Diploid Genomic Architecture of *Nitzschia Hildebrandi*, An Elite Biomass Production Diatom

Aaron Oliver

University of California, San Diego

Sheila Podell (✉ spodell@ucsd.edu)

University of California, San Diego

Agnieszka Pinowska

Global Algae Innovations (United States)

Jesse C. Traller

Global Algae Innovations (United States)

Sarah R. Smith

J. Craig Venter Institute

Ryan S McClure

Pacific Northwest National Laboratory

Alexander S Beliaev

Pacific Northwest National Laboratory

Pavlo Bohutskyi

Pacific Northwest National Laboratory

Eric A. Hill

Pacific Northwest National Laboratory

Ariel Rabines

J. Craig Venter Institute

Hong Zheng

J. Craig Venter Institute

Lisa Zeigler Allen

J. Craig Venter Institute

Alan Kuo

Joint Genome Institute

Igor V. Grigoriev

Joint Genome Institute

Andrew E. Allen

University of California, San Diego

David Hazlebeck

Global Algae Innovations (United States)

Eric E. Allen

University of California, San Diego

Research Article

Keywords: *Nitzschia hildebrandi*, *Phaeodactylum tricornutum*, *Fragilariopsis cylindrus*, allelic heterozygosity, bioproduct production

Posted Date: April 8th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-396293/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Version of Record: A version of this preprint was published at Scientific Reports on August 2nd, 2021. See the published version at <https://doi.org/10.1038/s41598-021-95106-3>.

Abstract

A near-complete diploid nuclear genome and accompanying circular mitochondrial and chloroplast genomes have been assembled from the elite commercial diatom species *Nitzschia hildebrandi*. The 50 Mbp haploid size of the nuclear genome is nearly double that of model diatom *Phaeodactylum tricornutum*, but 30% smaller than closer relative *Fragilariopsis cylindrus*. Diploid assembly, which was facilitated by low levels of allelic heterozygosity (2.7%), included 14 candidate chromosome pairs composed of long, syntenic contigs, covering 93% of the total assembly. Telomeric ends were capped with an unusual 12-mer, G-rich, degenerate repeat sequence. Predicted proteins were highly enriched in strain-specific marker domains associated with cell-surface adhesion, biofilm formation, and raphe system gliding motility. Expanded species-specific families of carbonic anhydrases suggest potential enhancement of carbon concentration efficiency, and duplicated glycolysis and fatty acid synthesis pathways across cytosolic and organellar compartments may enhance peak metabolic output, contributing to competitive success over other organisms in mixed cultures. The *N. hildebrandi* genome delivers a robust new reference for future functional and transcriptomic studies to illuminate the physiology of benthic pennate diatoms and harness their unique adaptations to support commercial algae biomass and bioproduct production.

Introduction

Nitzschia is a globally distributed genus of pennate diatoms found in benthic and planktonic habitats from both freshwater and marine environments. *Nitzschia* species have been identified as especially promising candidates for production of algae-derived biofuels and bioproducts via large-scale aquaculture, due to their prolific biomass production, high intracellular lipid content, and robust growth characteristics over a wide range of environmental conditions (Abdel-Hamid et al 2013, Hildebrand et al 2012). However, molecular optimization of photosynthetic *Nitzschia* candidate strains for large-scale commercial operations suffers from a lack of genomic sequence data, hindering the effective application of genetic engineering tools and interpretation of experimentally generated gene and protein expression data. Not knowing the extent to which desirable characteristics might be encoded by homozygous or heterozygous alleles also makes it difficult to predict the potential stability of these traits to recombination during sexual reproduction and selective breeding.

Although diatom nuclear genomes are generally obtained from diploid vegetative cells rather than haploid gametes, historical sequencing and assembly technologies often collapse closely related allelic duplicates into single consensus sequences. Aside from well-characterized model organisms like *Thalassiosira pseudonana* (Armbrust et al 2004) and *Phaeodactylum tricornutum* (Rastogi et al 2018), most currently available diatom genomes contain hundreds to thousands of small contigs that cannot be assigned to individual chromosomes (Basu et al 2017, Guillory et al 2018, Mock et al 2017, Osuna-Cruz et al 2020, Tanaka et al 2015). These short fragments can be difficult to distinguish from sequencing and assembly errors, requiring extensive transcriptome data for validation (Paajanen et al 2017) and potentially leading to omission of rarely expressed genes.

Recently developed long read sequencing technologies coupled with modern, diploid-aware assembly programs provide an opportunity to address many of these issues, including enhanced abilities to resolve diploid alleles and low complexity repeat elements (Jayakumar and Sakakibara 2019, Rhoads and Au 2015). Here we present the first photosynthetic genome representative from genus *Nitzschia*. This near-complete diploid nuclear genome and accompanying organelle assemblies provide new insights into the genomic capabilities of pennate diatoms in the context of allelic diversity, and robust new references for future functional and transcriptomic studies.

Materials And Methods

Nucleic acid isolation, library construction, and sequencing and assembly

An axenic *N. hildebrandi* str. GAI-293 culture was obtained from a single colony on a streaked agar plate, then scaled up to a 500 mL liquid culture in an artificial brackish growth medium similar to that previously described (Chagoya et al 2014, Jiang et al 2015). After 36 hours of growth under continuous light, cells were harvested at late exponential phase via centrifugation.

High molecular weight DNA was extracted using CTAB-chloroform:isoamyl alcohol (Porebski et al 1997). A large-insert (10-20 kb) genomic library for PacBio sequencing was constructed with the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences) and sequenced on a single SMRT Cell on the PacBio Sequel platform at the University of California Davis Genome Center DNA Technologies Core.

PacBio sequencing produced 646,344 reads that were cleaned, trimmed, and assembled with CANU v.1.8 (Koren et al 2017), using default program parameters except for genomeSize=175m. This assembly produced contigs ranging in size from 3,477 - 6,574,884 nucleotides, with an N50 of 3.7 MB. Read mapping to assembled contigs, coverage statistics, and predicted circularity were obtained from CANU program output files. Nucleotide compositions for individual contigs (percent G+C) were obtained by custom perl script. Contigs supported by only a single PacBio read (1X coverage) were discarded.

Nuclear contigs were distinguished from those derived from mitochondria and chloroplasts on the basis of higher G+C nucleotide composition, lower coverage depths, and predicted linearity versus circularity. Length weighted average G+C values for nuclear, chloroplast, and mitochondrial contigs were 45.1%, and 33.4%, and 29.3%, respectively, while coverage depths were 59X, 364X, and 222X. Contig assignments to nuclear, chloroplast, or mitochondrial bins were confirmed by BLASTX searches versus Genbank reference sequences for chloroplast and mitochondrial genomes. Randomly selected, mapped reads from circular organelle contigs were also used for re-assembly at a range of lower coverage depths (12-90X) with more stringent CANU program parameters (correctedErrorRate=0.01, minOverlapLength=1000, min_read_length=20000).

