Resolving a bioindicator diatom species complex using genomic approaches for freshwater biomonitoring

Ciftci, O.

CHAPTER 1

General Introduction

1.1 Origins of diatoms and their life cycle

Diatoms are unicellular, microscopic eukaryotes that are important players in the biogeochemical cycles of carbon and silicon. At least 20% of all carbon fixed through photosynthesis globally each year, a comparable scale to all terrestrial rainforests combined, is fixed by diatoms (Field et al., 1998; Mann, 1999; Falkowski et al., 2000). Over geological time, diatoms may have influenced global climate by changing the flux of atmospheric carbon dioxide into the oceans (Brzezinski, 2002). Diatoms also take up and use available silicic acid in the environment to synthesize their mineralized cell walls which makes them key players in the silicon cycling in the world's oceans (Tréguer et al., 1995). Furthermore, diatoms form the primary food source for many organisms, including copepods, fish, and filter feeders such as mollusks and tunicates (Arrigo, 2005). Therefore, a reduction in diatom diversity may have great repercussions for higher trophic levels.

Diatoms belong to the stramenopiles group, also called heterokonts, many of which contain plastids that are rich in chlorophylls a and c. There are two major diatom groups based on their symmetry, the so-called centrics (radially symmetrical) and the pennates (bilaterally symmetrical). The first diatom fossils are centrics that appear after the Permian–Triassic mass extinctions and date to around 180 million years ago (Sims et al., 2006; Armbrust, 2009). Although there are uncertainties surrounding these oldest fossil records of diatoms, molecular clocks estimate that diatom microalgae originated near the Triassic–Jurassic boundary (200 Ma) (Nakov et al., 2018a; Bryłka et al., 2023). Early Cretaceous (100-145 million years ago) fossil record contains centric diatoms only, while pennate diatoms are first recorded in the late Cretaceous period (65-75 million years ago) (Hajos & Stradner, 1975; Harwood, 1988; Harwood & Gersonde, 1990). These ancient pennate diatoms are araphid (non-motile), whereas the raphid (motile) pennate diatoms first appear in the Paleocene (55.8-65.5 million years ago) and reach reasonable numbers in the middle Eocene (48-38 million years ago) (Medlin et al., 1993; Sims et al., 2006). Therefore, centric diatoms grade into araphid pennates, and araphid pennates further grade into the raphid pennate clade, which is the only monophyletic group among these three (Alverson & Theriot, 2005). We can trace back the appearance of many modern diatoms to around 30 million years ago at the Eocene–Oligocene boundary when the populations of these major divisions expanded and diversified enormously (Bowler et al., 2010).

In addition to the difference in their pattern centers (Mann, 1984), centric and pennate diatoms also differ in their modes of sexual reproduction (Drebes, 1977), and types of plastids (Mereschkowsky, 1902-1903). Centric diatoms release their sperm into the aquatic environment, whereas pennate diatoms (both araphid and raphid) initiate sexual pairing via physical contact or production of mucilage envelopes or copulation tubes (Round et al., 1990; Edlund & Stoermer, 1997). In addition to this period of sexual reproduction, the diatom life cycle comprises a period of vegetative, mitotic cell division. In this vegetative period, new cell wall elements (valves and girdle bands) are produced within existing wall elements of the parent cell and are therefore smaller (Round et al., 1990). Sexual reproduction in diatoms can occur only after the cells reach a minimum size range, typically 30- 40% of their maximum size (Edlund & Stoermer, 1997). A second condition that must be met to induce sexual reproduction is the presence of the correct environmental conditions, which include specific combinations of temperature, light, nutrients, trace metals, organic growth factors, and osmolarity (Edlund & Stoermer, 1997; Chepurnov et al., 2004).

