Survivors of the Sea: Investigating the genomics and survival strategies of the diatom *Skeletonema marinoi*

Matthew Pinder

Doctoral thesis

UNIVERSITY OF GOTHENBURG

Department of Marine Sciences

Faculty of Science

2023

Survivors of the Sea: Investigating the genomics and survival strategies of the diatom *Skeletonema marinoi*

ISBN 978-91-8069-399-8 (PRINT)
ISBN 978-91-8069-400-1 (PDF)

Electronic version available at http://hdl.handle.net/2077/78323
Printed by Stema Specialtryck AB, Borås, Sweden 2023

Abstract

Diatoms are an ecologically important group of phytoplankton, responsible for around 20% of global primary production. One of the features contributing to their success is their ability to form resting stages, a response to adverse conditions in which they enter a dormant state and sink to the sediment. While some resting stages may germinate upon the return of favourable conditions and become resuspended, others can become buried in the dark, anoxic sediment for long periods, retaining their viability. Resting stages of some species have been revived in the lab after over a century, yet the precise mechanisms behind this ability to survive so long in this state is not well understood. One such species capable of forming long-lived resting stages is *Skeletonema marinoi*. The draft reference genome of this marine diatom was recently assembled and annotated, and thus my thesis has involved using this genome as a tool to explore the species and investigate its resting stage survival.

In **paper I**, the use of ancient diatom DNA (i.e. DNA obtained from revived diatom resting stages) in research was reviewed, with particular attention to *S. marinoi*. While much insight has already been gained using this resource, developments in several fields have led to exciting prospects for future research.

In **papers II** and **III**, I compared gene expression between vegetative cells of *S. marinoi*, and resting stages induced by exposure to darkness, anoxia, and nitrogen starvation. After re-exposure to nitrate, the resting stages' gene expression was measured using RNA sequencing over the course of six months. In **paper II**, I performed a differential expression analysis to determine which processes were active in this 'dormant' life stage. Much of the resting stages' central metabolism underwent shifts in expression, and genes involved in protein synthesis were upregulated throughout the resting stage. In **paper III**, I examined the RNA sequencing results from the perspective of allelic expression bias, and identified a handful of genes showing significant shifts in such bias between vegetative and resting cells. Of these, I noted several that formed complete pathways related to the formation of diatoms' silica cell walls.

Lastly, in **paper IV**, I developed a bioinformatic tool – Bamboozle – to identify novel, intraspecific genomic barcoding loci, capable of tracking the relative abundance of multiple strains of *S. marinoi* in co-culture over time. This tool was initially applied to 54 strains of *S. marinoi*, revealing loci that enabled subsequent tracking of strains during an artificial evolution experiment. Bamboozle has been

further developed to accommodate haploid organisms, and was used to identify intraspecific barcoding loci in the model green alga *Chlamydomonas reinhardtii*.

The results of this thesis provide insights and highlight further questions regarding diatom resting stages, as well as providing a novel tool for studying these fascinating phytoplankton. Application of additional omics methods to diatom resting stages in future, and testing of Bamboozle on natural diatom populations, should both lead to further understanding of this taxon.

Keywords: Diatoms, *Skeletonema marinoi*, resting stages, bioinformatics, genomics, gene expression, metabarcoding

Populärvetenskaplig sammanfattning

Kiselalger är en ekologiskt viktig grupp av fytoplankton, och är ansvariga för 20% av den globala primärproduktionen. En egenskap som bidrar till deras framgång är förmågan att bilda vilstadier som en respons på ogynnsamma miljöförhållanden. Dessa vilceller sjunker ner till botten varifrån vissa kan återupplivas när villkoren blir gynnsamma igen, medan andra kan bli begravda i det mörka, anoxiska sedimenten under långa perioder. Vilceller från vissa arter har kunnat återuppväckas i lab efter århundraden, men mekanismerna bakom hur denna förmåga att överleva fungerar är ännu så länge okända.

En av arterna med förmåga att bilda långlivade vilceller är *Skeletonema marinoi*. Ett referensgenom har nyligen tagits fram för denna art, och sålunda har många av studierna i min avhandling använt sig av detta verktyg för att utforska artens förmåga att överleva som vilcell.

I **artikel I**, presenteras en genomgång av hur DNA från kiselalger som återuppväckts från vilceller har använts inom forskning, med speciellt fokus på arten *S. marinoi*. Samtidigt som mycket forskning som involverar denna art och dess genom redan har gjorts, har utvecklingen inom fler fält lett till spännande utsikter för framtida forskning.

I **artikel II** och **III**, jämför jag genuttryck mellan vegetativa celler av *S. marinoi*, och vilceller som inducerats genom att exponeras för mörker, anoxi, och kvävesvält. Efter åter-exponering av kväve, mättes vilostadiernas genuttryck med hjälp av RNA-sekvensering vid sex tidpunkter över sex månaders tid. I **artikel II**, utförde jag en undersökning av differentiellt genuttryck för att bestämma vilka processer som var aktiva i de "vilande" cellerna. Mycket av vilcellernas metabolism visade en förändring i uttryck, och även gener involverade i proteinsyntes var uppreglerade. I **artikel III**, undersökte jag RNA-sekvensresultaten på allel-nivå, och identifierade en handfull gener som visade signifikanta skillnader i uttryck mellan vegetativa och vilande celler. Av dessa, noterar jag att flera av generna är inblandade i bildandet av cellernas intrikata väggar av kiseldioxid, kallad frustule.

Slutligen, i **artikel IV**, utvecklade jag ett bioinformatiskt verktyg – Bamboozle – för att ta fram nya, intraspecifika genomiska barkodningsmarkörer, och visar att dessa kan användas för att kvantitativt analysera olika genetiska stammar av *S. marinoi* som odlas tillsammans, men skiftar i abundans över tid. Detta verktyg användes för analys av 54 stammar av *S. marinoi*, som ingick i ett artificiellt evolutions-experiment. Bamboozle har även utökats med funktionalitet för att

kunna analysera haploida organismer, och användes för att identifiera intraspecifika barkodningsmarkörer för grönalgen *Chlamydomonas reinhardtii*.

Resultatet av denna avhandling ger insikt och lyfter fram ytterligare frågor om kiselalgers vilceller, samt tillhandahåller nya verktyg för att studera dessa fascinerande fytoplankton. Användningen av ytterligare omics-metoder på kiselalgers vilceller i framtiden, och av Bamboozle på naturliga kiselalgpopulationer, bör båda leda till ytterligare förståelse för denna fascinerande organismgrupp.

Table of Contents

Abstract ..... **II**
Populärvetenskaplig sammanfattning ..... **IV**
Acknowledgements **1**
List of papers **2**
My contribution 2
Other publications not in this thesis 3
Background **4**
Diatom genomics 4
Diatom resting stages 6
Skeletonema marinoi 8
Allele-specific expression 8
DNA barcoding 10
Methods employed in thesis **12**
Data used 12
Reference genome – functional annotation 12
Differential expression analysis 13
Variant calling 14
Allele-specific expression analysis 15
Bamboozle 16
Summary of papers **18**
Paper I 18
Paper II 19
Paper III 20
Paper IV 21
Conclusion and future directions **23**
Funding **24**
References **25**

Acknowledgements

Science is a field bolstered by cooperation and collaboration, and as such I owe a great deal of thanks to the people who have played a part in my journey thus far.

To **Mats Töpel**, supervisor of both my Master's project and the PhD thesis presented here, whose patience and positivity have helped guide me through the world of bioinformatics.

To **Adrian Clarke**, my co-supervisor, in whose lab the *Skeletonema marinoi* microbiome that formed the basis of my Master's project was studied.

To **Helle Ploug**, my examiner, for helping me to navigate the necessary paperwork and study plans.

To **Tomas Larsson**, first as a colleague at GU, and later as my mentor through NBIS's Bioinformatics Advisory Program.

To all of my instructors on the Master's Programme for Genomics and Systems Biology, through which I decided that bioinformatics was indeed the path I wanted to pursue. In particular I would like to thank **Magnus Alm Rosenblad**, the coordinator of the programme's Bioinformatics course.

To **Anna Godhe**, the original PI for the Survivors of the Sea project that formed the core of my PhD studies.

To **Rickard Stenow**, whose own work on diatom resting stages has led to many fruitful discussions.

To **Björn Andersson**, whose work on diatoms' response to metal pollution led to the creation of Bamboozle.

To **Olga Kourtchenko** and **Elizabeth Robertson**, for their vital roles in the Survivors of the Sea project.

To my parents, **Lorraine** and **Ian**, and my brother **Brad**, for their constant encouragement.

And to my partner **Maria**, for the emotional support that kept me sane through both a PhD and a pandemic.