Transcriptome data to support gene model predictions was obtained using Illumina PE-150 reads from cultures obtained from six *N. hildebrandi* str. GAI-293 cultures grown under conditions of varying light and nutrient availability (Supplementary Table S3). The cells were collected by centrifugation (2,000 x g for 4 min, 4°C), flash frozen in liquid N₂ and stored at -80°C prior to RNA isolation. Total RNA was collected from each sample using TRIzol followed by a Zymo RNA Clean and concentrate kit. Quality was confirmed with an Agilent Bioanalyzer and RNA was treated with DNase to remove contaminating DNA. RNA sequencing was carried out by GENEWIZ (South Plainfield, NJ) using the Illumina HiSeq platform. Raw Illumina reads were cleaned using Trimmomatic version 0.36 (Bolger et al 2014) with the following program parameters: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10; LEADING:3; TRAILING:10; HEADCROP:15; SLIDINGWINDOW:4:15; MINLEN:120. Cleaned, trimmed transcriptome reads were assembled using Trinity v2.8.4 (Haas et al 2013) using default parameters to produce 452,784 contigs.

Gene models and functional annotation

Organelle genomes were annotated using GeSeq (Tillich et al 2017), supported by reference sequences from *P. multiseriis* (NC_027265.1), *Halamphora coffeaeformis* (NC_037727.1), *P. tricornutum* (NC_016739.1), and *F. cylindrus* (NC_045244.1). Nuclear genome models were obtained using BRAKER2 (Bruna et al 2021), with the *F. cylindrus* genome as a seed training model and the Trinity-assembled *N. hildebrandi* transcriptome contigs as reference sequences for two subsequent iterations. The number of gene model coding sequences beginning and ending with canonical start (ATG) and termination (TAG, TAA, TGA) codons was tallied using a custom perl script. BLASTN searches were used to measure agreement between predicted coding regions and the original assembled transcriptome sequences used as BRAKER2 input.

Nuclear genome completeness was assessed using BUSCO v.4.06 (Seppey et al 2019, Simao et al 2015) under the taxonomic setting "stramenopiles". Functional gene descriptions were assigned to predicted nuclear proteins based on BLAST matches to reference database sequences at an e-value cutoff of 1e-3 and HMM pattern matches above model-specified gathering cutoffs. Selection of product descriptive names was prioritized in the following order: blast matches to COGs, 2014 update (Galperin et al 2015), followed by HMM pattern matches to TIGRFAM v. 15.0 (Haft et al 2013) and PFAM v. 32 (Finn et al 2016). Candidate transporter proteins were identified based on BLAST searches against the Transporter Classification Database (Saier et al 2021) and PFAM v.32 models bearing the following descriptive terms: transport, permease, efflux, export, antiporter, channel, and porin. KEGG metabolic pathway associations were assigned using KoFAM Koala (Aramaki et al 2020) and EnrichM v. 0.2.1 (Boyd et al 2020). Biosynthetic gene cluster searches were performed using AntiSMASH v5 (Blin et al 2019) with both plant (plantSMASH) and bacterial options.

RNA genes were annotated using RNAmmer v. 1.2 (Lagesen et al 2007), Infernal v. 1.1.2 (Kalvari et al 2018) with RFAM.cm database v. 14.1, and tRNAscanSE v 2.07 (Chan and Lowe 2019). Subcellular localizations of nuclear proteins were predicted using Hectar (Gschloessl et al 2008), ASAFind (Gruber et al 2015), MitoFates (Fukasawa et al 2015), SignalP v. 5.0 (Almagro Armenteros et al 2019), SignalP v. 4.1 (Nielsen 2017), SignalP v. 3.0 (Bendtsen et al 2004), and TMHMM v. 2.0 (Krogh et al 2001). Numbers of transposon and repetitive DNA sequences were annotated using all curated models from the DFAM database of repetitive DNA families, release 3.2 (Hubley et al 2016), together with transposase-associated patterns from PFAM v. 32 (Finn et al 2016). Total percentage of genomic nucleic acid residues present as low-complexity repeats were determined using RepeatModeler2 (Flynn et al 2020). Tandemly repeated telomere sequences were identified using BioSerf (Somanathan and Baysdorfer 2018).

Candidate bicarbonate transporters and carbonic anhydrases (CA) were retrieved using PhycoCosm gene catalogs (Grigoriev et al 2021, Saier et al 2021), HMM searches with models PF00955, PF00194, and KOG0382, and BLAST searches against previously annotated CCM components (Hopkinson et al 2016, Jensen et al 2019, Kikutani et al 2016, Matsuda et al 2017, Nakajima et al 2013, Samukawa et al 2014, Tachibana et al 2011, Tsuji et al 2017). Full-length sequence alignments and maximum likelihood trees were constructed for bicarbonate transporter and α CA sequences using the CLC Genomics Workbench v. 11.0, set at bootstrap = 100 and protein substitution model =WAG.

Taxonomic classification

18S rRNA gene trees were created using FastTree (Price et al 2010) from sequence alignments created with MUSCLE v3.8.31 (Edgar 2004). Average amino acid identity (AAI) scores were calculated using the AAI calculator website (Rodriguez-R and Konstantinidis 2016). Horizontal gene transfer candidates were identified using DarkHorse version 2.0 (Podell and Gaasterland 2007, Podell 2019) based on the GenBank nr database, downloaded January 2019. Allelic duplicates, alternative transcript predictions, and low complexity sequences were removed from the 38,601 predicted proteins of the whole diploid assembly using the JGI Eukaryotic Annotation Pipeline (Grigoriev et al 2021), to create a set of 17,968 haploid representatives. These representatives were clustered together with predicted proteins from other sequenced diatom relatives to identify, compare and visualize protein families having two or more members using OrthoVenn (Xu et al 2019) with an e-value cutoff of $1e-5$ and an inflation value of 1.5.

Genome polymorphism, diploidy and allelic variation

GAI-293 nuclear genome heterozygosity was estimated based on 17mer counts obtained using Jellyfish version 2.2.10 (Marcais and Kingsford 2011) from both assembled contigs and raw reads, and visualized by the method of Kajitani (Kajitani et al 2014), as implemented on GenomeScope (Vurture et al 2017). Diploid contig pairs were identified by self-BLASTN nucleotide searches and mapping using the NUCMER module of MUMmer version 3.23 (Kurtz et al 2004).

Results

Genome sequencing and assembly

Nitzschia diatom strain GAI-293, originally collected from a stream near Lihue on the island of Kauai, Hawaii, USA, was isolated in axenic culture by Global Algae Innovations. This strain has been given the species name of *Nitzschiahildebrandi*, to honor the memory of diatom biologist Dr. Mark Hildebrand (1958-2018). At the Global Algae Innovations Kauai algae facility, *N.*

hildebrandi has demonstrated average productivity of $22 \text{ g/m}^2/\text{day}$ on an ash-free dry weight basis over the span of a 2.5 month cultivation period. This species is capable of growth using a CO_2 supply derived from power plant flue gas and using recycled water in large-scale outdoor facilities. Vegetative cells have a typical raphid pennate morphology, with gliding motility and prominent lipid droplets visible under conditions of silica depletion in the media (**Figure 1**).

DNA extracted from actively growing, diploid vegetative cells was used to generate 646,344 PacBio reads, ranging in size from 1,000 to 117,697 nt, with an average length of 31,140 nt. Long reads were assembled using the diploid-aware program CANU, producing 123 ungapped, nuclear contigs plus individual circular contigs for mitochondrial and chloroplast genomes (**Table 1**). Mitochondrial and chloroplast contigs were recognizable by their divergent nucleotide compositions of 29.3% and 33.4% G+C, respectively, versus an average of $45.1\% \pm 0.9\%$ for nuclear sequences.

