SORBONNE UNIVERSITÉ

ECOLE DOCTORALE 227

SCIENCES DE LA NATURE ET DE L'HOMME

UMI 3614 Biologie évolutive et écologie des algues - Station Biologique de Roscoff

Equipe DYNECO Pelagos - Ifremer centre de Bretagne

From gene expression to genetic adaptation : insights into the spatio-temporal dynamics of *Alexandrium minutum* species complex

Auteur:
Gabriel METEGNIER

Sous la direction de :
Pr. Christophe DESTOMBE
Dr. Mickael LE GAC

THÈSE DE DOCTORAT D'EVOLUTION ET ECOLOGIE
DANS LE BUT D'OBTENIR LE GRADE DE DOCTEUR DE SORBONNE UNIVERSITÉ

Présentée et soutenue publiquement le 29 Octobre 2018, devant un jury composé de :

Pr. Allan CEMBELLA, Institut Alfred Wegener	Rapporteur
Dr. Mariella FERRANTE, Station Zoologique Anton Dohrn	Rapportrice
Pr. Eric THIÉBAUT, CNRS - Station Biologique de Roscoff	Examinateur
Pr. Philippe VANDENKOORNHUYSE, CNRS - Université de Rennes 1	Examinateur
Pr. Christophe DESTOMBE, CNRS - Station Biologique de Roscoff	Directeur de thèse
Dr. Mickael LE GAC, Ifremer - Centre de Bretagne	Co-directeur de thèse

Remerciements

En premier lieu, je souhaite adresser mes plus sincères remerciements à mes deux directeurs de thèse : Mickael Le Gac et Christophe Destombe. Tout d'abord, merci de m'avoir fait confiance il y'a trois ans, moi qui ne connaissait presque rien au milieu marin mais que vous avez réussi à faire rêver avec votre algue microscopique (qu'on attend toujours dans la rade, soit dit en passant), et merci de votre confiance renouvelée à chaque instant tout au long de ce projet. Plus que de la supervision, votre encadrement a été une réelle collaboration, et pour moi exactement la vision que je me faisais de la recherche : une réflexion commune, nourrie de nos sensibilités respectives. Sachez que j'ai conscience de la chance que j'ai eu de vous avoir comme encadrants, et j'espère que vous garderez un aussi bon souvenir de moi que moi de vous ! Alors encore une fois, merci pour tout. Christophe pour ton insatiable curiosité, ta bonne humeur, et pour toutes ces stimulantes discussions : l'évolution, c'est génial, hein ?! Mickael, merci de ta patience sans faille, de n'avoir jamais arrêté de m'ouvrir les yeux, de ta confiance permanente, et de m'avoir toujours demandé "juste une toute petite dernière question".

Les résultats obtenus lors de cette thèse doivent aussi beaucoup aux discussions menées lors de mes différents comités de suivi, dont je remercie vivement les membres : Pierre Boudry, Erwan Corre (merci aussi pour tes conseils si utiles en bio-info), Laure Guillou, Gwenaël Piganeau et Frédérique Viard. Je tiens à adresser une pensée toute particulière pour toi Frédérique, qui a toujours eu à coeur de s'assurer que tout aille bien, sur tous les plans.

Bien évidemment, j'adresse mes plus sincères remerciements aux membres de mon jury de thèse : notamment Mariella Ferrante et Allan Cembella, qui ont accepté le rôle de rapporteur ainsi qu'à Philippe Vandenkoornhuyse et Eric Thiébaut d'avoir accepté le rôle d'examinateur.

Bien sûr, il me faut aussi remercier les soutiens financier (50 % région Bretagne et 50 % Ifremer) qui m'ont permis de réaliser cette thèse au sein des laboratoires de l'UMI 3614 de la station biologique de Roscoff et de Dyneco-Pelagos de l'Ifremer centre de Bretagne. Merci à Myriam Valéro et Cédric Bacher pour leur accueil dans ces structures.

Dans chacun de ces environnements de travail j'ai fait de nombreuses rencontres, et je souhaiterai remercier tout particulièrement les équipes techniques, entre autres Julien, Sophie, Marie, Florian, Agnès pour leur bonne humeur et leur constante disponibilité. Merci aussi à Marc, Raffaele, Martin (mention spéciale pour le super coup de pouce sur la fin !), Annie, Luis et tant d'autres... Un merci tout particulier à Anne-Laure, toi la sauveuse, la magicienne de l'administration, qui fut en quelques sorte une maman, attentive, tout au long de ces trois années. Merci aussi à Francis pour ton insondable mine de connaissance sur la Bretagne, et pour nos discussions autour de rénovations, maçonneries, gestion des forêts, etc etc. Merci aussi aux autres membres présents ou passés de l'unité Dyneco, notamment Mickael V., Touria, Tristan, Kim, Tania, Auriane, Laurie, Sam (courage pour la suite !), ainsi qu'aux nouveaux arrivés : Stéphane, Gaspart, Aurore et Cécile. Merci Mlle Jauzein pour ta gentillesse et ton soutien. Ah et, une pensée toute particulière pour Amelia : merci pour le chauffage d'appoint qui nous a maintenu au chaud pendant la rude période hivernale (Enfin.. ces automnes-qui-durent-8-mois, pas évident hein, Pierre ?).

Ahh Pierre, qu'aurait été cette thèse sans toi comme colocataire de bureau ? Entre discussions musicales (musclées), diatribes philosophiques, calculs improbables et "tu veux voir le plus beau graph de tous les temps ?", j'ai le sentiment d'avoir gagné bien plus qu'un collègue. J'ai bien hâte d'aller grimper avec toi dans ton beau pays dont tu nous chantes tant les louanges ! Merci aussi pour toute cette période où tu as été d'un soutien sans faille, et que je n'oublierai jamais.

Et aussi bien sûr, merci aux deux vaillants Benthologues du bas : Thibault et Bastien, bon courage a vous pour la suite, tenez bon, et merci pour tous ces bons moments de décompression !

Merci aussi aux deux stagiaires de M2 que j'ai eu la chance de côtoyer, notamment Sauvann, qui m'a permis de tester mes capacité d'encadrement.

Je souhaite aussi remercier les copains de partout, qui ont toujours eu un mot sympa pour me permettre de rester la tête dans le guidon, notamment Jimmy, Camille, Anne-So, Michel & Sylvie, Caro et Quentin entre autres.

Merci à tous les copains du Dojo (je pense tout particulièrement à Virginie, Anthony, Nicolas, Tonyo, Franck, Jean-Louis...), pour leur bonne humeur et leur gentillesse. Tout particulièrement, merci à toi André, qui m'a initié sans compter à l'Art de la Paix. Tu n'imagines pas à quel point tes cours m'ont aidé.

Je remercie aussi ma famille, d'un soutien sans faille et d'une écoute permanente, à vous mes petites soeurs, qui grandissent si vite... J'adresse une pensée chaleureuse à vous qui n'êtes plus là, mais qui je le sais, seriez bien fier de leur 'petit'. Je souhaite aussi adresser une pensée toute particulière à mon frère Adrien, qui m'étonne chaque jour un peu plus par son mental d'acier, qui n'a d'égal que sa bienveillance.

Evidemment, je ne peux pas écrire ces remerciements sans te remercier, toi, qui te reconnaîtras. Merci pour tout.

"Soyez reconnaissant envers les difficultés, les retours en arrière;
L'échec est la clef du succès;
Chaque erreur nous apprend quelque chose."

Morihei UESHIBA
L'Art de la Paix.

Contents

List of Figures

List of Tables

Chapter 1

General introduction

1.1 Living in a changing environment

Throughout their life, species are subjected to numerous environmental fluctuations. As a consequence, different types of responses evolved by natural selection. Although many environmental changes are cyclic (such as seasons, tide or sunrise and sunset), other perturbations are unpredictable and sometimes rapidly appearing. In the same way that environmental perturbations occur within potentially wide action range, amplitude and rates, at both time and space scales, biotic responses run across many different process scales, from temporal intra-cellular biochemical buffering to longer term population-wide genetic adaptation (see figure 1.1; left side). Phenotypic plasticity is one of these possible responses. It is the ability of a single genotype to produce more than one phenotype under different environmental conditions (see review in Kelly, Panhuis, and Stoehr (2012). Therefore, observable traits in an individual (*i.e.* the phenotype, from Greek phainein, 'to show' and typos 'type') results from both the expression of the genotype, influence of the environment and interactions between the two. The role of phenotypic plasticity (through alterations of gene expression, *i.e.* molecular physiology) in survival to environmental changes can therefore be preponderant. On a middle to long term basis, gene expression levels can be modified by epigenetic modifications (typically referring at DNA methylations and histone modifications) that arise from environmental stimulations. Such modifications affect gene expression in a durable manner, with some epigenetic modifications being transgenerationally inherited (see reviews in Heard and Martienssen (2014) and Trerotola, Relli, Simeone, and Alberti (2015) for instance). Epigenetic modifications are one durable response to environmental stimuli through long term but reversible changes in gene expression, without DNA sequence modifications. Modulating gene expression levels is a temporary possibility to handle environmental perturbation, and such process is highly dynamic, easily reversible, and allows to quickly respond to environmental fluctuations in a short term basis. In this case, gene expression also has a long lasting potential, and can allow an individual to survive durable environmental changes (up to a generation).

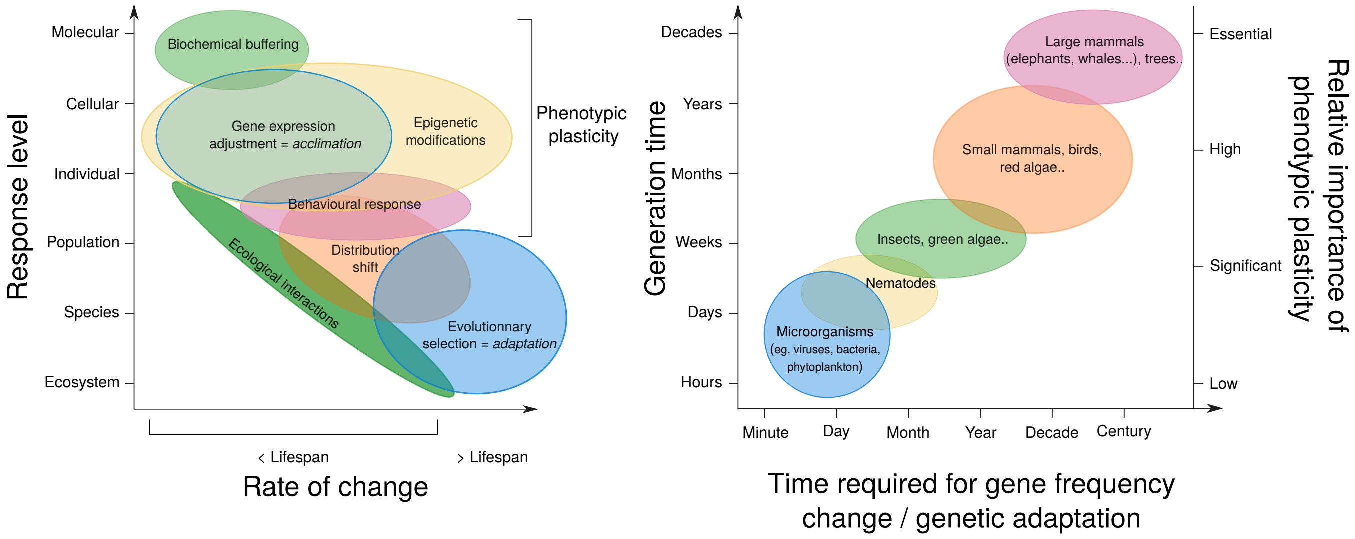
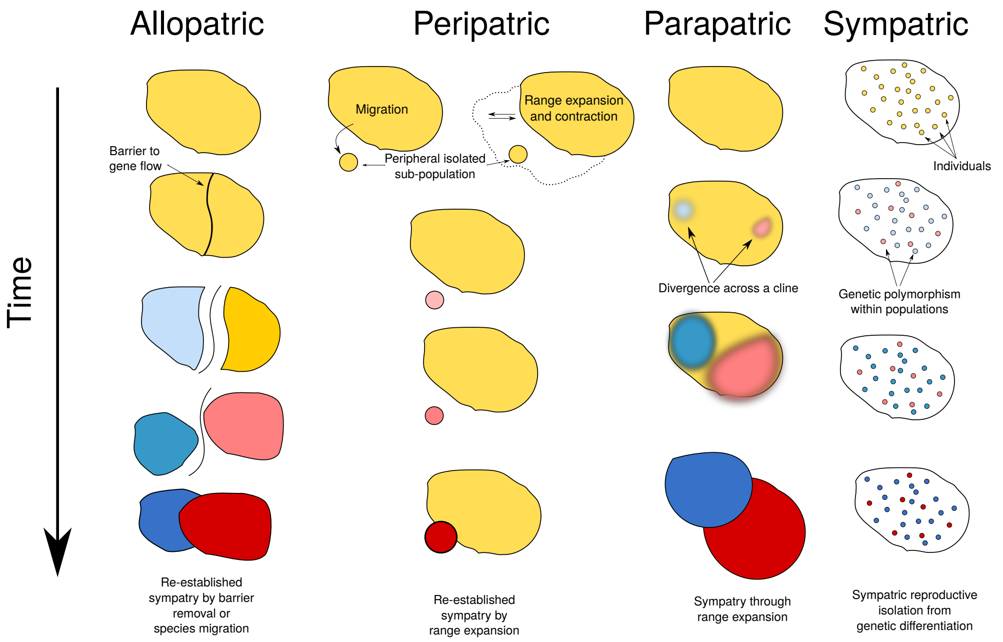


FIGURE 1.1: *Numerous response possibilities toward environmental changes were evolutionary selected, and act at different levels. Left : Schematic representation of the various biological responses at several levels (from molecular to ecosystem) in function of the duration of the experienced environmental perturbation. Right : Schematic representation of the trade-off between the relative importance of phenotypic plasticity (through gene expression levels acclimatization, right axis) and genetic adaptation (through populations-wide mutation fixation, x axis) in function of the generation time (left axis) of several groups of taxa. Inspired and modified from Peck (2011)*

Second type of response : genetic adaptation, which refers to the selection of advantageous mutations that increase in frequency in the populations. Genetic adaptation through allelic frequency changes at the population level is a long term process (it exceeds the lifetime of the individual). When organisms of a particular population undergo differential selective pressures (imposed by their local environment) as compared to other populations within their species, they experience local genetic adaptation (Kawecki & Ebert, 2004; Williams, 1966). This genetic differentiation between populations through the accumulation of mutations (neutral and/or locally beneficial) may, in some cases, lead to speciation (see box 1), that can be seen as an ultimate form of adaptive response to long term environmental conditions.

Box 1

Speciation processes According to the "biological concept", a species is defined as "groups of actually or potentially interbreeding natural populations that are reproductively isolated from other such groups" (Dobzhansky, 1937; Mayr, 1942, 1963). In this context, reproductive isolation can be rather pre-mating, post-mating–prezygotic or post-zygotic. The evolutionary mechanisms leading to such reproductive isolation are of primary interest and were extensively studied (see Coyne and Orr (2004) for a comprehensive description of the different mechanisms underlying speciation processes). Hereafter is schematized four main speciation processes : Allopatric (divergence following the population split into two geographical isolated subpopulations), Peripatric (divergence of two geographically isolated sub-populations following migration or range expansion / contraction), Parapatric (divergence through a gradient of phenotypic or genetic frequencies that are subjected to differential selection) and Sympatric (divergence within populations).


Allopatric
Peripatric
Parapatric
Sympatric
Time
Barrier to gene flow
Migration
Range expansion and contraction
Peripheral isolated sub-population
Individuals
Divergence across a cline
Genetic polymorphism within populations
Re-established sympatry by barrier removal or species migration
Re-established sympatry by range expansion
Sympatry through range expansion
Sympatric reproductive isolation from genetic differentiation

Of course, these different responses are not exclusive, and numerous investigations of the phenotypic plasticity - genetic evolution continuum in long term adaptation are now available (for instance through genetic assimilation; see box 2). Hereafter, we will focus on the two extreme side of this continuum : phenotypic plasticity and genetic adaptation.

Box 2

The genetic assimilation concept : If new environmental conditions last, the populations facing such change may experience an adaptation process, and the new phenotype expressed through phenotypic plasticity can be evolutionarily selected. This process is referred as "genetic assimilation" Pigliucci and Murren (2003); Schlichting, Pigliucci, et al. (1998); Shapiro (1976); Waddington (1953); WADDINGTON (1961). In a study in the early 1950's, Waddington (Waddington, 1953) formally introduced the "genetic assimilation" process of a character induced by an environmental stress in *Drosophila melanogaster*. Briefly, following the previous observation that a cross-veinless phenotype could be obtained at low frequencies when *D. melanogaster* pupae were subjected to heat stress at certain developmental stages, he artificially selected, at each generation, flies displaying this apparent phenotypic novelty (mimicking a natural selection processes) in order to raise their prevalence in the population. It was after as few as 16 generations that he observed what he referred as "genetic assimilation" : even without any heat shock treatment, the flies still displayed the cross-veinless phenotype. This illustrated how (relatively) long lasting selective pressure to new environmental conditions could shift the observed phenotype from being the result of temporary phenotypic plasticity to a genetically encoded adaptation. Although there are some experimental demonstrations of genetic assimilation (see for instance Suzuki and Nijhout (2006) in *Manduca sexta*, Ghalambor et al. (2015) in *Trinidadian guppies*, Van Tienderen (1990) in the grassland *Plantago lanceolata*, Sword (2002) in the grasshoppers *Schistocerca emarginata*), it remains poorly described in nature (but see Aubret and Shine (2009).

Investigating ecological divergence and its repercussions in terms of phenotypic plasticity and genetic adaptation are of primary interest to understand the mechanisms allowing populations to survive in their environment. Such insights can be achieved by answering questions as : how do populations respond to environmental fluctuations ? What is the interplay between the short term molecular physiological response and long term genetic adaptation ? And finally, to what extent does genetic divergence impact molecular physiology ? To investigate these questions, and as schematized in figure 1.1, long living organisms are unlikely to represent model species because their reproduction rates are too slow to allow the observation of genetic adaptation in action. With their large effective population sizes and high reproduction rates, microbial species constitute predisposed organisms to study responses to environmental perturbations at a time scale compatible with experimental surveys.

1.2 Microorganisms *in natura* : biodiversity and ecological importance

The existence of life forms too small to be seen was mentioned as early as the 6th century BC, and scarce scripts (On Agriculture by Varro - 1st century BC; The Canon of Medicine by Avicenna in 1020) discuss the implications of unseen creatures in disease spreadings. After a slowly increasing amount of observations, Antoni van Leeuwenhoek (1632 - 1723) opened the gate to the microorganism world with his homemade microscope. In 1878, the French surgeon Charles Sédillot proposes to regroup all the invisible living entities under the term 'microbe' (literally mikrós, « small » and bíos, « life »; Sedillot (1972). Microbiology was born and microorganisms started to be considered in the biodiversity landscape.

Nowadays, it is well known that the observable biodiversity does not truly reflect the extent of actual biodiversity, due to the estimated large amount of still undescribed microorganisms on earth (Kallmeyer, Pockalny, Adhikari, Smith, & D'Hondt, 2012; Locey & Lennon, 2016; Whitman, Coleman, & Wiebe, 1998). To date, two of the three domains of life (bacteria and archaea) proposed by Carl Woese and George Fox (Woese & Fox, 1977) are only composed of microorganisms, and the third domain of life (eukaryotic) is also composed of numerous microscopic species (some microscopic fungi and protists). The taxonomic diversity of microorganisms reflects their ubiquitous distribution, morphological diversity, trophic level positions (some are primary producers, grazers, parasites etc) and importance in ecological dynamics. Indeed, they are key players in planetary-scales biogeochemical fluxes by recycling the majority of the six fundamental elements that sustain life (Hydrogen, Carbon, Nitrogen, Oxygen, Sulfur, and Phosphate; Falkowski, Fenchel, and Delong (2008); Prosser (2012).

In this context, answering the questions of rather who is in a given community, what is the community doing and how it responds to environmental fluctuations in terms of functional activity are of primary interest, but are challenging. Numerous works attempted to investigate such questions for decades, in all kind of environments (Upton, Nedwell, and Wynn-Williams (1990) in lake sediment; Bobbie and White (1980) in benthic microbial community; Federle, Dobbins, Thornton-Manning, and Jones (1986) in subsurface soils for instance). However, the lack of microbial taxonomic knowledge, resource consuming available methods (such as strain cultivation in Upton et al. (1990) or using fatty-acid extractions from environmental samples in Federle et al. (1986) for instance) and furthermore the prevalence of unculturable microbial species in natural communities undoubtedly limited possibilities. Recently, however, the advances of "omic" technologies drastically pushed forward the studying of microbial community biology (in terms of community structure and functional activity) at multiple levels (see Muller, Glaab, May, Vlassis, and Wilmes (2013)).

1.3 *Who* and *What* : on the use of "omics" to study microbial ecology *in situ*

1.3.1 An "omic" approach for each informational level

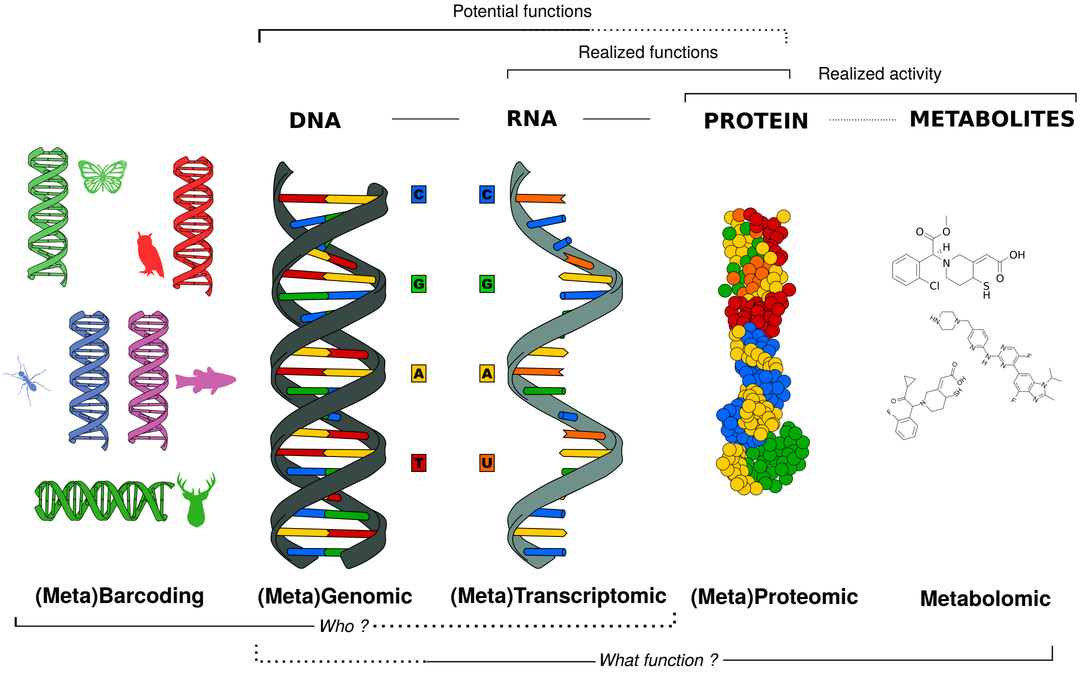


FIGURE 1.2: *Representation of several "-omic" methods used to ask several questions ("who" and "what", bottom), using various molecular entities that inform on different levels of bioogical functions or activity (upper part).*

In molecular biology, the use of the "-omics" suffix refers to the study of the totality of an entity, and encompasses a wide range of disciplines, from genomics to proteomics, transcriptomics, glycomics to lipidomics, etc (figure OMIC). Together with the preposition "Meta", methods from the “meta[...]omic” family aim at investigating molecular biology at the level of an entire community.

Metabarcoding and metagenomic studies typically ask the “*Who is there*” question, and allowed the extensive description of numerous communities from various environments (see for instance Mahé et al. (2017) in soils; Le Calvez, Burgaud, Mahé, Barbier, and Vandenkoornhuyse (2009); Xie et al. (2011) in deep-sea hydrothermal ecosystems; Delong et al. (2006) in seawater; Guarner and Malagelada (2003) in human gut; Pernice et al. (2016), Le Bescot et al. (2016) and Morard et al. (2018) in ocean-wide communities), greatly expanding our knowledge in environmental microbiology through extensive taxonomic diversity cataloging. By the genome wide scale of metagenomics, one can also study the functional potential of a given community, and therefore aims to answer the “*What the community could (potentially) do*” question (see for instance Tyson et al. (2004) in an acid mine drainage community or Warnecke et al. (2007) in a termite gut microbiota). “(Meta)transcriptomic” is the analysis of the expressed set of genes in a given sample and, contrary to metagenomic, gives access to the realized functions in an entire cell/organ/individual/community (see for instance Crovadore et al. (2017) in bacterial communities; Jiang, Xiong, Danska, and Parkinson (2016) in various microbial communities; Abu-Ali et al. (2018) in human microbial communities composition and activity; Hewson et al. (2014) in plankton communities). (Meta)proteomic and metabolomic studies are the forward steps in the identification of active functions in the given sample. While metabolomic aims to study the small molecules implicated in several critical functions (metabolic functions, signaling molecules..), metaproteomic studies have shown to successfully describe community wide metabolic activity (seeRam et al. (2005) in a microbial biofilm; Bargiela et al. (2015) in polluted sediments; Zampieri, Chiapello, Daghino, Bonfante, and Mello (2016) in soil). Together with metabolomic, metaproteomic face the common-garden limitation of respectively functional annotation and identification of protein and metabolic components. As reviewed in Muth, Benndorf, Reichl, Rapp, and Martens (2013), proteomic and metabolomic methods applied to entire communities face additional challenges such as the inherent difficulty to reliably extract proteins from complex biological assemblages or the correct taxonomic assignation of homologous proteins.

1.3.2 The transcriptomic level : a fair compromise ?

Schematically, molecular physiology relies on three cornerstones : in functions of the cell's needs, DNA (which carries the genetic information) is transcribed into RNA, which is modified (RNA maturation), transported outside the nucleus and then finally, translated into proteins. Messenger RNAs (mRNAs; molecules that convey the translated genetic information from DNA to the ribosome for protein traduction), can therefore be considered as a "proxy", linking gene expression responses to the functional need in protein. However, the validity of mRNA as a direct measure of protein level has been challenged and several authors underlined the need to consider it with caution. For instance, Gygi, Rochon, Franza, and Aebersold (1999) investigated the correlation between mRNA concentration and the respective abundance of more than 150 proteins in the yeast *Saccharomyces cerevisiae*, and showed very contrasting situations. For some genes, mRNA levels remained constant, but protein levels varied by more than 20 fold. Conversely, some mRNA levels were shown to vary by a 30 fold-change while no protein level variations were observed. Although their result were very pessimistic 20 years ago, numerous tools were developed since then, and now allow high-quality and accurate mRNA and protein levels detection (Y. Liu, Beyer, & Aebersold, 2016).

Among transcriptomic analysis, microarray and RNA sequencing (RNA-seq) are the two most common approaches to date. While microarray quantifies the expression of a predetermined set of sequences, RNA-seq relies on high-throughput sequencing to monitor all transcripts without any *a priori*.

The RNA-seq revolution

Nowadays, RNA sequencing allows to accurately quantify transcript abundance variations at a transcriptome-wide scale. In a recent review, Y. Liu et al. (2016) investigate how the relationship between protein and mRNA levels holds depending on the question asked, and discuss the powers and limitations of mRNA levels as an estimator of protein levels. Altogether, they conclude that we now dispose of sufficiently accurate methods and techniques to describe a cell molecular physiological state using mRNA levels (at least globally and while one is not trying to uncover highly dynamic and / or small time-scale protein levels adjustments). Transcriptomic analyses were successfully applied in a wide range of contexts, in order to monitor rather basal molecular physiological states or dynamic response to perturbations at both community scale (see for instance Poretsky et al. (2005) in marine and freshwater bacterioplankton; Poretsky et al. (2009) in marine community functional response change between day and night; Urich et al. (2008) in soil microbiota; Frias-Lopez et al. (2008) in ocean surface water microbial community; Gosalbes et al. (2011) in human gut microbiota), or intra and inter-species levels (see for instance Salazar-Jaramillo et al. (2017) in *Drosophila* spp. response to parasitoid wasp attack and Parkinson et al. (2016) who showed both inter and intra gene expression variations in the coral-associated dinoflagellate genus *Symbiodinium*; Han et al. (2017) in the petal development of the China rose *Rosa chinensis* or Coolen et al. (2016) in *Arabidopsis* transcriptome response to stresses). This approach can be of primary interest for studying inter-individual divergence in terms of molecular physiology, and can also be great allies to investigate the mechanisms underlying speciation processes and species adaptation capacities (see H. X. Xu et al. (2015) for an analysis of how two closely related whiteflies species transcriptionally respond to different host; Busby et al. (2011) in yeast species).

The possibility to take a blind approach : Unlike hybridization-based methods that are based on assumptions on a priori biologically relevant sequences, RNA-seq approach relies on the comprehensive analysis of the complete set of transcripts present in a given sample at a given time, and isn't therefore limited to the screening of transcripts with previously known genomic sequence. This makes RNA-seq particularly interesting when studying non model species, for which we typically lack this kind of information, and was extensively used for de novo transcriptome assembly (see for instance Vera et al. (2008) in the Glanville fritillary butterfly; Chen et al. (2010) in Locusts; Parchman, Geist, Grahnen, Benkman, and Buerkle (2010) in the Lodgepole pine; Marchant et al. (2015) in the blood-sucking bug *Triatoma brasiliensis* and so on).

Analyzing sequence variation : As RNA-seq relies on methodologies generating sequences of previously reverse transcribed complementary DNA (cDNA) fragments, they can be applied in a sequence variation screening purpose, such as SNP discovery, spatial and temporal genetic structuration analyses and genome wide association studies (*i.e.* hypothesis free methods for linking genetic markers with traits; see Romiguier et al. (2014); van Belleghem, Roelofs, van Houdt, and Hendrickx (2012); Takahagi et al. (2016) and Lopez-Maestre et al. (2016) for instance).

No (theoretical) detection limit : In microarrays, transcript abundance is quantified by visual detection of a concentration-dependent probe-target hybridization. Sophisticated methods are required for analysing microarray data, and many factors (including noise level control, physical damage of the microarray chip, normalization procedures, etc; see review in Jaksik, Iwanaszko, Rzeszowska-Wolny, and Kimmel (2015)) are known to challenge measures reliability. Also, microarrays suffer from known failure to detect both low (lower limit to signal detection) and highly (signal saturation) expressed genes. Although several methods were developed to circumvent such limitations, they add significant loads to the experimental procedure. As in RNA-seq transcript abundance is directly monitored through its relative abundance in the total extracted mRNA, such limitations do not hold and transcript quantification is much more straightforward (Marioni, Mason, Mane, Stephens, & Gilad, 2008; Z. Wang, Gerstein, & Snyder, 2009).

As every technical methodology, RNA-seq suffer from pitfalls (see box 3). As described in Conesa et al. (2016), the application range of RNA-seq is wide and constantly evolving, and therefore face multiple limitations and ambushes. However, using the right pipeline according to the right question allows to extract the best from this technology. Therefore, a growing amount of studies using an RNA-seq investigate gene expression variations and functional activity in a broad range of contexts (see examples above).

Box 3

Why not to fall too fast for RNA-seq : Some of the major pitfalls encountered in RNA seq, for in-deep review of RNA-seq limitations, see for instance Sendler, Johnson, and Krawetz (2011) and Z. Wang et al. (2009).

Longer transcripts are more likely to be called differentially expressed : As reviewed in Oshlack and Wakefield (2009), it appears that longer transcripts will generate more reads than shorter ones, leading to a better coverage of such transcripts, and positive correlation between transcript length and their probability of being called differentially expressed between analysed samples. Several methods were developed to circumvent this issue, including FPKM (Fragments Per Kilobase Million) and RPKM (Reads Per Kilobase Million) and TPM (Transcripts Per kilobase Million).

On the importance of sequencing depth : Assuming that, as gene length does not change from sample to sample, transcript-length effects on differential expression results between samples is low, and seems much less important than sequencing depth differences (Tarazona, García-Alcalde, Dopazo, Ferrer, & Conesa, 2011). In this context, approaches were developed (such as DEseq2 (Love, Huber, & Anders, 2014), edgeR (Robinson, Mccarthy, Chen, & Smyth, 2011) or limma (Ritchie et al., 2015) and were given much attention. Such methods use "raw counts" (*i.e.* raw numbers of reads aligning on each transcript) and rather focus on adjusting for library size differences.

Struggle in quantifying low expressed transcripts : In a comparative study of multiple methods for quantifying RNA splice variants, Mehta et al. (2016) illustrated the difficulties encountered by RNA-seq data for detecting expression levels of lowly expressed transcripts (with a restriction for studies with very good coverage of the targeted loci).

1.4 *Alexandrium minutum* : a model species among dinoflagellates ?

1.4.1 The dinoflagellates

Ecological significance

Dinoflagellates are a large group of eukaryotic unicellular organisms, with more than 4,000 described species (half being extinct species) among 500 living or extinct genera, and their diversity as high as their ubiquity (see for instance Lin, Zhang, Hou, Zhuang, and Miranda (2009); Stern et al. (2010); Mukherjee et al. (2015)). The recurrent discoveries of new species and community-wide metabarcoding studies suggest that dinoflagellate diversity is widely underestimated (Le Bescot et al., 2016). The extensive fossil records left by dinoflagellate species (one of the most important among microbial eukaryotes) is notably due to the rigid cellulosic plates that surround the cell of armoured dinoflagellates, and also to the ability of numerous species to form resting cysts (Sarjeant, 1974). These cysts are resistance forms that can be separated in two main types : temporary (typically formed when cells experience rapid and short term stresses) and resting cysts (generally formed by sexual reproduction and which allows a prolonged survival under harsh conditions). As dinoflagellate species are widespread, cysts can be found in virtually any sediment around the globe (see Zonneveld et al. (2013) for an extensive study on cysts distribution). Numerous dinoflagellate species are known for their ability to produce dense blooms following the rapid proliferation of their unicellular haploid cells, and potentially impact human activities when they form "Harmful Algal Blooms" (HAB). Dinoflagellates are among the most important group of toxin producing microalgal species (Cembella, 2003; Van Dolah, 2000; D. Z. Wang, 2008), and produce three major types of toxins : Diarrheic Shellfish Poisoning (DSP) caused by Okadaic Acid and Dinophysistoxins in species of *Prorocentrum* and *Dinophysis* genera for instance, Neurotoxic Shellfish Poisoning (NSP) caused by brevetoxins and Paralytic Shellfish Poisoning (PSP) mainly due to intoxication from saxitoxin consumption (Cembella, 2003). The later is produced by several species, such as *Alexandrium minutum*. Since 1970, the distribution, frequency and intensity of PSP events increased worldwide (as shown in figure 1.3), leading to significant social and economic concerns in impacted areas.

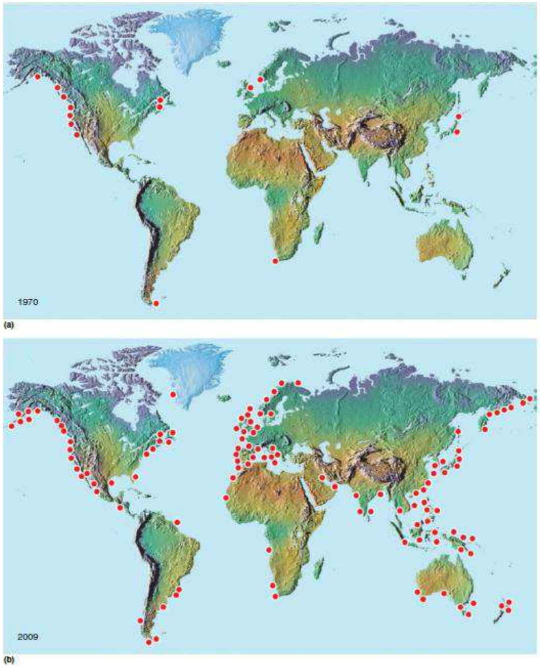


FIGURE 1.3: *Evolution of known presence of Paralytic Shellfish Poisoning events between 1970 (**a**) and 2009 (**b**). From Medlin and Cembella (2013).*

A phytoplankton bloom is globally defined as an important increase in cell density in a short period of time, and the understanding of such phenomenon is still evolving. First, such events were conceptually considered to be the result of the rapid division of few clonal strains that exponentially divide through mitotic division. However recent studies on dinoflagellate populations clearly indicate that blooms are composed of genetically different individuals, and that such genetic diversity changes through both time and space (Dia et al., 2014; Erdner, Richlen, McCauley, & Anderson, 2011; Lebret, Kritzberg, Figueroa, & Rengefors, 2012; Nagai et al., 2007; Rynearson, Newton, & Armbrust, 2006). Therefore, phytoplankton blooms can be considered as highly dynamic populations of genetically distinct individuals that divide rapidly until they reach high cell densities. As they are composed of rapidly dividing individuals in large effective population size, blooms have the potential to quickly adapt to selective pressures, leading to rapid changes in genetic diversity through both time and space.

Genetic particularities

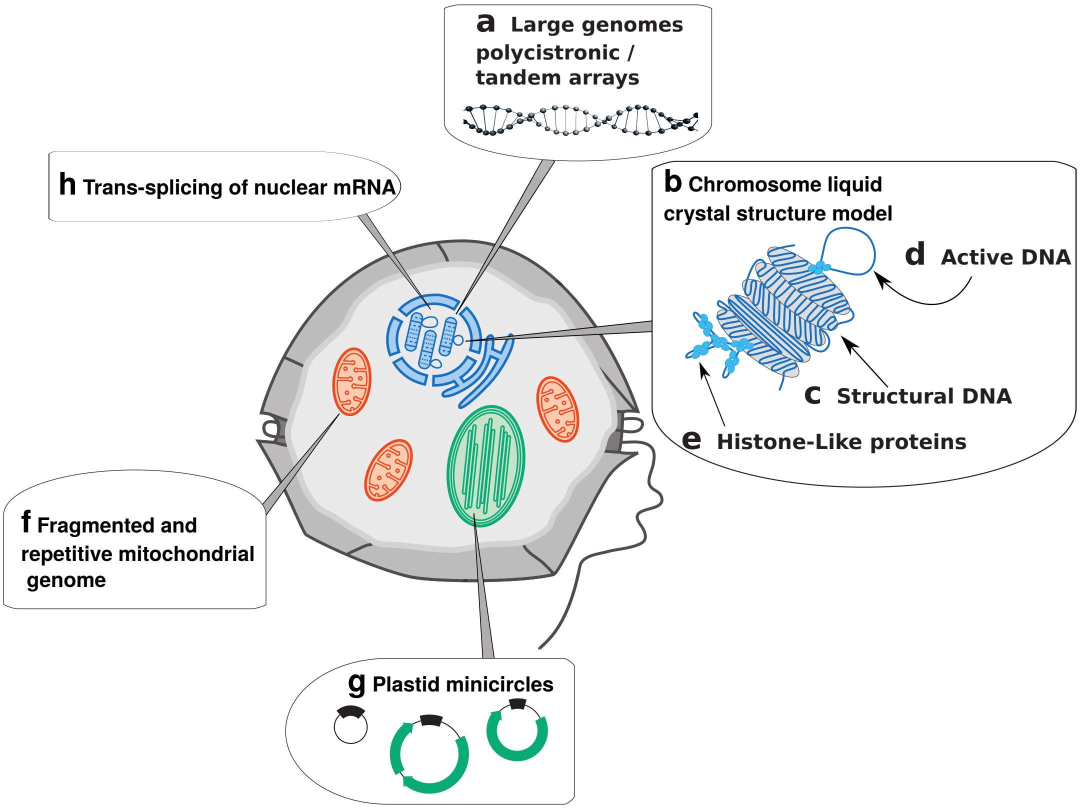

a Large genomes polycistronic / tandem arrays
h Trans-splicing of nuclear mRNA
b Chromosome liquid crystal structure model
d Active DNA
c Structural DNA
e Histone-Like proteins
f Fragmented and repetitive mitochondrial genome
g Plastid minicircles

FIGURE 1.4: *Few genomic characteristics of a typical dinoflagellate. Modified and extended from Wisecaver and Hackett (2011).*

Genetically, the dinoflagellates show remarkably uncommon characteristics for eukaryotic lineages, that, until molecular phylogenetics confirmed them as Alveolata, made them considered as mesokaryotes (an intermediate between eukaryotes and prokaryotes; Dodge (1965). Briefly, the main genomic features of dinoflagellates are : some of the largest eukaryotic genomes (figure 1.4 **a**) (from 1.5Gb in *Symbiodinium* to up to 185 Gb in *Lingulodinium polyedrum*; (LaJeunesse, Lambert, Andersen, Coffroth, & Galbraith, 2005; Spector, 1984)), tremendous chromosome copy number variations (from 24 to 220 ; (Sigee, 1986; Wisecaver & Hackett, 2011) and reference therein) that are (for most of the dinoflagellates), permanently condensed into liquid crystals (figure 1.4 **b**) (see reviews in RIZZO (1991, 2003); Wisecaver and Hackett (2011). Along their numerous chromosomes, genes seem to be arranged in multiple copies into polycistronic or tandem arrays (Bachvaroff & Place, 2008; Lin, 2011; Moreno Díaz de la Espina, Alverca, Cuadrado, & Franca, 2005). In these species, DNA is thought to be present in two distinct forms : (i) the structural DNA (figure 1.4 **c**), which is genetically inactive but crucial for chromosome stabilization (it is condensed into liquid crystals and therefore likely too dense to allow transcription) and (ii), the genetically active DNA (figure 1.4 **d**), in the form of uncondensed, extended extrachromosomal loops (Sigee, 1986). The accessibility of these DNA loops are thought to be permitted, in a concentration-dependant manner, by Histone-Like Proteins (figure 1.4 **e**) (Chan & Wong, 2007; Sala-Rovira et al., 1991). Until recently, dinoflagellates were thought to completely lack histones, however, histone transcripts were identified in various species (Hackett et al., 2005; Lin, 2011; Okamoto & Hastings, 2003). Some authors suggest that previous research have overlooked histones in dinoflagellates either because of their very low expression and/or stage specific presence (*e.g.* in cysts; see discussion in Lin (2011); Wisecaver and Hackett (2011). The large nuclear genome of dinoflagellate contrasts with their organellar genomes : respectively only 3 and 16 genes coding for proteins in the mitochondria (only 2 in *Oxyrrhis marina* due to the fusion of *cob* and *cox3* genes) and in the most common plastid respectively (see Howe, Nisbet, and Barbrook (2008); Waller and Jackson (2009)). Organellar genomes organization in dinoflagellates reflect their highly dynamic and particular nuclear genomic characteristics. Indeed, their mitochondrial genome displays experienced numerous fragmentations, re-arrangement and important reduction (figure 1.4 **f**) (Jackson et al., 2007; Nash et al., 2007), and the genomes of their plastid also shows high rearrangement into minicircles, that encode only one or few genes (figure 1.4 **g**; Koumandou, Nisbet, Barbrook, and Howe (2004); Z. Zhang, Green, and Cavalier-Smith (1999). Furthermore, recurrent horizontal gene transfer events were shown to have profoundly impacted dinoflagellate genomes. For instance, the genes coding for the histone-like proteins described previously were shown to have been acquired by horizontal gene transfer (movement of DNA between organisms) from proteobacteria (Hackett et al., 2005; Wong, New, & Hung, 2003). Finally, dinoflagellates also display uncommon transcriptional particularities, that are exposed in the next subsection.

Transcriptional particularities

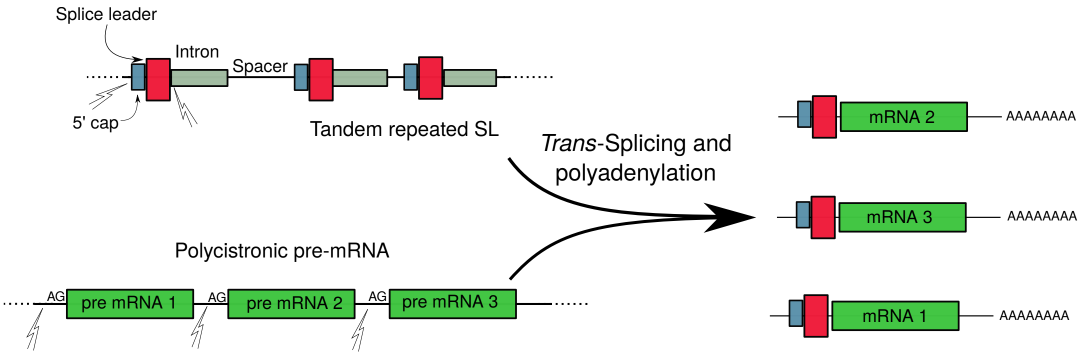


FIGURE 1.5: *Representation scheme of the mRNA maturation process in dinoflagellates.*

In eukaryotes, several steps are required to transform pre-mRNA into mature mRNA : typically, these are addition of a cap structure at the 5′ end and of a poly-A tail at the 3′ end (that are both required for nuclear mRNA export and translation efficiency enhancement; Baum et al. (1975); Bechler (1997); Hamm and Mattaj (1990); Shatkin (1976). Finally, maturation ends by RNA splicing (which ends pre-mRNAs maturation by removing introns when present) and joining together exons from a single pre-mRNA into mature, functional mRNAs. This type of RNA splicing is called *cis*-splicing. Unlike in the vast majority of eukaryotes where *cis*-splicing processes a single pre-mRNA into mature mRNA, some of them use a *trans*-splicing RNA maturation mechanism (see figure 1.5), which refers to the maturation of pre-mRNA through the joining of independently-generated primary RNA transcripts (Agabian, 1990; Bonen, 1993). Known in several groups of eukaryotes, and widespread in dinoflagellates, this pre-mRNA maturation process was first described in the trypanosome *Trypanosoma brucei* (Murphy, Watkins, & Agabian, 1986; Sutton & Boothroyd, 1986), and has been extensively studied in this species (Agabian, 1990; Huang & van der Ploeg, 1991; Laird, 1989; Ullu, Matthews, & Tschudi, 1993), see Preußer, Jaé, and Bindereif (2012) for a review). "Spliced Leader *trans*-splicing" refers to the slicing of a short (< 50 nt) non coding RNA sequence (the "Spliced Leader") to the 5′ end of an independently transcribed pre-mRNA. Following this, polycistronic transcripts turn into mature, translatable, monocistronic mRNAs (H. Zhang et al., 2007). H. Zhang et al. (2007), characterized dinoflagellate Spliced Leader as a

22 nt long sequence located at the beginning of the cDNA 5′ UTRs (DCCGTAGCCATTTTGGCTCAAG [D = T, A, or G]) (figure 1.4 **h**).

Box 4

Gene expression monitoring through RNA-seq in SL-*Trans*-splicing species Due to the uncommon pre-mRNA maturation process encountered in dinoflagellates, interrogations toward the ability to monitor gene expression dynamics with RNA-seq methods are common. If genes are constitutively transcribed with protein levels being only post-transcriptionally regulated, such interrogations may be legitimate, and further emphasized by the discrete mRNA level variations observed in dinoflagellates compared to other groups of protist for instance (see Alexander, Jenkins, Rynearson, and Dyhrman (2015). However, as the SL *trans*-splicing mRNA maturation process includes the addition of a poly-A tail, and given that the RNA-seq library preparation includes a rRNA removing step by poly-A selection of mature mRNA, monitoring expression of mature mRNA should be possible in SL-*trans*-splicing species in general, and in dinoflagellates in particular. We note that the monitoring of mRNA levels through mRNA sequencing do not specifically reflects gene expression levels in SL-*trans*-splicing species but rather to post-transcriptional mRNA levels, which may even be a better indicator of the physiological status of the cells for such organisms. However, hereafter, "gene expression" will be used for reading clarity.

1.4.2 *Alexandrium minutum*

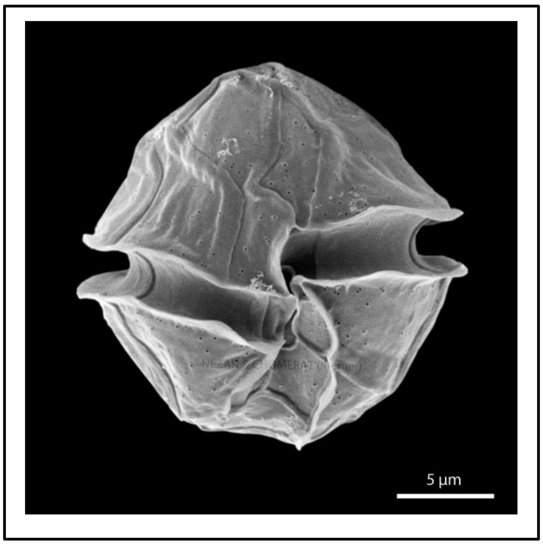


FIGURE 1.6: Alexandrium minutum *under Scanning Electron Microscopy. Note that the two flagella are absent here.*

Alexandrium minutum Halim 1960 is a spherical to ellipsoidal microscopic algae (15 - 30 μm long and 13 - 24 wide) that produces saxitoxin, a Paralytic Shellfish Poisoning (PSP) toxin. Composed of rigid cellulosic plates, *A. minutum*'s protective theca is divided into two parts : the epicone (upper part) and the hypocone (lower part), that are separated by two central furrows (cingulum) where flagella are inserted. Exclusively marine, this species has however been shown to tolerate rather moderate salinities and can therefore be found in upper parts of the estuaries, as long as waters are nutrient rich (see for instance Grzebyk et al. (2003) in a Breton strain; Vila et al. (2005) in a Mediterranean strain; Lim et al. (2011) in a Vietnamese strain).

Worldwide situation

The number of recorded *A. minutum* HAB events increased worldwide in the past decades (Anderson (1989); Hallegraeff, Bolch, Blackburn, and Oshima (1991); Lilly, Halanych, and Anderson (2005, 2007); see figure DISTRIBUTION). Since its first detection in Alexandria harbor, Egypt (Halim, 1960), this species is nowadays observed in Spain (Franco, Fernández, & Reguera, 1994), France (Belin, 1993), North Sea (Elbrachter, 1998; Hansen, Daugbjerg, & Franco, 2003; Nehring, 1998), Ireland (Gross, 1989), Sweden (Persson, Godhe, & Karlson, 2000), Taiwan (Hwang & Lu, 2000), New Zealand (F. H. Chang et al., 1999) and so on. Along the European coasts, *A. minutum* has even been classified as an invasive species. Anderson (1989) proposed several factors contributing to *A. minutum* expansion worldwide. Despite that improvement in toxin detection methods and intensifying HAB events monitoring programs undoubtedly favoured this species detection, Anderson (1989) noted that shellfish farming might ease bloom development, notably by the local nutrient enrichment induced by food intake and biological excretions from such cultures. More generally, global increases in nutrient loading, together with natural dispersal and human-facilitated transport, are the main factors thought to be responsible for the worldwide distribution of *Alexandrium* species nowadays (Anderson, 1989; Hallegraeff et al., 1991; Vila et al., 2005 (in *A. catanella*); Persich, Kulis, Lilly, Anderson, & Garcia, 2006 (in *A. tamarense*)).

French situation

The presence of *A. minutum* along the French coasts was recorded since mid 1980's (Belin, 1993; Lassus & Bardouil, 1988, in the bay of Vilaine)). In the bay of Brest in particular, its presence is recorded since 1990 by the REPHY network (Réseau de surveillance et d'observation du Phytoplancton et des Phycotoxines) and until the recent years, *A. minutum* densities were relatively low (few thousands cells.L^{-1}), but they drastically increased, reaching up to 42 million cells per liter during 2012 summer. High concentrations of HAB induce toxin bio-accumulation in shellfish and then sometimes leads to shellfish farms closure. These recent events led to the classification of the bay of Brest as a high-risk zone for *A. minutum* blooms (Chapelle et al., 2015a).

Life cycle

A. minutum has a heteromorphic life cycle, and alternates both planktonic and benthic stages (see figure 1.7). The benthic phase is composed of cells resting in the sediment into resistant cysts, which germinates into vegetative free living cells after excystment. This particular life cycle, with the alternation of two main different forms (resistant cysts vs vegetative, free-living cells), results in the absence of *A. minutum* from the water column for more than two third of the year, followed by a rapid growth (sometimes in very important proportions - more than 40 million cells.L^{-1} in the bay of Brest in 2012) and then a disappearance until the next blooming season. Following the recruitment of cells through excystment, the asexual division of vegetative cells and gametic cells fusion form altogether what is called a bloom.

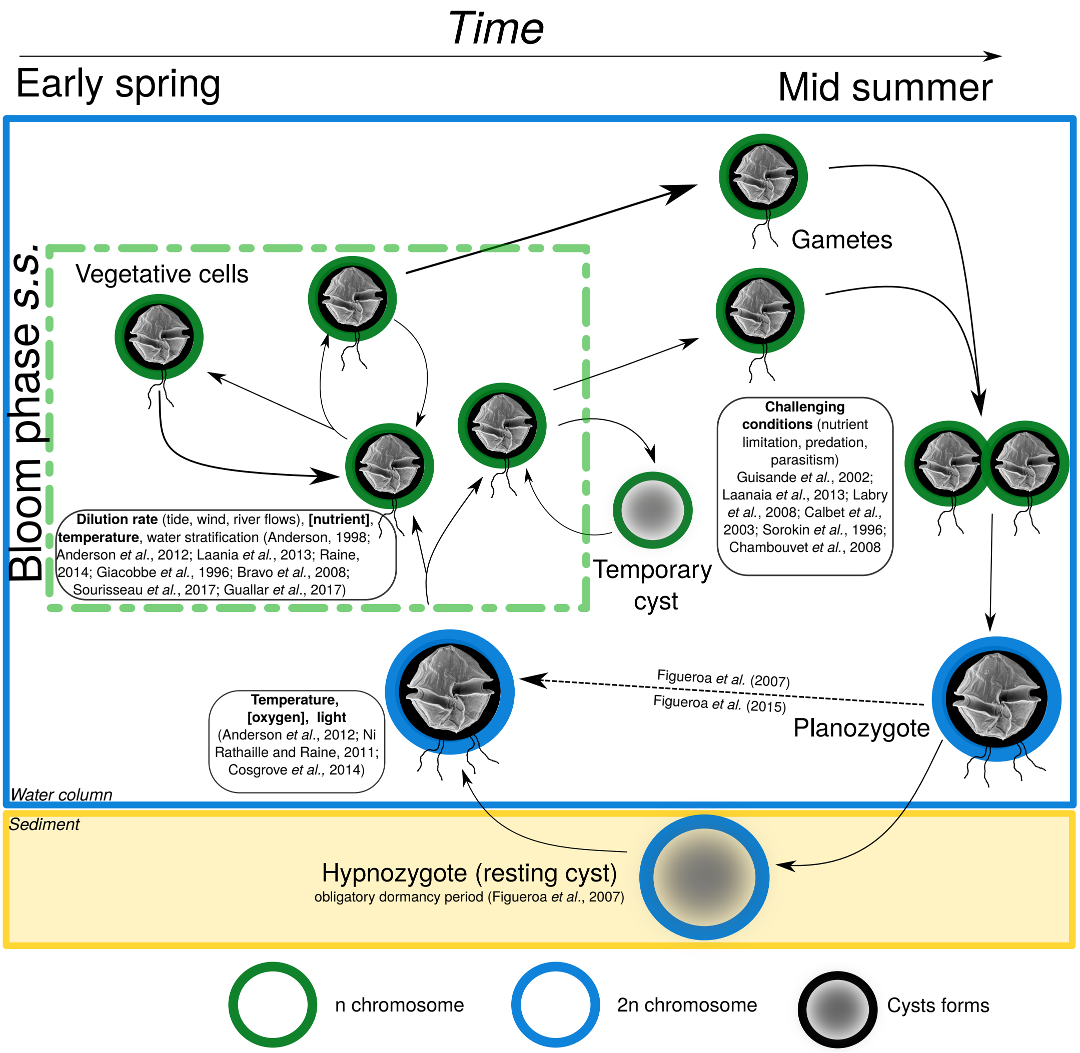

Time
Early spring
Mid summer
Bloom phase s.s.
Vegetative cells
Gametes
Challenging conditions (nutrient limitation, predation, parasitism) Guisande et al., 2002; Laanaia et al., 2013; Labry et al., 2008; Calbet et al., 2003; Sorokin et al., 1996; Chambouvet et al., 2008
Dilution rate (tide, wind, river flows), [nutrient], temperature, water stratification (Anderson, 1998; Anderson et al., 2012; Laania et al., 2013; Raine, 2014; Giacobbe et al., 1996; Bravo et al., 2008; Sourisseau et al., 2017; Guallar et al., 2017)
Temporary cyst
Figueroa et al. (2007)
Figueroa et al. (2015)
Planozygote
Temperature, [oxygen], light (Anderson et al., 2012; Ni Rathaille and Raine, 2011; Cosgrove et al., 2014)
Water column
Sediment
Hypnozygote (resting cyst)
obligatory dormancy period (Figueroa et al., 2007)
n chromosome
2n chromosome
Cysts forms

FIGURE 1.7: *Scheme of the major stages experienced by* A. minutum *cells through their heteromorphic cycle. Major environmental factors influencing each transition are indicated with some references.*

Drivers of bloom initiation, growth and decline are not well understood. Initiation is supposed to begin through excystment of resting cells. This excystment is possible after a dormancy period (Figueroa, Garcés, & Bravo, 2007) and may be regulated by temperature, light and oxygen concentration (Anderson et al., 2012; Ní Rathaille & Raine, 2011). Cosgrove, Rathaille, and Raine (2014) showed that blooms of *A. minutum* in the Cork Harbour initiates after the first important spring tide with water column temperatures superior to 15°C. Recently, based on numerical modeling, Sourisseau, Le Guennec, Le Gland, Plus, and Chapelle (2017) showed that water temperature and dilution rate may be important factors regulating *A. minutum* bloom initiations. The importance of these factors was confirmed using phenology and threshold analysis (Guallar, Bacher, & Chapelle, 2017). Bloom development seems to be regulated mainly by flushing rate (characterized by both tide, wind and river flow) and water stratification (thermal and/or haline) (Anderson, 1998; Anderson et al., 2012; Guallar et al., 2017; Laanaia et al., 2013; Raine, 2014; Sourisseau et al., 2017). Highly correlated with river flows, nutrient concentration is a key factor that controls bloom dynamics trough growth rate (Chapelle et al., 2015a; Guallar et al., 2017; Maguer, Wafar, Madec, Morin, & Erard-Le Denn, 2004; Romero et al., 2013). Again, temperature is definitely a keystone in bloom development dynamics by regulating growth rate and bloom phase timing (Bravo, Vila, Masó, Figueroa, & Ramilo, 2008; Chapelle et al., 2015a; M. G. Giacobbe, Oliva, & Maimone, 1996; Guallar et al., 2017; Sourisseau et al., 2017). Bloom termination factors are certainly the less evident to isolate. Among possible processes involved, several studies pointed toward significant roles of nutrient limitations (Guisande, Frangópulos, Maneiro, Vergara, & Riveiro, 2002; Laanaia et al., 2013; Labry et al., 2008), predation (Calbet et al., 2003; Sorokin, Sorokin, & Ravagnan, 1996) - even if some studies show that grazers have limited impact on *A. minutum* blooms (Probert, 1999) - and parasitism (Chambouvet, Morin, Marie, & Guillou, 2008). During this bentho-pelagic life cycle, *A. minutum* undergoes various genetic states : it's assumed that, after a haploid stage as a pelagic vegetative cell mostly asexually proliferating (at the beginning of the bloom), gametic cells arise (somehow at the end of the blooming period, when the conditions are getting challenging) and fuse to give a mobile planozygote. This planozygote sinks toward the sediment and encysts as a hypnozygote. After a dormancy period of various time, cyst germinate as planomeiocysts. Meiosis is generally accepted to occur during of after the germination, when the planozygote divides into vegetative cells (Anderson, 1998; Brosnahan et al., 2015; Figueroa, Dapena, Bravo, & Cuadrado, 2015; Figueroa et al., 2007; Figueroa, Vazquez, Massanet, Murado, & Bravo, 2011; Garcés et al., 2004; Probert, 1999).

This global scheme has been refined at multiple levels, and is still conceptual. For instance, Figueroa et al. (2007) showed that *A. minutum*'s pelagic planozygotes can give vegetative cells without encystment. Based on experimental observations, Figueroa et al. (2015) showed that sexual reproduction in *A. minutum* occurs throughout the entire bloom while in other species such as in *A. tamarense*, Anglès, Garcés, Hattenrath-Lehmann, and Gobler (2012) showed that it rather occurs during the exponential phase of the bloom. Physiological dynamics of *A. minutum* were investigated both *in vitro* and *in situ*. For example, basic growth rates of *A. minutum* strains were estimated under various experimental conditions (temperature, salinity, nutrient depletion, etc; see for instance Frangópulos, Guisande, DeBlas, and Maneiro (2004); Grzebyk et al. (2003); Guisande et al. (2002); Lim et al. (2006); Ní Rathaille and Raine (2011) in excystment experiments). In the field, many studies were conducted to investigate the correlation between environmental factors with bloom growth and decline (Laanaia et al., 2013; Maguer et al., 2004; **?**). Although there are a growing amount of studies that investigated *A. minutum*'s life cycle transitions both under culture and natural conditions (see for instance the review in Anderson (1998); Anglès et al. (2012); Figueroa et al. (2015, 2007)), the way *A. minutum* molecular physiology changes through its life cycle and response to environmental stimuli is still unknown. Therefore, the global understanding of this species dynamics now hurts the critical step of an integrative, *in situ* analysis of the how *A. minutum* populations respond to the multiple environmental disturbances they face through their life cycle, both in terms of genetic adaptation and molecular physiology. The general aim of this work is to bring back together pieces of understanding on the molecular physiology and genetic diversity dynamics of this species in its natural environment, and therefore investigate the link between genetic, ecological and molecular physiology divergence.

1.5 Objectives of the Ph.D project

The "omic" revolution opened the gates to large scale, in situ and extensive description of microbial and functional diversity. During this Ph.D project, we took the advantages of one of these molecular advances to investigate several aspects around a global question : **from gene expression variation to genetic adaptation : what is the biological response to environmental fluctuations ?**

The present manuscript is divided in four main parts, each of them addressing a particular aspect of this problematic.

Chapter 2 is divided in two parts and aims to investigate **the gene expression divergence between two cryptic species living in the same environment.** Therefore, the first part presents the patterns and processes of divergence of two incipient *A. minutum* cryptic species (Le Gac et al., 2016). The second part of this first chapter is dedicated to investigate further the link between genetic and molecular physiology divergence between these two cryptic species. This work was submitted as a scientific publication in the Molecular Ecology journal (Metegnier, Destombe, Quéré, & Le Gac, 2018b *Submitted*).

In chapter 2, one population surveyed was shown to be only composed of one of the two cryptic species. Thanks to this, we were able to gain insights into its gene expression over a three year time scale and to ask the following question in **chapter 3 : what is the molecular physiology temporal response to environmental fluctuations ?** This chapter presents the transcriptional dynamic of *A. minutum* in its natural environment, and investigates the effect of abiotic environmental factors fluctuations on gene expression levels. The results we obtained were submitted as a second research paper to the Molecular Ecology journal (Metegnier, Destombe, Quéré, & Le Gac, 2018a *Submitted*).

Chapter 4 addresses another level of biological response, and aims to answer the following question : **how does intraspecific genetic diversity changes over space and time, and how do populations adapt to environmental fluctuations ?** Chapter 4 presents our attempt to characterize genetic differentiation of *A. minutum* blooms at both a small spatial scales and throughout three consecutive blooms. Also, we aim to isolate the environmental factors (both biotic and abiotic) explaining them. These analyses give encouraging preliminary results, offering great opportunities for further investigations.

Finally, in **chapter 5**, we took the advantages of our metatranscriptomic dataset to develop the methodology presented in chapter 3, and to investigate the molecular physiology dynamics at the community level and examine whether distantly related species belonging to the same community react similarly to environmental fluctuations. A manuscript is under preparation.

Chapter 2

Alexandrium minutum cryptic species in blooms : a spatio-temporal survey of both assemblage and gene expression divergence *in vitro* vs *in situ*

2.1 General context

Carl Von linné (1707 - 1778) proposed what was the first widely accepted scheme of biological diversity classification, consisting in seven hierarchical categories : Kingdom, Phylum, Class, Order, Family, Genus and Species. As the prevalent principle of creation at Linné's living time did not include species evolution, species were considered as constant and invariable. During the next century however, both Jean-Baptiste Lamarck (1744 - 1829), and more particularly Alfred Russel Wallace (1823 - 1913) and Charles Darwin (1809 - 1882) dramatically shakened biology by introducing the concept of species evolution under natural selection ("On The Origin of Species by means of Natural Selection", Darwin (1859)). To date, numerous definitions and concepts of the "species" unit are available (see reviews in Mayden (1997); Wheeler and Meier (2000) and reference therein), and more than defining species boundaries, they also address the questions relative to species divergence underlying processes (*i.e.* speciation). Speciation can be seen as a gradual continuum of reproductive barriers appearance between two diverging populations (de Queiroz, 2005). Reproductive barriers can gradually arise as a result of restricted gene flow between sub-populations, due to rather physical factors (allopatric speciation; habitat fragmentation for instance), or differential evolution of groups of individuals within a population (sympatric speciation; through different selection across an environmental cline for instance). While allopatric speciation is relatively intuitive, the mechanisms underlying sympatric speciation are much more arduous to isolate (Bolnick & Fitzpatrick, 2007; Coyne & Orr, 2004; box 3).

Although the concept of cryptic species (different species mis-classified as one due to lack of diagnostic morphological characters) is not new, molecular technologies recently shed a new light on their prevalence. Indeed, they are found in every phylum investigated (see for instance Oliver, Adams, Lee, Hutchinson, and Doughty (2009) in an Australian lizard; Pinceel, Jordaens, van Houtte, de Winter, and Backeljau (2004) in slug; Forlani, Tonini, Cruz, Zaher, and de Sá (2017) in a Brazilian frog; Cruz, Suárez, Kottke, and Piepenbring (2014) in a patch-forming fungi). Some marine species has also been shown to gather cryptic species (Colborn, Crabtree, Shaklee, Pfeiler, and Bowen (2007) in bonefishes; Meyer, Geller, and Paulay (2005) in gastropods; Bakker, Olsen, Stam, and van den Hoek (1992) in green algae; Leliaert et al. (2014); see reviews in Bickford et al. (2007) and Fišer, Robinson, and Malard (2018))). Several hypotheses were proposed to explain the lack of morphological divergence between cryptic species. Briefly, it is proposed that cryptic diversity occurs (i) in recently diverged species complex (where no morphological differentiation appeared yet, but see Fišer et al. (2018)), (ii) under strong selections forces on characters such as behaviors and physiology ("morphological stasis", Bickford et al. (2007); as in parasitic systems, see Schonrogge et al. (2002)), or (iii) from common responses of already divergent species toward similar selective pressures (morphological convergence).

One approach to study the mechanisms underlying population divergence and more particularly incipient speciation, is to investigate how such processes translate in terms of molecular physiology divergence.

The present chapter investigates **the gene expression divergence between two cryptic species living in the same environment**. First, it starts with a study of how an incipient speciation process traduces in terms of the molecular physiology and genetic divergence in *A. minutum* complex of cryptic species. The second part presents an in depth description of the gene expression divergence between these two cryptic species, and how they both adjust their gene expression levels when facing the same environmental conditions, therefore investigating the link between genetic and molecular physiology divergence.

2.2 Evolutionary processes and cellular functions underlying divergence in *Alexandrium minutum*

Author contributions : In this collaborative research project, I performed a part of the data analysis, mainly evolutionary model simulations, and I participated to manuscript writing and review.

MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 5129–5143 doi: 10.1111/mec.13815

Evolutionary processes and cellular functions underlying divergence in *Alexandrium minutum*

MICKAEL LE GAC,* GABRIEL METEGNIER,*† NICOLAS CHOMÉRAT,‡ PASCALE MALESTROIT,* JULIEN QUÉRÉ,* OLIVIER BOUCHEZ,§ RAFFAELE SIANO,* CHRISTOPHE DESTOMBE,† LAURE GUILLOU¶ and ANNIE CHAPELLE*

**Ifremer, DYNECO PELAGOS, 29280 Plouzané, France, †CNRS, PUCCh, UACH, UMI 3614, Evolutionary Biology and Ecology of Algae, Station Biologique de Roscoff, Université Pierre et Marie Curie – Paris 6, Sorbonne Universités, Place Georges Teissier, CS90074 29688, Roscoff Cedex, France, ‡Station de Biologie Marine, IFREMER, 29900 Concarneau, France, §GeT PlaGe, Genotoul, INRA Auzeville, Castanet Tolosan, France, ¶CNRS, UMR 7144, Station Biologique de Roscoff, Université Pierre et Marie Curie – Paris 6, Sorbonne Universités, Place Georges Teissier CS90074, 29688 Roscoff Cedex, France*

Abstract

Understanding divergence in the highly dispersive and seemingly homogeneous pelagic environment for organisms living as free drifters in the water column remains a challenge. Here, we analysed the transcriptome-wide mRNA sequences, as well as the morphology of 18 strains of *Alexandrium minutum*, a dinoflagellate responsible for harmful algal blooms worldwide, to investigate the functional bases of a divergence event. Analysis of the joint site frequency spectrum (JSFS) pointed towards an ancestral divergence in complete isolations followed by a secondary contact resulting in gene flow between the two diverging groups, but heterogeneous across sites. The sites displaying fixed SNPs were associated with a highly restricted gene flow and a strong overrepresentation of nonsynonymous polymorphism, suggesting the importance of selective pressures as drivers of the divergence. The most divergent transcripts were homologs to genes involved in calcium/potassium fluxes across the membrane, calcium transduction signal and saxitoxin production. The implication of these results in terms of ecological divergence and build-up of reproductive isolation is discussed. Dinoflagellates are especially difficult to study in the field at the ecological level due to their small size and the dynamic nature of their natural environment, but also at the genomic level due to their huge and complex genome and the absence of closely related model organism. This study illustrates the possibility to identify the traits of primary importance in ecology and evolution starting from high-throughput sequencing data, even for such organisms.

Keywords: Dinoflagellates, harmful algal blooms, Populations Genomics, Pseudocryptic species, speciation

Received 24 February 2016; revision accepted 4 August 2016

Introduction

The high number of unicellular eukaryote species coexisting in an apparently homogeneous pelagic environment has long puzzled ecologists (the paradox of the plankton, Hutchinson 1961). At the ecological scale, the paradox may be resolved, at least partly, by invoking out of equilibrium dynamics (Roy & Chattopadhyay 2007). However, at the evolutionary scale, the paradox remains extremely puzzling. In the marine environment, numerous species have a pelagic stage often associated with long-range dispersal creating high gene flow, opposing local adaptation and the speciation process (Palumbi 1992). For plants and animals with a benthic phase or animals able to swim against currents to remain in specific habitats, adaptive divergence for specific environmental conditions seems nevertheless

Correspondence: Mickael Le Gac, Fax: (33) 298224548; E-mail: Mickael.Le.Gac@ifremer.fr.

possible (Bierne *et al.* 2003). For organisms remaining as free drifters in the water column, such as phytoplankton, the forces that may drive such divergence are virtually unknown. Theoretical works taking into account the huge population sizes, specific life history traits, such as the ability to form resting cysts embedded in the sediment, and the dependency on hydrodynamics not only as a dispersive force of propagules but also as a force potentially impeding the organisms to remain in favourable environmental conditions during active growth are extremely scarce (but see Shoresh *et al.* 2008). Empirically speaking, a growing number of population genetic studies have highlighted that phytoplankton species may be spatially and temporally structured (Rynearson & Armbrust 2004; Iglesias-Rodríguez *et al.* 2006; Masseret *et al.* 2009; Casteleyn *et al.* 2010; Casabianca *et al.* 2011; Dia *et al.* 2014). Moreover, some works investigating ecological divergence between closely related species have highlighted vertical niche partitioning in foraminifer (Weiner *et al.* 2012), specialization for different light intensities and utilization of different parts of the light spectrum in cyanobacteria (Rocap *et al.* 2003; Stomp *et al.* 2004), as well as divergence in terms of metal usage in chlorophytes (Palenik *et al.* 2007) and diatoms (Peers & Price 2006).

Dinoflagellates constitute an enigmatic group of mainly marine unicellular eukaryotes with lifestyles ranging from mixotrophic (autotrophic and predator) to fully heterotrophic for half of the species, sometimes producing toxins that have ecological, economic and sanitary impacts (Anderson *et al.* 2012a) and displaying many original genomic characteristics, including genome sizes among the largest of any organisms (up to 60 times the size of the human genome, Wisecaver & Hackett 2011). The species belonging to the genus *Alexandrium* (Anderson *et al.* 2012b) are responsible for paralytic shellfish poisoning caused by the production of several toxins including saxitoxin (Cusick & Sayler 2013), a molecule classified as schedule 1 substance, in the sense of the Chemical Weapons Convention due to its very low lethal dose.

Thanks to the recent development in sequencing technologies and bioinformatics tools, it is now becoming possible to investigate the genome-wide patterns of divergence (Seehausen *et al.* 2014). These developments are not only transforming our understanding of divergence from an individual gene to a whole-genome perspective (Feder *et al.* 2013), but also enabling the investigation of genomic divergence in a wide variety of organisms spanning the entire tree of life, including organisms that are not closely related to any model organism (Ellegren 2014). So-called reverse ecology approaches where genomic data are the starting point to identify the traits of ecological and evolutionary interest (Li *et al.* 2008) are especially appealing for organisms that are difficult to study in the field, such as plankton species, to gain insight into the evolutionary processes at play during divergence and the affected cellular functions.

Here by sequencing and analysing the mRNA sequences, as well as characterizing the morphology of 18 strains of *A. minutum* isolated from natural populations, we highlight a divergence event. We investigated (i) the model of divergence most likely to explain the observed joint site frequency spectrum among seven models of divergence, (ii) whether this event is driven by selective pressures and (iii) what are the underlying divergent cellular functions.

Material and methods

RNA extraction, library preparation and sequencing

Starting from environmental samples, each *A. minutum* strain was identified by micropipetting a single cell into fresh medium under inverted microscope. Following isolation, the strains are maintained in the laboratory by biweekly dilution into fresh media. Under culture conditions, the cells are haploid and divide by mitosis, and each strain is thus composed of clonal individuals. A total of 18 strains isolated from various localities and time (Fig. 1) were grown to mid-exponential phase in 100 mL of K medium at 18 °C, 12/12-h photoperiod and 80 μE/s/m of irradiance. Cell densities ranged from 5.10^6 to $2.5.10^7$ cell/L. Cultures were centrifuged at 4500 *g* for 8 min and sonicated on ice for 20 s in RLT lysis buffer (Qiagen) containing β-mercaptoethanol. Extraction was performed using RNeasy plus mini kit (Qiagen) following the manufacturer's protocol. Extracted RNA was quantified using a Biotek Epoch spectrophotometer and the quality estimated on RNA 6000 nanochips using a Bioanalyzer (Agilent). Reverse transcription of 4 μg of total RNA into cDNA and library preparation were performed at the GeT-PlaGe France Genomics sequencing platform (Toulouse, France) using the Illumina truseq RNA V2 kits. One library was generated per *A. minutum* strain. Library quality was assessed on a Bioanalyzer using high-sensitivity DNA analysis chips and quantified using Kappa Library Quantification Kit. Paired-end sequencing was performed using 2 × 100 bp cycles. The 18 libraries were sequenced on two Illumina HiSeq lanes.

Reads quality assessment and filtering

Galaxy interface (Giardine *et al.* 2005) was used to visualize the sequencing outputs and filter out low-quality reads. Visualization was performed using FastQC.