1.2 Diatom biomonitoring

Diatoms have high specificity towards environmental parameters, such as nutrients (Borchardt, 1996), pH (Smol et al., 1986), salinity (Snoeijs, 1999), light and temperature (Hill, 1996; DeNicola, 1996), and hydrological conditions (Dixit et al., 1993; Fritz et al., 2010). Relying on their enormous ecological importance and global abundance, including the fossil record, diatoms are applied as ecological indicators of biotic and abiotic conditions, as well as of anthropogenic and natural impacts (Smol & Stoermer, 2010; Stevenson et al., 2010). Aquatic biodiversity is particularly affected by the increased release of chemicals from agricultural, industrial, and domestic sources over the past decades (Dudgeon et al., 2006; Vörösmarty et al., 2010). Therefore, large-scale monitoring programs have been established such as the European Union Water Framework Directive (Directive 2000/60/EC) to assess the resulting degradation. Biofilms are one of the biological compartments recognized by the WFD as a necessary target which are attached communities of microorganisms on surfaces. These organisms have small sizes, grow rapidly, and are physiologically diverse. Therefore, biofilms respond quickly to changes in environmental conditions and integrate such effects over long periods (Sabater et al., 2007). Diatoms are important constituents of biofilms as primary producers. Several studies demonstrate that diatoms show higher sensitivity compared to other aquatic bioindicator organism groups, especially when eutrophication/organic pollution is assessed (Hering et al., 2006; Johnson et al., 2006). Accordingly, several diatom indices have been developed to estimate freshwater quality, such as "Indice de Polluosensibilité Spécifique" (Cemagref, 1982) or the Biological Diatom Index (BDI) (Coste et al., 2009), that are routinely used for biomonitoring applications in several European countries.

The end product of these surveys is a taxonomically explicit estimate of the species composition and their relative abundances, which are used to calculate ecological indices or compared to those expected in the absence of anthropogenic pressures (Benfield et al., 2007; Leese et al., 2018). The task of diatom identification and quantification is performed by experts via examination of their tiny frustules and requires special sample preparation, high-quality microscopes, and taxonomic expertise (Keck et al., 2017; Apothéloz-Perret-Gentil et al., 2017; Apothéloz-Perret-Gentil et al., 2021). Therefore, more rapid and cost-effective methods of species identification and quantification are needed for diagnostic monitoring purposes, and a methodological shift to molecular approaches is strongly advocated and anticipated (Carew et al., 2013; Apothéloz-Perret-Gentil et al., 2021).

In the following sections, I will (i) provide a short review of diatom systematics, (ii) summarize how recent advancements in molecular biology affected diatom systematics and improved our understanding of diatom evolution, with a specific focus on cryptic species, (iii) outline the advantages and limitations of currently available molecular methods for diatom biomonitoring, and (iv) argue that genome-scale approaches can provide more effective tools for diatom biomonitoring and inform diatom systematics at the same time.

1.3 Diatom systematics

Diatom identification traditionally relies on the examination of morphological features of the siliceous cell wall. The application of the scanning electron microscope (SEM) revealed ultrastructural features of the cell wall that are not visible under light microscopy, and consequently, SEM data became foundational for the diatom classification system (Round et al., 1990). Most species boundaries in diatoms are based on multiple lines of evidence, but reproductive isolation is often considered the decisive criterion consistent with the biological species concept (Mayr, 1942). In this concept, the discontinuities in morphological variation are assumed to originate from restrictions to gene flow between populations (i.e., reproductive isolation), allowing divergence through adaptation and genetic drift (Mann, 2010). Therefore, these discontinuities are often used as a proxy for reproductive isolation and species limits in diatoms, without much insight into their adaptive significance or ontogenic development (Kociolek et al., 1989; Mann, 1999). Most recently, there are more studies investigating multiple lines of evidence such as reproductive isolation (i.e., mating experiments), morphology, and genetics to infer species limits in diatoms (Mann et al., 2004; Amato et al., 2007; Vanormelingen et al., 2007; Trobajo et al., 2009). These can be considered in line with the 'separately evolving metapopulation lineage` concept by de Queiroz (2007), which proposes that congruent lines of evidence are required to obtain stronger support for lineage separation so the species hypothesis becomes less likely to be rejected (Alverson, 2008).

