexandrium tamarense* (Dinophyceae) in Japanese coastal waters. *J Phycol* 2007;**43**:43–54. <https://doi.org/10.1111/j.1529-8817.2006.00304.x>
- Kashtan N, Roggensack SE, Rodrigue S et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* 2014;**344**:416–20. <https://doi.org/10.1126/science.1248575>
- Vyverman W, Verleyen E, Sabbe K et al. Historical processes constrain patterns in global diatom diversity. *Ecology* 2007;**88**: 1924–31. <https://doi.org/10.1890/06-1564.1>
- Rengefors K, Logares R, Laybourn-Parry J. Polar lakes may act as ecological islands to aquatic protists. *Mol Ecol* 2012;**21**:3200–9. <https://doi.org/10.1111/j.1365-294X.2012.05596.x>
- Van den Wyngaert S, Möst M, Freimann R et al. Hidden diversity in the freshwater planktonic diatom *Asterionella formosa*. *Mol Ecol* 2015;**24**:2955–72. <https://doi.org/10.1111/mec.13218>
- Rengefors K, Gollnisch R, Sassenhagen I et al. Genome-wide single nucleotide polymorphism markers reveal population structure and dispersal direction of an expanding nuisance algal bloom species. *Mol Ecol* 2021;**30**:912–25. <https://doi.org/10.1111/mec.15787>
- Rengefors K, Kremp A, Reusch TBH et al. Genetic diversity and evolution in eukaryotic phytoplankton: revelations from population genetic studies. *J Plankton Res* 2017;**39**:165–79. <https://doi.org/10.1093/plankt/fbw098>
- Epstein SS. The phenomenon of microbial uncultivability. *Curr Opin Microbiol* 2013;**16**:636–42. <https://doi.org/10.1016/j.mib.2013.08.003>
- Rappé MS, Giovannoni SJ. The uncultured microbial majority. *Annu Rev Microbiol* 2003;**57**:369–94. <https://doi.org/10.1146/annurev.micro.57.030502.090759>
- Gollnisch R, Wallenius J, Gribble KE et al. SAG-RAD: a method for single-cell population genomics of unicellular eukaryotes. *Mol Biol Evol* 2023;**40**:msad095. <https://doi.org/10.1093/molbev/msad095>
- Karosiene J, Kasperovičienė J, Koreivienė J et al. Assessment of the vulnerability of Lithuanian lakes to expansion of *Gonyostomum semen* (Raphidophyceae). *Limnologia* 2014;**45**:7–15. <https://doi.org/10.1016/j.limno.2013.10.005>
- Rengefors K, Weyhenmeyer GA, Bloch I. Temperature as a driver for the expansion of the microalga *Gonyostomum semen* in Swedish lakes. *Harmful Algae* 2012;**18**:65–73. <https://doi.org/10.1016/j.hal.2012.04.005>
- Hagman CHC, Ballot A, Hjermand DØ et al. The occurrence and spread of *Gonyostomum semen* (Ehr.) Diesing (Raphidophyceae) in Norwegian lakes. *Hydrobiologia* 2015;**744**:1–14. <https://doi.org/10.1007/s10750-014-2050-y>
- Lepistö L, Antikainen S, Kivinen J. The occurrence of *Gonyostomum semen* (Ehr.) Diesing in Finnish lakes. *Hydrobiologia* 1994;**273**:1–8. <https://doi.org/10.1007/BF00126764>
- Le Cohu PR, Guitard J, Comoy N et al. *Gonyostomum semen* (Raphidophycées), nuisance potentielle des grands réservoirs français? L'exemple du lac de Pareloup. *Arch Hydrobiol* 1989;**117**:225–36. <https://doi.org/10.1127/archiv-hydrobiol/117/1989/225>
- Paulino S, Vilares A, Valério E. Ocorrência e disseminação da microalga *Gonyostomum semen* em albufeiras portuguesas. Instituto Nacional de Saúde Doutor Ricardo Jorge, IP, Lisbon Portugal; Boletim Epidemiológico Observações 2015 julho-setembro;**4**(13):23–5. 2015.