Taxonomic classification

Precise taxonomic placement of *N. hildebrandi* within diatom family Bacillariaceae is limited by inconsistencies in the types of molecular marker data available for its potential relatives. Based on genome-wide average amino acid identity of predicted proteins (**Figure 2A**), the closest relatives to *N. hildebrandi* with sequenced genomes are *F. cylindrus* (57%, (Mock et al 2017)), *P. multistriata* (55%, (Basu et al 2017)) and *P. tricornutum* (52%, (Rastogi et al 2018)). An 18S rRNA gene tree (**Figure 2B**) suggests *Nitzschia inconspicua*, *Nitzschia amphibia*, *Nitzschia hantzschiana*, and *Denticula kuetzingii*, all at 98.5% nucleotide identity, are more closely related to *N. hildebrandi* than *F. cylindrus* (94.8% nucleotide identity), but no nuclear genomes are available for closely related *Nitzschia* or *Denticula* species, and no 18S rRNA sequences are available for *P. multistriata*.

Evidence for near-complete diploid assembly

Overall quality of the *N. hildebrandi* assembly is higher than most currently available diatom genomes, based on sizes and numbers of contigs, completeness of conserved BUSCO marker genes, and percentage of non-fragmented protein predictions, defined as coding sequences that begin and end with canonical start and stop codons (**Table 1**). Six assembled contigs, ranging in size from 2.2-4.1 Mbp, were capped with telomeres of 355-414 nt at both 5' and 3' ends, suggesting complete chromosomes. Exceptional assembly quality was made possible by high-coverage, long-read PacBio read technology and the ability of the CANU assembler to resolve allelic ambiguities (Garg et al 2018, Koren et al 2017, Wang et al 2020).

The 99.7 Mbp size of the nuclear assembly and its 38,601 predicted proteins were larger than expected based on previously published numbers for other diatom species. However, diatom genome size statistics are conventionally reported as haploid values, even when obtained from diploid organisms, making an estimated haploid size of 49.9 Mbp and 17,988 predicted proteins for *N. hildebrandi* unremarkable. Assembly diploidy was first demonstrated by MUMmer dot plots of paired contigs (**Figure 3, Supplementary Figures S1A-S1M**). 42 contigs were resolved into 14 highly syntenic paired alignments, at approximately 94% average nucleotide identity per pair and nearly equal assembly coverage depths between partner contigs (**Supplementary Table S1**). Paired contig alignments account for 93% (92,897,432 nt) of the nuclear assembly. At least eleven of the matched contig sets have few insertions, deletions, or re-arrangements, but it is not clear whether incompleteness of the remaining alignments, covering multiple contigs, are due to biologically relevant sequence differences, assembly errors, or some combination of these factors.

Additional evidence supporting diploid completeness of the assembly was obtained using predicted protein sequences based on BRAKER2 (Bruna et al 2021) gene models. Not only were 100% of the conserved BUSCO markers from Stramenopiles present in predicted *N. hildebrandi* gene models, but 92% of these markers were also present in a second, unfragmented copy, consistent with the 92.8% genome alignment coverage observed in MUMmer matched contig pairs (**Supplementary Table S2**). Duplicated BUSCO gene markers were not observed in any other previously published diatom genome at greater than 1%, except for putatively allopolyploid *Fistulifera solaris* (Tanaka et al 2015).

Diploid assembly and gene model correctness were further validated by comparing coding sequences for all 38,601 predicted proteins with 452,784 transcriptome sequences that were generated from growth under a variety of environmental conditions and *de novo* assembled (**Supplementary Table S3**). 98.8% of the coding sequences matched assembled transcriptome sequences at an average of $98.5\% \pm 2.5\%$ nt identity (median 99.4%) (**Supplementary Figure S2**). These results are consistent with expression of both diploid alleles, supporting previous reports suggesting the absence of genome-wide silencing mechanisms in diatoms (De Riso et al 2009, Hildebrand et al 2012, Hogue et al 2021).

To determine the level of genomic variability between diploid alleles, predicted coding sequences and corresponding translations were compared to each other to identify matched pairs (**Supplementary Figure S3**). Average nucleotide sequence identity for paired coding sequences was $97.3 \pm 2.9\%$. 98% of the sequences were more than 90% identical to their partners, and 32% were 100% identical. These results are consistent with an independent estimate of 3.5% overall genome heterozygosity, obtained from unassembled raw reads using relative k-mer abundance (**Supplementary Figure S4**).

Non-protein coding genes

The *N. hildebrandi* genome contained multiple copies of large and small subunit rRNAs and tRNAs for all 20 standard amino acids plus tRNA-SeC, potentially enabling the incorporation of selenocysteine. Paired allelic copies of other well-known RNA genes were also present, including a THI element TPP riboswitch (thiamine biosynthesis regulator), a 5' *ureB* small RNA (urease regulator), U1 and U6 spliceosomal RNAs, and small nucleolar RNAs of types SNORD24 and SNORA11, potentially acting as guides for the methylation of RNA targets and the conversion of uridine to pseudouridine, respectively.

Non-protein coding regions comprised 41% of the *N. hildebrandi* nuclear genome, including 7% identified as low-complexity elements by RepeatModeler (Flynn et al 2020). These relative proportions were similar to levels previously reported in finished diatom chromosomes from *P. tricornutum*, and *T. pseudonana*, but much lower than draft genomes for *P. multistriata*, and *F. cylindrus* (**Table 1**, percent coding sequences). *N. hildebrandi* transposable elements were widely distributed across 44 different contigs (**Figure 4**). The most abundant elements were Class II (cut and paste) sequences belonging to the DDE Insertion and Eukaryotic Terminal Interspersed Repeat Mutator (MULE) subtypes, followed by Class I (copy and paste) elements of the LTR-Gypsy subtype. Class II elements are rare in previously sequenced diatom genomes, and expansion of the MULE subtype is unique to *N. hildebrandi*. Transposable elements were most abundant in contigs tig00044028, tig00000004, tig00000012 and tig00000120, potentially contributing to the lower alignment synteny shown in **Figure 3** for contig pairs 13 and 14.

The most frequently repeated sequence in the *N. hildebrandi* nuclear genome (630 exact matches) was the 12-mer TTAGGGTTGGGG, a G-rich extension of the conserved eukaryotic telomere motif TTAGGG. The second half of this sequence (TTGGGG), originally described in the macronucleus of ciliate *Tetrahymena thermophila* (Eisen et al 2006, Larson et al 1987), is quite different from the consistent 6-mer telomere pattern used by *T. pseudonana* and *P. tricornutum* (TTAGGG), and the more T-rich sequences found in *Chlamydomonas reinhardtii* (TTTTAGGG) and most land plants (TTTAGGG) (Eberhard et al 2019, Fajkus et al 2019, Falcatore et al 2020, Fulneckova et al 2013). Tandem repeats containing multiple copies of *N. hildebrandi* telomere motifs were found exclusively at the ends of contigs, including thirteen of the fourteen matched pair sets shown in **Figure 3**.

N. hildebrandi telomeres differ from those of *T. pseudonana* and *P. tricornutum*, whose 6-mer repeats are all identical, in containing additional degenerate, truncated subsequences of varying lengths interspersed between full-length 12-mers (**Supplementary Figure S5**). Although degenerate *N. hildebrandi* telomere repeat patterns were supported by PacBio read assembly depths of 12 to 31-fold in multiple contigs (**Supplementary Table S4**), excluding all potential artifacts that might have been introduced during sequencing and assembly would require verification using an alternative technology (e.g. Illumina).