Like SEM, the application of molecular biological techniques to diatom systematics has also a marked impact because a hidden variation has been revealed (Alverson, 2008). An increasing number of molecular studies have shown that diatom diversity based on cell morphology is highly underestimated and many diatom morphospecies harbor distinct molecular variation, likely corresponding to species-level differentiation (i.e., cryptic species) (Vanormelingen et al., 2008; Quijano-Scheggia et al., 2009; Kermarrec et al., 2013; Pinseel et al., 2019). Still, the role of valve morphology will very likely remain central to diatom taxonomy because it is easily described, always available (i.e. in contrast to methods that require culturing), and most importantly, provides continuity with past taxonomies (Mann, 1999). However, molecular methods come with their own merits: (i) cryptic species can be distinguished where morphology is by definition inadequate for species discovery and identification, and (ii) valuable clues about the biogeography, phylogeny, and reproductive history of natural populations can be obtained by interpreting gene flow and dispersal mechanisms (Alverson, 2008; Medlin, 2018). Genome analyses of a few dozen diatom species examined so far revealed evidence for rapid divergence rates, natural hybrids, and genome duplication in diatoms (Bowler et al., 2008; Casteleyn et al., 2009; Tanaka et al., 2015; Parks et al., 2018; Pinseel et al., 2020).

Molecular systematic studies of diatoms are mostly based on different portions of the nuclear ribosomal DNA (rDNA), mitochondrial *cox*1 and a few plastid genes due to the availability of universal primers and the growing databases of these sequences which facilitate broad comparative analyses. Small subunit (SSU or 18S) rDNA has been widely used for reconstructing higher-level taxonomic relationships in diatoms, because its phylogenetic signal rises slowly and steadily to deeper nodes, whereas plastid *psb*C and *rbc*L provide more resolution towards the tips (Alverson et al., 2006; Theriot et al., 2015). Large subunit (LSU or 28S) and internal transcribed spacer (ITS) regions of rDNA can also resolve species and sometimes population-level relationships (Beszteri et al., 2005; Vanormelingen et al., 2007). However, high amounts of intragenomic polymorphisms were reported along the entire length of the rDNA cistron (i.e. more than 7%) which can obscure species boundaries and biodiversity estimates (Beszteri et al., 2005; Alverson, 2008). Plastid and mitochondrial markers are usually preferred over nuclear markers due to the advantages associated with their uniparental inheritance. However, the transmission of plastids was shown to be biparental in some diatoms, suggesting that plastid DNA might not always confer this advantage (Ghiron et al., 2008; Rimet et al., 2014). The use of plastid markers together with rDNA had been suggested to provide a most complete picture of phylogenetic relationships in diatoms (Alverson & Theriot, 2005; Fazekas et al., 2009). Accordingly, there have been several studies in the last decade that used multiple markers (Theriot, 2010; Abarca et al., 2014; Pinseel et al., 2019). The utility of a four-gene set derived from both nuclear and plastid parts of the genome (i.e. nSSU+nLSU+*rbc*L+*psb*C) was demonstrated to uncover several important findings on diatom evolution and diversity (Alverson et al., 2007; Stepanek & Kociolek, 2019; Mann et al., 2021).