List of papers

I: Pinder MIM, Töpel M (2022) Ancient Diatom DNA. In: Falciatore A, Mock T (eds) The Molecular Life of Diatoms. Springer, Cham. https://doi.org/10.1007/978-3-030-92499-7_4

II: Pinder MIM, Kourtchenko O, Robertson EK, Stenow R, Larsson T, Maumus F, Johansson ON, Osuna-Cruz CM, Vandepoele K, Brüchert V, Clarke AK, Ploug H, Godhe A, Töpel M. Survivors of the Sea: Transcriptome response of *Skeletonema marinoi* to long-term dormancy. *Manuscript in preparation for Nature Communications.*

III: Pinder MIM, Töpel M. Shifts in allele-specific expression in resting cells of the marine diatom *Skeletonema marinoi*. *Manuscript in preparation for Scientific Reports.*

IV: Pinder MIM*, Andersson B*, Rengefors K, Blossom H, Svensson M, Töpel M. Bamboozle: A bioinformatic tool for identification and quantification of intraspecific barcodes. *Submitted to Molecular Ecology Resources; also available at bioRxiv* (https://www.biorxiv.org/content/10.1101/2023.03.16.532925v2).

* shared first authorship

My contribution

I: Researched and wrote the chapter, alongside MT.
II: Performed the differential expression analysis; interpreted the results alongside coauthors; polished the reference genome, alongside MT; performed functional annotation and manual curation of the gene models; responsible for the first draft of the manuscript, figures and tables; revised the manuscript alongside coauthors.
III: Designed the experiment; performed the analysis; interpreted the results, alongside MT; responsible for the first draft of the manuscript; revised the manuscript alongside MT.
IV: Developed the Bamboozle program; performed downstream analysis alongside BA; drafted the manuscript alongside BA; revised the manuscript alongside coauthors.

Other publications not in this thesis

Work on the following papers was also performed during my PhD.

Peer-reviewed

Cheregi O, **Pinder MIM**, Shaikh KM, Andersson MX, Engelbrektsson J, Strömberg N, Ekendahl S, Kourtchenko O, Godhe A, Töpel M, Spetea C (2023) Transcriptome analysis reveals insights into adaptive responses of two marine microalgae species to Nordic seasons. Algal Research 103222. https://doi.org/10.1016/j.algal.2023.103222

Andersson B, Berglund O, Filipsson H, Kourtchenko O, Godhe A, Johannesson K, Töpel M, **Pinder M**, Hoepfner L, Rengefors K (2023) Strain-specific metabarcoding reveals rapid evolution of copper tolerance in populations of the coastal diatom *Skeletonema marinoi*. *Accepted in Molecular Ecology*. https://doi.org/10.22541/au.168069094.46659870/v1

Pereyra RT, Rafajlović M, De Wit P, **Pinder M**, Kinnby A, Töpel M, Johannesson K (2023) Clones on the run: The genomics of a recently expanded partially clonal species. Molecular Ecology 32:4209–4223. https://doi.org/10.1111/mec.16996

Manuscripts

Bachimanchi H, **Pinder MIM**, Robert C, De Wit P, Havenhand J, Kinnby A, Midtvedt D, Selander E, Volpe G (2023) Deep-learning-powered data analysis in plankton ecology. *Manuscript in preparation for Limnology and Oceanography Letters.*

Pinseel E, Ruck EC, Nakov T, Jonsson PR, Kourtchenko O, Kremp A, **Pinder MIM**, Roberts WR, Sjöqvist C, Töpel M, Godhe A, Hahn MW, Alverson AJ. Local adaptation of a marine diatom is governed by genome-wide changes in diverse metabolic processes. *Manuscript in preparation.*

Other

Töpel M, **Pinder M** (2022) [Reproducible analysis of environmental DNA in national monitoring programs]. Swedish Environmental Protection Agency. ISBN: 978-91-620-7084-7 https://www.naturvardsverket.se/publikationer/7000/978-91-620-7084-7

Background

Diatoms (Bacillariophyceae) are an important class of microalgae, responsible for around 20% of global primary production, and 40% of oceanic primary production (Nelson *et al.*, 1995; Field *et al.*, 1998). These unicellular organisms possess an intricate silica cell wall (frustule), are found throughout the world, and can have a substantial impact on the environment. On the one hand, they are a large component of the base of marine food webs, and are able to sequester vast amounts of carbon when aggregates and resting stages sink to the seafloor and are buried, thus contributing to the biological carbon pump (Tréguer *et al.*, 2018). On the other hand, under conditions of high light and nutrients, diatoms are capable of forming immense blooms, with the potential for toxin production in some species (such as in the genus *Pseudo-nitzschia* [Bates *et al.*, 2018]).

In addition to their environmental importance, diatoms have also been investigated for their potential uses by humans. A distinctive feature of diatoms is their silica frustule, which is of interest in biotechnology for applications such as electronic devices (Jeffryes *et al.*, 2011). The potential for some species to produce high amounts of lipids also makes them a potentially valuable source of biofuel (Hildebrand *et al.*, 2012).

Given all of the above factors, diatoms are a very interesting and relevant group of organisms for study.

Diatom genomics

The first whole diatom genomes to be sequenced were those of *Thalassiosira pseudonana* (Armbrust *et al.*, 2004) and *Phaeodactylum tricornutum* (Bowler *et al.*, 2008), representatives of the centric and pennate diatoms, respectively. In the years since, numerous other diatom genome sequences have become available, with many more in the works, thanks to initiatives such as the 100 Diatom Genomes Project (https://jgi.doe.gov/csp-2021-100-diatom-genomes/). Through such studies, various phenomena of diatom genomes have been investigated.

Diatom genomes have been reported to contain many horizontally-transferred genes from multiple sources, with estimates of between 1-5% of a diatom's genes being derived from other taxa (Van Etten & Bhattacharya, 2020; Vancaester *et al.*, 2020). This has granted diatoms access to additional, unexpected capabilities, such as the urea cycle, which aids in intracellular nitrogen recycling and recovery from nitrogen limitation (Allen *et al.*, 2011), and the production of antifreeze proteins in cold-adapted diatoms (Sorhannus, 2011).

Diatom genomes can be of a variety of ploidy levels, in addition to the diploidy found in the two genomes initially sequenced. The oleaginous species *Fistulifera solaris* was identified as an allodiploid (i.e. it contains two subgenomes, each derived from a different parent in a hybridisation event) (Tanaka *et al.*, 2015), which may contribute to the species' ability to accumulate lipids (Nomaguchi *et al.*, 2018). Triploidy (three copies of each chromosome) and aneuploidy (differing numbers of chromosomes) have also been reported in several species, including *Fragilariopsis cylindrus* (Mock *et al.*, 2022; suspected triploidy) and *F. solaris* (Maeda *et al.*, 2022; suspected aneuploidy). In *Skeletonema marinoi*, depending on the strain, diploidy and triploidy (**paper IV**), as well as aneuploidy (**paper II**), have been observed.

Despite the low heterozygosity reported in some diatom species (e.g. 0.75% of the genome of *T. pseudonana* contains polymorphisms [Armbrust *et al.*, 2004]), other species display a high degree of haplotype divergence (i.e. many variations between alleles). Sequencing of the cold-adapted species *F. cylindrus* revealed that nearly a quarter of its genes had highly divergent alleles, which displayed differential expression under environmental conditions relevant to life in the Southern Ocean, such as darkness and freezing temperatures (Mock *et al.*, 2017). Such variability also extends to the intraspecific level, with different strains of the same species possessing high levels of variation (Godhe & Rynearson, 2017). This results in substantial standing genetic diversity upon which selection can act in response to challenges, contributing to diatoms' success. I make use of this high intraspecific diversity in **paper IV** to identify novel DNA barcodes for identification of diatom strains in co-culture.

In addition to whole genome sequencing of diatoms, sequencing of shorter, traditional DNA barcodes (the principles of which are described in the section *DNA barcoding* below) has also proven useful. For example, given the highly similar morphologies of some diatom species that confound identification via microscopy (e.g. *Skeletonema marinoi* and *Skeletonema dohrnii* [Ellegaard *et al.*, 2008]), the use of such DNA sequences can aid in distinguishing between them.

In **paper II** of this thesis, the draft genome sequence of *Skeletonema marinoi* strain R05AC is presented. Furthermore, the intraspecific diversity of diatoms is harnessed in **paper IV** to identify sequences enabling quantification of a mixture of strains, using the software tool Bamboozle.

Diatom resting stages

Diatoms live in a shifting environment, subject to frequent changes in conditions such as light, temperature, and nutrient availability. One way in which diatoms are able to survive such shifts is through the formation of *resting stages*, a dormant life stage characterised by a severely-reduced metabolism, greater abundance of storage vesicles, and altered plastid structure (McQuoid & Hobson, 1996). These resting stages then sink through the water column into the sediment, from which they have the potential to germinate and become resuspended once favourable conditions return. There are two commonly-accepted types of resting stage: resting *cells*, and resting *spores*. Resting cells are morphologically similar to their vegetative counterparts, whereas resting spores are much more heavily silicified. A third type of resting stage – a *winter stage* – is also proposed for some Antarctic diatoms, but differs from resting cells and resting spores in being able to divide, and lacking storage bodies (Fryxell & Prasad, 1990).

Diatoms can remain as resting stages for varying periods of time. On one hand, resting stages allow diatoms to survive short-term nutrient depletion, before becoming resuspended and potentially seeding a new bloom (McQuoid & Godhe, 2004). On the other hand, if the cells become buried in the sediment, revival is possible after significant periods, which helps to anchor a population to a given location (Sundqvist *et al.*, 2018). Multiple studies have reported revival of diatom resting stages in the lab after decades, centuries (Härnström *et al.*, 2011), and even millennia (Sanyal *et al.*, 2021), after isolation of these resting stages from sediment cores.