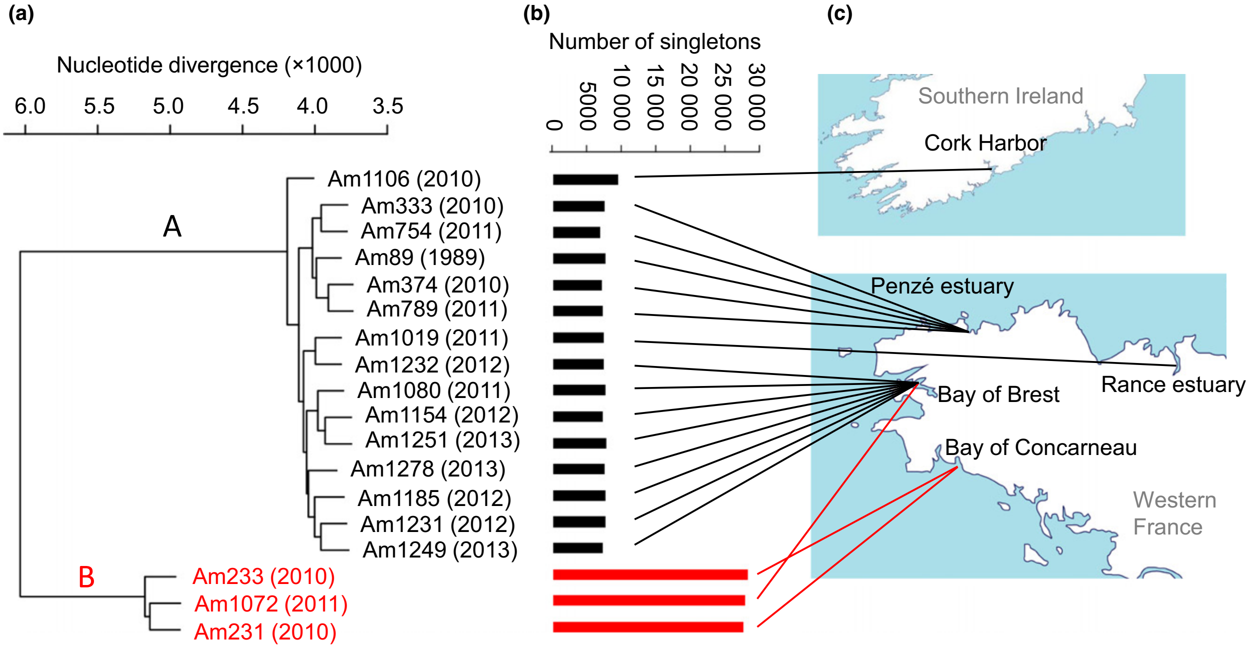


Fig. 1 Genetic divergence. (a) Hierarchical clustering analysis displaying the genetic distance among *A. minutum* strains based on nucleotide divergence, names of the strains and year of isolation is indicated; (b) the number of singletons par strain; (c) origin of the strains. The strains from group A are in black, and the ones from group B in red.

Reads were truncated until the last nucleotide displayed a Phred score of at least 25. Reads shorter than 70 bp or with an average Phred score lower than 25 were removed. Cutadapt was used to remove the sequences corresponding to the TruSeq indexed adapter, TruSeq Universal Adapter, dinoflagellate spliced leader (Zhang *et al.* 2007), as well as poly-A tails. For the 18 *A. minutum* strains sequenced, more than 68.10^9 bases were generated, of which about 4.10^9 (~6%) were discarded after quality filtering.

Obtaining A. minutum reference transcriptome

After the initial quality filtering, overlapping paired-end reads were merged using Flash (Magoc & Salzberg 2011). Sequences shorter than 70 bp were removed. Merged paired-end reads, as well as nonoverlapping paired-end and orphan reads, from the 18 strains were used to perform a *de novo* assembly of *A. minutum* transcriptome using Trinity (Haas *et al.* 2013) after pooling the reads of the 18 strains. Only transcripts longer than 200 bp were retained. A total of 216 203 transcripts were generated representing more than 178.10^6 bases of sequence and an average sequence length of 824 bases. When several isoforms were detected, only the longest one was retained for the analyses, representing 153 222 transcripts for a total of 117 601 765 bp with an average transcript size of 767 bp. Sequence similarity of the transcripts with genes of identified function in the UNIPROT databank was investigated using the bank-to-bank sequence similarity search tool NGKLAST v4.3 using the KLASTX algorithm (Nguyen & Lavenier 2009) with *E*-value $<10^{-3}$ (32 948 transcripts with homologs). The transcripts were classified into various Gene Ontology categories (GO; http://geneontology.org/) based on this result. Independently from this annotation, Transdecoder (Haas *et al.* 2013) was used to determine coding sequences (CDS) from the transcripts (76 698 transcripts with CDS). When more than one possible frame was detected (17 492 CDS, i.e. ~23%), the CDS were not considered unless it contained mutations, in which case the frame minimizing the number of nonsynonymous mutations was retained (9032 CDS). The effect of this choice on the ratio of nonsynonymous mutations per transcript is illustrated in Fig. S1, Supporting information. The analyses were also performed after excluding all the transcripts with more than one possible frame, without any impact on the conclusions (data not shown). As *A. minutum* is not closely related to any model organism, transcript annotation has to be taken with great caution. As a mean to both evaluate at what point annotations may be meaningful, and decrease the amount of wrongly annotated transcripts, the frames assigned by the NGKLAST annotation and the ones inferred from

Transdecoder were compared. A total of 26 487 transcripts had frames assigned by both Transdecoder and ngklast, of which 17 235 did match (65%). This is about four times more than expected if the annotation was biologically irrelevant (as there are six possible frames random matches are expected for 1/6 of the transcripts). When the two frames did not match, the frame inferred using Transdecoder was conserved, but the annotation was discarded.

Alignment to the reference transcriptome

The 18 strains were then individually aligned to the reference consisting of 153 222 transcripts with Bowtie2 (Langmead & Salzberg 2012) using paired-end reads. Only reads with a mapping score >10 were retained. Alignments were sorted and duplicates removed using Samtools (Li *et al.* 2009). Taking into account all strains together, sites had an average sequencing depth of 462. Individually, the strains had an average sequencing depth ranging from 11 to 49.

Mutation analyses

For variant analyses, only transcripts with more than 100 sites covered more than ten times in each of the 18 strains were considered. Single nucleotide polymorphisms (SNPs) were detected using FreeBayes (Garrison & Marth 2012). In culture conditions, *A. minutum* cells are in a vegetative, haploid stage. We took advantage of this to remove spurious SNPs and more specifically SNPs that may be identified because of genetic polymorphism within a single genome (in the case of paralogy) and not between genomes. To do so, FreeBayes was run with three sets of parameters: (i) haploidy enforced, (ii) diploidy enforced and (iii) diploidy enforced with a minimal allele count supported by at least five reads to call a genotype. Mutations identified by freebayes were then filtered using vcftools (Danecek *et al.* 2011), only keeping positions involved in SNPs, with two alleles, a quality criterion >40, and covered more than 10 times in each of the 18 sequenced strains. Because cultures are composed of haploid clones, diploid enforced genotypes must be homozygote. After filtering, the results of the three FreeBayes runs were compared and only positions identified in the haploid enforced run and identified as homozygous in the two diploid enforced runs were considered. Genotypes identified as heterozygotes in the diploidy enforced runs were discarded. Genetic distance among any two strains was calculated as the proportion of variant sites. Hierarchical clustering analysis with complete linkage was performed in r using hclust.

To investigate the divergence between group A and group B, the demographic history was analysed from their joint site frequency spectrum (JSFS) using δaδi v1.7.0 (Gutenkunst *et al.* 2009). As proposed by Tine *et al.* (2014), we tested seven alternative models of historical divergence: strict isolation (SI), isolation with migration (IM), ancient migration (AM), secondary contact (SC), as well as a version of IM, AM and SC including a restricted migration rate for a subset of SNPs (IM2 m, AM2 m and SC2 m). As the ancestral states of the SNPs could not be determined with confidence, we used folded joint site frequency spectrum, that is the frequency spectrum based on minor allele count. The demographic history was inferred using all polymorphic sites, as well as using five subsets composed of a single randomly chosen synonymous polymorphic site per transcript. For each demographic model and each data set, more than 30 runs were performed to identify the maximum-likelihood and the corresponding parameter estimates. Using this modelling approach, the SNPs observed as fixed (one allele in all members of group A and the alternative allele in all members of group B) were identified as displaying a highly restricted gene flow between the two groups. We note that when we refer to fixed polymorphism, we considered the observed pattern in the 18 strains studied and do not extrapolate the fixation at the entire group level.

Fisher's exact tests were used to investigate the deviation from random accumulation of fixed SNPs in the transcripts. False discovery rate (FDR) correction for multiple testing with a significance threshold set at *q*-value = 0.05 was used.

Following a McDonald & Kreitman (1991) approach, we used Fisher's exact tests to investigate whether NS mutations are overrepresented in the fixed differences.

Overrepresentation of (i) SNPs, (ii) nonsynonymous (NS) SNPs, (iii) fixed SNPs and (iv) fixed NS SNPs in GO categories was tested for GO categories represented by at least five transcripts, using Fisher's exact tests followed by FDR correction for multiple testing with a significance threshold set at *q*-value = 0.0001 (a very stringent FDR was set to balance the uncertainty of the GO annotations due to the absence of a closely related model organism). Only GO categories containing >five mutated transcripts were considered. Overrepresentation analyses were based on SNPs rather than on mutated transcripts to add more weight to the transcripts carrying multiple SNPs.

Saxitoxin, COI and rRNA genes

Two forms homolog to the cyanobacteria *sxtA* gene, named long and short forms, as well as one homologous of the *sxtG* cyanobacteria gene known to be involved in saxitoxin production, were identified in *Alexandrium* (Stüken *et al.* 2011). We searched the

A. minutum reference transcriptome generated above for the *A. fundyense sxtA* short (JF343238) and long (JF343239) forms as well as *sxtG* (JX995121) using BLASTN 2.2 (Zhang *et al.* 2000). Similarly, we searched for published COI (AB374235) and rRNA (AY831408) sequences in the *A. minutum* reference transcriptome generated above using BLASTN 2.2.

Intergroup differential expression

Differential expression analyses were performed using the packages, DESEQ2 (Love *et al.* 2014), EDGER (Robinson *et al.* 2010) and LIMMA (Ritchie *et al.* 2015) in R. Only transcripts with a total read count higher than 200 were considered, representing 100 797 transcripts with a median coverage per transcript ranging from 42 to 188 reads for the different strains (mean range: 108–505). Hierarchical clustering was performed using HCLUST (R) based on the Euclidean distance calculated by the dist function (R) on the rlog-transformed count matrix. Differential expression between the two groups of strains was tested with a significance FDR threshold set at q-value = 0.05, with rlog (DESEQ2), TMM (EDGER) and voom (LIMMA) normalization. The transcripts significant with the three methods (intersection) were considered as differentially expressed. We note that differentially expressed transcripts may be the result of differential regulation of gene expression in the two groups of strains, but also of deletion of the encoding genes in one of the two groups. Overrepresentation of GO categories was tested for GO categories represented by at least five transcripts, using Fisher's exact test followed by a false discovery rate (FDR) correction for multiple testing with a significance threshold set at q-value = 0.01. Only GO categories containing >five differentially expressed transcripts were considered. We note that the presence of a conserved spliced leader in 5′ of all dinoflagellate mRNA might indicate important post-transcriptional regulation of gene expression in these organisms (Zhang *et al.* 2007).

Morphological analyses of the strains

Thecal plate pattern and the presence of a ventral pore on the first apical plate (1′) of the different strains were analysed after staining thecae with Fluorescent Brightener 28 (Sigma Aldrich) according to the method of Fritz & Triemer (1985). Strains were observed on a slide covered with a coverslip in epifluorescence microscopy after adding a drop of 1% (w/v) of the fluorophore and using a BX41 (Olympus, Tokyo) upright microscope fitted with a 100-W mercury lamp and epifluorescence (U-MWU2 filter cube).

Results

Genetic diversity

To investigate genetic diversity, we only considered transcripts with more than 100 sites covered more than ten times in each of the 18 sequenced strains, representing a total of 24 630 108 sites in 45 089 transcripts, and identified a total of 457 368 polymorphic sites (~1.9% of the sites) in 41 698 transcripts (~92.5% of the transcripts, Table 1). We performed a hierarchical clustering analysis based on the nucleotide divergence among any two strains (Fig. 1a). Two groups of strains may be distinguished. The first group, hereafter named group A, of 15 strains composed of a slightly divergent strain isolated from Cork (Ireland), and two subclades grouping, on the one hand, all the strains isolated from the Penzé estuary (France) and, on the other, strains isolated from the Bay of Brest (France) and one strain isolated from the Rance estuary (France). In this group, the median

Table 1 Summary of transcripts and mutations analysed, considering the entire data set (total), the transcripts displaying mutations (mutated), the transcripts displaying mutations excluding singletons (mutated no singleton) and the transcript displaying fixed mutations (fixed)

	The number of transcripts	Length	Transcripts with CDS	Length CDS (NS)	Transcripts with homolog	Length annotated (NS)
Total	45 089	24 630 108[a]	32 797	20 396 618[a]	10 454	7 703 971[a]
Mutated	41 698	457 368[b]	31 111	376 242[b] (85 923[b])	10 029	139 286[b] (26 725[b])
Mutated no singleton	38 116	264 573[b]	29 089	221 489[b] (44 880[b])	9508	82 805[b] (14 007[b])
Fixed	6215	12 188[b]	4916	9551[b] (3818[b])	1670	3408[b] (1183[b])

[a]Length of the transcripts.

[b]The number of mutations.

number of variable sites among any two strains is 99 224 (~22% of the variable sites), representing a nucleotide divergence of ~0.004, reflecting a high level of genetic diversity (Fig. 1a, black). The second group, hereafter named group B, is composed of three strains, one isolated from the Bay of Brest and two from the Bay of Concarneau. Within this group B, the three strains are also very divergent genetically, with a median of 127 407 variable sites among strains (~28%), representing a nucleotide divergence of ~0.005. The intergroup median number of variable sites is 147 913 (~32%), representing a nucleotide divergence of ~0.006. A total of 193 325 variants are singletons; that is, they were identified in a single strain, representing more than 42% of the identified variants. Looking at the repartition of these singletons in the 18 strains, the two groups of strains previously identified are again clearly visible. Within group A, the median number of singletons is 7303 (Fig. 1b, black). Within group B, there is almost four times more singletons per strain (median = 27 532) (Fig. 1b, red).

Divergence between group A and group B

To replace the divergence between group A and group B in a classical phylogenetic context, we note that in the transcript corresponding to the ribosomal RNA, two SNPs observed as fixed (one allele in all members of group A and the alternative allele in all members of group B) are identified in the 5′ external transcribed spacer, but none in the region corresponding to the 18S, ITS1, 5.8S, ITS2 and LSU. Similarly, no SNP was identified in the transcript corresponding to the cytochrome c oxidase subunit I (COI), another gene often used to identify closely related species (Table 2).

To better grasp the patterns of divergence between strains belonging to group A and group B and gaining insights into the underlying evolutionary processes, we analysed the demographic history of groups A and B using their joint site frequency spectrum (JSFS), exploring seven scenarios of divergence (Fig. 2). The simplest model involving divergence without any gene flow (SI, Fig. 2, Table S1, Supporting information) did not explain the data as well as models involving some amount of gene flow after the split. Of these, the secondary contact (SC) model had the best likelihood, especially because it explained the low occurrence of minor allele only observed in group A (Fig. 2). The only part of the JSFS not correctly explained by the SC model was an excess of observed fixed polymorphism compared to the model prediction (lowest residual values, Fig. 2). The observed fixed polymorphism was correctly estimated when a heterogeneous migration rate across SNPs, with a fraction of the sites displaying a highly restricted gene flow between groups, was introduced in the divergence model (SC2 m model, Fig. 2, Table S1, Supporting information). Similarly, when considering subsets of the entire data set composed of a single randomly chosen synonymous polymorphic site per transcript, the model with the highest likelihood was the SC2 m model (Table S2, Supporting information). These analyses indicated an ancient divergence of the groups A and B in total isolation, followed by a secondary contact resulting in gene flow between the two groups that is heterogeneous across the genomes, with a fraction of the SNPs displaying highly restricted gene flow. As seen in Fig. 2, the polymorphic sites that are observed as fixed between the two groups are the ones displaying restricted gene flow (part of the folded JSFS requiring a heterogeneous migration rate across genomes to be correctly explained, Fig. 2).

In the data set, 12 188 variant sites (5% of the variable sites, excluding singletons) display a fixed difference between groups A and B (Table 1). We focused on the fixed differences between groups A and B to determine whether these SNPs are restricted to a few transcripts or randomly distributed in the transcripts. The 12 188 fixed differences occur in 6215 transcripts, but are overrepresented, compared to the other differences, in 927 transcripts (Fisher's exact test, q-value <0.05, Fig. 3a, red dots), representing 4616 fixed mutations (38% of the fixed differences). This result clearly points towards a preferential accumulation of the fixed differences in some transcripts.

Next, we investigated whether mutations are synonymous (S) or nonsynonymous (NS). Excluding singletons, a total of 44 880 NS and 176 609 S mutations were identified in 29 089 transcripts (Table 1). Focusing on the fixed differences, 3818 NS and 5733 S mutations were detected in 4916 transcripts, indicating that nonsynonymous mutations are 2.77 times more frequent in the fixed differences compared to the other mutations (Fisher's exact test, P-value $<2.2e^{-16}$). More precisely, the frequency of nonsynonymous mutations in the transcripts is higher when considering fixed mutations (Fig. 3b grey), not only in the transcripts where fixed mutations are overrepresented (Fig. 3b red), but also in the transcripts only displaying a few fixed mutations (Fig 3b blue). This indicates that the potential modification of protein functions associated with the divergence is not only linked to transcripts displaying numerous fixed mutations, but also to the transcripts only displaying a few fixed mutations.

Functional genetic divergence

We used two approaches to investigate the functions of the genes displaying fixed differences. In the first one,

Table 2 Most divergent transcripts between A and B, and loci classically used in phylogenetic studies

	Name	Fixed mutations	Not fixed mutations	Homologs		*E*-value	Identity (%)
25 most divergent transcripts between A and B	**comp60373_c0_seq1**	**22**	**2**	**CALL6_HUMAN**	**Calmodulin-like protein 6**	**1.10^{-10}**	**38.1**
	comp98959_c0_seq1	18	0	TGS1_HUMAN	Trimethylguanosine synthase	3.10^{-41}	39.1
	comp102434_c0_seq1	**18**	**3**	**PKD2_MOUSE**	**Polycystin-2**	**2.10^{-19}**	**35.4**
	comp124736_c0_seq1	15	0	NA			
	comp95518_c0_seq1	**16**	**2**	**NAC3_HUMAN**	**Sodium/calcium exchanger 3**	**2.10^{-120}**	**33.3**
	comp86525_c0_seq1	13	0	NA			
	comp101280_c0_seq1	15	4	NA			
	comp101305_c0_seq1	13	1	NA			
	comp75832_c0_seq1	13	2	CMBL_RAT	Carboxymethylenebutenolidase homolog	6.10^{-12}	22.8
	comp96757_c0_seq1	13	2	NA			
	comp96807_c0_seq1	12	1	NEK5_HUMAN	Serine/threonine protein kinase Nek5	5.10^{-05}	26.5
	comp124661_c0_seq5	14	5	PGMC2_ARATH	Glucose phosphomutase 2	1.10^{-164}	48.9
	comp78930_c0_seq1	11	0	NA			
	comp94714_c0_seq1	15	8	NA			
	comp104352_c0_seq1	12	2	NAAA_MOUSE	N-Acylethanolamine-hydrolysing acid amidase	8.10^{-33}	29.9
	comp82584_c0_seq1	11	1	NA			
	comp115853_c0_seq1	13	5	PAMO_THEFY	Phenylacetone monooxygenase	5.10^{-05}	34
	comp95265_c0_seq1	12	3	NA			
	comp105111_c2_seq1	10	0	EF1A_CRYPV	Elongation factor 1-alpha	3.10^{-96}	46.2
	comp106635_c0_seq1	**10**	**0**	**CDPKD_ARATH**	**Calcium-dependent protein kinase 13**	**1.10^{-31}**	**24.9**
	comp119140_c0_seq2	10	0	WIPF1_MOUSE	WAS/WASL-interacting protein family member 1	2.10^{-05}	34.3
	comp86654_c0_seq1	11	2	NA			
	comp121041_c0_seq1	**15**	**13**	**F5BWX9_ALEFU**	**SxtA short isoform precursor**	**0.0**	**63**
	comp117520_c0_seq1	14	14	MSL7_MYCMM	Beta-ketoacyl-acyl-carrier-protein synthase I	7.10^{-22}	30.4
	comp66739_c0_seq1	10	1	ATAD3_BOVIN	ATPase family AAA domain-containing protein 3	7.10^{-91}	40
COI	comp126209_c0_seq1	0	0	AB374235	A. catenella cox1	0.0	99
rRNA	comp93300_c0_seq1	2 (ETS)	0	AY831408	*A. minutum* CCMP113 ETS-18S-ITS1-5.8S-ITS2-LSU	0.0	99

Upper part, transcripts displaying the highest level of divergence between groups A and B (KLASTX against UNIPROT/SWISSPROT). Transcripts with homologs involved in saxitoxin production and calcium signal transduction are indicated in violet and red, respectively. Lower part, loci classically used in phylogenetic studies (BLASTN).

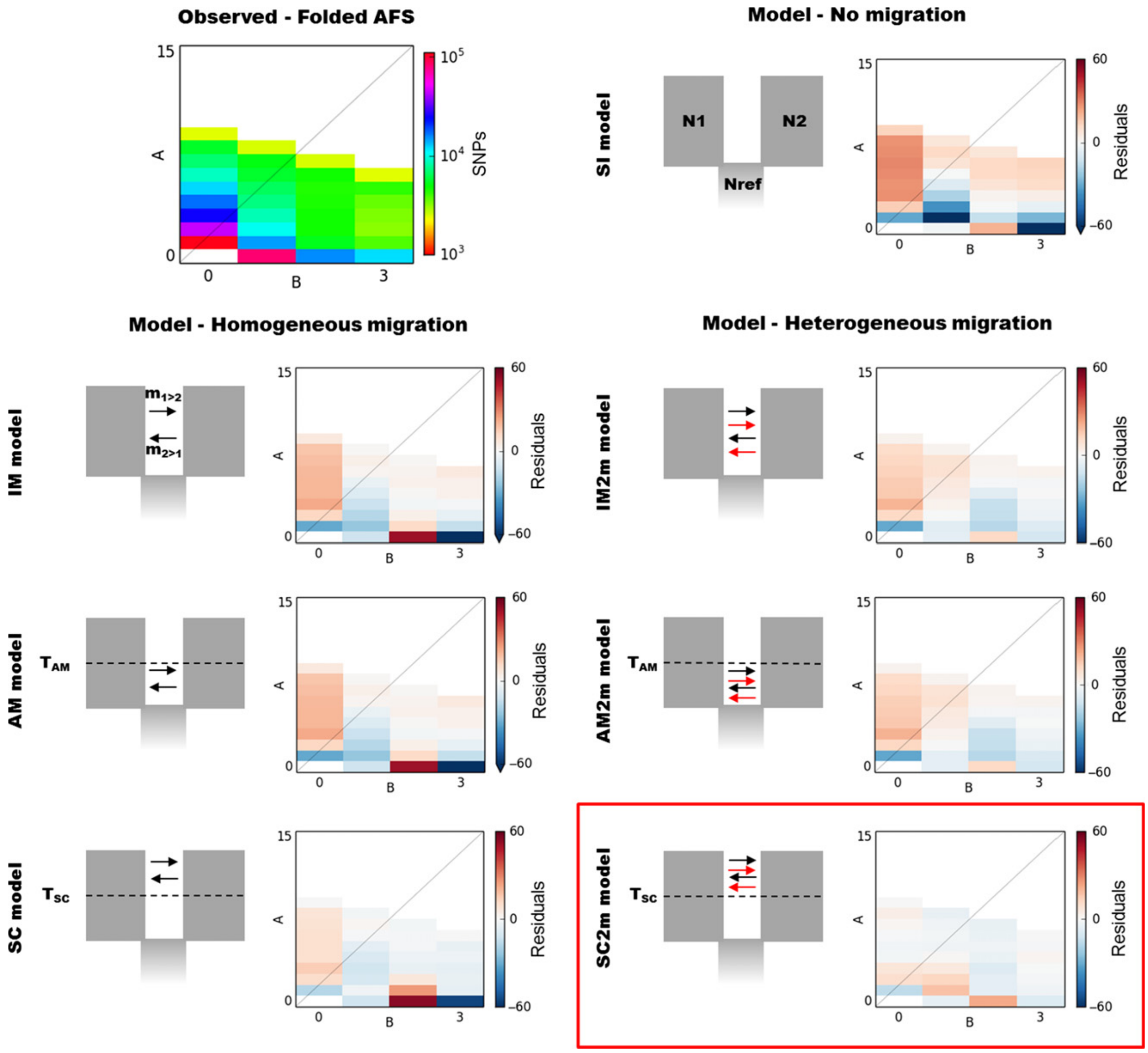


Fig. 2 Results of model fitting for seven alternative models of divergence. The observed folded allele frequency spectrum (AFS), as well as for each model, the residuals of the modelled AFS are presented. SI is the strict isolation model. IM is the isolation-with-migration model; AM, the ancient migration model; and SC, the secondary contact model. All three models of divergence with gene flow were implemented using one shared migration rate in each direction (m1 > 2, m2 > 1) across the genome (homogeneous migration), or with two categories of migration rates in each direction across the genome (heterogeneous migration). The data are best explained by the SC2 m model.

we analysed the repartition of SNPs associated with the different gene product properties, as defined by Gene Ontology (GO). A total of 9508 transcripts representing 82 805 mutations could be associated with GO categories (Table 1). We tested whether mutations are over- or underrepresented in the different GO categories. Considering the entire data set, 24 GO categories display an excess of mutations and 147 display less mutations than expected (Fig. 4, 171 GO categories overall). The nonsynonymous mutations were overrepresented in six categories and underrepresented in six (Fig. 4, 12 GO categories overall nonsynonymous). Focusing on the fixed differences, mutations were overrepresented in 33 categories (nonsynonymous mutations, four categories) and not underrepresented in any GO category (Fig. 4 fixed and fixed nonsynonymous). These fixed differences are found in a total of 328 transcripts. Of special interest are 130 transcripts involved in five GO

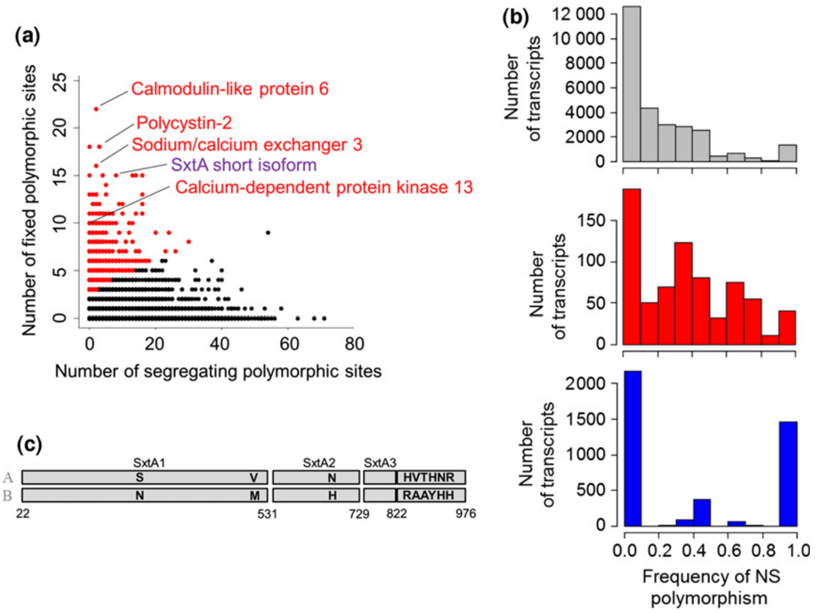


Fig. 3 Fixed polymorphism. (a) Repartition of the transcripts based on the number sites displaying fixed and segregating polymorphism. Red dots indicate overrepresentation of fixed polymorphism (*q*-value <0.05). For the 25 most divergent transcripts, homology with genes involved in calcium transduction signal (red) and saxitoxin production (violet) is indicated. (b) Frequency of NS polymorphism considering segregating polymorphism (grey), fixed polymorphism in transcripts where fixed polymorphism is overrepresented (red), and fixed polymorphism in transcripts without overrepresentation of fixed polymorphism (blue). (c) Fixed amino acid substitutions in SxtA.

categories related to calcium binding and fluxes across membranes (Fig. 4 red) and 44 in four GO categories related to potassium fluxes across membranes (Fig. 4 blue).

In a second approach to grasp the functional bases of the divergence, we focused on the 25 transcripts displaying most fixed genetic divergence between the divergent groups (lowest *q*-value, Fig. 3a, i.e. ~0.5‰ (25/45 089) transcripts displaying the highest level of genetic divergence (Table 2)). Of these 25 transcripts, 14 were identified as homologs to genes encoding for proteins with known functions. Of extreme interest was the presence of four transcripts homologs to genes involved in calcium-mediated transduction signals: two involved in calcium transport (polycystin-2 and sodium/calcium exchanger 3), one intermediate messenger transducing calcium signals by binding calcium ions (calmodulin-like protein 6) and one calcium-dependent protein kinase thought to function in signal transduction pathways that utilize the changes in cellular Ca^{2+} concentration to couple cellular responses to extracellular stimuli (calcium-dependent protein kinase 13). Even more interesting was the genetic divergence of a transcript corresponding to the short form of the *sxtA* gene, a gene known to be involved in saxitoxin production in cyanobacteria. It contains domains 1–3 homologous to the *sxtA* genes found in cyanobacteria and a last translated region that has no homolog in databases except the end of the short *sxtA* form from *A. fundyense* (Stüken *et al.* 2011). Moreover, these fixed differences include numerous NS mutations (9), two of them being in the first domain (sxtA1, corresponding to the amino acids 28-531), one in the second domain (sxtA2, amino acids 535-729), none in the third (sxtA3, amino acids 750-822) and six in the last translated part of the transcript (amino acids 822-976) (Fig. 2C).

Differential gene expression

We analysed the mRNA sequences to investigate the differential gene expression *in vitro*. First, a clustering analysis based on the expression levels clearly indicates that the two groups of strains identified above using genetic information are also identified using global expression data (Fig. 5a). Differential expression was analysed between group A and group B strains, and a total of 1518 transcripts were identified as differentially expressed (*q*-value <0.05; Fig. 5b; Table S3, Supporting information), but no gene ontology category was identified as over- or underrepresented in the differentially expressed transcripts at a FDR level <0.1.

Morphology

The 18 strains were stained with Fluorescent Brightener 28 and observed blindly, i.e. without knowing which strains belonged to group A and B group in epifluorescence microscopy to analyse the thecal plate pattern. No difference in the thecal organization was found among strains that all possessed the typical plate pattern of *A. minutum*. However, the presence of a ventral pore on the right side of the 1′ plate was found on the three strains belonging to group B, while the 15 strains belonging to group A lacked this feature (Fig. 6).

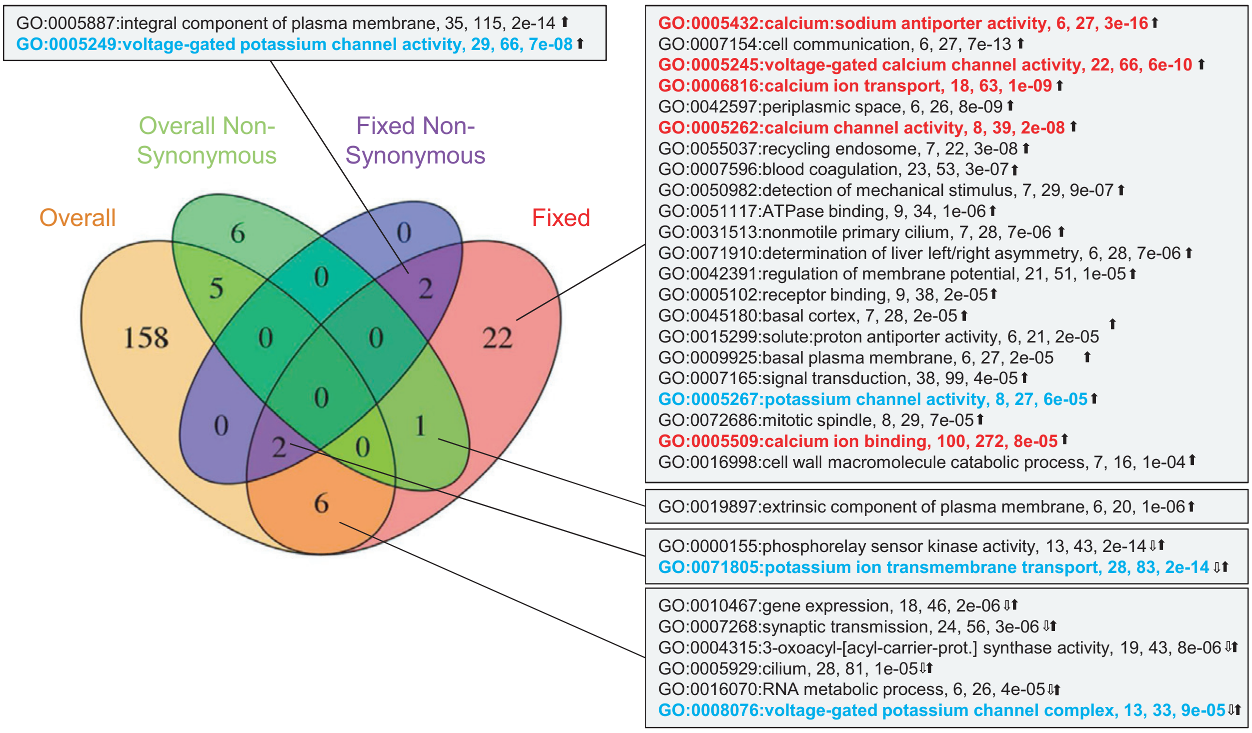

GO:0005887:integral component of plasma membrane, 35, 115, 2e-14 ⬆
GO:0005249:voltage-gated potassium channel activity, 29, 66, 7e-08 ⬆
Overall Non-Synonymous
Fixed Non-Synonymous
Overall
Fixed
6
0
5
0
2
158
0
0
22
0
0
1
2
0
6
GO:0005432:calcium:sodium antiporter activity, 6, 27, 3e-16 ⬆
GO:0007154:cell communication, 6, 27, 7e-13 ⬆
GO:0005245:voltage-gated calcium channel activity, 22, 66, 6e-10 ⬆
GO:0006816:calcium ion transport, 18, 63, 1e-09 ⬆
GO:0042597:periplasmic space, 6, 26, 8e-09 ⬆
GO:0005262:calcium channel activity, 8, 39, 2e-08 ⬆
GO:0055037:recycling endosome, 7, 22, 3e-08 ⬆
GO:0007596:blood coagulation, 23, 53, 3e-07 ⬆
GO:0050982:detection of mechanical stimulus, 7, 29, 9e-07 ⬆
GO:0051117:ATPase binding, 9, 34, 1e-06 ⬆
GO:0031513:nonmotile primary cilium, 7, 28, 7e-06 ⬆
GO:0071910:determination of liver left/right asymmetry, 6, 28, 7e-06 ⬆
GO:0042391:regulation of membrane potential, 21, 51, 1e-05 ⬆
GO:0005102:receptor binding, 9, 38, 2e-05 ⬆
GO:0045180:basal cortex, 7, 28, 2e-05 ⬆
GO:0015299:solute:proton antiporter activity, 6, 21, 2e-05 ⬆
GO:0009925:basal plasma membrane, 6, 27, 2e-05 ⬆
GO:0007165:signal transduction, 38, 99, 4e-05 ⬆
GO:0005267:potassium channel activity, 8, 27, 6e-05 ⬆
GO:0072686:mitotic spindle, 8, 29, 7e-05 ⬆
GO:0005509:calcium ion binding, 100, 272, 8e-05 ⬆
GO:0016998:cell wall macromolecule catabolic process, 7, 16, 1e-04 ⬆
GO:0019897:extrinsic component of plasma membrane, 6, 20, 1e-06 ⬆
GO:0000155:phosphorelay sensor kinase activity, 13, 43, 2e-14 ⇩⬆
GO:0071805:potassium ion transmembrane transport, 28, 83, 2e-14 ⇩⬆
GO:0010467:gene expression, 18, 46, 2e-06 ⇩⬆
GO:0007268:synaptic transmission, 24, 56, 3e-06 ⇩⬆
GO:0004315:3-oxoacyl-[acyl-carrier-prot.] synthase activity, 19, 43, 8e-06 ⇩⬆
GO:0005929:cilium, 28, 81, 1e-05 ⇩⬆
GO:0016070:RNA metabolic process, 6, 26, 4e-05 ⇩⬆
GO:0008076:voltage-gated potassium channel complex, 13, 33, 9e-05 ⇩⬆

Fig. 4 Venn diagram indicating the number of Gene Ontology (GO) categories displaying deviation from random accumulation of mutations (q-value <0.0001), considering all mutations (overall), the NS mutations (overall nonsynonymous), the fixed mutations (fixed) and the NS fixed mutations (nonsynonymous fixed). For the analyses focusing on the fixed mutations, the name of the GO categories is given, as well as the number of transcripts mutated, the number of mutations and q-value s. Black arrows indicate over-representation of fixed mutations, and white arrows indicate underrepresentation of mutations overall.

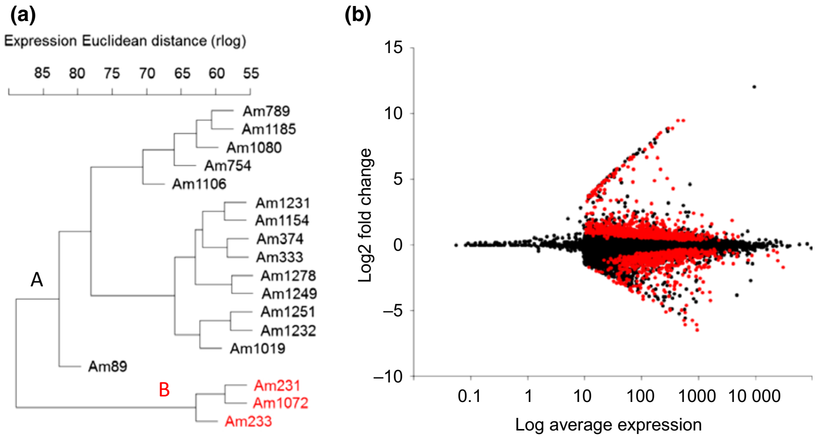


Fig. 5 Gene expression. (a) Hierarchical clustering based on the expression Euclidean distance (rlog). The strains from group A are in black and the ones from group B in red. (b) MA plot showing for each transcript the fold change (group B/group A) as a function of the average expression. Transcripts identified as differentially expressed are in red (q-value <0.05).

Discussion

Analysing mRNA sequences in 18 *A. minutum* strains, we identified the divergence of two groups, represented by 15 and 3 strains, respectively. The identification of these two groups was incidental, explaining the unbalanced sampling and illustrating the possibility for reverse ecology approaches to uncover cryptic diversity. This divergence was not detectable using the classical barcoding loci ITS and COI, but the analysis of a transcriptome-wide SNPs data set pointed towards the presence of two distinct evolutionary units. A genetic distance analysis clearly indicated the presence of the two groups. A consistent observation is the difference in the number of singletons identified in the strains belonging to the groups A and B, clearly pointed

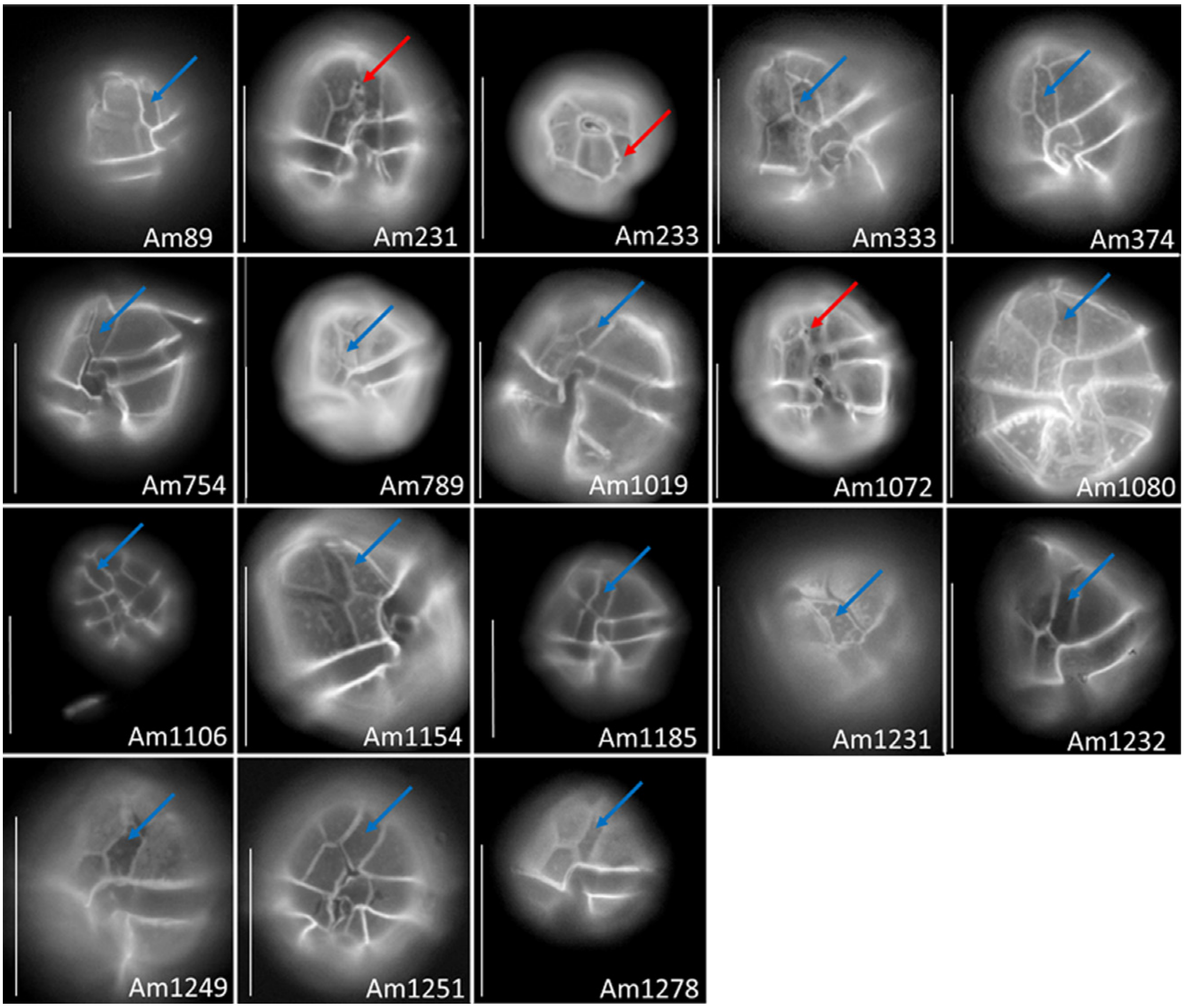


Fig. 6 Epifluorescence micrographs of the 18 strains showing the presence (red arrow) or the absence (blue arrow) of a ventral pore on the first apical plate of the theca. Scale bars: 20 μm.

towards the sampling of two independent genetic entities. Differential expression, although less dramatic than genetic divergence, also goes in the same direction, with the strains belonging to the two groups displaying the most difference in terms of global expression profile. One of the strains was isolated from the natural environment in 1989 (Am89) and maintained ever since (i.e. for 24 years, corresponding to ~3000 generations of cellular division) in batch culture involving biweekly transfer in the culture media used in the present work. The other strains were isolated from 2010 to 2013 and maintained in the same culture regime (6 months–3 years of laboratory culture, 60-350 generations). Despite the difference in the time spent in the laboratory environment and thus experiencing the associated strong selective pressures, the strain Am89 is genetically indistinguishable from the other strains belonging to group A. It is illustrated that compared to the standing genetic variation encountered in natural populations, there are very few mutations that occurred during the long-term maintenance in culture. In terms of gene expression, Am89 clusters at the base of group A, pointing towards a more extensive evolution of gene expression profile, but is still insufficient to overcome the difference in global expression profile occurring between groups A and B.

Following the analysis of the mRNA sequences and the identification of the two diverging groups, a morphological difference, the presence/absence of a ventral pore, was identified. This morphological character seems diagnostic of the two groups, suggesting the occurrence of two pseudocryptic (or pseudosibling) species (Knowlton

1993), but caution must be taken due to the very limited sampling of one of the two groups. Nonetheless, this morphological character is especially interesting to replace our study in a biogeographical context. Indeed, this morphological feature has been reported in *A. minutum* studies with some indication that the morphotype with ventral pore may be more frequent in southern Europe and the one lacking the ventral pore more frequent in northern Europe (Hansen *et al.* 2003). Interestingly, the two types have also been reported in mixed communities (Western Ireland, Hansen *et al.* 2003; in the present study, Am1072 (group B) and Am1080 (group A) were isolated from the same day and locality), which rules out the complete allopatry.

Using the SNPs data set, we investigated the process of divergence between groups A and B. We compared the joint site frequency spectrum of these two groups to the patterns expected following seven models of divergence. The most likely scenario involves an ancient divergence in complete isolation followed by a secondary contact involving gene flow between the two groups. Quite interestingly, the introduction of a heterogeneous migration rate across the genome, with a fraction of the genome displaying a highly restricted gene flow, considerably improved the likelihood of the models. So far, only a handful of studies have considered heterogeneous gene flow across the genomes when investigating the divergence of population/species. We note that models of divergence in isolation followed by a secondary contact allowing gene flow between diverging populations/species, but at different rates across the genome, are, so far, almost always the best at explaining the observed allelic frequencies in ascidian (Roux *et al.* 2013), mussels (Roux *et al.* 2014), fishes (Tine *et al.* 2014; Rougemont *et al.* 2015; Le Moan *et al.* 2016) and Ascomycota (Gladieux *et al.* 2015). Here, an extremely low migration rate at a fraction of the genome is required to explain the observed pattern of fixed polymorphism, that is of polymorphism with all members of group A displaying one allele and all members of group B displaying the alternative allele. This fixed polymorphism corresponds to about 5% of the SNPs displaying a heterogeneous distribution in the various transcripts. This is similar to the pattern reported in studies investigating recent or ongoing speciation events at the genome scale (Seehausen *et al.* 2014). However, one of the caveats of using transcriptome- and not genome-wide data is that the information regarding the physical linkage between the genes encoding for the transcripts is lacking. As a result, we do not know whether the transcripts displaying high levels of genetic divergence are physically linked in a few genomic islands of divergence (Turner *et al.* 2005) or whether they are spread out in the genome.

We compared the proportion of nonsynonymous polymorphism segregating and fixed between the two groups and identified a strong excess of nonsynonymous polymorphism in the fixed mutations. SNPs fixation within each group, but divergence between groups associated with overrepresentation of NS SNPs is difficult to explain with demographic fluctuations or relaxed selection and points towards the importance of selection as a driving force of the divergence between the two groups. This pattern could reflect classic selective sweeps, that is the fixation of adaptive mutations in either group, and the associated hitchhiking of physically linked neutral mutations (Nielsen 2005). Interestingly, an excess of fixed nonsynonymous mutations was also identified in transcripts only displaying a few fixed polymorphic sites, often associated with segregating polymorphism. This excess of NS mutations suggests that mutations associated with the functional divergence of the two divergent groups are not systematically associated with a selective sweep, that is may get to fixation without a drastic reduction in diversity at neighbouring sites. Indeed, the pattern of linkage disequilibrium associated with an adaptive mutation is influenced by numerous factors including the strength of selection, local levels of recombination and whether adaptive mutations are *de novo* mutations or were segregating in ancestral populations before becoming adaptive (Fay & Wu 2000; Przeworski *et al.* 2005; Lee *et al.* 2014).

The selective pressures responsible for restricted and heterogeneous gene flow may be directly linked to the ecological divergence of the two groups. It could, for example, be the case, if the two groups occupy geographically and ecologically distinct habitats and only encounter each other and exchange genes at localized hybrid zones. In this case, introgression of neutral SNPs from one group to the other would occur more or less freely, while the introgression of the SNPs responsible for local adaptation of each group would be counterselected. An alternative scenario could involve the build-up of reproductive isolation between the two groups. For example, gene flow could be restricted overall if members of the two groups are not likely to recognize each other as proper mates, and negative epistasis between sets of SNPs could lead to reduced hybrid fitness depending (hybrid maladaptation) or not (genetic incompatibilities) on the environmental conditions. Distinguishing between these different scenarios (none of them being mutually exclusive) would require extensive sampling from the natural environment, crossing experiments and fitness assays that are beyond the scope of the present work. However, investigating the cellular functions of the transcripts displaying restricted gene flow between the two groups could help pointing in one direction.

Transcripts related to potassium and calcium fluxes across membranes were identified as carrying more fixed polymorphism than expected. Moreover, among the transcripts displaying the highest levels of divergence, four could be related to calcium-mediated transduction signals, and one was homologous to *sxtA*, a gene involved in saxitoxin production (Stüken *et al.* 2011). Two genetically divergent forms of *sxtA* have been identified in *Alexandrium* transcriptomes (Stüken *et al.* 2011). Here, the *sxtA* identified as highly divergent between the two groups corresponds to the short form, which is probably not involved in saxitoxin production (Murray *et al.* 2015; the long form was also identified in all strains, but without displaying a pattern of divergence, data not shown). As a result, we hypothesize that the molecule of interest associated with the divergence of the two groups might not be the saxitoxin itself but another compound synthesized via the saxitoxin biosynthesis pathway. There may be a direct link between *sxtA*, genes related to calcium and potassium fluxes, and calcium-mediated signal transduction. Indeed, although the saxitoxin toxicity occurs through the blocking of mammal sodium channels, it is also known to bind to mammal calcium and potassium channels, modifying calcium and potassium fluxes without entirely blocking them (Cusick & Sayler 2013). This analysis points towards a molecular mechanism that may be at play during the divergence of the two groups, but does not indicate whether it is related to ecological divergence or the build-up of reproductive isolation. In favour of the build-up of reproductive isolation, saxitoxin has been proposed to act as a sex pheromone in natural environment (Wyatt & Jenkinson 1997; Cusick & Sayler 2013) and another guanidine alkaloids marine toxin, the tetrodotoxin, has been shown to act as sex pheromone (Matsumura 1995). However, some *Alexandrium* strains do not produce the toxin and would thus be unable to attract proper mates, but as discussed above, the molecule at play here is probably not the saxitoxin itself but a related molecule. In favour of an ecological divergence, we may cite the proposed role of saxitoxin as a grazer deterrent (Cusick & Sayler 2013), but it would require a specialized relationship to exert a selective pressure responsible for the observed divergence. Finally, unicellular motility is often linked to calcium fluxes across the membrane (Verret *et al.* 2010), with the potential implications in both ecological divergence and reproductive isolation.

To conclude, using a reverse ecology approach based on the mRNA sequencing and morphology analysis of several strains of the dinoflagellate *A. minutum*, two diverging groups, co-occurring in nature, were identified. The most likely scenario of divergence involved ancient divergence in complete isolation followed by a secondary contact resulting in gene flow, heterogeneous across the genome, between the diverging groups. The SNPs subjected to restricted gene flow also display an overrepresentation of fixed nonsynonymous polymorphism. This highlights the importance of the functional aspect of the divergence and identifies selection as a potential major evolutionary force driving this event. At the molecular level, the functions associated with the divergence are especially related to toxin production and calcium/potassium fluxes with the potential implications in terms of ecological divergence and build-up of reproductive isolation that remain to be tested.

Acknowledgements

This work was supported by a financial support from the Région Bretagne and the ANR Hapar (ANR-14-CE02-0007-01). We acknowledge RIC and Abims facilities for bioinformatics support, the laboratories LSEM and PFOM for providing access to some of their equipment, Elisabeth Nézan and Pierre-Alexandre Gagnaire for discussion, four anonymous reviewers, Tatiana Giraud and Myriam Valero for comments on the manuscript.

References

Anderson DM, Alpermann TJ, Cembella AD *et al.* (2012a) The globally distributed genus Alexandrium: multifaceted roles in marine ecosystems and impacts on human health. *Harmful Algae*, **14**, 10–35.

Anderson DM, Cembella AD, Hallegraeff GM (2012b) Progress in understanding harmful algal blooms: paradigm shifts and new technologies for research, monitoring, and management. *Annual Review of Marine Science*, **4**, 143–176.

Bierne N, Bonhomme F, David P (2003) Habitat preference and the marine-speciation paradox. *Proceedings of the Royal Society B-Biological Sciences*, **270**, 1399–1406.

Casabianca S, Penna A, Pecchioli E, Jordi A, Basterretxea G, Vernesi C (2011) Population genetic structure and connectivity of the harmful dinoflagellate *Alexandrium minutum* in the Mediterranean Sea. *Proceedings of the Royal Society B-Biological Sciences*, **279**, 129–138.

Casteleyn G, Leliaert F, Backeljau T *et al.* (2010) Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proceedings of the National Academy of Sciences, USA*, **107**, 12952–12957.

Cusick KD, Sayler GS (2013) An overview on the marine neurotoxin, saxitoxin: genetics, molecular targets, methods of detection and ecological functions. *Marine Drugs*, **11**, 991–1018.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Dia A, Guillou L, Mauger S *et al.* (2014) Spatiotemporal changes in the genetic diversity of harmful algal blooms caused by the toxic dinoflagellate *Alexandrium minutum*. *Molecular Ecology*, **23**, 549–560.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.

Feder JL, Flaxman SM, Egan SP, Comeault AA, Nosil P (2013) Geographic mode of speciation and genomic divergence. *Annual Review of Ecology Evolution and Systematics*, **44**, 73–97.

Fritz L, Triemer RE (1985) A rapid simple technique utilizing Calcofluor White M2R for the visualization of dinoflagellate thecal plates. *Journal of Phycology*, **21**, 662–664.

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. Preprint arXiv:1207.3907

Giardine B, Riemer C, Hardison RC *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, **15**, 1451–1455.

Gladieux P, Wilson BA, Perraudeau F *et al.* (2015) Genomic sequencing reveals historical, demographic and selective factors associated with the diversification of the fire-associated fungus Neurospora discreta. *Molecular Ecology*, **24**, 5657–5675.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP data. *PLoS Genetics*, **5**, e1000695.

Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.

Hansen G, Daugbjerg N, Franco JM (2003) Morphology, toxin composition and LSU rDNA phylogeny of *Alexandrium minutum* (Dinophyceae) from Denmark, with some morphological observations on other European strains. *Harmful Algae*, **2**, 317–335.

Hutchinson G (1961) The paradox of the plankton. *The American Naturalist*, **95**, 137–145.

Iglesias-Rodríguez MD, Schofield OM, Batley J, Medlin LK, Hayes PK (2006) Intraspecific genetic diversity in the marine coccolithophore *Emilianta huxleyi* (Primnesiophyceae): the use of microsatellite analysis in marine phytoplankton population studies. *Journal of Phycology*, **42**, 526–536.

Knowlton N (1993) Sibling Species in the Sea. *Annual Review of Ecology and Systematics*, **24**, 189–216.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Le Moan A, Gagnaire P-A, Bonhomme F (2016) Parallel genetic divergence among coastal–marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Molecular Ecology*, **25**, 3187–3202.

Lee YCG, Langley CH, Begun DJ (2014) Differential strengths of positive selection revealed by hitchhiking effects at small physical scales in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **31**, 804–816.

Li YF, Costello JC, Holloway AK, Hahn MW (2008) "Reverse ecology" and the power of population genomics. *Evolution*, **62**, 2984–2994.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.

Magoc T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.

Masseret E, Grzebyk D, Nagai S *et al.* (2009) Unexpected genetic diversity among and within populations of the toxic dinoflagellate *Alexandrium catenella* as revealed by nuclear microsatellite markers. *Applied and Environment Microbiology*, **75**, 2037–2045.

Matsumura K (1995) Tetrodotoxin as a pheromone. *Nature*, **378**, 563–564.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the adh locus in *Drosophila*. *Nature*, **351**, 652–654.

Murray SA, Diwan R, Orr RJS, Kohli GS, John U (2015) Gene duplication, loss and selection in the evolution of saxitoxin biosynthesis in alveolates. *Molecular Phylogenetics and Evolution*, **92**, 165–180.

Nguyen VH, Lavenier D (2009) PLAST: parallel local alignment search tool for database comparison. *BMC Bioinformatics*, **10**, 329.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Palenik B, Grimwood J, Aerts A *et al.* (2007) The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences, USA*, **104**, 7705–7710.

Palumbi SR (1992) Marine speciation on a small planet. *Trends in Ecology & Evolution*, **7**, 114–118.

Peers G, Price NM (2006) Copper-containing plastocyanin used for electron transport by an oceanic diatom. *Nature*, **441**, 341–344.

Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. *Evolution*, **59**, 2312–2323.

Ritchie ME, Phipson B, Wu D *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47.

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rocap G, Larimer FW, Lamerdin J *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, **424**, 1042–1047.

Rougemont Q, Gaigher A, Lasne E *et al.* (2015) Low reproductive isolation and highly variable levels of gene flow reveal limited progress towards speciation between European river and brook lampreys. *Molecular Ecology*, **28**, 2249–2263.

Roux C, Tsagkogeorga G, Bierne N, Galtier N (2013) Crossing the species barrier: genomic hotspots of introgression between two highly divergent *Ciona intestinalis* species. *Molecular Biology and Evolution*, **30**, 1574–1587.

Roux C, Fraïsse C, Castric V, Vekemans X, Pogson GH, Bierne N (2014) Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? A case study in a *Mytilus* hybrid zone. *Journal of Evolutionary Biology*, **27**, 1662–1675.

Roy S, Chattopadhyay J (2007) Towards a resolution of 'the paradox of the plankton': a brief overview of the proposed mechanisms. *Ecological Complexity*, **4**, 26–33.

Rynearson TA, Armbrust VE (2004) Genetic differentiation among populations of the planktonic marine diatom *Ditylum brightwellii* (Bacillariophycae). *Journal of Phycology*, **40**, 34–43.

Seehausen O, Butlin RK, Keller I *et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176–192.

Shoresh N, Hegreness M, Kishony R (2008) Evolution exacerbates the paradox of the plankton. *Proceedings of the National Academy of Sciences, USA*, **105**, 12365–12369.

Stomp M, Huisman J, de Jongh F *et al.* (2004) Adaptive divergence in pigment composition promotes phytoplankton biodiversity. *Nature*, **432**, 104–107.

Stüken A, Orr RJS, Kellmann R *et al.* (2011) Discovery of nuclear-encoded genes for the neurotoxin saxitoxin in dinoflagellates. *PLoS ONE*, **6**, 12.

Tine M, Kuhl H, Gagnaire P-A *et al.* (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, **5**, 5770.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, 1572–1578.

Verret F, Wheeler G, Taylor AR, Farnham G, Brownlee C (2010) Calcium channels in photosynthetic eukaryotes: implications for evolution of calcium-based signalling. *New Phytologist*, **187**, 23–43.

Weiner A, Aurahs R, Kurasawa A, Kitazato H, Kucera M (2012) Vertical niche partitioning between cryptic sibling species of a cosmopolitan marine planktonic protist. *Molecular Ecology*, **21**, 4063–4073.

Wisecaver JH, Hackett JD (2011) Dinoflagellate genome evolution. *Annual Review of Microbiology*, **65**, 369–387.

Wyatt T, Jenkinson IR (1997) Notes on *Alexandrium* population dynamics. *Journal of Plankton Research*, **19**, 551–575.

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7**, 203–214.

Zhang H, Hou Y, Miranda L *et al.* (2007) Spliced leader RNA trans-splicing in dinoflagellates. *Proceedings of the National Academy of Sciences, USA*, **104**, 4618–4623.

M.L.G., C.D. and L.G. designed research; M.L.G., G.M., N.C., P.M., J.Q. and O.B. performed research; M.L.G. and N.C. analysed the data; M.L.G., G.M., N.C., R.S., C.D., L.G. and A.C. wrote the manuscript.

Data accessibility

Raw reads and Reference transcriptome: European Nucleotide Archive http://www.ebi.ac.uk/ena/data/view/PRJEB15046

SNP and differential expression information: SEANOE database http://doi.org/10.17882/45445

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Selecting the reading frame minimizing the proportion of non-synonymous mutations when several reading frames are possible.

Tables S1–S2 Results of model fitting for seven alternative models of divergence.

Table S3 Annotated transcripts displaying differential expression between groups A and B.

2.3 Linking *in vitro* vs *in situ* gene expression variations to genetic background : investigating species divergence under the molecular physiology scope

Title : Linking *in vitro* vs *in situ* gene expression variations to genetic background : investigating species divergence under the molecular physiology scope.

Running title : Species and gene expression divergence

Authors : METEGNIER Gabriel[1,2], QUERE Julien[1], DESTOMBE Christophe[2], LE GAC Mickael[1]

[1] Ifremer, DYNECO PELAGOS, 29280 Plouzané, France
[2] CNRS, Sorbonne Université, Pontificia Universidad Catolica de Chile, Universidad Austral de Chile, UMI 3614, Evolutionary Biology and Ecology of Algae, Station Biologique de Roscoff, Place Georges Teissier, CS90074 29688, Roscoff Cedex, France

Corresponding author : METEGNIER Gabriel ;
Ifremer, DYNECO PELAGOS, 29280 Plouzané, France ;
Phone number : +33 2 98 22 42 60 ;
E-mail: gabriel.metegnier@gmail.com

Keywords : cryptic species, adaptation, transcriptomic, reverse ecology

Submitted to Molecular Ecology

Author contributions : Results presented here are based on environmental data, sampled by numerous persons from the DYNECO-PELAGOS laboratory before the start of the Ph.D. Wet laboratory work was led by Mickael Le Gac and Julien Quéré for 2013 and 2014 samples while I prepared 2015 samples for sequencing. I performed data analysis and manuscript writing, Mickael Le Gac and Christophe Destombe participated to result analysis, discussion construction and manuscript corrections.

2.3.1 Abstract

More than only describing species diversity, understanding their divergence is a cornerstone in biological sciences, and replacing species divergence into their ecological context is another challenging dimension. The development of NGS techniques opened the gate of molecular physiology and gene expression responses to both internal and external stimuli. Here, we used the advantages of *in situ* transcriptomics to follow two *Alexandrium minutum* cryptic species in two separated area, and we tested how both their genetic divergence and environment impact their gene expression. First, using fixed SNP positions, we identified two types of populations : one is constantly composed of a single cryptic species over a period of three years while one other is composed of both species co-occurring together. Second, we investigated the factors responsible for their expression divergence. One of the key result is the dual impact of growing conditions and genetic background on gene expression : the molecular physiology differences observed between the two cryptic species is not the same when cells are living in their natural environment or under controlled conditions. These results reflect the difficulty to study ecological divergence of species outside their natural environment, and highlights the potential of reverse ecology to identify the genes and mutations underlying adaptive traits of sibling species, and to follow their dynamics at multiple levels, both *in vitro* and *in situ*.

2.3.2 Introduction

Hutchinson (1961) describes the plankton paradox as the apparent conflict between the high plankton diversity observed and the principle of competitive exclusion that does not allow two species to compete for the same resource without one excluding the other at the equilibrium. Recently, the use of molecular tools has shown that morphology does not always reflect species boundaries and consequently, many morphological species are actually complexes of cryptic or pseudo-cryptic species (different species that have been overlooked due to lack of clear diagnostic morphological character; see Adams, Raadik, Burridge, and Georges (2014);

Fišer et al. (2018) for cryptic species prevalence and impact on biodiversity assessment). While an upstream molecular-based identification is necessary to find diagnostic characters and to uncover pseudo-cryptic species a posteriori, the careful observation of discrete morphological characters can still be a powerful identification method (see the review of Lajus, Sukhikh, and Alekseev (2015) in *Eurytemora affinis* species complex taxonomy). This disparity between genetic diversity and morphology, although observed in all groups of organisms (Bickford et al., 2007),becomes more problematic when organisms decrease in morphological complexity and size (De Clerck, Guiry, Leliaert, Samyn, & Verbruggen, 2013; Verbruggen et al., 2009; Whittaker et al., 2005). Indeed, levels of cryptic diversity are very important in some groups of algae (e.g. several red algal genera such as *Portieria* (Payo et al., 2012) and *Gracilaria* (Destombe, Valero, & Guillemin, 2010); brown algal genus *Ectocarpus* (Montecinos et al., 2017)and *Pylaiella* (Geoffroy, Mauger, De Jode, Le Gall, & Destombe, 2015); diatom genus *Pseudonitzschia* (Lundholm et al., 2012; Percopo et al., 2016). The recent advances in molecular approaches and Next Generation Sequencing techniques are great allies in the complex species / cryptic species characterization process (Chomérat, Sellos, Zentz, & Nézan, 2010; Laza-Martinez, Orive, & Miguel, 2011; Le Gac et al., 2016; Murray, Garby, Hoppenrath, & Neilan, 2012). They may have huge impacts on both conservation problematics and interspecific eco-evolutionary assumptions (Bickford et al., 2007) as these genetically distinct species can display different ecological preferendum. For instance, Henderiks et al. (2012)observed two cryptic species in *Emiliania huxleyi* as part of a succession mostly governed by environmental gradients : one species living in late-stage upwelling waters (balanced nutrient concentrations, warm waters) while the other species lives in oceanic waters (high phosphate and low nitrogen, cold waters), highlighting the importance of describing new cryptic diversity for understanding species eco-evolutionary dynamics. However, besides phenology studies, more research is needed to gain insights on these microbial communities cycles, both in terms of gene expression variations and genetic dynamics, because the links and interactions between ecological and genetic divergence are still poorly understood. Particularly, what is the degree of divergence in terms of ecology, genetics and molecular physiology between sympatric cryptic species ? And how do these levels of divergence influence each other : do they use the same metabolic pathway, or do they rather use different strategies in order to avoid direct competition ? When considering expression divergence, what are the respective proportions of constitutive and dynamic responses ? These are fundamental questions that are still to be investigated to understand the eco-evolutionary framework of microbial species communities. *Alexandrium minutum* is a unicellular toxic dinoflagellate regularly causing Harmful Algal Blooms in temperate waters. It has been classified as an invasive species along the European coasts. In the bay of Brest, HAB became more frequent and intense in the past few decades, drastically increasing recently, with cell concentrations reaching up to tens of millions of cells per liters in 2012. Since 1990, the REPHY (a French phytoplankton monitoring network) records *A. minutum* cell concentrations and alerts about paralytic shellfish poisoning risks, regularly leading to administrative closure of shellfish farms. Dia et al. (2014) showed significant spatio-temporal genetic differentiations during and between *A. minutum* bloom events along Brittany coast. Two distinct morphotypes are described in *A. minutum*, distinguishable by the presence/absence of a ventral pore. Cells displaying this ventral pore are more frequently encountered in Southern Europe (M. Giacobbe & Maimone, 1994; Montresor, Marino, Zingone, & Dafnis G., 1990), Asia (Hwang & Lu, 2000; Su & Chiang, 1991; Yoshida et al., 2000) and along Australian (Hallegraeff et al., 1991)and New Zealand coasts (F. Chang, Anderson, Kulis, & Till, 1997; MacKenzie & Berkett, 1997), contrary to cells without ventral pore, mostly found in Northern Europe. Interestingly, these two morphotypes aren't in a complete allopatry and are observed in mixed populations in Ireland (Hansen et al., 2003), Sweden (Kuylenstierna & Karlson, 2000)and France (Belin, 1993; Le Gac et al., 2016; Ledoux et al., 1993; see Hansen et al., 2003, for review). Using transcriptome wide SNP markers, Le Gac et al. (2016)have recently demonstrated that these morphotypes correspond to different genetic entities, suggesting the occurrence of two pseudo cryptic species under the name of *A. minutum* living in the same area. In this context and taking the advantage of Brittany waters as known sympatry areas of the two cryptic species of *A. minutum* described in Le Gac et al. (2016), we tested our ability to track them in the environment, and then to analyse the extent of their gene expression divergence.

Until recently, following cryptic microbial species or intra-specific dynamics usually required isolating single cells from their natural environment before obtaining clonal cultures (Dia et al., 2014; Lebret et al., 2012; Sjöqvist & Kremp, 2016; Toulza et al., 2010). However, more than only being time consuming, this cultivation step only allows for an extremely partial sampling of the microbial populations of interest (a few tens of cells at best for populations that may be composed of several billions of cells). Another potential strong bias is introduced by the frequently low percentage of the isolated cells actually turning into cultures. Applying NGS technologies to *in situ* samples offers the promise to follow these species and their expression dynamic divergence at a population wide scale in their natural environment. First, using a set of fixed SNP positions, we analyzed the respective specific allele frequencies of the two *A. minutum* cryptic species over three consecutive blooms of the bay of Brest and two other geographically distant sampling sites. In a second time, we used transcriptomic data to analyse the causes of gene expression divergences : both from growth conditions (between cells blooming *in situ* and living *in vitro*) and from genetic background (between cryptic species). Using these approaches, we addressed the following questions: Are the two cryptic species present at the same time in natural *A. minutum* blooms ? Are their proportions varying in time and/or space ? To what extent does these species diverge in terms of molecular physiology in their natural environment ? Are gene expression divergence patterns constitutive between the two species or do they correspond to response to environmental variation ?

2.3.3 Material and methods

Sample collection

***In vitro* samples :** *In vitro* samples come from 18 strains founded after micropipetting individual cells. Strains were grown to mid exponential phase in K medium at 18°C, 12/12 photoperiod (80μ E.s^{-1}). The presence of a ventral pore was assessed, RNA was extracted, sequenced and assembled into a de novo transcriptome that was then annotated using the UNIPROT/Swissprot database and classified in Gene Ontology categories (GO), as described in Le Gac et al. (2016). Hereafter, cells without ventral pore will be referred as “cryptic species A“ and cells that have one as "cryptic species B". Among the 18 strains, 15 belong to cryptic species A and 3 to cryptic species B.

Environmental samples : We sampled three consecutive blooms during 2013, 2014 and 2015 summers. Twice a week, when *A. minutum* densities reached 10,000 cells.L^{-1} and until they dropped below this threshold in two consecutive samples, the micro-phytoplankton community was sampled at a single point in the bay of Brest (called "Daoulas", see figure figure 1; 48.350543, -4.292507) (see Metegnier et al. (2018a *Submitted*) for details). In addition, ephemeral blooms were sampled the 07-01-2015 in the Penzé river (48.644633, -3.950772) and the 07-06-2015 in the south of the bay of Brest at the "Sillon des Anglais" (SDA) site (48.299141, -4.288425), resulting in 43 environmental samples (41 temporal samples at the single Daoulas site - 10 in 2013, 19 in 2014 and 12 in 2015 - and two spatially distributed samples respectively at Sillon des Anglais and Penzé river; see figure 2.1 and details in Metegnier et al. (2018a *Submitted*)).

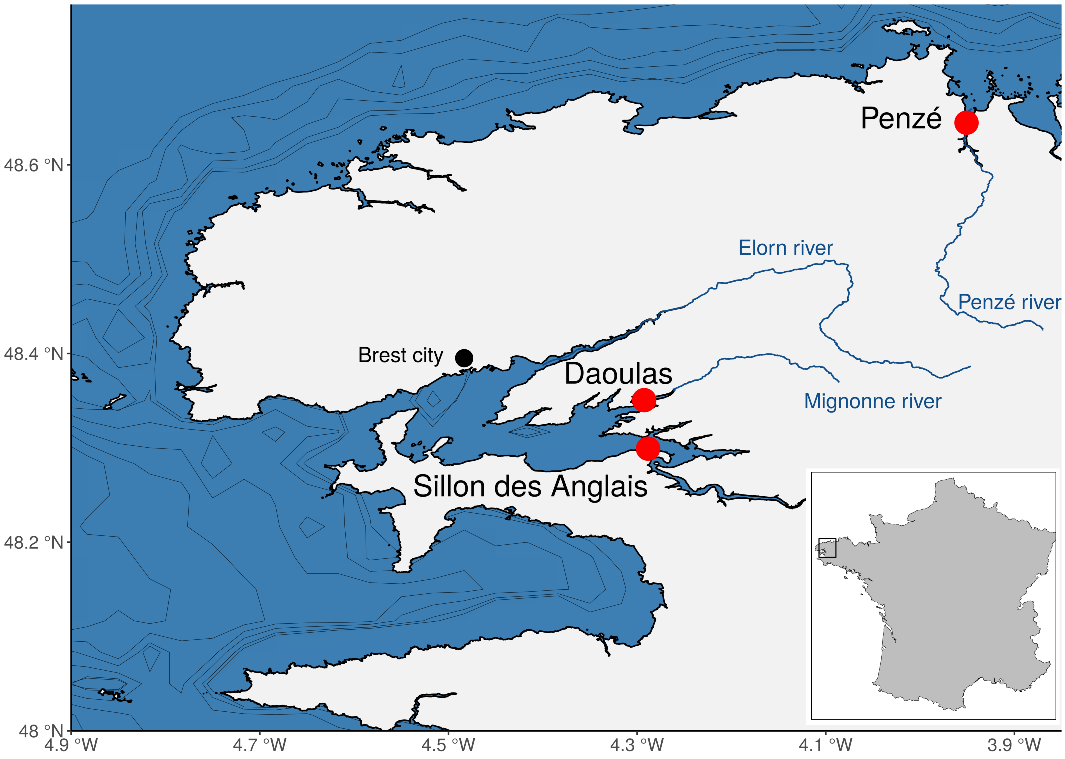


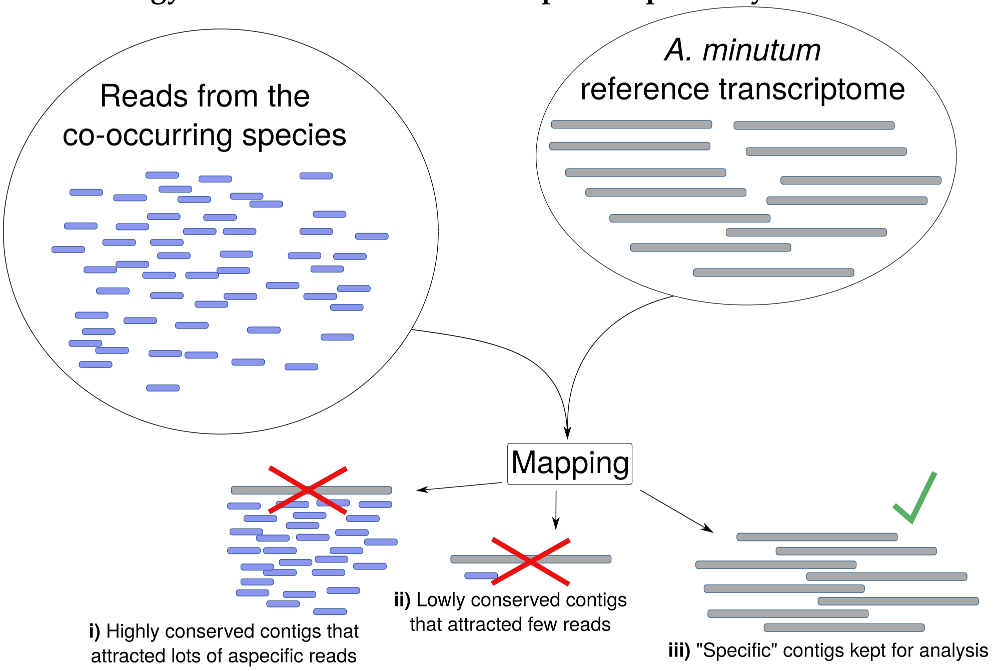
FIGURE 2.1: *Sampling map*

Specificity of the alignment and read mapping

As detailed and described in Metegnier et al. (2018a *Submitted*) (see chapter 3, material and method), all the transcripts from the *A. minutum* reference transcriptome (Le Gac et al., 2016)that attracted at least one read, using BOWTIE2 with sensitive set of parameters (V. 2.2.4; Langmead and Salzberg (2012)), from 24 species (13 dinoflagellates, 10 diatoms (respectively 3 and 7 from classes Bacillariophyceae and Coscinodiscophyceae) and 1 Labyrinthulomycetes) often co-occurring with *A. minutum* near our study site (48.3334355, -4.3264413) were removed to assure *A. minutum* specific gene expression monitoring (to exclude the *A. minutum* transcripts to which reads from other species could align due to sequence similarity, see box 5). After this step, and prior to *A. minutum*'s read mapping, low quality reads and sequencing adaptors were removed using TRIMMOMATIC (V. 0.33; Bolger, Lohse, and Usadel (2014)) with parameters ILLUMINACLIP:Adapters.fasta:2:30:10, LEADING:3, TRAILING:3, MAXINFO:80:0.8 and MINLEN:70. High quality mRNA reads from both strains and environmental samples were then realigned to the reference transcriptome (Le Gac et al., 2016) using BOWTIE2 with sensitive set of parameters. Gene expression abundances were assessed by counting the number of reads mapping to each transcripts using the SAMTOOLS idxstats tool (V. 0.1.19; H. Li et al. (2009)).

Box 5

Methodology used to control for transcriptome specificity :


Reads from the co-occurring species
A. minutum reference transcriptome
Mapping
i) Highly conserved contigs that attracted lots of aspecific reads
ii) Lowly conserved contigs that attracted few reads
iii) "Specific" contigs kept for analysis

The strategy used to remove from the reference transcriptomes contigs that may attract unwanted reads (*i.e.* not from *A. minutum*) from the environmental metatranscriptomic samples. Three situations were encountered : (i) some contigs attracted high numbers of reads from the co-occurring species, potentially indicating sequences that had important cross-species sequence homologies, maybe due to the highly sequence conservation of genes involved in fundamental functions; (ii) some contigs attracted very few reads, sometimes only one (contigs to which reads were mapped in (i) and (ii) were discarded) and (iii) the majority of the contigs did not attract any reds from the co-occurring species, and were therefore kept for the analysis.