1.4 Genome-scale approaches in diatom systematics

These few markers, however, are still unlikely to resolve groups that have recently diverged, which very likely comprise a large proportion of cryptic diatom species identified to date (Behnke et al., 2004; Evans et al., 2007; Alverson, 2008). The problem in these cases is that individual gene trees can disagree with the underlying species tree due to processes such as incomplete lineage sorting (ILS) or gene flow (Maddison, 1997; Degnan & Rosenberg, 2006; Mann, 2010). Therefore, the use of multiple unlinked loci from faster-evolving nuclear regions is necessary to investigate these processes (Funk & Omland, 2003; Alverson, 2008; Vanormelingen et al., 2013; Parks et al., 2018). There are different high-throughput sequencing (HTS) based methods for sequencing a large number of organellar and/or nuclear loci, including transcriptome sequencing (RNA-seq), genome skimming, restriction site-associated DNA sequencing (RAD-seq), and targeted capture (Hyb-seq) (Yu et al., 2018). Among these methods, only Hyb-seq requires a priori knowledge of the target genome for the design of capture probes, whereas other methods allow *de-novo* assembly of the organellar and/or nuclear DNA from thousands of loci in non-model species as well. Genome-scale data obtained by such methods have been successfully used to reveal evidence of ILS and gene flow and provided valuable insights into the reproductive history of cryptic diatom species (Mallet, 2005; Beszteri et al., 2007; Parks et al., 2018). This knowledge may have significant implications for diatom systematics because the evidence for gene flow (i.e., the lack of reproductive isolation) would require updates to species circumscriptions that are based on morphological data.

The unique silica metabolism of diatoms, which is a key factor in their ecological success, directed the attention of diatom genomics research toward silica biomineralization. Consequently, several gene families associated with silica metabolism have been characterized, such as silaffins and silacidins (Poulsen & Kröger, 2004; Wenzl et al., 2008; Wieneke et al., 2011; Kotzsch et al., 2017). Silica-associated proteins that are encoded by these genes may be responsible for the differences in silica morphology between species, and comparative analyses or gene knockout/knockdown studies have been performed to demonstrate this link (Görlich et al., 2019; Skeffington et al., 2022). HTS data are a valuable resource in this sense because they also allow the identification of orthologs of such morphologically or ecologically important genes in non-model organisms which can further be used as markers for a range of phylogenetic utility, as demonstrated for plants (Li et al., 2017; Valderrama et al., 2018; Choi et al., 2019) and microbial eukaryotes (Tekle & Wood, 2018).

1.5 Application of molecular methods to diatom biomonitoring

In the last decade, the application of DNA/RNA sequencing methods to environmental samples (e.g. soil, sediment, or water) revealed critical information on biodiversity and provided a complementary framework for biomonitoring applications. These methods usually involve assigning taxonomic names to DNA/RNA sequences obtained from bulk environmental samples through comparisons with public or local databases. They are fast, efficient, and cost-effective, and their sample collection is simple and non-destructive which are critical advantages over conventional morphology-based methods for aquatic biomonitoring (Sigsgaard et al., 2015; Qu & Stewart, 2019). Therefore, these methods have been successfully used to detect overall changes in community composition for a broad taxonomic range of organisms (Groendahl et al., 2017; Thomsen & Sigsgaard, 2019) and to design management strategies based on invasive, rare, and cryptic species (Scriver et al., 2015a; Levi et al., 2019; Qu & Stewart, 2019). However, several different natural processes can influence the composition and quantity of detectable DNA in environmental samples, and current literature highlights a need for further research on DNA shedding, transport, and degradation rates in aquatic environments, especially for studies focusing on plants (Thomsen & Willerslev, 2015; Barnes & Turner, 2016).

Most of the currently available molecular methods to sequence DNA/RNA in environmental samples are targeted approaches, which amplify taxonomically informative marker genes (e.g. barcoding markers). Due to the availability of quantitative PCR-based methods, some of these approaches are suitable for the quantification of species abundances or biomass (e.g. qPCR and ddPCR). However, these quantitative methods are not HTS-based and only a few species can be analyzed in a single run, so they provide limited scalability. Metabarcoding, on the other hand, is an HTS-based targeted approach that allows the detection of a large number of species including those that are rare or have low biomass (Alsos et al., 2018). In the last decade, it has been frequently used for assessments of biodiversity and ecological impact (Yu, D. W. et al., 2012; Bik et al., 2012), and in studies on ecosystem dynamics and identification of invasive species (Bott et al., 2010; Andersson et al., 2010). For diatoms, markers that are commonly used for systematic studies are usually preferred in metabarcoding studies as well (i.e., portions of rDNA, *cox*1 and *rbc*L genes, see section 1.3).