While the longevity of diatom resting stages is well-documented, the precise mechanisms used by diatoms to survive for such long periods, particularly under darkness and anoxia (where they cannot photosynthesis or use oxygen for respiration), are poorly understood. Some species of diatoms are capable of luxury uptake of nitrate, and can store it in their vacuole at concentrations far exceeding those of the surrounding environment (Lomas & Glibert, 2000). In the case of *S. marinoi*, for example, the intracellular concentration of nitrate can exceed the extracellular concentration by 2-3 orders of magnitude (Kamp *et al.*, 2016). Some species can then use this stored nitrate to perform dissimilatory nitrate reduction to ammonium (DNRA), a form of anaerobic respiration where nitrate is used as an electron acceptor in place of oxygen (e.g. Kamp *et al.*, 2011; 2013; 2016; Merz *et al.*, 2021). However, while this may help fuel the transition of cells to the resting

stage and provide energy for short periods thereafter, upon transition to darkness and anoxia these nitrate stores are depleted on a timescale of days.

In addition to DNRA, other processes have also been observed or proposed to explain diatom survival under such conditions. Dark fermentation has been demonstrated under darkness and anoxia in the genus *Fragilariopsis* (Bourke *et al.*, 2017), and further fermentation pathways have been detected in *T. pseudonana* (Gain *et al.*, 2021). Resting stages of *Skeletonema marinoi* have also been discovered to be capable of uptake and assimilation of carbon and nitrogen from organic compounds (urea and acetate) (Stenow, 2023). As noted in the previous section, given diatoms' capacity for horizontal gene transfer, yet other unexpected metabolic capacities could be present. In anoxic sediments, there are a number of anaerobic pathways employed by various microorganisms – reduction of nitrate, manganese, iron, or sulfate, or formation of methane (Middelburg, 2019) – that might be relevant to diatoms.

In addition to their ecological relevance, diatoms taken from sediment cores have been used to answer research questions across various fields. While previous studies have made use of dead diatoms, analysing changes in diatom assemblages over time by looking at their frustules (e.g. Emeis *et al.*, 2003), more recently the DNA of living diatoms, revived from resting stages, has been harnessed. In **paper I**, an overview of the types of studies undertaken using DNA obtained from revived resting stages is presented, along with speculation regarding the types of studies that could be embarked upon given advances in sequencing and bioinformatic techniques. In **papers II** and **III**, resting stages are investigated in the context of their gene expression. In **paper II**, I examine the differential gene expression between vegetative and resting cells of *S. marinoi*, while in **paper III**, I compare the occurrence of allele-specific expression between these two states.

Skeletonema marinoi

Skeletonema marinoi is a pelagic, chain-forming centric diatom species in the order Thalassiosirales. While initially believed to be part of the species *Skeletonema costatum* (Sarno *et al.*, 2005), it was later classified as its own species, a part of the *S. marinoi-dohrnii* species complex (Ellegaard *et al.*, 2008; Yamada *et al.*, 2010). This species has been found in temperate waters around the world, is a major component of spring blooms in the Baltic Sea, and is easy to maintain in culture, making it an attractive model species (Sarno *et al.*, 2005; Saravanan & Godhe, 2010). Another appealing aspect of *S. marinoi* as a model species is its ability to form long-lived resting stages. Resting cells of this species have reportedly been revived from sediments dating back at least one hundred years (Härnström *et al.*, 2011).

Of particular relevance to this thesis is the *S. marinoi* strain R05AC. This strain was revived from a resting cell isolated from top layer sediments in Öresund, Sweden (14m depth) in May 2010. The strain has been used in numerous studies, looking at the aforementioned resting stages (Stenow *et al.*, 2020; **papers II** and **III** of this thesis), as well as mutagenesis (Johansson *et al.*, 2019a), metal tolerance (Andersson *et al.*, 2020), and its microbiome (Johansson *et al.*, 2019b). A draft reference genome has been generated from this strain, and is presented in **paper II** of this thesis.

Skeletonema marinoi is the main subject of all four papers presented in this thesis. In **paper I**, it is a major example of a species whose revived resting stages have been used in research; in **papers II**, **III**, and **IV**, I make use of its aforementioned draft genome, as well as both DNA and RNA sequencing data derived from multiple strains.

Allele-specific expression

As non-haploid organisms contain multiple alleles of a given gene, it is possible that these alleles may be expressed in unequal quantities. This phenomenon – *allele-specific expression* – allows these organisms to fine-tune their gene expression, thus leading to phenotypic diversity. Examples in multicellular organisms include diversity in production of immune system and olfactory genes, giving a wide range of antigen and odorant receptors (e.g. Shykind, 2005; Cedar & Bergman, 2008). A number of mechanisms can be responsible for such expression bias. Genomic imprinting, for example, is a well-known phenomenon in mammals that can potentially silence one parental copy of a gene through epigenetic modification of

one allele (Tucci *et al.*, 2019). In terms of nucleotide sequence, genetic variants in non-coding regions of a gene can also result in expression bias. This can occur through, for example, sequence-specific epigenetic modifications, or through altered ability of transcription factors to bind to the DNA (Cleary & Seoighe, 2021).

As well as studies in multicellular organisms, the phenomenon of allele-specific expression has also been studied in a number of model diatoms. RNA sequencing was performed on the cold-adapted diatom *F. cylindrus* under five sets of conditions relevant to life in the Southern Ocean, such as low iron and prolonged darkness (Mock *et al.*, 2017). Of the 6,071 genes noted as being divergent (i.e. being different enough that they were placed on different scaffolds during genome assembly), around 45% (2,730 genes) displayed over fourfold bias in expression between the two alleles in at least one experiment (versus growth under optimal conditions). In the allodiploid *F. solaris*, around 61% of homoeologous gene pairs displayed biassed expression towards one of the two parental subgenomes, with around half showing a bias towards each subgenome, and with most of these showing the same direction of bias across all three timepoints of the experiment (before, during, and after oil accumulation) (Nomaguchi *et al.*, 2018). At the pathway level, the authors noted some patterns to this bias – chloroplast- and ER-localised fatty acid (FA) synthesis genes generally showed a bias towards one subgenome at most timepoints, while mitochondria- and peroxisome-localised FA degradation and elongation genes tended to be biassed towards the other subgenome. In the diploid model diatom *P. tricornutum*, under standard culture conditions around 1% of genes showed a strong bias in allelic expression, with around 22% showing a more moderate bias, and around 11% showing little to no bias (Hoguin *et al.*, 2021). Those displaying a strong bias showed some enrichment in lipid metabolism and hydrolase activity, while those with moderate bias were enriched for genes related to catabolism and protein transport.

All of the aforementioned diatom examples involve pennates, and to the best of my knowledge, a similar study has not yet been carried out in a centric species. Thus, in **paper III**, I sought to investigate allele-specific expression in the centric *S. marinoi*, to both determine its extent and to examine shifts in the genes showing allelic bias in the resting stages.

DNA barcoding

DNA sequencing can enable identification of organisms to varying taxonomic levels. By sequencing specific loci – DNA barcodes – one can identify a sample potentially to species level, as genes and other genomic regions have been identified whose sequence is species-specific. Applied to complex samples, a technique called metabarcoding can be used, whereby a locus is sequenced in order to identify multiple species within the sample. This technique is useful in, for example, analysis of environmental samples, if one wishes to assess the presence of many species in a single sequencing run. Common barcoding loci include genes for ribosomal RNAs, chloroplastic *rbcL*, and mitochondrial *cox1* genes, depending on the taxa being studied (Hebert *et al.*, 2003; Hollingsworth *et al.*, 2009; Guo *et al.*, 2015).

To be useful as a DNA barcode, a locus must fulfil a number of criteria (Kress & Erickson, 2008):

1. The locus must be short enough to allow PCR amplification and subsequent sequencing.
2. The sequence must contain sufficient interspecific variability to differentiate between taxa in a sample.
3. Flanking regions should be conserved between taxa to allow for binding of broad-specificity primers.

While many of the most commonly-used barcodes are effective at differentiating between *species*, they are sufficiently conserved within that species to render them useless for differentiating between *strains*. Short, repeated sequences – microsatellites – can be used for this purpose, with different strains containing different copy numbers, but as high-throughput sequencing methods are less able to handle such highly-repetitive sequences, such an approach cannot easily be applied to a mixed sample. Thus, in the context of a study where we wished to follow the evolution of a mixed-strain culture of microalgae over time in response to a stressor, a different locus was required for the quantitative analysis of strains.

In **paper IV**, we used a genome-scanning technique to identify sequences that fulfil the necessary criteria for barcodes, at the intraspecific level, and tested one of these novel barcodes in *S. marinoi*.

Knowledge gaps and thesis aims

The overall aim of this thesis has been exploration of the recently-sequenced draft genome of *Skeletonema marinoi* strain R05AC, and the investigation of its resting stage gene expression, in order to shed light on its survival strategies under long-term darkness and anoxia.

Paper I is a review of the use of DNA derived from resting stages – how it has been used in the past, best practices for its use in the present, and the potential for studies in the future.