Cryptic species assemblages in the blooms

To monitor the composition and dynamics of the two *A. minutum* cryptic species throughout the bloom events, we analysed the respective allelic frequencies at the SNP positions described as fixed between cryptic species A and B in Le Gac et al. (2016). These fixed SNPs occured at 12,188 positions where all strains from one cryptic species had a different allele from the strains of the other cryptic species. Among these 12,188 positions, less than 160 (1.3%) are expected to be false positives, based on the maximum number of fixed SNPs observed for all possible combinations of three and fifteen strains (supplementary figure A.1). In order to be able to calculate allelic frequencies, a minimum depth of 10 at each SNP in each sample analysed was set as a depth threshold. Allelic frequencies were reported as the proportion of reads with allele from cryptic species A and B respectively, covering each SNP position considered.

Expression divergence

Two levels of gene expression divergence in *A. minutum* were investigated : first, a differential gene expression analysis at the inter-species level (between the cryptic species A and B) *in vitro* was performed (as in Le Gac et al. (2016)). Second, in order to investigate the effect of growth culture (*in vitro* vs *in situ*) on gene expression, a differential gene expression analysis was performed between cells from cryptic species A (i) cultivated *in vitro* and (ii) blooming *in situ*. Of the 43 *in situ* samples, only the ones which had at least 10 % of their transcripts covered at least 50 times were considered as informative enough and kept for the analysis. Expression levels were assessed using the same protocol as for the strains *in vitro*. The variance stabilising transformation method from DESeq2 package (V 1.12.4; Love et al. (2014))was used to turn count data into homoscedastic values. First, after correcting for year specific batch effect between the three sampling years using ComBat (Leek, Johnson, Parker, Jaffe, & Storey, 2012), a Principal Component Analysis was conducted to visualize expression pattern divergence between the analysed samples. Second, we analysed differential expression patterns between both cryptic species and between strains of cryptic species A and *in situ* samples collected at the Daoulas sampling site using DESeq2 at q.value = 0.001 as significance threshold. Over and under representation of GO terms was conducted using two tailed fisher exact tests with a 0.1 FDR threshold for multiple test type 1 error correction. The GO enrichment analysis may be subjected to redundancies due to the hierarchical Gene Ontology classification system. To mitigate such redundancy, GO categories with an overlap coefficient (OC) higher than 75 % of their transcripts were clustered, and only the GO category displaying the lowest q-value was reported. Given the number of transcripts A and B in sets A and B, OC is calculated as :

$$OC = \frac{A \cap B}{Min(A, B)}$$

2.3.4 Results

Transcriptome specificity, read mapping and sample filtering

To ensure the alignment specificity of the environmental sequences on the reference transcriptome, we excluded from the *A. minutum* reference the transcripts that may attract environmental sequences from other species. With this perspective in mind, raw reads of 24 microphytoplankton species detected in the water at the time of sampling were aligned against *A. minutum* reference transcriptome and transcripts to which at least one read from the tested species mapped were removed, resulting in 142,797 transcripts left for the analysis (detailed results are presented in Metegnier et al. (2018a *Submitted*)). After this step, we aligned the raw reads from 18 strains cultivated in the lab and 43 environmental samples as described in Le Gac et al. (2016). After filtering out samples with less than 10 % of their transcripts covered with at least 50 reads, 29 samples (6 in 2013; 14 in 2014; 7 in 2015, the Penzé and the Sillon des Anglais samples) were considered as informative enough and kept in the analysis. Following the filtering of the SNP positions that were shown to be markers of A and B cryptic species (Le Gac et al., 2016), 918 SNPs had a sufficient depth of 10 in each of the previously selected samples and were used to monitor allelic frequencies variations at both time and space scales through the respective proportions of the respective alleles of these two cryptic species.

Cryptic species assemblages into the blooms

To investigate the co-occurence of the two cryptic species *in situ*, the allelic proportion at SNPs positions having at least a depth of 10 in each sample (918 SNPs, 584 transcripts) previously identified as diagnostic of the two species (Le Gac et al., 2016), were analyzed during the blooms (figure 2.2). Both intra and inter years, this transcriptome wide screening shows that cryptic species A greatly dominates and might even be the only species in the populations at the Daoulas sampling site, with mean allelic proportion recorded constantly above 95 %. Across the time series in Daoulas, several positions show genetic variability : 793 SNPs displayed two allelic states at least once (only 35 SNPs systematically displayed more than one allele over

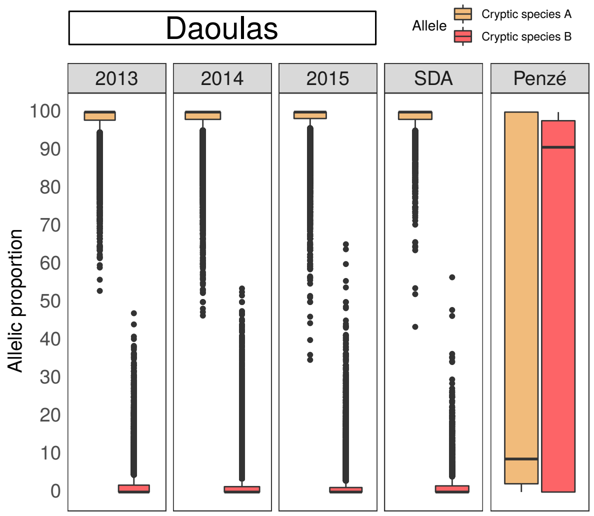


FIGURE 2.2: *Proportion of alleles from both cryptic species at the Daoulas (average over all samples from this site per year), Sillon des Anglais (SDA) and Penzé sampling site (918 SNP positions analysed). Are represented the median (black line), 25 and 75 percentiles (lower and upper hinges) and 1.5 IQR from the respective hinge (lower and upper hinges). Outliers (SNPs beyond the end of the whiskers) are represented in light gray)*

the three sampled years; see figure 2.2, supplementary figure A.2 and supplementary figure A.5). The number of SNP positions showing more than one allele ranged from 183 (08-05-13) to 442 (06-10-14; supplementary figure A.4), and is linked to sample coverage (see supplementary figure A.4). Indeed, the SNP positions showing multiallelic states are not conserved across the samples, as their shared number decreases with the addition of new samples (see supplementary figure A.5). This can suggest the co-occurence of both cryptic species (with a small to undetectable amount of cryptic species B in the samples) or simply noise due to sequencing errors (however, although they represent low allele prevalences, they are always at least 2.7 times higher than basal Illumina error rates (Glenn (2011); data not shown)). All in all, we conservatively consider that Daoulas sampled blooms are only composed of cells from cryptic species A. At the bay level, figure 2.2 and supplementary figure A.2 shows that this pattern is the same at the SDA sampling site. In the Penzé bloom however, the mean allelic proportion of alleles from cryptic species A and B are respectively 36.71 % (± 43.42), and 62.16 % (± 43.83), suggesting the co-occurence of both cryptic species in this sample. This contrasting situation is investigated in more details and at different levels in the next subsections. To do this, the allelic frequencies of both cryptic species are re-analysed in the Penzé sample by using a larger amount of SNP positions (*i.e.* all SNPs displaying a depth of 10 in this population; resulting in 8,815 positions).

Gene expression divergence in *A. minutum*

A genetic and environmental interaction : Figure 2.3A shows that the respective proportions of alleles from cryptic species A and B in the Penzé bloom displays an extremely bimodal distribution. This result is highly surprising, as for the analyzed SNPs, depending on the transcripts, almost only alleles corresponding to species A or to species B are sequenced. This suggests an extremely strong gene differential expression between the two species *in situ*. This pattern is independent from the amount of analysed SNPs and conserved whether we consider positions displaying a depth of 10 in all sampled populations (918 SNPs, see supplementary figure A.3) or SNPs with a depth of 10 only in the Penzé sample (8,815; see figure 2.3A). In a previous study, we compared the differential expression between strains belonging to species A and B *in vitro* (Le Gac et al., 2016). By re-analysing this dataset, it was clear that the transcripts that could be assigned to either species and displaying strong differential expression *in situ* (figure 2.3A) did not display such a pattern *in vitro* (figure 2.3B, y-axis). Similarly, these same transcripts did not display a pattern of differential expression when comparing cryptic species A *in vitro* and *in situ* (Daoulas samples) (figure 2.3B, x-axis). The bimodal distribution observed in figure 2.3A suggests a differential expression pattern of genes between cryptic species A and B. However, the first results on gene expression divergence observed in figure 2.3B (only conducted on a restricted set of transcripts - the ones presenting fixed genetic differences between species A and B -, figure 2.3A) show that the expression divergence observed *in situ* and *in vitro* are different. To obtain a broader view of gene expression divergence induced by both genetic background and growing conditions (*in vitro* vs *in situ*) in *A. minutum*, we used the Daoulas sampling site as an in situ growing-condition experimental group for species A (as we characterized it as only composed of cells belonging to this species; see figure 2.2 and supplementary figure A.2), and analysed expression divergence at a transcriptome wide scale (142,797 transcripts).

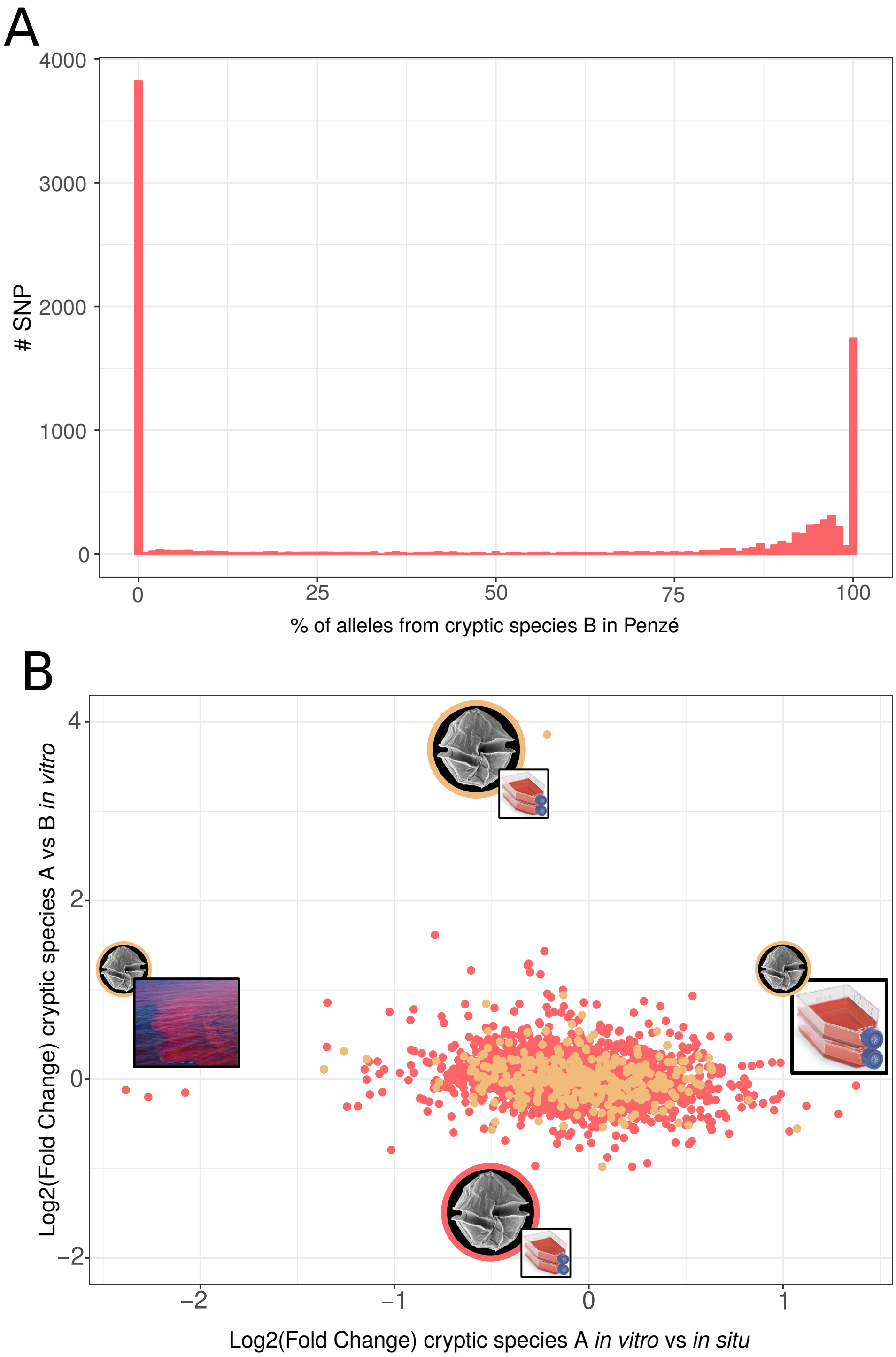


FIGURE 2.3: ***A*** *: Histogram representing the distribution of alleles from cryptic species B among 8,815 SNP positions studied (having a depth of 10 in the Penzé sample).* ***B*** *: Expression divergence of the transcripts displaying excesses of alleles from both cryptic species in the Penzé bloom. Yellow and red dots are transcripts showing respectively only alleles from cryptic species A or B in the Penze sample. X axis represents Log2 (fold change) of cryptic species A expression levels grown* in situ *(in Daoulas) vs* in vitro. *Y axis represents Log2 (fold change) expression levels of cryptic species A vs B grown* in vitro.

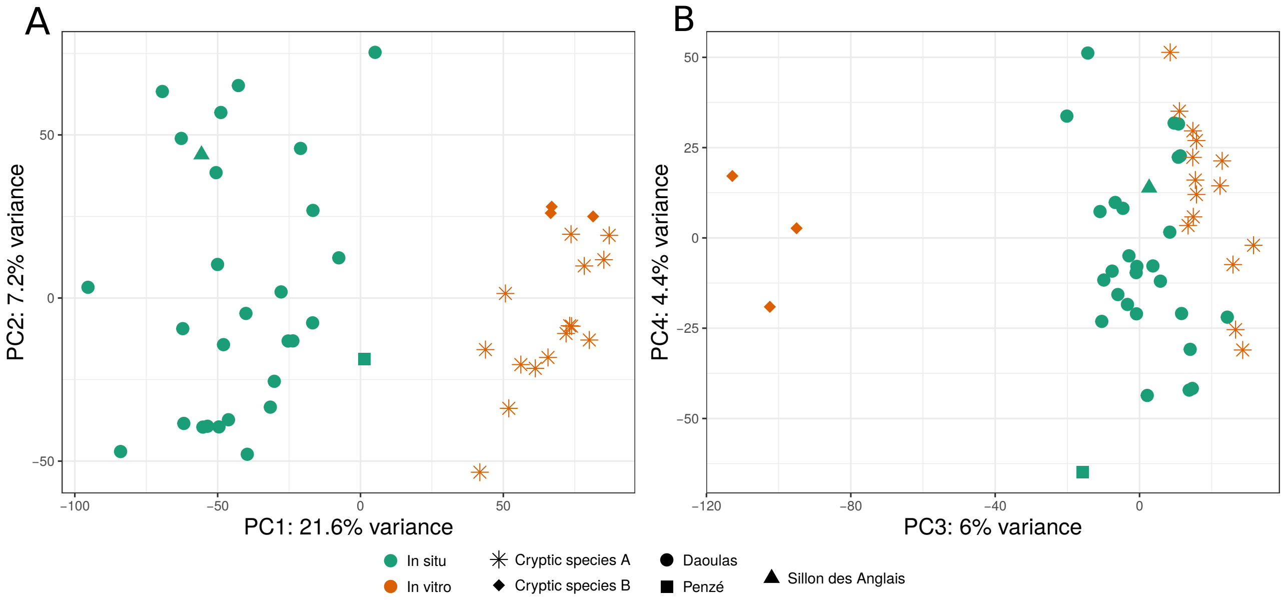


FIGURE 2.4: ***A, B*** *: First four axes of the PCA led on gene expression levels of both* A. minutum *populations sampled* in situ *and strains cultivated* in vitro.

Figure 2.4A and B represent the first four axes of the PCA analysis led on the gene expression levels between *in vitro* (18 strains; 15 from cryptic species A and 3 from cryptic species B) and *in situ* samples (27 samples from a single sampling site, one from Sillon des Anglais area and one from the Penzé river). The first axis explains more than 21% of the variance and mainly separates *in situ* from *in vitro* samples. The third axis explains 6 % of the variance and mainly separates strains of cryptic species B cultured in the lab from the other samples. The second and fourth axis explain respectively 7.2 and 4.4% of the observed variance and don't clearly separate the groups of samples.

Inter species gene expression divergence : Genetic divergence effect on gene expression is represented by the third axis of the PCA presented in figure 2.4B. The differential expression analysis we conducted between the two cryptic species grown *in vitro* uncovered 1,469 transcripts (1.03 % of the transcriptome; 108 annotated) at our significance threshold, 985 upregulated in cryptic species A (62 annotated) and 484 in cryptic species B (46 annotated) (data not shown). No GO term was over or under represented in the differentially expressed transcripts in both cryptic species. Using slightly different sets of parameters and other differential expression analysis methods, Le Gac et al. (2016) showed the same patterns.

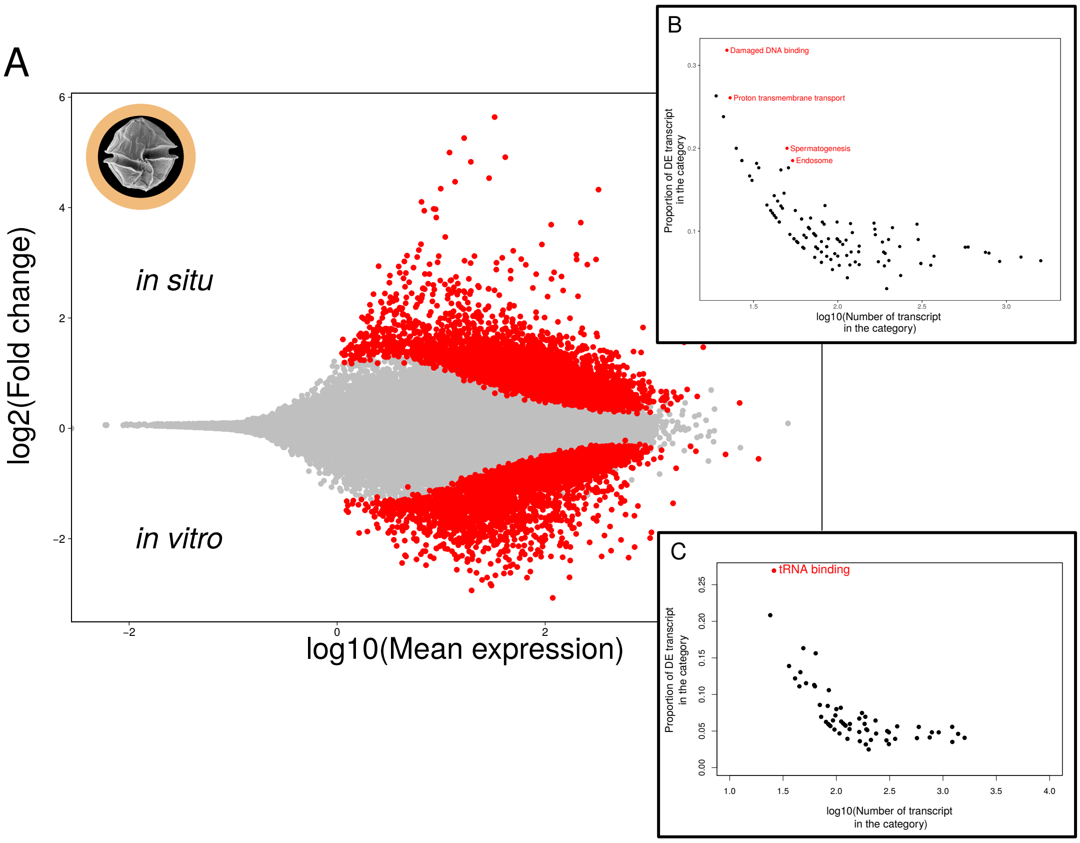


FIGURE 2.5: ***A** : Differential gene expression analysis of cells from cryptic species A sampled* in situ *and* in vitro*. **B, C** : Gene ontology terms respectively over represented in transcripts more expressed* in situ *(B, up) and* in vitro *(C, down).*

Condition specific gene differential expression : As shown by the first axis of the PCA presented in figure 2.4A, the main effect on gene expression divergence comes from the distinction between *in vitro* and *in situ* growing conditions. Out of 8,361 transcripts (5.86 % of the transcriptome) differentially expressed between strains from cryptic species A cultivated in the lab and *in situ* samples, 4,456 transcripts (908 annotated) were upregulated in the lab samples and 3,905 (564 annotated) *in situ* (figure 2.5A). At FDR = 0.1, five GO terms were over represented in the DE transcripts : four in the genes overexpressed *in vitro* (figure 2.5C) and one in the genes overexpressed *in situ* (figure 2.5B) . The genes more expressed *in vitro* are implicated in two main functions. The first group is related to cell division, cell maturation and DNA maintenance (“damaged DNA binding” (GO:0003684), “spermatogenesis” (GO:0007283)), with differentially expressed transcripts related to analogs of adenylate cyclase type 10 protein (that produces the cAMP which regulates cAMP-responsive nuclear factors indispensable for sperm maturation in mammalian spermatogenesis), ATP-dependent RNA helicase spindle-E (implicated in germline integrity during gametogenesis by repressing transposable elements and preventing their mobilization), Centrin-2 protein (regulates cytokinesis and genome stability, involved in global genome nucleotide excision repair and DNA damage repair) and numerous other DNA damage repair related proteins (UV excision repair protein RAD23, DNA mismatch repair protein MSH1, DNA polymerase iota and kappa, DNA repair protein REV1). Also *in vitro* are functions related to cellular intake and transmembrane transport (GO:0005768 (endosome) and GO:0015992 (proton transmembrane transport)), with transcripts homologs to peroxisomal adenine nucleotide transporter 1, K(+)-insensitive pyrophosphate-energized proton pump, Inositol transporter 1, Proton-coupled amino acid transporter 1, Kinesin-like protein, Phospholipid-transporting ATPase, Rab11 family-interacting protein 3, EH domain-containing protein 1. There are also two transcripts implicated in photosynthesis processes : one related to a sulfhydryl oxidase 1 protein (reduction of oxygen to hydrogen peroxide) and one to phosphatidylinositol transfer protein CSR1 (resistance to oxidative stress). The transcripts more expressed in situ are enriched in rRNA binding functions (GO:0019843), that gather transcripts implicated in the translation machinery. 3 of these transcripts (comp106026_c0_seq1, comp87794_c0_seq1, comp22514_c0_seq1) have homologies with the large subunit (respectively protein L1, L4 and L16) of the ribosomal machinery, and one have homology with the small subunit (comp98082_c0_seq1, protein S13). Among these transcripts are also two transcripts showing homologies with proteins involved in RNA processing (Ribonuclease 3 and an U3 small nucleolar ribonucleoprotein), such as ribonuclease and ribosomal proteins (see supplementary table A.6).

2.3.5 Discussion

To survey bloom dynamics and expression divergence of two cryptic species (A and B) over long periods of time *in situ*, we used the advantages of metatranscriptomics to sequence mRNA from entire populations of *A. minutum* in their natural environment. Using fixed SNPs markers, we monitored the proportion of two species at different scales: first at a single sampling site over a period of three years in the bay of Brest, and second at a spatial scale by comparing the previous sampling site and two other areas, respectively located in the south of the bay (at the Sillon des Anglais site) and in the Penzé river. In a complement, we investigated the gene expression divergence in *A. minutum* at several levels : between species, between living conditions *(in situ* vs *in vitro*) and both of them.

Bloom compositions : a contrasting situation

In the bay of Brest, the sampled blooms are systematically composed of more than 95 % of the cryptic species A, if not only composed of this species. Indeed, the small amount of other alleles recorded, their unconstant distribution across the analysed SNPs and their direct link with sequencing depth make their analysis somehow difficult. Without being possible to completely rule out the presence of the other cryptic species (species B), its proportions and dynamics are below analysis thresholds. Le Gac et al. (2016)however, sampled one strain from this species (species B) in the bay of Brest in 2011. Excluding the hypothesis that the sampling of this species was purely accidental, the bloom assemblages observed here may be the result of a strong competitive exclusion between the two cryptic species, cryptic species A ruling out cryptic species B in the bay of Brest. For instance, Gabaldón, Carmona, Montero-Pau, and Serra (2015)showed that two cryptic species of Rotifers dynamically coexisting in lakes of the Iberian peninsula outcompete each other under different environmental conditions. However, the apparent stability of bloom composition across such long periods of time might rather suggests durable installation and local adaptation of species A in the bay of Brest environment. Unfortunately, the present study does not allow to favour one or another hypothesis, but encourages more investigations of both species ecological preferendum and interactions. Contrary to the bay of Brest where blooms are composed of a single cryptic species, the situation observed in the Penzé bloom is different. Cells from both cryptic species A and B are present at the same time in the water in apparent much greater proportions. However, their respective proportions can't be determined precisely due to the transcription-level dependant nature of our detection method (see below). Ecological controls of cell type distributions and dynamics were studied in few plankton species, including the bloom forming coccolithophore *Emiliania huxleyi* (Hagino, Okada, & Matsuoka, 2005; Henderiks et al., 2012) and the diatom *Ditylum brightwellii*

(Rynearson et al., 2006). In *A. minutum*, the global distribution of cells from cryptic species with ventral pore (species B) is generally accepted to be restricted to southern waters (from southern Europe to New Zealand) and cryptic species without ventral pore (species A) to northern Europe waters (see review in Hansen et al. (2003)) with few mixed populations, supporting ecological divergence. The complex assemblage of these two cryptic species in the Penzé population may be the result of biological controls. Indeed, population structures and compositions can be greatly modeled by species specific interactions, as shown in *A. minutum*, that harbors complex strain specific relationships with different species of the parasite *Parvilucifera sinerae* (Figueroa, Garcés, Massana, & Camp, 2008).

Gene expression divergence analysis

As reviewed in Wray et al. (2003), evolutionary changes rely on the genotype-phenotype interplay, with gene expression acting as a cornerstone. Although we know that our transcriptomic approach might led us to miss trans-regulatory indices of inter species divergence, we believe that our methodology can still permit us to track the interaction between genetic and expression variations through the analysis of cis-acting regulatory elements (Pai, Pritchard, & Gilad, 2015).

The identification of transcripts gathering an almost perfect exclusivity of alleles from the two cryptic species in the Penzé bloom suggested specific preferential gene expression. In this context, we first analysed gene expression divergence between strains from the two cryptic species cultivated in the lab and in a second time the effect of *in situ* and *in vitro* growth conditions on their molecular physiology. Between the two cryptic species *in vitro*, we observed a relatively low amount of gene expression divergence (1.03 % of the analysed transcripts). This result was previously showed by Le Gac et al. (2016)and, even if the number of differentially expressed genes is quite small, it might still be important in the light of the low gene expression responses observed in dinoflagellates (Alexander, Jenkins, et al., 2015; Parkinson et al., 2016). In other studies, I. Yang et al. (2010)(between toxic and non toxic *A. minutum* strains) and Wohlrab, Tillmann, Cembella, and John (2016)(between two *A. fundyense* strains) also showed a relatively small (respectively 192 genes in *A.*

minutum and around 5% of the tested genes in *A. fundyense*) constitutive molecular physiology divergence in *Alexandrium* species. In many other lineages, several studies also showed that expression levels were affected by genetic variation (see for instance Deng et al. (2018)in bacteria; Francesconi and Lehner (2014)in nematoda, Ackermann, Sikora-Wohlfeld, and Beyer (2013)in murine, Monks et al. (2004)in humans). However, the extent of genetically-based effects on global transcription levels are still poorly known. For instance, Turk et al. (2004)analysed mouse strains transcriptome profiles, and estimated that genetic background effect on gene expression was approximately ten-fold lower than a disease-causing mutation. Here, we studied the expression divergence between these two cryptic species to test the hypothesis whether they would have specific gene expression signatures, as suggested by the almost exclusive presence of their respective alleles at several SNP positions. Although we acknowledge our inability to analyse cryptic species B specific gene expression divergence *in vitro* vs *in situ*, due to the lack of populations only composed of this cryptic species (contrary to cryptic species A), our sampling design allowed us to investigate other levels of gene expression divergence in *A. minutum*.

Interestingly, we did not find clear evidence of exclusive specific gene expression patterns : the transcripts gathering in majority alleles from each cryptic species in the Penzé bloom were not specifically or even differentially expressed by the respective cryptic species *in vitro*. We also note that some transcripts that only display one cryptic species alleles *in situ* are seen to be more expressed by the other cryptic species *in vitro*. In *Daphnia*, Herrmann, Ravindran, Schwenk, and Cordellier (2018)studied the links between sequence and expression variations and showed their dual effect on divergence between populations responding to thermal variations. Without being able to precisely evaluate their respective contributions, they show that local selection acts at the genes sequence level, without constitutively changing their expression levels. They also suggest condition specific effects of sequence divergence on expression levels. Here, these observations rule out the completely exclusive expression of genes between the two cryptic species, but rather suggest a preferential expression of genes, and then the use of different metabolic pathways, in response to specific conditions and underlines molecular physiology plasticity under contrasting conditions. Cohen et al. (2017)investigated gene expression responses divergence in two cosmopolitan diatom genera (*Pseudo-nitzschia* and *Thalassiosira*) subjected to environmental stresses. Firstly, they showed that gene expression is significantly impacted by the nature of their environment (oceanic versus coastal) and secondly, that different species are able to use diverse response strategies to similar changes. Similarly, Alexander, Jenkins, et al. (2015) showed niche partitioning in two diatom species showing opposite gene expression activities under the same environment, which could be a response to direct interspecies resource competition and/or local adaptation. Differential tolerance and responses to environmental constraints has been investigated in various contexts between cryptic species, such as, for instance, temperature and salinity stress life-history responses in nematode (De Meester, Derycke, Rigaux, & Moens, 2015) or tolerance to pollution in a meiobenthic copepod (Rocha-Olivares, Fleeger, & Foltz, 2004). Not surprisingly, such different biological responses can also be characterized using a molecular approach, as in the differential gene expression response observed between two whiteflies cryptic species to host plants (H. X. Xu et al., 2015). In two cryptic species of the earthworm *Lumbricus rubellus*, Liebeke et al. (2014)did not distinguish them with obviously distinct metabolic levels but rather with discrete differential metabolite use. Conversely, Rose, Bay, Morikawa, and Palumbi (2018)studied the genomic basis of thermal adaptation in of the tabletop coral *Acropora hyacinthus* and showed widespread divergence in gene expression between two cryptic species, associated with widespread allele frequency divergence but few fixed differences between species. These studies highlight the difficulty to discriminate the genomic mechanisms that drive species ecological divergence. Here, the analysis we conducted on spatio temporal transcript abundances variations between 18 *in vitro* cell cultures and *in situ* sampled bloom revealed that molecular physiology variations is influenced by two main variables (living conditions -*in situ* or *in vitro*- and genetic background of the cells), acting as sets of factors interacting together. In terms of ecological significance, considering the observation of divergent gene expression levels under controlled laboratory conditions and physiological responses under natural environmental conditions (Penzé population), it underlines the potential ability of species belonging to the *Alexandrium* genus to colonize new areas and cause harmful algal blooms in diverse, changing environments. Although *A. minutum* is known to be present in the bay of Brest for at least a century (Klouch et al., 2016), we can hypothesise that the recruitment of one of the two cryptic species studied here is responsible for the increase in bloom frequencies and intensities observed in the past decades. This highlights the importance of studying cryptic species, as it their inaccurate distinction and monitoring can greatly alter our ability to conserve or manage them (see Bickford et al. (2007)for the importance of considering cryptic species diversity in conservation problematics). The functional enrichment analysis revealed that cells from cryptic species A growing *in vitro* are over expressing transcripts involved global maintenance functions such as internalization processes, transmembrane transport and DNA management. On the other hand, cryptic species A cells living *in situ* are experiencing high expression of functions involved in rRNA binding proteins machinery, suggesting high gene expression regulation. The most important gene expression divergence observed here is between *in vitro* and *in situ* conditions. This result is especially interesting, as the *in situ* samples responded to a wide variety of environmental conditions, ranging from oceanic-like conditions (high salinity and moderate nutrient concentrations) to more estuarine-influenced waters (fluctuating temperatures, high nutrient concentrations, Metegnier et al. (2018a *Submitted*)). It clearly highlights the fact that the *in vitro* environment is drastically different from natural conditions and that it has a very strong impact on global expression profiles. Together with the results of Le Gac et al. (2016), the present study illustrates the mixed effects of both genetic background and growth context on gene expression levels. Indeed, while Penzé population is composed of both cryptic species, and that these two cell types display gene expression differences, the analysis we performed here did not allow us to conclude (and quantify their respective effects) whether this expression divergence comes from abiotic (different environmental characteristics between the Penzé and Daoulas samples) or biotic effects (that can arise from different community compositions, or from the indirect and / or direct interactions between the two species), but still underlines the importance of ecological conditions on transcription levels. Altogether, the results we present here also emphasize the need to consider the potential drawback of using only one or few clonal strains grown *in vitro* in transcriptomic studies, as culture origins of the samples appear to be the main source of variations in gene expression. Genetic difference between field populations and cultivated strains were also hypothesized as responsible for the gene expression discrepancies observed between *in situ* and *in vitro* cultivated *Prochlorococcus* cells in Ottesen et al. (2014). Also, it suggests that investigating the functional divergence of two cryptic species using strictly *in vitro* approaches would be difficult, if not impossible. Although efforts were made (however still rare and mostly restricted to medical purpose) to uncover *in vitro* vs *in vivo* gene expression response (see for instance Heise et al. (2012); Knijn et al. (2005); M. G. Liu, Li, Xu, Barnstable, and Zhang (2008); Zaitseva, Vollenhoven, and Rogers (2006)), to our knowledge, only few studies investigated gene expression divergence between *in vitro* and *in situ* living organisms. Smith and Oliver (2006), studied gene expression divergence between free living and cultivated cells of the gram-negative bacterium *Vibrio vulnificus*. In a study from Ottesen et al. (2013)is compared *in vitro* and *in situ* diel gene expression patterns in several microbial species, including *Ostreococcus* and *Synechococcus*. In both species, they show that the majority of diurnal gene expression patterns observed in laboratory studies were conserved in wild populations. However, they note some discrepancies, with 24 genes that were not observed as dynamic in laboratory cultures while they were in natural *Synechococcus* populations. In the same way, they founded more genes dynamically expressed in the field than in the laboratory in *Ostreococcus*. As in our study, one of their hypothesis is that these discrepancies result from specific field condition responses that laboratory cultures fail to reproduce. More than only highlighting the importance of considering both *in situ* vs *in vitro* approaches to understand molecular physiology adaptations, these studies, together with the present work, give a uncommon but crucial insight on the dual effects of both environment and genetic background on molecular physiology responses. We believe further research on this topic should be done to give a better understanding of the forces driving species gene expression responses, assemblage and dynamics in response to their environment.

2.3.6 Conclusion

In the will to study a cryptic species complex, we monitored both their co-occurrence at the population level, and how their genetic divergence translates in terms of gene expression. Using the advantages of metatranscriptomic data, we tracked the presence of two distinct cryptic species in natural blooms using a set of fixed SNP markers between them. In one population, we showed that cells from cryptic species A are widely dominating, with no outbreaking of the B cryptic species, describing the stability of *A. minutum* blooms composition across long periods of time in this population. In another spatially distant population, we uncovered a new mixing zone for these two cell types, located in the Penzé estuary. Not only monitoring the presence/absence of species in their natural environment, we also investigated their gene expression divergence. One of the key result presented here is dual effect of both genetic background and living conditions on molecular physiology : although both species do express different transcripts *in situ*, this divergence is completely different under controlled conditions, with *in vitro* living cells over expressing functions involved in global cell maintenance while *in situ* growing cells experience high gene expression regulation processes. These apparent discrepancies in interspecies gene differential expression between *in situ* and *in vitro* growing conditions highlight the mixed effect of both genetic background and environmental conditions on gene expression levels. Here, we present an additional evidence for the potentials of reverse ecology in the understanding of species biological dynamics, and a rare example describing the need to use the advantages of both *in vitro* and *in situ* approaches (as hypothesis rising, testing and confirmation tools under both controlled designs and natural conditions) to lead large-scaled, comprehensive molecular physiology divergence studies.

2.3.7 Acknowledgements

The authors want to thank the Region Bretagne for funding the Daoulex program, GetPlage sequencing platform for technical advices and help. We also thank Julien Coïc, Hélène Hégaret as well as members of the Pelagos and LER/BO labs for helps provided during the sampling and members of the UMI 3614 for discussions. We also thank ABIMS and RIC bioinformatic platforms for the technical support.

Chapter 3

Gene expression dynamics *in natura*

3.1 General context

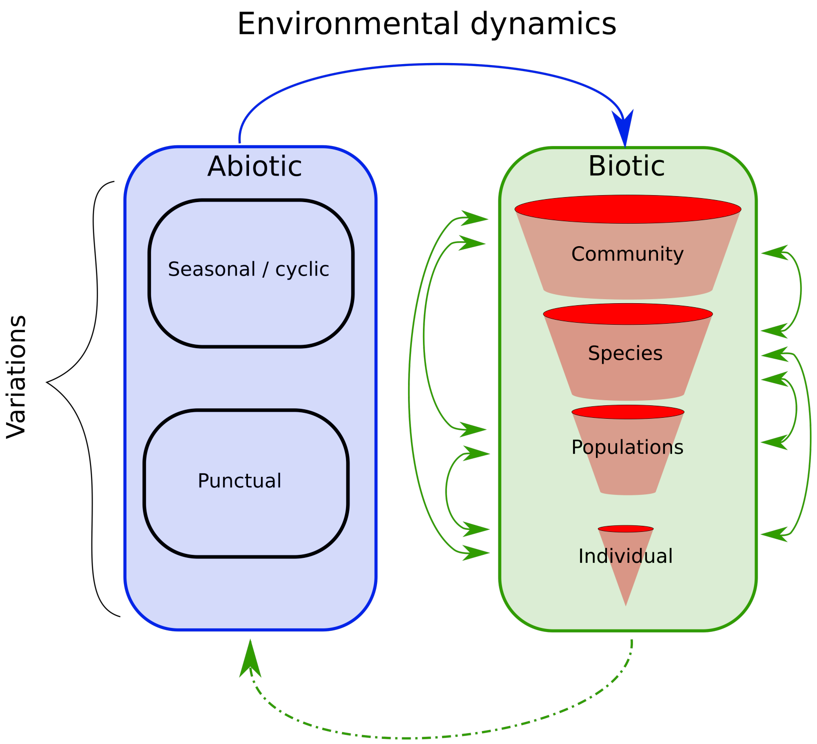


FIGURE 3.1: Schematic representation of environmental dynamics. Here, the ecosystem is separated between abiotic and biotic environmental compartments. Blue arrow represents the variations of both seasonally cyclic dynamics or punctual variations of abiotic factors that will directly or indirectly affect the biotic environment. The biotic environment is represented as layered from entire communities, to species / intra-species / individual scale. Green arrows represent the respective inter-layer effects of each respective biotic entities dynamic.

POPULATIONS constantly face external stimuli, that can arise both from biotic interactions and abiotic factors fluctuations. While populations face external dynamics, they impose in return their own dynamics on the rest of the complex as a feedback loop, all together referred as 'environmental dynamics' (see figure 3.1).

As exposed in the introduction of the present manuscript, populations can respond to a changing environment through two main processes : phenotypic plasticity and genetic adaptation. However, deciphering the relative contribution of each of these processes to population response, and investigating the complex interactions of this response continuum are far from being straightforward.

In the context of phytoplankton blooms, numerous works were carried out to describe the environmental factors impacting bloom formation, growth and decline. Even if their relative contribution to each bloom phase transition isn't fully understood, numerous models and studies described the major factors altering phytoplankton dynamics (combined effects of temperature and nutrients (Maguer et al., 2004), dispersion and dilution (Lehahn et al., 2017), wind (Y. Xu, Cahill, Wilkin, & Schofield, 2012), cell mortality (Choi, Brosnahan, Sehein, Anderson, & Erdner, 2017), grazing (Calbet et al., 2003)). What is less known, however, is the different cellular states (in terms of molecular physiology) phytoplankton populations experience throughout the blooms. Therefore, the aim of the net chapter is to answer the question : **what is the molecular physiology temporal response to environmental fluctuations ?** To do this, we investigated the temporal gene expression dynamics of a microbial species at a population scale, in order to characterize the molecular physiology response to environmental factors fluctuations of entire blooms of *A. minutum* in its natural environment, during three consecutive bloom events (2013, 2014 and 2015).

Author contributions : Water was sampled and filtered by numerous persons from the DYNECO-PELAGOS laboratory before the start of the Ph.D. Sequencing library for prelevements sampled in 2013 and 2014 were prepared by Mickael Le Gac and Julien Quéré, and by myself for 2015 samples. I performed data analysis and manuscript writing, Mickael Le Gac and Christophe Destombe participated to result analysis, discussion construction and manuscript corrections.

3.2 Challenging microorganisms surveys *in situ* : new highlights into the gene expression dynamics of a marine phytoplankton

Title : Challenging microorganisms surveys *in situ* : new highlights into the gene expression dynamics of a marine phytoplankton.

Running title : *In situ* phytoplankton gene expression survey

Authors : METEGNIER Gabriel[1,2], QUERE Julien[1], DESTOMBE Christophe[2], LE GAC Mickael[1]

[1] Ifremer, DYNECO PELAGOS, 29280 Plouzané, France
[2] CNRS, Sorbonne Université, Pontificia Universidad Catolica de Chile, Universidad Austral de Chile, UMI 3614, Evolutionary Biology and Ecology of Algae, Station Biologique de Roscoff, Place Georges Teissier, CS90074 29688, Roscoff Cedex, France

Corresponding author : METEGNIER Gabriel ;
Ifremer, DYNECO PELAGOS, 29280 Plouzané, France ;
Phone number : +33 2 98 22 42 60 ;
E-mail: gabriel.metegnier@gmail.com

Submitted to Molecular Ecology

3.2.1 Abstract

Although transcriptomic analyses now provide easy access to physiological dynamics in experimental environments, understanding the physiological dynamics of microbial species in their natural habitat remains a major challenge. Here, a meta-transcriptomic dataset from three blooms of the toxic dinoflagellate *Alexandrium minutum* was generated and analysed to understand its gene expression dynamics in the field. Special attention was given to ensure the quality of the species-specific molecular physiology dynamics monitored (from the alignment specificity to the separation of gene expression from relative abundance). We were able to identify distinct sets of genes showing *in situ* dynamic expressions and to correlate them to environmental factors fluctuations. Functions that may have seem interesting *a priori*, such as nutrient metabolism, were not among the ones displaying extensive expression dynamics. Of special interest were two major sets of genes gathering unexpected functional categories, both of them showing high expression in cold, low irradiance and salinity environments. One is related to motility behaviors and one suggests increasing cell-to-cell interactions toward the end of the blooming period. This study highlighted the physiological dynamics of *A. minutum* in its environment and gives an unprecedented overview of the possibilities to monitor molecular physiological dynamics of non-model species *in situ*.

3.2.2 Introduction

Background

Understanding what a particular microbial species is doing in its natural habitats and how it reacts to environmental variations remains a major challenge in microbial ecology. Transcriptome-wide analysis has become a popular technique to describe physiological responses of individuals towards various stresses. Often used in controlled designs *in vitro* to characterise fine and specific physiological variations, these experiments are often hard to connect to the natural environment. *In situ* meta-transcriptomic approach based on RNA sequencing is a useful tool to track the metabolic activity of entire microbial communities as they occur in nature and provides information regarding the interactions between microbes and their environments (Gosalbes et al., 2011; Moran et al., 2013; Radax et al., 2012). However, conducting such analysis at the species level remains virtually unexplored, but is potentially promising, especially when few species represent a major fraction of entire communities, such as during micro-algal blooms (Alexander, Jenkins, et al., 2015). Several key points need to be taken into account. First, the alignment of the environmental sequences to the reference transcriptome of the species of interest has to be species-specific. Since meta-transcriptomic datasets are composed of short sequencing reads (typically 100 to 150 bp), this is especially critical when orthologous sequences display a high level of homology. Second, even when considering species dominating entire communities, the relative abundance of the focal species in the community may fluctuate widely. Given a constant sequencing effort across samples, this will translate into wide variation of sequencing effort for the species of interest. As a result, the proportion of transcriptome that can be used to quantify gene expression for the species of interest can be highly variable across sampling points, and special attention has to be paid to only focus on the transcripts displaying quantitative gene expression in all the samples considered. Third, contrary to the classical *in vitro* approach where differential gene expression is investigated between two or more experimental conditions following a pre-established experimental design, investigating *in situ* gene expression dynamics is exploratory and requires the ability to detect expression patterns without grouping the samples *a piori*. Finally, in order to be able to generate testable hypotheses regarding the factors driving gene expression dynamics *in situ*, it is essential to correlate them to environmental factors.

Studied species

The dinoflagellate *Alexandrium minutum* (Halim, 1960)is involved in paralytic shellfish poisoning (PSP) blooms worldwide. This microalga produces saxitoxins that can accumulate along the trophic food web. They produce PSP events that cause cases of human poisoning, typically by ingestion of contaminated shellfish. This species shows an alternation of asexual and sexual phases. The vegetative cell is haploid, and sexual reproduction takes place when the environmental conditions become unfavourable (Figueroa et al., 2007). Sexual reproduction is the result of the fusion between two gametes, forming a planozygote (Figueroa et al., 2015, 2007; Probert, 1999), and finally a resting cyst that sinks to the sediment (like seeds in a seed bank; Wyatt and Jenkinson (1997). After a dormancy period of several months (Figueroa et al., 2007; Garcés et al., 2004), cysts may germinate and give rise to vegetative haploid cells that reproduce asexually in the water column. Drivers of blooms initiation, growth and decline are not well understood. Bloom initiation is assumed to begin with the excystment of resting cells. This excystment is possible after a dormancy period (Figueroa et al., 2007)and may be regulated by temperature, light and oxygen concentration (Anderson et al., 2012; Ní Rathaille & Raine, 2011). Bloom development seems to be regulated mainly by flushing rate (characterised by both tide, wind and river flow) and water stratification (thermal and/or haline) (Anderson, 1998; Anderson et al., 2012; Guallar et al., 2017; Laanaia et al., 2013; Raine, 2014; Sourisseau et al., 2017). Nutrient concentration, which is highly dependent on fluvial flow, is a key factor controlling bloom dynamics (Chapelle et al., 2015a; Guallar et al., 2017; Maguer et al., 2004; Romero et al., 2013). On the other hand, the factors responsible for the ending of the bloom are more difficult to define. Among possible processes involved, several studies have highlighted significant roles of nutrient limitations (Guisande et al., 2002; Laanaia et al., 2013; Labry et al., 2008), predation (Calbet et al., 2003; Sorokin et al., 1996), even if some studies show that grazers have a limited impact on *A. minutum* blooms (Probert, 1999), and parasitism (Chambouvet et al., 2008). Since its first detection in Alexandria harbor, Egypt (Halim, 1960), *A. minutum* blooms have been observed in Spain (Franco et al., 1994), France (Belin, 1993), North Sea (Elbrachter, 1998; Hansen et al., 2003; Nehring, 1998), Ireland (Gross, 1989), Sweden (Persson et al., 2000), Taiwan (Hwang & Lu, 2000) and New Zealand (F. H. Chang et al., 1999). *A. minutum* has even been classified as an invasive species along the European coasts. In the bay of Brest, *A. minutum* cells have been recorded since 1990 by the REPHY (a French phytoplankton monitoring network). Until the recent years, recorded cell densities were relatively low (few thousands cells.L^{-1}), but it drastically increased, reaching up to 42 million cells per liter in 2012. In this context, we sequenced the mRNA of the entire microphytoplanktonic community during three consecutive *A. minutum* summer blooms (2013, 2014, 2015) in the bay of Brest (France). An *in situ* transcriptomic approach was conducted to answer the following questions: Is it possible to monitor the *in situ* gene expression dynamics of a single species of interest ? Can we associate molecular functions with these dynamics? And finally, can we correlate environmental factors to these gene expression patterns and generate hypotheses regarding the environmental factors driving gene expression dynamics *in situ* ?

3.2.3 Material and Methods

Material

Sample collection and RNA preparation : During three consecutive summers (2013, 2014 and 2015), temporal sampling was conducted in the bay of Brest (48.350543, -4.292507). When the water temperature reached 15°C during spring, a weekly survey of *A. minutum* densities was performed by the national phytoplankton network (REPHY). When *A. minutum* densities reached 10,000 cells.L^{-1} and until densities dropped below this threshold in two consecutive samples, the microphytoplanktonic community was sampled twice a week. At the same time, salinity and sea surface temperature were measured *in situ*. The concentration of ammonia, nitrate, nitrite, phosphate and silicate were determined using a Seal Analytical AA3 HR automatic analyser following the method of Aminot and Kerouel (Aminot & Kerouel, 2007). In addition, the time elapsed between sunrise and sampling, the number of days since the beginning of the bloom period (> 10,000 cells.L^{-1}), the flow of the adjacent Mignonne river (from Hydro database), the tidal coefficient (from SHOM database) and the irradiance at sea surface (from MeteoFrance database) at the time of the sampling were also recorded.

As quickly as possible after the sampling, 4.8 ± 1.3 (mean ± SD) liters (supplementary table B.1) were filtered (20 μm) using a peristaltic pump and the filters were frozen into liquid nitrogen and stored at -80°C until RNA extraction. For all samples, the different steps from sampling to freezing took around 2 hours. To prevent RNA degradation during filter thawing, RNA Later Ice® was added before thawing for samples collected in 2013, or samples were directly frozen with RNA Later® in 2014 and 2015. Filters were ultra-sonicated on ice (Bioblock Scientific, Vibra-cell

75115) for 10 seconds at 20% intensity. RNA was extracted following the Quiagen Rneasy Plus Mini Kit® recommendations and library prepared with the Illumina Truseq mRNA V2 kit®. Samples were sequenced at Get-PlaGe France Genomics sequencing platform (Toulouse, France) on Illumina HiSeq 2000/2500 2*100 pb (2013 and 2014; Multiplex 10) and on HiSeq 3000 2*150 pb for 2015 samples (Multiplex 24). Overall, 41 libraries were sequenced (10 in 2013, 19 in 2014 and 12 in 2015, see details in supplementary table B.1).

Data preparation and read mapping : Prior to read mapping, raw reads were initially characterised with FASTQC and TRIMMOMATIC (V. 0.33, Bolger et al. (2014)) was used to trim ambiguous, low quality reads and sequencing adapters with parameters ILLUMINACLIP:Adapters.fasta:2:30:10, LEADING:3, TRAILING:3, MAXINFO:80:0.8 and MINLEN:70. We used a reference transcriptome previously assembled using 18 *A. minutum* strains (Le Gac et al., 2016). This 153,222 transcripts reference, for a total of 117,601,765 bp, was annotated using UNIPROT/Swissprot and classified in Gene Ontology categories (GO) as described in Le Gac et al. (2016).

Preliminary precautions were taken before aligning the environmental reads to *A. minutum* raw transcriptome. First, in order to quantify specific gene expression of *A. minutum* and minimize contamination by reads from other species present in the community, only transcripts showing species-specific alignment were considered for the analyses. Specific transcripts of *A. minutum* were selected as follows: species often co-occurring with *A. minutum* next to our study site (48.3334355, -4.3264413) were identified using the microphytoplanktonic flora obtained by the Veliger network. Raw reads from 24 co-occuring species were obtained from the MMETSP database ((Keeling et al., 2014), see supplementary table B.2). These 24 raw read datasets were trimmed as described above and aligned to the *A. minutum* reference transcriptome using BOWTIE 2 (V. 2.2.4, Langmead and Salzberg (2012)) algorithm for paired reads using the sensitive set of parameters. Very conservatively, all the transcripts attracting at least one read from the co-occurring species were discarded from the *A. minutum* reference transcriptome and not considered for the following analyses (see box 5). As a second preliminary step, we aligned the environmental reads to this transcriptome using BOWTIE 2 as previously described. When several isoforms were possible for a single transcript, only the isoform attracting the greatest number of our environmental reads was retained for analysis. After aligning the trimmed reads of the 41 environmental samples to this transcriptome, reads with mapping quality lower than 10 and paired reads mapping to different transcript were removed. The number of reads mapping to each transcript was counted using the SAMTOOLS idxstats tool (V. 0.1.19, H. Li et al. (2009)).

Gene expression analysis

Normalization : When investigating species-specific gene expression dynamics *in situ* using a community-wide sequencing approach, deciphering true expression changes from relative abundance fluctuation of the focal species in the community is challenging. For the 41 environmental samples considered in the present study, between 7 and 33 million PE reads were obtained, corresponding to a maximum fourfold difference in sequencing depth between samples. Such a difference in sequencing depth may be corrected by the normalization procedures classically used in RNA-seq studies (rlog, VST, Love et al. (2014)). However, in these same samples, and probably due to relative abundance fluctuation in the community, the number of PE reads specifically mapping to the *A. minutum* reference transcriptome varied from 0.1 to 17 million, which corresponds to a maximum of 150 fold difference in sequencing depth. In such cases, gene expression levels might be highly influenced by the direct relative abundance of *A. minutum* in the community. To circumvent this problem, a two steps procedure was applied. First, only samples with at least 10 % of *A. minutum* transcripts covered by at least 40 reads (see supplementary table B.1) were analysed. This step reduced the number of samples to 29 and the difference in *A. minutum* sequencing depth to tenfold. Second, in the same effort to keep only the most biologically relevant information in the downstream analyses, transcripts displaying less than 40 reads in average across the remaining samples were discarded. Various sets of filtering parameters (every transcripts, 10, 20, 40 and 120 in average) were tested, resulting in unchanged behaviors from the dataset and the resulting biological information (data not shown). Raw counts of the remaining transcripts were then normalized using the variance stabilizing transformation from DESEQ2

R package (Love et al., 2014), in order to transform the count data into homoskedastic values and to normalize with respect to library size. As libraries were prepared in three separate batches (at the end of each sampling year), batch effects were corrected using COMBAT algorithm in the SVA R package (Leek et al., 2012). Following this normalization step it was not possible to study year specific changes in gene expression. Using this approach, we were able to analyse the gene expression level variations free from technical biases (both from the relative abundance of *A. minutum* in the sampled community, library size differences and year-specific batch effects).

Gene expression dynamic analysis : In contrast to classical RNA-seq studies, the purpose of the present work was not to compare expression profiles between contrasting experimental conditions, but rather to identify the transcripts displaying dynamic expression *in situ* without any preconceptions. To do so, transcripts with similar expression dynamics were grouped into modules of co-expressed transcripts using the Weighted Gene Co-Expression Network Analysis (WGCNA; V 1.51, Langfelder and Horvath (2008)) R package. A soft-thresholding power of 10 and a minimum module size of 30 transcripts were chosen as cut-offs. The gene expression modules were finally separated using a 0.2 height cut-off.

Gene Ontology enrichment analyses were performed for each module using a one sided Fisher exact test, with a false discovery rate set to 0.05 for multiple test type 1 error control. The hierarchical Gene Ontology classification system leads to redundancies in the GO enrichment analysis. To mitigate such redundancy, GO categories displaying an overlap coefficient superior to 0.75 were clustered, and for each cluster only the GO category displaying the lowest q-value was reported. Given the number of transcripts A and B in sets A and B, the overlap coefficient (OC) is calculated as :

$$OC = \frac{A \cap B}{Min(A, B)}$$

When GO categories were shared between gene expression modules, the shared GO category and not the one displaying the lowest q-value was considered as representative of the GO cluster. For each gene expression module, the functions of the

10% of transcripts showing the highest difference between their lowest and maximal expression among samples were inspected individually.

Environmental factors analysis

The correlation between the dynamics of gene expression and environmental fluctuations was explored using a Redundancy analysis (RDA; ADE4 R package, Dray and Dufour (2007)). For each expression module, WGCNA was used to extract eigengenes, which can be defined as the first principal component of the expression matrix of the detected modules. The linear relationships between the dynamic of these eigengenes and twelve centered reduced environmental factors were summarized using a RDA. The twelve environmental factors considered were : the progression through the bloom (the sampling date normalized by the total duration of the bloom), the outflow of the Mignonne river, the tidal coefficient, temperature, salinity, ammonium, nitrogen oxide, phosphate and silicate concentrations, the sea surface irradiance, the time from sunrise to sampling, and the cell density of *A. minutum* (cells.L^{-1}). The dynamics of these factors in our time series are illustrated in figure 3.2.

3.2.4 Results

To investigate *A. minutum* gene expression dynamics *in situ*, the mRNA corresponding to the microphytoplanktonic community during three consecutive *A. minutum* blooms was sequenced from 41 environmental samples. The sampling took place during spring and summer 2013, 2014, and 2015, representing respectively (10, 19 and 12 samples), in a single site located in western France. During these blooms, the absolute *A. minutum* abundances ranged from 1 500 (07/02/2015) to more than one million cells per liter (06/06/2014). According to cell density estimates, the three blooms surveyed lasted 31, 80 and 49 days respectively.

Transcriptome specificity

To exclude from the *A. minutum* reference the transcripts that may attract environmental sequences from other species, raw reads of 24 microphytoplanktonic

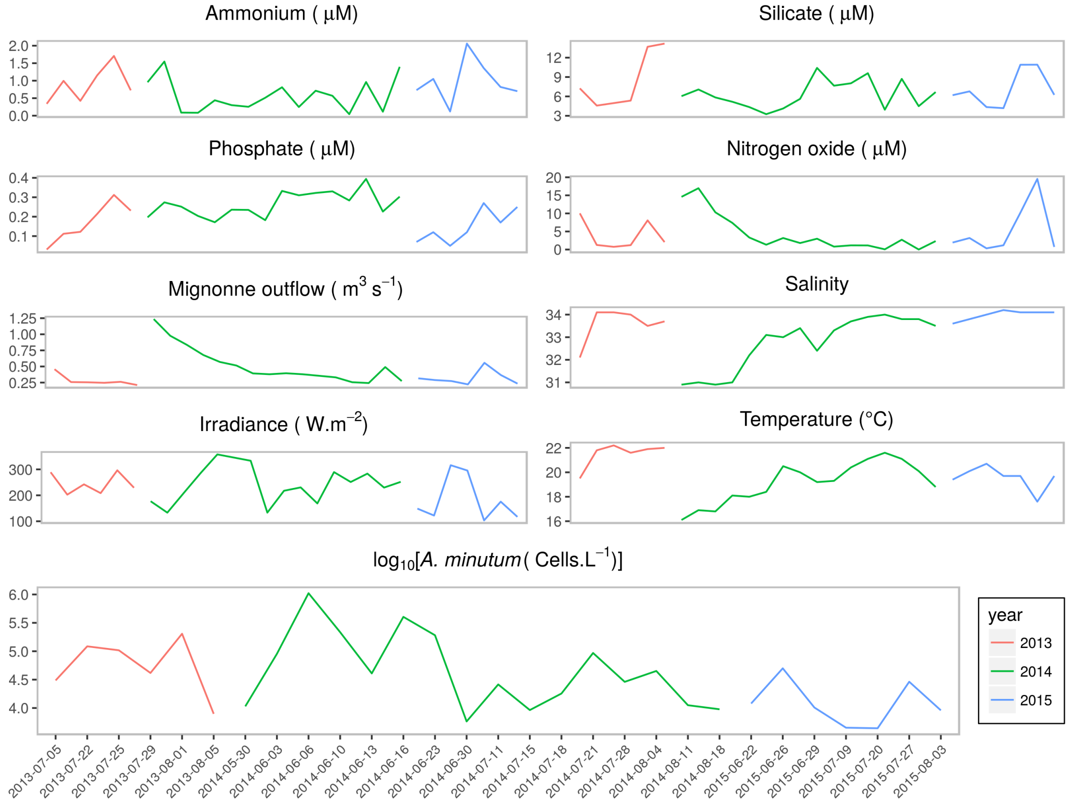


FIGURE 3.2: *Dynamics of nine abiotic factors during the sampling period. Sea surface salinity and temperature were recorded* in situ. *Ammonium, silicate, phosphate and nitrogen oxyde concentrations and A. minutum cell densities were calculated in the laboratory. Mignonne outflow and sea surface irradiance were collected respectively at http://hydro.eaufrance.fr/ and http://www.meteofrance.com/. Sampling dates are displayed in x axis and y axis represents each respective factor scales.*

species (13 dinoflagellates, 10 diatoms (from classes Bacillariophyceae and Coscinodiscophyceae, respectively 3 and 7) and 1 Labyrinthulomycetes) detected in the water at the time of sampling were aligned against *A. minutum* reference transcriptome (see supplementary Table B.2). About 1.1% of the 1 281,725,910 tested reads aligned to a total of 10,425 transcripts, representing 6.8% of the *A. minutum* transcripts (see details in supplementary table B.2 and supplementary figure B.1). Of these non-specific transcripts, 3,362 had an annotation (32.25%, See supplementary figure B.2), which is significantly enriched regarding the proportion of annotated transcripts in the reference (Fisher exact test, Odds Ratio (OR) = 4.4, p = 2.2e^{-16}). In addition, 35 Gene Ontology categories were over-represented within the non specific transcripts (see supplementary table B.3).

The remaining 142,797 transcripts did not attract any aspecific read and were used as the *A. minutum* reference in the present study.

Selecting samples and transcripts

A total of 176,340,078 paired-end reads from 41 environmental samples aligned to the *A. minutum* transcriptome. The number of reads aligned to the reference transcriptome varied from 10^5 to 1.6 10^7 representing from 1.5% to 57% of the reads for the various samples (supplementary table B.1). Although, this extremely wide range probably reflected the variation of *A. minutum* in terms of relative abundance within the community, we note that it followed, at least partially, its absolute abundance fluctuations in the sampled water (see supplementary figure B.3). An immediate consequence of this difference in *A. minutum* sequencing depth is that the number of transcripts whose expression may be quantified varies greatly among the samples. For example, there were 857 fold differences, in terms of number of transcripts covered by at least 40 reads, between the samples displaying the lowest and highest *A. minutum* sequencing depth (see supplementary figure B.4). To circumvent this problem a double filter was applied. First, samples with a shallow *A. minutum* sequencing depth that prevented transcriptome wide quantification of gene expression were rejected. After this first step, 29 samples (6 from 2013, 16 from 2014 and 7 from 2015) were taken into account for the analyses, reducing the difference in terms of number of transcripts covered by at least 40 reads to less than six fold (see supplementary table B.1). Second, the transcripts with low level of expression across samples were rejected, reducing the number of transcripts considered to 36,626 (23% of the transcriptome), representing 167,422,299 paired-end reads (95% of the aligned reads).

Identifying transcripts displaying dynamic expression *in situ*

Classically, gene expression responses to stimuli are studied using differential gene expression algorithm such as DESEQ2 (Love et al., 2014), LIMMA (Ritchie et al., 2015)or EDGER (Robinson et al., 2011). These approaches compare the levels of gene expression between groups of samples and require the assignment of samples into pre-defined groups. They are ideally suited to study the differential gene expression following a predefined experimental design, but are not adequate for exploring and detecting patterns of differential gene expression without without any preconceptions. An alternative is to use clustering techniques to identify sets of transcripts displaying similar expression dynamics across samples (modules, hereafter). Using such approach, eight modules regrouping between 229 and 3,007 transcripts for a total of 9,527 transcripts (26 % of the 36,626 analysed transcripts) were identified (figure 3.3). The observed changes in gene expression were moderate, with eigengenes maximum absolute log2 fold changes between minimal and maximal expression across the time series for modules A to H of respectively 1.17, 1.12, 1.17, 1.19, 1.27, 1.17, 1.38 and 1.13 but nevertheless indicate global shifts in gene expression patterns.

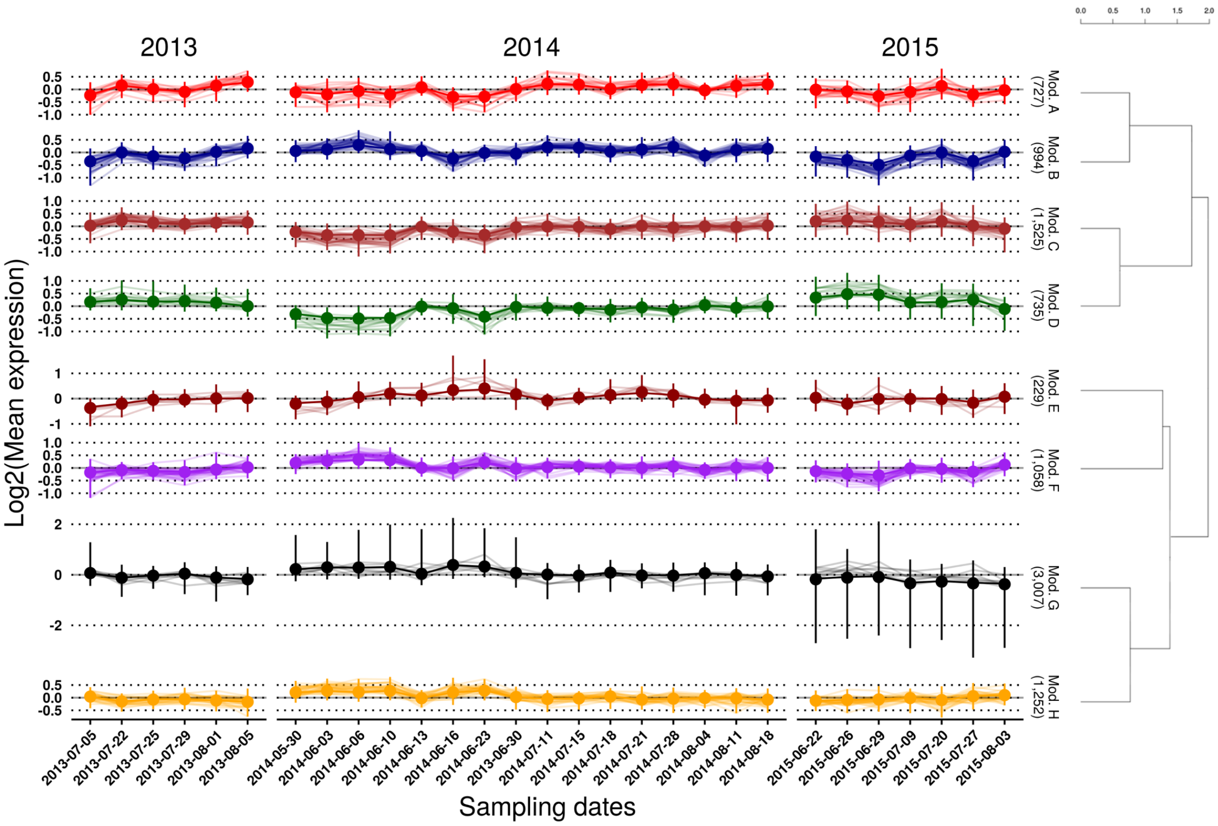


FIGURE 3.3: *Representation of the expression dynamics (relative to the Log2(mean expression) over all samples) of the transcripts in each expression modules, presented with a dendrogram representing the euclidean distance of their respective eigengenes. Modules were detected using* WGCNA *on 36,626 transcripts with a sufficient expression, previously variance-stabilized (*DESEQ2*) and batch effect free (*COMBAT*). Expression modules were separated using a 0.2 height cut-off. Number in brackets under each expression module is the number of transcripts falling into them.*

Functions of the transcripts displaying dynamic expression

In order to understand the functional implications of the observed gene expression dynamics *in situ*, transcripts were grouped into Gene Ontology (GO) categories based on their homology to genes with known functions. Nineteen GO terms were over represented among the transcripts considered as expressed, with numerous functions associated with ionic functions, membrane depolarization and the production of action potentials (9/19). There were also many underrepresented functions (82) in these transcripts (see supplementary table B.4). Out of the 36 626 transcripts used in the analyses (7,592 were annotated; 20.73%), 9 527 transcripts (1,639 annotated; 17.2%) were grouped into modules. This first observation showed a depletion in annotated transcripts for the transcripts clustered into expression modules (See supplementary figure B.2, Fisher exact test, Odds Ratio (OR) = 0.79, p = 8.931e^{-15}). The approach developed in the present paper aims at identifying cellular functions displaying dynamic gene expression *in situ* without any preconceptions. Nevertheless, several functions may *a priori* be of special interest (for instance central functions in an algae life cycle such as nitrogen and phosphorous metabolism, TCA cycle, photosynthesis), so we checked at what point these specific functions are represented in the expressed transcripts (see supplementary figure B.5). None of the tested functions were significantly over-represented (although some were significantly under-represented in the expressed transcripts such as chloroplast and stress related functions). Even though these GO categories were present in the analysed transcriptome, many of them are rather small (implying a low statistical power to detect over or under-representation). Considering each of the eight transcript modules separately, the proportion of annotated transcripts was extremely variable, going from a very strong under representation of annotated transcripts in module G (Fisher exact test, Odds Ratio (OR) = 0.09, p <2.2e^{-16}) to a strong over representation in module F (Fisher exact test, Odds Ratio (OR) = 1.9, p <2.2e^{-16}, see supplementary figure B.2).

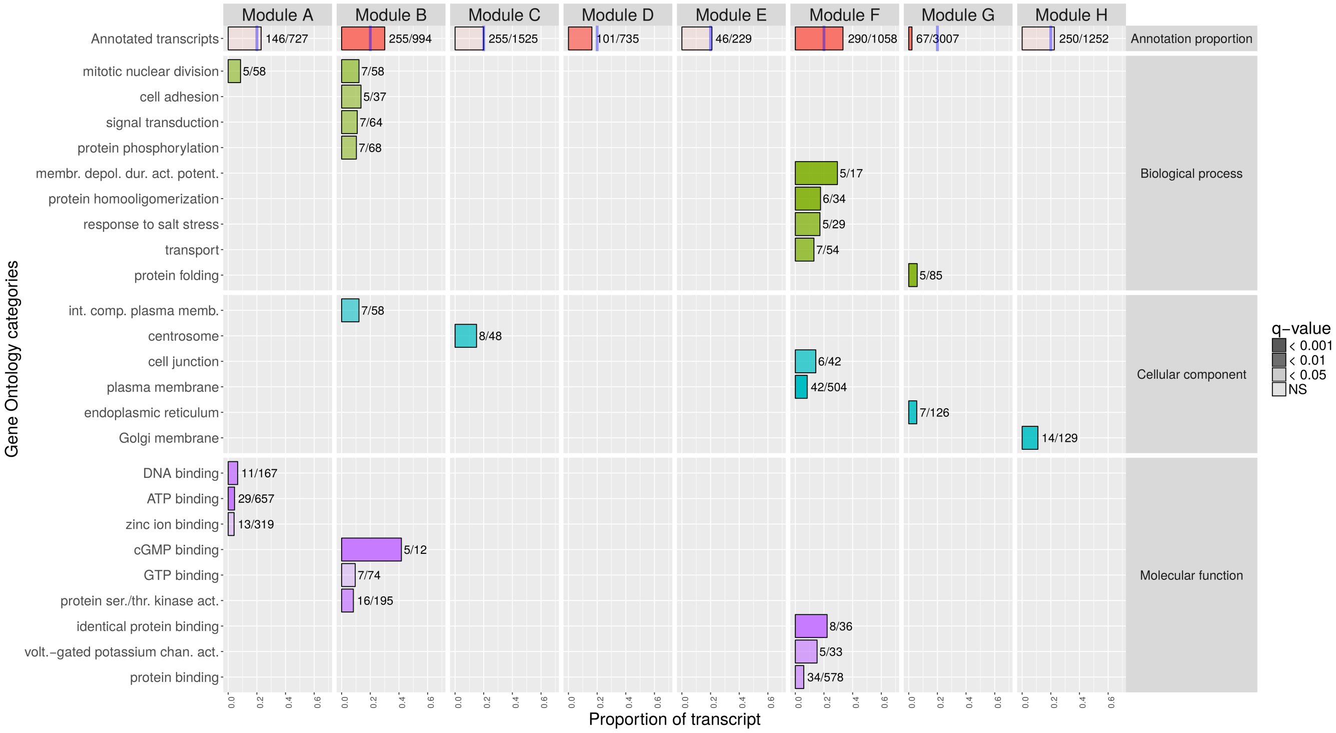


FIGURE 3.4: *Over-represented Gene Ontology functional categories in the transcripts gathered in each expression modules (vertical panels). For each GO category (y axis), the proportion of transcripts annotated in each category in the form "Number of transcript in module / Total number of transcript in the category" is represented in the x axis. The color of each GO indicates the broader biological process / cellular component / molecular function to which they belong (horizontal panels). Color transparency represents the q-value of Fisher's enrichment test. Enrichment/Depletion of annotated transcripts per module compared to the proportion of annotated transcripts in the overall tested transcripts (blue shaded line) is shown in panel "Annotation proportion".*

A total of 24 GO categories were identified as displaying an excess of transcripts in at least one of the eight modules (figure 3.4). Between 0 and 9 GO categories were enriched per module (Module A : 4, Module B : 8, Module C : 1, Module D : 0, Module E : 0, Module F : 9, Module G : 2, Module H : 1). Out of the 24 enriched GO categories, only one (mitotic nuclear division) is enriched in 2 modules. Module A is characterised by a slight excess of transcripts involved in mitosis, but also DNA and ATP binding. Module B is also characterised by an excess of transcripts involved in mitosis, but also in cGMP mediated signal transduction and in cell adhesion (cell-cell or cell-matrix). Module C presents an excess of transcripts associated with the cell cycle and more precisely of transcripts located in the centrosome. Modules D and E do not display any enriched GO category. Module F regroups numerous transporters and especially with an over-representation of voltage gated channels associated with action potentials, but also of transcripts involved in cell to cell junctions (link between cytoskeletal proteins), protein binding and responses to stress. Transcripts associated with protein maturation in the endoplasmic reticulum are over-represented in module G. Finally, in module H, transcripts linked to the Golgi apparatus are over-represented. As a complement to the previous approach based on the over-representation of functional categories, for each transcript module, we also considered the transcripts displaying the most variable expression (top 10%). Their expression dynamics are displayed in figure 3.5. Of special interest in module B, are five transcripts (out of 25 annotated), which homologs are related to the synthesis of antimicrobial products (Gramicidin synthase, Mycocerosic acid synthase and 3 polyketide synthase (PpsD, PksL and PksM). In module C, six transcripts are linked to the cytoskeleton and three more to the cilia (out of 30 transcripts). In module F, there are 17 transporters (out of 40 transcripts), among which four voltage dependent/gated transporters and one cGMP related transcript. Out of these 40 most variable transcripts of module F, there is also the presence of two light driven exchangers (Sodium/potassium/calcium exchanger 2 and Cruxrhodopsin-1, a light-driven proton pump) (see supplementary table B.6).