Examples of telomere pattern degeneracy are well-known in *Paramecium* (McCormick-Graham et al 1997) and *Saccharomyces* (Shampay and Blackburn 1988), but have not previously been reported in diatoms. Telomere sequence inconsistencies are believed to result from site specific nucleotide mis-incorporation in the telomerase enzyme of *Paramecium*, and slippage with premature termination of reverse transcriptase activity in *Saccharomyces* (Forstemann and Lingner 2001, Prescott and Blackburn 1997). The variable lengths of partial repeats in *N. hildebrandi* telomeres suggest greater consistency with the yeast degeneracy model. Amino acid sequences of the two *N. hildebrandi* telomerase alleles are 95% identical to each other but only 37% identical to their closest diatom match (*P. multistriata*), consistent with potential differences from previously sequenced diatoms in enzyme processivity.

Mitochondrial and chloroplast genomes

The circular *N. hildebrandi* mitochondrial genome is larger than many other pennate diatoms due to a 19,055 nt intragenic spacer (**Figure 5A**), consistent with the observation of intragenic spacers of widely varying sizes in mitochondrial genomes from other diatoms (Crowell et al 2019). In addition to small and large mitochondrial rRNA subunits (rrnS, rrnL), the mitochondrial genome encodes ATP synthase complex components (atp 6 and 9), cytochrome B, cytochrome oxidase subunits (cox 1-3), NADH dehydrogenase subunits (nad 1-7, 9, and 11), large and small ribosomal subunit proteins (rpl 2, 5, 7, 14, and 16, plus rps 2, 3, 4, 7, 8, 10, 11, 13, 14, and 19), a tatC transporter, and 24 tRNA genes. COX1 and ribosomal subunit genes flanking the large intergenic spacer are interrupted by multiple group II introns matching RFAM patterns RF02012 and RF00029, but no sequences matching these motifs were detected within the spacer region itself. Group II introns have previously been reported to occur sporadically within COX1 and ribosomal rRNA genes in the mitochondrial sequences of other diatoms (Crowell et al 2019, Gastineau et al 2019, Oudot-Le Secq and Green 2011, Pogoda et al 2019), consistent with potential historical contributions to organelle genome plasticity (Lambowitz and Zimmerly 2011).

The *N. hildebrandi* chloroplast genome architecture is typical of that found in other diatoms, as well as many plants and algae, with two inverted repeats separating long single copy (LSC) and short single copy (SSC) regions (**Figure 5B**). The inverted repeat sections are each 13,169 nt long and contain three ribosomal RNA genes and two tRNAs. Chloroplast genes encoded in non-duplicated regions include those associated with photosystem I, photosystem II, cytochrome b/f complex, ATP synthase, RuBisCO, thiamin synthesis, tRNAs, RNA polymerase, large and small subunit RNA proteins, TatC and SecA/SecY type transporters, molecular chaperones, antioxidants, and quality control protease FtsH. Five conserved open reading frames of unknown function (ycf39, ycf45, ycf66, ycf89, ycf90) were shared with the *F. cylindrus* chloroplast genome at 56-85% amino acid sequence identity. No group II introns were found in the *N. hildebrandi* chloroplast genome, despite their abundance in the mitochondrial genome and a recent report describing their discovery in the chloroplast of pennate diatom *Toxarium undulatum* (Ruck et al 2017). However, the GAl23 chloroplast genome does contain a Tn3 transposon recombinase related to those typically found in bacteria, but also present in the *Semanivis robusta* chloroplast sequence (Brembu et al 2014) at 76% amino acid identity.

Like other diatoms, many sequences from the *N. hildebrandi* nuclear genome contain targeting signals for intracellular localization to mitochondrial and chloroplast compartments. 1600 mitochondrial target candidates were predicted by MitoFates (Fukasawa et al 2015), and 2172 chloroplast target candidates were predicted by ASAFind (Gruber et al 2015). Some of these genes may have been laterally transferred from ancestral organelles and their symbiont predecessors to the nucleus, a process that is believed to be ongoing even in current times (Timmis et al 2004, Vancaester et al 2020). However, no evidence of recent lateral transfer was detected in *N. hildebrandi* genes predicted to contain organelle-targeting sequences; the closest GenBank sequence matches for these genes were all to nuclear genomes from other diatoms.

Shared, unique and expanded protein families

Predicted protein sequences from *N. hildebrandi* were clustered together with those of four other diatom genomes (*F. cylindrus*, *P. multistriata*, *P. tricornutum*, and *T. pseudonana*) to identify shared versus unique families (**Figure 6**). 3,922 protein families were shared among all five species, a number similar to the 3,742 families previously reported as shared between three *Pseudonitzschia* transcriptomes and the genomes of *P. tricornutum*, and *T. pseudonana* (Di Dato et al 2015). Shared families encoded complete enzyme sets for many well-characterized, highly conserved metabolic pathways, including glycolysis, gluconeogenesis, citric acid cycle, C-4 metabolism, oxidative and reductive pentose phosphate pathways, urea cycle, mevalonate and non-mevalonate pathways, and assimilatory sulfate reduction. Complete, paralogous sets of glycolysis and fatty acid pathway enzymes were predicted to be localized in both cytosolic and chloroplast cellular compartments of *N. hildebrandi*, along with mitochondrial duplication of the second half of the glycolysis pathway, similar to other diatom genomes where compartmental redundancies have been proposed to enable distributed cellular mechanisms for increasing total metabolic capacity (Kroth et al 2008, Smith et al 2012, Traller et al 2016).

Functions for 46% of the shared protein families and 67% of those unique to *N. hildebrandi* were annotated as hypothetical, while many others were described only by general similarity to PFAM domain patterns. While some cases of apparent

divergence within families of indeterminate function might be due to sequencing or assembly errors, transcriptome sequence matches and the existence of closely related allelic copies on syntenic contigs for > 99% of the *N. hildebrandi*-unique loci provides strong evidence supporting their validity.

A large number of *N. hildebrandi* protein sequences contained domains likely to be involved in the processes of cell-surface adhesion, biofilm formation, and raphe system gliding motility. Many of these adhesive domain families were expanded by the inclusion of *N. hildebrandi*-unique members annotated as fibronectins, ankyrins, laminins, outer membrane adhesins, fasciclins, lectins, von Willebrand factors, and villin headpiece domains. Conserved diatom loci annotated as capsular and exopolysaccharide biosynthesis proteins, exostosins, and sulfotransferases may participate in the synthesis of sulfated polymer exudates supporting raphid motility. However, no predicted proteins from *N. hildebrandi* included the adhesion-specific GDPH pattern recently described in *Semanavis robusta* and the transcriptomes of other benthic diatoms (Osuna-Cruz et al 2020).

N. hildebrandi tolerates ambient chemistries that are often high in pH and low in dissolved CO₂, conditions known to limit productivity of aquatic photosynthetic organisms. To understand the molecular basis of this tolerance, we investigated the identities and abundance of several putative components of the biophysical carbon concentrating mechanism (CCM), including bicarbonate transporters and carbonic anhydrases (CAs).