While diatom resting stages have been used in myriad studies, the precise mechanisms behind diatoms' ability to survive long periods as resting stages are poorly understood. While gene expression studies have been performed previously on resting stages, many of these were focused on the short period after induction of the resting stages (i.e. focusing on the *transition* to resting stages, rather than *long-term* survival). To the best of my knowledge, the only comparable study looking at differential gene expression on a similar timescale was performed in *Fragilariopsis cylindrus*, under continuous darkness but not anoxia, reflective of winter darkness in the Southern Ocean rather than resting stage burial in sediment (Joli *et al.*, 2023). Thus, in **paper II** I sought to investigate this question of differential expression in long-term resting cells of *S. marinoi*, to determine which processes are active in these 'dormant' diatoms.

As an extension of this, in **paper III** I looked at the shifting trends in allele-specific expression between the vegetative cells and the resting stages. While such biases in allelic expression have been examined in diatoms previously, those studies have been performed in pennates (Mock *et al.*, 2017; Nomaguchi *et al.*, 2018; Hoguin *et al.*, 2021), so I aimed to both perform a similar analysis in a centric diatom, and determine whether shifts in allele-specific expression could grant further insight into resting stage survival.

In **paper IV**, by using the *S. marinoi* reference genome alongside WGS data from additional *S. marinoi* strains, I sought to make use of diatoms' intraspecific variability to identify intraspecific barcodes that would enable quantification of strains in co-culture. The barcodes identified using this novel method were then tested on a population of *S. marinoi* exposed to toxic copper stress.

Methods employed in thesis

Data used

The primary data source for this thesis comprises genomic and transcriptomic data from various strains of *Skeletonema marinoi*.

- R05AC – this strain, revived from top layer sediments isolated from Öresund, Sweden in 2010, was used to generate the draft reference genome that is used throughout this thesis. RNA sequencing data from both vegetative and resting cells of this strain was used in **papers II** and **III**.
- GF0410J – this strain was isolated from Gullmar Fjord, Sweden in 2004, and RNA sequencing data from both vegetative and resting cells of this strain was used in **paper II**.
- In **paper IV**, WGS data from 56 strains of *S. marinoi* was used for developing Bamboozle. These strains were isolated from surface sediments collected from two Baltic Sea inlets – Gåsfjärden, which was exposed to discharges from a local copper mine for several centuries (Söderhielm & Sundblad, 1996), and Gropviken, which was not exposed.

In addition, in order to test the applicability of the Bamboozle pipeline to haploid organisms in **paper IV**, I downloaded publicly-available WGS data from seventeen strains of *Chlamydomonas reinhardtii*. These strains are described in Flowers *et al.* (2015) and Ness *et al.* (2016), and were originally isolated from across North America.

Reference genome – functional annotation

Prior to beginning my PhD, I was involved in the assembly of the *S. marinoi* strain R05AC reference genome. However, while gene prediction had been performed during this process, the functional annotation had not, thus this was included in my project work. Numerous tools exist for the functional annotation of genes, of which I employed several.

- eggNOG mapper version 2.0.1 (Cantalapiedra *et al.*, 2021) – this tool uses a database of orthologs (i.e. genes of common descent that are expected to share a function) to assign functions to gene models. Where Gene Ontology (GO) functional enrichment analysis was performed in **papers II** and **III**, the analysed GO terms were obtained from eggNOG mapper.

- InterProScan version 5.47-82.0 (Jones *et al.*, 2014) – this tool makes use of several different algorithms to search for various features in a given input sequence, which are then used to assign an annotation.
- BLASTp version 2.8.1+ (Camacho *et al.*, 2009) – this tool identifies similar protein sequences by aligning the query to sequences in NCBI's databases.
- Conserved Domain Database (CDD) search (Lu *et al.*, 2020) – this webtool hosted by NCBI uses a BLAST algorithm to search a database for conserved domains within a protein sequence.

Annotations were primarily derived from eggNOG mapper and InterProScan. When there was uncertainty, such as when the tools gave differing annotations, data from BLASTp and CDD was considered in order to give the most appropriate annotation.

Differential expression analysis

In **paper II**, differential expression analysis was used in order to investigate the activity of pathways in the resting stages of *S. marinoi*.

Differential expression analysis involves extraction and sequencing of mRNA from an organism under different conditions in order to compare how its gene expression changes. After the RNA was sequenced (in this case, using Illumina's NovaSeq6000 platform to produce 2x151 bp paired end reads) and quality control performed to remove low-quality sequences, I mapped the reads to the reference genome using the program HISAT2 version 2.1.0 (Kim *et al.*, 2019). This is a 'splice-aware' aligner – as genes can contain introns which would constitute a potentially long gap in the genome sequence, it is able to align RNA reads that would otherwise be heavily penalised by a splice-unaware aligner and thus map either incorrectly or not at all.

Using this file of mapped reads, the `featureCounts` function of Rsubread version 2.0.1 (Liao *et al.*, 2014) was used to quantify how many reads mapped to the coordinates of each gene in the *S. marinoi* genome. I then used DESeq2 version 1.26.0 (Love *et al.*, 2014) to analyse these values and obtain the level of differential expression between conditions.

While it is possible to perform a differential expression analysis by assembling the experimental RNA sequencing (RNAseq) reads into a *de novo* transcriptome assembly, and then mapping the reads back to this assembly to obtain read counts, where a reference genome sequence is available the reads can be mapped to this

instead. *De novo* transcriptome assembly is a non-trivial process, particularly when assembled from short reads, such as Illumina. Issues can arise when, for example, genes with low expression are assembled incorrectly, or genes adjacent on the chromosome are merged inadvertently (Grabherr *et al.*, 2011). These problems can be somewhat addressed through the use of an annotated reference genome. In this way, additional information can also be obtained, such as a given gene's genomic context, and the presence of introns.

An important caveat to bear in mind when performing differential expression analysis, however, is that gene expression is not a 1:1 proxy for protein levels. Post-transcriptional regulation can lead to protein levels below those suggested by the corresponding transcript levels, for example if mRNAs are degraded before they can be translated (Mata *et al.*, 2005). In the context of *S. marinoi*, Zhang *et al.* (2021) performed both transcriptomic and proteomic analysis of this species under conditions of increased CO_2 concentrations, and found that only 60.8% of protein levels detected in their study correlated with the transcriptomics data. Thus, where possible, multiple omics approaches should be harnessed to obtain a more complete picture.

Variant calling

For **papers III** and **IV**, variant calling was required to differentiate between diploid alleles, as the *S. marinoi* R05AC reference genome represents only a single haplotype. While the initial variant calling in **paper IV** was performed using BCFtools version 1.10.2 (Danecek *et al.*, 2021), further testing was done with the Genome Analysis Toolkit (GATK) version 4.1.8.0 (Van der Auwera & O'Connor, 2020), which was found to produce better results in this case, and so was the variant calling pipeline of choice for both **papers III** and **IV**.

The pipeline functions by aligning reads to a reference, and then analysing where differences between the read sequence and the reference sequence occur. If these differences are supported by enough good-quality reads (i.e. they represent true variation and not sequencing errors), they are reported as variants.

In **papers III** and **IV**, genomic reads from *S. marinoi* R05AC were aligned to the reference genome using the (splice-unaware) aligner Bowtie2 version 2.3.4.3 (Langmead & Salzburg, 2012) in order to, respectively, assess whether genes were in non-diploid regions (due to suspected aneuploidy in the genome), and to distinguish between alleles. In **paper III**, variant calling was also applied to RNAseq reads, to determine the ratios between expression of two alleles and thus

identify biases in expression between them, as detailed in the section *Allele-specific expression analysis* below. As recommended by the GATK best practice guidelines for variant calling of RNAseq data, the initial alignment was performed using STAR version 2.7.10b (Dobin *et al.*, 2013).

Allele-specific expression analysis

In their study on allele-specific expression in *P. tricornutum*, Hoguin *et al.* (2021) used the following formula to estimate allelic bias in their dataset, which I also applied in my study of this phenomenon in *S. marinoi*:

$$AFB\ (\%) or\ AEB\ (\%) = \frac{\sum_{i=1}^{n}\left(\left|\frac{AD(Ref)}{RD} - \frac{AD(Alt)}{RD}\right| \times 100\right)}{N}$$

where AD(Ref) is the allelic depth of the reference allele; AD(Alt) is the allelic depth of the alternate allele; RD is the total read depth at a given position; and N is the total heterozygous positions in the gene.

This formula is used to estimate two types of bias:

- Allelic frequency bias (AFB) represents the deviation from the expected 1:1 ratio of variants at a given position in a diploid organism. Genes with a high AFB (≥ 20%) were not considered when measuring levels of allele-specific expression; as *S. marinoi* strain R05AC is an aneuploid, genes on the duplicated contigs would have a 2:1 ratio of variants, and would have complicated downstream analysis.
- Allelic expression bias (AEB) represents the deviation from the expectation that both alleles of a gene will be expressed equally. This measure was used to classify each gene into one of three categories, as defined by Hoguin *et al.* (2021):
    - Genes with an AEB ≤ 20% (i.e. alleles are expressed in almost equal proportions) are classified as biallelic expressed genes (BAE).
    - Genes with a moderate level of bias (20% < AEB ≤ 60%) are classified as allele-specific expressed genes (ASE).
    - Genes with a high level of bias (AEB > 60%) are classified as monoallelic expressed genes (MAE).