- Negro AI, De Hoyos C, Vega JC. Phytoplankton structure and dynamics in Lake Sanabria and Valparaíso reservoir (NW

- Spain). *Hydrobiologia* 2000;**424**:25–37. <https://doi.org/10.1023/A:1003940625437>
26. Watson SB, Whitton BA, Higgins SN et al. Harmful algal blooms. In: Wehr J.D., Sheath R., Kociolek J.P. (eds.), *Freshwater Algae of North America*. Boston: Academic Press, 2015, 873–920.
 27. Lebet K, Kritzbeg ES, Rengefors K. Population genetic structure of a microalgal species under expansion. *PLoS One* 2013;**8**:e82510. <https://doi.org/10.1371/journal.pone.0082510>
 28. Sassenhagen I, Sefbom J, Sall T et al. Freshwater protists do not go with the flow: population structure in *Gonyostomum semen* independent of connectivity among lakes. *Environ Microbiol* 2015;**17**:5063–72. <https://doi.org/10.1111/1462-2920.12987>
 29. Johansson KSL, Lüthig K, Klaminder J et al. Development of a quantitative PCR method to explore the historical occurrence of a nuisance microalga under expansion. *Harmful Algae* 2016;**56**: 67–76. <https://doi.org/10.1016/j.hal.2016.04.012>
 30. Folmer O, Black M, Hoeh W et al. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol* 1994;**3**: 294–9.
 31. Camacho C, Coulouris G, Avagyan V et al. BLAST+: architecture and applications. *Bmc Bioinformatics* 2009;**10**:421. <https://doi.org/10.1186/1471-2105-10-421>
 32. Rochette NC, Rivera-Colón AG, Catchen JM. Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol Ecol* 2019;**28**:4737–54. <https://doi.org/10.1111/mec.15253>
 33. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biol* 2019;**20**:257. <https://doi.org/10.1186/s13059-019-1891-0>
 34. Ewels P, Magnusson M, Lundin S et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**:3047–8. <https://doi.org/10.1093/bioinformatics/btw354>
 35. Cerca J, Maurstad MF, Rochette NC et al. Removing the bad apples: a simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms. *Methods Ecol Evol* 2021;**12**:805–17. <https://doi.org/10.1111/2041-210X.13562>
 36. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2020. <https://www.R-project.org/>
 37. Wickham H, Averick M, Bryan J et al. Welcome to the tidyverse. *J Open Source Softw* 2019;**4**:01686. <https://doi.org/10.21105/joss.01686>
 38. Kamvar ZN, Tabima JF, Gruenwald NJ. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2014;**2**:e281. <https://doi.org/10.7717/peerj.281>
 39. Pebesma E. Simple features for R: standardized support for spatial vector data. *R J* 2018;**10**:439–46. <https://doi.org/10.32614/RJ-2018-009>
 40. Ritz C, Spiess A-N. qpcR: an R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. *Bioinformatics* 2008;**24**:1549–51. <https://doi.org/10.1093/bioinformatics/btn227>
 41. Malinsky M, Trucchi E, Lawson DJ et al. RADpainter and fineRAD-structure: population inference from RADseq data. *Mol Biol Evol* 2018;**35**:1284–90. <https://doi.org/10.1093/molbev/msy023>
 42. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;**155**: 945–59. <https://doi.org/10.1093/genetics/155.2.945>
 43. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003;**164**:1567–87. <https://doi.org/10.1093/genetics/164.4.1567>
 44. Earl DA, vonHoldt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* 2012;**4**:359–61. <https://doi.org/10.1007/s12686-011-9548-7>
 45. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 2005;**14**:2611–20. <https://doi.org/10.1111/j.1365-294X.2005.02553.x>
 46. Kopelman NM, Mayzel J, Jakobsson M et al. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour* 2015;**15**:1179–91. <https://doi.org/10.1111/1755-0998.12387>
 47. Meirmans PG. Genodive version 3.0: easy-to-use software for the analysis of genetic data of diploids and polyploids. *Mol Ecol Resour* 2020;**20**:1126–31. <https://doi.org/10.1111/1755-0998.13145>
 48. Jombart T, Ahmed I. Adegnet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 2011;**27**:3070–1. <https://doi.org/10.1093/bioinformatics/btr521>