Phylogenetic analysis of bicarbonate transporter sequences from *N. hildebrandi* (**Supplementary Figure S6**) reveals two loci belonging to a metazoan-type clade orthologous to plasma membrane-localized and low CO₂-sensitive bicarbonate transporters from *P. tricornutum*, as well as a single transporter ortholog predicted to be localized to the *P. tricornutum* chloroplast (**Supplementary Figure S6**). Additional bicarbonate transporter family proteins detected using HMM pattern PF00955 belong to a more distant clade, along with other diatom sequences annotated as boron transporters (TC 2.A.31.3, (Saier et al 2021)), leading to uncertainty about a potential role in the CCM. Excluding these, *N. hildebrandi* has slightly fewer (n = 3 per genome) bicarbonate transporters overall than diatoms on average (n = 4 per genome) suggesting there has been little need to expand bicarbonate acquisition capabilities through genetic diversification in this species.

N. hildebrandi encodes 21 CA genes from the α , γ , δ , θ , and LCIP63 subclasses but does not encode genes from the β , or ζ types (**Supplementary Dataset SRS**, (Jensen et al 2019)). LCIP63s and α CAs are particularly enriched relative to other diatoms. *N. hildebrandi* has six putative LCIP63 candidate loci, three of which are chloroplast-localized, as compared to a single copy each in *P. tricornutum* and *T. pseudonana* (Jensen et al 2019), and more than twice the number of α CA loci (n = 11) found in *P. tricornutum* or *T. pseudonana* (Matsuda et al 2017). Phylogenetic analysis shows this is largely the result of major independent expansions of α CAs belonging to two distinct clades (**Figure 7A**). The first expansion (clade GAI-EC1) likely occurred prior to the divergence of *N. hildebrandi*, *P. multiseriata*, and *F. cylindrus*, giving rise to five putatively periplastid space/ER lumen/secretion-targeted genes, while the second expansion, specific to *N. hildebrandi* (GAI-EC2), led to four paralogous loci each possessing a transmembrane domain in the N-terminus. The only other sequence with an N-terminal transmembrane domain that clades with GAI-EC2 is from closely related *F. cylindrus* (Frcy1_244027), making this gene family expansion unique for *N. hildebrandi*. A previously described N-terminal transmembrane domain in diatom CAs TpaCA2, Tp δ CA1, Tp δ CA4 (Samukawa et al 2014, Tachibana et al 2011) has been shown experimentally to anchor Tp δ CA1 in the plasmalemma, supporting the idea that *N. hildebrandi* clade GAI-EC2 α CA proteins may be similarly targeted to the cell surface (Samukawa et al 2014).

Predicted identities and subcellular localizations of *N. hildebrandi* CCM components were used to construct a cellular model for comparison with existing models of *T. pseudonana* and *P. tricornutum* (**Figure 7B**, (Matsuda et al 2017, Tsuji et al 2017)). The most notable differences between *N. hildebrandi* and other diatoms are i) extra copies of putative chloroplast-localized LCIP63 genes, and ii) independent gene family expansions in *N. hildebrandi*, giving rise to extra α CA copies that either decorate the cell surface, surround the chloroplast, or are possibly secreted extracellularly. Operating as a CCM, these extra CAs should capture and concentrate CO₂ towards the pyrenoid, the site of RuBisCO and carbon fixation, in order to sustain high productivity at low concentrations of dissolved CO₂ (Hopkinson et al 2016). Future experimental work is needed to conclusively determine the localization and roles of these CCM components. Nonetheless, these expansions provide good genomic evidence for the presence of a unique and powerful CCM and intracellular pH buffering system relative to what is observed in other diatoms,

which is likely to be one of the features underlying tolerance and high productivity over a spectrum of salinities and dissolved CO₂ levels.

Based on reports of antimicrobial activity in some *Nitzschia* strains (Binea et al 2009, Munir et al 2018, Shannon and Abu-Ghannam 2016), bioinformatic searches were performed to identify potential biosynthetic gene clusters that might encode secondary metabolite pathways. AntiSMASH (Blin et al 2019) detected a number of individual genes similar to those commonly associated with biosynthetic gene clusters, including Type III polyketide synthases, non-ribosomal peptide synthetase-like proteins, cytochrome P450s, and isoprenoid synthesis enzymes. However, none of these protein families were identified as unique to *N. hildebrandi*, and none of these genes were recognized as belonging to multi-gene clusters similar to those of known secondary metabolites. A separate blast search for genes involved in the production of domoic acid in many strains of *Pseudonitzschia* (Brunson et al 2018) produced no matches in *N. hildebrandi*.

Discussion

This work describes a new reference-quality, diploid nuclear genome for *Nitzschia hildebrandi* str. GAI-293. The haploid size of this genome is nearly double that of well-characterized model diatom *P. tricornutum*, but 15-60% smaller than closer pennate relatives *P. multistriata* and *F. cylindrus*. An estimate of 14 chromosome pairs, supported by syntenic alignments of contig pairs with telomer-capped ends, will need to be verified by physical measurements. However, this number is similar to cytological observations of 15-17 chromosomes for other *Nitzschia* species, narrowing down the much wider range (8-130) previously reported for more distantly related pennate taxa (Kociolek and Stoermer 1989, Parks et al 2018).

The expansion of Class II versus Class I transposable elements in *N. hildebrandi* suggests lineage-specific activities promoting enhanced genomic plasticity. This plasticity may be expanded in candidate chromosome pairs 13 and 14, where these elements are particularly abundant. Group II introns like those found in the especially large mitochondrial genome could serve a similar plasticity-enhancing function in this organelle, consistent with widely varying mitochondrial genome sizes reported among other sequenced diatoms (Guillory et al 2018, Pogoda et al 2019). The *N. hildebrandi* chloroplast genome conforms to the more uniform size characteristic of other diatoms, including a well-conserved overall structure similar to that found in many plants.

Some types of information provided by the *N. hildebrandi* diploid genome could not have been obtained from more fragmented, less complete assemblies. Paired coding sequences covering 93% of the genome enabled a direct calculation of allelic heterozygosity at 2.7% (97.3% nucleotide identity), contrasting sharply with the estimate of 25% heterozygosity reported for *F. cylindrus*, although the latter value was calculated from only 5,400 allelic pairs, representing less than 30% of predicted coding sequences (Paajanen et al 2017). Differences between relatively complete and highly fragmented assemblies may also affect coding sequence percentages, which were calculated at 59% for *N. hildebrandi*, 54% for *P. tricornutum*, and 50% for *T. pseudonana*, versus 34% for *P. multistriata* and 27% for *F. cylindrus*. It is not clear if these disparities represent true allelic variability or potential assembly artifacts, but the latter explanation is favored by the lower percentages of complete coding sequences (bounded by canonical start and stop codons) found in diatom genomes with larger numbers of short contigs and lower N50 values (**Table 1**).

Telomer sequences have long been the subject of intense interest with respect to vertebrate cellular aging, but have not yet been systematically studied in diatoms, and only recently investigated in algae (Eberhard et al 2019). The unusually G-rich, degenerate pattern of telomer repeats found in *N. hildebrandi* was surprising in light of the unremarkable, uniform telomeres found in model diatoms *P. tricornutum* and *T. pseudonana*. The altered nucleotide composition of the *N. hildebrandi* repeat pattern should increase secondary structure stability in telomer caps protecting single-stranded chromosome ends, especially at elevated temperatures or under high pH conditions. It will be interesting to learn whether similar degenerate, G-rich motifs occur in other diatoms, and whether they might serve an adaptive function.