Bamboozle

Paper IV involved development of the Bamboozle pipeline, designed to identify intraspecific barcoding loci in a set of strains within a given species. To achieve this, several popular bioinformatics tools – Bedtools (Quinlan & Hall, 2010), Samtools, and BCFtools (Danecek *et al.*, 2021) – were used. An overview of the Bamboozle pipeline is shown in Figure 1, and is described in detail here.

The input for this Python pipeline consists of three sets of files – a reference genome in FASTA format, read alignment files for each strain to be investigated (BAM format), and the variant-calling files derived from these alignment files (VCF format). If the input organism is a diploid (as is generally the case in *S. marinoi*), some additional files are also required – phased BAM and VCF files, obtained by splitting ('phasing') the two haplotypes of the diploid. In this case, this was achieved using the `phase` function of SAMtools version 1.10 (Danecek *et al.*, 2021).

The main steps in the Bamboozle pipeline are as follows:

1. Regions of the genome where the mapped reads have a coverage greater than double or less than half of the contig median coverage are identified. These can correspond to issues such as large indels or problems in the underlying reference, and thus can complicate the downstream analysis.
2. The VCF files, containing information on allelic variants from each strain, are parsed, and the list of all variants across all strains is saved.
3. Bamboozle steps through the genome (step size of 1), and assesses every window of the user-defined length. If there is at least one variant in a window, and no variants within the flanking regions (length also defined by the user), then that window is saved for step 4. Window and flanking region length are user-definable to account for different sequencing technologies and primer sizes, but the defaults are set to a 500 bp window size and 21 bp flanking regions. The respective values were chosen, because, in **paper IV**, the subsequent amplicon sequencing was to be performed using 2x300 bp Illumina MiSeq, and because PCR primers are commonly around 20 bp.
4. Windows where the flanking regions overlap are merged, provided this doesn't introduce a variant position into the conserved flanking regions. This lowers the number of windows to process.
5. Windows that contain a region of irregular coverage (as identified in step 1) are filtered out.

6. The allele sequences in each remaining window are compared across all strains. If each strain has at least one unique allele within the window, that window is reported to the user.

The pipeline thus reports a list of loci that fulfil the criteria of a good barcode, namely suitable length, sufficient diversity, and conserved flanking regions.

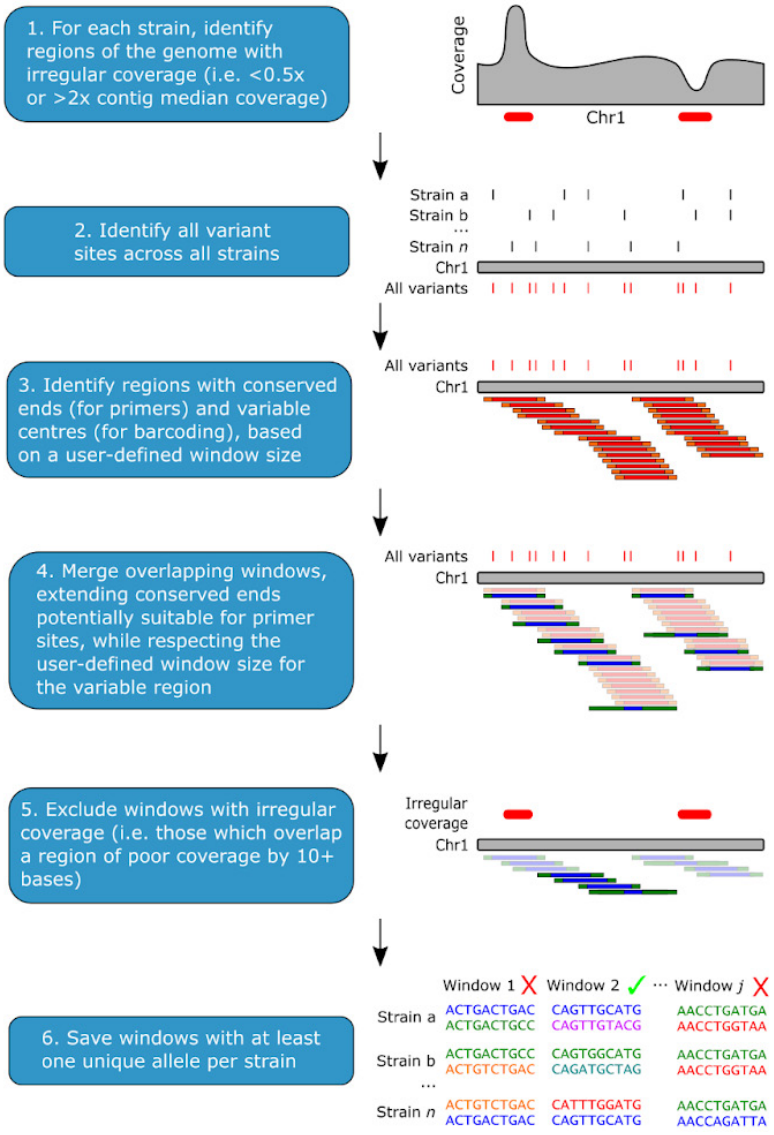

Figure 1: Workflow of the Bamboozle pipeline

Summary of papers

Paper I

In **paper I**, we sought to highlight the potential of diatom resting stages in molecular studies, with a particular focus on *Skeletonema marinoi*.

Diatoms have been used in historical studies for many years, although initially, it was their silica frustules that were of interest. By analysing and comparing the composition of frustules found in different sediment layers, one can draw conclusions about past environmental changes. In addition, traces of diatom DNA can be found, in the form of ancient DNA preserved in the sediment. However, given the ability of some diatom species to enter a long-lived resting stage, studies on past conditions using diatoms can be taken a step further. Whereas studies using the remains of dead diatoms can give us only limited information, and only a limited amount of (likely degraded) genetic material to work with, resting stage diatoms that have been revived can often be cultured, resulting in both a larger amount of DNA for genetic/genomic studies, as well as the potential to compare these revived cells with their contemporary counterparts.

In this paper, we identified numerous diatom studies that have made use of both diatom environmental DNA, as well as DNA derived from revived diatom resting stages. Those in the former category sequenced DNA barcodes from sedimentary DNA in order to follow lineages over time or determine taxonomic composition, while those in the latter category were primarily involved with inferring on population structure across time and space.

As nucleotide databases improve in terms of both DNA barcodes and whole genomes for diatoms, and as sequencing technologies and associated bioinformatic methods continue to evolve, we also discuss the future potential of diatom resting stage studies. Along with a best practice workflow for isolating, reviving, culturing, sequencing, and analysing resting stage diatoms, we discuss the benefits of applying these more recent techniques to diatom studies. These include improved read depth for supporting studies of variation between populations, and the use of machine learning to correct ancient diatom DNA sequences that have undergone post-mortem DNA mutation.

Lastly, we speculate on the types of ancient diatom DNA-based studies that could be undertaken in light of these developing technologies. For example, in water bodies that have been subjected to phenomena such as increased temperatures or eutrophication, diatom resting stages from before the onset of

that event could be revived, and, alongside their contemporary counterparts, sequenced in tandem with phenotypic studies in order to assess the loss or gain of protein function, or indeed genomic regions, that has occurred in the intervening years. Another example is the potential to perform comparative metagenomics of diatom microbiomes, to determine how their associated bacteria change during their host's dormancy. With the continued improvement of sequencing technology and bioinformatic techniques, we stand to be able to answer more and more questions about diatoms through their resting stages.

Paper II

In **paper II**, gene expression in the resting stages of two strains of *S. marinoi* was examined, to try and shed light on the processes keeping these dormant diatoms alive, potentially for a century or more. In particular, given recent findings that resting stages of *S. marinoi* can assimilate nitrate from the environment (Stenow *et al.*, 2020), we were interested in differential expression of nitrogen metabolism. Strains R05AC and GF0410J were exposed to darkness, anoxia, and nitrate starvation for 49 days to induce resting stages, at which point isotopically-labelled nitrate was resupplied to the resting stages. RNA was extracted from both strains at day 0 (before induction of resting stages), day 49 (a few hours after re-addition of nitrate), day 56, day 72, day 91, day 126, and day 189.

As expected, resting stages of *S. marinoi* exposed to long-term darkness and anoxia showed major shifts in expression of myriad metabolic pathways compared to their vegetative counterparts. Many of these genes are consistent with previous findings on diatom resting stages. We also observed a number of genes with opposing trends between the strains (e.g. resting stage upregulation in GF0410J, but downregulation in R05AC), highlighting the differences in gene expression between strains of the same diatom species, as noted in *S. marinoi* previously (Pinseel *et al.*, 2022).

Many genes pertaining to protein synthesis – ribosomal protein genes, nucleotide and amino acid synthesis genes – were consistently upregulated throughout the resting stage timepoints, in line with observations that some nitrogen assimilation takes place in the resting stages. Upregulation of genes involved in these processes appeared to be stronger in GF0410J than in strain R05AC, consistent with the findings of Stenow *et al.* (2020) that GF0410J resting stages had a higher survival rate than R05AC when provided with nitrate under darkness and anoxia.