 49. Gollnisch R, Alling T, Stockenreiter M et al. Calcium and pH interaction limits bloom formation and expansion of a nuisance microalga. *Limnol Oceanogr* 2021;**66**:3523–34. <https://doi.org/10.1002/lno.11896>
 50. Dray S, Dufour A-B. The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw* 2007;**22**:v22i04. <https://doi.org/10.18637/jss.v022.i04>
 51. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* 2010;**11**:94. <https://doi.org/10.1186/1471-2156-11-94>
 52. Coleman AW. Biogeography and speciation in the Pandorina/Volvulina (Chlorophyta) superclade. *J Phycol* 2001;**37**:836–51. <https://doi.org/10.1046/j.1529-8817.2001.01043.x>
 53. Weisse T. Distribution and diversity of aquatic protists: an evolutionary and ecological perspective. *Biodivers Conserv* 2008;**17**: 243–59. <https://doi.org/10.1007/s10531-007-9249-4>
 54. Tesson SVM, Weißbach A, Kremp A et al. The potential for dispersal of microalgal resting cysts by migratory birds. *J Phycol* 2018;**54**:518–28. <https://doi.org/10.1111/jpy.12756>
 55. Coleman AW. Are the impacts of events in the earth's history discernable in the current distributions of freshwater algae? *Hydrobiologia* 1996;**336**:137–42. <https://doi.org/10.1007/BF00010827>
 56. Proctor VW, Malone CR, Devlamin.VI. Dispersal of aquatic organisms - visibility of disseminules recovered from intestinal tract of captive killdeer. *Ecology* 1967;**48**:672–6. <https://doi.org/10.2307/1936517>
 57. Sassenhagen I, Sefbom J, Godhe A et al. Germination and colonization success of *Gonyostomum semen* (Raphidophyceae) cysts after dispersal to new habitats. *J Plankton Res* 2015;**37**:857–61. <https://doi.org/10.1093/plankt/fbv067>
 58. Tesson SVM, Šantl-Temkiv T. Ice nucleation activity and Aeolian dispersal success in airborne and aquatic microalgae. *Front Microbiol* 2018;**9**:2681. <https://doi.org/10.3389/fmicb.2018.02681>
 59. Kristiansen J. Dispersal of freshwater algae - a review. *Hydrobiologia* 1996;**336**:151–7. <https://doi.org/10.1007/BF00010829>
 60. Tesson SVM, Skjøth CA, Šantl-Temkiv T et al. Airborne microalgae: insights, opportunities, and challenges. *Appl Environ Microbiol* 2016;**82**:1978–91. <https://doi.org/10.1128/AEM.03333-15>
 61. Hagman CHC, Skjelbred B, Thrane JE et al. Growth responses of the nuisance algae *Gonyostomum semen* (Raphidophyceae)

- to DOC and associated alterations of light quality and quantity. *Aquat Microb Ecol* 2019;**82**:241–51. <https://doi.org/10.3354/ame01894>
62. Drouet F, Cohen A. The morphology of *Gonyostomum semen* from woods hole, Massachusetts. *Biol Bull* 1935;**68**:422–39. <https://doi.org/10.2307/1537563>
 63. Münzner K, Gollnisch R, Koreiviene J et al. High iron requirements for growth in the nuisance alga *Gonyostomum semen* (Raphidophyceae). *J Phycol* 2021;**57**:1309–22. <https://doi.org/10.1111/jpy.13170>
 64. Lebet K, Ostman O, Langenheder S et al. High abundances of the nuisance raphidophyte *Gonyostomum semen* in brown water lakes are associated with high concentrations of iron. *Sci Rep* 2018;**8**:13463. <https://doi.org/10.1038/s41598-018-31892-7>
 65. Kawecki TJ, Ebert D. Conceptual issues in local adaptation. *Ecol Lett* 2004;**7**:1225–41. <https://doi.org/10.1111/j.1461-0248.2004.00684.x>
 66. Lee CR, Mitchell-Olds T. Quantifying effects of environmental and geographical factors on patterns of genetic differentiation. *Mol Ecol* 2011;**20**:4631–42. <https://doi.org/10.1111/j.1365-294X.2011.05310.x>
 67. Rochette NC, Catchen JM. Deriving genotypes from RAD-seq short-read data using stacks. *Nat Protoc* 2017;**12**:2640–59. <https://doi.org/10.1038/nprot.2017.123>
 68. Sabina J, Leamon JH. Bias in whole genome amplification: causes and considerations. In: Kroneis T. (ed.), *Whole Genome Amplification: Methods and Protocols. Methods in Molecular Biology*. New York, NY: Humana Press, 2015, 15–41.