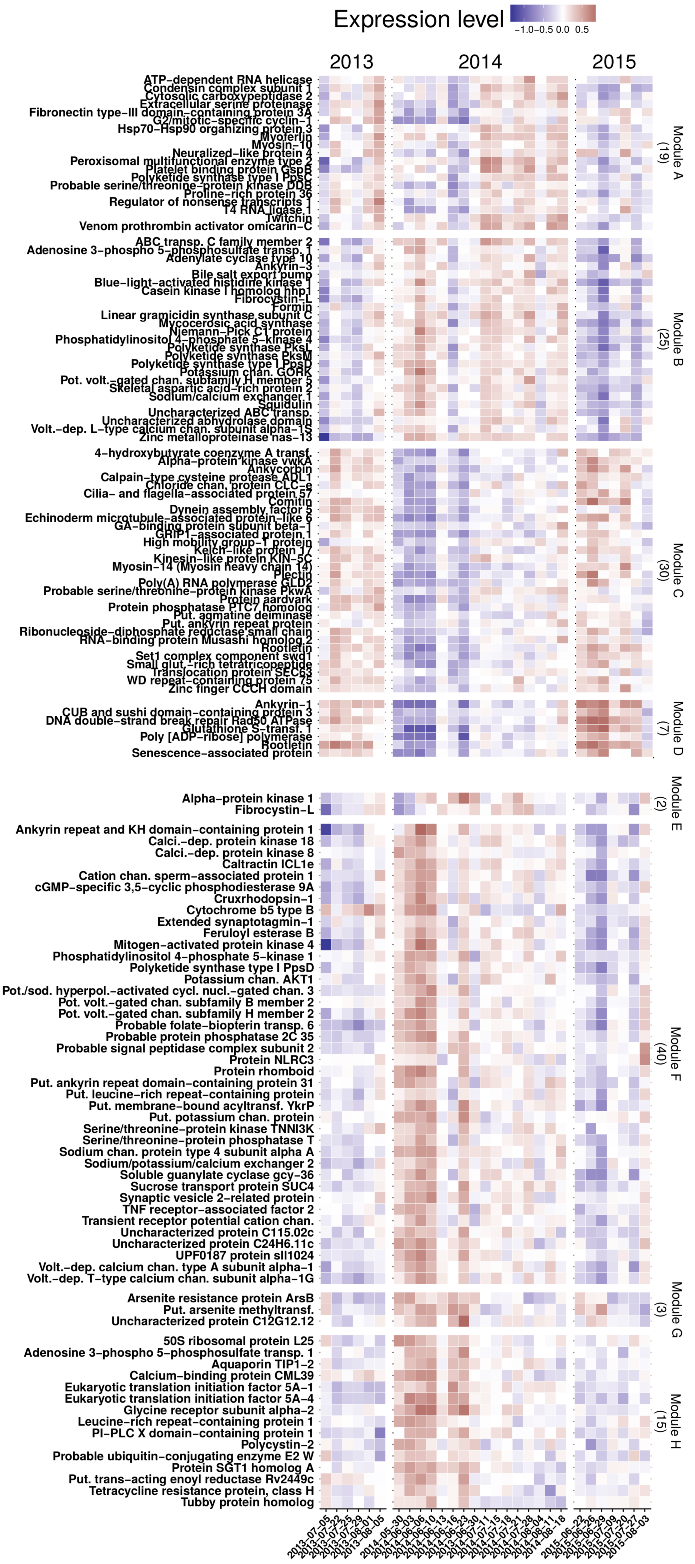


FIGURE 3.5: *Expression dynamics (relative to the Log2(mean expression) over all samples) of the annotated transcripts among the top 10% most dynamic transcripts (maximum expression fold change across all samples) in each expression module (each panel), across the time series (x axis).*

Linking gene expression profiles to abiotic environmental factors

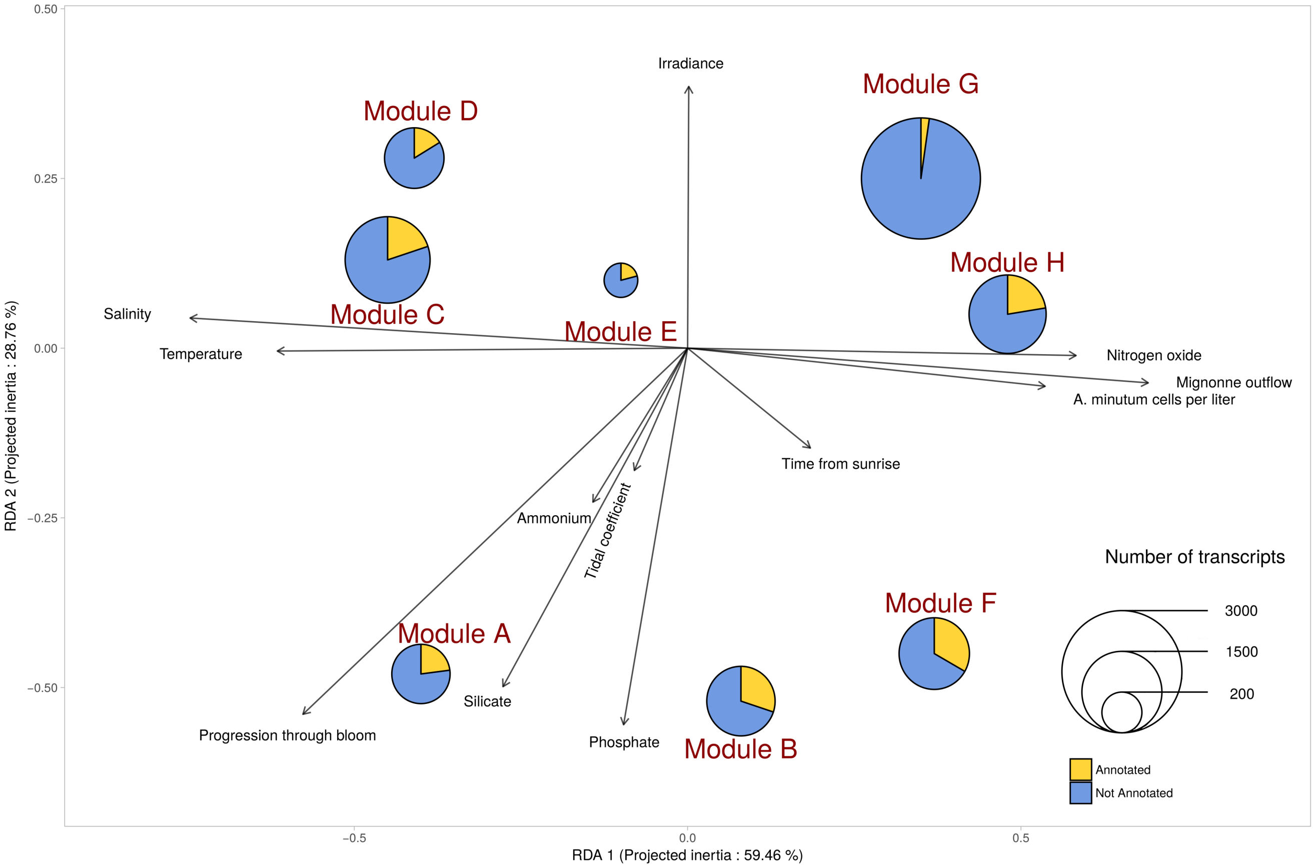


FIGURE 3.6: *Representation of the two first axis of the RDA led between the gene expression modules and environmental factors. The expression modules (represented in red) are plotted with the 12 explicative environmental variables (in black). Arrows represent their contribution to the two first axis. The two first axis explain 88 % of the variance (respectively 59.46 and 28.76%).*

Twelve environmental variables were quantified for each sample. In order to understand how the dynamics of gene expression may be related to the variation of the abiotic environment, the expression profile of each module was summarized as a single eigengene (see method), and the linear relationships between the expression of these eigengenes and the environmental variables were summarized using a RDA analysis (figure 3.6). The first RDA axis explains 59% of the variation. It separates the modules A, C and D displaying high expressions in nitrogen poor, warm and high salinity environmental conditions associated with sparse *A. minutum* populations from the modules F, G, and H displaying high expressions at the beginning of *A. minutum* blooms, in cold, low salinity and nitrogen rich conditions. The second axis explains 29% of the variations and mainly separates modules A, B and F, expressed mostly at the end of the bloom season in low irradiance, phosphate and silicate rich environments, from modules D and G, that are highly expressed in high irradiance and nutrient depleted environments (except for Nitrogen oxide which is not explained by the second axis). The third axis explains 7.9% of the variations and separates module E from the other modules (see supplementary figure B.6). Transcripts belonging to module E are expressed in nitrogen and ammonium rich environments, in low irradiance, salt depleted, cold and low *A. minutum* densities waters.

3.2.5 Discussion

This study used meta-transcriptomic datasets to uncover the molecular physiology dynamics of a single microbial species *in situ*. Meta-transcriptomic approaches have been applied to investigate several community level aspects such as community wide expression dynamics, microbiome, soil and marine bacterial communities assemblages, or active diversity of protists communities (Gosalbes et al., 2011; Lejzerowicz, Voltsky, & Pawlowski, 2013; Radax et al., 2012; R. Sharma & Sharma, 2018). Such approach has nevertheless rarely been applied to investigate species specific expression dynamics *in situ* (but see Alexander, Jenkins, et al. (2015) for a study at the class level). However, for communities dominated by a few major species, such as algal blooms, it offers the promise to scale down the *in situ* monitoring of molecular physiology at the species level. Here, the species of interest was *A. minutum*, a dinoflagellate responsible for toxic algal blooms in coastal waters worldwide.

We first want to highlight the fact that we faced a classic drawback of studying non model organisms : the reference transcriptome we used here is annotated using a sequence homology method against a database composed of more or less closely related species. We expect that the transcriptome lack functions specific to *A. minutum*, and is enriched in core functions, more likely to have sequence homologies in the databases used for the annotation. The relatively low proportion of annotated transcripts clearly reflects this effect. Moreover, the transcriptome used here comes from 18 strains of *A. minutum* collected in the field but cultivated in the lab (Le Gac et al., 2016). Then, our reference transcriptome is likely to reflect the *in vitro* physiological states of *A. minutum* and some *in vivo* specific transcripts might be missing in this assembly.

Isolation of *A. minutum* in our data

As we worked with environmental samples, the sequenced reads were composed of mRNA transcribed by various members of the sampled community. The first step was to refine the reference transcriptome of *A. minutum* to ensure the specificity of the alignment of the short environmental sequences. The nonspecific transcripts found in the transcriptome display a strong enrichment in annotated sequences, reflecting that highly conserved sequences are easier to annotate. Moreover, core functions such as photosynthetic functions are over-represented in the non specific transcripts, reflecting that key functions are more likely to be highly conserved and then more likely to attract reads from other species. As a drawback, such key functions are found under-represented in the expressed set of transcripts in regard to the overall transcriptome (see supplementary figure B.5). Of course, more species could be tested and the number of nonspecific transcripts might then be refined, but the rather low amount of nonspecific transcripts and the very conservative approach we used insure that the remaining transcriptome is *A. minutum* specific. As sequence homologies increase with phylogenetic relatedness, this approach could be problematic if applied to environments with several abundant and closely related species co-occurring with the species of interest -which is not the case here-, as it would strongly affect alignment specificity.

Summarizing complex gene expression dynamics

Specific samples and transcripts were selected to mitigate the massive *A. minutum* sequencing depth variations among samples due to relative abundance fluctuations in the natural community. About a quarter of the transcripts considered were then grouped into modules displaying similar expression patterns across samples. It seems important to note that these overall moderate but significant expression dynamics reflect population wide dynamics. *In vitro* experiments are classically conducted using one or a few strains (thus completely removing the potential effect of genetic background diversity on physiological responses) cultivated under drastically different conditions, which might overstate natural gene expression responses.

The *in situ* expression signals variations observed here are of lower magnitude than those observed with *in vitro* surveys. This is not surprising considering that it corresponds to the population wide average response of thousands of genetically diverse *A. minutum* cells subjected to natural environmental variations. Another hypothesis to explain these moderate variations is the uncommon Trans-splicing protein level regulation mechanism in dinoflagellates. In this group, it is controlled by a short (< 50 nt) non coding RNA sequence spliced to the 5′ end of an independently transcribed pre-mRNA. This is supposed to occur in many if not most of dinoflagellate protein coding genes and be one of the most important way of regulating protein levels (Lidie & Van Dolah, 2007; Moustafa et al., 2010; H. Zhang et al., 2007). In this scheme, genes are constitutively transcribed and protein levels are mainly post-transcriptionally regulated. However, a few other studies showed that dinoflagellates exhibit transcriptional regulation in response to environmental changes (Erdner & Anderson, 2006; Leggat et al., 2011; Moustafa et al., 2010), although they are less pronounced than in other groups of protists (see for instance Alexander, Rouco, et al. (2015)).

Environmentally triggered functions

The different physiological states experienced by micro-eucaryotes *in situ* remain difficult to study. Here, we recorded transcript abundance fluctuations over three consecutive *A. minutum* blooms and classified them into 8 expression modules. Following this, we assigned biological functions to 6 expression modules using a Gene Ontology term enrichment analysis. Several key functions were expected to be found dynamically expressed. This is for instance the case of transcripts related to nutrient metabolism, and especially nitrogen metabolism or phosphate uptake, two major nutrients often considered as involved in resource competition between phytoplankton species and as a result as important drivers of phytoplankton community dynamics. More generally, nitrogen metabolism is an important process involved in the synthesis of nucleic acids, chlorophylls and toxins, and which the presence under different forms in the water is often correlated with the development of HAB (Anderson et al., 2008; Lee, 2006; San Diego-McGlone, Azanza, Villanoy, & Jacinto, 2008; Zhou, Shen, & Yu, 2008), but we may also consider functions related to tricarboxylic acid cycle and photosynthesis for instance (see supplementary figure B.5). Interestingly, none were over represented in the expressed transcripts, even rather several of them were found to be under represented on the contrary. Although some GO categories belonging to these functions were small (reducing the statistical power to detect them over represented), the presence of underrepresented key functions is noteworthy. It would tend to illustrate that cellular functions identified a priori are not the one displaying extensive expression dynamics *in situ*. Interestingly, only one out of the three saxitoxin production related transcript was expressed enough to be considered in the analysis, and showed a constant expression through the blooms (data not shown). Although we did not find such *a priori* interesting functions, the main object of the study was to identify without *a priori* biologically processes that are relevant *in situ*. Of particular interest are two sets of transcripts, both over enriched in transcripts involved in signal transduction functions.

Motility in *A. minutum* : Ionic fluxes in micro-eukaryotes are well documented and were shown to be involved in motile and sensory responses in *Chlamydomonas* for instance (Goodenough et al., 1993; Harz & Hegemann, 1991; Holland, Braun, Nonnengässer, Harz, & Hegemann, 1996; Quarmby, 1994). In dinoflagellates of the genus *Noctiluca*, tentacle movements were shown to be initiated and regulated by calcium (and also Na^+) transporters (Nawar & Sibaoka, 1987; Oami, Naitoh, & Sibaoka, 1995). More generally, ionic transporters are known to be involved in numerous modes of cell motility (see the review by Stock, Ludwig, Hanley, and Schwab (2013)). Here, we show that a set of transcript (Module F) gathers homologs of sperm motility proteins, membrane depolarization due to action potential and numerous voltage gated transporters (both in terms of over-represented functions and most highly dynamic transcripts) that are more expressed in cold environments, which could reflect active cell movements in such unfavorable environments. Motility behaviors in dinoflagellates are also affected by irradiance in a two-way response. These autotrophic organisms display positive phototropism in low lights but are repelled by high lights and long exposure times (Cullen & MacIntyre, 1998; Ekelund, 1991; Tirlapur, Scheuerlein, & Hader, 1993). Here the transcripts implicated in ionic transporter activities are less expressed in high irradiance environments, which is congruent with the previous observations of lower motility under higher light exposure.

Interspecific interactions : Inter-specific interactions are supposed to play a key role in controlling bloom dynamics. During bloom expansion, recruited bacteria, grazers and parasites tend to moderate, stop or reduce bloom growth and participate to the transition between bloom development and termination phases (Calbet et al., 2003; Chambouvet et al., 2008; Labry et al., 2008; Legendre & Rassoulzadegan, 1995; Sorokin et al., 1996). Numerous interaction-related compounds were showed to have similar expression dynamics over the sampled blooms. For instance, among the most dynamic transcripts in one set of transcripts (module B) are products related to anti-bacterial and polyketide synthase homologs, a large group of secondary metabolites that appears to be important mediators of allelopathic and anti-predators responses in eukaryotic marine micro-algae (Cembella, 2003; Kohli, John, Van Dolah, & Murray, 2016). Interactions with bacterias are supposed to increase as the bloom approaches its end, as they were shown to stimulates cyst formation in dinoflagellates, and can then be one of the factor regulating termination of the bloom (Adachi et al., 1999; Mayali, Franks, & Azam, 2007). Here, genes involved in such functions are more expressed toward the end of the blooming period. Their particularly highly dynamic expression characteristics may also reflect a rapidly adaptive physiological response from *A. minutum* to other species. While calcium signaling might not be extensively studied in dinoflagellates, a few reports in *Crypthecodinium cohnii* show that Ca^{2+} fluxes modulation play a role in cell cycle regulation in response to mechanical stimulation (Lam, New, & Wong, 2001; Yeung, Lam, Ma, Wong, & Wong, 2006). Calcium concentration regulatory channels are known for transducing organized and controlled spatial and temporal sensory signals. In diatoms, calcium concentration elevations were shown to respond to mechanical stimuli (Falciatore, D'Alcalà, Croot, & Bowler, 2000). In two divergent *A. minutum* groups of strains, genes involved in various calcium related functions were shown to be among the most divergent genetically, highlighting their particularly central role in *A. minutum* life cycle and biology (Le Gac et al., 2016). Here, we highlight numerous co-expressed sets of transcripts involved in ionic (calcium, sodium, potassium) channels activity and ionic (trans-membrane) transports, and notably one (module B), that may suggest important cellular exchanges or inter-cellular communications and support their important role in *A. minutum* life cycle. Interestingly, the ionic-related enriched functions are not necessarily correlated with *A. minutum* cell concentrations, suggesting that genes encoding the proteins involved in these functions are transcribed regardless of the cell density of the bloom. Moreover, it's noteworthy that transcripts related to cell adhesion processes (either to a substrate or to another cell) are over-represented and more expressed toward the end of the bloom. All these results point toward increasing inter-specific interactions late in bloom, and in nutrient rich with low irradiance environments.

3.2.6 Conclusion

Using an *in situ* transcriptomic approach, we overpassed the challenges of studying microscopic, non-sessile species living in a highly dynamic environment to monitor *A. minutum* genes expressions dynamics in its natural habitat over long periods of time. In this study we first extracted species-specific transcriptomic information from environmental-level mRNAs, second we analysed their transcript abundance dynamics, third we performed functional over-representation in these expression patterns and finally we replaced them in a global environmental framework. Our results point out that, over a three year consecutive bloom physiology monitoring, *A. minutum* shows major groups of co-expressed genes coding for transcripts involved in specific functions such as ionic-related mechanisms including motility and cell-cell interactions, cell division or cellular maintenance. We also showed that the expression dynamic is strongly correlated to environmental fluctuations, and more specifically that, as the bloom approaches its end, inter-specific interactions and responses to stimulus related functions are more expressed, suggesting active cellular activities prior to encystment. This study shed a new light on the possible use of meta-transcriptomic data. In particular, this approach easily stimulates the emergence of novel hypothesis regarding physiological responses of importance *in situ*. As expression data may be studied both *in situ* and *in vitro*, parallel studies investigating processes of interest in the field and in the lab are especially promising to better understand the physiology and ecology of unicellular eukaryotes.

3.2.7 Acknowldgments

The authors want to thank the Region Bretagne for funding the Daoulex program, GetPlage sequencing platform for technical advices and help and the Rephy network for giving us access to their data. We also thank Julien Coïc, Hélène Hégaret as well as members of the Pelagos and LER/BO labs for helps provided during the sampling and members of the UMI 3614 for discussions. We also thank ABIMS and RIC bioinformatic platforms for the technical support.

Chapter 4

Spatio-temporal population genetic structure

4.1 General context

As presented in the introduction, one of the possible response toward environmental changes is genetic adaptation. When such changes impose differential selective pressures to groups of individuals within the same species, local adaptation may appear, leading to population genetic differentiation (along spatial or temporal gradients for instance). Such genetic differentiation (through accumulation of beneficial -*i.e.* local adaptation- or neutral -drift- mutations) may lead to reproductive barrier appearance and finally to speciation which can, in this context, appear as an ultimate evolutionary response to environmental change. At the population scale, investigating genetic structure (especially local changes and inter-population gene flows for instance) is of primary interest to understand population response (in terms of genetic adaptation) to environmental fluctuations.

Populations dynamics are notably influenced by recruitment (the formation of new organisms by sexual or asexual reproduction) and migration between subpopulations. In some plant species, the recruitment system of new individuals from seed banks is of great interest both in maintaining species diversity (see for instance Honnay, Bossuyt, Jacquemyn, Shimono, and Uchiyama (2008) and Stocklin and Fischer (1999) who showed that grassland remnants with longer-lived seeds have lower extinction rates) and also habitat restoration (Augusto, Dupouey, Picard, & Ranger, 2001; Rydgren, Hestmark, & Okland, 1998). In cyst-forming species - such as dinoflagellates -, resting cysts buried in the sediment can be considered as a ‘cyst bank’, as plant seeds in seed banks (Wyatt & Jenkinson, 1997), and therefore constitute a reservoir of genetic diversity in these species as cyst can germinate after up to decades of encystment (Klouch et al., 2016). Immigration is also an important factor influencing population genetic dynamics (see for instance Dilmaghani et al. (2012) in the oilseed rape pathogen *Leptosphaeria maculans*; Wollebæk, Heggenes, and Røed (2011) in the arctic char; He, Lamont, Krauss, and Enright (2010) in the *Banksia hookeriana* shrub). The spatial distribution of an organism is the result of the interaction between its behaviour and its environment. In the oceans, the fluid nature of the marine environment contributes greatly to the dispersal of individuals within and between populations. Even sessile species have generally at least one highly dispersive life stage (*i.e.* larvae, spores, gametes). For planktonic species, dispersal capabilities mainly rely on three factors : the physical characteristics of the water mass, cell's life duration and floatability (McManus & Woodson, 2012). In the will to investigate population connectivity - *i.e.* to estimate the gene flow between subpopulations (through the migrants participating to the reproduction at the next generation; K. R. Cowen and Sponaugle (2009); R. K. Cowen, Gawarkiewicz, Pineda, Thorrold, and Werner (2007); Pineda, Hare, and Sponaugle (2007) - in natural environments, numerous estimations of marine species dispersal were performed (see for instance Levin, Huggett, Myers, Bridges, and Weaver (1993) with tagging techniques or Chevaldonné, Jollivet, Vangriesheim, and Desbruye (1997) in hydrothermal vents polychaete larval dispersal using models). Dispersal capacities of marine species can also be investigated using population genetic tools (see review in Hedgecock, Barber, and Edmands (2007); Taylor and Bruenn (2009) in reef fish larvae, or Chust et al. (2016) in various marine groups). Due to the apparent ever-changing three dimensional environment in which planktonic species evolve, it has long been proposed that marine microbial species populations must have little genetic structure, even at large spatial scales, as individual cells are supposed to disperse at high rates. However recent studies on dinoflagellate populations clearly indicate that blooms are structured in space and time (Dia et al., 2014; Nagai et al., 2007).

Then, population genetic dynamics will be greatly influenced by several factors, including recruitment (through selective reproduction or seed bank preferential germination), individual dispersal, but also local adaptation to their environment. Indeed, both biotic and abiotic factors can act as selective agents on adapted genotypes, and therefore change spatio-temporal allelic frequencies within and between populations. As presented in figure 1.1 and discussed in the introduction of the present manuscript, the rate of population evolutionary response to environmental selection differs greatly depending on the taxa considered and, as allelic frequencies might change at a decadal scale in long living species, genetic evolution might be observable at a day to week scale in some microbial species (mainly due to their large effective population size and short generation time (see figure 1.1 and chapter 1.1)). In this context, we can consider genetic adaptation in microbial populations as a complementary rapid response, together with phenotypic plasticity, toward environmental change.

The aim of the next chapter is to investigate **how does intraspecific genetic diversity change over space and time, and how do populations adapt to environmental fluctuations ?** As we investigated the interplay between genetic divergence and molecular physiology in the first chapter, and showed environmentally correlated gene expression dynamics in chapter 3, we hypothesize that *A. minutum* blooms also experience genetic structure. To test this, we use *in situ* metatranscriptomic data to describe the spatio-temporal genetic structure between and within bloom events occuring in the bay of Brest.

Author contributions : Environmental samples were collected and filtered by numerous laboratory members. After filtration, RNA was extracted and libraries were prepared by Mickael Le Gac and Julien Quéré (2013 and 2014 samples). I prepared 2015 samples for sequencing. Martin Plus performed cell dispersal simulations. I performed data mining and analysis, Mickael Le Gac and Christophe Destombe participated to result analysis and general discussions.

4.2 *A. minutum*'s populations spatio-temporal genetic structure

4.2.1 Introduction

Populations can be separated both by space and/or time. Along these two levels, restricted exchange of individuals may promote an increase in genetic divergence (by accumulation of neutral or beneficial mutations) between the separated populations. Some populations may live in sympatry, but reproduce at different time periods, resulting in an isolation-by-time divergence pattern. This can be the result of differential recruitment waves within and between years, a process that was shown in several species (see for instance Hendry, Berg, and Quinn (1999); Maes, Pujolar, Hellemans, and Volckaert (2006); Rolshausen, Hobson, and Schaefer (2010); Ribolli et al. (2017)). If reproduction time depends on geographical location, isolation-by-time might confound with another divergence process that is "isolation-by-distance". According to Wright (1943), dispersal probability and mixing of migrants decline as distance from the source increases, inducing an increasing genetic dissimilarity between populations with distance (Wright, 1943). Isolation by distance is a common-garden explanation for the population differentiation observed in population genetic studies (see for instance Wirth and Bernatchez (2001) in the European eel; Mims, Hauser, Goldberg, and Olden (2016) in the Arizona Treefrog; Aguillon et al. (2017) the Florida Scrub-Jay; Pusadee, Jamjod, Chiang, Rerkasem, and Schaal (2009) in rice). Of course, the magnitude of such isolation by distance effect on population differentiation will be influenced by several factors, such as dispersion capacities (with the lesser dispersive ones tending to have higher inter-population genetic structures), or even life cycles and life history traits of the species of interest. Several studies even showed isolation-by-distance effects discrepancies between sexes of the same species, with the less dispersive sexe displaying higher genetic structure patterns (see for instance Temple, Hoffman, and Amos (2006) in the white-breasted thrasher; Roy et al. (2014) in gorilla; M. H. Li and Merilä (2010) in the Siberian jay). In conservation studies strategy optimizations, assessments of population connectivity were shown to be important parameters for species management, and were extensively investigated (see for instance Cushman, Landguth, and Flather (2013) in three American Great Plains species; Marie et al. (2016) in the wedge clam *Donax trunculus*; Treml and Halpin (2012) in species from the Coral Triangle region).

In the marine environment, currents and fronts influence species dispersion (for at least one life stage for fixed species, or throughout the entire lifespan for non sessile species such as plankton), and genetic connectivity can therefore be greatly altered depending on the local oceanographic features surrounding them (see for instance Gilg and Hilbish (2003) and Mitarai, Siegel, Watson, Dong, and McWilliams (2009)). In the bay of Brest, the toxic dinoflagellate *A. minutum* gained attention in the past few decades, following the increases in bloom frequencies and magnitudes (reaching up to 42 million cells per liter in 2012) that led to the classification of this area as a high-risk zone for *A. minutum* toxic events (Chapelle et al., 2015a). This bloom forming species encyst in the sediment at the end of the blooming period, leading the sediment to be considered as a 'seed bank' that will serve as an inoculum for the next blooming season. Few studies investigated large spatial scale population genetic structure and connectivity in this species. One of them (Casabianca et al., 2012) showed the Mediterranean sea to be composed of 4 main clusters of genetically homogeneous groups, exchanging only a small fraction of their respective individuals. Dia et al. (2014) studied the spatio temporal genetic diversity and population structure within and between *A. minutum* blooms occurring in estuaries separated by 150 km for two consecutive years along the Northern Brittany coasts. They showed significant genetic differentiation at both spatial and temporal levels, suggesting high genetic differentiation between two consecutive bloom events, and limited exchanges between two bloom events occurring in two separated estuaries. Also, they showed that each bloom experienced genetic differentiation through time, with increasing genetic differences between cells sampled at different moments of the bloom event (between the beginning and the end of the bloom). The studies mentioned above, however, suffer from two main drawbacks : (i) they used a small number of genetic markers (12 microsatellite markers) and (ii), even if numerous monoclonal strains (360) were used, it is likely that they were only able to analyse a portion of the genetic variability encountered in natural populations composed of billions of cells. Then, the questions of population connectivity and genetic structure at a fine scale (few kilometers and high throughput temporal dynamics) using a population wide approach (and not monoclonal cultivated strains) remains unexplored. Here, we take advantage of large numbers of transcriptome wide SNP markers to investigate both long term temporal (but high frequency : 3 years, 41 time points) and small scale spatial survey (9 sampling points in an area of less than 40 km^2) to investigate genetic differentiation among and between *A. minutum* populations blooming in the bay of Brest. At the spatial scale, we hypothesize that the semi-enclosed, highly dynamic hydrological characteristics of the bay promote intense cell mixing, and therefore high connectivity among spatially distant blooms. If this hypothesis holds, over years, blooms are supposed to be genetically homogeneous, as such strong hydrodynamic will easily permit every cells to rapidly colonize every area of the bay. At the temporal scale, we can suppose that this short lived species, living in large populations will rapidly adapt to a changing environment, and therefore we expect an increasing genetic differentiation through time. To test these hypotheses, we will answer the following questions : what is *A. minutum* spatio-temporal genetic structure of the bay of Brest ? Is it composed of a single homogenized population, or are spatially closely situated (few kilometers) *A. minutum* blooms disconnected ? How does hydrodynamic and local environmental conditions impact small scale population structure ? As we are looking at population differentiation patterns using metatranscriptomic data, we also discuss the potentials and drawbacks of using mRNA to conduct population genetic analysis.

4.2.2 Material and Method

Sampling area

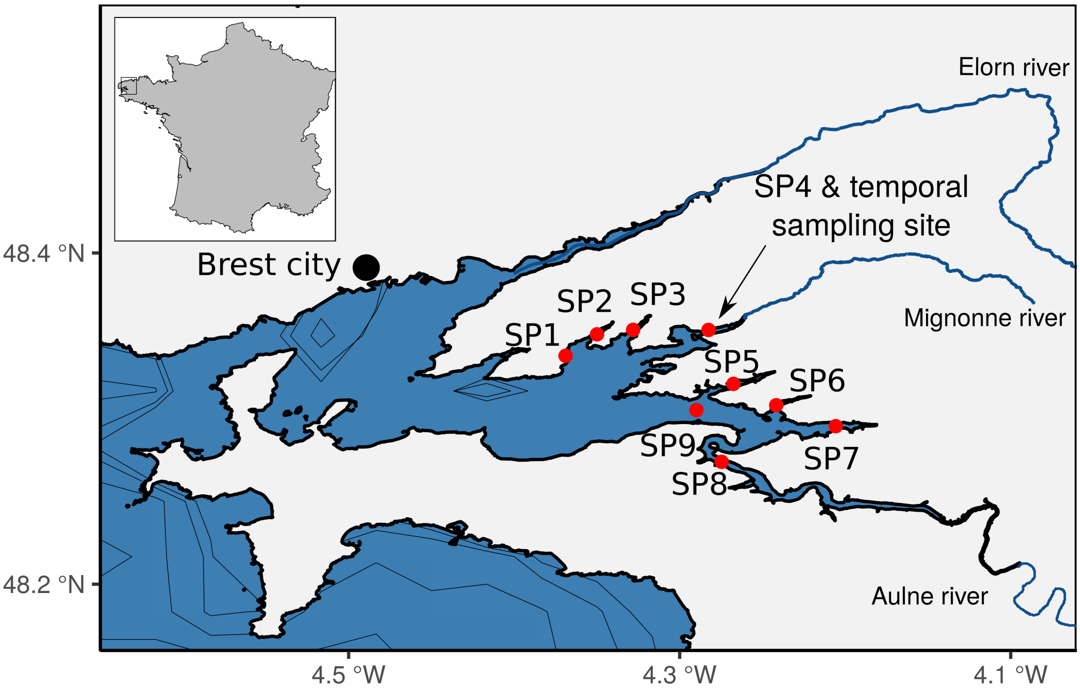


FIGURE 4.1: *Map of the sites sampled on June, 14th 2014 for the spatial sampling of the bay of Brest. SP4 corresponds to the temporal sampling site.*

The bay of Brest (Brittany, France; Figure 4.1) is a semi-enclosed, shallow (with half of its surface not deeper than 5m, average depth 8m; Chauvaud, Jean, Ragueneau, and Thouzeau (2000) coastal ecosystem, that is connected with the Atlantic Ocean by a narrow (2km) but deep (50m) strait. Two main rivers (Aulne and Elorn) are responsible for 80% of the total freshwater inputs, and high tidal variations (almost 8m during spring tides, representing an oscillating volume of 40% of the high tide volume; Chauvaud et al. (2000)) assure important mixing of the water masses.

The bay of Brest also experiences drastically different tide regimes, ranging from 1.22 meters for the lowest coefficients (20) to 7.32 for exceptionally hide tide coefficients (120; Auffret (1981))

Sampling strategy

Two strategies were used to investigate population spatio-temporal genetic structure across the bay of Brest : (i) a spatial survey : 9 populations were sampled across the bay on June, 18th 2014, four hours around high tide (from 2 hours before to 2 hours after high tide) (see figure 4.1), and (ii) a temporal survey : as described in chapter 3, a single time point (corresponding to SP4 in figure 4.1) was sampled 41 times across 3 consecutive years.

For both types of survey, as quickly as possible after sampling, water was filtered (20 μm) using a peristaltic pump and the filters were frozen into liquid nitrogen and stored at -80°C until RNA extraction. To prevent RNA degradation during filter thawing, RNA Later Ice® was added before thawing for samples collected in 2013, or samples were directly frozen with RNA Later® in 2014 (temporal and spatial samples) and 2015. Filters were ultra-sonicated on ice (Bioblock Scientific, Vibra-cell 75115) for 10 seconds at 20% intensity. RNA was extracted following the Quiagen Rneasy Plus Mini Kit® recommendations and library prepared with the Illumina Truseq mRNA V2 kit®. Samples were sequenced at Get-PlaGe France Genomics sequencing platform (Toulouse, France) on Illumina HiSeq 2000/2500 2*100 pb (2013 and 2014 (both temporal and spatial samples); Multiplex 10) and on HiSeq 3000 2*150 pb for 2015 samples (Multiplex 24).

Data preparation and read mapping

Raw reads were first characterized with FASTQC and second trimmed using TRIMMOMATIC (V. 0.33, Bolger et al. (2014)), in order to to remove ambiguous, low quality reads and sequencing adapters with parameters ILLUMINACLIP:Adapters.fasta:2:30:10, LEADING:3, TRAILING:3, MAXINFO:80:0.8 and MINLEN:70. We used a reference transcriptome previously assembled from the mRNA of 18 *A. minutum* strains (Le Gac et al., 2016). Before aligning the environmental reads to *A. minutum* raw transcriptome, preliminary precautions were taken. First, to specifically monitor *A.*

minutum mRNA reads, and to minimize contamination by reads from other species present in the community, only transcripts showing species-specific alignments were considered for the analyses. Selection for specific transcripts in *A. minutum* reference transcriptome was conducted as follows : species often co-occurring with *A. minutum* at a microphytoplanktonic community monitoring site (48.3334355, -4.3264413), next to SP4 sampling point were identified using the flora obtained by the Veliger network. Raw reads from 24 of these co-occurring species were obtained from the MMETSP database (Keeling et al., 2014), see supplementary table B.2). These 24 raw read datasets were trimmed as described above and aligned to the *A. minutum* reference transcriptome using BOWTIE 2 (V. 2.2.4, Langmead and Salzberg (2012)) algorithm for paired reads using the sensitive set of parameters. Very conservatively, all the transcripts attracting at least one read from the co-occurring species were discarded from the *A. minutum* reference transcriptome and therefore not further considered for analyses (see box 5). As a second preliminary step, we selected the more informative isoform as follows : trimmed environmental reads were aligned to the aspecific-transcripts free transcriptome previously acquired using BOWTIE 2 using the same parameters. When several isoforms were possible for a single transcript, only the isoform attracting the greatest number of environmental reads was retained for analysis. After re-aligning the trimmed reads of the 50 (41 temporal and 9 spatial) environmental samples to this transcriptome, reads with mapping quality lower than 10 and paired reads mapping to different transcript were removed.

As described in chapter 2, we monitored the presence of both cryptic species based on the fixed SNPs uncovered in Le Gac et al. (2016). As shown in figure 4.2, at the spatial level, blooms are only composed of cells belonging to cryptic species A, which allows the analysis of intra species populations genetic structure at the bay of Brest level. The temporal stability of bloom composition was previously assessed in chapter 2.

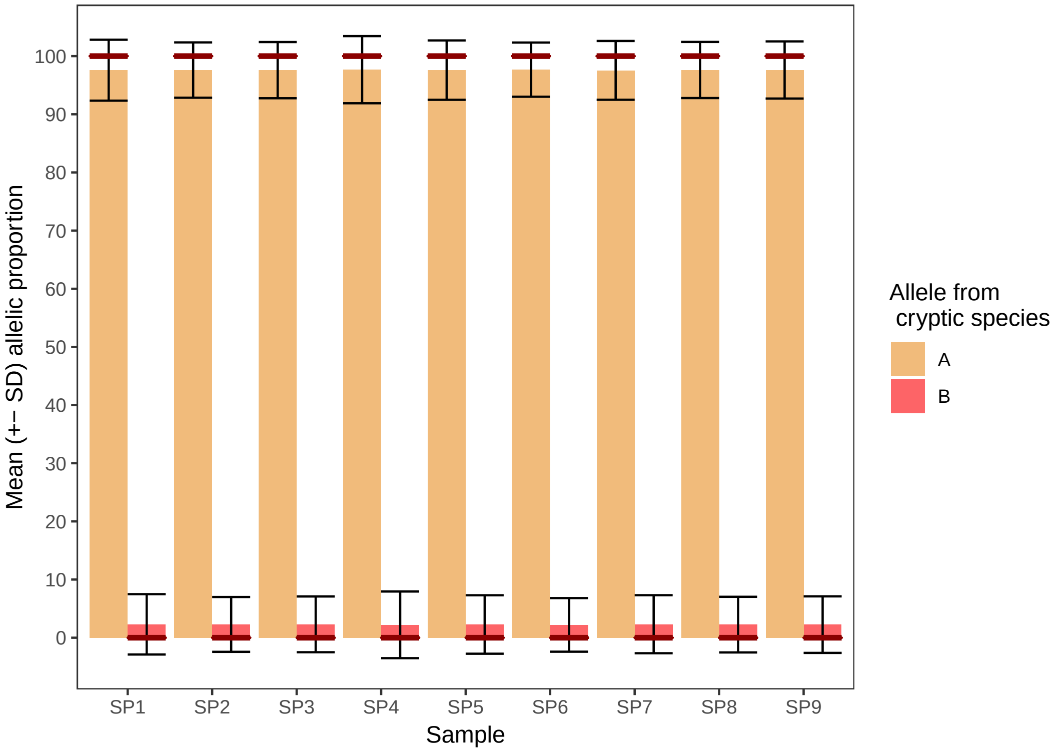


FIGURE 4.2: *Proportion of each cryptic species based on the proportions of their respective alleles (Le Gac et al., 2016) over 7,190 positions (with a depth of at least 10 in each samples).*

Modelling cell dispersion

Alexandrium minutum cell dispersion in the bay of Brest were simulated by a lagrangian model using the ICHTHYOP tool (Lett et al., 2008) on the MARS3D model configured for the bay of Brest (50 m resolution). Briefly, cell dispersion simulations were run for 5 days, between June, 13th 2014 and and June, 18th 2014, with a simulation timestep set to 6 seconds, and a recording timestep set to 5 minutes. Around each sampling station (see table 4.1), 10,000 particles were released as an uniform patch having a 50 meters radius and 1 meter depth on June, 13th 2014 at 8h30. The model used considered constant floatability and basal horizontal dispersion (set to 1 e^{-9} m^2/sec^3). For each sampling site, simulations were stopped at the time corresponding to water sampling (see table 4.1), and the respective number of particles coming from each site was retrieved using a custom R script. For each sampled population, the origin of the particles in a 1 km^2 area around the samping site was recorded. The hydrodynamic distance between sampling sites was calculated as the the euclidean distance based on the relative composition of each population.

TABLE 4.1: Sampling site location, starting / ending date and time for cell dispersal simulations

Sampling site	Latitude	Longitude	Starting date	Starting time	Ending date	Ending time
SP1	48.33653	-4.3681	13/06/2014	8h30	18/06/14	8h40
SP2	48.35012	-4.35152	13/06/2014	8h30	18/06/14	9h01
SP3	48.35448	-4.3256	13/06/2014	8h30	18/06/14	9h19
SP4	48.35363	-4.27832	13/06/2014	8h30	18/06/14	9h37
SP5	48.32252	-4.25962	13/06/2014	8h30	18/06/14	10h12
SP6	48.30932	-4.24275	13/06/2014	8h30	18/06/14	10h36
SP7	48.29582	-4.20705	13/06/2014	8h30	18/06/14	10h54
SP8	48.27767	-4.28137	13/06/2014	8h30	18/06/14	11h10
SP9	48.30448	-4.29143	13/06/2014	8h30	18/06/14	11h26

Selecting samples and Fst calculation

To account for differences in sequencing depth across the temporal samples and to ensure a robust estimation of pairwise population genetic structure, only samples with at least 150,000 SNPs (of the 457,358, *i.e.* more than 30%) with a depth of at least 10 was conserved for further analysis. Pairwise Fst values between the selected samples were calculated using the POOLFSTAT R package (Hivert, Leblois, Petit, Gautier, & Vitalis, 2018). The R function cor.test was used to test the correlation (using a spearman correlation coefficient) between genetic differentiation and five types of distances : geographical (by using population connectivity as a proxy of the "hydrological distance"), temporal (in number of days separating each sample), ecological (both from biotic and abiotic variables) and "physiological" (see below). Ecological distances were estimated using the euclidean distance between samples in a PCA geographical projection. For the abiotic factors, PCA was led using eight variables (tidal coefficient, ammonium, nitrogen oxide, phosphate and silicate concentrations (μM), sea surface temperature (°C) and irradiance (W m^{-2}) and salinity), and euclidean distances calculated on the two first axis (explaining 57.6 % of the observed variance). For the "ecological biotic distance", euclidean distances were calculated on the two first axis of a PCA (explaining 46 % of the observed variance) conducted using the relative abundance of the 12 most abundant species (number of reads in the environmental samples : *Alexandrium minutum*, *Chaetoceros curvisetus*, *Prorocentrum micans*, *Chaetoceros* cf. *neogracile*, *Lankesteria abbotti*, *Micromonas* CCMP1646, *Ostreococcus lucimarinus*, *Helicotheca tamesis*, *Leptocylindrus aporus*, *Skeletonema dohrnii*, *Gonyaulax spinifera* and *Thalassiosira punctigera*) in the samples (see chapter 5 for details). The "physiological" distance refers to the divergence in expression levels of eight major groups of genes co-expressed during the respective blooms (see chapter 3 for details). This distance was calculated based on the euclidean distance between the time points in the geometrical projection of the two first axis of a PCA (explaining 82.8 % of the observed variance).

4.2.3 Results

Among the 50 environmental samples, only the ones showing a depth of at least 10 in at least 30% of the SNP positions were considered informative enough for analysis. After this conservative threshold, 34 samples were kept for analysis (the 9 spatial samples and 23 temporal samples : 5 from 2013; 11 from 2014 and 7 from 2015).

Spatial genetic structure

TABLE 4.2: Composition of each sampling site (in column) according to the number of cells coming from the other sampling sites (in row)

	SP1	SP2	SP3	SP4	SP5	SP6	SP7	SP8	SP9
SP1	111	8	3	4	1	1	0	7	50
SP2	547	287	15	79	1	2	0	5	37
SP3	320	69	190	73	3	4	2	20	48
SP4	70	12	8	1078	5	5	8	50	52
SP5	13	19	8	38	117	13	1	72	208
SP6	12	6	0	3	10	31	37	206	112
SP7	12	5	4	2	18	26	45	194	91
SP8	16	5	4	4	15	22	25	158	95
SP9	20	1	2	8	2	1	3	28	42

TABLE 4.3: Pairwise Fst values between the spatial samples

	SP1	SP2	SP3	SP4	SP5	SP6	SP7	SP8	SP9
SP1	0	0.0021	0.0016	0.0028	0.0017	0.0025	0.003	0.0033	0.0024
SP2		0	0.0022	0.0033	0.0021	0.0032	0.0034	0.0041	0.0032
SP3			0	0.0023	0.002	0.0028	0.0033	0.0038	0.0028
SP4				0	0.0029	0.0042	0.0043	0.0051	0.0039
SP5					0	0.0027	0.0031	0.0038	0.0028
SP6						0	0.0041	0.0047	0.0035
SP7							0	0.0049	0.0039
SP8								0	0.0045
SP9									0

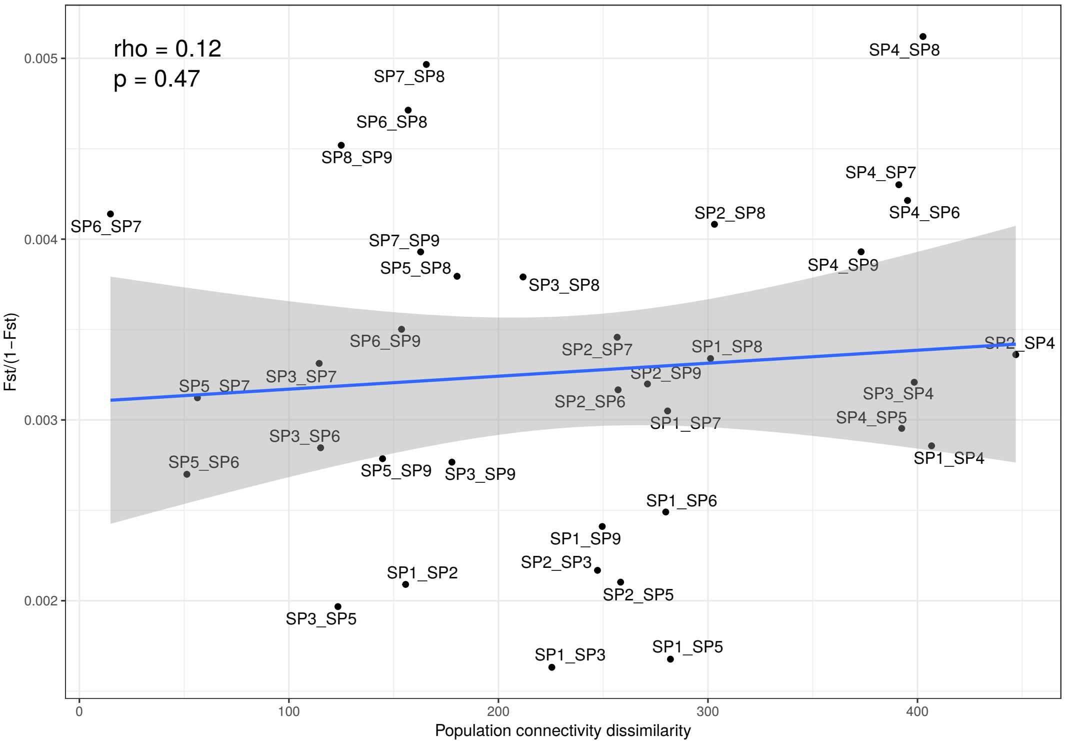


FIGURE 4.3: *The points represent the genetic distance (Fst/(1-Fst)) between the 9 spatial samples in function of the their connectivity (based on simulations of cell dispersal from each sampling site). Spearman's rank correlation test values (rho and p-value) are indicated.*

Following the filters applied for SNP selection, 297,136 positions were included in the pairwise Fst calculation between the 9 spatially distributed samples. Pairwise Fst values are presented in the table 4.3. They range from 0.0016 (SP1 vs SP3) to 0.005 (SP8 vs SP4 are the most divergent populations). To investigate whether such genetic structure might be the result of isolation-by-distance, we estimated the connectivity between the sampled sites. To do so, we simulated the dispersion of 10,000 particles (that represent A. minutum cells) released at each sampling site (SP1 to SP9, see figure 4.1). At the end of the simulation, the position of each particle was recorded, and each sampling site was characterized according to the number of particles from each other respective sites present on a 1km^2 area around the sampling point. Results of these simulations are presented in table 4.2 and figure 4.4. First, we want to indicate that the composition in particle SP3 is rather different from the other sites. Overall, this site gathers few particles, mostly due to auto connectivity (*i.e.* the vast majority of the particles in the sampling area are from this area). Although such pattern is also observed in other site (SP2 and SP4 for instance), it appears that SP3 connectivity estimation site particularly suffers from this. Therefore, it lowers the reliability of the comparisons involving SP3. As an example, SP3 and SP7 appear to be highly connected in figure 4.3, while their composition is very different (table 4.2. However, removing SP3 from the isolation-by-distance analysis does not change the results, and was therefore left on the figure and table before in-depth investigations that was beyond the scope of the present study (but is not further discussed).

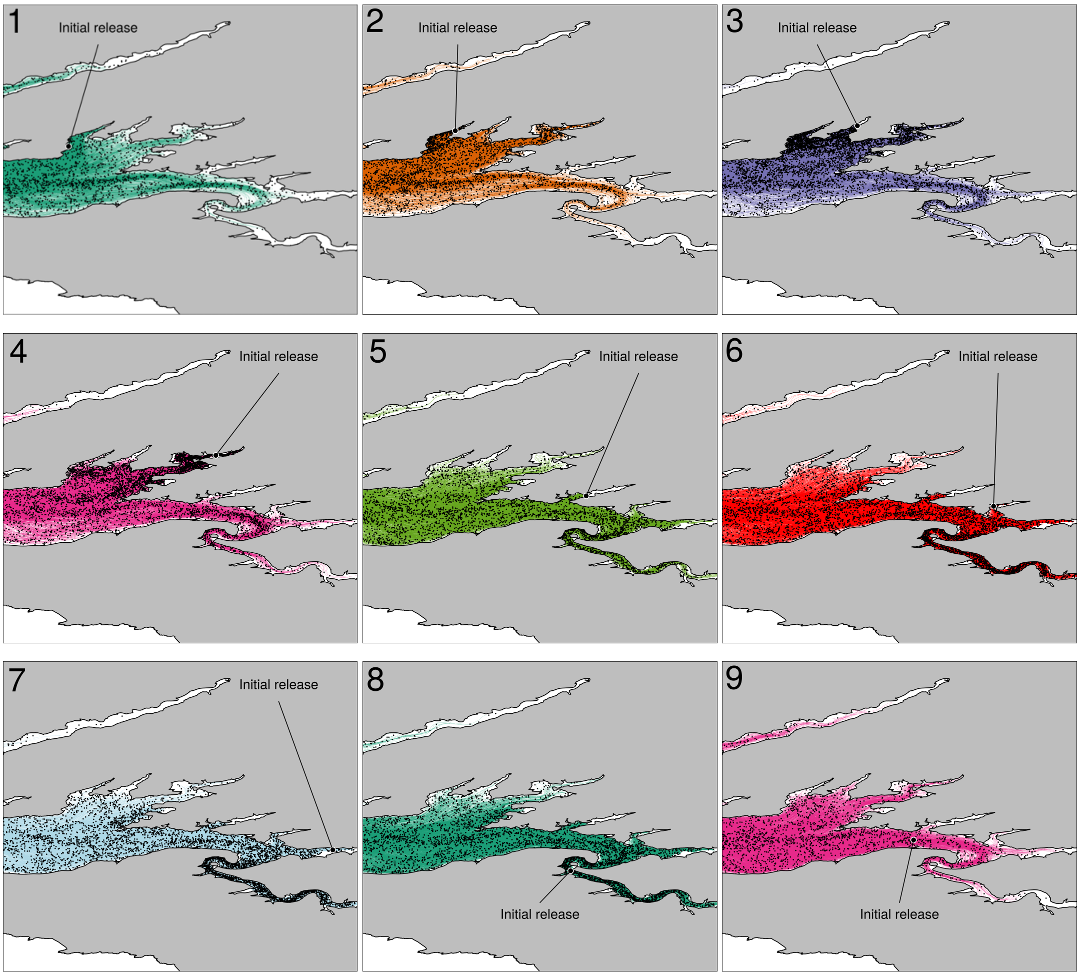


FIGURE 4.4: *Results of the cell dispersal simulation. Each plot (zoomed on the area of tha spatial sampling : bottom area of bay) represent the dispersion, from each sampling site (**1** - **9**), of 10,000 particles in the bay of Brest after 5 days of simulation between June, 13th and June, 18th 2014*

The figure 4.3 shows the correlation between genetic distance and connectivity dissimilarity between the sampled points. Overall, it shows that there is no correlation between inter sample connectivity (x axis) and their genetic distance (y axis). Indeed, although closely situated samples generally show similar population community patterns (*i.e.* according to the dispersal simulation, the cells composing these populations come from different area, see for example SP6 vs SP7 and SP5 vs SP6) and vice versa (see SP1 vs SP8, SP2 vs SP9 for instance), genetic distance does not follow the same pattern. For instance, populations sampled in SP6 and SP7 show high hydrodynamic connectivity (x axis), but relatively high genetic distance (y axis). On the other side, populations sampled in SP1 and SP4 show lower genetic distance but low hydrodynamic connectivity (on the right side of the x axis). Interestingly, it appears that SP4 sampling site is relatively disconnected from the other populations sampled (every comparison involving the SP4 sample is on the extreme right side of the x axis, which can also be seen on figure 4.4 **4**). Also, they show high variability in genetic distance values with the other populations (from 0.0029, SP1 vs SP4 to 0.0051 for SP1 vs SP8, table 4.3). This indicates that, despite being highly disconnected, some populations can show very variable levels of genetic differences. In reverse, we can see that some populations display very different connectivity patterns but also small genetic structure. For instance, the population sampled in SP7 shows similar genetic distance with populations sampled in SP5 and SP1, but very different connectivity patterns (high similarity with SP5, left side of the x axis but high dissimilarity with SP1, right side of the x axis).

Temporal genetic structure

We investigated the effect of various factors on the temporal genetic structure at a single sampling point (SP4, see figure 4.1).

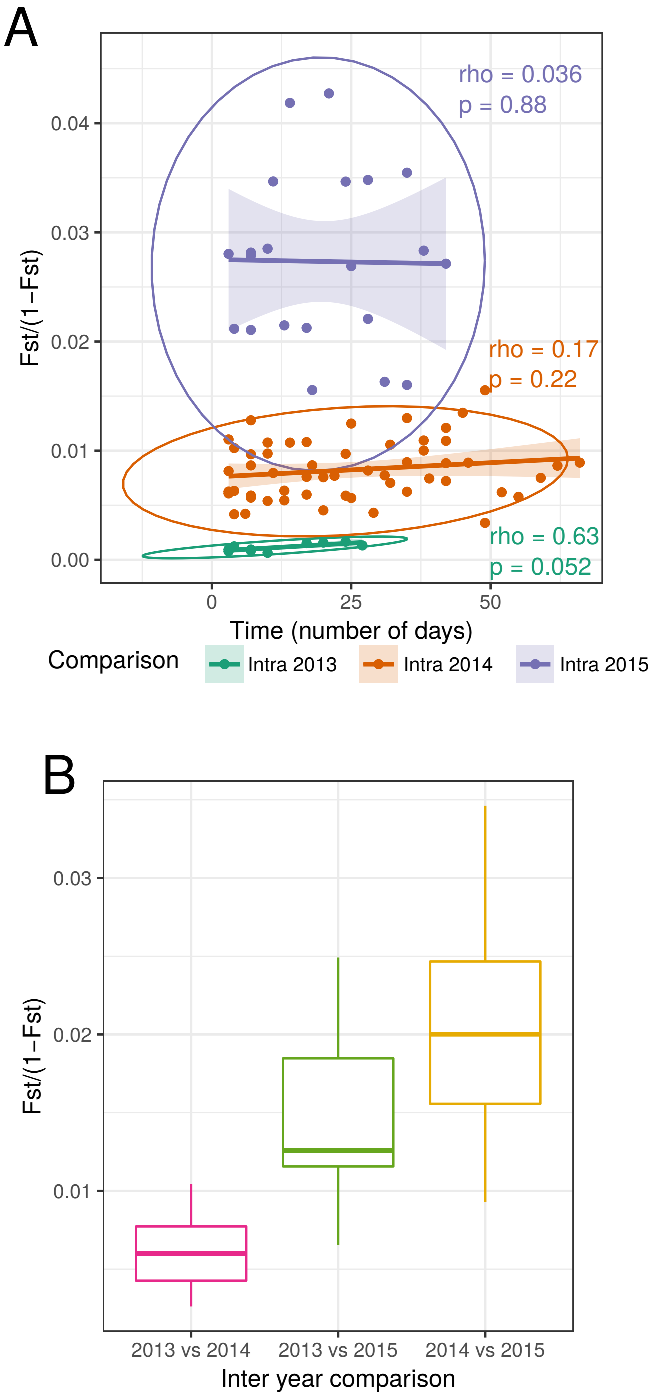


FIGURE 4.5: ***A*** *: Relationship between the genetic distance (Fst/(1-Fst)) between the temporal samples (each point) in function of the distance (in days of bloom) between the sampled time-points.* ***B*** *: Distribution of the genetic distance inter-bloom.*

For each bloom sampled respectively in 2013, 2014 and 2015, we tested whether they showed indices of "isolation-by-time" by correlating their intra-year (intra-bloom) genetic distance to the time between each sample (see supplementary table B.1). The figure 4.5 **A** represents this correlation (see supplementary table C.1 for details), and shows that the different blooms display various patterns of intra-years genetic structure patterns. Within a given year, differentiation among samples dramatically increase from 2013 to 2014 and then 2015, with the latter showing both the higher intra year genetic structure (both in terms of intra bloom structure and variability, ranging from 0.016 between 09/07/15 and 27/07/15 to 0.042 between 29/06/15 and 20/07/15 samples; figure 4.5 **A**). Both 2013 and 2014 blooms show lower intra-year genetic structure relative to 2015 bloom, but they display moderate indices of temporal isolation. In particular, 2013 (despite low Fst values), shows the higher genetic differentiation correlation with time (rho = 0.63, p = 0.052), a pattern less observed in 2014 (rho = 0.17, p = 0.22), and not observed at all in 2015 (rho = 0.036; p = 0.88). Between years (figure 4.5 **B**), genetic distance increases between the comparison 2013 vs 2014, 2013 vs 2015 and finally 2014 vs 2015. The distribution of inter-year genetic distance values reflects the patterns observed in figure 4.5 **A**, with higher genetic differences (and variability) when it comes to comparison involving 2015 bloom, further underlining the high genetic structure variability within and between blooms. Finally, we note that genetic distance within 2015 tend to be higher than between years.

Ecological genetic structure

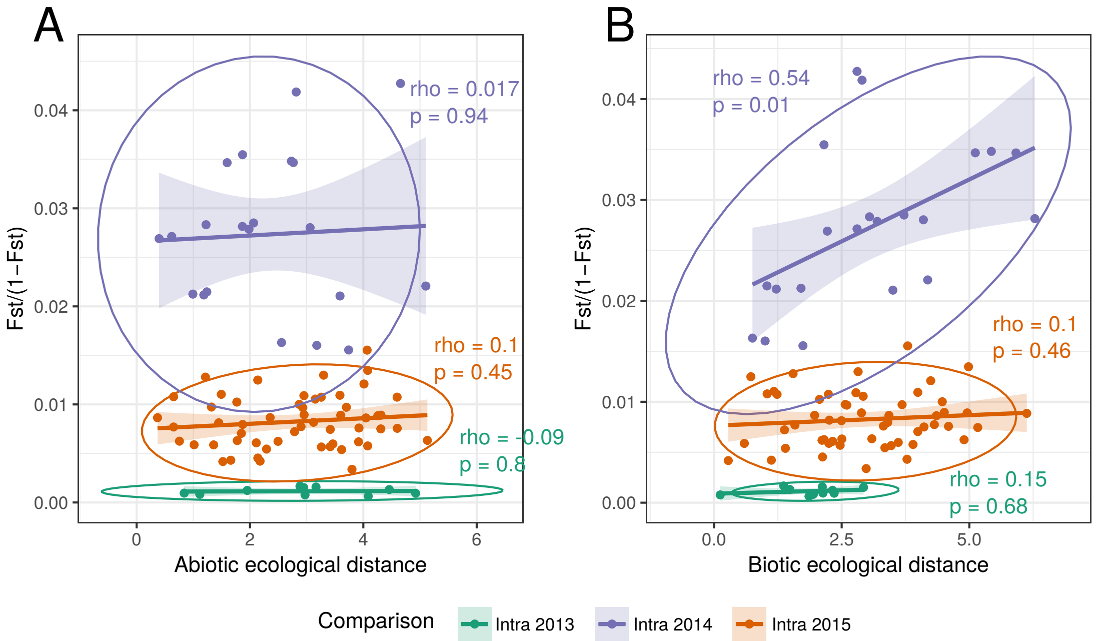


FIGURE 4.6: *Relationship between the genetic distance (Fst/(1-Fst)) between the temporal samples (each point) in function of their ecological dissimilarity (euclidean distance) between the sampled time-points.* ***A*** *: abiotic variables* ***B*** *: biotic variables.*

To explore whether these genetic structure patterns changes could be explained by environmental factors, we investigate the effects of both biotic and abiotic factors fluctuations on genetic structure level changes within blooms. The figure 4.6 **A** & **B** represents the correlation between the within bloom genetic distance and ecological dissimilarity (based on the euclidean distance from a PCA analysis, see method). Within every bloom, genetic differentiation variations show no correlation with abiotic factors fluctuations (respectively rho = -0.09, p = 0.8; rho = 0.1, p = 0.45; rho = 0.017, p = 0.94; figure 4.6 **A**). Interestingly, even populations living under very different environmental conditions (when considering abiotic factors only) show no indice of genetic structure. However, when considering biotic ecological distance (*i.e.* dissimilarity in species community), it appears that genetic differentiation increases, at least within 2015 bloom (figure 4.6 **B**; rho = 0.54, p = 0.01). It does not seem to be the case in 2013 and 2014 (respectively rho = 0.15, p = 0.68 and rho = 0.1 (similar as in abiotic factors, see figure 4.6 **A**), p = 0.46). Very interestingly, within 2014 and 2015 blooms show similar magnitudes of biotic dissimilarity (figure 4.6 **B**, x axis), that are higher than within 2013 bloom (small intra bloom differences in biotic variables, figure 4.6 **B**, x axis).

Gene expression divergence

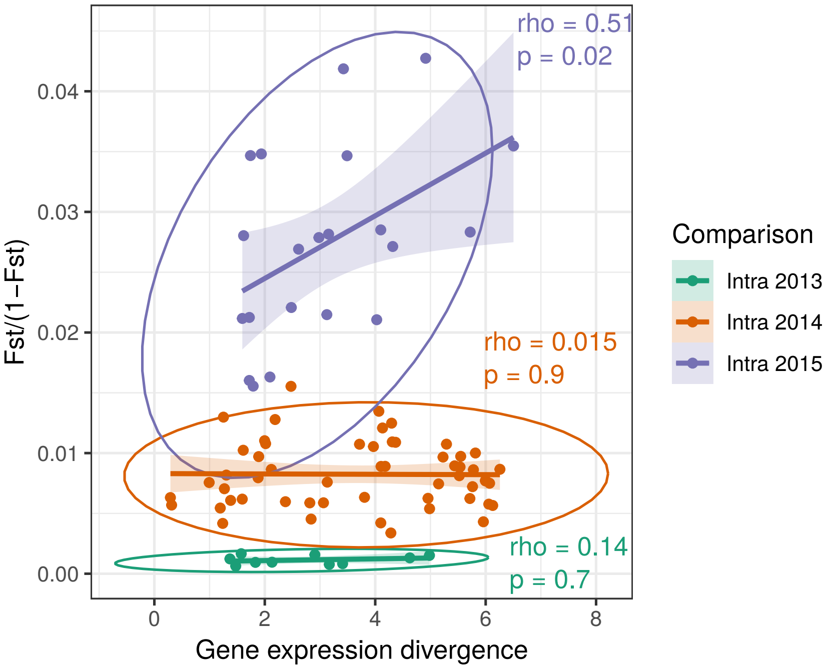


FIGURE 4.7: *Relationship between the genetic distance (Fst/(1-Fst)) between the temporal samples (each point) and their gene expression levels dissimilarity (euclidean distance).*

To investigate whether genetic changes are linked with changes in gene expression levels, which could indicate the traduction of phenotypic plasticity into genetic adaptation (*i.e.* genetic assimilation, see box 2), we correlated within blooms genetic distance with gene expression divergence of their major expression modules (described in chapter 3). Again, the different blooms show very different patterns of genetic distance changes (figure 4.7). We note that it is impossible to directly compare inter year gene expression levels due to normalization for year specific batch effects (see chapter 3.2.3). It appears that within 2015 bloom, the genetic differentiation between the samples is positively correlated (rho = 0.51, p = 0.02) to changes in gene expression. Interestingly, this pattern is not the observed at all in 2013 and 2014 blooms (respectively rho = 0.014, p = 0.7; rho = 0.015, p = 0.9).

Link between gene expression divergence and biotic ecological distance

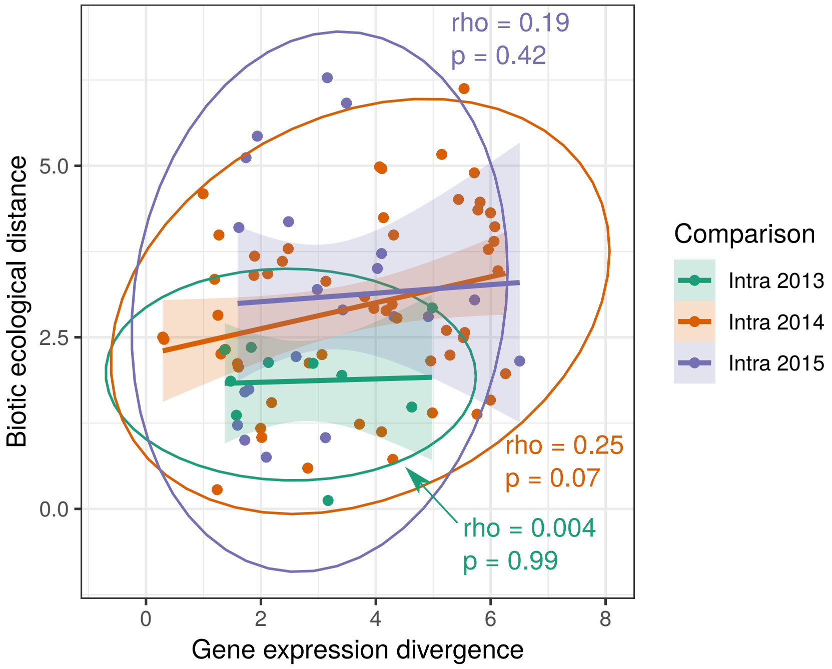


FIGURE 4.8: *Relationship between gene expression divergence between the temporal samples (each point) and their biotic ecological distance (euclidean distance).*

To investigate the interplay between gene expression and changes in the community co-occurring with *A. minutum*, we correlated gene expression divergence and the biotic ecological distance within bloom. As presented in figure 4.8, the biotic ecological distance between the samples is not correlated with their respective changes in gene expression levels of their major expression modules (described in chapter 3). However, each bloom displays particular patterns. 2013 bloom for instance, experience small changes in biotic factors, and moderate changes in gene expression. On the other side, 2015 shows medium gene expression divergence, but high variations in biotic ecological similarity (figure 4.8, y axis). 2014 bloom displays an intermediate pattern, with both high variations in gene expression divergence and biotic ecological distance. Also, 2014 bloom displays higher correlation between gene expression levels divergence and biotic factors dissimilarity (*i.e.* gene expression changes the most between samples that also experience the most dramatic changes in biotic community; rho = 0.25, p = 0.07).

4.2.4 Discussion

Phenotypic plasticity and genetic adaptation both contribute to the biological response to environmental fluctuations. However, separating the respective contribution of each of them is harduous, and greatly depends on the taxa considered. Although the time scale observation of these types of response is widely different for long living species (with fast appearing phenotypic plasticity but centennial scale genetic adaptation), this response continuum can theoretically be observed together in action for short lived species such as microorganisms. Here, using allelic frequencies differences at previously characterized variable SNP positions, we used metatranscriptomic dataset to investigate the spatio temporal genetic structure of blooms from a toxic dinoflagellate in the bay of Brest. As no assembled genome of *A. minutum* is available, environmental reads were mapped to a reference transcriptome coming from 18 strains, and the SNP positions analysed were previously curated from variable positions showing indices of paralogy (*i.e.* displaying a biallelic state in the haploid strains used for the de novo assembly, see Le Gac et al. (2016).

Using mRNA in population genetics

We are aware of the potential drawback of using transcriptomic datasets in a population genetic analysis, as there might be a confundant effect of mRNA levels (due to gene expression changes) and allelic frequencies variations. Indeed, the genetic structure observed using mRNA can result from (i) genetically different populations that have the same gene expression levels; (ii) genetically similar populations but with genotype-specific differential gene expression; or (iii) the interaction of the two : genetically different populations, in which some genotypes also experience phenotypic plasticity (through gene expression modulation). Here, the SNP analysed were discovered and selected without prior expectations, and Fst values calculated globally, which we believe moderates the effect of high gene expression levels fluctuations on genetic structure (if any). However, more than being only limitations, such characteristics of transcriptomic data could be of primary interest for investigating the result of genetic adaptation in gene expression levels. Indeed, we do show that a portion of the genetic structure observed can be explained by gene expression level fluctuations (figure 4.7, point (iii)). Hereafter, we quickly address this aspect, while knowing that it should require further in depth investigations. To separate this potential confundant effect inherent of mRNA-based approached, we could lead similar analysis on the genes that showed no changes in expression through time (*i.e.* "housekeeping genes"; see chapter 3).

Spatial genetic structure

Geographical distance between populations is one of the key factors directly influencing their ability to exchange individuals, and population connectivity highly depends on both individuals dispersal capacities and environmental characteristics. In the marine environment, currents and water masses shift by tide are known to greatly shape population connectivity and genetic structure patterns (see for example Gilg and Hilbish (2003) in mussel larvae dispersion by currents). To address this, we correlated the genetic differentiation observed between the populations sampled around the bay of Brest with their connectivity. To do this, we estimated an "hydrological distance" between the samples, based on dispersion models. The basic idea is that as more connected populations exchange more individuals, we should be able to estimate the composition of a particular population based on cell-dispersion patterns from other sites. Then, we simulated cell dispersion from each of our sample locations (by modeling the dispersion of 10,000 particles from each sampling site), and analysed their localization at the end of the simulation. Populations displaying similar compositions (*i.e.* with the same pool of particles) being more connected. Here we only tested the effect of hydrodynamic in population genetic structure, and showed that it does not explain the differences observed. We note that incorporating biological constraints such as cell mortality or division into the model would help refining population connectivity estimations. Also, only the final position of particles was recorded, and particle trajectories should be further investigated (for instance studying the time spent in each area, the number of times one particle passed through a given area etc). However, the hydrological model used here is very precise (50 m), and modulating the time of the simulation (from 3, 4 and 5 days) did not change the results (data not shown), which comforts the reliability of our connectivity distance estimations. Integrating hydrological parameters are of primary importance, as we are looking at very fine scale genetic structure. Previous studies on *Alexandrium* genetic diversity dynamics were investigated at very large scales (150 km in Dia et al. (2014), hundreds to thousands kilometers in Nagai et al. (2007)), which might prevent too dramatic impacts of hydrodynamics. Nevertheless, even in the case of very distant populations, considering that isolation-by-distance processes in the marine environment is not impacted by hydrologically-constrained cell dispersal is dubious. For instance, Casabianca et al. (2012) studied genetic structure between *A. minutum* populations distant from several hundreds of kilometers and, helped with oceanic circulation models, highlighted the necessity to incorporate environmentally-triggered cell dispersal to population genetic investigations. When scaling down to a highly dynamic semi enclosed marine area as the bay of Brest Auffret (1981); Chauvaud et al. (2000), it seems therefore of primordial importance. Here, we investigated the spatial structure of 9 *A. minutum* populations, distributed in an area of less than 40 km^2. Overall, we show no pattern of isolation-by-distance, but our result address several aspects of population connectivity. Noteworthy, we show that distance between two populations is not synonym of genetic differentiation, and very closely situated populations (for example SP6, 7, 8 and 9) can experience higher genetic structure than distant populations. This suggests that local conditions have significant impact on genetic differentiation. However, we may hypothesize that biologically connected areas (*i.e.* that exchange individuals) are also ecologically connected (high environmental homogeneity), but our results suggest very different situations. Indeed, the genetic differentiation observed between populations is not only explained by hydrological connectivity nor local abiotic environmental conditions. Another factor that could be responsible for such population differentiation would be a "priority effect", which gives an advantage to species or genotypes early establishing in vacant niches over later-arriving ones (Connell & Slatyer, 1977). Therefore, it can structure populations on a long time scale after the rapid growth of a small number of individuals locally well-adapted, that monopolize resources and hamper other colonizers to establish in the populations. Despite that it might be difficult for this "priority effect" to apply in microbial species, due to large population sizes (several thousands of cells per liter) and (potentially) low competition for resources, it was proposed by Sefbom, Sassenhagen, Rengefors, and Godhe (2015) as a potentially important factor of population differentiation in strains of Skeletonema marinoi (a bloom forming marine diatom).

Temporal genetic structure

In two *Alexandrium* species (respectively *A. minutum* and *A. fundyense*), Dia et al. (2014) and Richlen, Erdner, McCauley, Liberal, and Anderson (2012) showed a significant genetic differentiation of cells through blooms. Here, using more than 290,000 SNPs over 23 temporal samples (5 from 2013; 11 from 2014 and 7 from 2015), we show contrasting results. Although there is no statistically significant isolation-by-time within the different sampled bloom, the moderate positive correlation between genetic difference and time separating samples in 2013 and 2014 blooms are still congruent with the observations of Dia et al. (2014) and Richlen et al. (2012). In this case, it reinforces the hypothesis whether bloom events are composed of groups of genetically divergent individuals succeeding each other through time, as a competition avoidance strategy between divergent populations for instance.

Over time, we observed very high variations in within blooms genetic structure. This result is surprising, and may be linked with large temporal scale *A. minutum* dynamics. In the bay of Brest, *A. minutum* blooms were shown to increase since 1990, a pattern that accelerated since 2010, with a peak in 2012 (more than 40 million cells per liter), before gradually decrease every year since then (Chapelle et al., 2014, 2015b), with barely no detectable blooms since 2015 (data not shown, REPHY network). It could be interesting to investigate whether the increasing variability in genetic differentiation observed between 2013 and 2015 is related to the overall less dense (and shorter) blooms observed since the inter-annual record year of 2012. Such observations may indicate that more dense blooms are formed by more genetically similar populations, while genetically divergent populations are less likely to form dense blooms. In this scenario, it could be due, for instance, as a trade-off between exponential growth of locally well adapted genotypes (promoting dense blooms of genetically homogeneous populations) and natural selection of different genotypes (promoting genetic structure) imposed by selective pressures arising from repeated environmental stresses (from both biotic or abiotic factor fluctuations).

Environmental genetic structure

Numerous studies investigated the factors underlying blooms initiation, growth and decline (see for instance Bravo et al. (2008); Guallar et al. (2017); Laanaia et al. (2013) among others). Although the effects of such environmental factors on *A. minutum* individual's growth are relatively well described, much less is known on their effect on genetic structure over the blooming period. Here, we present results that address this question. Previously, Rynearson et al. (2006) showed that environmental selection was a significant factors regulating the bloom dynamics of different populations of the diatom *Ditylum brightwellii*. Here, although we previously showed that changes in abiotic environmental factors are correlated with gene expression levels variations (chapter 3, Metegnier et al. (2018a *Submitted*)) our results show that changes in genetic structure are not significantly correlated with fluctuations of abiotic variables. However, the biotic composition of the microbial community is a factor explaining some of the changes in genetic composition observed (at least within 2015 bloom and between 2013 and 2015 and between 2014 and 2015 blooms). Such effects of biotic interactions on bloom dynamics were also showed by Chambouvet et al. (2008), who studied host specific interactions inside algal blooms, and showed that the temporal succession of the parasitoid species is -at least partially- a factor controlling bloom dynamics (see also Blanquart et al. (2016) in *A. minutum* blooms). Our results suggest that blooms are not only influenced by such potentially highly specific parasitic interactions, but rather that community-wide changes have the potential to influence *A. minutum* populations genetic structure.

Investigating the link between these observations and the effect of environmental factors on genetic differentiations should be of primary interest to understand bloom dynamics, especially in the context of toxic and potentially invasive species. As discussed before, *A. minutum* blooms show drastic inter-years variations in terms of bloom length and cell density (Chapelle et al. (2015b); Rephy network), which underlines how bloom dynamics might be triggered by numerous factors. For instance, we showed that genetic structure (at least within 2015 bloom) can be correlated with the fluctuations of biotic factors. Therefore, it clearly appears that bloom dynamics are the result of complex and various factor fluctuations, that deserve high attention and further investigations.

Link between phenotypic plasticity and genetic adaptation

The contributions of both phenotypic plasticity and genetic adaptation toward environmental fluctuations are part of a wide scale response continuum, and their respective proportions in population resilience to perturbations greatly depend on the taxa considered. In microbial species, due to their short life cycles and large effective population sizes, these two kinds of responses are however possibly acting at analytically compatible time scales. Here, we present evidences for this continuum to occur in the natural environment. First, we showed that changes in community composition (figure 4.6 **B**) are correlated with genetic structure during 2015 bloom, and that genetic structure is also correlated with gene expression levels variations (figure 4.7), indicating that some variations in molecular physiology result from such genetic changes. Second, in chapter 3, we showed that gene expression variations were correlated with abiotic environmental factors and that (although here genetic structure is not correlated with abiotic variables), some groups of genes were overrepresented in functions related to interspecific interactions (see chapter 3 and Metegnier et al. (2018a *Submitted*). Altogether, the result presented here might reflect the genetic adaptation to interspecific interactions, together accompanied by changes in molecular physiology. Such natural selection of gene expression variations were already studied in numerous organisms (see review in Fay and Wittkopp (2008); Gilad, Oshlack, and Rifkin (2006) or for instance Fraser, Moses, and Schadt (2010) in yeast, Fraser (2013) in humans and Nourmohammad et al. (2017) in flies, among others), and here, we present another example of how environmental fluctuations selection affects both genetic differentiation and gene expression dynamics.