1. *hildebrandi* genome sequencing has revealed a number of features that may contribute to successful exploitation of benthic habitats and suitability for commercial biofuel production. Like other diatoms, distributed duplication of glycolysis and fatty acid synthesis among different cellular compartments could provide an energy boost enhancing competitive

success against other organisms in mixed cultures. *N. hildebrandi*'s extensive repertoire of adhesive domain proteins may facilitate attachment to suspended particles during sediment mixing, increasing access to surface light and enhancing photosynthetic efficiency in the shallow, turbid environments where it thrives. Expansion and diversification of carbonic anhydrase paralogs may enable fine-tuning of carbon concentration activities under variable environmental conditions. Resistance to predation and bacterial pathogenesis may be achieved through currently uncharacterized secondary metabolite compounds, produced using genes encoding type III polyketide synthases, ribosomally-produced post-translationally modified proteins, cytochrome P450s, and isoprenoid synthesis enzymes.

The nearly complete set of diploid information delivered by the *Nitzschia hildebrandi* str. GAI-293 genome provides a robust new set of reference material for future functional and transcriptomic studies illuminating the physiology of benthic pennate diatoms and harnessing their unique adaptations to support commercial algae biomass and bioproduct production.

Declarations

Data availability

DNA-sequencing data associated with this study have been deposited in NCBI GenBank under accession number PRJNA675887 and the JGI Algal Genomics Resource PhycoCosm (Grigoriev et al 2021) at <https://mycocosm.jgi.doe.gov/Nithil2/Nithil2.info.html>.

Acknowledgements

We thank the following investigators for permission to use unpublished genomic sequence data in our analyses: Dr. E. Virginia Armbrust (*Pseudonitzschia multiseriata*) and Dr. Alexandra Z. Worden and Dr. Charmaine Yung (*Minidiscus trioculatus*). This study was supported by the United States Department of Energy Office of Energy Efficiency and Renewable Energy (EERE) under the BioEnergy Technologies Office (BETO) Energy Productivity Enhanced Algae and Tool-Kits (PEAK) Program (DE-EE0008121), Genomic Sciences Program (DE-SC00006729 and DE-SC0008593) to A.P., L.Z.A., D.H., E.E.A., R.M., A.B., P.B., E.H. and Dr. Matthew Posewitz, and Enhanced Carbon Uptake from Algal Systems (ECUAS) Program (DE-EE0008639) to A.E.A. and D.H. PNNL is operated for the U.S. DOE by Battelle Memorial Institute under Contract DE-AC05-76RLO 1830. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Author contributions

Genomic data acquisition, analysis, and interpretation were performed by AO, SP, AP, JT, SRS, AK, and HZ. Transcriptome data collection and sequencing were performed by LZA, AR, RM, AB, PB, and EH. IVG, AEA, DH, and EEA contributed to the conception and design of the study. The manuscript was drafted by SP with input from other authors. All authors read and approved the final manuscript.

References

1. Abdel-Hamid, M. I., El-Refaay, D. A., Abdel-Mogib, M. & Azab, Y. A. Studies on biomass and lipid production of seven diatom species with special emphasis on lipid composition of *Nitzschia palea* (Bacillariophyceae) as reliable biodiesel feedstock. *Algal. Stud.* **143**, 65–87 (2013).
2. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol.* **37**, 420–423 (2019).
3. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics.* **36**, 2251–2252 (2020).
4. Armbrust, E. V. *et al.* The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science.* **306**, 79–86 (2004).
5. Basu, S. *et al.* Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. *New Phytol.* **215**, 140–156 (2017).
6. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol.* **340**, 783–795 (2004).
7. Binea, H. K., Kassim, T. I. & Binea, A. K. ANTIBACTERIAL ACTIVITY OF DIATOM NITZSCHIA PALEA (KUETZ.)W. SM. EXTRACT. *Iraqi J Biotech.* **8**, 562–566 (2009).
8. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
9. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* **30**, 2114–2120 (2014).
10. Boyd, B., Woodcroft, B. J. & Tyson, G. W. (2020). Comparative genomics using EnrichM.
11. Brembu, T. *et al.* The chloroplast genome of the diatom *Seminavis robusta*: new features introduced through multiple mechanisms of horizontal gene transfer. *Mar Genomics.* **16**, 17–27 (2014).
12. Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP + and AUGUSTUS supported by a protein database. *NAR Genom Bioinform.* **3**, lqaa108 (2021).
13. Brunson, J. K. *et al.* Biosynthesis of the neurotoxin domoic acid in a bloom-forming diatom. *Science.* **361**, 1356–1358 (2018).
14. Chagoya, J. C. *et al.* Media optimization and lipid formation of two native diatoms for cultivation in the Southwest Texas desert. *Journal of Applied Phycology.* **26**, 2075–2085 (2014).
15. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol Biol.* **1962**, 1–14 (2019).
16. Crowell, R. M., Nienow, J. A. & Cahoon, A. B. The complete chloroplast and mitochondrial genomes of the diatom *Nitzschia palea* (Bacillariophyceae) demonstrate high sequence similarity to the endosymbiont organelles of the dinotom *Durinskia baltica*. *J Phycol.* **55**, 352–364 (2019).
17. De Riso, V. *et al.* Gene silencing in the marine diatom *Phaeodactylum tricornutum*. *Nucleic Acids Res.* **37**, e96 (2009).
18. Di Dato, V. *et al.* Transcriptome sequencing of three Pseudo-nitzschia species reveals comparable gene sets and the presence of Nitric Oxide Synthase genes in diatoms. *Sci Rep.* **5**, 12329 (2015).
19. Eberhard, S. *et al.* (2019). Molecular characterization of *Chlamydomonas reinhardtii* telomeres and telomerase mutants. *Life Sci Alliance* **2**.
20. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
21. Eisen, J. A. *et al.* Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.* **4**, e286 (2006).
22. Fajkus, P. *et al.* Telomerase RNAs in land plants. *Nucleic Acids Res.* **47**, 9842–9856 (2019).