In terms of pathways pertaining to energy production, genes in the TCA cycle were strongly downregulated. We found a mixture of both up- and down-regulation in genes pertaining to beta-oxidation and glycolysis, the latter displaying opposing expression changes between isozymes predicted to be localised to different cellular compartments. This leads us to believe that these two processes may be relevant in the resting stages, although the implications of the differing cellular locations of the upregulated isozymes for glycolysis is unknown. Both nitrate and nitrite reductase were upregulated, although whether these particular reductases are the ones responsible for DNRA would require further functional studies. When searching the genome for genes relevant to other anaerobic respiratory pathways, we identified two putative lactate dehydrogenases, both upregulated, suggesting the possibility of lactic acid fermentation. While we could not identify any genes pertaining to the potential dissimilatory reduction of sulfate or iron, we did observe upregulation of genes pertaining to iron uptake, suggesting its importance in the resting stages, potentially due to its role as a cofactor in nitrate assimilation-related genes.

The results of this paper provide a look at the processes active within the *S. marinoi* resting stages, and highlight several pathways that bear further investigation in future omics studies.

Paper III

In **paper III**, I investigated shifts in allele-specific expression (ASE) in *S. marinoi* strain R05AC under changing conditions. Specifically, using the RNAseq data generated in **paper II**, I examined the extent of the phenomenon in both vegetative cells and resting cells, and then performed a comparison between the two.

When assessing each timepoint individually, proportions of genes falling under the various categories of bias level are similar to those reported for *P. tricornutum* – 9-19% of genes were biallelically expressed (BAE), 20-23% of genes showed moderate ASE, and 3-4% of genes were strongly biassed (monoallelic expression; MAE).

When comparing the vegetative and resting cell datasets, only a small number of genes underwent a shift in bias that was both large (≥ 30% change in allelic expression bias) and consistent (observed across resting stages for which we had data in at least three of the six timepoints). I identified only 111 genes showing a large decrease in bias in the resting stages versus the vegetative cells, and 239 that

showed a similarly large increase in bias. In each case, only around half of the genes had any functional annotation.

While those genes showing a strong shift in bias between the vegetative and resting cells belonged to many different pathways, we noted that some of them formed complete or near-complete pathways with potential relevance to diatoms' silica cell wall (frustule). Three genes involved in the synthesis of polyamines – a major organic component of the frustule – displayed increased bias in the resting stages, as did two genes in the biosynthesis pathway of UDP-xylose, a precursor to one of the polysaccharides associated with the frustule. Twenty-one transporter were also among the bias-shifted genes, showing a mixture of increased (14) and decreased (7) bias in the resting stages.

Taken together, allele-specific expression appears to be a relevant process in *Skeletonema marinoi* and its resting stages. Given the presence of various genes related to the frustule among those genes displaying a strong shift in bias, a comparison of frustule composition between vegetative and resting cells may provide further insights into the resting stages.

Paper IV

Given the high level of intraspecific diversity within diatoms, and the availability of the *S. marinoi* genome (as described in **paper II**), I sought to determine whether it was possible to identify loci within the genome that were sufficiently variable to enable distinction between strains within the species, while also possessing conserved flanking regions for the binding of PCR primers, thus allowing sequencing in the same manner as a more traditional DNA barcode.

In **paper IV**, I developed the Bamboozle pipeline for identifying such loci at the intraspecific level. The context for developing such a pipeline was an artificial evolution experiment regarding copper tolerance in *S. marinoi* (**paper IV**; Andersson *et al.*, 2023). Multiple strains of *S. marinoi* were mixed together in equal proportions, and subjected to copper stress, with the aim of quantifying fitness of all strains during co-cultivation. However, conventional barcodes cannot distinguish between strains in this species, thus a different barcode had to be identified for this purpose.

WGS data from 54 strains of *S. marinoi* was used to test the pipeline, resulting in the identification of four loci of around 500 bp that fulfilled the criteria of a suitable barcode. We then used amplicon sequencing to check the accuracy of Bamboozle's predictions (by sequencing each barcode in monocultures of each

strain), showing that the pipeline's predictions were 99.9% accurate. The 0.1% error was attributed to the unexpected presence of two triploid strains, which confounded the variant-calling software's expectations of diploidy.

One of the barcodes was then used to track the abundance of strains over the course of the copper tolerance study. Merging and denoising the amplicon reads resulted in substantial loss of data, as quality trimming of the sequencing data led to a very short overlap between the paired reads. However, the remaining data largely showed the expected result – a 1:1 ratio between each strain's two alleles across the time series. The four remaining false-positive sequences were identified as chimeras, based on comparisons with known sequences in the dataset.

In order to broaden the applicability of Bamboozle, I made adjustments to the initial diploid-oriented pipeline to allow it to function with haploid genomes. This involved accounting for a smaller number of input files (one BAM and VCF file per strain, rather than three each), and ensuring that each allele reported per window is entirely unique within the dataset. This was tested on seventeen strains of the model green alga *Chlamydomonas reinhardtii*. In order to identify suitable barcodes in this species, the barcode length parameter was shortened to 300 bp, and the pipeline was run separately on each mating type (in line with the intended experimental design involving no mixing of mating types). Thus, we identified fifteen barcodes for the eight mt+ strains, and seven barcodes for the nine mt- strains. Sanger sequencing of one barcode from each mating type showed prediction accuracy of 99.98% and 99.88%, respectively.

Our results suggest that the intraspecific diversity of diatoms can indeed be harnessed to identify novel barcoding loci, allowing quantitative analysis of mixed-strain cultures. Our method can also be applied to non-diatom organisms, including haploids, which will hopefully make Bamboozle a useful tool for the wider scientific community.

Conclusion and future directions

In **paper I**, we discussed the potential research directions for future diatom resting stage studies. Given the increasing length and accuracy, and decreasing costs, of high-throughput sequencing, adding a genomic dimension to such studies becomes more and more viable, as evidenced by the sequencing of resting stages (**papers II** and **III**) and revived resting stages (**paper IV**) in the other parts of this thesis. Adding to this the improvements in bioinformatic tools, the diversity of potential studies in this field is immense.

In **papers II** and **III**, we gained insight into the gene expression of *S. marinoi* resting stages, and the pathways that are active in this 'dormant' life stage. Much of the data from **paper II** is consistent with previous findings in resting stages, however given the short timeframe of many other resting stage gene expression studies, we show that these trends continue into the long-term. Some of our findings, such as the differential expression of differently-targeted isozymes of glycolytic genes, certainly require further study into their significance. The findings in **paper III** also provide novel insight into the prevalence of allele-specific expression in both vegetative and resting stages of *S. marinoi*. In particular, the potential phenotypic consequences of the observed shifts in allelic expression bias in genes related to the frustule would be an interesting research question.

A major observation in these two studies was the sheer number of gene models only annotated as 'predicted protein' in the datasets of interest. Thankfully, protocols for mutagenesis of diatoms have been established (e.g. Stukenberg *et al.*, 2018; Johansson *et al.*, 2019a), and further investigation into *S. marinoi* mutagenesis is currently underway. Considering the potential of diatoms in biotechnology, such as production of biosilica and alternative fuels, such methods are very important, and are aided by the increasing number of available diatom genomes. Thus, the outlook is positive for improved functional annotation of diatom genomes in the years to come, and I hope that such developments aid in answering the questions raised in **papers II** and **III**.

In **paper IV**, we determined that the intraspecific diversity of diatoms can be used to follow changing frequencies of strains within a mixed sample, and present a tool for identifying such barcoding loci. While in its current state, Bamboozle has proven able to identify multiple potential strain-specific markers in both *S. marinoi* and *C. reinhardtii*, there are additional developments I hope to make that should further improve its potential. Chief among these are 1) testing on more strains of *S. marinoi* to determine whether the identified barcodes remain

informative across additional genotypes; 2) testing the pipeline across additional taxa to determine how widely the pipeline can be applied; 3) improving the flexibility of the scanning window sizing, which could potentially identify further novel barcoding loci; and 4) include additional steps such as scanning for off-target primer binding, to minimise the downstream manual filtering required. With such improvements, I hope that the pipeline may prove useful in a wider array of studies, perhaps including those aimed at natural populations.

Funding

This thesis was supported by the University of Gothenburg, the Swedish Research Council for Sustainable Development Formas (2017-00466), and the Swedish Research Council Vetenskapsrådet (2018-04555).