4.2.5 Conclusion

Genetic structure can arise from both genetic drift or natural selection from selective pressures imposed by the environment. As a consequence, disconnected populations, or populations experiencing differential pressures may diverge. Understanding such processes and the factors underlying genetic divergence is not straightforward, and difficulties add up when considering marine microbial populations, living in close and dynamics areas. Here, we investigated the spatio-temporal genetic structure of a marine bloom forming species using allele frequency changes at almost 300,000 SNP positions. Using cell dispersion simulation, we tested the effect of hydrological dynamics on population connectivity and genetic divergence. As expected, hydrological characteristics have primary impact on population connectivity. What was less expected, however, was that genetic distance between spatially closely distant populations (around 40km^2) was not only influenced by dispersion from water mass. To investigate further the factors that may influence genetic differentiation patterns, we looked at the effect of various factors on genetic structure through time. Very interestingly, the three bloom sampled show widely different genetic structure patterns. Among such difference are the increasing in both genetic differentiation between blooms, and increasing in within bloom genetic differentiation between 2013, 2014 and 2015 blooms. Also, these bloom-specific genetic structure patterns were correlated to different factors. Noteworthy, 2013 bloom shows increasing genetic structure with time, while genetic changes observed in 2015 were correlated to changes in the community, a pattern that was not observed for the other blooms. We can not exclude the priority effect, which allows the first genotype arriving in a given area to monopolize ressource, and therefore may give an advantage to the first cells that excyst at the beginning of the blooming period. Also, we showed that genetic structure is linked with changes in gene expression levels (in particular in 2014 bloom). This result is of primary interest, as it points out how phenotypic plasticity may be linked with genetic adaptation. Altogether, these results underline how microbial populations can experience genetic structure on a (relativelly) short timescale, and how numerous factors (biotic, abiotic, hydrological) can interfere and influence population divergence.

4.3 Addendum to chapter 4 and technical considerations

During the temporal survey (*i.e.* twice a week as long as *A. minutum* cell concentrations were superior to 10,000 cells.L^{-1} and until this density dropped below this threshold for two consecutive samples) water was sampled twice, the microphytoplanktonic community collected on two separate filters, and stored in the same conditions before RNA extraction. For each sample, these second filters (hereafter called "XX/XX/XXXXBis") typically served to re-extract mRNA in cases where technical difficulties didn't allow to use the first ones. At the same time as 2015 samples, mRNA from six samples from the previous sampling campaign were re-extracted, sequenced (22/07/2013Bis, 01/08/2013Bis, 16/06/2014Bis, 27/06/2014Bis, 30/06/2014Bis, 25/07/2014Bis), and compared to the samples coming from their "pair" filters (respectively 22/07/2013, 01/08/2013, 16/06/2014, 27/06/2014, 30/06/2014, 25/07/ 2014), as a mean to estimate the "intra-population" structure (*i.e.* the observed genetic structure inside a water mass sampled twice few minutes apart). The same protocol as described in chapter 4 was used to calculated allelic frequencies and estimate genetic structure. The background hypothesis was that these "Bis" samples would show no or small genetic differentiation with their pair. In the case where a small genetic structure between the two filters was observed, it would have indicated rather batch effect, or that the sampling of 10 liters of water would not have been sufficient to capture the entire genetic diversity of the bloom (as the two samples would have gathered genetically different populations of cells). However, the situation is different, and the table 4.4 present the results of the pairwise Fst estimates between the temporal samples used in chapter 4 and the "Bis" samples corresponding. Using the same threshold as for the sample selection (at least 30% of their SNPs displaying a minimal number of 25 reads), only two "Bis" samples could be used (22/07/2013Bis and 01/08/2013Bis).

TABLE 4.4: Pairwise Fst values computed on 296,470 SNP positions using the POOLFSTAT R package. Fst values for comparison involving "Bis" samples are indicated in bold

	05/07/13	22/07/13	**22/07/13Bis**	01/08/13	**01/08/13Bis**	06/06/14	10/06/14	23/06/14	27/07/15
05/07/13	0	0.0015	**0.0188**	0.0013	**0.0158**	0.007	0.006	0.007	0.0073
22/07/13		0	**0.0175**	0.0006	**0.0145**	0.0062	0.0054	0.0064	0.0065
22/07/13Bis			**0**	**0.0181**	**0.0294**	**0.0231**	**0.0219**	**0.0229**	**0.022**
01/08/13				0	0.0149	0.0065	0.0052	0.0063	0.0067
01/08/13Bis					**0**	**0.02**	**0.0178**	**0.0197**	**0.0174**
06/06/14						0	0.0041	0.0059	0.0115
10/06/14							0	0.0054	0.0122
23/06/14								0	0.0122
27/07/15									0

The pairwise Fst values presented in table 4.4 show that every comparison involving a “Bis” sample shows indices of genetic structure. Moreover, it appears that such Fst values between any sample and the “Bis” samples are higher than any other comparison. This result is highly surprising, and requires some attention. Taken alone, these results suggest that two water samples taken few minutes apart are genetically more different than samples coming from blooms separated by two years, suggesting that, by chance, we sampled more genetically similar cells (several hundreds of thousands) in 22/07/13 and in 27/07/2015 (two years apart; Fst = 0.0065) than between the two 22/07/13 (22/07/13 and 22/07/13Bis) samples (few minutes apart; Fst = 0.0294).

Aside from the “chance factor”, we tried to uncover other factors that may be responsible for these results. In particular, among the differences existing between the pair of samples can be cited : (i) the sequencing technology (“Bis” samples were sequenced on Illumina 3000 while their paired samples were sequenced on Illumina 2000/2500), (ii) the person who prepared the samples (“Bis” samples were prepared by me while their paired samples were prepared by Mickael Le Gac and Julien Quéré), and (iii) the storage time (mRNA from the “Bis” samples were extracted in 2015, more than two years after their paired samples). Giving the results presented here, the two first possibilities can be excluded. Indeed, sample 27/07/15 shows similar genetic structure with sample 22/07/13 (Fst = 0.0065) than sample 22/07/13 with 23/06/14 (Fst = 0.0064) for instance, and if this genetic structure was artifactually due to technical biases (such as the person who prepared the samples or the sequencing technology leading to preferential mRNA extraction or sequencing of certain cDNA fragments), high Fst values would be expected for every samples prepared in the “2015 batch”, which is not the case.

The third possibility is the different storage time sample experienced, that could have altered a preferential part of the mRNA in the “Bis” samples, leading to biased fragment sequencing and therefore false allelic frequencies calculations. Although storage at -80°C is a common practice and was shown to be a fairly good procedure for RNA conservation, Seyhan and Burke (2000) showed that some ribozyme molecules are still active at -70°C. Therefore, we can not exclude that a small amount of such molecules or more generally ribonucleases (called RNases, *i.e.* nucleases that catalyse the degradation of RNA molecules), were still active during the storage phase, and slowly degraded preferential parts of the mRNA samples over time. However, numerous studies stated that storage a -80°C. allows long term conservation of RNA samples (see for instance Gorokhova (2005) in *Artemia* spp. : 8 months; (Auer et al., 2014) in human tissues : 8 years; (Duale et al., 2014) in blood samples : 6 years; (Andreasson, Kiss, Juhlin, & Höög, 2013) in human tumor tissues : almost 30 years). Olivieri et al. (2014) showed that RNA concentration is an important factor impacting its preservation, with a significant degradation of RNA after only 8 month storage when concentrated at only 25 ng/μL (while purified RNA with a 250 ng/uL concentration could conserved for up to 4 years at -80°C. Nowadays, numerous protocols for storing RNA samples are proposed, and one of them (Fabre, Colotte, Luis, Tuffet, & Bonnet, 2014), describes the potential of RNA encapsulation to store extracted RNA in a long time range. More than only permitting long time conservation of RNA samples, the technology described allows to keep the samples at room temperature, and therefore diminishes resource allocation to storage under freezing conditions. Another possibility would be to quickly convert mRNA to cDNA (which has a higher stability) after sampling to allow long term conservation. Finally, the maximal storage time proposed in the previously mentioned studies were mostly estimated using human tissues or at least in single species. The filters stored in our study, however, are composed of the entire microphytoplanktonic community, and therefore of potentially numerous and diverse types of RNases, that may promote faster mRNAs degradation.

Chapter 5

Investigating species specific gene expression dynamics from metatranscriptomic datasets

5.1 Introduction

In this last chapter, we used the same metatranscriptomic datasets described in chapter 2 to 4, to analyze the expression profile dynamics of the most abundant species co-occuring with *A. minutum* during the blooms. The goal of this work was twofold. The first one was rather technical, and aimed at identifying the best approach to investigate the species specific expression dynamics of the most abundant species in a given community. This was addressed during the M2 internship of Sauvann Dubois that I mentored. The second one, was to investigate whether comparing *in situ* expression profiles among several species may gave us insights on how different species belonging to the same community adjusted their expression levels to cope with the fluctuations of their natural environment. Considering that global expression somehow reflects how a given species acclimates to its environment, one may compare global expression patterns between the species co-occurring in a community, to investigate the ecological niche differentiation among the members of a given community. Considering the ecological niches as a highly multidimensional hypervolume, where the dimensions are environmental conditions and resources (Hutchinson, 1957), we may try to use the global expression profiles to identify the dimensions where the ecological niches of two species overlap, from the dimensions where the niches are differentiated. Indeed, cellular functions for which expression patterns are similar among members of the community might be related to the overlapping dimensions of the ecological niches, while differential expression might point toward the partitioned dimensions of the niches. The main difficulty relying on being able to translate the expression patterns into ecologically relevant functions. Using this approach, we investigated the following questions: 1. Can we characterize species specific dynamic expression patterns for several members of the community ? 2. Can we relate the expression dynamics to environmental fluctuations ? 3. Can we compare the expression dynamics between species and relate them to ecological similarity/differentiation ?

A manuscript is currently under preparation.

Author contributions : Environmental samples were collected and filtered by numerous laboratory members. After filtration, RNA was extracted and library prepared by Mickael Le Gac and Julien Quéré (2013 and 2014 samples). I prepared 2015 samples for sequencing. Metabarcoding data were obtained by Pierre Ramond, who participated to general discussions and manuscript corrections. Preliminary investigations on metatranscriptomic data were performed by Sauvann Dubois, as part of her M2 internship that I mentored. Mickael Le Gac led in-depth analysis of the data and prepared the manuscript. I participated to results analysis, general discussions and manuscript correction.

5.2 Material and Methods

5.2.1 Metatranscriptome datasets and reads preprocessing

During three consecutive summers (2013, 2014 and 2015), the microphytoplanktonic community was sampled at a single site (48.350543, -4.292507) in the bay of Brest. Water was sampled twice a week, when *Alexandrium minutum* concentration was superior to 10 000 cells.L^{-1} (and until densities were inferior to this threshold in two consecutive samples). Several environmental conditions were monitored for each sampling, including salinity, sea surface temperature (both measured *in situ*), concentration of ammonia, nitrate, nitrite, phosphate and silicate (determined in the laboratory using a Seal Analytical AA3 HR automatic analyzer following the method of Aminot and Kerouel (2007)). Also, the number of days since the beginning of the bloom period (> 10 000 cells.L^{-1}), the time elapsed between sunrise and sampling, the tidal coefficient (SHOM), the flow of the adjacent Mignonne river (HYDRO), and the irradiance at sea surface (MeteoFrance) at the time of the sampling were also recorded (supplementary table B.1).

As quickly as possible after the sampling, water (4.8 ± 1.3 (mean ± SD) liters); supplementary table B.1) were filtered (20 μm) using a peristaltic pump, filters were frozen into liquid nitrogen and stored at -80°C until RNA extraction (these steps took around 2 hours). To prevent RNA degradation during filter thawing, RNA Later Ice® was added before thawing for samples collected in 2013. In 2014 and 2015, samples were directly frozen with RNA Later®. RNA was extracted following the Quiagen Rneasy Plus Mini Kit® recommendations, after ultra-sonication of the samples on ice (Bioblock Scientific, Vibra-cell 75115) for 10 seconds at 20% intensity. Library were prepared with the Illumina Truseq mRNA V2 kit®, and samples were sequenced at Get-PlaGe France Genomics sequencing platform (Toulouse, France) on Illumina HiSeq 2000/2500 2*100 pb (2013 and 2014; Multiplex 10) and on HiSeq 3000 2*150 pb for 2015 samples (Multiplex 24). Finally, 41 libraries were sequenced (10 in 2013, 19 in 2014 and 12 in 2015, see details in supplementary table B.1).

Reads from sequencing were first characterized with FastQC, and Trimmomatic (V. 0.33, Bolger et al. (2014)) was used to trim ambiguous, low quality reads and sequencing adapters with parameters ILLUMINACLIP : Adapt.fasta : 2 :30 :10 LEADING : 3 TRAILING : 3 MAXINFO : 135 : 0,8 MINLEN : 80.

5.2.2 Metabarcoding sampling and bioinformatic procedures for OTU construction

At the same time as the metatranscriptomic sampling, a seawater differential filtration approach was used to separate the communities of micro- (> 20 μm), nano- (20-3 μm) and pico- (3-0.2 μm) plankton. Carbonate membrane filters of 47 mm in diameter (Main Manufacturing, Michigan, USA) were used for each pore sizes. Particles of the two first size fractions were separated by consecutive filtration with a peristaltic water pump and swinnex filter supports, due to filter clogging the volume filtered ranged in between 1.5 and 5L. The residual filtrate was used for separate filtration onto the 0.2 μm filters. Only samples of micro- and nano- plankton were collected in duplicates, resulting into 242 water filters, in total for the three years. After filtration, filters were flash-frozen in liquid nitrogen and stored at -80°C until DNA extraction.

A metabarcoding approach was adopted to characterize the genetic diversity of the protistan community associated with the blooms of *A. minutum*. The hypervariable V4 domain of the 18S rDNA region was chosen as a barcode for its conservative character within the eukaryotic microbial community and its relatively high length (230-520bp; Nickrent and Sargent (1991)) which allows a good genetic distinction of marine protists (Behnke et al., 2011; Stoeck et al., 2010). Genomic DNA, issued from the cells collected on water filters, was isolated following the protocol of DNA extraction kit Nucleospin Plant II (Macherey-Nagel, Hoerdt, France). In parallel, some blank extractions (Millipore filtered water) were carried out to check and validate the extraction procedure. DNA quality (proteins/DNA absorbance: A260/A280) and concentration of purified products were respectively measured using a BioTek FLX 80 spectrofluorophotometer and a Quant-iT PicoGreen dsDNA quantification kit (Invitrogen, Cralsbad, CA, USA) following the manufacturer's instructions. Final DNA concentration of all extracts was normalized to 5-10 ng/μL. PCR was then ran with V4 markers assembled with the GeT-PlaGe adapters of the Genotoul sequencing platform (Forward : V4f-PlaGe: 5′CTT-TCC-CTA-CAC-GAC-GCT-CTT-CCG-ATC-TCC-AGC-A(C/G)C-(C/T)GC-GGT-AAT-TCC′3, Reverse: V4f-PlaGe 5′GGA-GTT-CAG-ACG-TGT-GCT-CTT-CCG-ATC-TAC-TTT-CGT-TCT-TGA-T(C/T)(A/G)-A′3). The process of PCR amplification was carried out three times for each DNA extract (representing a unique filter). The amplification protocol consisted of a denaturation step at 98°C for 30s, followed by two set of cycles 1) 12 x [98°C (10s), 53°C (30s), 74°C (30s)] and 2) 18 x [98°C (10s), 48°C (30s), 74°C (30s)]. The cycles were followed by a final elongation at 72°C for 10 min. Amplification results were verified by gel electrophoresis, triplicate reactions were pooled and purified using NucleoSpin Gel and PCR Clean-up (Macherey-Nagel, Hoerdt, France). Purified products were diluted to obtain equimolar concentrations before library construction at Genotoul for Illumina MISeq (2x250) sequencing. A single library was assembled; sequencing results are available on the sextant catalog

Bioinformatics were carried out on a larger sequencing dataset comprising (7 libraries, see (Ramond et al., 2018 *Submitted*)) to increase the number of sequences which allows a refined OTU construction and error detection. The cleaning steps and the rest of bioinformatics are the same as in (Ramond et al., 2018 *Submitted*). After cleaning steps, sequences were annotated taxonomically with PR2 (Guillou et al., 2013) and clustered into OTUs with SWARM2 (Mahé, Rognes, Quince, de Vargas, & Dunthorn, 2015). Each OTUs was then given the taxonomic reference and the nucleotide sequence of its most abundant metabarcode. The dataset used in this study (protistan communities from the Daoulas river in 2013, 2014, 2015) contains 38,227 OTUs, annotated to 1,167 distinct taxonomic references and cumulating into 7.5 10^6 reads.

5.2.3 Reference transcriptomes and alignments

All the MMETSP reference transcriptomes (Keeling et al., 2014) were downloaded from Cyverse. All reference corresponding to unknown species were discarded. The reference corresponding to *Alexandrium minutum* was replaced by the one obtained by Le Gac et al. (2016), and a custom reference corresponding to *Scrippsiella trochoidea* was added. Within each transcriptome, for each transcript, only the longest isoform was considered. When several transcriptomes per species were available (several strains or culture conditions), they were merged into a single reference, and CD-HIT-EST (Fu, Niu, Zhu, Wu, & Li, 2012) was used to remove homologous sequences within each reference. Briefly, CD-HIT-EST compares and clusters similar sequences based on a similarity threshold. It is an incremental algorithm that takes the longest sequence of a set as the first cluster representative sequence, and compares each other sequences from the set to classify them as redundant, or representative sequence of a new cluster. A total of 313 species specific reference transcriptomes, representing 213 unique genus, were considered and concatenated in several ways to build meta-references. In the first one (hereafter MetaRef1), all (313) the species specific reference transcriptomes were concatenated. Reads from the meta-transcriptomic datasets were aligned to the meta-reference using the BWA-MEM aligner (H. Li & Durbin, 2009). SAMTOOLS (H. Li et al., 2009) was used to discard reads displaying low quality alignments (MapQ<10) as well as the ones mapping to several transcripts (FLAGS<1,000).

For each sample ***i*** (each meta-transcriptomic dataset), the relative abundance of the transcripts (RAT) for species ***j*** in sample ***i*** was calculated as

$$RAT_{\mathrm{ij}} = \frac{N_{\mathrm{ij}}}{L_{\mathrm{j}}} \;/ \sum_{j} \frac{N_{\mathrm{ij}}}{L_{\mathrm{j}}}$$

where $\boldsymbol{N_{ij}}$ corresponds to the number of reads from sample ***i*** mapping to species ***j*** reference transcriptome, $\boldsymbol{L_j}$ to the total length, in base, of the reference transcriptome of species ***j***.

The second meta-reference (MetaRef2) was composed of 219 species specific transcriptomes selected as follows. First, when a genus was represented by a single species in the MMETSP reference transcriptomes, the species was added to MetaRef2

(162 species). Second when several species belonged to the same genus, the Spearman rank correlation of the RAT obtained using MetaRef1 for each pair of species across the time points were computed. When the Spearman rank correlations were > 0.75 for all pairwise comparisons within a genus, only the species with the highest maximum RAT was added to MetaRef2 (40 species). Within the genus displaying at least one Spearman rank correlation > 0.75, were added to MetaRef2: 1. The species with the maximum RAT for group of species displaying at least one Spearman rank correlation > 0.75 among them, and 2. All the species not displaying a single Spearman rank correlations > 0.75 (25 species from 11 genus). Finally, all the reference transcriptomes displaying a total length < 2.10^6 bases were discarded (8 species). Alignment and filtering were performed as described above. After these alignments, species displaying a maximum RAT > 0.03 (hereafter named species of interest) were analyzed for species specific gene expression dynamics. Similarly, larger phylum displaying a maximum RAT > 0.03 were analyzed for phylum specific gene expression dynamics. To estimate how the number of species in the meta-reference influenced the alignments of the reads on the species of interest references, eight other meta-references were built using the references of species of interest and 0, 25, 50, 75, 100, 125, 150, and 175 other species specific references randomly chosen from the species composing MetaRef2. Alignment and filtering were performed as described above.

5.2.4 Determining gene expression dynamics *in situ*

The matrix of reads aligning to MetaRef2 was used to investigate gene expression dynamics. After preliminary analyses, one sample (15/07/13) was excluded from the analyses due to poor quality. Transcripts covered by less than 10 reads on average across samples were discarded from the analyses. The dataset was normalized using DESEQ2 Variance Stabilizing Transformation (Love et al., 2014). After preliminary PCA and clustering analyses, year specific batch effects were considered as negligible and were not corrected for. Transcripts with similar expression dynamics were grouped into modules of co-expressed transcripts using the WGCNA (V 1.51, Langfelder and Horvath (2008) R package. A soft-thresholding power of 6, a maximum block size of 35,000 transcripts, and bicor correlation were used as parameters. Module identification was performed using dynamic tree cut with minimum cluster size of 1,000 transcripts, and modules displaying a Pearson correlation higher than 0.75 were merged. For each expression module, WGCNA was used to extract eigengenes, which can be defined as the first principal component of the expression matrix of the detected modules.

5.2.5 Annotation and Gene Ontology Enrichment analyses

Sequence similarity of the transcripts with genes of identified function in the UniProt databank was investigated using BLASTX with E-Value < 10^{-3}. The transcripts were classified in various Gene Ontology categories (GO) based on this result. For each of the species of interest displaying more than 450 transcripts in a given expression module, GO enrichment analyses were performed for GO involved in Biological Processes and displaying at least 20 transcripts. Two sided Fisher exact test, with a false discovery rate set to 0.05 for multiple test type 1 error control were used and GO displaying an odd-ratio > 3 and a q-value < 0.01 were considered as significantly enriched. The hierarchical Gene Ontology classification system leads to redundancies in the GO enrichment analysis. To take such redundancy into account, a distance between GO categories was calculated as

$$1 - \frac{GO_{\mathrm{i}} \cap GO_{\mathrm{j}}}{Min(GO_{\mathrm{i}}, GO_{\mathrm{j}})}$$

where $\boldsymbol{GO_i}$ is the size of the GO category $\boldsymbol{i}$.

5.3 Results

During three consecutive years, an *in situ* survey of *A. minutum* blooms was performed in the bay of Daoulas (France), resulting in 41 meta-transcriptomic datasets (a total of $1.6.10^9$ reads). The aim of the present publication was to assess at what point such datasets may be used to follow species specific gene expression dynamics *in situ*.

5.3.1 Defining a meta-reference transcriptome

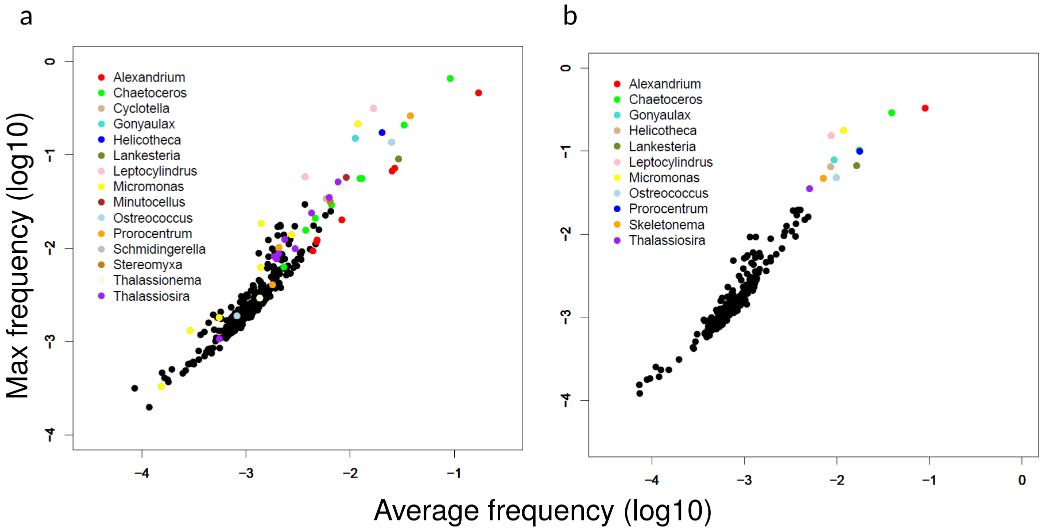


FIGURE 5.1: *Alignment of environmental reads on **a.** MetaRef1 and **b.** Metaref2. Each point corresponds to a species. Positions of the points are given based on the maximum proportion (y axis) and average proportion (x-axis), across time points, of reads aligning to the species reference transcriptome within the metareference. Colors indicate **a.** the members of genus for which at least one species displays a maximum proportion > 0.03, **b.** the twelve species of interest considered for the subsequent analyses.*

The reads resulting from the sequencing of the meta-transcriptomic datasets were aligned to several meta-references built using reference transcriptomes developed during the MMETSP sequencing project. In a first step, the meta-reference was composed of all the species specific reference transcriptomes from the MMETSP project (MetaRef1). The output of this first alignment is summarized figure 5.1a. A strong signal of this first alignment is that among the species attracting numerous environmental reads, there are several members of the same genus. This is for instance the case of the dinoflagellate genus *Alexandrium*, but also of the diatom genus *Chaetoceros* and *Thalassiosira*. Two distinct possibilities may explain this pattern. It could be observed because several species of the same genus tend to co-occur in the sampled communities. In this case, the observed pattern is ecologically relevant and should be further analyzed. Or it could be an artefact resulting from the misalignment of part of the environmental reads belonging to a given species on the reference transcriptome of closely related species. To distinguish between these two possibilities, the dynamics of the RAT across samples were compared for the species belonging to the same genus using Spearman rank correlation coefficients. This comparison was performed for the 51 genus represented by more than one species in MMETSP (supplementary table D.1). For 40 genus, the dynamics of the RAT across samples were extremely similar (Spearman rank correlation coefficients>0.75) and only the species displaying the highest maximum RAT was considered. For 11 genus, several species displayed different dynamics (Spearman rank correlation coefficients<0.75) and were considered for the next steps of the analyses (supplementary table D.1; An illustration is given in supplementary figure D.1 for the genus *Alexandrium* where a single species was selected and the genus *Chaetoceros* where three species were selected). Based on these results, the environmental reads were aligned to a second meta-reference (MetaRef2) composed of 219 species specific transcriptomes. This second alignment is summarized in figure 5.1b, and 12 species displayed a maximum RAT higher than 0.03. Out of these 12 species, there were six large diatoms (*C. curvicetus*, *C.* cf. *neogracile*, *Helicotheca tamesis*, *Leptocylindrus aporus*, *Skeletonema dorhnii*, and *T. punctigera*), three dinoflagellates (*A. minutum*, *Gonyaulax spinifera*, *Prorocentrum micans*), two small (< 2 μm) but extremely abundant chlorophytes (*Micromonas pusilla*, and *Ostreococcus lucimarinus*) and one apicomplexan parasite of free-swimming tunicates (*Lankesteria abbotti*). In the previous part, the impact of the presence of several closely related species in the meta-reference on the alignment of environmental reads was investigated. However, relatedness among species is not the only source of potential misalignment. Other potential sources of misalignment may not be directly linked to species relatedness but rather to potential horizontal transfers or to differences in the speed of molecular evolution across genes and taxa. As a global way to assess the robustness of the alignments on the reference transcriptomes of the 12 species identified above, the environmental reads were aligned to eight meta-references composed of the 12 species of interest and 0, 25, 50, 75, 100, 125, 150, and 175 species randomly chosen from the species composing MetaRef2. The results clearly illustrate the importance of including numerous species in meta-references when dealing with environmental reads (figure 5.2). Interestingly, the magnitude of the response was extremely different across species, the two extremes being *G. spinifera* and *A. minutum*. The former species reference attracting six times more reads when the metareference is composed of 12 species than when it is composed of 219 species (MetaRef2) while for the latter the difference is below 0.2. We note that this huge difference can neither be explained by differences in the length of the reference transcriptomes of the species of interest (a short reference might reflect a poor quality, or display an overrepresentation of highly expressed and often slowly evolving housekeeping genes) nor by a difference between the number of reads aligning to a given species (a rare species would only be represented by reads corresponding to highly expressed and often slowly evolving housekeeping genes).

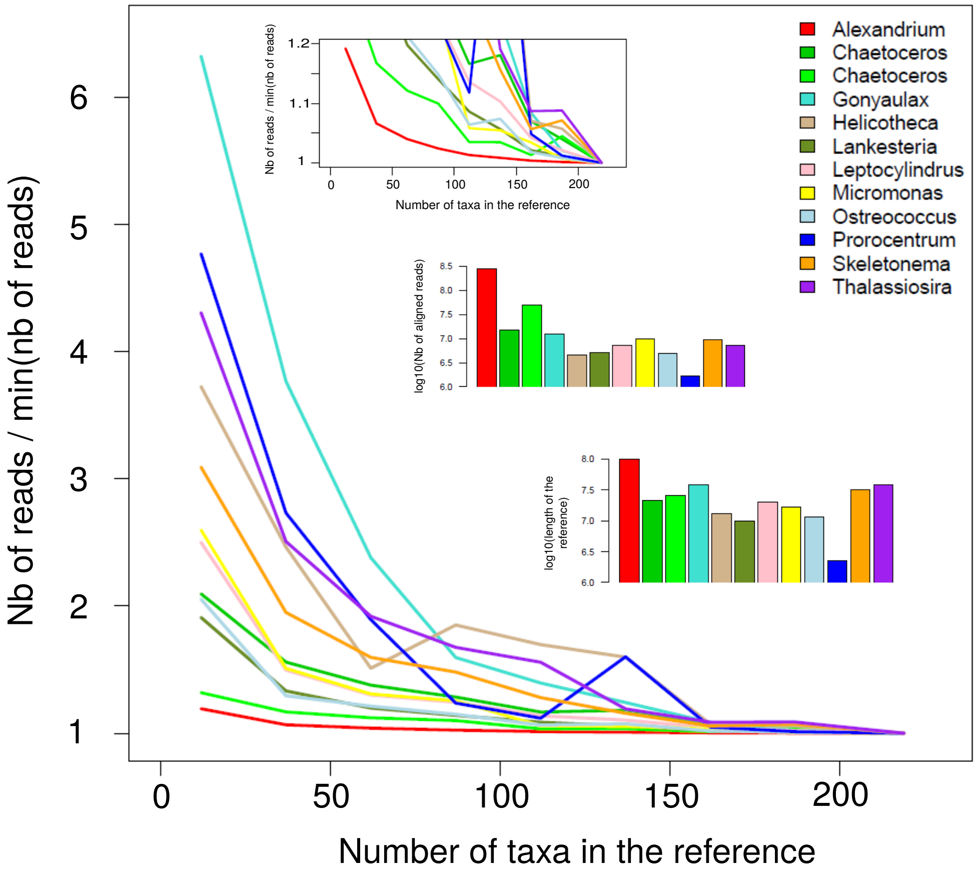


FIGURE 5.2: *The number of species in the meta-reference transcriptome strongly affects the specificity of the alignment on the reference transcriptome of the 12 species of interest. For each species of interest, is indicated the number of reads aligning to the reference transcriptome divided by the number of reads aligning to MetaRef2 (219 species) as a function of the number of species in the meta-reference. The three insets correspond from top to bottom to:* ***1.*** *The magnification of the lower part of the main graph,* ***2.*** *For each species, the total number of aligned reads when aligning against MetaRef2,* ***3.*** *For each species, the length of the references in base.*

5.3.2 RAT for the 12 species of interest

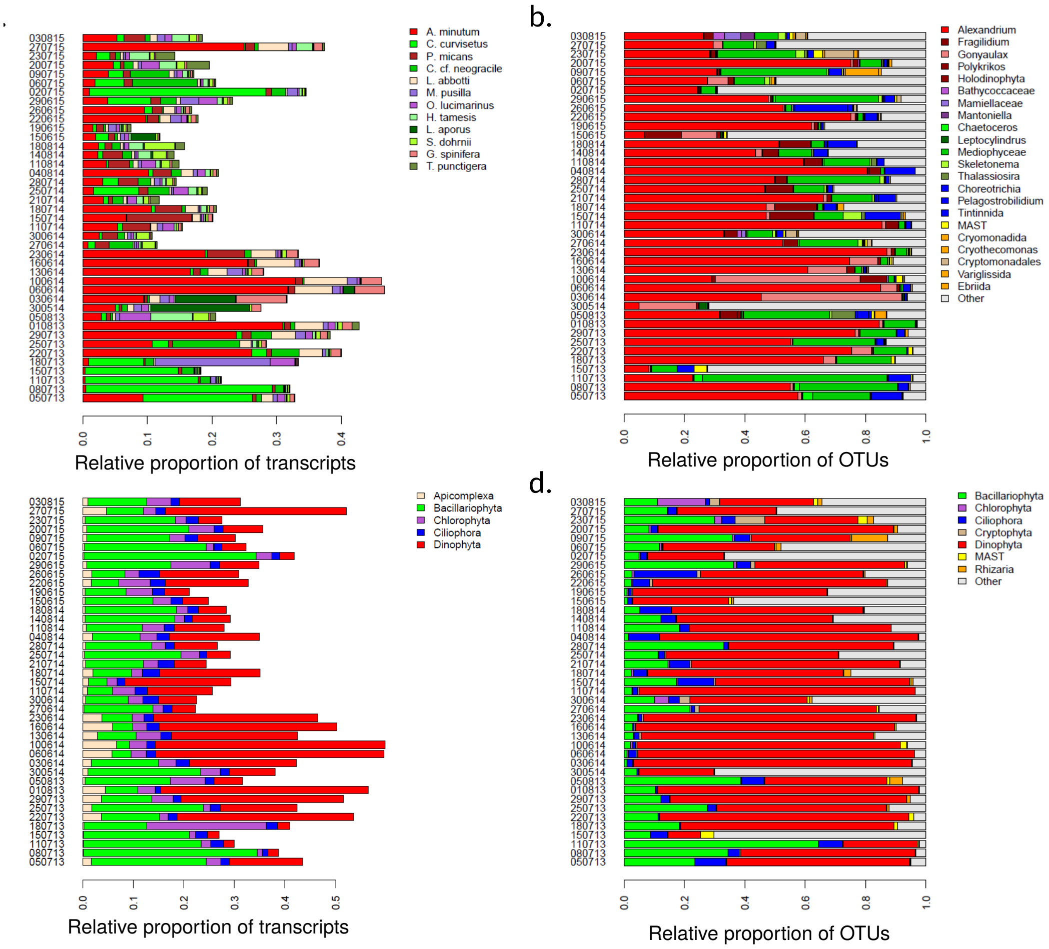


FIGURE 5.3: *Relative abundances of the members of the sampled communities for each sampled time point. Panels **a.** and **c.** indicate the relative abundance of transcripts (RAT, see methods) for each **a.** species, and **b.** phylum. Panels **b.** and **d.** indicate the relative abundance of OTU based on metabarcoding data for each **b.** genera, and **c.** phylum.*

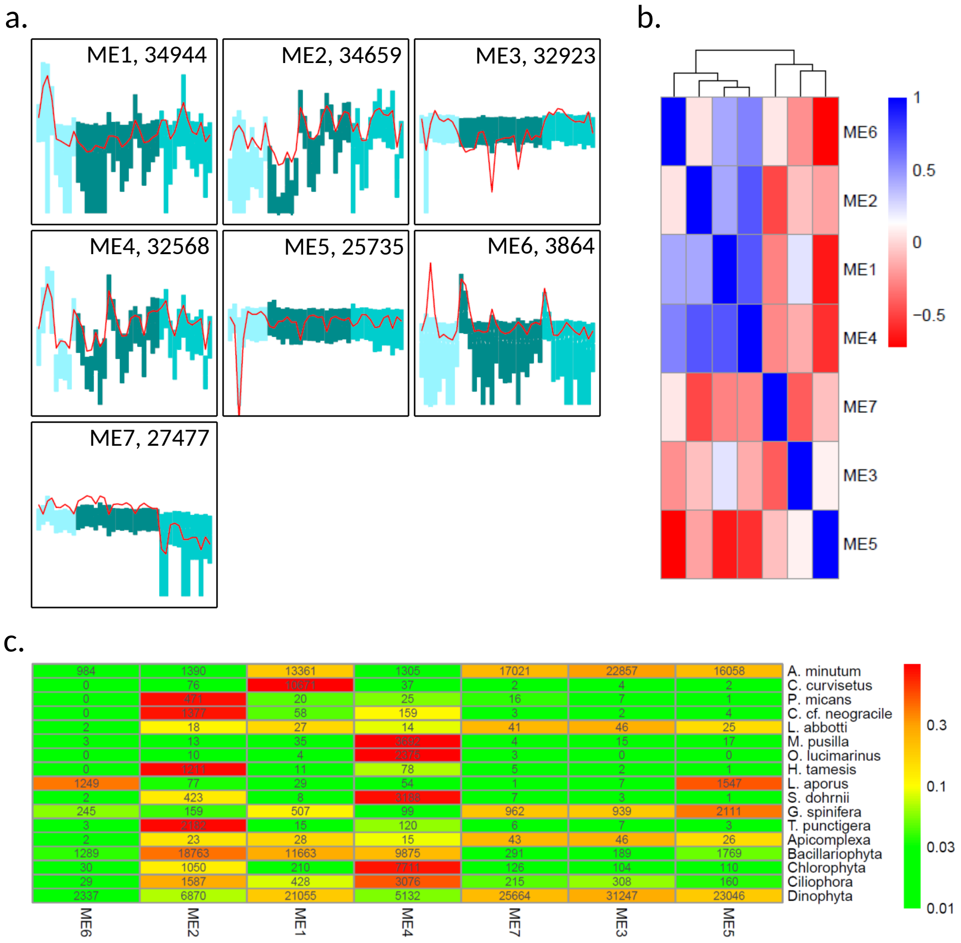
After aligning the environmental reads to MetaRef2, the total number of reads aligning to the 12 species ranged from 1.7.10^6 to 2.9.10^8, representing between 1% and 17.4% of the analyzed reads for *P. micans* and *A. minutum*, respectively (supplementary table D.2). Depending on the dates considered between 7.4% and 50.8% of the sequenced reads aligned to the 12 species (supplementary table D.3). After correcting for differences in species specific reference transcriptome lengths and number of analyzed reads across time points, we analyzed the RAT for the 12 species of interest across the 41 time points (figure 5.3a). The first thing to note is the highly dynamic fluctuations across time points. Without surprise, the transcripts of *A. minutum*, the dinoflagellate species blooming during the survey, were especially abundant with 31 sampling times displaying a RAT>0.02 and a maximum RAT=0.33. Then three species often displayed a RAT > 0.02 and often reached RAT > 0.1. This is the case for the diatoms *C. curvicetus* (14 samples with a RAT > 0.02, max RAT = 0.29) and *C.* cf. *neogracile* (9, 0.1), as well as for the dinoflagellate *P. micans* (11, 0.1). Two species, the diatom *L. aporus* (3, 0.15) and the chlorophyte *M. pusilla* (3, 0.18) very punctually displayed high RAT. Three others, the diatom *H. tamesis* (7, 0.07), the chlorophyte *O. lucimarinus* (6, 0.05), and the apicomplexan *L. abbotti* (10, 0.07) often displayed RAT > 0.02, but never reached high RAT. Finally the last three species considered, the diatoms *S. dorhnii* (3, 0.05), and *T. punctigera* (2, 0.04), and the dinoflagellate *G. spinifera* (4, 0.08) displayed a RAT > 0.02 in only a few sampled and never reached high RAT. Performing the RAT analysis at the scale of the major phylum clearly indicated that the aligned reads mainly belonged to diatoms (*Bacillariophyta*) and Dinoflagellates (*Dinophyta*) in variable proportions and to a lesser extend to ciliates (*Ciliophora*), chlorophytes (*Chlorophyta*), and *Apicomplexa* (figure 5.3c). The 12 species of interest explained a very large proportion of the RAT observed at the phylum level, suggesting that the expression dynamics at the phylum level is mainly due to a few abundant species rather than to a multitude of rare species.

5.3.3 Comparison with metabarcoding

In parallel to the meta-transcriptomic approach, the composition of the phytoplanctonic community was characterized using a metabarcoding approach (figure 5.3b and 5.3d). The comparison of the two datasets is interesting in several ways. First and reassuringly, the same genera are often identified with the two approaches. This is indeed the case, whithin the *Dinophyta* for the genera *Alexandrium*, but also *Gonyaulax*, and within the *Bacillariophyta* for the genera *Chaetoceros*, *Leptocylindrus*, *Skeletonema*, *Thalassiosira*. Some of the results that may, at first sight, appear as not entirely compatible among the two datasets probably simply reflect a difference in taxonomic resolution. This is for instance the case for the holodinophyta OTU that probably corresponds to *P. micans* in the meta-transcriptomic dataset, but also for the mediophyceae OTU that may in part corresponds to the two *Chaetoceros* species in the meta-transcriptomic. Second, a major difference between the two datasets is the amount of relative abundance of transcripts and of OTU captured by the two datasets. While the 12 species of interest correspond on average to a RAT of 0.24 across samples (considering the raw data instead of the normalized proportion, the average proportion of aligned reads is 0.43), the average relative proportion of OTU that could be attributed to phytoplankton is 0.78 (the category Other corresponding to a mix of metazoa, land plant, unresolved eukaryotes and unassigned reads). This difference is mainly explained by the same species representing a very different relative abundance of transcripts and OTU. This is extremely clear for *Alexandrium*, representing an average RAT of 0.09 (average proportion of aligned reads of 0.17) while displaying an average relative proportion of OTU of 0.50. Third, another important difference is highlighted when considering the phylum level. While in term of expression, the average of the ratio between RAT of dinoflagellates and diatoms is 1.10 (2.4 considering raw reads), when considering the metabarcoding data, the average of the ratio rises to 4.2, indicating that in the same community the amount of rRNA is four times higher for dinoflagellates compared to diatoms, while the amount of mRNA is about the same.

5.3.4 Gene expression dynamics *in situ*

A total of seven modules regrouping between 3 864 and 34 944 transcripts displaying similar expression dynamics across samples were identified (figure 5.4a). The expression dynamics of these modules across samples range from relatively stable (Modules ME3, ME5, ME7) to highly dynamics (ME2, ME4). The eigengenes of four modules have a slightly correlated expression dynamics (ME1, ME2, ME4, ME6) while the three others appear completely independent (ME3, ME5, ME7) (figure 5.4b). The distribution of the transcripts of the 12 species of interest in the seven modules is highly heterogeneous, and is of course strongly linked to the RAT dynamics described above (figure 5.4c). Five of the modules are to great extend composed of transcripts belonging to the 12 species of interest (> 64% of the transcripts of ME1, ME3, ME5, ME6, ME7); while in the two remaining modules their proportion is considerably lower (21% and 34% of ME2, and ME4). Eight of the species of interest have more than 85% of their transcripts in a single module. This is the case of *C.*



ME6	ME2	ME1	ME4	ME7	ME3	ME5	
984	1390	13361	1305	17021	22857	16058	A. minutum
0	76	10671	37	2	4	2	C. curvisetus
0	471	20	25	16	7	1	P. micans
0	1377	58	159	3	2	4	C. cf. neogracile
2	18	27	14	41	46	25	L. abbotti
3	13	35	3692	4	15	17	M. pusilla
0	10	4	2375	3	0	0	O. lucimarinus
0	1211	11	78	5	2	1	H. tamesis
1249	77	29	54	1	7	1547	L. aporus
2	423	8	3188	7	3	1	S. dohrnii
245	159	507	99	962	939	2111	G. spinifera
3	2182	15	120	6	7	3	T. punctigera
2	23	28	15	43	46	26	Apicomplexa
1289	18763	11663	9875	291	189	1769	Bacillariophyta
30	1050	210	7711	126	104	110	Chlorophyta
29	1587	428	3076	215	308	160	Ciliophora
2337	6870	21055	5132	25664	31247	23046	Dinophyta

FIGURE 5.4: *Modules of transcripts displaying similar expression dynamics across samples.* ***a.*** *the 7 identified expression modules. The 40 samples are on the x-axis, the shade of the bars represent the three years (2013, 2014, 2015). The y-axis indicates the level of expression (vst transformed raw-counts, range (-7,12)), the red lines indicate the expression dynamics of the eigengenes, the bars range from the first to the third quartile of the expression levels. The number of transcripts belonging to each module is indicated,* ***b.*** *Heatmap representing the Pearson correlation coefficient among the 7 expression modules.* ***c.*** *Heatmap indicating how the transcripts of the 12 species of interest as well as of the main phylum are distributed in the 7 expression modules. The color scale is based on the proportion of transcripts from each species/phylum assigned to each module. Numbers indicate the number of transcripts in each module.*

curvicetus (99% of the transcripts in ME1), *P. micans* (87% of the transcripts in ME2), *C.* cf *neogracile* (86% of the transcripts in ME2), *M. pusilla* (98% of the transcripts in ME4), *O. lucimarinus* (99% of the transcripts in ME4), *H. tamesis* (93% of the transcripts in ME2), *S. dohrnii* (88% of the transcripts in ME4), and *T. punctigera* (93% of the transcripts in ME2). For three species, a non-negligible proportion of transcripts are found in more than one module. This is the case of *A. minutum* displaying more than 900 transcripts in the seven modules and more than 20% of its transcripts in four modules (ME1, ME3, ME5, ME7). To a lesser extent this is also the case of *L. aporus displaying* 42% and 52% of its transcripts in ME5, and ME6, but also of *G. spinifera* with more than 18% of its transcripts in four modules (ME1, ME3, ME5, ME7). The last species, *L. abbotti* displays very few transcripts in all the modules and is not further analyzed. Focusing on the transcripts at the phylum level (figure 5.4c), we note that three modules are mainly composed of transcripts from dinoflagellates and diatoms (ME1, ME2, ME6) in mixed proportions. One module is composed of transcripts from diatoms and dinoflagellates but also from chlorophytes and ciliates (ME4). The last three modules are mostly composed of transcripts from dinoflagellates (ME5, ME3, ME7; >89% of the transcripts). As stated above, two modules (ME2, ME4) are composed of transcripts not belonging to the 12 species of interest. In ME2, more than 40% of the transcripts aligned to other diatom reference transcriptomes. In ME4, the transcripts not corresponding to the species of interest aligned to other diatoms (20%), dinoflagellates (11%), but also ciliates (10%).

5.3.5 Relationships between gene expression dynamics and environmental factors

A set of 11 abiotic factors were used to characterize the abiotic environment encountered by the sampled phytoplanktonic communities (see methods). The expression dynamics of the seven modules was correlated with this abiotic factor as well as with the RAT (taken as an indicator of the relative abundance of the species within the community). As already suggested by the analyses presented above, a large portion of the expression variance is linked to high relative abundance of diatoms (ME1, ME2, ME6, ME4) and dinoflagellates (ME3, ME5, ME7; first axis, figure 5.5). The expression patterns were also correlated with abiotic factors and especially to salinity variation (ME3) and to the river outflow (ME7; second axis, figure 5.5). Axis 3 mainly separates expression modules ME3 (ME1 and ME6 in a less extent) from ME2 (ME5 and ME4 in a less extent), respectively displaying low and high expression in low phosphate, cold environments, mostly at the end of the blooming period. ME6 displays high expression levels in waters greatly influenced by river flow (nitrogen rich and low temperature), while transcripts gathered into module ME7 have low expressions in such environments.

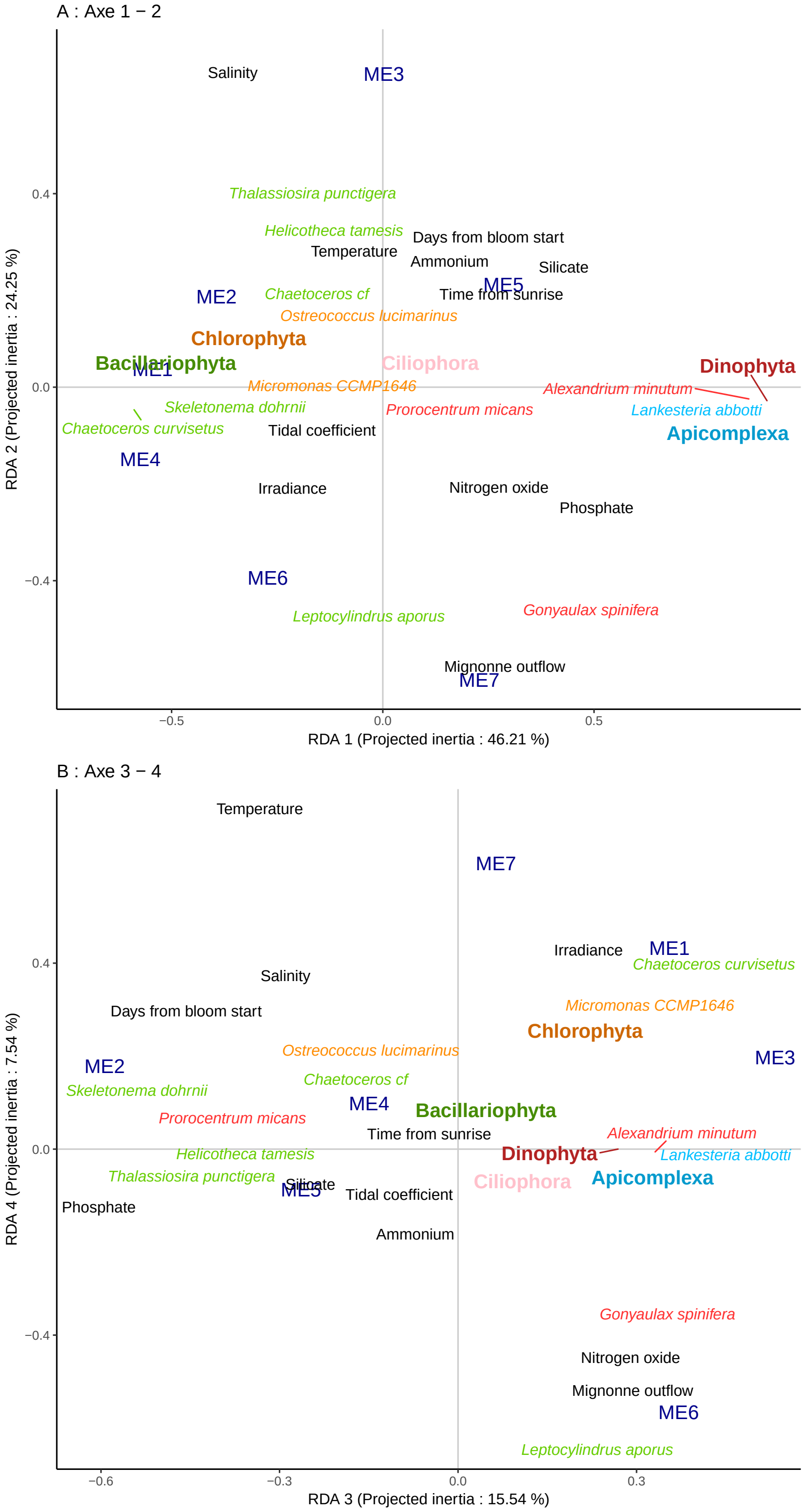


FIGURE 5.5: *Representation of the four first axis of the RDA led between the seven metatranscriptomic gene expression modules (represented in dark blue) and both abiotic variables (in black) and the relative abundance of transcripts (RAT; colored).*

5.3.6 Functions of the transcripts displaying dynamic expression

The functions of the transcripts displaying dynamic expression were investigated as species specific expression modules (SSEMs, hereafter). To do so, a GO enrichment analysis was performed for each species in each module (see Material and Methods for details; figure 5.6). A total of 76 biological process GO were identified as displaying a species specific transcript enrichment in at least one expression module. Several types of questions maybe asked following this analysis. First, we ask if for different species, the transcripts displaying similar expression dynamics (i.e. belonging to the same expression module) are related to the same biological processes. Second, if for a given species, the transcripts displaying different expression dynamics (*i.e.* belonging to different expression modules) are related to different biological processes. Third, if for species belonging to the same phylum, expression dynamics tend to be related to the same biological process. There are two broad categories of SSEMs (see figure 5.6 clustering by columns). The first one regroups diatom, chlorophytes and dinoflagellate SSEMs, and the second diatom and dinoflagellates SSEMs. SSEMs from the same species belong to a single broad category, but SSEMs from the different expression modules are found in the two broad categories. The first category tends to display an overrepresentation of transcripts involved in photosynthesis as well as related to central metabolism (TCA cycle, glucose metabolism, glycolysis, gluconeogenesis, pentose phosphate shunt). Within this first broad category, four SSEMs display very similar pattern of biological processes over-representation (*C.* cf *neogracile* ME2, *H. tamesis* ME2, *T. punctigera* ME2, *L. aporus* ME5). These modules tend to regroup numerous transcripts involved in photosynthesis. Interestingly, all four correspond to diatoms and three belongs to the same expression module (ME2). There are three other SSEMs corresponding to diatoms. Two of them display a slightly lower overrepresentation of transcripts involved in photosynthesis and central metabolic processes. This is the case of the transcripts of *S. dohrnii* belonging to ME4, but also of the transcripts of *L. aporus* belonging to ME6. Interestingly, *L. aporus*, display numerous transcripts in two modules (ME5, and ME6). In addition to the moderate differences noted above, the two main differences between these *L. aporus* specific expression modules are linked to glucose metabolism, and digestion.

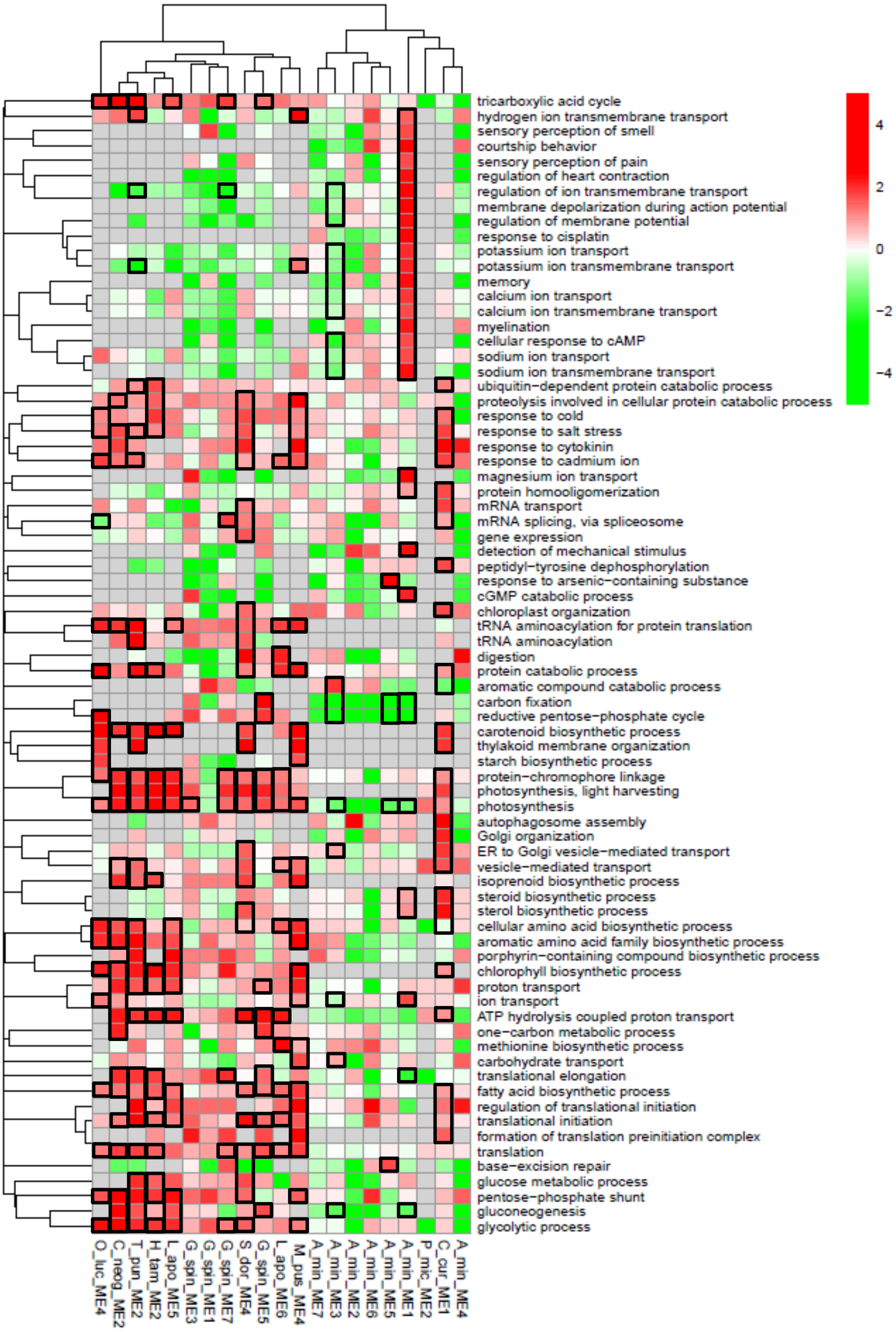


FIGURE 5.6: *Over-representation of Biological Process GO terms in each species in each expression module. Heatmap colors indicate the odd-ratio for each GO term (GO displaying less than 20 transcripts were not analysed and are in grey). GO terms were clustered based on overlap between GO (see Material and Methods for details). Black squares indicate q-values < 0.1).*

The last diatom C regroups *C. curvicetus* transcripts belonging to module ME1. It belongs to the second broad category of SSEMs and thus displays a very different GO category overrepresentation pattern. This mainly translates into an absence of overrepresentation of central metabolism related biological processes (TCA cycle, glucose metabolism, glycolysis, gluconeogenesis, pentose phosphate shunt), and an overrepresentation of transcripts linked to sterol/steroid metabolism, to autophagosome assembly and Golgi organisation. The two SSEMs corresponding to chlorophytes (*M. pusilla* and *O. lucimarinus*, both from ME4) display similar biological processes overrepresentation patterns with some differences related to carbohydrate and vesicle-mediated transport. We also note that these two SSEMs are the only one displaying a strong overrepresentation of transcripts involved in starch biosynthetic process. Turning toward dinoflagellates, two species, *A. minutum* and *G. spinifera*, are mainly considered (a third species, *P. micans* is also analyzed, but its reference transcriptome is extremely short and probably partial, making the functional analysis extremely limited). The SSEMs of these two species are split into the two broad categories of SSEMs and thus display extremely different overrepresentation patterns mainly concerning functions related to photosynthesis and central metabolism (figure 5.6). Besides these major differences, it is interesting to focus on a few groups of biological processes displaying marked overrepresentation differences across species and or modules. The first notable group of biological processes is linked to ion, but also calcium, potassium, and sodium transport. It displayed a dynamic expression pattern in a single species (*A. minutum*), with modules either displaying strong over- or under-representation (ME1 and ME3, respectively). In the other species, this same group of biological processes tends to be globally underrepresented, with a few exceptions related to a few specific kind of transport rather than to the whole category: overrepresentation of ion transport in the two chlorophytes in ME4, and in *C.* cf. *neogracile* in ME2; overrepresentation of potassium ion transmembrane transport in *M. pusilla* ME2. Second, several functions related to primary production (photosynthesis, carbon fixation...) tend to be strongly overrepresented in SSEMs corresponding to diatoms and chlorophytes, but display a contrasting pattern in the two dinoflagellates. They are underrepresented in all *A. minutum* SSEMs, and display a dynamic pattern across modules in SSEMs corresponding to

G. spinifera (underrepresented in ME1, overrepresented in the other modules). Third, biological processes related to central metabolism (glucose metabolism, glycolysis, gluconeogenesis, pentose phosphate shunt) also display a notable pattern. For example, transcripts related to glucose metabolic process are strongly overrepresented in three diatoms SSEMs (*T. punctigera* ME2, *H. tamesis* ME2, and *S. dohrnii* ME4), but tend to be underrepresented in P. micans ME2, and tend to display a dynamic pattern across expression modules in *L. aporus* and *A. minutum*. Fourth, transcripts related to aromatic compound catabolic process are strongly underrepresented in *C. curvicetus* ME1, *S. dohrnii* ME4, and display a dynamic pattern across expression modules in *A. minutum*. Fifth transcripts related to sterol/steroid biosynthesis are overrepresented in *A. minutum* ME1, *C. curvicetus* ME1, *S. dohrnii* ME4.

5.4 Discussion and conclusion

During a three year *in situ* survey of the toxic dinoflagellate *A. minutum*, metatranscriptomic datasets were used to investigate the species specific *in situ* gene expression dynamics of the species dominating the phytoplanktonic community. The first question investigated was rather technical: At what point is it possible to follow species specific gene expression dynamics *in situ* for eukaryotic microbial species ? Thanks to the next generation sequencing revolution, it is now easy to generate community wide transcriptomic datasets, and several such studies have already been published, mainly concerning prokaryotic communities (Aylward et al., 2015; Berg et al., 2018; Frias-Lopez et al., 2008; Ottesen et al., 2011, 2013, 2014; A. K. Sharma et al., 2014; Shi, Tyson, Eppley, & DeLong, 2011; Stewart, Sharma, Bryant, Eppley, & DeLong, 2011). Similar approaches focusing on marine eukaryotic communities have developed more recently. One of the main obstacles had been the lack of reference genomes or transcriptomes for the vast majority of marine unicellular eukaryotes (Caron et al., 2017). Thanks to the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP, Keeling et al. (2014), that allowed the generation of reference transcriptomes for > 200 species, a major step has been taken. A few marine community wide transcriptome projects have started to get published (Alexander, Jenkins, et al., 2015; Alexander, Rouco, et al., 2015; Gong, Paerl, & Marchetti, 2018;

Hu et al., 2018; Marchetti et al., 2012). A major challenge to be able to infer species specific gene expression dynamics is to ensure the specificity of the alignment of the environmental reads on the reference transcriptomes. Strong homology between orthologous genes may interfere with the robustness of the alignment. To cope with such issue, we propose to align environmental reads to a meta-reference composed of a maximum of species specific transcriptomes, only keeping one species per genera when several species belonging to the same genera display extremely similar relative abundance dynamics, and discarding environmental reads aligning to several reference transcripts. The second question was also a technical one, but with strong biological implications. With around 200 species specific reference transcriptomes, are we able to follow species specific expression dynamics of entire microeukaryote communities ? As a matter of comparison, meta-barcoding approaches aiming at characterizing the composition of communities use reference databases corresponding to several thousands of OTU. Reassuringly, after comparing metatranscriptomic and meta-barcoding approaches, results were congruent regarding the species identified as dominating the sampled communities. However, there were striking differences regarding the proportion of the environmental sequences aligning to the references. While the proportion of meta-barcoding reads that could be assigned to the abundant OTU was often higher than 0.8, the proportion of metatranscriptomic reads assigned to the abundant species was often lower than 0.3. Making our alignment parameters less strict (and especially allowing for a single read to align to several transcripts) had little influence (a few percent) on this proportion (data not shown). Several kinds of explanations maybe possible, the first one involves the quality of some reference transcriptomes. For example, *P. micans* reference transcriptome is extremely short and incomplete, precluding global expression profiling or this species. Second, intraspecific diversity may also play a role. Reference transcriptomes often result from the sequencing of a single strain for a given species, precluding gene expression analysis at the pan-genome (the total set of genes found in a given species) scale. For example, about 60% of the baker yeast (*Saccharomyces cerevisiae*) genes belong to the core genome (the gene set shared by all strains of a given species; (Peter et al., 2018). For the haptophyte *Emiliania huxleyi*, the core genome corresponds to $\sim$ 80% of the genome (Read et al., 2013).

For the other marine micro-eucaryote this proportion is virtually unknown. Third, the number of cryptic species within the phytoplankton is probably extremely high. As an example, following the discovery of the toxicity in the diatom genus *Pseudo-nitzschia*, and the subsequent interest that this genus attracted from the researcher, the number of species rose from ∼15 in the 80's (Trainer et al., 2012) to close to 50 nowadays (Guiry, M.D and Guiry, G.M, 2014). If closely related but nevertheless distinct species are present in the sampled community and in the meta-reference, the expression profiling would be extremely partial, as it would only be possible for the ortholog genes (the genes represented in the genomes of the two closely related species). For example, in *Pseudo-nitzschia*, the number of ortholog transcripts identified in three species represented between 20% and 40% of each species reference transcriptome (Lema et al., 2018 *Submitted*). Finally, a last but potentially extremely important factor that could affect the expression profiling in natural communities is the presence of transcripts expressed *in situ* but barely if at all in the artificial experimental environments that were used to obtain the reference transcriptomes. It would, of course, be impossible to characterize the expression dynamics of such transcripts, and solving this problem is far from being straightforward. Indeed, if solving the first three issues presented above is "only" a matter of improving the quality of the sequencing or sequencing more strains or more species; characterizing transcripts only expressed *in situ* for organisms displaying a genome too big to be sequenced is highly problematic. Some of our results tend to show that this may be a non-negligible problem. Considering *A. minutum*, for which the reference transcriptome was obtained using several strains isolated locally, and often representing about 50% of the OTU, the RAT was at time very low (see for instance the second half of the season 2014, figure 5.1), pointing toward a partial characterization of *A. minutum* expression. Third, as discussed above, performing species specific gene expression dynamics within natural communities is not without causing problems, and analyzing expression dynamics at the phylum level may seem like an interesting alternative (Alexander, Jenkins, et al., 2015; Gong et al., 2018; Hu et al., 2018; Marchetti et al., 2012). However, such approaches make important implicit assumptions which may not hold for all communities. The first implicit assumption is that the various species belonging to the same phylum are more or less ecologically redundant (Mutshinda, Finkel, Widdicombe, & Irwin, 2016), that they respond more or less similarly to environmental fluctuations, and that as a result global expression patterns are much more similar within than between phylum. Here we see it is not necessarily the case, as diatom expression profiles are sometimes more similar among diatom, dinoflagellate and/or chlorophyta species than within each phylum. The second assumption is that the community is composed of a multitude of co-occurring species and that the community wide expression dynamics is mainly driven by changes in expression within the co-occurring species and not by major modifications in terms of community composition. This second assumption does not hold in the community investigated in the present study as a great proportion of the community is composed of a few dominant species displaying major relative abundance fluctuations across samples. For instance, it appeared that diatom transcripts are abundant in three modules of expression (ME1, ME2, ME4). However, the species mainly contributing to the diatom level expression are different from one module to the other, meaning that shifts in expression patterns is driven by the succession of different species.

As global expression is adjusted to respond to fluctuating environmental conditions, the expression of a given species is expected to display variable expression patterns in different environmental conditions. Indeed, we saw that the expression modules are strongly correlated with environmental conditions. Interestingly there are three species for which expression is investigated in different modules. This is the case for *L. aporus* in ME5 and ME6. The transcripts belonging to ME6, tend to be highly expressed when silicate is scarce and *L. aporus* relatively abundant while it is the opposite for the transcripts belonging to ME5 and we note that a few specific biological processes display different expression patterns (digestion, glucose metabolism). In *G. spinifera* transcripts were analyzed in four expression modules highly expressed in different environmental conditions. A few specific differences were observed but globally the enriched biological processes are not dramatically different (see Results). The last species is *A. minutum*, sometimes displaying marked differences in terms of biological processes related to the highly expressed transcripts. This has already been extensively documented in chapter 3, but we may nevertheless highlight marked differences in enriched biological processes between ME1 (environment with a lot of diatoms, and relatively few *A. minutum*) and ME3 (high salinity), especially related to ion (Na, Ca, K) transporters. So far very few studies have investigated species specific dynamic of expression *in situ* (Alexander, Rouco, et al., 2015).

Coming back to the three questions highlighted at the end of the introduction, we saw that our approach enabled the characterization of species specific dynamic expression patterns for several members of the community. Nevertheless, the quality of the species specific reference transcriptomes is a key point to allow such characterization. We also saw that the expression dynamics strongly correlates with environmental fluctuations, suggesting that expression patterns may be used to investigate how species respond to environmental fluctuations *in situ*. One key point when performing *in situ* transcriptomic approaches is the ability to translate expression data into ecologically relevant functions. In the present case several biological processes identified as displaying dynamic expression patterns *in situ* may easily be translated into ecological terms (photosynthesis, digestion, fatty acid biosynthesis...). However, one should keep in mind that the identification of the biological processes is based on homology between genes from distantly related organisms. In addition to the homology based approach, like others (Keeling & del Campo, 2017), we would like to advocate for an experimentally based approach where expression patterns related to specific ecologically relevant processes would be analyzed under a controlled setting and then used as molecular markers of the ecological processes *in situ*. Finally, the present work shows that expression patterns may be compared between species, and that there are marked differences between species. However, we would like to add a word of caution. Indeed, contrary to what is done *in vitro*, natural populations are not composed of clones, but of genetically diverse cells, and we saw in the previous chapters focusing on *A. minutum* the importance of considering the cryptic and intraspecific diversity to fully understand dynamics of expression patterns *in situ*.

Chapter 6

General discussion, conclusions and perspectives

ONE can investigate responses to environmental fluctuations using different approaches. In marine microbial populations, due to the highly dynamic characteristics of the environment and to the large population sizes involved, using only few strains cultivated under controlled conditions seems dubious. Although the importance of conducting large scale surveys *in situ* is therefore obvious, such investigations remained challenging for long time, mainly due to technical difficulties. Recently, the "omic" revolution opened the gate to such investigations, and now allow, as exposed in the present manuscript, to monitor both gene expression and genetic differentiation of entire populations living in their natural environment.

6.1 Few considerations on the methodology

6.1.1 On the use of environmental data

One of the early reflexion that arose during this project was the importance of the specificity of the alignment of our environmental reads. As the approach relies on environmental data (and more precisely to the specific monitoring of *A. minutum* mRNA into the midst of the total mRNA from the entire community that was retained on the 20 μm filter), how can we be sure to only specifically study *A. minutum* mRNA ? Indeed, given that globally, nucleotide divergence is positively linked with phylogenetic distance (one of the basic assumptions of molecular phylogenetics; Brown (2002); Z. Yang and Rannala (2012), we can fairly hypothesize that (at least for conserved genes), transcript sequence in a given transcriptome won't be divergent enough between two closely related species to differentiate the two sequences. Therefore, the question asked here was : when we monitor reads from environmental mRNA aligning to *A. minutum* transcriptome, are we sure not to count also reads from other species present in the community ? Therefore, we decided to discard from the *A. minutum* transcriptome every transcript on which, simply by sequence homology, reads from other species align. Although the set of species we selected for this was certainly not fully extensive, it was the result of a balance between both data availability (only a limited, albeit consistent, number of marine microbial species transcriptome availables : MMETSP database (Keeling et al., 2014)) and representativity of these species in the co-occurring community within *A. minutum* blooms.

Of course, it is also possible to control for such bias (that are inherent to data coming from entire communities) by acting on reads themselves (prior to mapping on the reference transcriptome of interest), for instance by aligning the environmental reads to a set of references (as aligning methods are based on sequence homology, each reads are supposed to align to their respective reference). This method was for instance used in Alexander, Jenkins, et al. (2015), where they aligned their environmental reads to a sequence library containing all assembled sequences from the MMETSP database. It could be a complementary approach to the one we used.

6.1.2 *In situ* gene expression monitoring using a transcriptome from *in vitro* cultivated strains

First, although the results presented here are, to our knowledge, unprecedented insights into the molecular physiology of a non model species, a potentially important piece of information is lacking. Indeed, we were not able to study alternatively spliced transcripts, as no assembled genome is available for *A. minutum*. As we showed that two cryptic species differentially express the same transcript (which could be a way to avoid direct competition), we can hypothesize that they also evolved specific isoform expression. Few methods for expression level estimations can be used for isoform differential expression analysis, such as CUFFDIFF2 Trapnell et al. (2013). However, these approach classically require replicates that, in the case of the *in situ* survey we led here, are not available. Second, the transcriptome used here is coming from 18 individual strains grown under laboratory conditions. Therefore, if some functions are only expressed *in situ*, they are lacking from our assembly, resulting in a potential detection bias toward *in vitro* only expressed genes. Furthermore, as we showed that expression patterns between *in vitro* and *in situ* growing conditions are different, it appears crucial to find alternative strategies to capture these missing pieces of information. Again, this could be done using an assembled genome, which would serve as a reference toward which environmental reads could be aligned, completing the present analysis.

6.2 Conclusions and perspectives

Understanding organisms responses to their environment, which includes a wide continuum of both reversible and irreversible mechanisms (from biochemical buffering to genetic adaptation), is a cornerstone in biological sciences. Two kinds of responses may arise when facing a changing environment : gene expression variation (phenotypic plasticity) and genetic adaptation (through the natural selection of advantageous mutations in the populations). These responses are part of a continuum, and their respective contribution to population resilience from environmental perturbations is not fully understood and appears to greatly depend on the taxa considered. Microorganisms, by their high reproductive rates and large effective population sizes, clearly constitute predisposed model organisms to study the interplay between these two types of response. However, because their causes and consequences are tightly linked and may overlap, understanding their respective contribution to population survival toward environmental changes is still far from being straightforward.

All along this Ph.D project, my aim was to investigate the interplay between these responses, and how they interact with each other. To do this, I used a dual *in vitro* - *in situ* approach on entire populations and strains of *A. minutum*, and was able to study 1) the mechanisms underlying divergence in an incipient cryptic speciation; 2) how genetic divergence between two cryptic species traduces in terms of molecular physiology when facing similar environmental conditions; 3) how gene expression changes over a three year time scale in response to environmental fluctuations; 4) the factors responsible for genetic structure over time and space and 5) how the molecular physiology of different taxa co-occurring with *A. minutum* diverges and fluctuates through time.

6.2.1 On the interplay between genetic and molecular physiology divergence

In chapter 2, we started by posing the molecular basis of genetic divergence in a complex of incipient cryptic species in *A. minutum*. Overall, the most divergent transcripts were homologous to genes coding for proteins involved in ionic transducing functions (membrane fluxes or signal transduction) and saxitoxin production (a compound that is thought to act as a sex pheromone; Cusick and Sayler (2013); Wyatt and Jenkinson (1997). These results investigated some of the molecular mechanisms underlying the build up of reproductive barriers, and showed how sequencing data can be used to address major ecology and evolution questions, even in non model organisms. Following this, we investigated the link between genetic divergence and molecular physiology response. By focussing on the gene expression level differences between these two closely related species, we showed how they preferentially use different metabolic pathways under similar growing conditions. *In situ*, we showed that the molecular physiology of both cryptic species can be drastically different, suggesting that they evolved different kinds of gene expression patterns, maybe as a way to avoid direct competition. Altogether, in chapter 2, we decrypted how genetic divergence induces profound molecular changes, even in the case of incipient speciation. In addition to potentially enhance the appearance of reproductive barriers, such divergence may also impact the ability of populations to change their gene expression levels in response to their changing environment. The *in vitro* experiments led in this study were restricted to monoclonal strains grown under controlled conditions, and still, molecular physiology divergence was observable. Under natural conditions (with interspecific interactions, environmental fluctuations etc), we can hypothesize that the molecular physiology of these two cryptic species may be widely different. Further investigating the interactions between genetic divergence and molecular physiology is therefore of primary interest to understand ecological divergence and the molecular basis of speciation for instance. Leading similar investigations in other microbial taxa (for instance occupying different trophic niches : from primary producers to parasites) could greatly help understanding the mechanisms underlying speciation processes and ecological divergence.

6.2.2 Gene expression dynamics *in situ*

In chapter 3, we were interested in the molecular physiology dynamics and response to environmental fluctuations of populations living in their natural environment. Using a large scale dataset, we showed several major groups of genes co-expressed through time, enriched in specific functions, and correlated with abiotic factors variations. Of primary interest was two sets of genes respectively gathering functions related to motility behaviors and cell-to-cell interactions, that were more expressed in cold, low irradiance and salinity conditions. More than only being an unprecedented insight into the physiological dynamics of a microbial species in its natural environment, this study shed a new light on how populations can change their molecular physiology in a highly dynamic manner. It also gives an *in situ* example of the buffering potential given by phenotypic plasticity toward environmental changes. Among the different factors influencing gene expression, we showed a low but still existent effect of the time between sunrise and sampling, suggesting some amount of circadian-triggered gene expression levels. In other organisms, some studies showed that gene expression levels vary across the time of the day (see Akman, Carlson, Holsinger, and Latimer (2016) in the South African shrub *Protea repens* for instance). Then, it would be interesting to analyse the gene expression dynamics of *A. minutum* with a high sampling effort on a small timescale, such as one sampling every hour for instance. Of course, the sampling methodology should be adapted to incorporate the effect of tide on water mass shift, and two possibilities are feasible : (i) using techniques to follow the water mass, and therefore potentially sample the same groups of cells independently from the effect of tide or (ii) keeping a fixed sampling point, which will lead to sampling different groups of cells depending on water mass displacement. The latter approach, by potentially sampling different groups of individuals closely situated but still subjected to fine hydrological constraints would be a great opportunity to analyse fine scale genetic diversity dynamics. With this kind of survey, we could test how the interplay between genetic divergence and phenotypic plasticity translates a such small scale. More than only being an interesting complement of the analysis led in both chapter 3 and 4, it would give further information on fine scale bloom dynamics.