23. Falciatore, A., Jaubert, M., Bouly, J. P., Bailleul, B. & Mock, T. Diatom Molecular Research Comes of Age: Model Species for Studying Phytoplankton Biology and Diversity. *Plant Cell*. **32**, 547–572 (2020).
24. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–285 (2016).
25. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. **117**, 9451–9457 (2020).
26. Forstemann, K. & Lingner, J. Molecular basis for telomere repeat divergence in budding yeast. *Mol Cell Biol*. **21**, 7277–7286 (2001).
27. Fukasawa, Y. *et al.* MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Mol Cell Proteomics*. **14**, 1113–1126 (2015).
28. Fulneckova, J. *et al.* A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. *Genome Biol Evol*. **5**, 468–483 (2013).
29. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–269 (2015).
30. Garg, S. *et al.* A graph-based approach to diploid genome assembly. *Bioinformatics*. **34**, i105–i114 (2018).
31. Gastineau, R. *et al.* Complete mitochondrial genome of a rare diatom (Bacillariophyta) *Proschkinia* and its phylogenetic and taxonomic implications. *Mitochondrial DNA Part B-Resources*. **4**, 25–26 (2019).
32. Grigoriev, I. V. *et al.* PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res.* **49**, D1004–D1011 (2021).
33. Gruber, A., Rocap, G., Kroth, P. G., Armbrust, E. V. & Mock, T. Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *Plant J*. **81**, 519–528 (2015).
34. Gschloessl, B., Guermeur, Y. & Cock, J. M. HECTAR: a method to predict subcellular targeting in heterokonts. *BMC Bioinformatics*. **9**, 393 (2008).
35. Guillory, W. X. *et al.* Recurrent Loss, Horizontal Transfer, and the Obscure Origins of Mitochondrial Introns in Diatoms (Bacillariophyta). *Genome Biol Evol*. **10**, 1504–1515 (2018).
36. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. **8**, 1494–1512 (2013).
37. Haft, D. H. *et al.* TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.* **41**, D387–395 (2013).
38. Hildebrand, M., Davis, A. K., Smith, S. R., Traller, J. C. & Abbriano, R. The place of diatoms in the biofuels industry. *Biofuels-Uk*. **3**, 221–240 (2012).
39. Huguin, A., Rastogi, A., Bowler, C. & Tirichine, L. Genome-wide analysis of allele-specific expression of genes in the model diatom *Phaeodactylum tricornutum*. *Sci Rep*. **11**, 2954 (2021).
40. Hopkinson, B. M., Dupont, C. L. & Matsuda, Y. The physiology and genetics of CO₂ concentrating mechanisms in model diatoms. *Curr Opin Plant Biol*. **31**, 51–57 (2016).
41. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–89 (2016).
42. Jayakumar, V. & Sakakibara, Y. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Brief Bioinform*. **20**, 866–876 (2019).
43. Jensen, E. L., Clement, R., Kosta, A., Maberly, S. C. & Gontero, B. A new widespread subclass of carbonic anhydrase in marine phytoplankton. *ISME J*. **13**, 2094–2106 (2019).
44. Jiang, Y. L., Nunez, M., Laverty, K. S. & Quigg, A. Coupled effect of silicate and nickel on the growth and lipid production in the diatom *Nitzschia perspicua*. *Journal of Applied Phycology*. **27**, 1137–1148 (2015).
45. Kajitani, R. *et al.* Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res*. **24**, 1384–1395 (2014).
46. Kalvari, I. *et al.* Non-Coding RNA Analysis Using the Rfam Database. *Curr Protoc Bioinformatics*. **62**, e51 (2018).

47. Kikutani, S. *et al.* Thylakoid luminal theta-carbonic anhydrase critical for growth and photosynthesis in the marine diatom *Phaeodactylum tricornutum*. *Proc Natl Acad Sci U S A*. **113**, 9828–9833 (2016).
48. Kocielek, J. P. & Stoermer, E. F. Chromosome Numbers in Diatoms: A Review. *Diatom Res.* **4**, 47–54 (1989).
49. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
50. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* **305**, 567–580 (2001).
51. Kroth, P. G. *et al.* A model for carbohydrate metabolism in the diatom *Phaeodactylum tricornutum* deduced from comparative whole genome analysis. *PLoS One*. **3**, e1426 (2008).
52. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
53. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
54. Lambowitz, A. M. & Zimmerly, S. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol.* **3**, a003616 (2011).
55. Larson, D. D., Spangler, E. A. & Blackburn, E. H. Dynamics of telomere length variation in *Tetrahymena thermophila*. *Cell*. **50**, 477–483 (1987).
56. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* **27**, 764–770 (2011).
57. Matsuda, Y., Hopkinson, B. M., Nakajima, K., Dupont, C. L. & Tsuji, Y. (2017). Mechanisms of carbon dioxide acquisition and CO₂ sensing in marine diatoms: a gateway to carbon metabolism. *Philos Trans R Soc Lond B Biol Sci* **372**.
58. McCormick-Graham, M., Haynes, W. J. & Romero, D. P. Variable telomeric repeat synthesis in *Paramecium tetraurelia* is consistent with misincorporation by telomerase. *EMBO J.* **16**, 3233–3242 (1997).
59. Mock, T. *et al.* Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature*. **541**, 536–540 (2017).
60. Munir, N., Rafique, M., Altaf, I., Sharif, N. & Naz, S. Antioxidant and Antimicrobial Activities of Extracts from Selected Algal Species. *Bangladesh Journal of Botany.* **47**, 53–61 (2018).
61. Nakajima, K., Tanaka, A. & Matsuda, Y. SLC4 family transporters in a marine diatom directly pump bicarbonate from seawater. *Proc Natl Acad Sci U S A.* **110**, 1767–1772 (2013).
62. Nielsen, H. Predicting Secretory Proteins with SignalP. *Methods Mol Biol.* **1611**, 59–73 (2017).
63. Osuna-Cruz, C. M. *et al.* The *Seminavis robusta* genome provides insights into the evolutionary adaptations of benthic diatoms. *Nat Commun.* **11**, 3320 (2020).
64. Oudot-Le Secq, M. P. & Green, B. R. Complex repeat structures and novel features in the mitochondrial genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. *Gene.* **476**, 20–26 (2011).
65. Paaanen, P. *et al.* Building a locally diploid genome and transcriptome of the diatom *Fragilariopsis cylindrus*. *Sci Data.* **4**, 170149 (2017).
66. Parks, M. B., Nakov, T., Ruck, E. C., Wickett, N. J. & Alverson, A. J. Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). *Am J Bot.* **105**, 330–347 (2018).
67. Podell, S. & Gaasterland, T. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol.* **8**, R16 (2007).
68. Podell, S. (2019). DarkHorse2 GitHub Repository.
69. Pogoda, C. S. *et al.* Comparative analysis of the mitochondrial genomes of six newly sequenced diatoms reveals group II introns in the barcoding region of *cox1*. *Mitochondrial DNA A DNA Mapp Seq Anal.* **30**, 43–51 (2019).
70. Porebski, S., Bailey, L. G. & Baum, B. R. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Molecular Biology Reporter.* **15**, 8–15 (1997).