References

Allen AE, Dupont CL, Oborník M, Horák A, Nunes-Nesi A, McCrow JP, Zheng H, Johnson DA, Hu H, Fernie AR, Bowler C (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. Nature 473:203–207. doi: 10.1038/nature10074

Andersson B, Godhe A, Filipsson HL, Rengefors K, Berglund O (2020) Differences in metal tolerance among strains, populations, and species of marine diatoms – Importance of exponential growth for quantification. Aquat Toxicol 226:105551. doi: 10.1016/j.aquatox.2020.105551

Anderson B, Berglund O, Filippson H, Kourtchenko O, Godhe A, Johannesson K, Töpel M, Pinder M, Hoepfner L, Rengefors K (2023) Strain-specific metabarcoding reveals rapid evolution of copper tolerance in populations of the coastal diatom *Skeletonema marinoi*. Accepted in Molecular Ecology. doi: 10.22541/au.168069094.46659870/v1

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov V V, Kröger N, Lau WWY, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The Genome of the Diatom *Thalassiosira pseudonana*: Ecology, Evolution, and Metabolism. Science 306:79–86. doi: 10.1126/science.1101156

Bates SS, Hubbard KA, Lundholm N, Montresor M, Leaw CP (2018) *Pseudo-nitzschia*, *Nitzschia*, and domoic acid: New research since 2011. Harmful Algae 79:3–43. doi: 10.1016/j.hal.2018.06.001

Bourke MF, Marriott PJ, Glud RN, Hasler-Sheetal H, Kamalanathan M, Beardall J, Greening C, Cook PLM (2017) Metabolism in anoxic permeable sediments is dominated by eukaryotic dark fermentation. Nat Geosci 10:30–35. doi: 10.1038/ngeo2843

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret J-P, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, Fournet J, Haruta M, Huysman MJJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kröger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin–Jézéquel V, Lopez PJ, Lucas S, Mangogna M, McGinnis K, Medlin LK, Montsant A, Secq M-PO, Napoli C, Obornik M, Parker MS, Petit J-L, Porcel BM, Poulsen N, Robison M, Rychlewski L, Rynearson TA, Schmutz J, Shapiro H, Siaut M, Stanley M, Sussman MR, Taylor AR, Vardi A, von Dassow P, Vyverman W, Willis A, Wyrwicz LS, Rokhsar DS, Weissenbach J, Armbrust EV, Green BR, Van de Peer Y, Grigoriev I V (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature 456:239–244. doi: 10.1038/nature07410

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. doi: 10.1186/1471-2105-10-421

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 38:5825–5829. doi: 10.1093/molbev/msab293

Cedar H, Bergman Y (2008) Choreography of Ig allelic exclusion. Curr Opin Immunol 20:308–317. doi: 10.1016/j.coi.2008.02.002

Cleary S, Seoighe C (2021) Perspectives on Allele-Specific Expression. Annu Rev Biomed Data Sci 4:101–122. doi: 10.1146/annurev-biodatasci-021621-122219

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021) Twelve years of SAMtools and BCFtools. Gigascience 10:giab008. doi: 10.1093/gigascience/giab008

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. doi: 10.1093/bioinformatics/bts635

Ellegaard M, Godhe A, Härnström K, McQuoid M (2008) The species concept in a marine diatom: LSU rDNA–based phylogenetic differentiation in *Skeletonema marinoi/dohrnii* (Bacillariophyceae) is not reflected in morphology. Phycologia 47:156–167. doi: 10.2216/07-09.1

Emeis K-C, Struck U, Blanz T, Kohly A, Voβ M (2003) Salinity changes in the central Baltic Sea (NW Europe) over the last 10000 years. The Holocene 13:411–421. doi: 10.1191/0959683603hl634rp

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. Science 281:237–240. doi: 10.1126/science.281.5374.237

Flowers JM, Hazzouri KM, Pham GM, Rosas U, Bahmani T, Khraiwesh B, Nelson DR, Jijakli K, Abdrabu R, Harris EH, Lefebvre PA, Hom EFY, Salehi-Ashtiani K, Purugganan MD (2015) Whole-Genome Resequencing Reveals Extensive Natural Variation in the Model Green Alga *Chlamydomonas reinhardtii*. Plant Cell 27:2353–2369. doi: 10.1105/tpc.15.00492

Fryxell GA, Prasad AKSK (1990) *Eucampia antarctica* var. *recta* (Mangin) stat. nov. (Biddulphiaceae, Bacillariophyceae): life stages at the Weddell Sea ice edge. Phycologia 29:27–38. doi: 10.2216/i0031-8884-29-1-27.1

Gain G, Berne N, Corato A, Godaux D, Remacle C, Cardol P (2021) Fermentation pathways and their interactions with photosynthesis in the marine diatom *T. pseudonana*: proteomic and biophysics approaches. [poster presentation]. Journées Françaises de Photosynthèse. https://hdl.handle.net/2268/258719

Godhe A, Rynearson T (2017) The role of intraspecific variation in the ecological and evolutionary success of diatoms in changing environments. Philos Trans R Soc B Biol Sci 372:20160399. doi: 10.1098/rstb.2016.0399

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652. doi: 10.1038/nbt.1883

Guo L, Sui Z, Zhang S, Ren Y, Liu Y (2015) Comparison of potential diatom "barcode" genes (the 18S rRNA gene and ITS, COI, rbcL) and their effectiveness in discriminating and determining species taxonomy in the Bacillariophyta. Int J Syst Evol Microbiol 65:1369–1380. doi: 10.1099/ijs.0.000076

Härnström K, Ellegaard M, Andersen TJ, Godhe A (2011) Hundred years of genetic structure in a sediment revived diatom population. Proc Natl Acad Sci U S A 108:4252–4257. doi: 10.1073/pnas.1013528108

Hildebrand M, Davis AK, Smith SR, Traller JC, Abbriano R (2012) The place of diatoms in the biofuels industry. Biofuels 3:221–240. doi: 10.4155/bfs.11.157

Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, Graham SW, James KE, Kim K-J, Kress WJ, Schneider H, van AlphenStahl J, Barrett SCH, van den Berg C, Bogarin D, Burgess KS, Cameron KM, Carine M, Chacón J, Clark A, Clarkson JJ, Conrad F, Devey DS, Ford CS, Hedderson TAJ, Hollingsworth ML, Husband BC, Kelly LJ, Kesanakurti PR, Kim JS, Kim Y-D, Lahaye R, Lee H-L, Long DG, Madriñán S, Maurin O, Meusnier I, Newmaster SG, Park C-W, Percy DM, Petersen G, Richardson JE, Salazar GA, Savolainen V, Seberg O, Wilkinson MJ, Yi D-K, Little DP (2009) A DNA barcode for land plants. Proc Natl Acad Sci 106:12794–12797. doi: 10.1073/pnas.0905845106

Hoguin A, Rastogi A, Bowler C, Tirichine L (2021) Genome-wide analysis of allele-specific expression of genes in the model diatom *Phaeodactylum tricornutum*. Sci Rep 11:2954. doi: 10.1038/s41598-021-82529-1

Jeffryes C, Campbell J, Li H, Jiao J, Rorrer G (2011) The potential of diatom nanobiotechnology for applications in solar cells, batteries, and electroluminescent devices. Energy Environ Sci 4:3930–3941. doi: 10.1039/C0EE00306A

Johansson ON, Töpel M, Pinder MIM, Kourtchenko O, Blomberg A, Godhe A, Clarke AK (2019a) *Skeletonema marinoi* as a new genetic model for marine chain-forming diatoms. Sci Rep 9:5391. doi: 10.1038/s41598-019-41085-5

Johansson ON, Pinder MIM, Ohlsson F, Egardt J, Töpel M, Clarke AK (2019b) Friends With Benefits: Exploring the Phycosphere of the Marine Diatom *Skeletonema marinoi*. Front. Microbiol. 10:1828. doi: 10.3389/fmicb.2019.01828

Joli N, Concia L, Mocaer K, Guterman J, Laude J, Guerin S, Sciandra T, Bruyant F, Ait-Mohamed O, Beguin M, Forget M-H, Bourbousse C, Lacour T, Bailleul B, Tremblay J-E, Campbell D, Lavaud J, Schwab Y, Babin M, Bowler C (2023) Hypometabolism to survive the long polar night in the diatom *Fragilariopsis cylindrus*. bioRxiv 2023.01.14.524047. doi: 10.1101/2023.01.14.524047

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. doi: 10.1093/bioinformatics/btu031

Kamp A, de Beer D, Nitsch JL, Lavik G, Stief P (2011) Diatoms respire nitrate to survive dark and anoxic conditions. Proc Natl Acad Sci 108:5649–5654. doi: 10.1073/pnas.1015744108

Kamp A, Stief P, Knappe J, de Beer D (2013) Response of the ubiquitous pelagic diatom *Thalassiosira weissflogii* to darkness and anoxia. PLoS One 8:e82605–e82605. doi: 10.1371/journal.pone.0082605

Kamp A, Stief P, Bristow LA, Thamdrup B, Glud RN (2016) Intracellular Nitrate of Marine Diatoms as a Driver of Anaerobic Nitrogen Cycling in Sinking Aggregates. Front Microbiol 7:1669. doi: 10.3389/fmicb.2016.01669

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915. doi: 10.1038/s41587-019-0201-4

Kress WJ, Erickson DL (2008) DNA barcodes: Genes, genomics, and bioinformatics. Proc Natl Acad Sci U S A 105:2761–2762. doi: 10.1073/pnas.0800476105

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. doi: 10.1038/nmeth.1923

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930. doi: 10.1093/bioinformatics/btt656

Lomas MW, Glibert PM (2000) Comparisons of nitrate uptake, storage, and reduction in marine diatoms and flagellates. J Phycol 36:903–913. doi: 10.1046/j.1529-8817.2000.99029.x

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. doi: 10.1186/s13059-014-0550-8

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A (2020) CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48:D265–D268. doi: 10.1093/nar/gkz991