One of the advantage of the approach used here is that the monitoring of gene expression response was done without any pre-consideration and allowed to uncover biologically relevant functions that are not classically sought *in vitro*. Therefore, it has a great potential for non model species and should be more extensively used as an hypothesis rising tool, allowing to conduct targeted experiments that would not have been led otherwise. We developed this aspect in chapter 5, by investigating the gene expression dynamics of the major taxa present in the community during *A. minutum* bloom. Such approach led us to uncover species-specific gene expression patterns, that are correlated with environmental fluctuations. Scaling up such investigations at this level addresses important questions relative to community-wide dynamics, and is of great interest to any research field that aim to understand how species co-occur, interact with their environment and how such interactions translate in terms of molecular physiology. Also, as the biological functions showed to be differentially expressed (between taxa or between environmental conditions) can be easily translated into ecologically relevant processes, such investigations can lead to determinant insights into both niche partitioning, community coexistence questions (the mechanisms that allow species diversity coexistence), or eco-evolutionary divergence processes for instance.

6.2.3 Spatio-temporal genetic structure

In chapter 4, we investigated another level of biological response toward environmental fluctuations by focusing on genetic adaptation. To do so, we analysed genetic structure in *A. minutum* populations blooming in the bay of Brest at two levels : both spatially (9 points sampled simultaneously) and temporally (3 year survey, 2 sampling per week during the blooming period). Previous studies showed genetic structure patterns between distantly located *Alexandrium* populations (see for instance Dia et al. (2014) : two estuaries separated by 150 km or Casabianca et al. (2012) : strains sampled across the entire Mediterranean sea). Here, we showed indices of population structure at an unrivaled small spatial scale (few kilometers, in a highly dynamic hydrological context), and our results highlight how spatially closely situated populations can experience genetic differentiation. Very interestingly, we showed that population connectivity through hydrological dynamics is not necessarily the only factor responsible for the genetic structure observed. This suggests that marine microbial populations may experience sufficiently high selective pressures from other factors to counterbalance individual dispersal caused by water mixing. These results are of primary interest to understand how local adaptation can lead to population divergence, even in the marine environment which is supposed to promote high population genetic homogeneity. Also, in the context of harmful algal blooms, these results are of great importance for understanding the dynamic of such events, both in terms of eco-evolutionary dynamics and ecological management decision processes for instance. Second, we investigated the genetic structure of populations subjected to a fluctuating environment through time. Interestingly, the different blooms sampled showed different patterns of genetic structure. One striking result is how genetic differentiation changes between years (2013, 2014 and 2015) showed high increasing differences in genetic structure indices. Another major result is that intra-year genetic structure can be explained by different factors depending on the bloom considered. The bloom of 2013 for instance, displays some amount of increasing genetic differentiation with time, suggesting gradual and rapid temporal adaptation. Very interestingly, the two other sampled blooms (2014 and 2015) did not show such patterns, but in 2015 bloom for instance, we were able to correlate genetic differentiation with fluctuations in the composition of the microbial community co-occurring with *A. minutum*. Overall, these results are of special interest because they highlight how genetic structure can occur (at a week to month time scale) in *A. minutum* populations in response to different types of factors (time, biotic or abiotic). Moreover, we showed positive correlation between genetic differentiation and gene expression levels, highlighting the close interplay between genetic and molecular physiology divergence, and further underlining how the different biological responses to environmental perturbations (from gene expression to genetic adaptation) are part of a continuum, and how they interact with each other. One of the next steps in investigating the fine scale genetic adaptatation of *A.minutum* would be to conduct an outlier analysis, in order to look for genes under selection. This would help understanding the forces that shape genetic differentiation among *A. minutum* populations. On a general context, such observations may have substantial implications on our representation of microbial population dynamics, and should be incorporated in studies aiming to understand their evolution (for instance in the context of invasive species management or climate change adaptation).

As a general perspective, it would be of primary interest to develop alternative approaches to separate further the respective effects of genetic adaptation and phenotypic plasticity. Both *in situ* and *in vitro* approaches have their respective drawbacks (*in situ* with sampling problematics and *in vitro* with small number of strains grown under controlled conditions - therefore unrepresentative of the natural environment - for instance). A mesocosm (which is an outdoor experimental system that allow to conduct semi controlled experiments) based approach for instance could be promising, as it would allow to perform the same kind of methodologies that were conducted here, and therefore to take advantages of their great potential in hypothesis rising. With these kinds of devices, we could test the response (both in terms of genetic adaptation and phenotypic plasticity) of different blooms that would be, for instance, isolated in different area (both distant and close) or time. Also, such systems allow to monitor the response of an entire community to changes in one or few environmental factor fluctuations, and would, by their dual *in natura*-controlled characteristics, provide an additional overview of populations response mechanisms.

During my Ph.D, I investigated several aspects around a central question : how do populations respond to changes in their environment ? In particular, I focused on two kinds of response : gene expression and genetic adaptation. The results underlined how they may be tightly linked, but still fluctuate in a distinctive manner. They shed a new light on population responses to environmental changes, and can profoundly change our perspective of biological dynamics. Undoubtedly, the road is still long toward the complete resolving of such problematic, and I have the belief that it will only be achieved through transdisciplinary approaches and ambitious exploratory surveys, such as the one led here.

Appendix A

Supplementary material for chapter 2

A.1 Supplementary figure 1 :

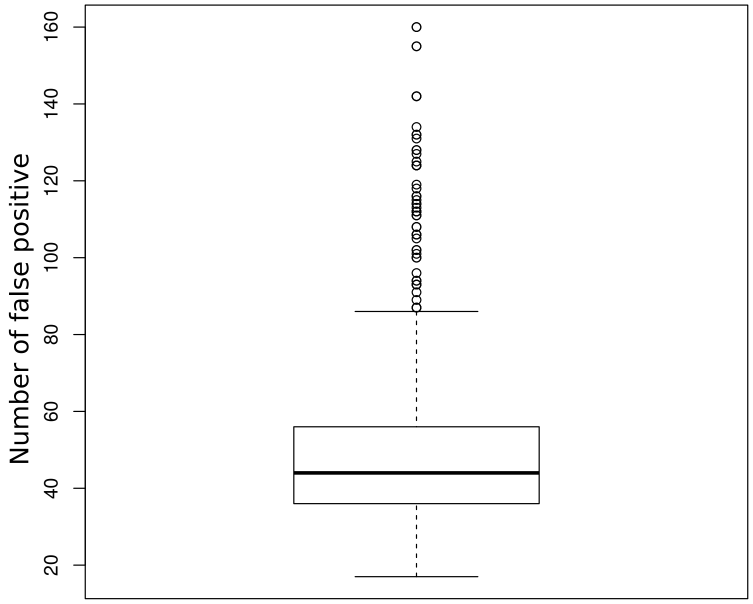


FIGURE A.1: Distribution of the number of SNPs observed as “fixed” at least once when permuting each strain from both cryptic species.

A.2 Supplementary figure 2 :

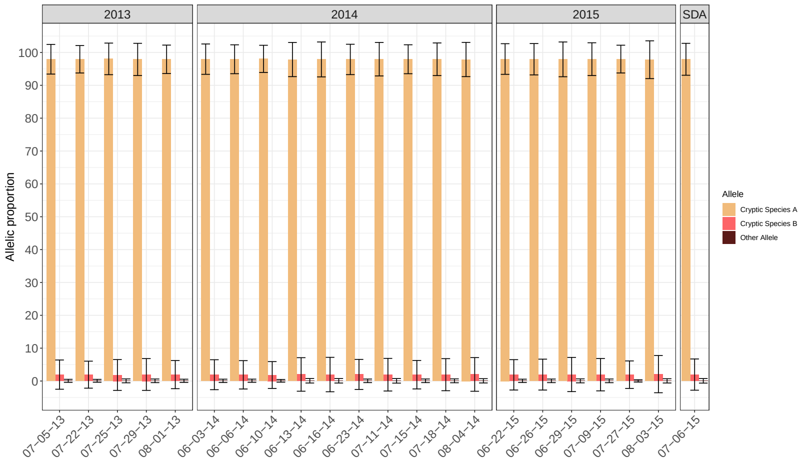


FIGURE A.2: *Proportions of alleles found at the 918 SNP positions analysed. Are represented allele proportions from alleles of the cryptic species A (yellow), B (red) in addition to the proportion of alleles from neither cryptic species (brown), here assimilated as background noise in the data.*

A.3 Supplementary figure 3 :

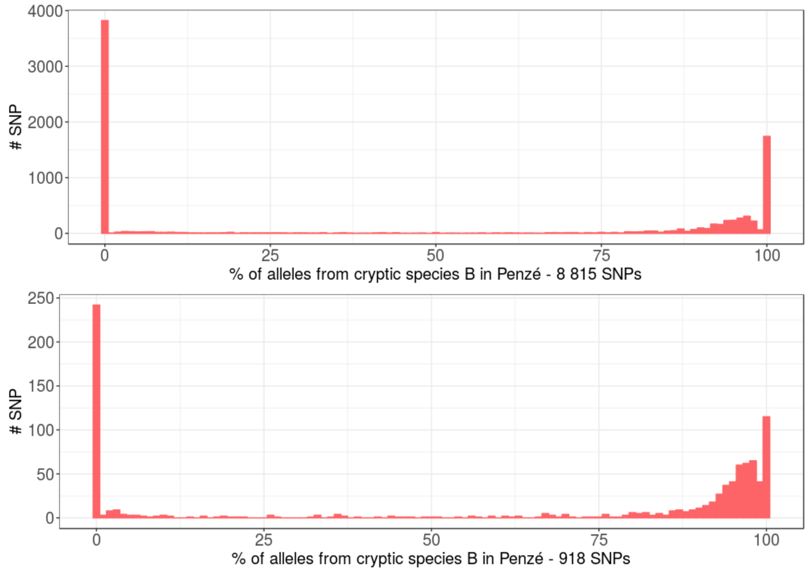


FIGURE A.3: *Comparison of allele frequencies distributions over both 8,815 SNP positions (Depth of 10 only in Penzé sample) and 918 SNP positions (depth of 10 in all samples considered in the present study).*

A.4 Supplementary figure 4 :

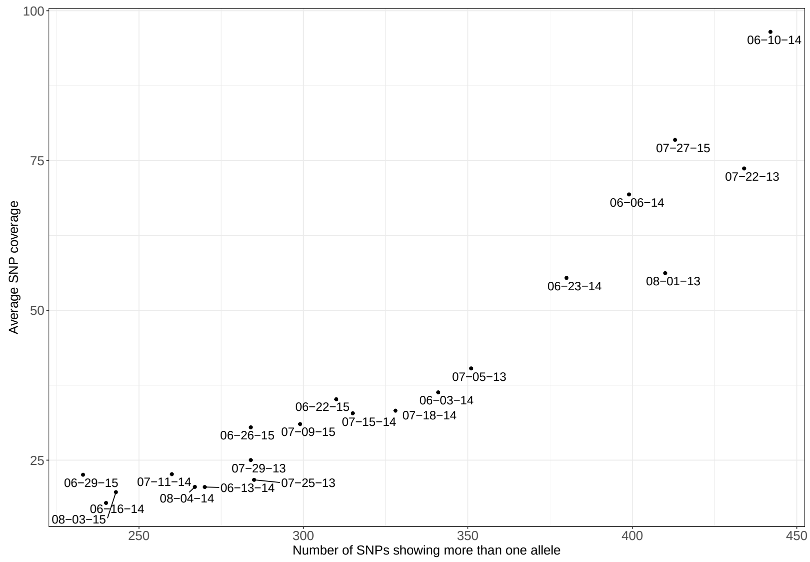


FIGURE A.4: *Number of SNPs showing more than one allele in function to the average SNP coverage over the populations sampled at the Daoulas sampling site.*

A.5 Supplementary figure 5 :

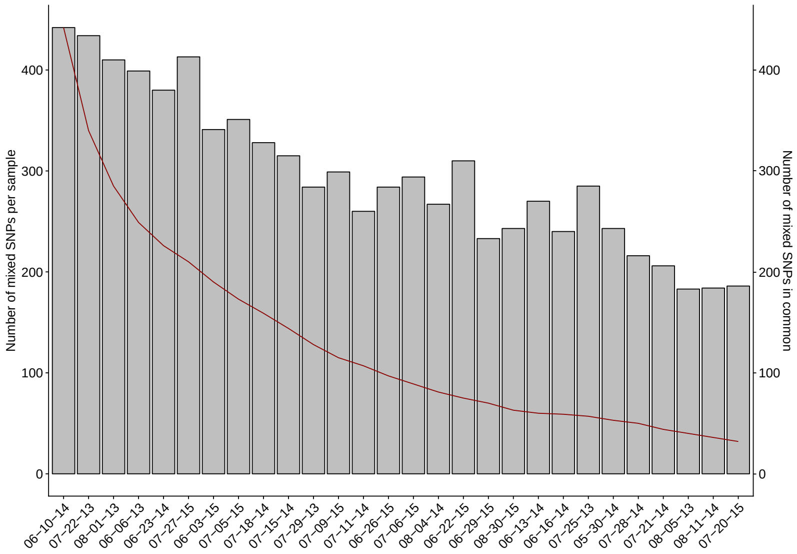


FIGURE A.5: *Number of SNPs showing mixed proportions of alleles from the different cryptic species per samples (gray bars) and evolution of the number of SNPs showing mixed proportions of alleles shared by the different samples (red line) at the Daoulas sampling site as adding new samples.*

A.6 Supplementary Table 1 :

GO over-representation (Fisher exact test) among transcripts differentially expressed by cells from cryptic species A grown *in situ* vs *in vitro*

	GO	GO description	Transcript
Enriched *in situ*	GO:0019843	rRNA binding	comp106026_c0_seq1
			comp108361_c0_seq1
			comp22514_c0_seq1
			comp87794_c0_seq1
			comp98082_c0_seq1
			comp117878_c0_seq1
			comp95353_c0_seq1
			Transcript
	GO	GO description	comp103809_c0_seq1
Enriched *in vitro*	GO:0003684	damaged DNA binding	comp111235_c0_seq1
			comp81048_c0_seq1
			comp106654_c0_seq2
			comp125547_c0_seq1
			comp126173_c0_seq1
			comp131872_c2_seq1
			comp114923_c0_seq2
	GO:0005768	endosome	comp125824_c0_seq1
			comp126973_c0_seq1
			comp108154_c0_seq1
			comp130540_c0_seq1
			comp123461_c1_seq2
			comp126460_c1_seq1
			comp93951_c1_seq1
			comp87755_c0_seq1
			comp129211_c0_seq1
			comp126450_c0_seq1
	GO:0007283	spermatogenesis	comp132640_c0_seq1
			comp131624_c0_seq1
			comp24608_c0_seq1
			comp58919_c0_seq1
			comp105801_c0_seq1
			comp90434_c0_seq1
			comp130398_c0_seq1
			comp97984_c0_seq1
			comp93801_c0_seq1
			comp106191_c0_seq1
	GO:0015992	proton transmembrane transport	comp118707_c0_seq1
			comp128086_c0_seq1
			comp130816_c0_seq1
			comp59646_c1_seq1
			comp97111_c0_seq1

Homolog	Transcript Description	p-value	Odd ratio
RL1_THELT	50S ribosomal protein L1	0.00013	7.77
MRA1_SCHPO	Ribosomal RNA small subunit methyltransferase mra1		
RL16_MOOTA	50S ribosomal protein L16		
RL4_SYNY3	50S ribosomal protein L4		
RS13_HERA2	30S ribosomal protein S13		
IMP3_MOUSE	U3 small nucleolar ribonucleoprotein protein IMP3		
RNC_AGRVS	Ribonuclease 3		
Homolog	Transcript Description	p-value	Odd ratio
RAD23_YEAST	UV excision repair protein RAD23	0.0004	6.55
PNKP_HUMAN	Bifunctional polynucleotide phosphatase/kinase		
MSH1_ARATH	DNA mismatch repair protein MSH1, mitochondrial		
POLI_MOUSE	DNA polymerase iota		
RECA_NOSS1	Protein RecA		
REV1_MOUSE	DNA repair protein REV1		
POLK_MOUSE	DNA polymerase kappa		
CATE_CAVPO	Cathepsin E	0.0028	3.2
ALA3_ARATH	Phospholipid-transporting ATPase 3		
QSOX1_ORYSJ	Sulfhydryl oxidase 1		
CSR1_ASHGO	Phosphatidylinositol transfer protein CSR1		
APY2_ORYSJ	Probable apyrase 2		
RFIP3_MOUSE	Rab11 family-interacting protein 3		
KI16B_HUMAN	Kinesin-like protein KIF16B		
EHD1_MOUSE	EH domain-containing protein 1		
CATE_RAT	Cathepsin E		
GUX3_ARATH	Putative UDP-glucuronate:xylan alpha-glucuronosyltransferase 3		
CTSR2_RAT	Cation channel sperm-associated protein 2	0.002	3.5
SPNE_ANOGA	Probable ATP-dependent RNA helicase spindle-E		
ALKB5_MOUSE	RNA demethylase ALKBH5		
ALKB5_DANRE	RNA demethylase ALKBH5		
TDRD9_DANRE	Putative ATP-dependent RNA helicase TDRD9		
TLK2_MOUSE	Serine/threonine-protein kinase tousled-like 2		
CETN2_HUMAN	Centrin-2		
DIAP3_HUMAN	Protein diaphanous homolog 3		
ADCYA_HUMAN	Adenylate cyclase type 10		
RN151_MOUSE	RING finger protein 151		
TCR4_SALOR	Tetracycline resistance protein, class D	0.003	4.95
S36A1_RAT	Proton-coupled amino acid transporter 1		
INT1_ARATH	Inositol transporter 1		
HPPA2_METAC	K		
HPPA_CLOTE	Putative K		
ANT1_YEAST	Peroxisomal adenine nucleotide transporter 1		

Appendix B

Supplementary material for chapter 3

B.1 Supplementary Table 1 :

TABLE B.1: Summary of the different data available for the different environmental samples. XX/XX/XXXX* : sample respecting the quality thresholds and then used in the study.

Sampling date	Sample name	Time from sunrise	Progression throught the bloom	Sampling time (decimal hours)	Tidal coefficient	Ammonium (μM)	Nitrogen oxide (μM)	Phosphate (μM)
05/07/2013*	050713_CTTGTA_L007_RX.fastq.gz	8.95	0	14.33	59	0.339	10.02	0.03
08/07/13	080713_GGCTAC_L007_RX.fastq.gz	2.41	9.6774193548	7.83	77	0.494	2.01	0.02
11/07/13	110713_TTAGGC_L007_RX.fastq.gz	3.86	19.3548387097	9.33	79	0.449	1.70	0.03
15/07/2013	150713_CGATGT_L007_RX.fastq.gz	6.22	32.2580645161	11.75	61	0.287	0.61	0.05
18/07/2013	180713_ATCACG_L007_RX.fastq.gz	6.17	41.935483871	11.75	53	0.276	0.94	0.04
22/07/2013*	220713_GGCTAC_L006_RX.fastq.gz	2	54.8387096774	7.67	93	0.995	1.26	0.11
25/07/2013*	250713__ACTGAT_L007_RX.fastq.gz	3.53	64.5161290323	9.25	107	0.424	0.75	0.12
29/07/2013*	290713_GAGTGG_L007_RX.fastq.gz	6.28	77.4193548387	12.08	61	1.152	1.22	0.21
01/08/2013*	010813_CGTACG_L007_RX.fastq.gz	6.46	87.0967741935	12.33	38	1.707	8.07	0.31
05/08/2013*	050813_AGTCAA_L007_RX.fastq.gz	9.48	100	15.43	70	0.725	2.06	0.23
30/05/2014*	300514_ATCACG_L001_RX.fastq.gz	2.88	0	8.25	85	0.956	14.59	0.20
03/06/2014*	030614_CGATGT_L001_RX.fastq.gz	2.67	5	8	61	1.547	16.97	0.27
06/06/2014*	060614_TTAGGC_L001_RX.fastq.gz	6.2	8.75	11.5	42	0.088	10.30	0.25
10/06/2014*	100614_TGACCA_L001_RX.fastq.gz	8.65	13.75	13.92	68	0.083	7.38	0.20
13/06/2014*	130614_GATCAG_L003_RX.fastq.gz	1.98	17.5	7.25	93	0.44	3.33	0.17
16/06/2014*	160614_ACAGTG_L001_RX.fastq.gz	2.97	21.25	8.22	98	0.301	1.34	0.24
23/06/2014*	230614_GCCAAT_L001_RX.fastq.gz	8.23	30	13.5	61	0.254	3.19	0.23
27/06/2014	270614_CAGATC_L001_RX.fastq.gz	2.03	35	7.33	78	0.341	0.52	0.21
30/06/2014*	300614_TAGCTT_L001_RX.fastq.gz	2.35	38.75	7.67	77	0.512	1.80	0.18
11/07/2014*	110714_GGCTAC_L001_RX.fastq.gz	1.2	52.5	6.67	81	0.812	3.00	0.33
15/07/2014*	150714_CTTGTA_L002_RX.fastq.gz	2.97	57.5	8.5	106	0.248	0.81	0.31
18/07/2014*	180714_AGTCAA_L002_RX.fastq.gz	3.59	61.25	9.17	78	0.713	1.16	0.32
21/07/2014*	210714_AGTTCC_L002_RX.fastq.gz	6.29	65	11.92	48	0.568	1.14	0.33
25/07/2014	250714_ATGTCA_L002_RX.fastq.gz	1.95	70	7.67	68	1.085	4.72	0.40
28/07/2014*	280714_CCGTCC_L002_RX.fastq.gz	11	73.75	16.78	81	0.043	0.06	0.28
04/08/2014*	040814_GGCTAC_L003_RX.fastq.gz	5.57	82.5	11.5	48	0.962	2.72	0.39
11/08/2014*	110814_AGTCAA_L003_RX.fastq.gz	10.4	91.25	16.5	110	0.114	0.01	0.23
14/08/2014	140814_AGTTCC_L003_RX.fastq.gz	2.83	95	9	94	0.282	0.54	0.22
18/08/2014*	180814_ATGTCA_L003_RX.fastq.gz	5.25	100	11.5	48	1.396	2.34	0.30
15/06/2015	150615B_CTTGTA_L006_RX.fastq.gz	9.75	0	15	87	0.24	0.85	0.16
19/06/2015	190615B_AGTCAA_L006_RX.fastq.gz	2.73	8.1632653061	8	85	0.59	1.91	0.09
22/06/2015*	220615B_AGTTCC_L006_RX.fastq.gz	4.48	14.2857142857	9.75	62	0.73	1.94	0.07
26/06/2015*	260615B_ATGTCA_L006_RX.fastq.gz	5.9	22.4489795918	11.18	40	1.05	3.19	0.12
29/06/2015*	290615_CCGTCC_L006_RX.fastq.gz	9.43	28.5714285714	14.75	64	0.12	0.33	0.05
02/07/15	020715_GTGAAA_L006_RX.fastq.gz	2.15	34.693877551	7.5	87	0.56	0.76	0.05
06/07/15	060715_GTGGCC_L006_RX.fastq.gz	3.87	42.8571428571	9.25	91	0.517	0.97	0.11
09/07/2015*	090715_GTTTCG_L006_RX.fastq.gz	4.07	48.9795918367	9.5	66	2.06	1.18	0.12
20/07/2015*	200715AB_GAGTGG_L006_RX.fastq.gz	3.13	71.4285714286	8.75	76	1.36	10.18	0.27
23/07/2015	230715_ATCACG_L006_RX.fastq.gz	2.82	77.5510204082	8.5	52	0.66	1.49	0.17
27/07/2015*	270715_CGATGT_L006_RX.fastq.gz	7.98	85.7142857143	13.75	47	0.82	19.54	0.17
03/08/2015*	030815_TTAGGC_L006_RX.fastq.gz	12.83	100	18.75	104	0.7	0.72	0.25

Sampling date	Silicate (μM)	Mignonne outflow (m^3.s^{-1})	Temp. (°C)	Salinity (W m^{-2})	Irradiance cells.L^{-1}	*A. minutum*	Vol. filtered (L)	Number of total reads	Number of mapped reads	Alignment percentage	Proportion of transcript > 40 reads
05/07/2013*	7.24	0.46	19.50	32.10	289.00	30631	4.60	33079735	8277944	25.02	37.56
08/07/13	0.56	0.37	19.00	32.90	350.60	20000	4.50	18192984	512967	2.82	0.64
11/07/13	1.59	0.32	19.40	33.60	348.20	12500	4.30	22623043	435687	1.93	0.35
15/07/2013	2.19	0.28	21.00	33.70	346.50	30800	4.20	7372924	110903	1.50	0.06
18/07/2013	2.89	0.27	21.60	33.90	338.80	252500	4.00	19841307	967026	4.87	4.18
22/07/2013*	4.58	0.26	21.80	34.10	202.60	122000	3.00	31619145	15898291	50.28	48.05
25/07/2013*	4.96	0.25	22.20	34.10	242.30	103900	3.25	15512632	4592918	29.61	26.94
29/07/2013*	5.35	0.25	21.60	34.00	208.40	41400	5.00	10884836	5120818	47.05	30.05
01/08/2013*	13.68	0.26	21.90	33.50	296.40	203224	2.75	21585379	11955464	55.39	42.86
05/08/2013*	14.17	0.21	22.00	33.70	229.20	7900	5.50	20281772	2198723	10.84	15.82
30/05/2014*	6.04	1.24	16.10	30.90	177.10	10700	7.25	19328092	3195933	16.54	20.98
03/06/2014*	7.07	0.98	16.90	31.00	133.20	90000	4.90	24931017	6277695	25.18	30.19
06/06/2014*	5.84	0.84	16.80	30.90	210.00	1052002	4.62	22413077	11529143	51.44	40.54
10/06/2014*	5.17	0.68	18.10	31.00	285.90	216500	2.13	28604242	16482841	57.62	43.83
13/06/2014*	4.35	0.57	18.00	32.20	357.60	40700	5.50	11257091	3979672	35.35	24.46
16/06/2014*	3.25	0.51	18.40	33.10	345.40	403000	4.20	6992704	3171918	45.36	22.03
23/06/2014*	4.13	0.39	20.50	33.00	333.20	191100	3.00	23660321	9710098	41.04	37.42
27/06/2014	3.27	0.37	20.60	33.60	219.50	5200	4.25	22208403	667027	3.00	3.52
30/06/2014*	5.62	0.38	20.00	33.40	133.30	5800	7.00	20413348	1660251	8.13	12.05
11/07/2014*	10.41	0.39	19.20	32.40	217.60	26000	6.25	27615900	4442894	16.09	25.49
15/07/2014*	7.66	0.38	19.30	33.30	230.50	9200	3.48	31945265	6372299	19.95	30.89
18/07/2014*	8.03	0.36	20.40	33.70	168.70	18000	1.72	23089263	6361295	27.55	30.95
21/07/2014*	9.59	0.33	21.10	33.90	289.40	93100	6.00	30080717	3006895	10.00	20.15
25/07/2014	7.81	0.28	21.70	33.20	315.30	5700	6.00	10592028	693535	6.55	3.46
28/07/2014*	3.94	0.26	21.60	34.00	251.30	28900	5.00	28148606	2843598	10.10	18.61
04/08/2014*	8.71	0.24	21.10	33.80	283.60	44900	5.50	15972564	3994959	25.01	24.63
11/08/2014*	4.49	0.49	20.10	33.80	229.80	11200	4.50	14977588	1787267	11.93	12.64
14/08/2014	4.79	0.36	19.30	33.60	170.30	18000	4.00	13372263	1116116	8.35	7.22
18/08/2014*	6.67	0.27	18.80	33.50	251.90	9500	6.50	20670929	1762263	8.53	12.32
15/06/2015	5.69	0.40	17.60	33.40	224.50	2100	6.25	14938302	1044330	6.99	7.41
19/06/2015	4.71	0.36	16.90	33.30	360.30	3200	7.00	10279204	498081	4.85	1.68
22/06/2015*	6.19	0.32	19.40	33.60	148.70	12000	7.00	21584555	4971001	23.03	28.41
26/06/2015*	6.78	0.29	20.10	33.80	122.10	50000	5.00	20042093	4339344	21.65	26.94
29/06/2015*	4.35	0.27	20.70	34.00	316.00	10200	5.00	25472682	3218067	12.63	22.63
02/07/15	1.83	0.25	20.70	34.00	267.10	1500	5.20	21836649	883173	4.04	5.61
06/07/15	3.71	0.23	20.70	34.00	281.30	3800	5.10	12645201	876213	6.93	5.85
09/07/2015*	4.18	0.22	19.70	34.20	295.50	4500	5.00	32796886	4285038	13.07	23.97
20/07/2015*	10.90	0.56	19.70	34.10	103.30	4400	5.00	14845608	2245562	15.13	16.15
23/07/2015	8.54	0.26	20.20	34.10	133.20	30000	5.00	15788227	1112721	7.05	8.12
27/07/2015*	10.91	0.37	17.60	34.10	175.50	29000	5.20	23255353	11058978	47.55	34.79
03/08/2015*	6.24	0.23	19.70	34.10	117.00	9100	5.00	16200254	2681130	16.55	17.35

B.2 Supplementary Table 2 :

TABLE B.2: Table displaying information about the 24 species used for the specificity test of our transcriptome

Species	Class	MMETSP accession nb	Investigators	Number of reads	Nb. reads aligned on transcriptome	Alignement %
Amphora coffeaeformis	Bacillariophyceae (Diatom)	CAM_SMPL_002502	Peter Stief	63726418	71664	0.11
Aurantiochytrium limacinum	Labyrinthulomycetes	CAM_SMPL_002701	Jackie Collier	39811132	61982	0.16
Ceratium fusus	Dinophyceae	CAM_SMPL_002648	Susanne Menden-Deuer	54627688	55757	0.10
Chaetoceros curvisetus	Coscinodiscophyceae (Diatom)	CAM_SMPL_002847	Sarah Smith	110999750	8461923	7.62
Chaetoceros debilis	Coscinodiscophyceae (Diatom)	CAM_SMPL_002437	Bank Beszteri	52721386	34446	0.07
Cylindrotheca closterium	Bacillariophyceae (Diatom)	CAM_SMPL_002329	E. Virginia Armbrust	34126456	9536	0.03
Dactyliosolen fragilissimus	Coscinodiscophyceae (Diatom)	CAM_SMPL_003104	Robert Olson	39877006	20357	0.05
Dinophysis acuminata	Dinophyceae	CAM_SMPL_002811	Jeremiah Hackett	48685084	140863	0.29
Gambierdiscus australes	Dinophyceae	CAM_SMPL_002780	Shauna Murray	58256390	2870704	4.93
Gonyaulax spinifera	Dinophyceae	CAM_SMPL_003287	Mike Preston	52639524	363158	0.69
Gymnodinium catenatum	Dinophyceae	CAM_SMPL_000708	Jeremiah Hackett	73471400	595291	0.81
Heterocapsa arctica	Dinophyceae	CAM_SMPL_003289	Mike Preston	54674410	183562	0.34
Heterocapsa rotundata	Dinophyceae	CAM_SMPL_000705	George McManus	43256526	172610	0.40
Heterocapsa triquestra	Dinophyceae	CAM_SMPL_002577	Susanne Menden-Deuer	45359332	122942	0.27
Leptocylindrus danicus	Coscinodiscophyceae (Diatom)	CAM_SMPL_003132	E. Virginia Armbrust	39861406	5398	0.01
Licmophora paradoxa	Coscinodiscophyceae (Diatom)	CAM_SMPL_003130	E. Virginia Armbrust	31537530	8499	0.03
Lingulodinium polyedra	Dinophyceae	CAM_SMPL_002589	Mary Ann Moran	81266712	576642	0.71
Peridinium aciculiferum	Dinophyceae	CAM_SMPL_002534	Karin Rengefors	41244260	297255	0.72
Protoceratium reticulatum	Dinophyceae	CAM_SMPL_002473	Juan Saldarriaga	46869592	284792	0.61
Pseudo nitzschia australis	Bacillariophyceae (Diatom)	CAM_SMPL_002414	G Jason Smith	62072312	9905	0.02
Rhizosolenia setigera	Coscinodiscophyceae (Diatom)	CAM_SMPL_002791	E. Virginia Armbrust	53486746	58520	0.11
Scrippsiella hangoei	Dinophyceae	CAM_SMPL_002520	Karin Rengefors	54731718	54867	0.10
Scrippsiella trochoidea	Dinophyceae	CAM_SMPL_002516	Geoff Sinclair, Boris Wawrik	61631094	513093	0.83
Skeletonema costatum	Coscinodiscophyceae (Diatom)	CAM_SMPL_002340	E. Virginia Armbrust	36792038	51155	0.14

Species	Nb of *A. minutum* transcripts with 1 read aligned	Number of reads per aligned transcripts	Transcript with most read aligned	Number of reads on most aligned transcript	Divergence
Amphora coffeaeformis	152.00	471.47	comp93300_c0_seq1 (rRNA protein 2)	55866.00	Phylum
Aurantiochytrium limacinum	87.00	712.44	comp93300_c0_seq1 (rRNA protein 2)	59562.00	sub order Chromalveolata
Ceratium fusus	214.00	260.55	comp93300_c0_seq1 (rRNA protein 2)	47687.00	Famille
Chaetoceros curvisetus	62.00	136482.63	comp93300_c0_seq1 (rRNA protein 2)	5217670.00	Phylum
Chaetoceros debilis	24.00	1435.25	comp93300_c0_seq1 (rRNA protein 2)	21038.00	Phylum
Cylindrotheca closterium	23.00	414.61	comp93300_c0_seq1 (rRNA protein 2)	6036.00	Phylum
Dactyliosolen fragilissimus	52.00	391.48	comp93300_c0_seq1 (rRNA protein 2)	16299.00	Phylum
Dinophysis acuminata	2051.00	68.68	comp93300_c0_seq1 (rRNA protein 2)	52431.00	Ordre
Gambierdiscus australes	1665.00	1724.15	comp93300_c0_seq1 (rRNA protein 2)	1962123.00	Genre
Gonyaulax spinifera	2432.00	149.32	comp126209_c0_seq1 (Cytochrome c oxidase subunit 1)	110075.00	Famille
Gymnodinium catenatum	816.00	729.52	comp93300_c0_seq1 (rRNA protein 2)	346084.00	Ordre
Heterocapsa arctica	2274.00	80.72	comp93300_c0_seq1 (rRNA protein 2)	23698.00	Ordre
Heterocapsa rotundata	2338.00	73.83	comp93300_c0_seq1 (rRNA protein 2)	15413.00	Ordre
Heterocapsa triquestra	1888.00	65.12	comp93300_c0_seq1 (rRNA protein 2)	60031.00	Ordre
Leptocylindrus danicus	35.00	154.23	comp93300_c0_seq1 (rRNA protein 2)	3489.00	Phylum
Licmophora paradoxa	66.00	128.77	comp93300_c0_seq1 (rRNA protein 2)	6559.00	Phylum
Lingulodinium polyedra	5290.00	109.01	comp8132_c1_seq1 (Photosystem II protein D1)	189319.00	Genre
Peridinium aciculiferum	1059.00	280.69	comp93300_c0_seq1 (rRNA protein 2)	268307.00	Ordre
Protoceratium reticulatum	3302.00	86.25	comp8132_c1_seq1 (Photosystem II protein D1)	82273.00	Famille
Pseudo nitzschia australis	63.00	157.22	comp93300_c0_seq1 (rRNA protein 2)	7571.00	Phylum
Rhizosolenia setigera	51.00	1147.45	comp93300_c0_seq1 (rRNA protein 2)	51281.00	Phylum
Scrippsiella hangoei	405.00	135.47	comp93300_c0_seq1 (rRNA protein 2)	43113.00	Ordre
Scrippsiella trochoidea	1200.00	427.58	comp93300_c0_seq1 (rRNA protein 2)	423731.00	Ordre
Skeletonema costatum	29.00	1763.97	comp93300_c0_seq1 (rRNA protein 2)	28647.00	Phylum

B.3 Supplementary Table 3 :

TABLE B.3: Gene Ontology categories over-represented within the non specific transcripts

GO	Name	Category	Number of transcript in transcriptome	Number of transcript in aspecific transcripts	Proportion of transcripts	Fisher test q-value
GO:0016226	iron-sulfur cluster assembly	biological process	13	8	0.62	0.027
GO:0051539	4 iron 4 sulfur cluster binding	molecular function	38	17	0.45	0.026
GO:0009060	aerobic respiration	biological process	5	5	1	0.01
GO:0009535	chloroplast thylakoid membrane	cellular component	62	26	0.42	0.012
GO:0009570	chloroplast stroma	cellular component	70	31	0.44	0.0029
GO:0015995	chlorophyll biosynthetic process	biological process	14	10	0.71	0.0045
GO:0045494	photoreceptor cell maintenance	biological process	11	8	0.73	0.01
GO:0050660	flavin adenine dinucleotide binding	molecular function	69	29	0.42	0.0091
GO:0005759	mitochondrial matrix	cellular component	54	24	0.44	0.0099
GO:0000105	histidine biosynthetic process	biological process	10	7	0.7	0.022
GO:0000398	mRNA splicing via spliceosome	biological process	49	27	0.55	0.00016
GO:0000502	proteasome complex	cellular component	32	20	0.62	0.0002
GO:0004298	threonine-type endopeptidase activity	molecular function	6	6	1	0.0042
GO:0004843	thiol-dependent ubiquitin-specific protease activity	molecular function	32	18	0.56	0.0024
GO:0005525	GTP binding	molecular function	152	63	0.41	0.000064
GO:0005840	ribosome	cellular component	41	21	0.51	0.0029
GO:0006886	intracellular protein transport	biological process	58	24	0.41	0.018
GO:0007093	mitotic cell cycle checkpoint	biological process	6	5	0.83	0.026
GO:0008380	RNA splicing	biological process	101	47	0.47	0.000064
GO:0010467	gene expression	biological process	58	24	0.41	0.018
GO:0015031	protein transport	biological process	164	60	0.37	0.0029
GO:0015991	ATP hydrolysis coupled proton transport	biological process	10	8	0.8	0.0057
GO:0016485	protein processing	biological process	8	6	0.75	0.026
GO:0016607	nuclear speck	cellular component	48	21	0.44	0.017
GO:0017119	Golgi transport complex	cellular component	7	6	0.86	0.012
GO:0019288	isopentenyl diphosphate biosynthetic process methylerythritol 4-phosphate pathway	biological process	5	5	1	0.01
GO:0031514	motile cilium	cellular component	15	11	0.73	0.0024
GO:0032402	melanosome transport	biological process	6	5	0.83	0.026
GO:0033116	endoplasmic reticulum-Golgi intermediate compartment membrane	cellular component	8	6	0.75	0.026
GO:0034464	BBSome	cellular component	6	5	0.83	0.026
GO:0034504	protein localization to nucleus	biological process	7	6	0.86	0.012
GO:0036064	ciliary basal body	cellular component	22	13	0.59	0.0074
GO:0043001	Golgi to plasma membrane protein transport	biological process	6	5	0.83	0.026
GO:0045335	phagocytic vesicle	cellular component	36	16	0.44	0.032
GO:0050661	NADP binding	molecular function	24	13	0.54	0.013

B.4 Supplementary Table 4 :

TABLE B.4: Significantly over- and under- represented Gene Ontology categories in the expressed transcripts - 1/2

GO	Description	Number of transcripts		Enrichment in expressed transcripts		
		In all transcritome	Expressed transcripts	P value	Q value	Odd ratio
GO:0000128	flocculation	26	0	6.2948714378938E-05	0.0038001655	0.02
GO:0001403	invasive growth in response to glucose limitation	29	0	1.20845504891598E-05	0.0008175823	0.02
GO:0007124	pseudohyphal growth	27	0	3.49398901462131E-05	0.0022850688	0.02
GO:0008986	pyruvate water dikinase activity	48	0	7.47595317948863E-09	1.12829385678128E-06	0.02
GO:0017006	protein-tetrapyrrole linkage	22	0	0.0003056537	0.014807226	0.02
GO:0031362	anchored component of external side of plasma membrane	39	0	3.0777457643036E-07	3.01926859478183E-05	0.02
GO:0006090	pyruvate metabolic process	59	1	3.65453631710494E-09	5.97516687846657E-07	0.0350940948
GO:0042025	host cell nucleus	56	1	9.50479360766085E-09	1.38136333764671E-06	0.0370158615
GO:0048255	mRNA stabilization	74	1	7.32897007684542E-12	1.68353247186171E-09	0.0278545462
GO:0060689	cell differentiation involved in salivary gland development	30	1	0.0001346898	0.0071422006	0.0703260558
GO:0002244	hematopoietic progenitor cell differentiation	44	2	0.000007688	0.0005449841	0.0970434081
GO:0016984	ribulose-bisphosphate carboxylase activity	103	2	2.36079590282151E-15	7.71980260222635E-13	0.0401936465
GO:0030115	S-layer	35	2	0.0002091669	0.0109436118	0.1235833381
GO:0030447	filamentous growth	36	2	0.0001327564	0.0071361112	0.1199406837
GO:0043205	fibril	48	2	3.09570221436162E-06	0.0002402977	0.0885811385
GO:0005201	extracellular matrix structural constituent	103	3	4.48000111313446E-14	1.25568031199569E-11	0.060905626
GO:0005581	collagen trimer	64	3	6.35564401774672E-08	7.33516091930533E-06	0.100109351
GO:0006094	gluconeogenesis	61	3	1.67820908256109E-07	1.80092043928763E-05	0.1053086072
GO:0009508	plastid chromosome	167	3	1.40608325994507E-24	7.88210101717782E-22	0.0369755671
GO:0019253	reductive pentose-phosphate cycle	103	3	4.48000111313446E-14	1.25568031199569E-11	0.060905626
GO:0042793	transcription from plastid promoter	167	3	1.40608325994507E-24	7.88210101717782E-22	0.0369755671
GO:0033018	sarcoplasmic reticulum lumen	40	4	0.0011668301	0.04769418	0.2265641647
GO:0009295	nucleoid	171	5	1.39230702863627E-22	6.82926597546093E-20	0.0608915898
GO:0043206	extracellular fibril organization	71	5	3.45208007790743E-07	3.30389322578262E-05	0.1541929003
GO:0070742	C2H2 zinc finger domain binding	61	5	7.77755121406718E-06	0.0005449841	0.1818464527
GO:0003857	3-hydroxyacyl-CoA dehydrogenase activity	6	6	0.0012575147	0.0488563136	Inf
GO:0005882	intermediate filament	55	6	0.0002536089	0.012770741	0.2495858972
GO:0006621	protein retention in ER lumen	6	6	0.0012575147	0.0488563136	Inf
GO:0047485	protein N-terminus binding	73	6	0.000001062	9.26092985244328E-05	0.1822777865
GO:0009853	photorespiration	114	7	7.72262601771425E-12	1.68353247186171E-09	0.1328206993
GO:0045095	keratin filament	76	7	0.000001925	0.000157365	0.2064886752
GO:0015030	Cajal body	97	8	1.33150470279704E-08	0.000001866	0.1827366323
GO:0016266	O-glycan processing	96	8	2.05720804141462E-08	0.000002604	0.1848254069
GO:0030286	dynein complex	65	9	0.0007945869	0.0342632869	0.3275271692
GO:0030515	snoRNA binding	10	9	0.0003137629	0.0149227048	18.4078202072
GO:0045333	cellular respiration	92	9	3.06136320321447E-07	3.01926859478183E-05	0.2205517565
GO:0005796	Golgi lumen	97	11	1.20820311403617E-06	0.000103065	0.2602281274
GO:0008266	polyU RNA binding	14	11	0.0005993795	0.0267268763	7.5004562058
GO:0009277	fungal-type cell wall	172	11	5.90781690331571E-17	2.57580816984565E-14	0.1382775797
GO:0034081	polyketide synthase complex	82	11	8.06587948418403E-05	0.0047239569	0.3155036908
GO:0048037	cofactor binding	90	11	8.48871663052793E-06	0.0005843811	0.2834090332
GO:0071013	catalytic step 2 spliceosome	131	11	5.98729644547494E-11	1.23653427642335E-08	0.1860355812
GO:0071766	Actinobacterium-type cell wall biogenesis	84	11	5.69198923088861E-05	0.0034899009	0.3068181231
GO:0005227	calcium activated cation channel activity	18	13	0.0007272698	0.0317089654	5.3192247352
GO:0015386	potassium:proton antiporter activity	17	13	0.0002893736	0.0141937753	6.6495263739
GO:0051539	4 iron 4 sulfur cluster binding	81	13	0.0008292603	0.0353697551	0.3895038331
GO:0004497	monooxygenase activity	140	14	3.17527853025324E-10	5.41730128378856E-08	0.2254889853
GO:0031225	anchored component of membrane	248	14	1.8291867320048E-25	1.43554574727737E-22	0.1205326224
GO:0004386	helicase activity	105	15	2.34902291568682E-05	0.0015622993	0.3391538401
GO:0004315	3-oxoacyl-[acyl-carrier-protein] synthase activity	106	16	4.3596621864712E-05	0.0027592443	0.3618137901
GO:0043687	post-translational protein modification	113	16	0.000007599	0.0005449841	0.3355437888
GO:0003700	transcription factor activity sequence-specific DNA binding	131	17	2.88464209775356E-07	2.97877252410131E-05	0.3030414964
GO:0005315	inorganic phosphate transmembrane transporter activity	28	18	0.0008509117	0.0359029833	3.6838156271
GO:0010119	regulation of stomatal movement	33	20	0.0012494693	0.0488563136	3.1489754244
GO:0000398	mRNA splicing via spliceosome	125	21	7.61526335459072E-05	0.0045276202	0.4108452595
GO:0019432	triglyceride biosynthetic process	37	22	0.0012175734	0.0488563136	3.0024579845
GO:0005578	proteinaceous extracellular matrix	133	23	8.54043684018456E-05	0.0047875249	0.4253728317
GO:0030246	carbohydrate binding	140	24	4.08332686769666E-05	0.0026267172	0.4207947524
GO:0005272	sodium channel activity	43	26	0.0002439249	0.0125942293	3.1322137124
GO:0031177	phosphopantetheine binding	205	27	1.06584142386834E-10	2.09118087362968E-08	0.3073277827
GO:0031930	mitochondria-nucleus signaling pathway	479	27	1.76380331823251E-48	6.92116422074436E-45	0.1187483457
GO:0005871	kinesin complex	145	28	0.0003563735	0.0166477324	0.4869664435
GO:0010019	chloroplast-nucleus signaling pathway	485	29	1.8053115168962E-47	3.54202119615034E-44	0.126424527

TABLE B.5: Significantly over- and under- represented Gene Ontology categories in the expressed transcripts - 2/2

GO	Description	Number of transcripts		Enrichment in expressed transcripts		
		In all transcritome	Expressed transcripts	P value	Q value	Odd ratio
GO:0044267	cellular protein metabolic process	150	29	0.0003156433	0.0149227048	0.487619577
GO:0005618	cell wall	218	33	4.08088415875495E-09	6.40535577558176E-07	0.3615368154
GO:0001522	pseudouridine synthesis	194	35	4.87420863564051E-06	0.0003608754	0.4470680045
GO:0007018	microtubule-based movement	203	35	8.1588598363196E-07	7.27621954493594E-05	0.4228576682
GO:0009982	pseudouridine synthase activity	209	35	1.69811560279414E-07	1.80092043928763E-05	0.4081088068
GO:0000329	fungal-type vacuole membrane	69	37	0.0004372425	0.0201851707	2.3692509849
GO:0016607	nuclear speck	222	37	0.000000076	8.51919593745301E-06	0.4055835056
GO:0003777	microtubule motor activity	216	39	1.27018869135853E-06	0.0001060472	0.447197913
GO:0045893	positive regulation of transcription DNA-templated	332	45	6.23450364231013E-16	2.38734465366974E-13	0.3160937341
GO:0006364	rRNA processing	231	53	0.0011509167	0.04769418	0.6054489228
GO:0016887	ATPase activity	268	54	4.11813783516315E-06	0.000310761	0.5119064404
GO:0005245	voltage-gated calcium channel activity	128	60	0.0012317766	0.0488563136	1.8094855636
GO:0003743	translation initiation factor activity	385	62	1.08916407197369E-13	2.84925321228318E-11	0.3869193729
GO:0008380	RNA splicing	286	64	0.0001049032	0.0057172264	0.5853329232
GO:0045454	cell redox homeostasis	144	67	0.0006720956	0.0296326173	1.7850769451
GO:0008233	peptidase activity	123	71	1.91496872361632E-08	2.50477909049014E-06	2.8072257973
GO:0006950	response to stress	350	72	3.79568163544078E-07	3.46378017150456E-05	0.5244135277
GO:0006397	mRNA processing	360	78	3.12313467776315E-06	0.0002402977	0.5603705737
GO:0042391	regulation of membrane potential	173	79	0.0004498752	0.0205268638	1.7249890043
GO:0003924	GTPase activity	445	81	3.73637508940045E-12	9.1634599067546E-10	0.4484672743
GO:0005874	microtubule	365	88	0.0002538526	0.012770741	0.6447466218
GO:0000287	magnesium ion binding	430	90	4.35320342122469E-08	5.17635461360172E-06	0.5350322453
GO:0005249	voltage-gated potassium channel activity	196	91	8.79047567186626E-05	0.0048582854	1.7805550873
GO:0071805	potassium ion transmembrane transport	177	94	2.93156398234884E-08	3.59483033335526E-06	2.3311105719
GO:0005615	extracellular space	395	98	0.0005257192	0.0237117498	0.6696748682
GO:0044822	polyA RNA binding	422	108	0.0011623952	0.04769418	0.6982113437
GO:0006355	regulation of transcription DNA-templated	607	154	5.37344742286597E-05	0.0033468901	0.6879103624
GO:0005525	GTP binding	781	177	2.18656592086564E-10	4.08575460641751E-08	0.5886442893
GO:0006351	transcription DNA-templated	738	185	2.98421186986049E-06	0.0002389806	0.6751768206
GO:0009507	chloroplast	1277	207	2.50935630236404E-43	3.28223804349216E-40	0.3774136294
GO:0003677	DNA binding	1432	258	2.15612225830409E-38	2.11515593539631E-35	0.4285831756
GO:0005576	extracellular region	1371	318	6.69235249499671E-16	2.38734465366974E-13	0.5993630847
GO:0003723	RNA binding	1195	331	0.000084755	0.0047875249	0.772260874
GO:0005509	calcium ion binding	1096	416	0.0002813621	0.0139755065	1.2651150872
GO:0005634	nucleus	3755	1131	8.4525927898712E-05	0.0047875249	0.8584633044
GO:0005524	ATP binding	3852	1132	3.72528890208847E-07	3.46378017150456E-05	0.8224555434
GO:0016021	integral component of membrane	3846	1415	1.41139799588738E-08	1.90976749512485E-06	1.2352963734
GO:0005737	cytoplasm	4900	1428	2.65760060861215E-10	4.74019308554276E-08	0.8009273276

B.5 Supplementary Table 5 :

TABLE B.6: Table of the annotated transcripts displaying the top ten percent higher log Fold Change between their lower and higher relative expression in each module -1/2

Contig	Module	Homolog	Name
comp101507 c0 seq1	Module A	CCNB1 MEDSV	G2/mitotic-specific cyclin-1
comp122768 c0 seq1	Module A	CBPC2 XENTR	Cytosolic carboxypeptidase
comp94039 c1 seq1	Module A	CND1 MOUSE	Condensin complex subunit
comp114175 c0 seq1	Module A	DED1 ASHGO	ATP-dependent RNA helicase
comp131915 c1 seq1	Module A	DHB4 DROME	Peroxisomal multifunctional enzyme type
comp133927 c0 seq1	Module A	FA5V OXYMI	Venom prothrombin activator omicarin-C
comp133976 c0 seq1	Module A	FND3A MOUSE	Fibronectin type-III domain-containing protein 3A
comp118087 c0 seq1	Module A	GSPB STRGN	Platelet binding protein GspB
comp127519 c0 seq1	Module A	HSOP3 ARATH	Hsp70-Hsp90 organizing protein
comp132323 c1 seq1	Module A	MYH10 BOVIN	Myosin-10
comp133962 c0 seq1	Module A	MYOF MOUSE	Myoferlin
comp125588 c2 seq1	Module A	NEUL4 HUMAN	Neuralized-like protein 4
comp126007 c0 seq2	Module A	PPSC MYCBO	Phthiocerol/phenolphthiocerol synthesis polyketide synthase type I PpsC
comp131138 c1 seq1	Module A	RENT1 DICDI	Regulator of nonsense transcripts
comp132356 c3 seq1	Module A	RLIG BPT4	T4 RNA ligase
comp122760 c0 seq1	Module A	SEPR THESR	Extracellular serine proteinase
comp131191 c0 seq1	Module A	UNC22 CAEEL	Twitchin
comp133573 c0 seq1	Module A	Y2354 DICDI	Probable serine/threonine-protein kinase
comp122878 c1 seq1	Module A	YS027 HUMAN	Proline-rich protein 36
comp131433 c1 seq1	Module B	AB2C ARATH	ABC transporter C family member
comp133816 c1 seq1	Module B	ABCBB RABIT	Bile salt export pump
comp97984 c0 seq1	Module B	ADCYA HUMAN	Adenylate cyclase type 10
comp127476 c0 seq1	Module B	ANK3 HUMAN	Ankyrin-3
comp80689 c0 seq4	Module B	CABO DORPE	Squidulin
comp124869 c1 seq1	Module B	CAC1S RABIT	Voltage-dependent L-type calcium channel subunit alpha-1S
comp128375 c0 seq1	Module B	FMN CHICK	Formin
comp86039 c0 seq1	Module B	GORK ARATH	Potassium channel GORK
comp134604 c0 seq1	Module B	HHP1 SCHPO	Casein kinase I homolog hhp1
comp119744 c0 seq1	Module B	KCNH5 HUMAN	Potassium voltage-gated channel subfamily H member
comp123796 c0 seq1	Module B	LGRC BREPA	Linear gramicidin synthase subunit C
comp96876 c0 seq2	Module B	LVHK1 ERYLH	Blue-light-activated histidine kinase
comp127509 c2 seq1	Module B	MCAS MYCBO	Mycocerosic acid synthase
comp96681 c0 seq1	Module B	NAC1 CAVPO	Sodium/calcium exchanger
comp114097 c0 seq1	Module B	NAS13 CAEEL	Zinc metalloproteinase nas-13
comp68250 c0 seq1	Module B	NPC1 HUMAN	Niemann-Pick C1 protein
comp117512 c0 seq1	Module B	PI5K4 ARATH	Phosphatidylinositol 4-phosphate 5-kinase
comp120585 c1 seq1	Module B	PKHL1 MOUSE	Fibrocystin-L
comp109465 c4 seq1	Module B	PKSL BACSU	Polyketide synthase PksL
comp99315 c1 seq1	Module B	PKSM BACSU	Polyketide synthase PksM
comp111100 c0 seq1	Module B	PPSD MYCBO	Phthiocerol/phenolphthiocerol synthesis polyketide synthase type I PpsD
comp104887 c0 seq1	Module B	S35B2 HUMAN	Adenosine 3-phospho 5-phosphosulfate transporter
comp122550 c0 seq1	Module B	SAAR2 ACRMI	Skeletal aspartic acid-rich protein
comp115758 c0 seq3	Module B	Y182 SYNY3	Uncharacterized ABC transporter ATP-binding protein
comp117037 c0 seq3	Module B	Y9086 DICDI	Uncharacterized abhydrolase domain-containing protein
comp57677 c0 seq1	Module C	AARA DICDI	Protein aardvark
comp128041 c0 seq1	Module C	AGUA STRCO	Putative agmatine deiminase
comp135745 c0 seq1	Module C	C3H59 ORYSJ	Zinc finger CCCH domain-containing protein 59
comp71310 c0 seq1	Module C	CAT2 CLOK5	4-hydroxybutyrate coenzyme A transferase
comp129039 c0 seq1	Module C	CLCE ARATH	Chloride channel protein CLC-e
comp127883 c0 seq1	Module C	COMI DICDI	Comitin
comp122712 c0 seq1	Module C	CROCC MOUSE	Rootletin
comp126637 c0 seq1	Module C	DEK1 ORYSJ	Calpain-type cysteine protease

TABLE B.7: Table of the annotated transcripts displaying the top ten percent higher log Fold Change between their lower and higher relative expression in each module -2/2

Contig	Module	Homolog	Name
comp130954 c0 seq1	Module F	CDPKI ARATH	Calcium-dependent protein kinase 18
comp123725 c0 seq1	Module F	CTSR1 MOUSE	Cation channel sperm-associated protein
comp106095 c0 seq1	Module F	CYB5B HUMAN	Cytochrome b5 type B
comp103245 c0 seq1	Module F	ESYT1 RAT	Extended synaptotagmin-1
comp90220 c0 seq1	Module F	FAEB PIREQ	Feruloyl esterase B
comp89527 c0 seq1	Module F	FBT6 ARATH	Probable folate-biopterin transporter 6
comp106332 c1 seq1	Module F	GCY36 CAEEL	Soluble guanylate cyclase gcy-36
comp96525 c0 seq3	Module F	HCN3 HUMAN	Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 3
comp88700 c0 seq1	Module F	KCNB2 MOUSE	Potassium voltage-gated channel subfamily B member
comp76244 c0 seq1	Module F	KCNH2 HUMAN	Potassium voltage-gated channel subfamily H member
comp116500 c0 seq1	Module F	MPK4 ORYSJ	Mitogen-activated protein kinase
comp97292 c0 seq1	Module F	NCKX2 RAT	Sodium/potassium/calcium exchanger
comp72671 c0 seq1	Module F	NLRC3 HUMAN	Protein NLRC3
comp49403 c0 seq1	Module F	P2C35 ARATH	Probable protein phosphatase 2C 35
comp89657 c0 seq1	Module F	PDE9A HUMAN	High affinity cGMP-specific 3,5-cyclic phosphodiesterase
comp96181 c0 seq1	Module F	PI5K1 ARATH	Phosphatidylinositol 4-phosphate 5-kinase
comp96690 c1 seq1	Module F	PPSD MYCTO	Phthiocerol synthesis polyketide synthase type I PpsD
comp112514 c0 seq1	Module F	PPT1 SCHPO	Serine/threonine-protein phosphatase T
comp66508 c0 seq1	Module F	RHOM DROME	Protein rhomboid
comp101963 c0 seq1	Module F	SC4AA TETNG	Sodium channel protein type 4 subunit alpha A
comp108531 c0 seq1	Module F	SPCS2 ARATH	Probable signal peptidase complex subunit
comp128004 c0 seq1	Module F	SUC4 ARATH	Sucrose transport protein SUC4
comp57604 c0 seq1	Module F	SVOP HUMAN	Synaptic vesicle 2-related protein
comp107040 c0 seq1	Module F	TNI3K RAT	Serine/threonine-protein kinase
comp132048 c0 seq1	Module F	TRAF2 HUMAN	TNF receptor-associated factor
comp114154 c0 seq1	Module F	TRPM2 HUMAN	Transient receptor potential cation channel subfamily M member
comp111617 c0 seq1	Module F	Y1024 SYNY3	UPF0187 protein
comp129942 c0 seq1	Module F	Y1931 DICDI	Putative leucine-rich repeat-containing protein DDB
comp101309 c0 seq1	Module F	Y4233 RHOPA	Putative potassium channel protein
comp124621 c0 seq1	Module F	YA7B SCHPO	Uncharacterized protein C24H6.11c
comp40669 c0 seq1	Module F	YBQ2 SCHPO	Uncharacterized protein C115.02c
comp119194 c0 seq1	Module F	YKRP BACSU	Putative membrane-bound acyltransferase Ykr
comp89603 c0 seq1	Module G	ARSB BACSU	Arsenite resistance protein ArsB
comp30530 c0 seq1	Module G	ARSM HALSA	Putative arsenite methyltransferase
comp85921 c1 seq1	Module G	YAGC SCHPO	Uncharacterized protein C12G12.12
comp106096 c0 seq1	Module H	CML39 ARATH	Calcium-binding protein CML39
comp26328 c0 seq1	Module H	GLRA2 HUMAN	Glycine receptor subunit alpha-2
comp70238 c0 seq1	Module H	IF5A1 SOLLC	Eukaryotic translation initiation factor 5A-1
comp95598 c0 seq1	Module H	IF5A4 SOLTU	Eukaryotic translation initiation factor 5A-4
comp105736 c0 seq1	Module H	LRRC1 MOUSE	Leucine-rich repeat-containing protein 1
comp44323 c0 seq1	Module H	PKD2 ORYLA	Polycystin-2
comp120703 c0 seq1	Module H	PLCX1 MOUSE	PI-PLC X domain-containing protein 1
comp78143 c0 seq1	Module H	RL25 AZOC5	50S ribosomal protein L25
comp8281 c0 seq1	Module H	S35B2 DICDI	Adenosine 3-phospho 5-phosphosulfate transporter
comp86202 c0 seq1	Module H	SGT1A ARATH	Protein SGT1 homolog A (AtSGT1a)
comp15150 c0 seq1	Module H	TCR8 PASMD	Tetracycline resistance protein, class H
comp101989 c0 seq1	Module H	TIP12 ARATH	Aquaporin TIP1-2
comp77002 c1 seq1	Module H	TUB HUMAN	Tubby protein homolog
comp65795 c0 seq1	Module H	UBE2W DICDI	Probable ubiquitin-conjugating enzyme E2 W
comp123720 c0 seq1	Module H	Y2449 MYCTU	Putative trans-acting enoyl reductase Rv2449c
comp104003 c0 seq1	Module C	EMAL6 HUMAN	Echinoderm microtubule-associated protein-like
comp120339 c1 seq1	Module C	GABP1 BOVIN	GA-binding protein subunit beta-1
comp70790 c0 seq1	Module C	GLD2 BOVIN	Poly(A) RNA polymerase GLD2
comp73798 c0 seq1	Module C	GRAP1 MOUSE	GRIP1-associated protein
comp125297 c2 seq1	Module C	HEAT2 HUMAN	Dynein assembly factor
comp84522 c0 seq1	Module C	HMGT ONCMY	High mobility group-T protein
comp116234 c0 seq1	Module C	K125 ARATH	Kinesin-like protein KIN-5C
comp126891 c1 seq1	Module C	KLH17 HUMAN	Kelch-like protein 17
comp124538 c0 seq1	Module C	MSI2H HUMAN	RNA-binding protein Musashi homolog
comp126161 c0 seq1	Module C	MYH14 HUMAN	Myosin-14
comp131809 c0 seq2	Module C	PKWA THECU	Probable serine/threonine-protein kinase PkwA
comp132725 c0 seq1	Module C	PLEC MOUSE	Plectin
comp124122 c0 seq1	Module C	PPTC7 MOUSE	Protein phosphatase PTC7 homolog
comp113625 c0 seq1	Module C	RAI14 MOUSE	Ankycorbin
comp99952 c0 seq1	Module C	RIR2 NEUCR	Ribonucleoside-diphosphate reductase small chain
comp115595 c0 seq1	Module C	SEC63 HUMAN	Translocation protein SEC63 homolog
comp72790 c0 seq1	Module C	SGT2 USTMA	Small glutamine-rich tetratricopeptide repeat-containing protein 2
comp133882 c0 seq2	Module C	SWD1 SCHPO	Set1 complex component swd1
comp127017 c1 seq1	Module C	VWKA DICDI	Alpha-protein kinase vwkA
comp135115 c0 seq1	Module C	WDR65 MOUSE	Cilia- and flagella-associated protein 57
comp20140 c0 seq1	Module C	WDR75 DANRE	WD repeat-containing protein 75
comp93546 c0 seq1	Module C	Y045 METMA	Putative ankyrin repeat protein MM
comp134656 c0 seq1	Module D	ANK1 MOUSE	Ankyrin-1
comp127241 c1 seq1	Module D	CROCC HUMAN	Rootletin
comp99488 c0 seq1	Module D	CSMD3 HUMAN	CUB and sushi domain-containing protein
comp82023 c0 seq1	Module D	DIN1 RAPSA	Senescence-associated protein DIN1
comp95052 c0 seq1	Module D	GST1 ASCSU	Glutathione S-transferase
comp126538 c0 seq1	Module D	PARP SARPE	Poly [ADP-ribose] polymerase
comp98801 c0 seq1	Module D	RAD50 METKA	DNA double-strand break repair Rad50 ATPase
comp108880 c0 seq1	Module E	AK1 DICDI	Alpha-protein kinase
comp128295 c0 seq1	Module E	PKHL1 HUMAN	Fibrocystin-L
comp103033 c0 seq1	Module F	AKT1 ARATH	Potassium channel AKT1
comp101881 c0 seq4	Module F	ANKH1 HUMAN	Ankyrin repeat and KH domain-containing protein
comp78812 c1 seq1	Module F	ANR31 HUMAN	Putative ankyrin repeat domain-containing protein 31
comp116691 c0 seq22	Module F	BACR HALAR	Cruxrhodopsin-1
comp43977 c0 seq1	Module F	CAC1A APIME	Voltage-dependent calcium channel type A subunit alpha-1
comp25437 c0 seq1	Module F	CAC1G RAT	Voltage-dependent T-type calcium channel subunit alpha-1G
comp95837 c0 seq1	Module F	CATR5 PARTE	Caltractin ICL1e
comp123994 c0 seq1	Module F	CDPK8 ARATH	Calcium-dependent protein kinase

B.6 Supplementary figure 1 :

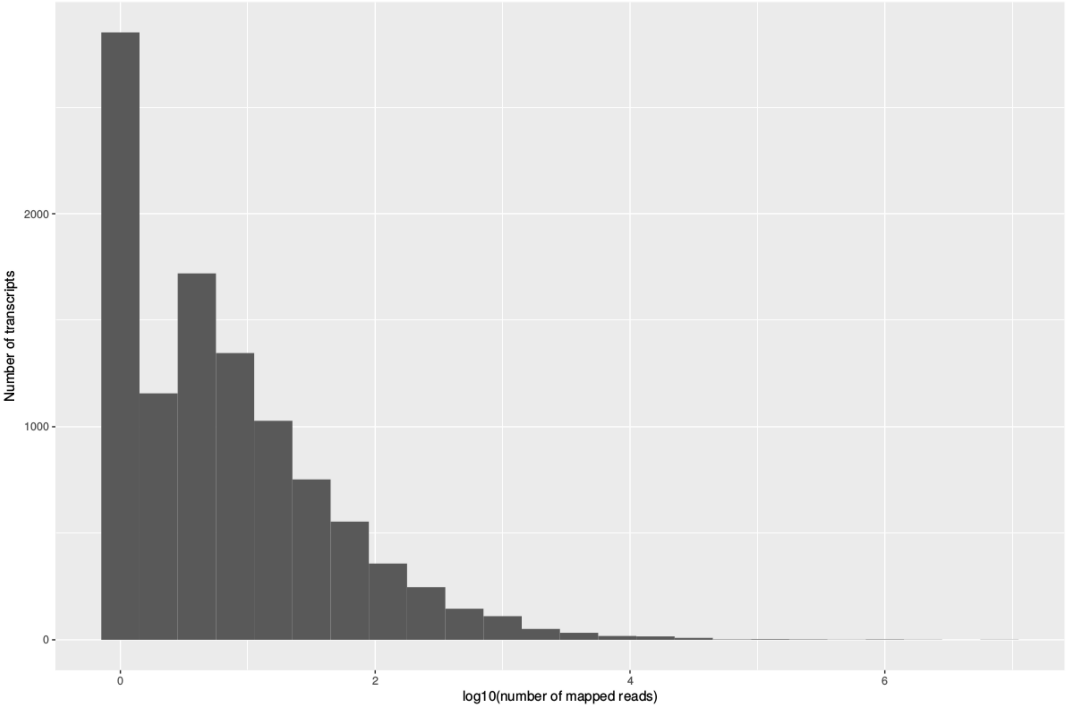


FIGURE B.1: *Histograms of the number of transcripts to which reads from 24 microphytoplanktonic species (13 dinoflagellates, 10 diatoms and 1 Labyrinthulomycetes) detected in the water at the time of the sampling and available in the MMETSP database aligned to* A. minutum *transcriptome. Reads were aligned using Bowtie 2 (V. 2.2.4) using the sensitive sets of parameters.*

B.7 Supplementary figure 2 :

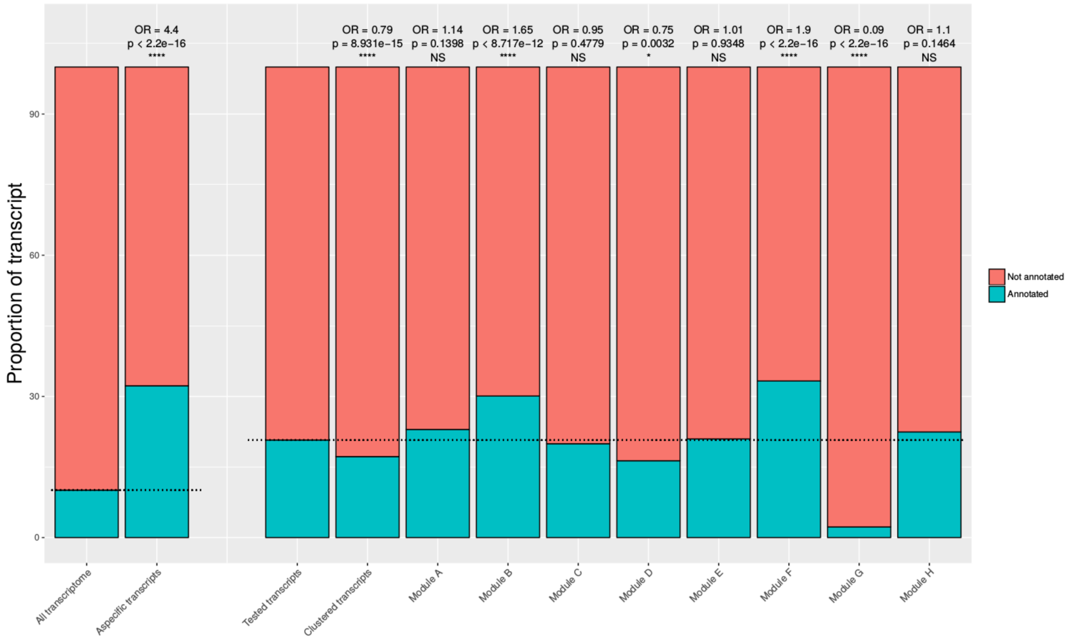


FIGURE B.2: *Barplot of the proportion of annotated transcripts in different sets of the dataset. Fisher exact test results are indicated on the top of each tested set. Over representation of annotation among the aspecific transcripts was tested against the whole transcriptome. Over representation of annotation among the clustered transcripts and transcripts in modules A to H was tested against the expressed transcripts. Stars represent the significance level.*

B.8 Supplementary figure 3 :

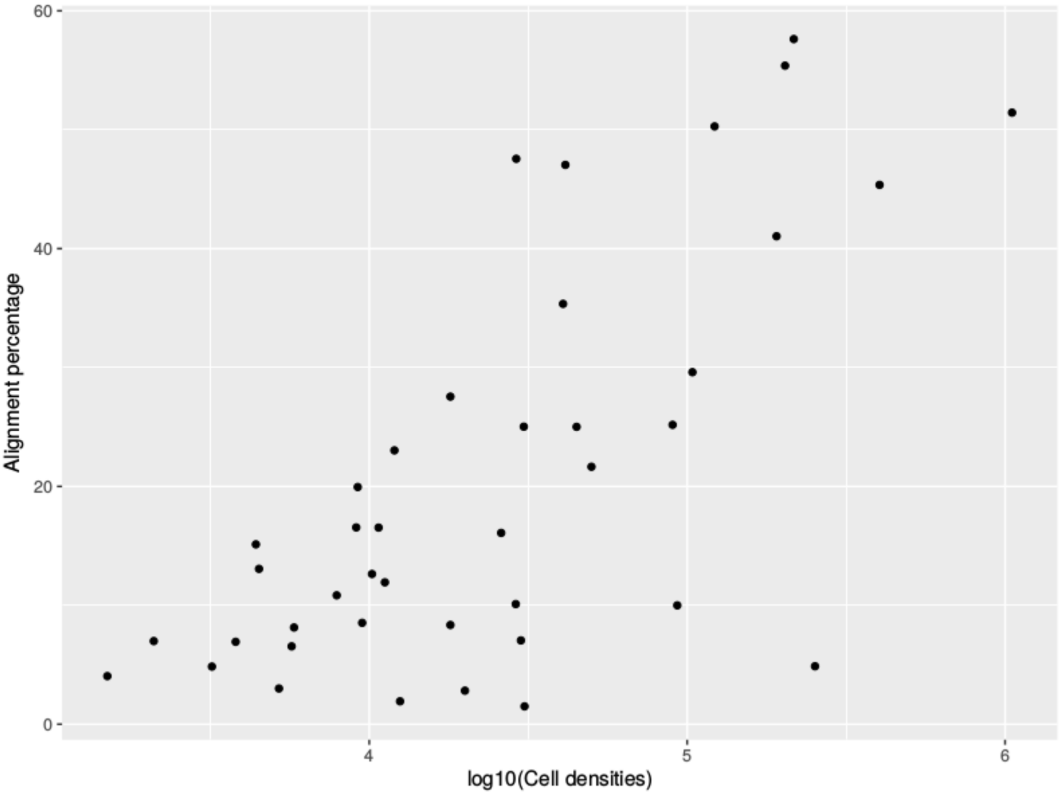


FIGURE B.3: *Plot representing the proportion of aligned reads to* A. minutum *transcriptome according to log10(cell density) in each sample (black dots).*

B.9 Supplementary figure 4 :

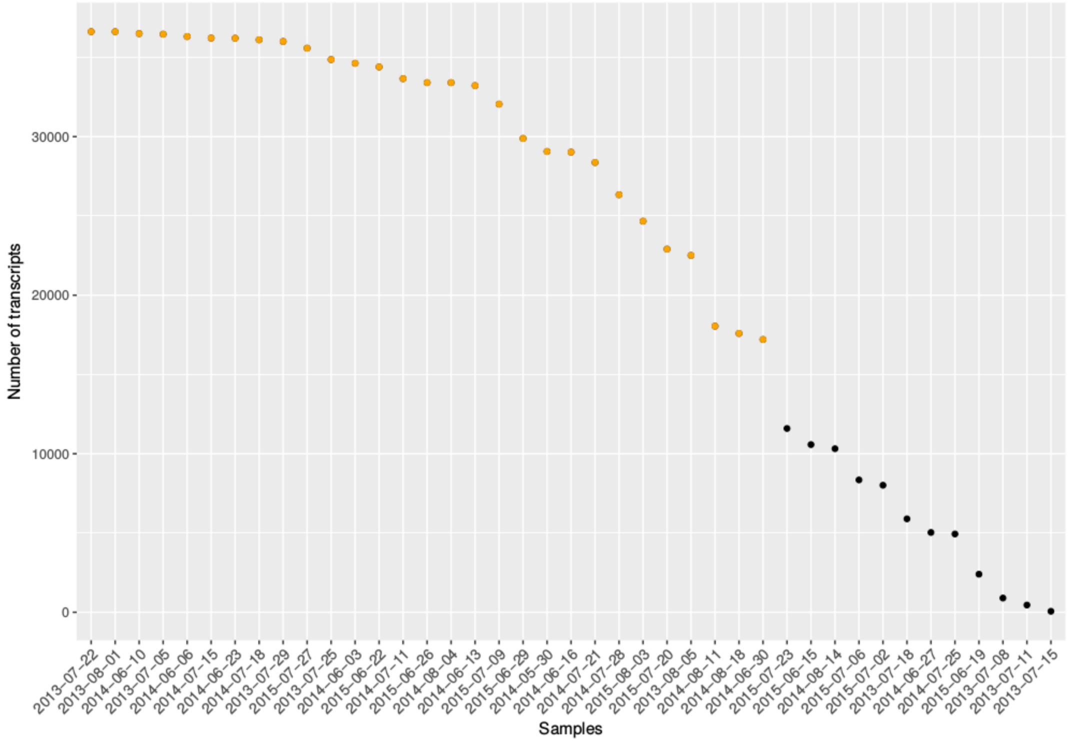


FIGURE B.4: *Number of transcripts to which at least 40 environmental reads aligned. Orange dots indicate samples with sufficient coverage and then considered for the analysis.*

B.10 Supplementary figure 5 :