In early metabarcoding studies, Amplicon Sequence Variants (ASVs), which represent the set of haplotypes in a sample, were always grouped together into Operational Taxonomic Units (OTUs) that were treated as proxies for species. Typically, a sequence similarity threshold (SST) is used for clustering. However, major concerns have been highlighted about such rigid threshold approaches because they always carry the risk of separating groups of sequences that originate from species with similar ecological preferences or merging those that originate from species with distinct preferences (Meyer & Paulay, 2005; Tapolczai et al., 2019). Recently, a methodological shift has occurred favoring the assignment of ASVs directly to taxa instead of using OTUs. Nevertheless, this assignment is also typically done using SSTs. This is a major challenge for diatom metabarcoding because shifts in diversification rates are expected to be common between genes of different diatom species due to their relatively long evolutionary history and vast diversity (Nakov et al., 2018b). Therefore, using a minimal SST for detecting species-level differentiation in diatoms has been considered optimistic at the very least (Pinseel et al., 2019). Similar issues related to the limited discriminatory power of single or multiple gene surveys have also been highlighted in other eukaryotic metabarcoding studies (Fraser et al., 2018; DiBattista et al., 2019).

In addition to the limitations related to the marker genes, there are additional challenges in the application of targeted sequencing approaches to environmental samples, such as reference database incompleteness and variations in gene copy numbers and DNA content (Prokopowich et al., 2003; Zimmermann et al., 2015; Groendahl et al., 2017; Pérez-Burillo et al., 2020). These are critical challenges for the quantification of diatom species because there is an estimated 1000-fold of rRNA gene copy number variation (Créach et al., 2006) and a 5000-fold range of DNA content variation per cell (Cavalier-Smith, 1978) in unicellular eukaryotic algae. Previous studies employing metabarcoding methods for the calculation of diatom indices highlighted several problematic species that cause significant abundance discrepancies, possibly owing to such issues (Vasselon et al., 2018; Bailet et al., 2019; Mora et al., 2019). Nevertheless, diatom-based indices used in European countries for freshwater quality monitoring require the identification and quantification of hundreds of species which include many morphologically similar and/or phylogenetically close taxa (Cemagref, 1982; Van Dam et al., 1994; Coste et al., 2009).

1.6 Genome-scale approaches in biomonitoring

Although it has been demonstrated that it might be possible to obtain more reliable abundance estimates by making methodological adjustments to the metabarcoding protocol, (e.g. by including mock multispecies mixtures in known relative abundances, Matesanz et al., 2019), the lack of resolution of the phylogenetic markers remains an important challenge for diatom metabarcoding. Therefore, non-targeted approaches that amplify random DNA fragments from the nuclear and/or organellar genomes can be more effective to delimit diatom species in biomonitoring applications. Although such quantitative, non-targeted sequencing approaches have not been extensively tested for mixed samples, there are a few promising exploratory studies on plants (Peel et al., 2019; Wagemaker et al., 2021). These methods do not require a reference genome and provide highly improved phylogenetic resolution and more accurate relative abundance estimates, so they might have great utility for diatom biomonitoring.