71. Prescott, J. & Blackburn, E. H. Telomerase RNA mutations in *Saccharomyces cerevisiae* alter telomerase action and reveal nonprocessivity in vivo and in vitro. *Genes Dev.* **11**, 528–540 (1997).
72. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One.* **5**, e9490 (2010).
73. Rastogi, A. *et al.* Integrative analysis of large scale transcriptome data draws a comprehensive landscape of *Phaeodactylum tricornutum* genome and evolutionary origin of diatoms. *Sci Rep.* **8**, 4834 (2018).
74. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics.* **13**, 278–289 (2015).
75. Rodriguez-R, L. M. & Konstantinidis, K. T. (2016).
76. Ruck, E. C., Linard, S. R., Nakov, T., Theriot, E. C. & Alverson, A. J. Hoarding and horizontal transfer led to an expanded gene and intron repertoire in the plastid genome of the diatom, *Toxarium undulatum* (Bacillariophyta). *Curr Genet.* **63**, 499–507 (2017).
77. Saier, M. H. *et al.* The Transporter Classification Database (TCDB): 2021 update. *Nucleic Acids Res.* **49**, D461–D467 (2021).
78. Samukawa, M., Shen, C., Hopkinson, B. M. & Matsuda, Y. Localization of putative carbonic anhydrases in the marine diatom, *Thalassiosira pseudonana*. *Photosynth Res.* **121**, 235–249 (2014).
79. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol.* **1962**, 227–245 (2019).
80. Shampay, J. & Blackburn, E. H. Generation of telomere-length heterogeneity in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A.* **85**, 534–538 (1988).
81. Shannon, E. & Abu-Ghannam, N. (2016). Antibacterial Derivatives of Marine Algae: An Overview of Pharmacological Mechanisms and Applications. *Mar Drugs* **14**.
82. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* **31**, 3210–3212 (2015).
83. Smith, S. R., Abbriano, R. M. & Hildebrand, M. Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways. *Algal Res.* **1**, 2–16 (2012).
84. Somanathan, I. & Baysdorfer, C. A bioinformatics approach to identify telomere sequences. *Biotechniques.* **65**, 20–25 (2018).
85. Tachibana, M. *et al.* Localization of putative carbonic anhydrases in two marine diatoms, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. *Photosynth Res.* **109**, 205–221 (2011).
86. Tanaka, T. *et al.* Oil accumulation by the oleaginous diatom *Fistulifera solaris* as revealed by the genome and transcriptome. *Plant Cell.* **27**, 162–176 (2015).
87. Tillich, M. *et al.* GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11 (2017).
88. Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet.* **5**, 123–135 (2004).
89. Traller, J. C. *et al.* Genome and methylome of the oleaginous diatom *Cyclotella cryptica* reveal genetic flexibility toward a high lipid phenotype. *Biotechnol Biofuels.* **9**, 258 (2016).
90. Tsuji, Y., Nakajima, K. & Matsuda, Y. Molecular aspects of the biophysical CO₂-concentrating mechanism and its regulation in marine diatoms. *J Exp Bot.* **68**, 3763–3772 (2017).
91. Vancaester, E., Depuydt, T., Osuna-Cruz, C. M. & Vandepoele, K. Comprehensive and Functional Analysis of Horizontal Gene Transfer Events in Diatoms. *Mol Biol Evol.* **37**, 3243–3257 (2020).
92. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics.* **33**, 2202–2204 (2017).
93. Wang, W. *et al.* (2020). The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for comparing de novo assemblies. *Gigascience* **9**.
94. Xu, L. *et al.* OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **47**, W52–W58 (2019).

Tables

Table 1. Genomic properties of *Nitzschia hildebrandi* str. GAI-293 and related diatoms. Genome size and number of predicted proteins shown for *N. hildebrandi* are for the entire diploid assembly, with estimated haploid values in parentheses. The *N. hildebrandi* haploid protein number is less than half of the diploid value because it was calculated using historical conventions typically applied to older diatom genomes, excluding allelic duplicates, alternative transcripts, and low complexity regions, as described in Materials and Methods. Additional information on BUSCO completeness scores is presented in **Supplementary Table 2**.

	<i>Nitzschia hildebrandi</i> GAI293	<i>Pseudo-nitzschia multistriata</i> B856	<i>Fragilariopsis cylindrus</i> CCMP1102	<i>Phaeodactylum tricornutum</i> v2.0	<i>Fistulifera solaris</i> JPCC DA0580	<i>Semanivis robusta</i> D6	<i>Thalassiosira pseudonana</i> v3.0
Nuclear genome size (Mbp)	99.7 (49.9)	56.8	80.5	26.1	49.7	125.6	31.3
Num nuclear contigs	123	700	271	33	295	4754	24
Max ctg length	6,574,884	679,566	5,926,375	2,535,400	904,706	318,117	3,042,585
N50	3.7 Mbp	141 Kbp	782 Kbp	423 Kbp	330 Kbp	51 Kbp	1.3 Mbp
GC content (%)	45.1	43.3	39.1	50.6	45.6	48.3	47.8
% coding sequences	59%	34%	28%	54%	61%	77%	50%
Predicted proteins	38,601 (17,968)	12,039	18,111	10,402	20,429	36,254	11,766
BUSCO completeness	100%	86%	95%	94%	97%	99%	94%
% CDS complete proteins (start/stop codons)	100%	99%	87%	84%	100%	98%	80%
Plastid genome (bp)	139,309	nd	123,275	117,369	134,918	150,240	128,813
Mitochondrial genome (bp)	69,563	nd	nd	77,356	39,648	44,018	43,827
Morphology	pennate	pennate	pennate	pennate	pennate	pennate	centric
Genome Sequencing technology	PacBio	Sanger	Sanger	Sanger	454	Illumina + PacBio	Sanger

Figures

Figure 1

Cellular morphology and commercial culture of *N. hildebrandi*. Panels A and B, differential interference contrast microscopy. Panel C, fluorescence micrograph of cells stained with Nile Red showing non-polar lipid droplets (yellow) and chlorophyll autofluorescence (red). Panel D, typical outdoor cultivation environment at the Kauai algae facility of Global Algae Innovations Inc.

Figure 2

Closest taxonomic relatives of *Nitzschia hildebrandi* str. GAI-293. A, Average Amino Acid Identity (AAI) for all predicted nuclear proteins from each genome. B, 18S rRNA gene tree.

Figure 3

MUMmer dot plots of syntenic contigs in *N. hildebrandi* diploid assembly. Enlarged view of pair #1 shows nucleotide positions for contigs along the x and y axes, with colored dots and lines indicating match direction (red = forward, blue = reverse). Full

sized images for all 14 thumbnail views in are presented in Supplementary Figure S1. Assembly properties of paired contigs are described in Supplementary Table S1.

Figure 4

Repetitive element types in *N. hildebrandi* versus other diatom genomes. Class designations refer to copy and paste (Class I) versus cut and paste (Class II) transposon replication mechanisms. Total numbers represent estimated haploid frequencies, so *N. hildebrandi* totals have been divided by two.

Figure 5

N. hildebrandi Organelle genome maps.

Figure 6

Orthologous protein families in *N. hildebrandi* versus other diatom genomes. Allelic duplicates and low-complexity repeat regions were removed from *N. hildebrandi* before clustering, to obtain 17,968 haploid representatives. Numbers indicate protein families with at least two members, excluding "singletons" that have no orthologous or paralogous matches.

Figure 7

Carbonic anhydrase genes supporting carbon concentration mechanisms. (A) Maximum-likelihood phylogeny of putative α-carbonic anhydrases (CAs) from diatom genomes. (B) Models of carbon concentrating mechanisms (CCMs) for *N. hildebrandi* str. GAI-293 compared to *P. tricornutum*, and *T. pseudonana* (after (Matsuda et al 2017)), showing predicted subcellular compartment distributions (cER = chloroplast endoplasmic reticulum, PPS = periplastid space). In (A), basal branches with higher bootstrap support (>30) are thicker and denoted with colored circles corresponding to colors in (B). Bootstrap values are shown only for nodes with >60% support. Sequence identifiers are JGI PhycoCosm accessions (Grigoriev et al 2021), with taxonomic abbreviations as defined in Supplementary Figure 6. Centric diatom species are colored grey, pennate species black, and GAI-293 sequences with the basal branch color to highlight placement within the diatom phylogeny. Both GAI-293 alleles are shown, when available. Targeting information (Supplementary Dataset SRS) is indicated by the following abbreviations: TM, transmembrane domains; SP, signal peptide; CP, chloroplast transit peptide. Asterisks indicate differences from previously proposed CCM models. Chloroplast low confidence predicted genes are assigned to the PPS. Other CA classes (δCA, γCA, ζCA) are not depicted.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [GAI293supplinfo031821.pdf](#)

- [SupplementaryDataset.xlsx](#)