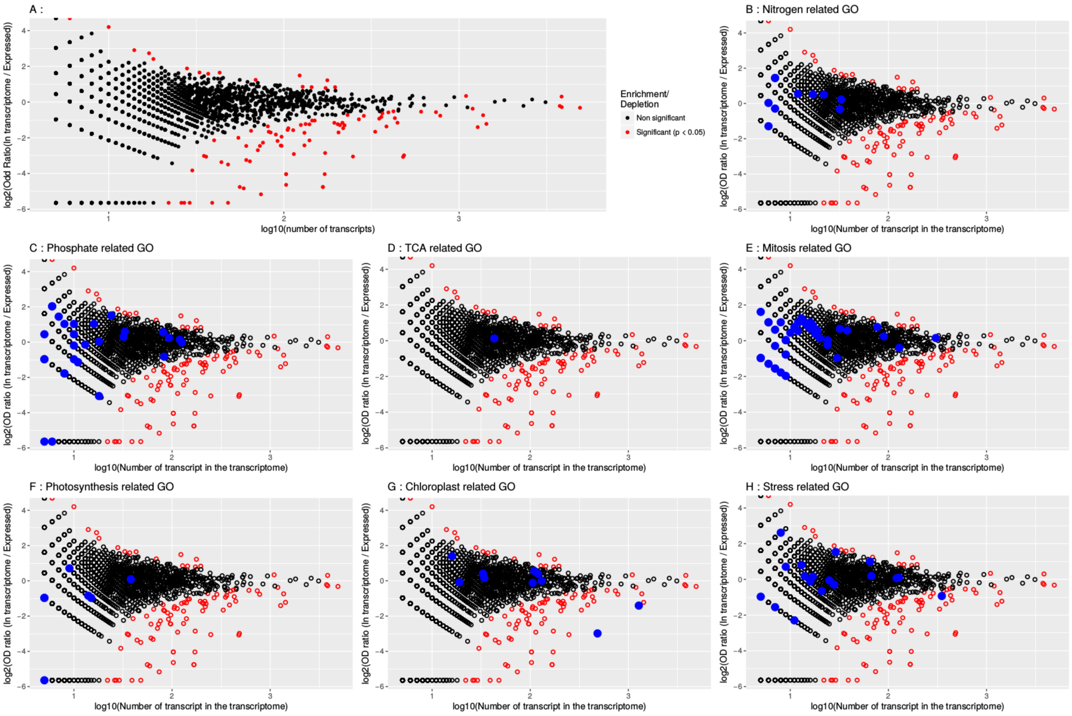


FIGURE B.5: *MA plot like representation of significantly (red dots) over and under represented functions in the expressed transcripts compared to the overall transriptome (A). B to H panels highlight a priori biologically relevant specific functions.*

B.11 Supplementary figure 6 :

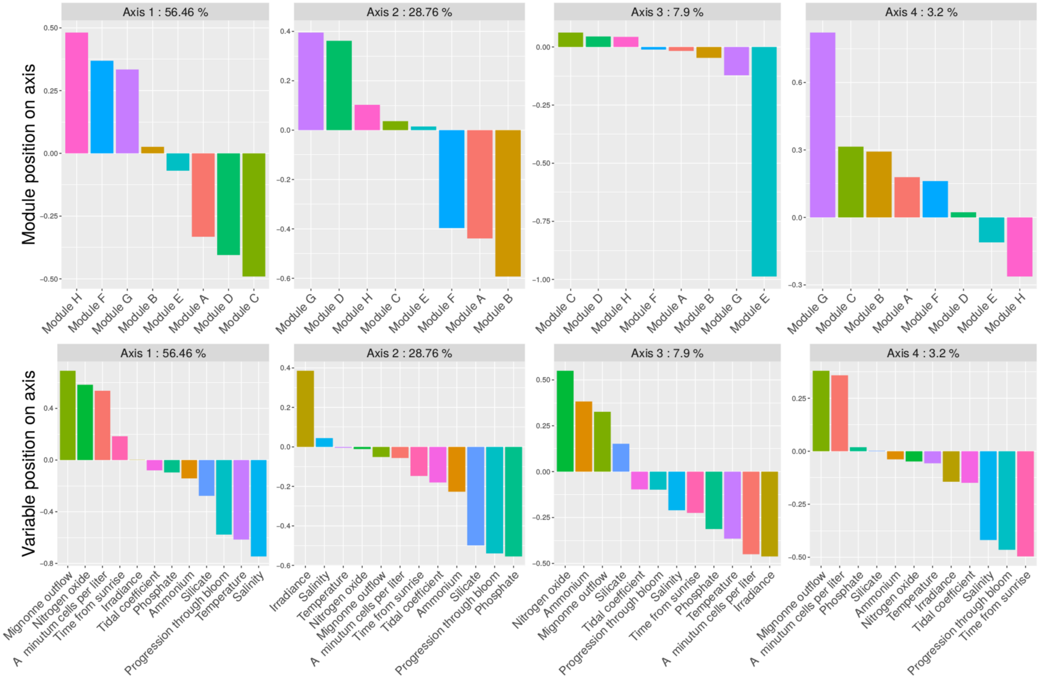


FIGURE B.6: *Figure representing the results of a RDA analysis led to explore the relations between expression module eigengenes dynamics and 12 environmental variables. Each panel represents the contribution of each expression module eigengene and environmental variable to each of the four first axis. These four axis represent more than 96% of the observed variance.*

Appendix C

Supplementary material for chapter 4

C.1 Supplementary table 1 :

TABLE C.1: Pairwise Fst values between the temporal samples

	05/07/13	22/07/13	25/07/13	29/07/13	01/08/13	30/05/14	03/06/14	06/06/14	10/06/14	13/06/14	16/06/14
05/07/13	0.0000	0.0015	0.0016	0.0017	0.0013	0.0082	0.0079	0.0071	0.0061	0.0055	0.0038
22/07/13		0.0000	0.0010	0.0008	0.0006	0.0080	0.0072	0.0063	0.0054	0.0052	0.0036
25/07/13			0.0000	0.0012	0.0009	0.0086	0.0084	0.0078	0.0068	0.0058	0.0036
29/07/13				0.0000	0.0008	0.0084	0.0084	0.0073	0.0060	0.0057	0.0035
01/08/13					0.0000	0.0079	0.0075	0.0066	0.0052	0.0049	0.0033
30/05/14						0.0000	0.0102	0.0086	0.0080	0.0107	0.0076
03/06/14							0.0000	0.0061	0.0057	0.0107	0.0063
06/06/14								0.0000	0.0042	0.0097	0.0054
10/06/14									0.0000	0.0081	0.0042
13/06/14										0.0000	0.0063
16/06/14											0.0000

	23/06/14	11/07/14	15/07/14	18/07/14	04/08/14	22/06/15	26/06/15	29/06/15	09/07/15	20/07/15	27/07/15	03/08/15
05/07/13	0.0070	0.0053	0.0042	0.0102	0.0030	0.0124	0.0129	0.0187	0.0123	0.0245	0.0074	0.0183
22/07/13	0.0065	0.0049	0.0039	0.0097	0.0030	0.0110	0.0113	0.0173	0.0112	0.0229	0.0065	0.0171
25/07/13	0.0077	0.0051	0.0043	0.0104	0.0031	0.0126	0.0124	0.0190	0.0121	0.0246	0.0073	0.0184
29/07/13	0.0070	0.0050	0.0040	0.0100	0.0029	0.0122	0.0128	0.0188	0.0123	0.0249	0.0072	0.0185
01/08/13	0.0063	0.0049	0.0036	0.0094	0.0026	0.0116	0.0120	0.0180	0.0115	0.0233	0.0068	0.0174
30/05/14	0.0097	0.0109	0.0089	0.0155	0.0089	0.0203	0.0208	0.0281	0.0200	0.0341	0.0145	0.0256
03/06/14	0.0076	0.0100	0.0089	0.0135	0.0086	0.0195	0.0203	0.0271	0.0196	0.0329	0.0140	0.0248
06/06/14	0.0060	0.0089	0.0074	0.0121	0.0075	0.0190	0.0191	0.0257	0.0180	0.0314	0.0128	0.0235
10/06/14	0.0054	0.0077	0.0062	0.0109	0.0058	0.0171	0.0177	0.0241	0.0169	0.0295	0.0118	0.0216
13/06/14	0.0097	0.0082	0.0070	0.0130	0.0062	0.0168	0.0175	0.0250	0.0163	0.0302	0.0115	0.0228
16/06/14	0.0059	0.0057	0.0043	0.0105	0.0034	0.0151	0.0153	0.0221	0.0151	0.0290	0.0099	0.0211
23/06/14	0.0000	0.0087	0.0077	0.0125	0.0072	0.0182	0.0186	0.0247	0.0176	0.0307	0.0125	0.0227
11/07/14		0.0000	0.0063	0.0128	0.0059	0.0165	0.0171	0.0236	0.0165	0.0294	0.0112	0.0224
15/07/14			0.0000	0.0110	0.0045	0.0152	0.0156	0.0224	0.0152	0.0276	0.0098	0.0203
18/07/14				0.0000	0.0108	0.0212	0.0222	0.0286	0.0211	0.0346	0.0155	0.0271
04/08/14					0.0000	0.0146	0.0144	0.0214	0.0140	0.0266	0.0093	0.0199
22/06/15						0.0000	0.0212	0.0279	0.0213	0.0348	0.0160	0.0271
26/06/15							0.0000	0.0280	0.0215	0.0347	0.0163	0.0283
29/06/15								0.0000	0.0285	0.0427	0.0221	0.0355
09/07/15									0.0000	0.0347	0.0156	0.0269
20/07/15										0.0000	0.0282	0.0419
27/07/15											0.0000	0.0211
03/08/15												0.0000

Appendix D

Supplementary material for chapter 5

D.1 Supplementary Table 1 :

TABLE D.1: Number of species per genus in the two meta-references

Genus	MetaRef1	MetaRef2
Alexandrium	7	1
Amphidinium	2	1
Amphiprora	2	1
Aplanochytrium	2	1
Bigelowiella	2	1
Bolidomonas	3	1
Cafeteria	2	1
Chaetoceros	8	3
Chlamydomonas	4	1
Chrysoreinhardia	2	1
Corethron	2	1
Cryptomonas	2	1
Erythrolobus	2	1
Euplotes	2	1
Florenciella	3	1
Geminigera	2	1
Goniomonas	2	1
Haptolina	2	1
Hemiselmis	4	1
Heterocapsa	3	2
Isochrysis	3	3
Leptocylindrus	2	1
Lotharella	2	1
Mantoniella	2	1
Micromonas	7	3
Ochromonas	3	1
Odontella	2	1
Ostreococcus	2	2
Paramoeba	2	1
Paraphysomonas	3	2
Pavlova	3	1
Perkinsus	2	1
Phaeocystis	3	1
Picochlorum	2	1
Prasinoderma	2	1
Proboscia	2	1
Prorocentrum	3	2
Prymnesium	2	1
Pseudo-nitzschia	6	2
Pseudokeronopsis	2	1
Pycnococcus	2	2
Pyramimonas	3	1
Rhodomonas	3	1
Scrippsiella	2	2
Skeletonema	6	1
Strombidinopsis	2	1
Strombidium	2	1
Symbiodinium	2	1
Tetraselmis	4	1
Thalassionema	2	2
Thalassiosira	9	1

D.2 Supplementary Table 2 :

TABLE D.2: Overall number of reads from the metatranscriptome aligning to the 12 selected species

	Tot aligned reads	prop aligned
Alexandrium minutum	285481571	0.17399194
Chaetoceros curvisetus	50158982	0.0305703046
Prorocentrum micans	1656583	0.0010096347
Chaetoceros cf. *neogracile*	15100529	0.0092032923
Lankesteria abbotti	5124594	0.0031232771
Micromonas CCMP1646	9718261	0.0059229711
Ostreococcus lucimarinus	4881082	0.0029748643
Helicotheca tamesis	4487332	0.0027348862
Leptocylindrus aporus	7214356	0.0043969206
Skeletonema dohrnii	9344981	0.0056954688
Gonyaulax spinifera	12427707	0.0075742922
Thalassiosira punctigera	7271933	0.0044320119

D.3 Supplementary Table 3 :

TABLE D.3: Per date number of reads from the metatranscriptome aligning to the 12 selected species

Species	05/07/13	08/07/13	11/07/13	15/07/13	18/07/13	22/07/13	25/07/13	29/07/13	01/08/13	05/08/13	30/05/14	03/06/14
Alexandrium minutum	13197561	605020	530960	162628	1447714	25248402	7140503	8270841	18861852	3398684	5310819	10131899
Chaetoceros cf. *neogracile*	259431	341163	477321	178734	446553	876899	1467285	232686	75087	77720	188591	59626
Chaetoceros curvisetus	6176379	9235205	6676511	1784890	3516401	555944	430260	82434	38916	228549	159899	45273
Gonyaulax spinifera	314012	51959	72059	31099	129158	594277	163704	154638	364994	100880	586109	3239768
Helicotheca tamesis	41940	30564	41679	14664	27804	38503	50878	19067	24783	1015126	42796	18689
Lankesteria abbotti	239630	13893	16116	5761	32114	348773	114066	128426	265056	50191	105274	181704
Leptocylindrus aporus	38711	27345	37763	15920	29356	35033	47617	11444	7760	26309	3269579	2015407
Micromonas CCMP1646	175629	22976	185564	8635	4828758	77701	23819	94716	78514	130490	213616	246716
Ostreococcus lucimarinus	185455	12639	24713	3780	693220	9866	6466	41574	81460	653206	140497	103298
Prorocentrum micans	18133	12250	11063	5909	10298	17423	9547	11286	13870	18585	11293	10538
Skeletonema dohrnii	213348	75578	109453	29562	39381	60005	94723	69006	136179	873538	97792	60269
Thalassiosira punctigera	48437	28932	37940	12391	24568	31568	34822	47600	250832	398037	34604	31115
Total reads	65303628	35984282	44752862	14585518	39148528	62247310	30628448	21454728	42469344	40086440	38092194	49147430

Species	06/06/14	10/06/14	13/06/14	16/06/14	23/06/14	27/06/14	30/06/14	11/07/14	15/07/14	18/07/14	21/07/14	25/07/14
Alexandrium minutum	20262052	26559454	6494453	5426511	15507035	1104038	2738753	7061925	10247754	9866022	4877916	1077703
Chaetoceros cf. *neogracile*	12226	25589	106183	14931	169947	1090002	239745	44629	50770	100126	511653	450003
Chaetoceros curvisetus	7899	14779	42691	6698	62806	390384	120508	279396	42912	128140	278338	1200255
Gonyaulax spinifera	1137664	972902	268120	189105	543734	77227	107824	215780	354868	238256	143888	38458
Helicotheca tamesis	4795	6501	9398	2870	16853	70895	37660	38526	36779	98705	189887	116168
Lankesteria abbotti	362966	534855	111145	124973	305135	25406	57575	122310	173551	188922	86680	23141
Leptocylindrus aporus	217540	12523	13510	3950	13263	52833	30283	26152	21985	26214	46337	23198
Micromonas CCMP1646	125526	210921	163542	28470	46791	73848	90986	194710	62509	24084	28509	31877
Ostreococcus lucimarinus	37761	42837	20024	3916	42509	54647	56864	274678	65761	27734	246203	176375
Prorocentrum micans	14384	19578	9597	4533	107975	65903	58887	125479	341421	88431	25060	21048
Skeletonema dohrnii	12823	28145	66957	16838	171125	833127	658289	136436	172672	91120	349636	108433
Thalassiosira punctigera	12282	15865	15920	5613	38499	116316	113005	72490	80050	159623	776109	200162
Total reads	43693870	56238032	22114290	13694868	46607830	43882724	40252144	54439528	62855938	45520194	59247556	20915104

Species	28/07/14	04/08/14	11/08/14	14/08/14	18/08/14	15/06/15	19/06/15	22/06/15	26/06/15	29/06/15	02/07/15	06/07/15
Alexandrium minutum	4549826	6395134	2815396	1761383	2734318	1690766	893932	8691934	7695705	5674073	1499925	1509661
Chaetoceros cf. *neogracile*	620082	200691	87683	221256	323014	93192	41829	34737	41564	756059	698605	1676682
Chaetoceros curvisetus	902686	104252	72872	147746	286558	444037	88958	68903	259586	2536532	10834055	903882
Gonyaulax spinifera	145672	145247	99074	64471	82457	62757	44317	277589	221276	202901	78929	60367
Helicotheca tamesis	95671	66272	66463	199968	319400	33426	17089	17079	41860	163060	82211	78486
Lankesteria abbotti	86430	117507	55563	34180	51410	35942	25093	150109	140080	106513	35052	30565
Leptocylindrus aporus	47424	18628	29842	38950	41696	673772	39929	18646	18199	39714	31325	40563
Micromonas CCMP1646	203692	65984	74357	33708	37595	91776	121939	252695	24840	696894	433280	19744
Ostreococcus lucimarinus	59903	84240	196954	24929	84700	69192	35266	87861	70508	475583	39890	24753
Prorocentrum micans	97735	40877	58195	52664	27376	18294	14886	34337	27724	48477	31127	23523
Skeletonema dohrnii	427516	53547	333401	502388	1672075	73430	89267	109492	51646	315046	157861	137493
Thalassiosira punctigera	183714	71283	324073	301633	593866	61661	131497	118877	97181	228405	65126	57754
Total reads	55467816	31549506	29645796	26460440	40927666	29980804	20887180	43840340	40772836	51669872	44018146	25538380

Species	09/07/15	20/07/15	23/07/15	27/07/15	03/08/15	Sum
Alexandrium minutum	6878598	3544639	1809893	17603625	4202264	285481571
Chaetoceros cf. *neogracile*	2211835	231093	161366	76058	127883	15100529
Chaetoceros curvisetus	1004459	218358	483170	59087	238374	50158982
Gonyaulax spinifera	206213	95143	67783	356601	126427	12427707
Helicotheca tamesis	184389	264640	358993	234412	268383	4487332
Lankesteria abbotti	130841	58632	39133	329829	80052	5124594
Leptocylindrus aporus	66107	16589	17671	7624	17645	7214356
Micromonas CCMP1646	130222	68710	45678	75123	173117	9718261
Ostreococcus lucimarinus	169950	237333	61702	50182	102653	4881082
Prorocentrum micans	42265	8638	16570	21810	59594	1656583
Skeletonema dohrnii	127722	273789	196704	44638	274531	9344981
Thalassiosira punctigera	193921	1015934	986095	90179	163954	7271933
Total reads	65895762	29914978	31669180	46672736	32500460	1640774688

D.4 Supplementary figure 1 :

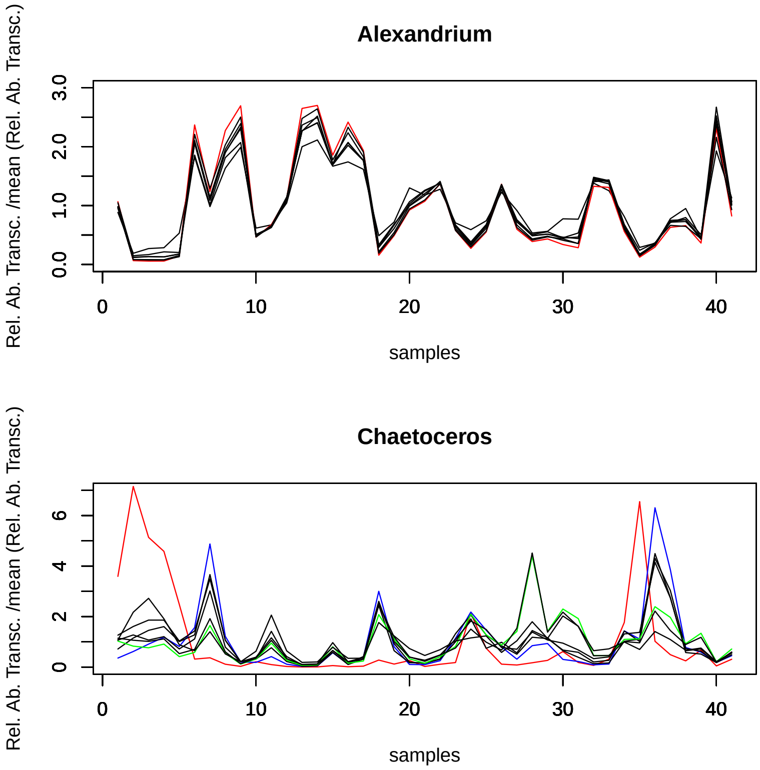


FIGURE D.1: *Representation of the relative abundance dynamics of species from the genus* Alexandrium *and* Chaetoceros *in the environmental data*

Appendix E

Other collaboration : Inter and intra specific transcriptional and phenotypic responses of *Pseudo-nitzchia* under different nutrient conditions

Author contributions : The next section regroups results obtained during a collaborative work aiming to investigate the transcriptional and phenotypic divergence, at both inter and intra specific levels in *Pseudo-nitzchia*. Cultures were grown by Mickael Le Gac, Kimberley Lema and Julien Quéré. I helped Mickael Le Gac, Kimberley Lema and Julien Quéré for RNA extraction and library preparation. Data analysis were mainly performed by Mickael Le Gac and Kimberley Lema, I helped in bio-informatic scripting and manuscript correction.

FEMS Microbiology Ecology

OXFORD UNIVERSITY PRESS | FEMS Microbiology Ecology

http://mc.manuscriptcentral.com/fems

Inter and intra specific transcriptional and phenotypic responses of Pseudo-nitzchia under different nutrient conditions

Journal:	*FEMS Microbiology Ecology*
Manuscript ID	FEMSEC-18-08-0438
Manuscript Type:	Research article
Date Submitted by the Author:	09-Aug-2018
Complete List of Authors:	Lema, Kimberley; Ifremer Centre de Bretagne, ODE/DYNECO/Pelagos Metegnier, Gabriel; Ifremer Centre de Bretagne, ODE/DYNECO/Pelagos; CNRS, Sorbonne Université, Pontificia Universidad Catolica de Chile, Universidad Austral de Chile, UMI 3614, Evolutionary Biology and Ecology of Algae Quéré, Julien; Ifremer Centre de Bretagne, DYNECO/Pelagos Latimier, Marie; Ifremer Centre de Bretagne, DYNECO/Pelagos Youenou, Agnès; Ifremer Centre de Bretagne, DYNECO/Pelagos Lambert, Christophe; University of Brest, LEMAR Fauchot, Juliette; Normandie Univ, UMR BOREA Le Gac, Mickaël; Ifremer Centre de Bretagne, DYNECO/Pelagos
Keywords:	phenotype, transcriptomics, divergence, nutrients, physiology, diatoms

The authors describe major interspecific but also intraspecific divergence in terms gene expression, toxin production, growth and nutrient consumptions among strains belonging to the diatom genus *Pseudo-nitzschia*.

TITLE

Inter and intra specific transcriptional and phenotypic responses of Pseudo-nitzchia under different nutrient conditions

Authors: Kimberley A. Lema[1], Gabriel Metegnier[1,2], Julien Quéré[1], Marie Latimier[1], Agnès Youenou[1], Christophe Lambert[3], Juliette Fauchot[4,5], Mickael Le Gac[1*]

Addresses of authors:

[1] IFREMER, Dyneco Pelagos, Plouzané 29280, France.

[2] CNRS, Sorbonne Université, Pontificia Universidad Catolica de Chile, Universidad Austral de Chile, UMI 3614, Evolutionary Biology and Ecology of Algae, Station Biologique de Roscoff, Place Georges Teissier, CS90074 29688, Roscoff Cedex, France.

[3] Laboratoire des Sciences de l'Environnement Marin (LEMAR), UMR 6539 CNRS UBO IRD IFREMER ; Institut Universitaire Européen de la Mer, Technopôle Brest-Iroise, Rue Dumont d'Urville, 29280 Plouzané, France.

[4] Normandie Univ, UNICAEN, 14000 Caen, France.

[5] UMR BOREA, CNRS-7208, IRD-207, MNHN, UPMC, UCBN, 14032 Caen, France.

*** Corresponding Author:**

Mickael Le Gac

Address: IFREMER, DYNECO PELAGOS

29280 Plouzané

France

Current Phone:

Current email: mickael.le.gac@ifremer.fr

Key Words: phenotype; transcriptomics; Pseudo-nitzschia; divergence; nutrients; physiology

ABSTRACT

Untangling the functional bases of divergence between closely related species is a step towards understanding species dynamics within communities. We investigated how cellular (i.e: growth, domoic acid production, nutrient consumption) and molecular (transcriptomics analyses) level phenotypes varied in response to varying nutrient concentrations across several strains belonging to three species of the diatom genus *Pseudo-nitzschia*. Profound inter-species divergences were observed. Under phosphate limiting conditions, *P.australis,* the most toxigenic species of all three, consumed rapidly all the phosphate available and expressed particular genes involved in phosphate transport and management. Under nitrogen limiting conditions, *P.pungens* did not consume all the nitrogen available and transcriptional responses reflected that this species was still expressing genes involved in photosynthesis. In contrast, *P.australis* and *P. fraudulenta* exhausted all nitrogen available and expressed genes involved in the TCA cycle and the nitrogen metabolism. At the intra-species level, differences were moderate. However, one strain showed extreme differences in terms of gene expression without extensive genetic divergence. We argue that these differences have a physiological

basis linked to its recent isolation from the natural environment prior experimentations. We therefore advocate for a deeper understanding of the influence of culture duration and life cycle on physiological responses.

1. INTRODUCTION

Pseudo-nitzschia is a cosmopolite and common genus of diatoms, with the singularity of producing domoic acid (DA), the toxin responsible for amnesic shellfish poisoning worldwide (Lelong *et al*. 2012; Trainer *et al*. 2012). Although numerous species within the genus are considered as DA producers (Teng *et al*. 2016; Zabaglo *et al*. 2016; Lema *et al*. 2017), in the various localities where toxicity events occur, only a few species among the several ones naturally occurring are really problematic. It is for example the case of *P. multiseries* and *P. australis*, often co-occurring with other less problematic species (Lelong *et al*. 2012; Trainer *et al*. 2012). Biotic and abiotic parameters also influence physiological responses such as the increase in DA concentrations under specific environmental parameters (reviewed in Lelong *et al.* 2012 ; Lelong, Hégaret and Soudant 2014; Sison-Mangus *et al.* 2014; Tammilehto *et al.* 2015; Lema *et al.* 2017).

Numerous studies have analysed cellular phenotypes of different *Pseudo-nitzschia* species in response to different environmental parameters in order to attempt understanding phenotypic divergence. Though, very few have also looked at what happens at molecular levels (e.g Boissenault *et al.* 2013),which is required to fully understand ecological divergence between closely related species and ideally predict their dynamics within communities. By monitoring simultaneously thousand traits with physiological and thus potential ecological implications, the sequencing and analyses of mRNA levels (i.e. the genes used in a given situation by an organism) through transcriptomics analysis is a powerful tool to determine to what extend closely related organisms diverge in their responses to environmental conditions. These approaches are especially promising because they allow to follow species specific responses to given environmental conditions *in vitro* for traits that are also measurable *in situ* via meta-transcriptomic approaches. .

Transcriptional responses in diatoms subject to varying environmental conditions have generally revealed divergent gene expression patterns and low numbers of orthologous genes between distantly related diatoms (Maheswari *et al.* 2010; Smith, Abbriano and Hildebrand 2012; Bender *et al.* 2014). Comparative transcriptional analysis of *Fragilariopsis cylindrus*, *Pseudo-nitzschia multiseries* (two pennate diatoms), and, *Thalassiosira pseudonana* (a centric diatom) under nitrogen limiting conditions showed differences in non-orthologous gene subsets and in transcription levels, with less than 5% of the shared orthologous gene clusters similarly transcribed (Bender *et al*. 2014). Previous studies comparing expression profiles under nitrogen limitation also found relatively few orthologs shared and highlighted differences in gene expression profiles between *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* (Maheswari *et al.* 2010; Hockin *et al.* 2012). Highly divergent expression patterns and the use of alternative sets of genes were found when comparing *Pseudo-nitzschia sp.* and *Thalassiosira sp.* under high and low Fe (Cohen *et al*. 2017). Overall, studies highlighted that distantly related diatoms may invoke alternative strategies when dealing with identical changes in the environment.

Only one previous study compared the transcriptomes of three closely related *Pseudo-nitzschia* species (*Pseudo-nitzschia arenysensis*, *Pseudo-nitzschia delicatissima* and *Pseudo-nitzschia multistriata*) (Di Dato *et al.* 2015). This study compared transcripts across species particularly searching for genes hypothesized to be involved in domoic production such as the ones involved in the synthesis of glutamate, geranyl type molecules and a SLC6 (Sodium and Chloride-dependent amino acid) transporter (Savage *et al.* 2012; Boissonneault *et al.* 2013). No major gene expression differences were found amongst the studied species for these particular genes except for transcripts related to the SLC6 transporter that were more expressed in *P. multistriata* (i.e the supposedly toxic species) (Di Dato *et al.* 2015). While relating the expression of SLC6 transport to DA requires a deeper analysis, the study was based on the collection of RNA at a single experimental condition (one medium and collection at exponential phase), for a single strain, and, cellular phenotypes (such as growth, DA and nutrient consumption) were not measured (Di Dato *et al.* 2015). Nonetheless, the major finding of this study was the strong inter-species divergence across the three studied species. Indeed, the transcriptome of *Pseudo-nitzschia multistriata* was quite different to the other two cryptic species studied displaying a large fraction of unique functions including a Nitric Oxide synthetase (PmNOS) (involved in cell signaling and usually found in metazoans) (Di Dato *et al*. 2015).

In a previous study we compared inter and intra-specific growth and domoic acid production in relation to nutrient ratios and concentrations for several strains of three Pseudo-nitzschia species: *P. australis*, *P. pungens* and *P. fradulenta* (Lema *et al*. 2017). We found that differences were governed by strong inter-species variation but that phosphate limitation enhanced similar responses with high DA concentrations and low growth across all strain /species studied. We also found that while intra-species variation was low, a recently isolated strain was quite different to its conspecifics suggesting the influence of time since isolation on the physiology and DA production.

As a further step towards the investigation of the potential ecological divergence between and within closely related species of *Pseudo-nitzschia* (*P. australis*, *P. pungens*, and

P. fraudulenta), the present study investigated phenotypic responses at both the cellular and molecular levels. We investigated growth, DA production, nutrient consumption as well as mRNA levels through transcriptomic analyses in three experimental conditions. Based on our previous study (Lema *et al*. 2017), we chose three mediums with contrasting limiting conditions differently affecting growth and domoic acid production: 1) a phosphate limited medium (X1); 2) A nitrate limited medium (X2) and; 3) A non-limiting environment (X3). To be able to appreciate whether the observed phenotypic variations are linked to difference between species, and not to any other kind of segregating polymorphism (i.e physiological, intraspecific genetic polymorphism), two to three strains (i.e. clones) were assayed per species (two *P. pungens* and *P. fraudulenta* strains and three *P. australis* strains). Using this experimental setting we ask: 1. How does nutrient availability affect physiology in terms of transcriptional and cellular responses? 2. To what extent are the physiological responses strain or species specific?

2. MATERIAL AND METHODS

2.1. Cellular phenotype

Pseudo-nitzschia cultures

Seven strains of three *Pseudo-nitzschia* species (3 strains of *P. australis*, 2 for *P. fraudulenta* and 2 for *P. pungens*) were collected and isolated from different locations on the north coast of France and selected for this study (Supplementary Table 1). To establish monoclonal strains, single cells were isolated using a sterile micropipette and washed thoroughly with filter-sterilized seawater. Cultures were maintained in sterile-filtered oligotrophic seawater amended with K/2 + Si medium (100.8 µM $Na_2SiO_3.5H_2O$; pH ~8;

salinity ~33.5) (modified from Keller *et al.* 1987) at 16 °C, under a 12:12 L:D cycle and 80 µmol.photons.m^{-2}.s^{-1}, in an algal incubators. *Pseudo-nitzschia* species were identified using transmission electron microscopy and/or genotyping.

Experimental setup

To test the effect of different nutrient concentrations (phosphate, nitrate and silicate) on cell densities and DA concentrations, strains were grown in three modified K/2 + Si media, in which concentrations of $NaNO_3$, $NaH_2PO_4.H_2O$ and $Na_2SiO_3.5H_2O$ varied (at 12:12 L:D cycle, temperature ~16 °C and pH ~8). From the results of a previous study on the same strains (Lema *et al.* 2017) we chose three mediums with contrasting effects on growth and domoic acid production: 1) a phosphate limited medium that enhanced highest DA production and lowest growth (X1: N=48µM, Si=48µM, P=3µM); 2) A nitrate limited medium (X2: N=48µM, Si=48µM, P=10µM) that enhanced high DA production but that allowed a better growth and; 3) A non-limiting environment (X3: N=480µM, Si=480µM, P=30µM) which enhanced the best growth and lowest DA production

All strains and species were grown in quadruplicates for each of the three experimental media, in 500ml flasks (T175, Sarstedt AG & Co, Nümbrecht, Germany) containing 200 ml of media and a starting cell concentration of ~1,500 cells.ml^{-1} (making a total of 84 cultures). In order to collect all samples simultaneously at stationary phase, differences in growth trends were initially monitored for each strain. The strains reached stationary phase eight days after inoculation in X1 and X2, and 13 days after inoculation in X3. Therefore for the experiment all strains in medium X3 were started 5 days prior to the other two mediums (X1 and X2). Growth was then followed for all treatments by subsampling 1 ml of homogenized culture every two days (for a total of 8 days for X2 and X1 mediums, and 13 days for X3 medium) into a 48-well plate and quantifying *in vivo* chlorophyll fluorescence (ex: 440/40, em: 680/20) in a FLX800 fluorescence microplate reader (Biotek Inc., VT, USA).

At late stationary phase, samples were simultaneously collected for the numerous subsequent analysis: 1) 5 ml of homogenized culture (cells and media) were collected and stored at -80 °C for domoic acid analysis; 2) 50ml of media (no cells) were collected and stored for nutrient analysis (i.e Nitrate, Phosphate and Silicate; nb: initial nutrient samples of each medium were also collected prior to the start of the experiment in order to measure nutrient consumption); 3) 1ml were fixed at a final concentration of 0.25 % glutaraldehyde for cell counting analysis (flow-cytometry) ;4) The remaining culture (~130ml) were used for RNA extractions which were immediately performed on samples.

Domoic acid quantification

Stored samples were subsequently thawed and centrifuged at ~2,000g for 20 min to separate cells from the medium and therefore to distinguish particulate DA (PDA) from dissolved DA (DDA). Domoic acid was subsequently quantified as in Lema *et al.* (2017) using the DA ELISA kit (Mercury Science, Durham, NC, USA).

Nutrient analysis

Nutrients concentrations were determined after sample dilutions, using segmented flow analysis (SFA) on a Seal Autoanalyseur AA3 and following classical methods detailed in (Aminot, Kerouel and Coverly 2009)

Cell quantification

Cell densities were quantified with a flow cytometer (FACSVerse, Becton Dickinson, San Jose, CA) equipped with three lasers (violet: 405 nm; blue: 488 nm; red: 640 nm) and eight filters (527/32 nm, 586/42 nm, 700/54 nm and 783/56 nm for the blue laser; 448/45 nm and 528/45 nm for the violet laser; and 660/10 nm and 783/560 nm for the red laser). Algal cells were discriminated by their natural red autofluorecence (linked to chlorophyll pigments) after excitation by the 488 nm laser detected on the red fluorescent detector of the flow-cytometer (700/54 nm BP) Cell quantification was calculated based on the number of events, gated as algal cells, and the volume of sample recorded by the coupled Flow-Sensor device. Analyses were run for 180 sec at medium flow rate (~80 µL min). Microscopy analyses of the samples revealed that the fixed cells were either isolated or forming two to three cell chains. As a result, the forward scatter (that can be considered as a proxy for size) of the flow cytometer displayed a bi- or three-modal distribution. Cell density estimates were corrected based on these multimodal distributions.

Statistical analysis

Anova were used to investigate how each of: cell density, domoic acid level, silicate, phosphate, and nitrate concentrations, at the end of exponential phase varied among strains and species in each of the three media. The models considered strains nested within species in interaction with media. Data were log transformed. Rather than putting emphases on p-values (all of them being < 0.0001), we focused on the effect sizes (η^2) of the explanatory variables and of their interaction. Effect sizes were calculated as the sum of square of the focal explanatory variable (or interaction) divided by the total sum of square and therefore indicate the proportion of the variance explained by the focal explanatory variable (or interaction).

2.2 Molecular phenotype

RNA extraction, library preparation and sequencing

For total RNA extraction, cells were pelleted at 8,500g during 8 min at 4°C. After addition of RLT buffer (Qiagen) supplemented with β-mercaptoethanol, samples were sonicated on ice, RNA was then purified using RNeasy Plus Minikit (Qiagen) following the manufacturer protocol. Extracted RNA was quantified using a Biotek Epoch spectrophotometer and the quality estimated on RNA 6000 nano chips using a Bioanalyzer (Agilent). RNA extraction failed for 15 samples (all, i.e eight, *P. fraudulenta* samples in medium X1; four *P. fraudulenta* samples in medium X2; two *P. pungens* samples in medium X1; and one *P. australis* sample in medium X1). Starting from 0.5µg of total RNA, poly-A selection, reverse transcription and library preparation was performed using the Illumina Stranded Truseq Total RNA Library Preparation Kit. One library was generated per sample, representing a total of 69 libraries. Library quality was assessed on a Bioanalyzer (Agilent) using high sensitivity DNA analysis chips and quantified using Kappa Library Quantification Kit. Paired-end sequencing was performed in 2 x 150bp on 3 lanes of a Hiseq3000 (Illumina).

To avoid batch effects, samples were randomized for RNA extraction, library preparation and sequencing.

Reads quality assessment and filtering

Adapter removal and quality filtering were initially performed on raw reads using Trimmomatic (v.0.32) (Bolger, Lohse and Usadel 2014) and the following parameters: ILLUMINACLIP: 2:30:10:8:TRUE, LEADING:3, TRAILING:3, MAXINFO:135:0.8, MINLEN:80

Obtaining reference transcriptomes and alignment

After initial quality filtering, for each species, the reads from the different samples were pooled, and de novo assembly of one reference transcriptome per species was obtained using Trinity (Haas *et al*. 2013). Only transcripts longer than 200bp were retained. When several isoforms were identified, only the longest one was retained. Sequence similarity of the transcripts with genes of identified function in the manually curated UniProt databank was investigated using blastx with E-Value < 10^{-3}. The transcripts were classified in various Gene Ontology categories (GO; http://geneontology.org/) based on this result. Transcripts resulting from carry-over rRNA contaminants were identified and removed after performing blastn against Silva database (Quast *et al*. 2013), corresponding to 37, 32, and 30 transcripts for *P. australis*, *P. pungens* et *P. fraudulenta*, respectively.

The 69 samples were individually aligned to their corresponding species reference transcriptome with Bowtie2 (Langmead and Salzberg 2012) using paired-end reads. Only reads with a mapping score > 10 were retained. Pairs for which the two reads did not map concordantly on the same transcript were removed. Samtools was used to sort, index and obtain raw read count tables from bam files (Li *et al*. 2009).

Intraspecific differential gene expression

For all differential gene expression (DE) analyses, only transcripts with an average read count per sample higher than ten were considered (Table 1).

Within each species, pairwise DE were analysed: 1. between media for each strain, as well as 2. between strains in each media. DE was tested using: 1. Negative binomial models on raw read counts (Wald tests as implemented in DESeq2; Love, Huber and Anders 2014) with a FDR threshold set q-value = 0.01 and 2. Linear models on voom normalized read counts (as implemented in limma; Ritchie *et al*. 2015) with a FDR threshold set q-value = 0.05. FDR thresholds were set based on preliminary analyses of the datasets. Only transcripts identified as significant using both approaches and displaying a log2 fold change >2 were considered as differentially expressed (Supplementary Table 2). Considering the transcripts identified as significant in at least one pairwise comparison, the expression profiles across samples were clustered following negative binomial models as implemented in MBCluster.Seq, using expectation-maximization algorithm for estimating the model parameters and cluster membership (Si *et al*. 2014). After preliminary analyses, 20, 10 and 5 clusters were investigated for *P. australis*, *P. pungens*, and *P. fraudulenta,* respectively. The expression profiles of the selected clusters were subsequently visually inspected and clusters displaying similar profiles were merged (see Results). Expression is reported as the log2 fold change of the median expression in the category of interest over the median expression considering all samples. It is calculated as $log2(\widetilde{2^{x_i}}/\widetilde{2^{X}})$ where $x_i$is the rlog transformation (as implemented in DESeq2) of the number of reads mapping to a given transcript for the samples belonging to category i and X is the rlog transformation of the number of reads mapping to a given transcript for all samples.

Over-representation of GO categories in the identified clusters were tested, for (GO) functional gene categories represented by at least five transcripts, using Fisher Exact tests followed by a False Discovery Rate (FDR) correction for multiple testing with a significance threshold set at q-value = 0.05. Only GO categories containing more than five DE transcripts in a given cluster were considered. GO categories may be redundant (as similar set of genes belong to different GO categories). To mitigate such redundancy, GO categories displaying an overlap coefficient $\frac{GO_i \cap GO_j}{min(GO_i,GO_j)} > 0.8$ (where $GO_i$is the size of the GO category i) were clustered, and for each cluster only the GO category displaying the lowest q-value was reported.

Interspecific differential gene expression

Orthology of the transcripts considered for DE analyses was investigated using reciprocal blastn with E-Value < 10^{-3}. Only transcripts with an ortholog in each of the three species were considered (see Table 1; Supplementary Figure 7). DE was investigated as above: i.e. investigating within each species, pairwise DE between media for each strain, as well as between strains in each media. Transcripts displaying similar expression profiles across tested samples were clustered (10, 10 and 5 clusters were considered for *P. australis*, *P. pungens*, and *P. fraudulenta* respectively). Interspecific overlap between the clusters was tested using Fisher Exact tests followed by FDR with a significance threshold set at q-value =

0.01 and absolute log2(odd ratio) > 0.5. Over-representation of GO categories in the identified clusters was tested as above.

Intraspecific nucleotide divergence

For each strain, the fastq files obtained after sequencing were concatenated. The reads were quality filtered using Trimmomatic (as above). Reads were then aligned to their corresponding reference transcriptome (as above) using Bowtie2 (Langmead and Salzberg 2012). Single nucleotide polymorphisms (SNPs) were detected using FREEBAYES (Garrison and Marth 2012) and filtered using VCFTOOLS (Danecek *et al*. 2011). Positions were considered when displaying two alleles, a quality criterion higher than 40, and when found more than 20 times in each strain. Pairwise nucleotide divergence was calculated as the number of SNPs divided by two times (diploid organisms) the number of positions covered more than 20 times in each strain.

3. RESULTS

3.1 Cellular phenotype

Growth and domoic acid

A total of seven strains from three *Pseudo-nitzschia* species (3 *P. australis*, 2 *P. pungens*, 2 *P. fraudulenta*) were grown in three media with variable nitrate, phosphate and silicate initial concentrations. At the onset of the stationary phase, cell densities were measured using flow cytometry, domoic acid was quantified, and nitrate, phosphate and silicate concentrations were measured (Figure 1). Considering cell densities and domoic acid, there were major and significant differences between species (Figure 2: Species effect). More precisely, *P. australis* strains produced more DA (i.e total DA: particulate and dissolved), while *P. fraudulenta* strains reached higher densities than the two other species (Figure 1a-b). Not surprisingly, for the three species, cell densities were higher in the medium with high initial nutrient concentrations (X3). As previously reported, the various strains produced more DA in the media X1 that inversely had lowest phosphate concentrations (Lema et al. 2017) (nb: Pa2 reached the highest DA with an average of 0,132 ng cell^{-1} in medium X1). Intra-specific differences were marginal as in general strains grew and produced DA in similar fashions (Figure 2, small strain effect size for growth and DA). There were also some less marked strain and media interaction effects (Figure 2). One illustrative example being the strain Pa2 displaying much more variable cell density and DA production patterns across media than the other strains (Figure 1 a-c).

Nutrient consumption

Differences in nutrient consumption at the end of stationary phase were majorly driven by differences in their initial nutrient concentration (Figure 2), though some species, strains and media interactions were also distinguishable (Figure 1 c-e; 2). In medium X1, phosphate was almost completely depleted at stationary phase (average P in X1= 0.06 µmol l^{-1} SE +/- 0.01) and therefore was potentially the nutrient limiting cell growth across all tested strains. In media X2, the limiting nutrient seemed to be nitrate except for *P.pungens* (see below for interspecific difference). In media X3, all three nutrients were still present at the onset of stationary phase.

The second most important explanatory factor was the species*media interaction, indicating that species consumed nutrients in different ways in the three media. This was particularly true for *P. pungens*, as both strains from this species consumed less P in X2 and more silicate in X3 than the other two species. The most important difference, and probably with strong implications in terms of gene expression (see below), was related to nitrate consumption in medium X2. While nitrate was entirely consumed by *P. australis* and *P. fraudulenta* strains, there was still a fair amount of nitrate left in *P. pungens* cultures. Finally, to a lesser extent, there were also some strain*media interaction effect. This was apparent in the Si consumption in medium X2 that was quite different across different strains (e.g Pa3 being the only strain that consumed all the silicate available in medium X2). Another example was the difference of phosphate consumption between Pf1 and Pf2 in X2.

3.2. Molecular phenotype

Intraspecific gene expression

Intraspecific differential gene expression (DE) was investigated in pairwise comparisons between media for each strain and between strains in each media. As described in the methods section, clusters were retained to describe the expression profiles of these DE transcripts across the samples within each species. After observing expression patterns across clusters, two general cluster trends could be recognized for the three studied species: 1) clusters reflecting a strain specific DE and 2) clusters reflecting media specific DE. The strain specific clusters contained transcripts barely expressed, if at all, in some strains but displaying variable expression levels in the other strains; whilst the media specific clusters reflected variable DE in relation to different medias. Below we develop these results for each species.

P. australis

Within *P. australis* differential gene expression (DE) was investigated in 18 pairwise comparisons (Supplementary Table 2), resulting in the identification of 20,663 DE transcripts. Ten clusters were retained to describe the expression profiles of these DE transcripts across the samples (Figure 3), five clusters reflected a strain specific DE (clusters Pa_C1, Pa_C7-10) and five clusters reflected a media specific DE (Pa_C2-6)

The main strain specific cluster (Pa_C1, Figure 3), in terms of both, fold change and number of transcripts, regrouped transcripts expressed in strain Pa3 and barely expressed, if at all, in the two other strains. More than 13,000 transcripts were found in this cluster (representing ~40% of the transcripts considered for DE analysis in *P. australis*) illustrating extensive divergence in terms of gene expression between Pa3 and the other two strains. More than 120 GO categories were over-represented in this cluster. Of special interest were numerous functions related to cell motility (Figure 4). Among the 50 transcripts displaying the highest fold change in gene expression between Pa3 and the other two strains and belonging to the cluster Pa_C1, we note the presence of tubulin and actin transcripts with very high expression levels (several 10,000th reads) in Pa3 and low expression levels in the other two strains (a few tens of reads), illustrating major differences in the cytoskeleton of these strains (Figure 5). There are also numerous ribosomal proteins, probably reflecting differences in the activity of the translation machinery (Figure 5). The presence of a phosphate carrier may also be noted (Figure 5). The other four strain specific clusters were much more anecdotal, as they only gathered a few tens of transcripts.

The five media specific clusters were of similar size (from 951 to 2,512 transcripts). They regrouped transcripts that were more expressed in the media X1 (Pa_C6), X2 (Pa_C2), and X3 (Pa_C5), as well as transcripts less expressed in X1 (Pa_C3) and X3 (Pa_C4). When looking at the GO categories over-represented in the clusters, cluster Pa_C2 (Figure 3) displayed an over-representation of transcripts involved in nitrate assimilation and nitrogen metabolism, but also in central metabolism (amino acid biosynthesis, glycolysis and TCA cycle) (Figure 4). Among the 50 transcripts displaying the highest fold change in gene expression between the medium X2 and the other two media, the presence of highly expressed nitrite reductase and nitrate transporters may be noted (Figure 5). The cluster Pa_C3 (Figure 3) displayed a slight over-representation of transcripts involved in photosynthesis and iron binding (Figure 4). Among the 50 transcripts displaying the highest fold change in gene expression between the medium X1 and the other two media, we note the presence of a phosphate transporter with extremely high expression levels (Figure 5). The cluster Pa_C4 displays an over-representation of transcripts that maybe involved in cell communication (Figure 4). Among the 50 transcripts displaying the highest fold change in gene expression between the medium X3 and the other two media, there are only 4 annotated transcripts, which may reflect that the cellular processes at play do not exist in model organisms (Figure 5). The cluster Pa_C5 displayed a strong over-representation of transcripts involved in photosynthesis, and to a lesser extends in glycolysis and amino acid synthesis (Figure 4). Among the 50 transcripts displaying the highest fold change in gene expression between the medium X3 and the other two media, there were fructose biphosphate aldolase (a key enzyme of the calcin cycle, gluconeogenesis and glycolysis), as well as several pigment encoding transcripts (especially fucoxanthin, but also violaxanthin) (Figure 5). Finally, the cluster Pa_C6 displayed an over-representation of transcripts involved in photosynthesis, pentose phosphate shunt, amino acid and fatty acid biosynthesis (Figure 4).

P. fraudulenta

We were unable to extract enough RNA from *P. fraudulenta* strains grown in the medium X1 to perform the RNA sequencing therefore DE in *P. fraudulenta* was only investigated across two media (X2 and X3) and two strains. Differential gene expression was investigated in 2 pairwise comparisons (between X2 and X3 for strain Pf2, and between strains Pf1 and Pf2 in media X3, as only one replicate was sequenced for Pf1 in media X2; no pairwise comparison could be done with Pf1 in X2; Supplementary Table2), resulting in the identification of 2,057 DE transcripts. Only two clusters were retained to describe the expression profiles of these DE transcripts across the samples (Supplementary Figure 1), and the clusters displayed media specific DE. The cluster Pf_C1 regrouped transcripts with higher expression in X2 than X3 (Supplementary Figure 1) and included an over-representation of transcripts related to the ribosomes and the transcription machinery in general (Supplementary Figure 2). Of special interest was also the over-representation of transcripts related to the nitrogen metabolism (Supplementary Figure 2). Among the transcripts displaying the highest DE between X2 and X3, there were nitrate and putative ammonium transporters, as well as highly expressed nitrate and nitrite reductase (Supplementary Figure 3). The cluster Pf_C2 regrouped transcripts with higher expression in X3 than X2 (Supplementary Figure 1) and included an over-representation of transcripts related to photosynthesis (Supplementary Figure 2) such as pigments encoding transcripts (fucoxanthin), as well as a transcripts involved in glycolysis and several transcripts involved in amino acid metabolism (Supplementary Figure 3).

P. pungens

Differential gene expression was investigated in 9 pairwise comparisons (between media for each strain and between strains in each media as for the other two species; Supplementary Table 2), resulting in the identification of 3,390 DE transcripts. Seven clusters were retained to describe the expression profiles of these DE transcripts across the samples (Supplementary Figure 4). Two of these clusters displayed strain specific DE patterns (Pp_C6-7) with a couple hundred transcripts in each cluster that were not related to any functional GO category. The five other clusters were media specific (Pp_C1-5). The media specific cluster Pp_C1 harboured transcripts with a higher expression in medium X2 (Supplementary Figure 4), and displayed an over-representation of transcripts involved in photosynthesis (Supplementary Figure 5). Among the 50 transcripts displaying the highest fold change between media X2 and the other two, only a few were annotated (Supplementary Figure 6). The cluster Pp_C2 regrouped transcripts with higher expression in X1 than in the other two media (Supplementary Figure 4) and displayed an over-representation of transcripts involved in central metabolism processes (TCA cycle, amino acid biosynthesis, oxido-reduction; Supplementary Figure 5). Among the 50 transcripts displaying the highest fold change between the media X1 and the other two, we note the presence of numerous transcripts with identified homologs in UNIPROT, including potential nitrate and phosphate transporters, as well as Fructose-bisphosphate aldolase and transcripts associated with the TCA cycle (Pyruvate dehydrogenase) and glyoxylate shunt (isocitrate lyase) (Supplementary Figure 6). The cluster Pp_C3 was characterized by transcripts displaying low expression in X3 compared to X1 and X2 (Supplementary Figure 4). Among the over-represented GO categories we note the presence of kinase activity, phosphorylation and lipid catabolism related functions (Supplementary Figure 5). The cluster Pp_C4 was characterized by transcripts displaying low expression in X2 compared to X1 and X3 (Supplementary Figure 4). We may note the over-representation of transcripts involved in protein synthesis in general (Supplementary Figure 5), the presence of fructose biphosphate aldolase, as well as of Glycerol-3-phosphate dehydrogenase (an enzyme linking carbohydrate and lipid metabolism; Supplementary Figure 6). The cluster Pp_C5 was characterized by transcripts displaying low expression in X1 compared to X2 and X3 (Supplementary Figure 4). We may note the over-representation of transcripts involved in photosynthesis (Supplementary Figure 5) with pigment associated transcripts (especially fucoxanthin, but also violaxanthin) among the most DE transcripts (Supplementary Figure 6).

Interspecific comparisons

Ortholog genes

To be able to compare DE profiles between species, only transcripts displaying orthologs in each of the three species were retained, reducing the number of transcripts considered for DE analyses to 6,557 (Table 1, Supplementary figure 7). Ortholog transcripts represented only ~20%, ~41% and ~32% of *P.australis, P. fraudulenta and P.pungens* overall transcripts respectively. Pairwise DE analyses were performed as described above for the intraspecific comparisons, (i.e. investigating, within each species, DE between media for each strain and between strains in each media; Supplementary Table2) and transcripts were clustered within each species to identify intraspecific expression profiles (Figure 6a). Overlaps between clusters of the different species were analysed.

A total of 2,325 transcripts were identified as DE in at least one pairwise comparison (Table 1). At the intraspecific level, contrary to the results obtained analyzing each species separately; no strain specific clusters could be identified. Even the extremely large cluster of transcripts highly expressed in the strain Pa1 (Pa_C1, Figure 3) was not identified, illustrating that the transcripts expressed in a strain specific manner had no homologs in all three species. In *P. fraudulenta*, transcripts of cluster fc3 (Figure 6a) displayed a more or less constant expression pattern across strains and media. Transcript clusters across the other two species were media specific and often similar, in terms of DE profiles, to the ones identified when investigating intraspecific gene expression profiles. A notable exception were two clusters displaying low expression in the medium X1 in *P. australis* (Pa_C3, Figure 3) and *P. pungens* (Pp_C5, Supplementary Figure 4) when considering each species separately, but not identified when focusing the orthologs.

In the medium X2 and X3 P. australis and P. fraudulenta display similar expression patterns

Four hundred and twenty of the transcripts considered as DE (Figure 6a) displayed high expression levels in media X2 in both *P. fraudulenta* (Figure 6a, fc2; 48% of the transcripts in fc2) and *P. australis* (Figure 6a, ac2; 62% of the transcripts in ac2), while 195 of these transcripts displayed, in *P. pungens*, low expression in X2 (Figure 6a, pc4; 50% of the transcripts in pc4 shared with both fc2 and ac2). Transcripts shared between these three clusters were associated with several broad functional GO categories linked to the transcription and translation machineries, but also to two specific GO categories: nitrogen metabolism and TCA cycle (Figure 6b).

Transcripts associated with nitrogen uptake and metabolism (Figure 7, nar, nir, at, gas, gad) were often more expressed in the nitrogen limited medium X2 within *P. australis* and *P.*

fraudulenta, while the same transcripts displayed a constant lower expression across media within *P. pungens*. For example, *P. pungens* displayed a very low expression of nitrate reductase (a key enzyme of nitrogen metabolism) in the three media, while its expression was higher and highly dynamics across the three media in the two other species (Figure 7, nar). Urease, a key enzyme linking the urea cycle to the general nitrogen metabolism, tended to be much more expressed in all media in *P. fraudulenta* than in the other two species (Figure 7, ure).

A similar pattern was observed for transcripts associated in the TCA cycle (Figure 7, pdh, cs, aco, ldh, od, scs, ss, fum) and gluconeogenesis (Figure 7, gpd, pgk, pgm) which were once again more expressed in X2 within *P. australis* and *P. fraudulenta*, while the same transcripts displayed a constant lower expression across media within *P. pungens*. Other species-specific differences were also observed. For example, malate dehydrogenase, the enzyme oxidizing malate into oxaloacetate during the TCA cycle, was barely expressed in *P. australis*, while displaying a high expression levels in the other two species.

Transcripts encoding for pigments (Fucoxanthin and Chlorophyll) and cytochromes, as well as transcripts associated with the photosystems I and II (Figure 6a and Figure7) displayed high expression levels in X3 within *P. fraudulenta* (fc1) and *P. australis* (Figure 6a, ac1, 403 shared transcripts, representing 52% of ac1 and 46% of fc1), and in contrast tended to be highly expressed in X2 in *P. pungens* (Figure 6a, ac1, 337 shared transcripts, representing 52% of pc1 and 43% of ac1).

In the medium X1, P. australis and P. pungens display moderately similar global expression patterns, but marked differences related to phosphate transporters

In the medium X1, where phosphate is a limiting resource, some similarities were noted between *P. australis* (Figure 6a, ac5) and *P. pungens* (pc5, 67 shared transcripts, representing 43% of ac5 and 20% of pc5) expression profiles though no clear functional roles were revealed. In the medium X1, transcripts linked to phosphate transport were a priori of specific interest, even if they were not grouped into specific GO categories. Interestingly, phosphate transporters identified in the three species displayed expression profiles that seemed to be mainly driven by species-specific differences and to a lesser extent by the experimental environments. A transcript annotated as a sodium-dependent phosphate transport protein was highly expressed by *P. australis*, particularly in X1, while this same transcript displayed extremely low expression levels across all media for the two other species (Figure 7, sdpt). Pointing towards divergent management of phosphate intracellular pools, on the contrary, transcripts annotated as mitochondrial phosphate carriers were highly expressed by *P. fraudulenta* and *P. pungens* in all three media, but barely expressed by *P. australis* (Figure 7, mpc). Also, PEP/PO_4^{3-} translocator and 5' bisphosphate nucleotidase displayed extremely different patterns of expression between species (Figure 7, bn, ppt).

Nucleotide divergence between strains

Intraspecific pairwise genetic divergence based on SNPs was calculated within each species (Table 2). The total length of sequences considered for this analysis was similar for the three species (~25.10^6 bases). Within *P. pungens*, the two strains displayed a divergence of 0.2%, corresponding to the presence of one SNP every ~480 bases. Within *P. fraudulenta*, the two strains displayed a divergence of 0.05%, corresponding to the presence of one SNP every ~2,100 bases. Within *P. australis*, the three strains displayed a pairwise divergence of ~0.03%, corresponding to the presence of one SNP every ~3,300 bases.

4. DISCUSSION

Intra and interspecific phenotypic divergence within diatoms of the genus *Pseudo-nitzschia* was investigated in three experimental conditions mimicking nitrate limitation, phosphate limitation, and phosphate/nitrate non-limiting conditions. Phenotypic diversity was assessed at the cellular (DA, cell density), molecular (transcriptomic DE analysis) levels, as well as in terms of nutrient consumption for a total of seven strains from three *Pseudo-nitzschia* species (*P.australis*, *P. fraudulenta*, and *P. pungens*). The results obtained may be summarized as follows. Species specific divergence not affected by media was observed in terms of DA production, and cell densities. For example, *P. australis* strains produced more DA than the two other species, whilst *P. fraudulenta* strains reached higher cell densities than the two other species across the three media. Species specific differences were also noted at the molecular level as only a low part of genes were shared across the three studied species (~20-40% orthologous genes), but also more specifically when considering transcripts associated with phosphate transport. Strain specific differences preserved across media were rather moderate at both the cellular and molecular levels, except for a *P. australis* strain recently isolated from the natural environment and displaying an extremely high strain specific DE. Depending on the phenotype considered, phenotypic differences varying across media were observed at the intra and/or interspecific levels. For example, the observed cellular densities were at least as variable across media when comparing strains of the same species than when comparing different species. Another example, the nitrate consumption as well as the DE patterns were different between *P. pungens* and the other two species. Some major points are discussed below.

Nutrient limitation and divergence among Pseudo-nitzchia species.

Under nitrogen limiting conditions (X2), *P. fraudulenta* and *P. australis* displayed similar cellular and molecular phenotypic responses, while *P. pungens* responses were different or opposed. Indeed, *P. australis* and *P. fraudulenta* expressed transcripts associated with the nitrogen metabolism, while *P. pungens* displayed constant expression for the same set of genes. These included the presence of nitrite reductase amongst the eleven most highly expressed transcripts in X2 for both *P. fraudulenta* and *P. australis*. Moreover, *P. pungens* displayed a very low expression of nitrate reductase in the three media, while it was highly expressed in the other two species. At the cellular phenotypic level, *P. australis* and *P. fraudulenta* entirely exhausted the nitrate available while around 10µM of nitrate were still available for *P. pungens* strains. These results reflect that the major differences in terms of gene expression among species in the media X2 were likely linked to the nitrate exhaustion by *P. australis* and *P. fraudulenta* but not by *P. pungens*. The general pattern of over expression of nitrogen recycling transcripts under nitrogen stress (and specifically of nitrogen reductase as in the present study) has also been documented in recent studies specifically investigating transcriptional responses to nitrogen limitation in diatoms (Bender *et al.* 2014; Levitan *et al.* 2015). Our results point towards major differences in nitrate usage by *P. pungens* compared to *P. australis* and *P. fraudulenta.* Strains or species of *Pseudo-nitzschia* have been shown to display differences in their ability to grow on different nitrogen sources (Lelong *et al.* 2012), with intraspecific variability being especially important (Thessen, Bowers and Stoecker 2009). At the molecular level, a couple of studies also pointed toward divergence in nitrogen metabolism expression patterns among distantly related diatoms both *in vitro* (Bender *et al*. 2014) and *in situ* (Alexander *et al*. 2015). Our results suggest that such divergence may also exist among closely related species.

Under phosphate limiting conditions (X1), all species exhausted the phosphate available highlighting that P was limiting across all samples at the time of RNA extraction. This media was also the one where highest DA concentrations were observed across all species although inter-specific variations governed absolute DA production. *P.australis* was the most toxigenic species, as previously observed for those same strains and conditions (Lema *et al.* 2017). At the molecular level, RNA extraction failed for *P. fraudulenta* in medium X1, therefore no data was available for this species under P limiting conditions. For *P. australis* and *P. pungens* some shared transcripts displayed high expression in this medium, though these transcripts could not be clearly related to specific functional categories. However, focusing on transcripts involved in phosphate transport (Alexander *et al.* 2015; Grossman and Aksoy 2015; Lin, Litaker and Sunda 2016), there were extreme interspecific differences in gene expression, with one low affinity phosphate transporter among the eight most highly expressed transcript in *P. australis*, but barely expressed in *P. pungens*. The intracellular management of phosphate also seemed to differ between the two species as illustrated by the expression of transcripts linked to intracellular phosphate transport (PEP/PO_4^{3-} translocators and mitochondrial phosphate carriers), and to the recycling of phosphorus from nucleic acids (5' bisphosphate nucleotidase). Such divergence between two centric diatoms has already been documented in the field (Alexander *et al*. 2015). The present work showed that even closely related species use different set of genes to manage the intracellular pool of phosphate under phosphate limiting conditions.

Another interesting difference among *Pseudo-nitzschia* species was related to the expression of genes involved in photosynthesis such as pigments (Fucoxanthin and Chlorophyll) and photosystem I and II associated transcripts. These transcripts were more expressed in *P. australis* and *P. fraudulenta* in the medium X3, but more expressed by *P. pungens* in the medium X2. The effect of nutrient limitation on photosynthesis has been documented for a long time, but whether limitation increases or decreases photosynthetic activity tends to vary from one study to another (Cullen, Yang and MacIntyre 1992). In the present study, high expression of transcripts related to photosynthesis was detected when cells did not exhaust nutrients, but nevertheless reached stationary phase. One rather simple explanation for such pattern could be that self-shading among the cells induced light limitation that in turns induced increased expression of photosynthesis related genes (Flynn 2010). However, it seems unlikely that self-shading limitation is the sole explanation. Indeed, in *P. pungens*, cell densities are higher in X3 than in X2 but photosynthesis related genes are more expressed in X2.

Recent findings highlight that diatoms can adjust their metabolic responses (i.e the expression of known N and P metabolic pathways) under the same environmental conditions suggesting niche partitioning when co-occurring and therefore high flexibility in their responses to nutrient limitation (Bender *et al*. 2014; Alexander *et al*. 2015). Functional diversity is also observed in closely related clades of *Symbiodinium* (coral holobionts dinoflagellates) when looking at transcriptional responses of photosynthesis related genes (Parkinson *et al*. 2016). The present study reflected some similarities in the nitrogen usage and molecular responses to nitrogen limitation for *P. fraudulenta* and *P. australis* and a distinct pattern for *P. pungens,* suggesting potential ecological divergence linked to nitrogen usage. It is interesting to note that this divergence in nitrogen usage may not simply be explained by the phylogeny, as *P. australis* and *P. pungens* appear to be more closely related than *P. fraudulenta* (Ajani *et al*. 2018). Although the link between *in vitro* and *in situ* environmental conditions is always difficult to draw, we note that in the area where the strains were isolated *P. fraudulenta* and *P. australis* display similar phenology (spring/summer blooms), while *P. pungens* is present more regularly throughout the year and tend to bloom in early summer when less nitrate is available (Klein *et al*. 2010, Thorel *et al*. 2017).

A strain recently isolated from the field displayed a highly divergent expression pattern

In general, at the cellular levels (DA, cell densities), as well as in terms of nutrient consumption, strain specific effects were moderate. At the molecular level, strain specific differential expression (transcript DE in one of the strain irrespective of the experimental environment) was even more limited except for one strain within *P. australis* (Pa3) that displayed an extremely divergent response (more than 1/3 of the transcripts considered were almost exclusively expressed in this strain). It seems important to note that Pa3 was neither extremely divergent from the other strains when considering the cellular level phenotypes assessed in the present study, the only exception being the exhaustion of silicate in the X2 environment, nor when considering transcriptome wide SNP genetic divergence. The differential expression of this particular *P. australis* strain strongly affected specific cellular functions related to cellular motility and chain formation. Although pennate diatoms do not carry cilia, they may glide across surfaces using a poorly characterized mechanism (Wang *et al*. 2013). Here, we identified numerous transcripts associated with motility and only expressed in the recently isolated *P. australis* strains. We may cite numerous transcripts related to dynein motor proteins (Roberts *et al*. 2013). The term “dynein” appears in 70 out of the 224 transcripts only expressed in the recently isolated strain and associated with the GO category “cilium”, including 36 transcripts homologs to the dynein heavy chain of the axonemes. We also note the presence of numerous homologs of the amoeba *Dictyostelium discoideum* and more precisely of myosin homologs (18 out of 50 transcripts in the GO category "cell leading edge": movement of plasma membrane at the cell front Ridley, 2011) and of a highly expressed homolog of Fimbrin (several thousand reads in the recently isolated Pa strain, 0 in the other strains). Microscopic observation of the strains showed that the recently isolated strain displayed motile chains of cells (data not shown). Environmental samples displaying long motile chains of *Pseudo-nitzschia* are often considered as indicative of healthy populations (Rines *et al*. 2002). We suggest that this strain specific DE pattern has rather a physiological basis since Pa3 was not genetically divergent from the two other *P. australis* strains. Indeed, the genetic divergence between the two *P. pungens* strains was ten times higher than between the *P. australis* strains, without any strain specific differential expression pattern. We note that in our previous experiment carried a few months before the present one, the same strain (Pa3) exhibited striking cellular level phenotypic differences for the traits of study (DA, cell density and size) that this time were not evidenced (Lema *et al*. 2017). Together, such observations contribute to the idea of progressive physiological modifications in culture. We may speculate that motility could be progressively shut down because it is energetically costly and confer no advantage in culture.

Conclusion

To conclude, this study highlighted strong species specific cellular and molecular phenotypic divergence among closely related species within the genus *Pseudo-nitzschia*. This divergence could be separated into two distinct forms. First, there were species specific modifications of the experimental environment leading to differential expression of co-expressed transcript clusters between species. This was exemplified by the different usage of nitrate by the three species. Second, the three species may use different genes to respond to similar environmental conditions. For example, when facing phosphate depletion, the species overexpressed distinct sets of phosphate transporters. Although less pronounced, intraspecific polymorphism was observed, especially considering the cellular phenotypic level. At the molecular level, one strain exhibited dramatic differential expression, and this difference probably has a physiological (potentially linked to the acclimation to the *in vitro* environment) rather than a genetic basis. Interestingly, this strong pattern involves only a fraction of the transcripts associated with specific cellular functions (and especially motility) while the rest of the transcriptome behaves like in the other strains. This result advocates for systematic multi-strain approaches, and highlights the risk of using strains maintained under laboratory conditions for a long time to study traits of ecological interest.

ACKNOWLEDGEMENTS

This work was supported by a “CNRS INSU EC2CO grant” (Pseudotox), a post-doctoral fellowship from “Région Bretagne” and “Ifremer” to AKL, as well as a PhD fellowship from “Région Bretagne” and “Ifremer” to GM. Sequencing has been done at the GeT-PlaGe Genotoul sequencing plateform (Toulouse, France). We thank the Ifremer RIC bio-informatics team for support.

DATA AVAILABILITY

Raw reads and reference transcriptomes are accessible : DOIXXXXX

REFERENCES

Ajani PA, Verma A, Lassudrie M *et al.* A new diatom species P. hallegraeffii sp. nov. belonging to the toxic genus Pseudo-nitzschia (Bacillariophyceae) from the East Australian Current. *PLOS ONE* 2018;**13**:e0195622.

Alexander H, Jenkins BD, Rynearson TA *et al.* Metatranscriptome analyses indicate resource partitioning between diatoms in the field. *Proc Natl Acad Sci* 2015;**112**:E2182–90.

Aminot A, Kerouel R, Coverly SC. Nutrients in seawater using segmented flow analysis. *Practical Guidelines for the Analysis of Seawater*. Wurl, O. CRC Press, 2009, 143–78.

Bender SJ, Durkin CA, Berthiaume CT *et al.* Transcriptional responses of three model diatoms to nitrate limitation of growth. *Front Mar Sci* 2014;**1**:1–15.

Boissonneault KR, Henningsen BM, Bates SS *et al.* Gene expression studies for the analysis of domoic acid production in the marine diatom Pseudo-nitzschia multiseries. *BMC Mol Biol* 2013;**14**:25.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.

Cohen NR, Ellis KA, Lampe RH *et al.* Diatom Transcriptional and Physiological Responses to Changes in Iron Bioavailability across Ocean Provinces. *Front Mar Sci* 2017;**4**, DOI: 10.3389/fmars.2017.00360.

Cullen JJ, Yang X, MacIntyre HL. Nutrient Limitation of Marine Photosynthesis. *Primary Productivity and Biogeochemical Cycles in the Sea*. Springer, Boston, MA, 1992, 69–88.

Danecek P, Auton A, Abecasis G *et al.* The variant call format and VCFtools. *Bioinformatics* 2011;**27**:2156–8.

Di Dato V, Musacchia F, Petrosino G *et al.* Transcriptome sequencing of three *Pseudo-nitzschia* species reveals comparable gene sets and the presence of Nitric Oxide Synthase genes in diatoms. *Sci Rep* 2015;**5**:12329.

Flynn KJ. Do external resource ratios matter? *J Mar Syst* 2010;**3**–**4**:170–80.

Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *ArXiv12073907 Q-Bio* 2012.

Grossman AR, Aksoy M. Algae in a phosphorus-limited landscape. In: Plaxton WC, Lambers H (eds.). *Annual Plant Reviews Volume 48*. John Wiley & Sons, Inc., 2015, 337–74.

Haas BJ, Papanicolaou A, Yassour M *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 2013;**8**:1494–512.

Hockin NL, Mock T, Mulholland F *et al.* The Response of Diatom Central Carbon Metabolism to Nitrogen Starvation Is Different from That of Green Algae and Higher Plants. *Plant Physiol* 2012;**158**:299–312.

Keller MD, Selvin RC, Claus W *et al.* Media for the Culture of Oceanic Ultraphytoplankton1,2. *J Phycol* 1987;**23**:633–8.

Klein C, Claquin P, Bouchart V *et al.* Dynamics of Pseudo-nitzschia spp. and domoic acid production in a macrotidal ecosystem of the Eastern English Channel (Normandy, France). *Harmful Algae* 2010;**9**:218–26.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.

Lelong A, Hégaret H, Soudant P *et al.* Pseudo-nitzschia (Bacillariophyceae) species, domoic acid and amnesic shellfish poisoning: revisiting previous paradigms. *Phycologia* 2012;**51**:168–216.

Lelong A, Hégaret H, Soudant P. Link between Domoic Acid Production and Cell Physiology after Exchange of Bacterial Communities between Toxic Pseudo-nitzschia multiseries and Non-Toxic Pseudo-nitzschia delicatissima. *Mar Drugs* 2014;**12**:3587–607.

Lema KA, Latimier M, Nézan É *et al.* Inter and intra-specific growth and domoic acid production in relation to nutrient ratios and concentrations in Pseudo-nitzschia: phosphate an important factor. *Harmful Algae* 2017;**64**:11–9.

Levitan O, Dinamarca J, Zelzion E *et al.* Remodeling of intermediate metabolism in the diatom Phaeodactylum tricornutum under nitrogen stress. *Proc Natl Acad Sci* 2015;**112**:412–7.

Li H, Handsaker B, Wysoker A *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.

Lin S, Litaker RW, Sunda WG. Phosphorus physiological ecology and molecular mechanisms in marine phytoplankton. *J Phycol* 2016;**52**:10–36.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.

Maheswari U, Jabbari K, Petit J-L *et al.* Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome Biol* 2010;**11**:R85.

Parkinson JE, Baumgarten S, Michell CT *et al.* Gene Expression Variation Resolves Species and Individual Strains among Coral-Associated Dinoflagellates within the Genus Symbiodinium. *Genome Biol Evol* 2016;**8**:665–80.

Quast C, Pruesse E, Yilmaz P *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;**41**:D590–6.

Ridley AJ. Life at the Leading Edge. *Cell* 2011;**145**:1012–22.

Rines JEB, Donaghay PL, Dekshenieks MM *et al.* Thin layers and camouflage: hidden Pseudo-nitzschia spp. (Bacillariophyceae) populations in a fjord in the San Juan Islands, Washington, USA. *Mar Ecol Prog Ser* 2002;**225**:123–37.

Ritchie ME, Phipson B, Wu D *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47–e47.

Roberts AJ, Kon T, Knight PJ *et al.* Functions and mechanics of dynein motor proteins. *Nat Rev Mol Cell Biol* 2013;**14**:713–26.

Savage TJ, Smith GJ, Clark AT *et al.* Condensation of the isoprenoid and amino precursors in the biosynthesis of domoic acid. *Toxicon* 2012;**59**:25–33.

Si Y, Liu P, Li P *et al.* Model-based clustering for RNA-seq data. *Bioinformatics* 2014;**30**:197–205.

Sison-Mangus MP, Jiang S, Tran KN *et al.* Host-specific adaptation governs the interaction of the marine diatom, Pseudo-nitzschia and their microbiota. *ISME J* 2014;**8**:63–76.

Smith SR, Abbriano RM, Hildebrand M. Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways. *Algal Res* 2012;**1**:2–16.

Tammilehto A, Nielsen TG, Krock B *et al.* Induction of domoic acid production in the toxic diatom Pseudo-nitzschia seriata by calanoid copepods. *Aquat Toxicol* 2015;**159**:52–61.

Teng ST, Tan SN, Lim HC *et al.* High diversity of Pseudo-nitzschia along the northern coast of Sarawak (Malaysian Borneo), with descriptions of P. bipertita sp. nov. and P. limii sp. nov. (Bacillariophyceae). *J Phycol* 2016;**52**:973–89.

Thessen AE, Bowers HA, Stoecker DK. Intra- and interspecies differences in growth and toxicity of Pseudo-nitzschia while using different nitrogen sources. *This Issue Contains Spec Sect Strains* 2009;**8**:792–810.

Thorel M, Claquin P, Schapira M *et al.* Nutrient ratios influence variability in Pseudo-nitzschia species diversity and particulate domoic acid production in the Bay of Seine (France). *Harmful Algae* 2017;**68**:192–205.

Trainer VL, Bates SS, Lundholm N *et al.* Pseudo-nitzschia physiological ecology, phylogeny, toxicity, monitoring and impacts on ecosystem health. *Harmful Algae-- Requir Species-Specif Inf* 2012;**14**:271–300.

Wang J, Cao S, Du C *et al.* Underwater locomotion strategy by a benthic pennate diatom Navicula sp. *Protoplasma* 2013;**250**:1203–12.

Zabaglo K, Chrapusta E, Bober B *et al.* Environmental roles and biological activity of domoic acid: A review. *Algal Res* 2016;**13**:94–101.

Figure legends

Figure 1: Cellular level phenotypes (a-b) and nutrient concentration (c-e) for each of the seven strains (x axis) in the three media (colour coded, see legend). Each dot corresponds to a replicate. Cell density (a) in cell.ml^{-1}, total domoic acid (b) in ng.cell^{-1}, nutrient concentration (c-e) is given in µmol.l^{-1}. All measures were taken at the time of RNA extraction.

Figure 2: Effect size (percentage of the variance explained) of the media, strain, species and their interaction on the cellular level phenotypes (Cell density, Total domoic acid) and nutrient concentration (Silicate, Phosphate, Nitrate) at the time of RNA extraction. The cumulated effect sizes do not reach 1, as a few percent of the variances (the residuals) are not explained by the explanatory variables of the models.

Figure 3: Expression profiles, across strains and media, of the ten clusters of transcripts summarizing the changes in expression observed within *P. autralis*. Expression is indicated using violin plots (rotated kernel density plots) based on the log2 fold change of the median expression in the category of interest over the median expression considering all samples (see methods). Black dots indicate median log2 fold change. Blue, Yellow and Red report to X1, X2, and X3 media, respectively. The number of transcripts belonging to each cluster is indicated.