When using diatoms as environmental indicators, the accuracy of demonstrated relationships in terms of species ecology is critical where closely related groups are expected to have similar trait values and habitat preferences (Carew et al., 2011; Keck et al., 2016). Although some studies found a link between morphological and ecological differentiation (Potapova & Hamilton, 2007; Kulichová & Fialová, 2016), the cases of cryptic species show that it is not always possible to distinguish species that differ in their growth responses to environmental factors using morphology (Wood & Leatham, 1992; Mann, 1999; Kelly et al., 2015; Pérez-Burillo et al., 2021). These cases might have significant implications for biomonitoring. For example, if a cryptic species pair with contrasting hypoxia tolerances occur in the same locality and are used as a bioindicator, an increase in the area of the anoxic bottom could remain undetected because more sensitive species will be replaced with the tolerant species without a change in the overall abundance estimates obtained by morphology-based methods (Bourlat, 2016). Similarly, if such a cryptic species occurs in different localities, the same human impact can affect species abundances at these localities differently. The quantification of relative abundances is also critical in this sense because different localities may have different proportions of the morphs with different tolerances. From an ecological view, understanding the distribution and dispersal of these populations with different eco-physiological characteristics would also provide insights into the settlement mechanisms, adaptation, and evolution of these species (Kim et al., 2017). Therefore, sampling many unlinked nuclear and/or organellar loci can also provide a more robust framework for investigating the functional, ecological, or biogeographic relationships among bioindicator diatoms.

1.7 Study design and the use of the *Nitzschia* genus

The genus *Nitzschia* sect. *Lanceolatae* has been central to the discussions on species concepts in diatoms because biological species could not be distinguished due to the continuous variation in the morphological characters within the group, suggesting extensive gene flow (Bonik & Lange-Bertalot, 1978; Mann, 2010). Moreover, taxonomic revisions of *Nitzschia* were primarily aimed to distinguish bioindicators of water quality because they are one of the most abundant diatoms in polluted waters and their morphological discrimination is usually problematic (Lange-Bertalot & Simonsen, 1978; Lange-Bertalot, 1980; Trobajo et al., 2009).

The classification of *Nitzschia* is largely based on cell shape and symmetry, the position of the raphe, extension of the fibulae, and folding of the valve. The cells usually are straight and generally have two plastids, one at each end of the cell. It is a relatively large and heterogenous genus split into several sections that comprise both freshwater and marine species. Currently, *Nitzschia* is the most speciose diatom genus with 899 taxonomically accepted species names in AlgaeBase (Guiry & Guiry, 2022), and recent analyses demonstrate the non-monophyly of the genus with species scattered over the entire Bacillariaceae tree, highlighting the need for further molecular systematic studies (Mann et al., 2021).

One *Nitzschia* species, *N. palea,* is highlighted as one of the most problematic species that cause significant abundance discrepancies in metabarcoding applications for diatom biomonitoring (Bailet et al., 2019; Mora et al., 2019). Furthermore, different varieties of *N. palea* have been described from waters with different pollution levels, but there is considerable overlap in their morphological characteristics, (i.e., in valve dimensions and pattern densities), and traditional barcoding markers do not differentiate these varieties, making the species a cryptic complex (Trobajo et al., 2009; Trobajo et al., 2010; Mora et al., 2019). Although *N. palea* is a common bioindicator species with global distribution, these issues are preventing a more complete understanding of the ecological niche differentiation within the species complex.

It is possible to obtain insights into the reproductive history of non-model species and resolve such cryptic cases through genome-scale analyses of many unlinked loci (see sections 1.3 and 1.4). However, such new knowledge on species boundaries consequently requires revisions in classification. The pace of these updates is rapid, especially in groups whose diversity is not fully recognized yet such as diatoms (Spaulding et al., 2021). Molecular methods are also increasingly being applied for biomonitoring applications that conventionally rely on the morphological identification of diatoms. Therefore, the advancements in molecular biology have a considerable impact both on diatom systematics and also on the applications that rely on it. Moreover, available molecular approaches for diatom identification and quantification have critical limitations and shortcomings (see sections 1.5 and 1.6). Genome-scale methods may be more effective tools for these quantitative applications because it is possible to obtain more reliable and accurate relative abundance estimates while distinguishing very closely related lineages, thus, informing diatom systematics as well.