Maeda Y, Kobayashi R, Watanabe K, Yoshino T, Bowler C, Matsumoto M, Tanaka T (2022) Chromosome-Scale Genome Assembly of the Marine Oleaginous Diatom *Fistulifera solaris*. Mar Biotechnol 24:788–800. doi: 10.1007/s10126-022-10147-7

Mata J, Marguerat S, Bähler J (2005) Post-transcriptional control of gene expression: a genome-wide perspective. Trends Biochem Sci 30:506–514. doi: 10.1016/j.tibs.2005.07.005

McQuoid MR, Godhe A (2004) Recruitment of coastal planktonic diatoms from benthic versus pelagic cells: Variations in bloom development and species composition. Limnol Oceanogr 49:1123–1133. doi: 10.4319/lo.2004.49.4.1123

McQuoid MR, Hobson LA (1996) Diatom Resting Stages. J Phycol 32:889–902. doi: 10.1111/j.0022-3646.1996.00889.x

Merz E, Dick GJ, de Beer D, Grim S, Hübener T, Littmann S, Olsen K, Stuart D, Lavik G, Marchant HK, Klatt JM (2021) Nitrate respiration and diel migration patterns of diatoms are linked in sediments underneath a microbial mat. Environ Microbiol 23:1422–1435. doi: 10.1111/1462-2920.15345

Middelburg JJ (2019) Carbon Processing at the Seafloor. In: Middelburg JJ (ed.) Marine Carbon Biogeochemistry: A Primer for Earth System Scientists. Springer, Cham. doi: 10.1007/978-3-030-10822-9_4

Mock T, Otillar RP, Strauss J, McMullan M, Paajanen P, Schmutz J, Salamov A, Sanges R, Toseland A, Ward BJ, Allen AE, Dupont CL, Frickenhaus S, Maumus F, Veluchamy A, Wu T, Barry KW, Falciatore A, Ferrante MI, Fortunato AE, Glöckner G, Gruber A, Hipkin R, Janech MG, Kroth PG, Leese F, Lindquist EA, Lyon BR, Martin J, Mayer C, Parker M, Quesneville H, Raymond JA, Uhlig C, Valas RE, Valentin KU, Worden AZ, Armbrust EV, Clark MD, Bowler C, Green BR, Moulton V, van Oosterhout C, Grigoriev I V (2017) Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. Nature 541:536–540. doi: 10.1038/nature20803

Mock T, Hodgkinson K, Wu T, Moulton V, Duncan A, van Oosterhout C, Pichler M (2022) Structure and Evolution of Diatom Nuclear Genes and Genomes. In: Falciatore A, Mock T (eds) The Molecular Life of Diatoms. Springer, Cham. doi: 10.1007/978-3-030-92499-7_5

Nelson DM, Tréguer P, Brzezinski MA, Leynaert A, Quéguiner B (1995) Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Global Biogeochem Cycles 9:359–372. doi: 10.1029/95GB01070

Ness RW, Kraemer SA, Colegrave N, Keightley PD (2016) Direct Estimate of the Spontaneous Mutation Rate Uncovers the Effects of Drift and Recombination in the *Chlamydomonas reinhardtii* Plastid Genome. Mol Biol Evol 33:800–808. doi: 10.1093/molbev/msv272

Nomaguchi T, Maeda Y, Yoshino T, Asahi T, Tirichine L, Bowler C, Tanaka T (2018) Homoeolog expression bias in allopolyploid oleaginous marine diatom *Fistulifera solaris*. BMC Genomics 19:330. doi: 10.1186/s12864-018-4691-0

Pinseel E, Nakov T, Van den Berge K, Downey KM, Judy KJ, Kourtchenko O, Kremp A, Ruck EC, Sjöqvist C, Töpel M, Godhe A, Alverson AJ (2022) Strain-specific transcriptional responses overshadow salinity effects in a marine diatom sampled along the Baltic Sea salinity cline. ISME J 16:1776–1787. doi: 10.1038/s41396-022-01230-x

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. doi: 10.1093/bioinformatics/btq033

Sanyal A, Larsson J, van Wirdum F, Andrén T, Moros M, Lönn M, Andrén E (2021) Not dead yet: Diatom resting spores can survive in nature for several millennia. Am J Bot 109:1–16. doi: 10.1002/ajb2.1780

Saravanan V, Godhe A (2010) Genetic heterogeneity and physiological variation among seasonally separated clones of *Skeletonema marinoi* (Bacillariophyceae) in the Gullmar Fjord, Sweden. Eur J Phycol 45:177–190. doi: 10.1080/09670260903445146

Sarno D, Kooistra WHCF, Medlin LK, Percopo I, Zingone A (2005) Diversity in the genus *Skeletonema* (Bacillariophyceae). II. An assessment of the taxonomy of *S. costatum*-like species with the description of four new species. J Phycol 41:151–176. doi: 10.1111/j.1529-8817.2005.04067.x

Shykind BM (2005) Regulation of odorant receptors: one allele at a time. Hum Mol Genet 14:R33–R39. doi: 10.1093/hmg/ddi105

Söderhielm J, Sundblad K (1996) The Solstad Cu-Co-Au mineralization and its relation to post-Svecofennian regional shear zones in southeastern Sweden. GFF 118:47. doi: 10.1080/11035899609546323

Sorhannus U (2011) Evolution of Antifreeze Protein Genes in the Diatom Genus Fragilariopsis: Evidence for Horizontal Gene Transfer, Gene Duplication and Episodic Diversifying Selection. Evol Bioinforma 7:EBO.S8321. doi: 10.4137/EBO.S8321

Stenow R, Olofsson M, Robertson EK, Kourtchenko O, Whitehouse MJ, Ploug H, Godhe A (2020) Resting stages of *Skeletonema marinoi* assimilate nitrogen from the ambient environment under dark, anoxic conditions. J Phycol 56:699–708. doi: 10.1111/jpy.12975

Stenow R (2023) Single cell carbon and nitrogen dynamics in chain forming diatoms, including their resting stages. *Doctoral thesis*. https://hdl.handle.net/2077/76760

Stukenberg D, Zauner S, Dell'Aquila G, Maier UG (2018) Optimizing CRISPR/Cas9 for the Diatom *Phaeodactylum tricornutum*. Front Plant Sci 9:740. doi: 10.3389/fpls.2018.00740

Sundqvist L, Godhe A, Jonsson PR, Sefbom J (2018) The anchoring effect—long-term dormancy and genetic population structure. ISME J 12:2929–2941. doi: 10.1038/s41396-018-0216-8

Tanaka T, Maeda Y, Veluchamy A, Tanaka M, Abida H, Maréchal E, Bowler C, Muto M, Sunaga Y, Tanaka M, Yoshino T, Taniguchi T, Fukuda Y, Nemoto M, Matsumoto M, Wong PS, Aburatani S, Fujibuchi W (2015) Oil Accumulation by the Oleaginous Diatom *Fistulifera solaris* as Revealed by the Genome and Transcriptome. Plant Cell 27:162–176. doi: 10.1105/tpc.114.135194

Tréguer P, Bowler C, Moriceau B, Dutkiewicz S, Gehlen M, Aumont O, Bittner L, Dugdale R, Finkel Z, Iudicone D, Jahn O, Guidi L, Lasbleiz M, Leblanc K, Levy M, Pondaven P (2018) Influence of diatom diversity on the ocean biological carbon pump. Nat Geosci 11:27–37. doi: 10.1038/s41561-017-0028-x

Tucci V, Isles AR, Kelsey G, Ferguson-Smith AC, Tucci V, Bartolomei MS, Benvenisty N, Bourc'his D, Charalambous M, Dulac C, Feil R, Glaser J, Huelsmann L, John RM, McNamara GI, Moorwood K, Muscatelli F, Sasaki H, Strassmann BI, Vincenz C, Wilkins J, Isles AR, Kelsey G, Ferguson-Smith AC (2019) Genomic Imprinting and Physiological Processes in Mammals. Cell 176:952–965. doi: 10.1016/j.cell.2019.01.043

Van der Auwera GA, O'Connor BD (2020) Genomics in the Cloud: Using Docker, GATK, and WDL in Terra, 1st edn. O'Reilly Media Inc.

Van Etten J, Bhattacharya D (2020) Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? Trends Genet 36:915–925. doi: 10.1016/j.tig.2020.08.006

Vancaester E, Depuydt T, Osuna-Cruz CM, Vandepoele K (2020) Comprehensive and Functional Analysis of Horizontal Gene Transfer Events in Diatoms. Mol Biol Evol 37:3243–3257. doi: 10.1093/molbev/msaa182

Yamada M, Katsuki E, Otsubo M, Kawaguchi M, Ichimi K, Kaeriyama H, Tada K, Harrison PJ (2010) Species diversity of the genus *Skeletonema* (Bacillariophyceae) in the industrial harbor Dokai Bay, Japan. J Oceanogr 66:755–771. doi: 10.1007/s10872-010-0062-4

Zhang M, Zhen Y, Mi T, Lin S (2021) Integrated Transcriptome Sequencing (RNA-seq) and Proteomic Studies Reveal Resource Reallocation towards Energy Metabolism and Defense in *Skeletonema marinoi* in Response to CO_2 Increase. Appl Environ Microbiol 87:e02614-20. doi: 10.1128/AEM.02614-20