1.8 Aims of this thesis

In this thesis, I aim to explore genome-scale methods for the identification and quantification of diatoms. I achieved this by (i) conducting a review of the literature on the use of DNA/RNA sequencing methods for aquatic biomonitoring applications and demonstrating their merits and limitations, (ii) inferring the evolutionary history of the non-model diatom, *Nitzschia palea*, using transcriptome data to inform diatom systematics, (iii) identifying homologs of silica genes in *N. palea* that can be used to obtain further insights into the ecology and evolutionary history of the species, and (iv) evaluating a genome-scale quantification method to estimate the relative abundances of conspecific strains in mock communities as an improved molecular approach for diatom biomonitoring.

1.9 Outline of the thesis

Chapter 1: General Introduction

This chapter provides general information on the biology, systematics, and ecological importance of diatoms in the context of freshwater biomonitoring with a specific focus on cryptic species, and it outlines the potential implications of genome-scale data in bridging the gap between diatom systematics and biomonitoring. The major research questions, objectives, and content of the thesis are outlined.

Chapter 2: DNA from water

This chapter provides a review of the use of DNA/RNA sequencing methods for aquatic biomonitoring applications and outlines the main strategies and key points for their experimental design. The literature survey shows that a very small proportion of these studies focus on aquatic macrophytes, possibly owing to issues with universal amplification and discriminatory power of single or multiple gene surveys, which also applies to diatoms. Moreover, the calculation of ecological indices usually requires obtaining relative abundance estimates for a large number of species, but currently used molecular methods have critical limitations in this sense. Therefore, we conclude that currently used DNA-based methods for aquatic biomonitoring of plants are still best coupled with conventional surveys.

Chapter 3: Phylotranscriptomics reveals the reticulate evolutionary history of a widespread diatom species complex

This chapter aims to obtain insights into the evolutionary and reproductive history of a cryptic diatom species complex, *Nitzschia palea*, using genome-scale data. Phylogenomic analyses of 183 unlinked nuclear loci revealed that *N. palea* is a recently diverged species complex, and there is a recent gene flow between clades with different morphologies with a resulting putative hybrid. This study is one of a growing number of phylogenomic studies in nonmodel organisms that has greatly increased the amount of phylogenetic signal compared to the use of few standard markers and revealed patterns of gene-tree discordance due to biological phenomena such as ILS, hybridization, and polyploidy with important implications for diatom taxonomy.

Chapter 4: Silica cell-wall associated proteins in *Nitzschia palea* mined from transcriptome data

This chapter aims to identify phylogenetic markers in *N. palea* that function in the unique silica metabolism of diatoms using the dataset generated in chapter two of this thesis and publicly available diatom genomes. We identified putative orthologous sequences for several known silica cell wall proteins of diatoms, one of which is specific to *N. palea*. Our results show that HTS-based approaches are a valuable resource to identify markers that can resolve cryptic species diversity and can further be used to test hypotheses on the ecology and evolutionary history of bioindicator diatom species.

Chapter 5: Genotyping by sequencing for estimating relative abundances of diatom taxa in mock communities

This chapter evaluates a genome-scale sequencing approach for the identification and quantification of diatom taxa in mock communities for freshwater biomonitoring. We obtained a highly improved taxonomic resolution compared to surveys targeting a few phylogenetic markers and more accurate relative abundance estimates compared to conventional light microscopic surveys. Our approach eliminates several limitations of single gene surveys in biomonitoring applications, such as gene copy number variation, the lack of phylogenetic resolution, and upscaling potential. Therefore, we highlight the methodological advantages of the method and outline its potential implications for biomonitoring.

Chapter 6: General discussion

This chapter synthesizes the principal findings of this thesis. It emphasizes the conceptual links between diatom systematics and biomonitoring and outlines the potential implications of the use of genome-scale data in these fields concerning the results obtained in the previous chapters.