Figure 4: GO categories with overepresented DE transcripts in the clusters identified within *P. australis*. Each bubble correspond to a GO category, x axes indicate the proportion of DE transcripts in each category, y axes indicate the probability of each GO category to be a false

positive (-log10(q-value)), the size of the bubble represent the number of transcripts in each GO category, and the names of the ten GO categories with the lowest probability to be a false positive (lowest q-value) are indicated . Clusters Pa_C7-10 did not display any overrepresented functional category.

Figure 5: For the six main transcript clusters identified within *P. australis*, heatmaps representing log2 fold change (see method) in gene expression for the 50 transcripts with the highest absolute fold change between the strain Pa1 and the others (Pa_C1), the media X2 and the others (Pa_C2), the media X1 and the others (Pa_C3), the media X3 and the others (Pa_C4), the media X3 and the others (Pa_C5), and the media X1 and the others (Pa_C6). Data for the small strain specific clusters Pa_C7-10 are not displayed. Transcript homologs are indicated (NA indicates that no homolog could be identified in the UNIPROT database).

Figure 6: Inter-specific comparison of gene expression profiles. a. figures on the outer circle represent the expression profiles, across strains and media, of the clusters of transcripts summarizing the changes in expression observed for P. fraudulenta (fc1-3), P. pungens (pc1-5), and P. australis (ac1-5), see legends of Fig 6 for more details. Widths of the ribbons in the inner circle correspond to the number of transcripts shared between clusters. Ribbon colors indicate the odd ratio of the overlap, with purple indicating excess of shared transcripts and orange indicating that less shared transcripts were observed than randomly expected (see color scale). The number of transcripts in each cluster is indicated. b. coloured dots indicate the overrepresented GO categories in each of the clusters, with coloured lines indicating overrepresented GO categories shared between clusters, the name of the shared categories are indicated.

Figure 7: Median fold change (log2) of the central metabolism transcripts for each of the three *Pseudo-nitzschia* species in the three media. Each heatmap corresponds to one transcript, transcript abbreviated names are in red, key intermediates are written in black.

Table 1: Reference transcriptomes

	Samples	Transcripts	Total length	Annotated	Considered for DE analyses	Mean coverage per transcript**	Considered for clustering
P. australis	35*	99,935	75,394,910	38,919	32,371	26-74-345	20,663
P. fraudulenta	12	50,449	43,415,564	17,265	16,045	64-375-1343	2,057
P. pungens	22	65,902	46,434,497	20,737	20,535	37-145-685	3,390
Interspecific	69	6,557	Pa :16,840,898 Pf :19,994,148 Pp :16,412,818	4,151	6,557	394-877-1,951	2,325

*For computing reasons, only 18 samples (2 replicates per strain and per condition) were used to obtain the reference transcriptome.

**First quartile, Median, and Third quartile of the mean coverage per transcripts considered for DE analyses, respectively.

Table 2: Transcriptome wide genetic distances between strains

Species	Strains	Number of SNPs	Pairwise genetic distance
P. australis	Pa1, Pa2	14,878	$3.00\ 10^{-4}$
	Pa1, Pa3	15,283	$3.08\ 10^{-4}$
	Pa2, Pa3	14,655	$2.95\ 10^{-4}$
P. fraudulenta	Pf1, Pf2	24,132	$4.77\ 10^{-4}$
P. pungens	Pp1, Pp2	103,723	$2.07\ 10^{-3}$

Supplementary Figure Legends:

Supplementary Figure 1: Expression profiles, across strains and media, of the seven clusters of transcripts summarizing the changes in expression observed within P. fraudulenta. Expression is indicated using violin plots (rotated kernel density plots) based on the log2 fold change of the median expression in the category of interest over the median expression considering all samples (see methods). Black dots indicate median log2 fold change. Yellow and Red report to X2, and X3 media, respectively. The number of transcripts belonging to each cluster is indicated. Note that for Pf1_X2 a single replicate was available, therefore the log2 fold change of this replicate and not the median log2 fold change is indicated.

Supplementary Figure 2: GO categories with overepresented DE transcripts in the clusters identified within P. fraudulenta. Each bubble correspond to a GO category, x axes indicate the proportion of DE transcripts in each category, y axes indicate the probability of each GO category to be a false positive (-log10(q-value)), the size of the bubble represent the number of transcripts in each GO category, and the names of the ten GO categories with the lowest probability to be a false positive (lowest q-value) are indicated .

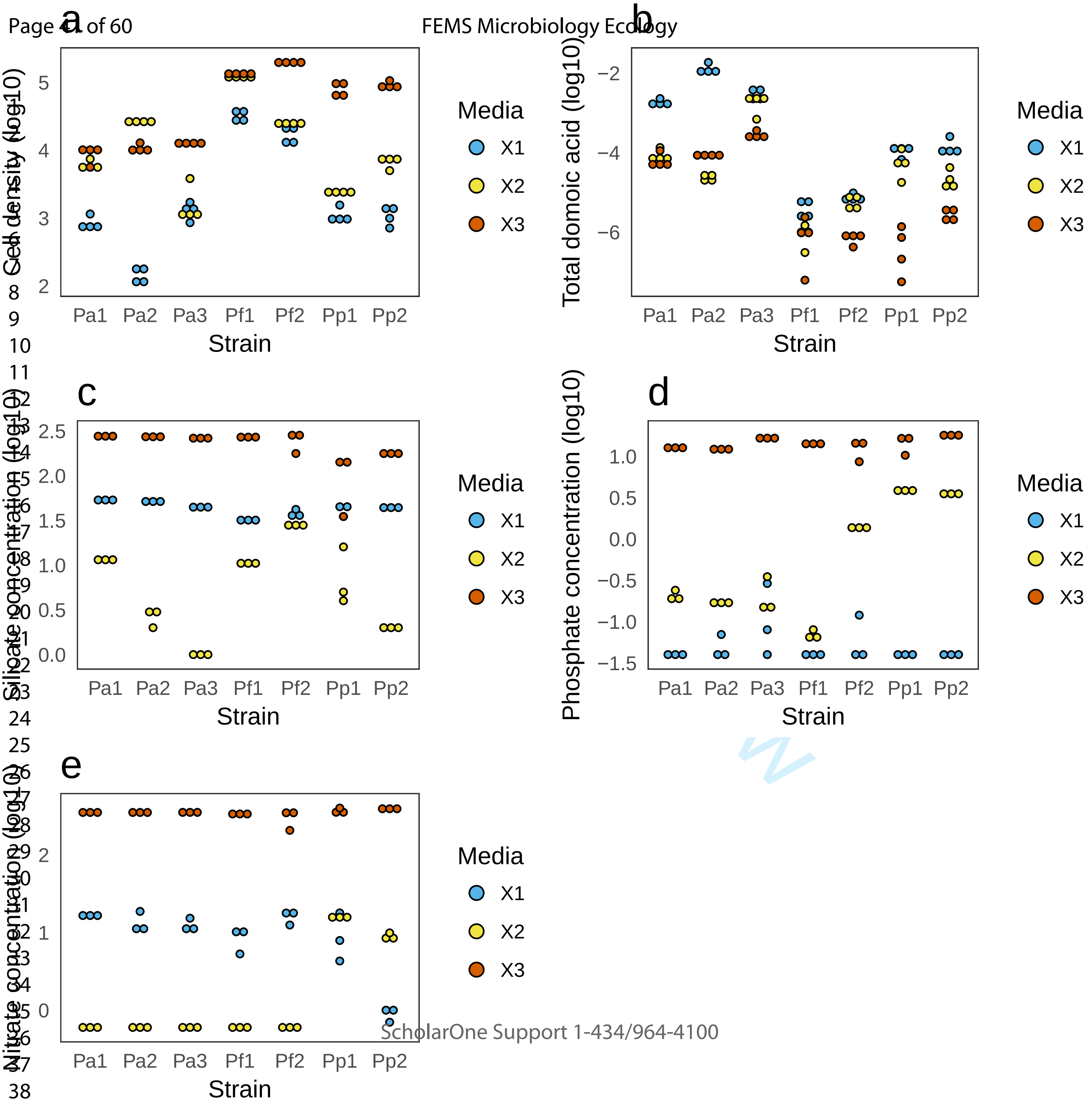
Supplementary Figure 3: For the two transcript clusters identified within *P. fraudulenta,* heatmaps representing log2 fold change (see method) in gene expression for the 50 transcripts with the highest absolute fold change between the media X2 and X3. Transcript homologs are indicated (NA indicates that no homolog could be identified in the UNIPROT database).

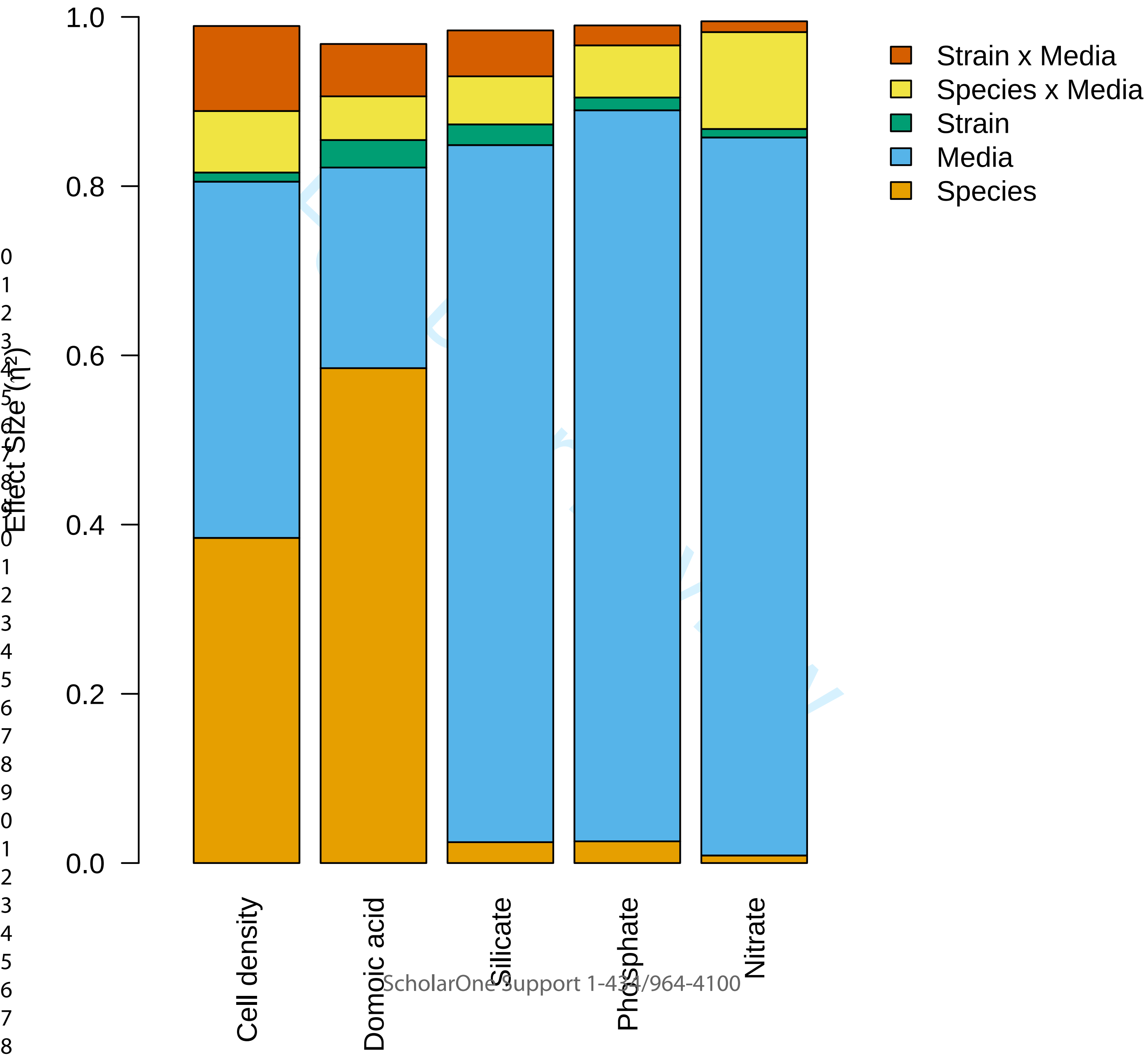
Supplementary Figure 4: Expression profiles, across strains and media, of the seven clusters of transcripts summarizing the changes in expression observed within P. pungens. Expression is indicated using violin plots (rotated kernel density plots) based on the log2 fold change of the median expression in the category of interest over the median expression considering all samples (see methods). Black dots indicate median log2 fold change. Blue, Yellow and Red report to X1, X2, and X3 media, respectively. The number of transcripts belonging to each cluster is indicated.

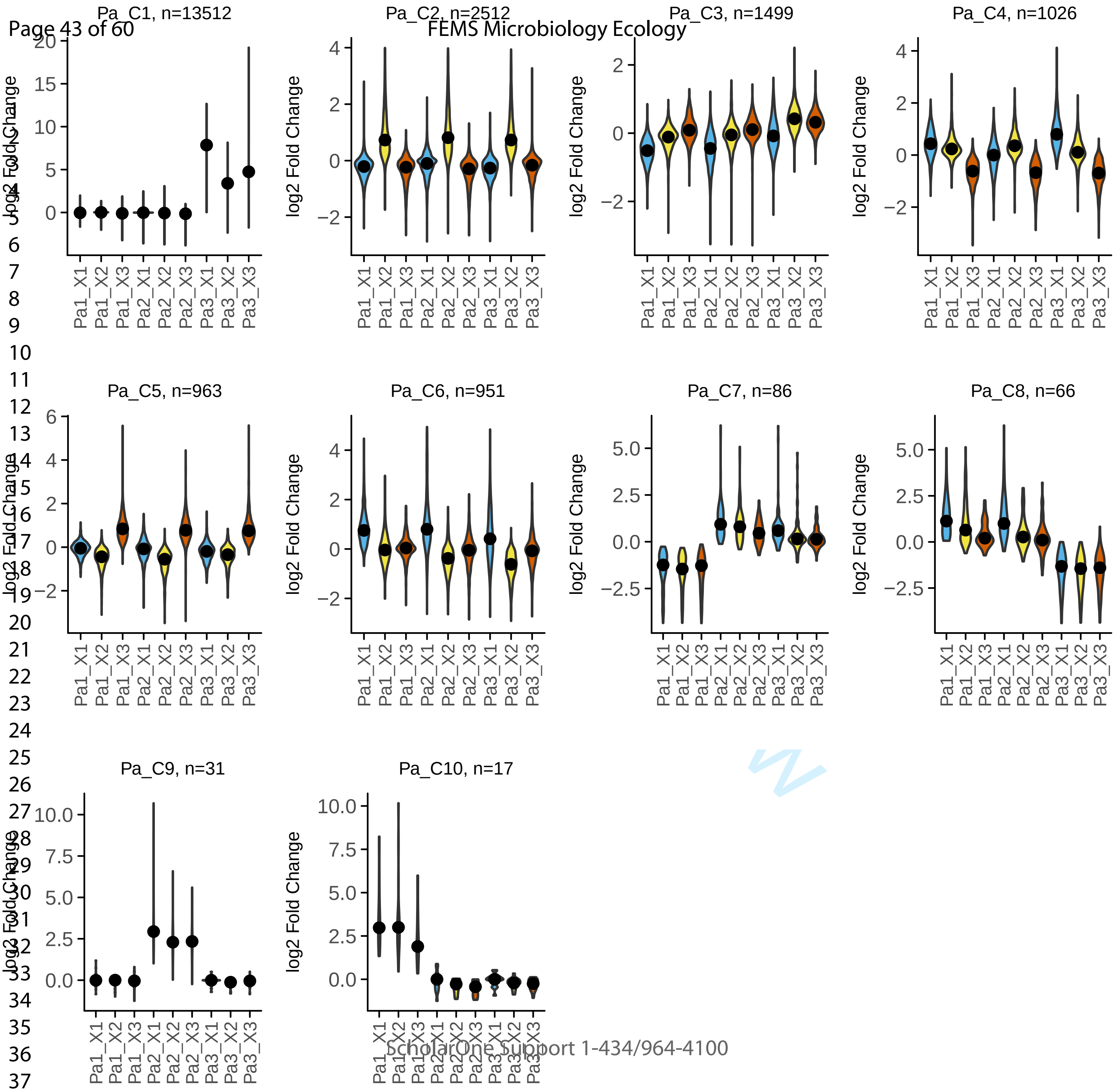
Supplementary Figure 5: GO categories with overepresented DE transcripts in the clusters identified within P. pungens. Each bubble correspond to a GO category, x axes indicate the proportion of DE transcripts in each category, y axes indicate the probability of each GO category to be a false positive (-log10(q-value)), the size of the bubble represent the number of transcripts in each GO category, and the names of the ten GO categories with the lowest probability to be a false positive (lowest q-value) are indicated . Clusters Pp_C6-7 did not display any overrepresented functional category.

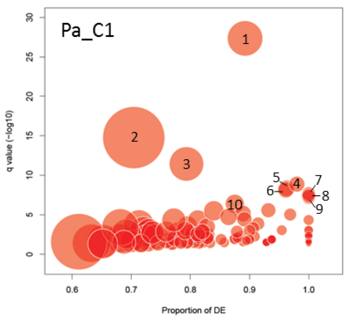
Supplementary Figure 6: For the five main transcript clusters identified within P. pungens, heatmaps representing log2 fold change (see method) in gene expression for the 50 transcripts with the highest absolute fold change between the media X2 and the others (Pp_C1), the media X1 and the others (Pp_C2), the media X3 and the others (Pp_C3), the media X2 and the others (Pp_C4), and the media X1 and the others (Pp_C5). Transcript homologs are indicated (NA indicates that no homolog could be identified in the UNIPROT database). Data for the small strain specific clusters Pp_C6-7 are not displayed.

Supplementary Figure 7: venn diagram representing the overlap in terms of homolog transcripts in the reference transcriptome among the three species.

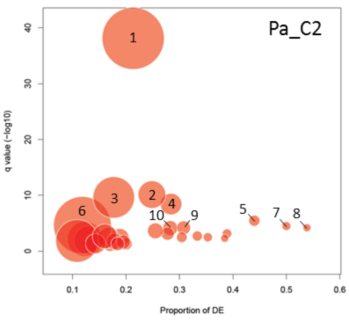

a
Cell density (log10)
2
3
4
5
Pa1
Pa2
Pa3
Pf1
Pf2
Pp1
Pp2
Strain
Media
X1
X2
X3
b
Total domoic acid (log10)
−2
−4
−6
Pa1
Pa2
Pa3
Pf1
Pf2
Pp1
Pp2
Strain
Media
X1
X2
X3
c
Silicate concentration (log10)
0.0
0.5
1.0
1.5
2.0
2.5
Pa1
Pa2
Pa3
Pf1
Pf2
Pp1
Pp2
Strain
Media
X1
X2
X3
d
Phosphate concentration (log10)
−1.5
−1.0
−0.5
0.0
0.5
1.0
Pa1
Pa2
Pa3
Pf1
Pf2
Pp1
Pp2
Strain
Media
X1
X2
X3
e
Nitrate concentration (log10)
0
1
2
Pa1
Pa2
Pa3
Pf1
Pf2
Pp1
Pp2
Strain
Media
X1
X2
X3


Effect Size (η^2)
1.0
0.8
0.6
0.4
0.2
0.0
Cell density
Domoic acid
Silicate
Phosphate
Nitrate
Strain x Media
Species x Media
Strain
Media
Species

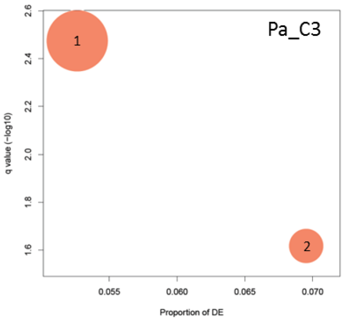

Pa_C1, n=13512
Pa_C2, n=2512
Pa_C3, n=1499
Pa_C4, n=1026
Pa_C5, n=963
Pa_C6, n=951
Pa_C7, n=86
Pa_C8, n=66
Pa_C9, n=31
Pa_C10, n=17
log2 Fold Change
Pa1_X1
Pa1_X2
Pa1_X3
Pa2_X1
Pa2_X2
Pa2_X3
Pa3_X1
Pa3_X2
Pa3_X3



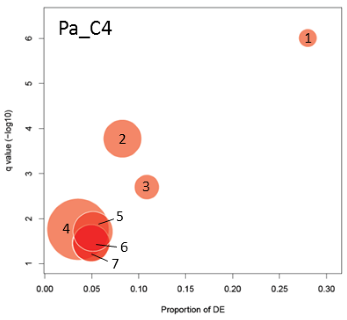
1. cilium, 2. cell. comp., 3. centrosome, 4. ciliary basal body, 5. Rho guanyl-nucl. ex. fact. act., 6. cell leading edge, 7. intr. comp. cytopl. side plasma memb., 8. reg. small GTPase signal transd., 9. pseudopodium, 10. phagocytosis



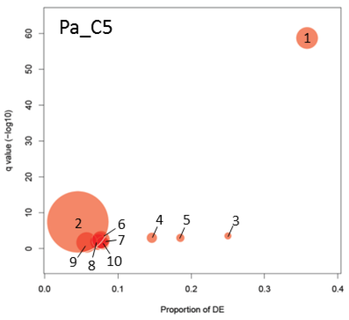
1. RNA binding, 2. cell. aa. biosynth. proc., 3. poly(A) RNA bind., 4. transl. init., 5. DNA-directed RNA pol. III, 6. intracell., 7. nitrate assim., 8. arginine biosynth. proc., 9. glutamine metab. proc., 10. ribos. LSU assembly



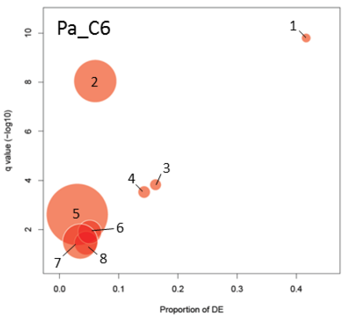
1. chloroplast, 2. iron ion bind.



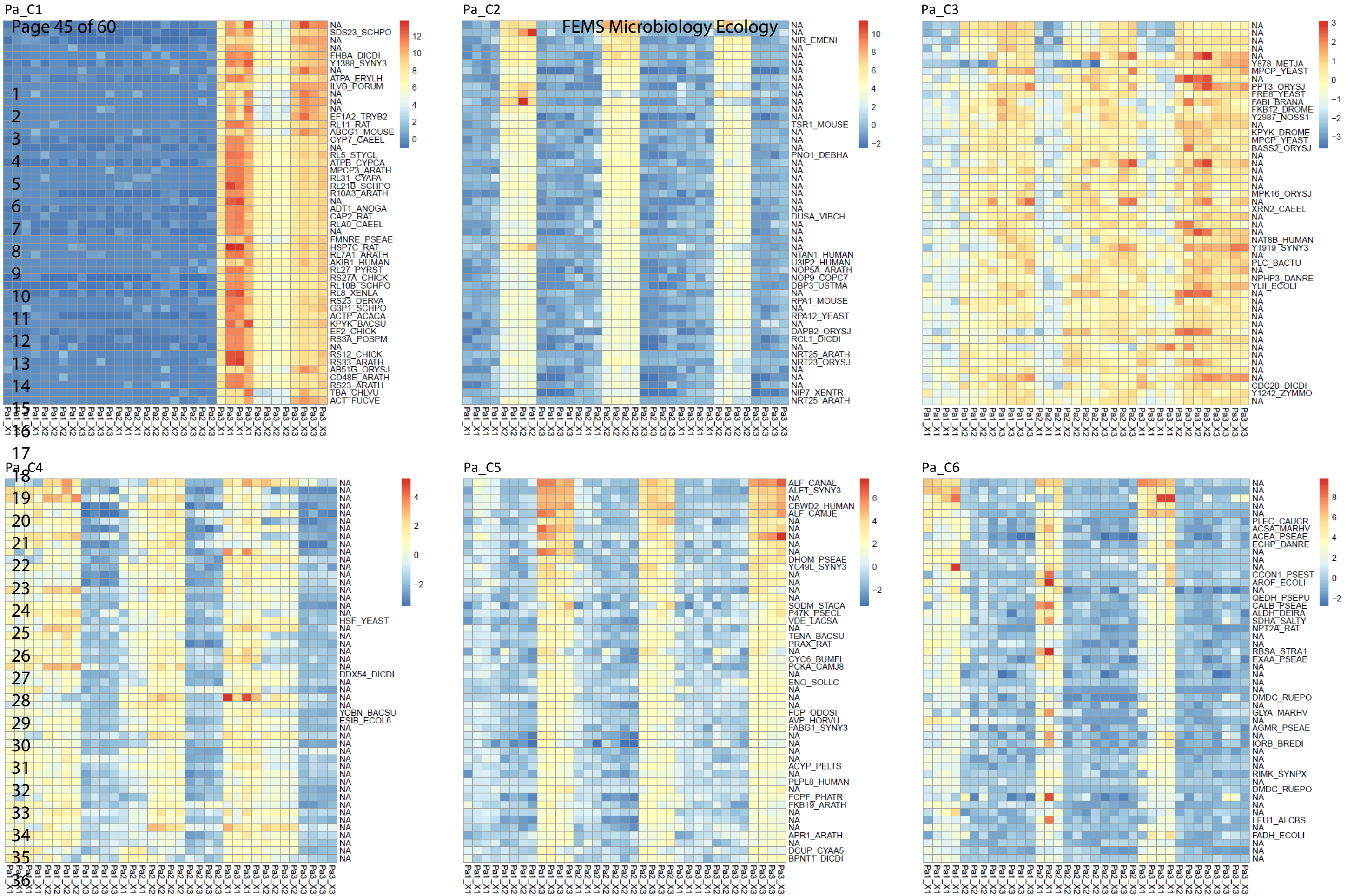
1. extracell. mat., 2. seq.-specific DNA bind., 3. Notch sign. path., 4. extracell. Space, 5. cell redox homeost., 6. nervous syst. dev., 7. cell surf.

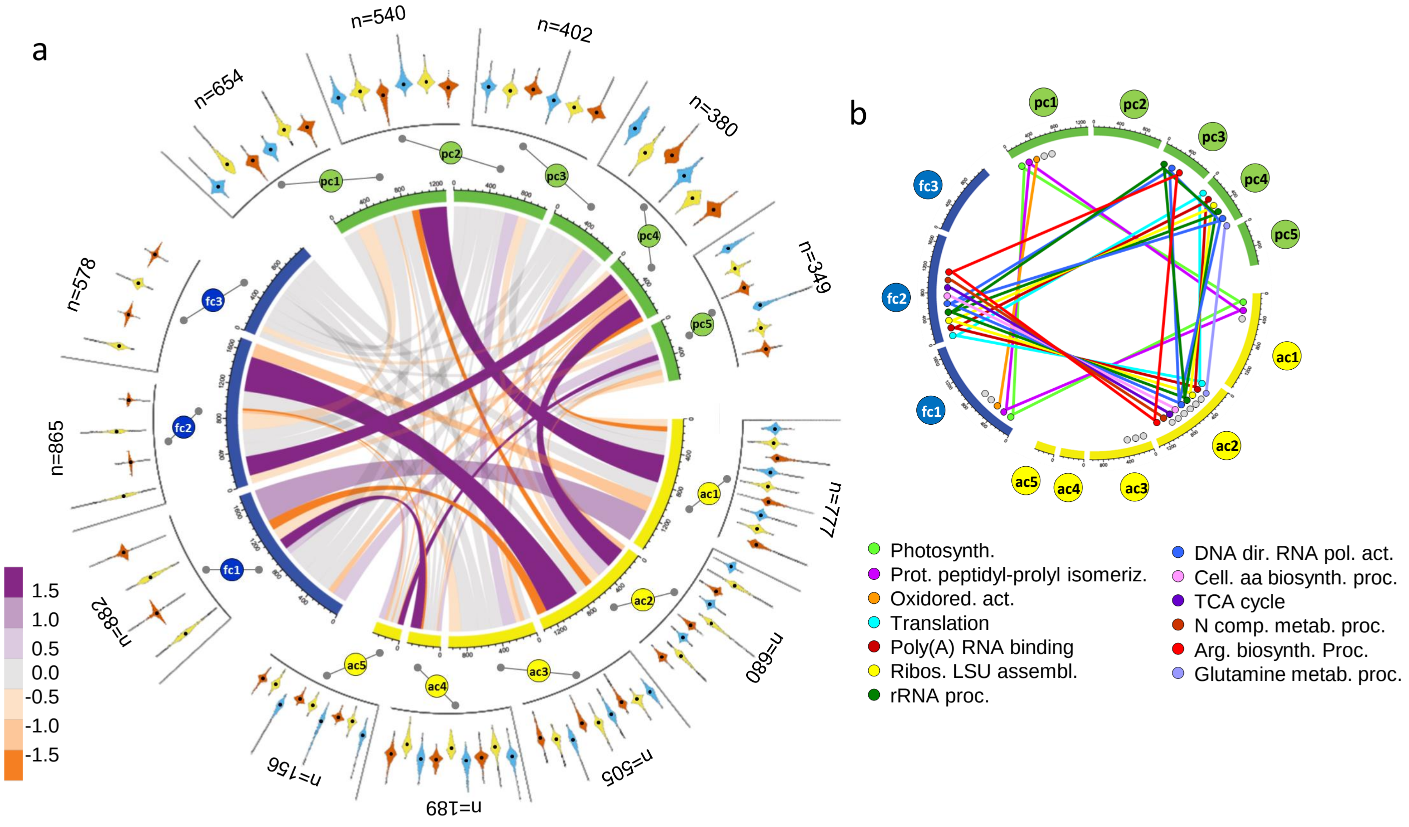


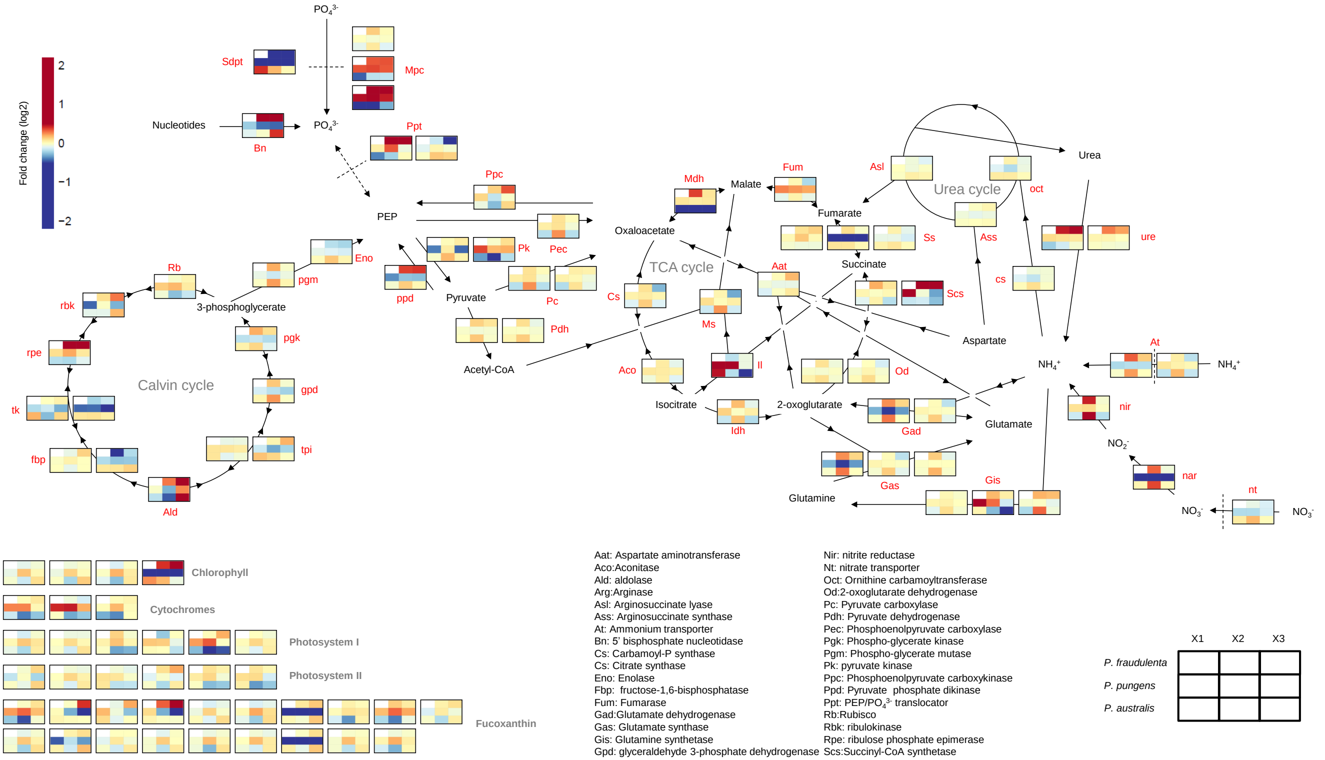
1. thylakoid, 2. oxi.-red. proc., 3. protoporph. IX biosynth. proc., 4. cell. resp. oxi. Stress, 5. chlorophyll biosynth. proc., 6. cell redox homeost., 7. prot. peptidyl-prolyl isomer., 8. glycolytic proc., 9. cell. aa biosynth. proc., 10. transaminase act.



1. DNA integr., 2. plastid, 3. pentose-phosphate shunt, 4. periplasm. space, 5. catalytic act. 6. cell. aa biosynth. proc., 7. lyase act., 8. fatty acid biosynth. proc.

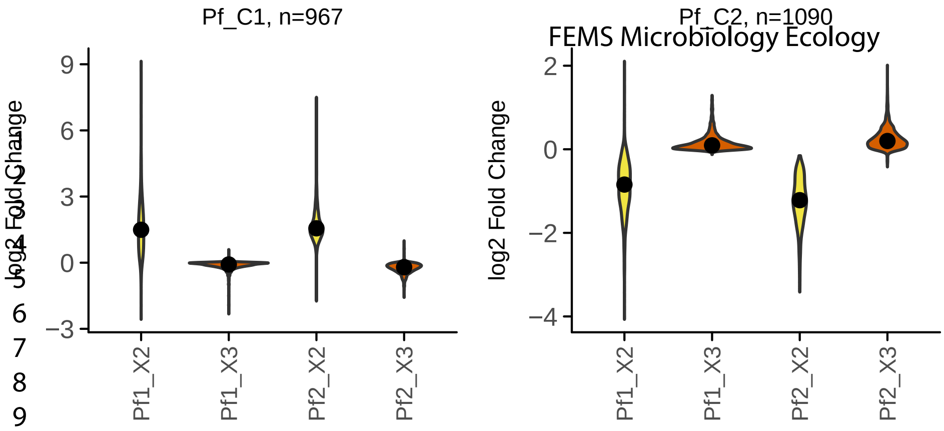

Pa_C1
SDS23_SCHPO
FHBA_DICDI
Y1388_SYNY3
ATPA_ERYLH
ILVB_PORUM
EF1A2_TRYB2
RL11_RAT
ABCG1_MOUSE
CYP7_CAEEL
RL5_STYCL
ATPB_CYPCA
MPCP3_ARATH
RL31_CYAPA
RL21B_SCHPO
R10A3_ARATH
ADT1_ANOGA
CAP2_RAT
RLA0_CAEEL
FMNRE_PSEAE
HSP7C_RAT
RL7A1_ARATH
AKIB1_HUMAN
RL27_PYRST
RS27A_CHICK
RL10B_SCHPO
RL8_XENLA
RS23_DERVA
G3P1_SCHPO
ACTP_ACACA
KPYK_BACSU
EF2_CHICK
RS3A_POSPM
RS12_CHICK
RS33_ARATH
AB51G_ORYSJ
CD48E_ARATH
RS23_ARATH
TBA_CHLVU
ACT_FUCVE
12
10
8
6
4
2
0
Pa_C2
NIR_EMENI
TSR1_MOUSE
PNO1_DEBHA
DUSA_VIBCH
NTAN1_HUMAN
U3IP2_HUMAN
NOP5A_ARATH
NOP9_COPC7
DBP3_USTMA
RPA1_MOUSE
RPA12_YEAST
DAPB2_ORYSJ
RCL1_DICDI
NRT25_ARATH
NRT23_ORYSJ
NIP7_XENTR
NRT25_ARATH
10
8
6
4
2
0
−2
Pa_C3
Y878_METJA
MPCP_YEAST
PPT3_ORYSJ
FRE8_YEAST
FABI_BRANA
FKB12_DROME
Y2987_NOSS1
KPYK_DROME
MPCP_YEAST
BASS2_ORYSJ
MPK16_ORYSJ
XRN2_CAEEL
NAT8B_HUMAN
Y1919_SYNY3
PLC_BACTU
NPHP3_DANRE
YLII_ECOLI
CDC20_DICDI
Y1242_ZYMMO
3
2
1
0
−1
−2
−3
Pa_C4
HSF_YEAST
DDX54_DICDI
YOBN_BACSU
ESIB_ECOL6
4
2
0
−2
Pa_C5
ALF_CANAL
ALF1_SYNY3
CBWD2_HUMAN
ALF_CAMJE
DHOM_PSEAE
YC49L_SYNY3
SODM_STACA
P47K_PSECL
VDE_LACSA
TENA_BACSU
PRAX_RAT
CYC6_BUMFI
PCKA_CAMJ8
ENO_SOLLC
FCP_ODOSI
AVP_HORVU
FABG1_SYNY3
ACYP_PELTS
PLPL8_HUMAN
FCPF_PHATR
FKB19_ARATH
APR1_ARATH
DCUP_CYAA5
BPNT_DICDI
6
4
2
0
−2
Pa_C6
PLEC_CAUCR
ACSA_MARHV
ACEA_PSEAE
ECHP_DANRE
CCON1_PSEST
AROF_ECOLI
QEDH_PSEPU
CALB_PSEAE
ALDH_DEIRA
SDHA_SALTY
NPT2A_RAT
RBSA_STRA1
EXAA_PSEAE
DMDC_RUEPO
GLYA_MARHV
AGMR_PSEAE
IORB_BREDI
RIMK_SYNPX
DMDC_RUEPO
LEU1_ALCBS
FADH_ECOLI
8
6
4
2
0
−2
Pa1_X1
Pa1_X2
Pa1_X3
Pa2_X1
Pa2_X2
Pa2_X3
Pa3_X1
Pa3_X2
Pa3_X3

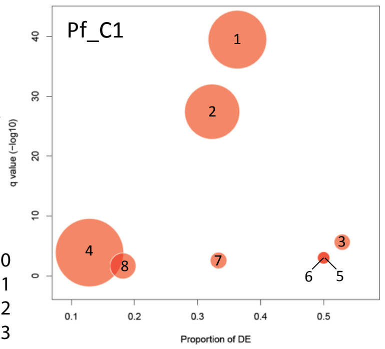

a
n=540
n=402
n=654
n=380
n=349
n=578
n=865
n=777
n=882
n=680
n=505
n=189
n=156
pc1
pc2
pc3
pc4
pc5
fc1
fc2
fc3
ac1
ac2
ac3
ac4
ac5
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
b
Photosynth.
Prot. peptidyl-prolyl isomeriz.
Oxidored. act.
Translation
Poly(A) RNA binding
Ribos. LSU assembl.
rRNA proc.
DNA dir. RNA pol. act.
Cell. aa biosynth. proc.
TCA cycle
N comp. metab. proc.
Arg. biosynth. Proc.
Glutamine metab. proc.



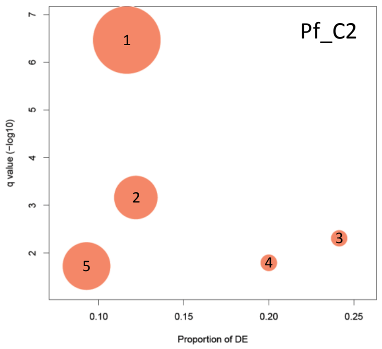
Aat: Aspartate aminotransferase
Aco:Aconitase
Ald: aldolase
Arg:Arginase
Asl: Arginosuccinate lyase
Ass: Arginosuccinate synthase
At: Ammonium transporter
Bn: 5' bisphosphate nucleotidase
Cs: Carbamoyl-P synthase
Cs: Citrate synthase
Eno: Enolase
Fbp: fructose-1,6-bisphosphatase
Fum: Fumarase
Gad:Glutamate dehydrogenase
Gas: Glutamate synthase
Gis: Glutamine synthetase
Gpd: glyceraldehyde 3-phosphate dehydrogenase
Idh: Isocitrate dehydrogenase
Il: Isocitrate lyase
Mdh: Malate dehydrogenase
Mpc: Mitochondrial phosphate carrier
Ms: Malate synthase
Nar:nitrate reductase

Nir: nitrite reductase
Nt: nitrate transporter
Oct: Ornithine carbamoyltransferase
Od:2-oxoglutarate dehydrogenase
Pc: Pyruvate carboxylase
Pdh: Pyruvate dehydrogenase
Pec: Phosphoenolpyruvate carboxylase
Pgk: Phospho-glycerate kinase
Pgm: Phospho-glycerate mutase
Pk: pyruvate kinase
Ppc: Phosphoenolpyruvate carboxykinase
Ppd: Pyruvate phosphate dikinase
Ppt: PEP/PO$_4^{3-}$ translocator
Rb:Rubisco
Rbk: ribulokinase
Rpe: ribulose phosphate epimerase
Scs:Succinyl-CoA synthetase
Sdpt: Sodium-dependent phosphate transport pn 2A
Ss: Succinate synthase
Tk: transketolase
Tpi: Triose phosphate isomerase
Ure:urease

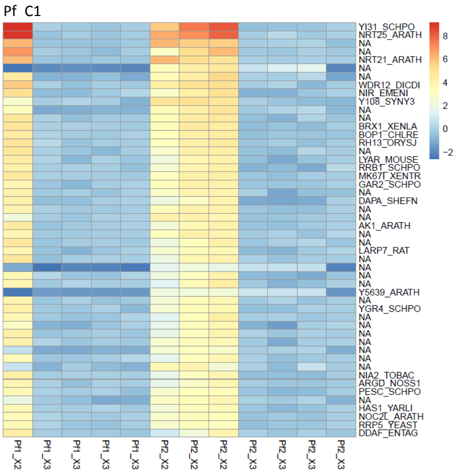

Pf_C1, n=967
Pf_C2, n=1090
log2 Fold Change
9
6
3
0
−3
2
0
−2
−4
Pf1_X2
Pf1_X3
Pf2_X2
Pf2_X3
Pf1_X2
Pf1_X3
Pf2_X2
Pf2_X3

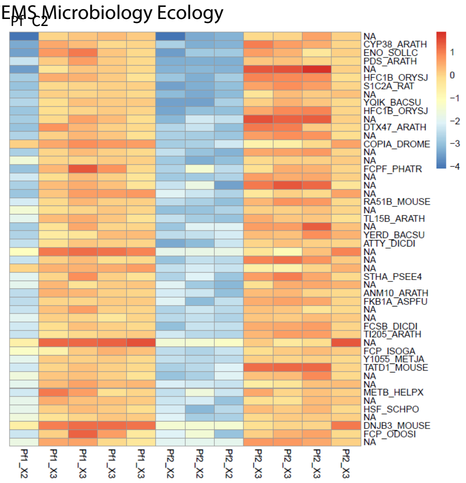


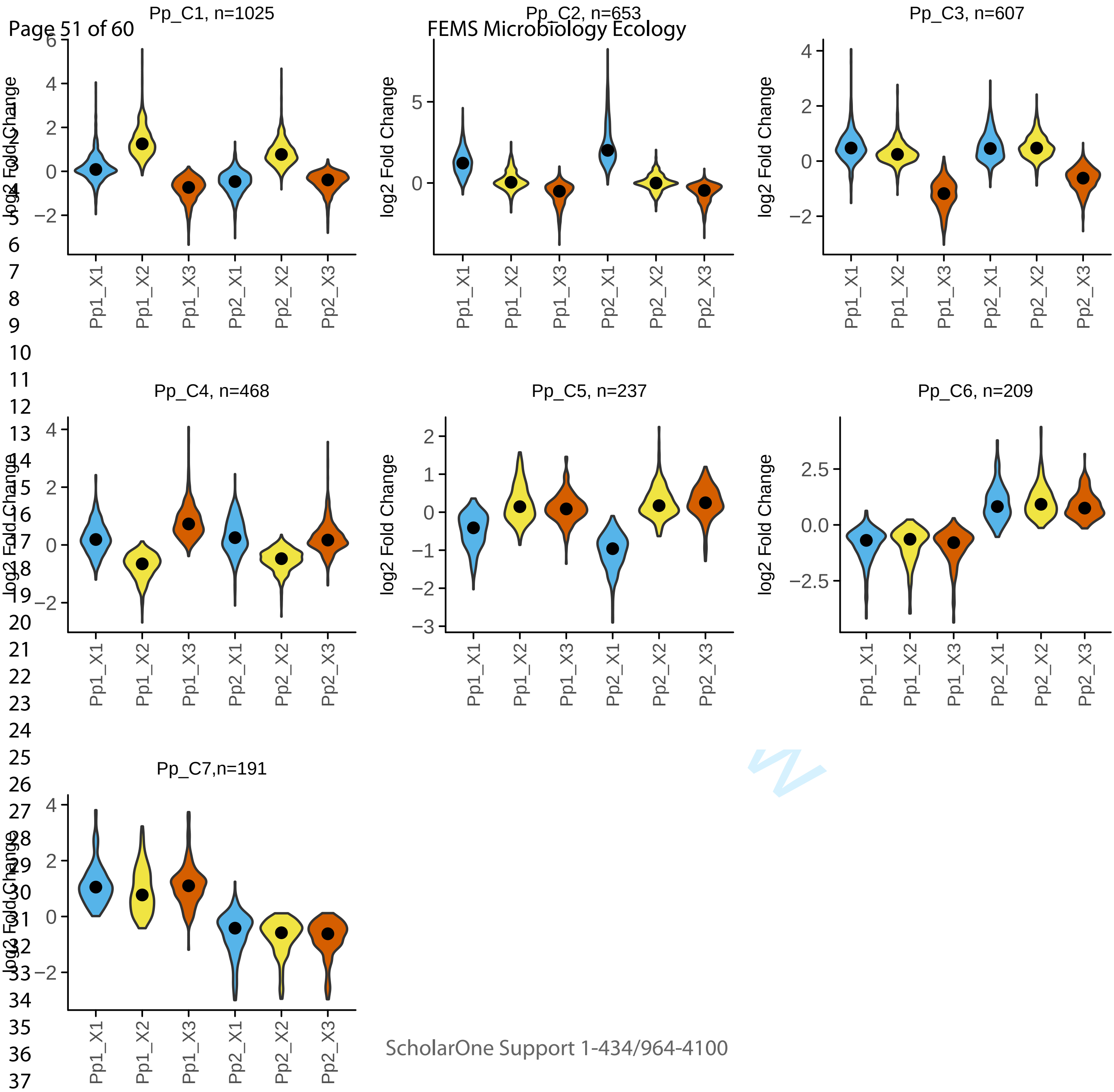
1. nucleol., 2. intracell. ribonucl. compl., 3. ribo. LSU assembly, 4. nucl. ac. bind., 5. nitrogen util., 6. maturat. SSU-rRNA, 7. nitrogen comp. metab. proc., 8. exonucl. act.

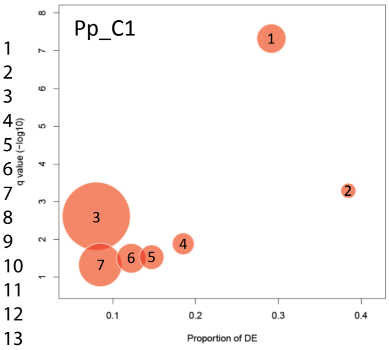


1. plastid., 2. isomer. act., 3. DNA integ., 4. drug transmenb. transp., 5. lipid meta. proc.

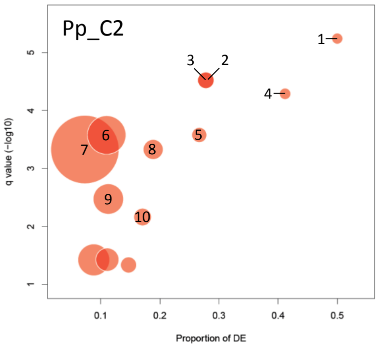

Pf_C1
YI31_SCHPO
NRT25_ARATH
NA
NA
NRT21_ARATH
NA
NA
WDR12_DICDI
NIR_EMENI
Y108_SYNY3
NA
NA
BRX1_XENLA
BOP1_CHLRE
RH13_ORYSJ
NA
LYAR_MOUSE
RRB1_SCHPO
MK67I_XENTR
GAR2_SCHPO
NA
DAPA_SHEFN
NA
NA
AK1_ARATH
NA
NA
LARP7_RAT
NA
NA
NA
NA
Y5639_ARATH
NA
YGR4_SCHPO
NA
NA
NA
NA
NA
NA
NA
NIA2_TOBAC
ARGD_NOSS1
PESC_SCHPO
NA
HAS1_YARLI
NOC2L_ARATH
RRP5_YEAST
DDAF_ENTAG
8
6
4
2
0
−2
Pf1_X2
Pf1_X3
Pf1_X3
Pf1_X3
Pf1_X3
Pf2_X2
Pf2_X2
Pf2_X2
Pf2_X3
Pf2_X3
Pf2_X3
Pf2_X3


Pf_C2
NA
CYP38_ARATH
ENO_SOLLC
PDS_ARATH
NA
HFC1B_ORYSJ
S1C2A_RAT
NA
YQIK_BACSU
HFC1B_ORYSJ
NA
DTX47_ARATH
NA
COPIA_DROME
NA
NA
FCPF_PHATR
NA
NA
NA
RA51B_MOUSE
NA
TL15B_ARATH
NA
YERD_BACSU
ATTY_DICDI
NA
NA
NA
STHA_PSEE4
NA
ANM10_ARATH
FKB1A_ASPFU
NA
NA
FCSB_DICDI
TI205_ARATH
NA
FCP_ISOGA
Y1055_METJA
TATD1_MOUSE
NA
NA
METB_HELPX
NA
HSF_SCHPO
NA
DNJB3_MOUSE
FCP_ODOSI
NA
1
0
−1
−2
−3
−4
Pf1_X2
Pf1_X3
Pf1_X3
Pf1_X3
Pf1_X3
Pf2_X2
Pf2_X2
Pf2_X2
Pf2_X3
Pf2_X3
Pf2_X3
Pf2_X3

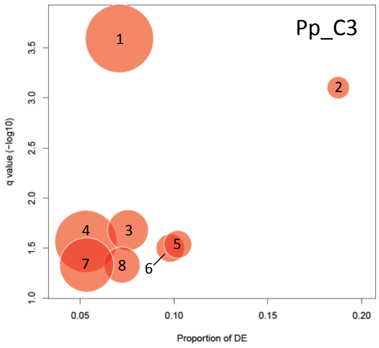

Pp_C1, n=1025
Pp_C2, n=653
Pp_C3, n=607
Pp_C4, n=468
Pp_C5, n=237
Pp_C6, n=209
Pp_C7,n=191
log2 Fold Change
Pp1_X1
Pp1_X2
Pp1_X3
Pp2_X1
Pp2_X2
Pp2_X3
6
4
2
0
−2
5
−2.5
0.0
2.5
1
−1
−3



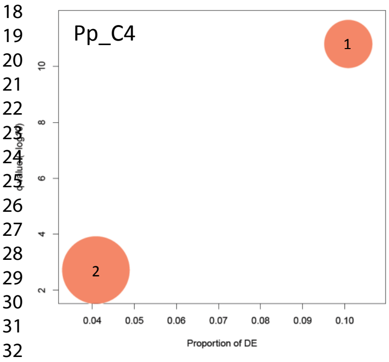
1. photosynth., 2. protein. extracell. mat., 3. prot. ser./thre. kinase act., 4. cilium morphogen., 5. cell prolif., 6. defense resp., 7. extracell. space



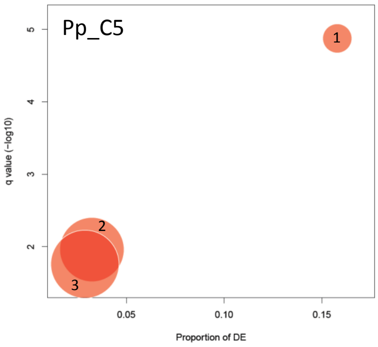
1. outer memb., 2. tricarbox. acid cycle, 3. periplasm. Space, 4. sigma fact. act., 5. oxidoreduct. act., 6. lyase act., 7. catalytic act., 8. rRNA bind., 9. Fe-S clust. bind., 10. aa transp.



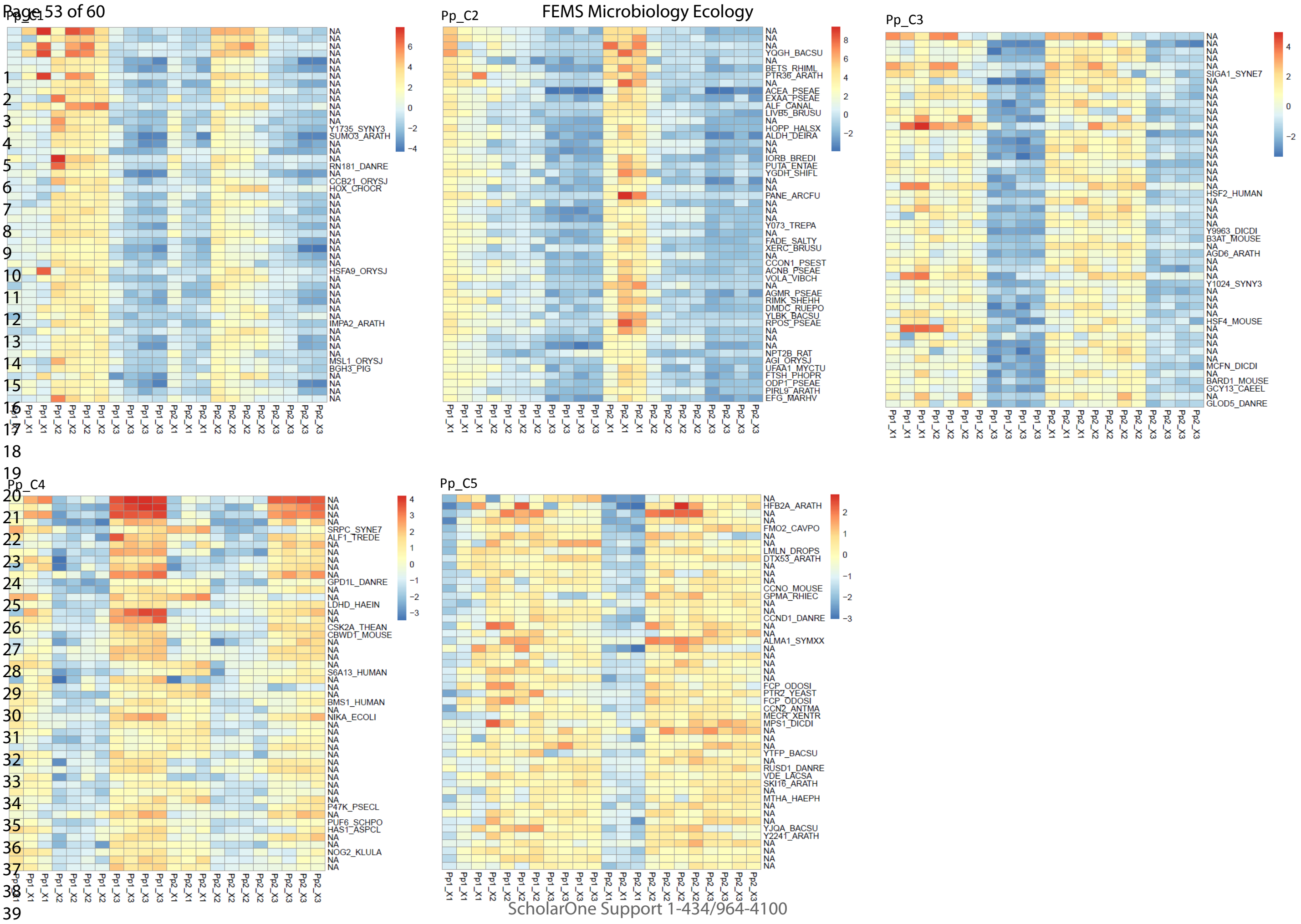
1. prot. kinase act., 2. signal transd. prot. phosphoryl., 3. extracell. space, 4. cell. comp., 5. cell wall organ., 6. lipid catab. proc., 7. extracell. exosome, 8. pos. reg. transcr. RNA pol. II promot.

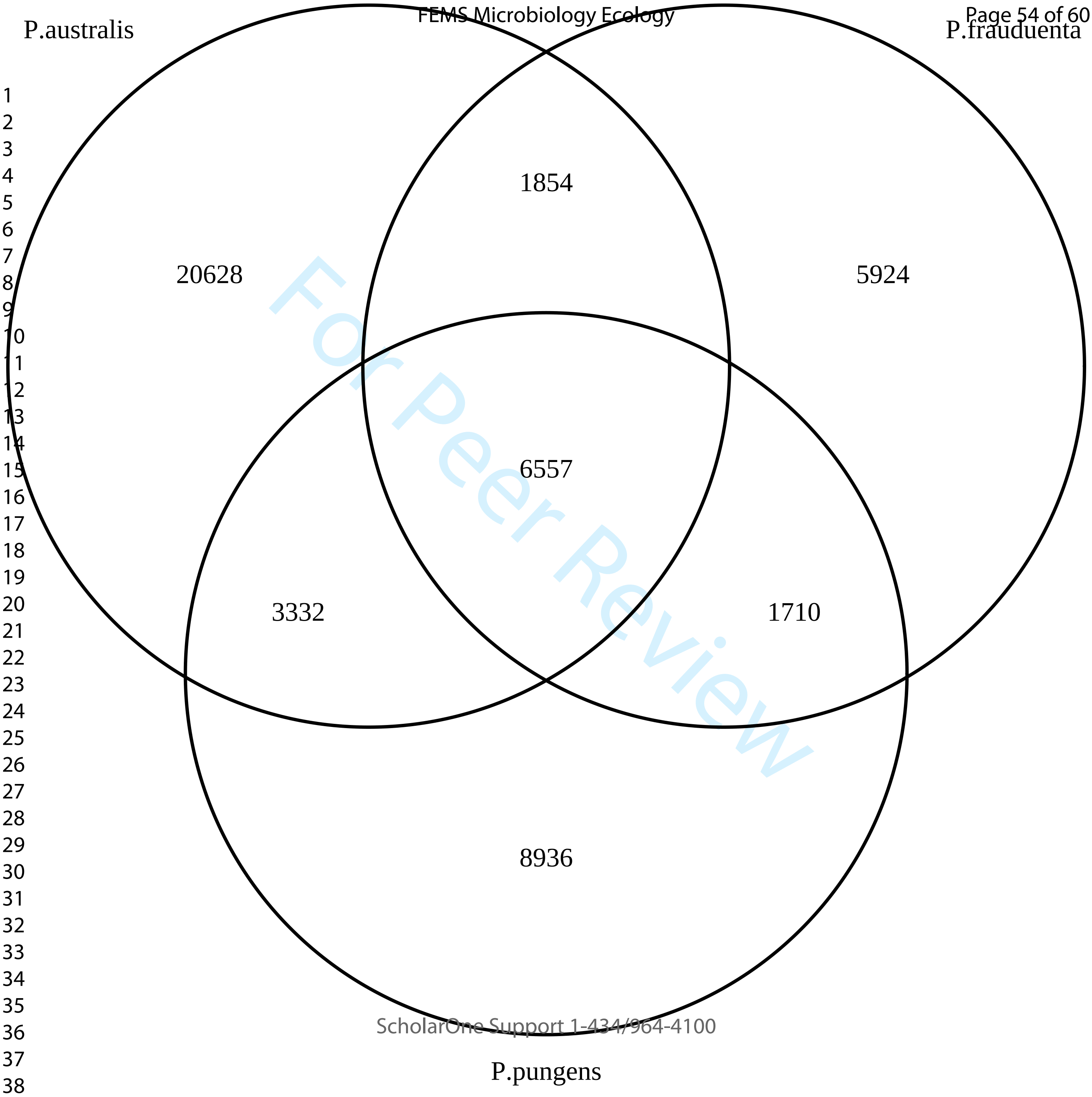


1.nucleolus, 2. RNA bind.



1. light-harvest. Compl., 2. extracell. reg., 3. cell cycle .


Pp_C1
Y1735_SYNY3
SUMO3_ARATH
RN181_DANRE
CCB21_ORYSJ
HOX_CHOCR
HSFA9_ORYSJ
IMPA2_ARATH
MSL1_ORYSJ
BGH3_PIG
6
4
2
0
−2
−4
Pp_C2
YOGH_BACSU
BETS_RHIML
PTR36_ARATH
ACEA_PSEAE
EXAA_PSEAE
ALF_CANAL
LIVB5_BRUSU
HOPP_HALSX
ALDH_DEIRA
IORB_BREDI
PUTA_ENTAE
YGDH_SHIFL
PANE_ARCFU
Y073_TREPA
FADE_SALTY
XERC_BRUSU
CCON1_PSEST
ACNB_PSEAE
VOLA_VIBCH
AGMR_PSEAE
RIMK_SHEHH
DMDC_RUEPO
YLBK_BACSU
RPOS_PSEAE
NPT2B_RAT
AGI_ORYSJ
UFAA1_MYCTU
FTSH_PHOPR
ODP1_PSEAE
PIRL9_ARATH
EFG_MARHV
8
6
4
2
0
−2
Pp_C3
SIGA1_SYNE7
HSF2_HUMAN
Y9963_DICDI
B3AT_MOUSE
AGD6_ARATH
Y1024_SYNY3
HSF4_MOUSE
MCFN_DICDI
BARD1_MOUSE
GCY13_CAEEL
GLOD5_DANRE
4
2
0
−2
Pp_C4
SRPC_SYNE7
ALF1_TREDE
GPD1L_DANRE
LDHD_HAEIN
CSK2A_THEAN
CBWDT_MOUSE
S6A13_HUMAN
BMS1_HUMAN
NIKA_ECOLI
P47K_PSECL
PUF6_SCHPO
HAST_ASPCL
NOG2_KLULA
4
3
2
1
0
−1
−2
−3
Pp_C5
HFB2A_ARATH
FMO2_CAVPO
LMLN_DROPS
DTX53_ARATH
CCNO_MOUSE
GPMA_RHIEC
CCND1_DANRE
ALMA1_SYMXX
FCP_ODOSI
PTR2_YEAST
FCP_ODOSI
CCN2_ANTMA
MECR_XENTR
MPS1_DICDI
YTFP_BACSU
RUSD1_DANRE
VDE_LACSA
SKI16_ARATH
MTHA_HAEPH
YJOA_BACSU
Y2241_ARATH
2
1
0
−1
−2
−3
NA
Pp1_X1
Pp1_X2
Pp1_X3
Pp2_X1
Pp2_X2
Pp2_X3


P.australis
P.frauduenta
1854
20628
5924
6557
3332
1710
8936
P.pungens

Supplementary Table 1. Cultures of *Pseudo-nitzschia* spp. used in the present study (nb: the collection references of the strains have also been added to enable comparisons with other studies). (A) stands for strains isolated in the Atlantic and (EC) for strains from the English Channel.

Species	Collection reference	Name present study	Collection date	Station	GPS Coordinates (approximative)
P. australis	P1D2	Pa1	28/03/14	Anse de Dinan, Camaret-sur-Mer (A)	48.225679; -4.563602
P. australis	P3B2	Pa2	04/04/14	Môle St Anne, Plouzané (A)	48.358573; -4.551193
P.australis	IFR-PAU-010	Pa3	14/07/15	Ouessant (A)	48.449511; -5.108088
P. fraudulenta	PNfra2	Pf1	24/08/11	Cabourg (EC)	49.302888; -0.103549
P. fraudulenta	PNfra31	Pf2	11/07/11	COMOR 41(EC)	49.414444; -0.408889
P. pungens	PNpun47	Pp1	24/08/11	Cabourg (EC)	49.302888; -0.103549
P. pungens	PNpun102	Pp2	21/08/12	Luc sur Mer (EC)	49.320109; -0.351193

Supplementary Table 2: For each pairwise gene expression analysis, number of transcripts identified

Intraspecific

P. australis

Pa1	Pa2	Pa3	X1	X2	X3	DESeq2
X			X	X		3497
X			X		X	4851
X				X	X	5240
	X		X	X		4115
	X		X		X	3616
	X			X	X	5952
		X	X	X		19443
		X	X		X	10358
		X		X	X	10204
X	X		X			824
X	X			X		731
X	X				X	740
X		X	X			16153
X		X		X		15324
X		X			X	16047
	X	X	X			15954
	X	X		X		15673
	X	X			X	15977

P. pungens

Pp1	Pp2	X1	X2	X3	DESeq2	limma
X		X	X		580	30
X		X		X	2169	1571
X			X	X	3394	3610
	X	X	X		1965	1292
	X	X		X	3006	2764
	X		X	X	2381	2502
X	X	X			501	12
X	X		X		606	276
X	X			X	1774	1804

P.fraudulenta

Pf1	Pf2	X1	X2	X3	DESeq2	limma
X		X	X		NA	NA
X		X		X	NA	NA
X			X	X	NA	NA
	X	X	X		NA	NA
	X	X		X	NA	NA
	X		X	X	5230	5848
X	X	X			NA	NA
X	X		X		NA	NA

X	X			X	83	0

Interspecific

P. australis

Pa1	Pa2	Pa3	X1	X2	X3	DESeq2
X			X	X		710
X			X		X	774
X				X	X	967
	X		X	X		762
	X		X		X	638
	X			X	X	1053
		X	X	X		788
		X	X		X	733
		X		X	X	951
X	X		X			78
X	X			X		58
X	X				X	50
X		X	X			180
X		X		X		167
X		X			X	196
	X	X	X			210
	X	X		X		171
	X	X			X	210

P. pungens

Pp1	Pp2	X1	X2	X3	DESeq2	limma
X		X	X		92	7
X		X		X	296	156
X			X	X	473	442
	X	X	X		245	130
	X	X		X	345	302
	X		X	X	285	242
X	X	X			29	0
X	X		X		32	0
X	X			X	190	160

P.fraudulenta

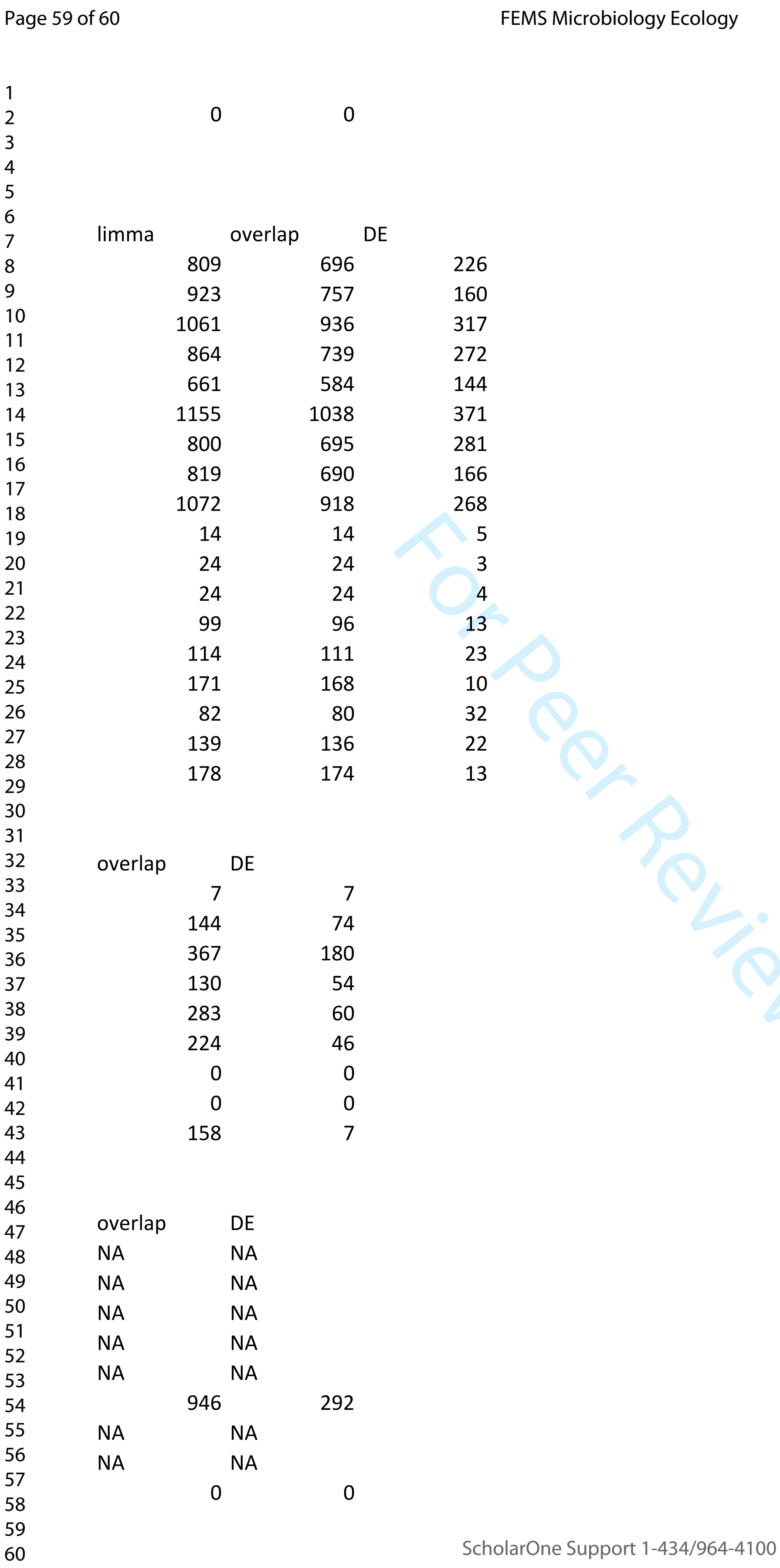
Pf1	Pf2	X1	X2	X3	DESeq2	limma
X		X	X		NA	NA
X		X		X	NA	NA
X			X	X	NA	NA
	X	X	X		NA	NA
	X	X		X	NA	NA
	X		X	X	987	1075
X	X	X			NA	NA
X	X		X		NA	NA
X	X			X	4	0

as DE using DESeq2, limma, considering the overlap between the two methods, and concidering tran

limma	overlap	DE
4079	3216	1190
5645	4446	1162
19866	5134	1792
4435	3577	1605
3878	3150	973
6833	5685	2262
23301	19321	12878
14386	9950	2270
13801	10105	3594
498	408	205
14072	727	221
669	578	199
16224	15742	14094
16033	14568	12436
17271	15958	13863
15674	15031	14318
16594	15248	12990
16936	15895	13945

overlap	DE
30	30
1438	913
2942	1801
1237	677
2554	948
2128	739
12	12
274	223
1557	386

overlap	DE
NA	NA
NA	NA
NA	NA
NA	NA
NA	NA
5010	2057
NA	NA
NA	NA



0	0

limma	overlap	DE
809	696	226
923	757	160
1061	936	317
864	739	272
661	584	144
1155	1038	371
800	695	281
819	690	166
1072	918	268
14	14	5
24	24	3
24	24	4
99	96	13
114	111	23
171	168	10
82	80	32
139	136	22
178	174	13

overlap	DE
7	7
144	74
367	180
130	54
283	60
224	46
0	0
0	0
158	7

overlap	DE
NA	NA
NA	NA
NA	NA
NA	NA
NA	NA
946	292
NA	NA
NA	NA
0	0

scripts with a log2 fold change > 2

References

Abu-Ali, G. S., Mehta, R. S., Lloyd-Price, J., Mallick, H., Branck, T., Ivey, K. L., … Huttenhower, C. (2018). Metatranscriptome of human faecal microbial communities in a cohort of adult men. *Nature Microbiology*, *3*(3), 356–366. Retrieved from http://dx.doi.org/10.1038/s41564-017-0084-4 doi: 10.1038/s41564-017-0084-4

Ackermann, M., Sikora-Wohlfeld, W., & Beyer, A. (2013). Impact of Natural Genetic Variation on Gene Expression Dynamics. *PLoS Genetics*, *9*(6). doi: 10.1371/journal.pgen.1003514

Adachi, M., Kanno, T., Matsubara, T., Nishijima, T., Itakura, S., & Yamaguchi, M. (1999). Promotion of cyst formation in the toxic dinoflagellate Alexandrium (Dinophyceae) by natural bacterial assemblages from Hiroshima Bay, Japan. *Marine Ecology Progress Series*, *191*, 175–185. Retrieved from http://www.int-res.com/abstracts/meps/v191/p175-185/ doi: 10.3354/meps191175

Adams, M., Raadik, T. A., Burridge, C. P., & Georges, A. (2014). Global biodiversity assessment and hyper-cryptic species complexes: more than one species of rlephant in the room? *Systematic Biology*, *63*(4), 518–533. doi: 10.1093/sysbio/syu017

Agabian, N. (1990). Trans splicing of nuclear pre-mRNAs. *Cell*, *61*(7), 1157–1160. doi: 10.1016/0092-8674(90)90674-4

Aguillon, S. M., Fitzpatrick, J. W., Bowman, R., Schoech, S. J., Clark, A. G., Coop, G., & Chen, N. (2017). Deconstructing isolation-by-distance: The genomic consequences of limited dispersal. *PLoS Genetics*, *13*(8), 1–27. doi: 10.1371/journal.pgen.1006911

Akman, M., Carlson, J. E., Holsinger, K. E., & Latimer, A. M. (2016). Transcriptome sequencing reveals population differentiation in gene expression linked to functional traits and environmental gradients in the south african shrub protea repens. *New Phytologist*, *210*(1), 295–309.

Alexander, H., Jenkins, B. D., Rynearson, T. A., & Dyhrman, S. T. (2015). Metatranscriptome analyses indicate resource partitioning between diatoms in the field. *Proceedings of the National Academy of Sciences*, *112*(17), E2182–E2190. Retrieved from http://www.pnas.org/lookup/doi/10.1073/pnas.1421993112 doi: 10.1073/pnas.1421993112

Alexander, H., Rouco, M., Haley, S. T., Wilson, S. T., Karl, D. M., & Dyhrman, S. T. (2015). Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. *Proceedings of the National Academy of Sciences*, *112*(44), E5972–E5979. Retrieved from http://www.pnas.org/lookup/doi/10.1073/pnas.1518165112 doi: 10.1073/pnas.1518165112

Aminot, A., & Kerouel, R. (2007). *Dosage automatique des nutriments dans les eaux marines : methodes en flux continu*. . Institut francais de recherche pour l'exploitation de la mer.

Anderson, D. M. (1989). Toxic Algal Blooms and Red tides: a Global Perspective. *Biology, environmental science and toxicology*(April), 11–16. doi: 10.1029/95rg00440

Anderson, D. M. (1998). *Physiology and bloom dynamics of toxic Alexandrium species, with emphasis on life cycle transitions.*

Anderson, D. M., Alpermann, T. J., Cembella, A. D., Collos, Y., Masseret, E., & Montresor, M. (2012). The globally distributed genus Alexandrium: Multifaceted roles in marine ecosystems and impacts on human health. *Harmful Algae*, *14*, 10–35. Retrieved from http://dx.doi.org/10.1016/j.hal.2011.10.012 doi: 10.1016/j.hal.2011.10.012

Anderson, D. M., Burkholder, J. M., Cochlan, W. P., Glibert, P. M., Gobler, C. J., Heil, C. A., ... Vargo, G. A. (2008, dec). Harmful algal blooms and eutrophication: Examining linkages from selected coastal regions of the United States. *Harmful Algae*, *8*(1), 39–53. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1568988308000978 doi: 10.1016/j.hal.2008.08.017

Andreasson, A., Kiss, N. B., Juhlin, C. C., & Höög, A. (2013). Long-Term Storage of Endocrine Tissues at -80°C Does Not Adversely Affect RNA Quality or Overall Histomorphology. *Biopreservation and Biobanking*, *11*(6), 366–370. Retrieved from http://online.liebertpub.com/doi/abs/10.1089/bio.2013.0038 doi: 10.1089/bio.2013.0038

Anglès, S., Garcés, E., Hattenrath-Lehmann, T. K., & Gobler, C. J. (2012). In situ life-cycle stages of Alexandrium fundyense during bloom development in Northport Harbor (New York, USA). *Harmful Algae*, *16*, 20–26. doi: 10.1016/j.hal.2011.12.008

Aubret, F., & Shine, R. (2009). Genetic Assimilation and the Postcolonization Erosion of Phenotypic Plasticity in Island Tiger Snakes. *Current Biology*, *19*(22), 1932–1936. Retrieved from http://dx.doi.org/10.1016/j.cub.2009.09.061 doi: 10.1016/j.cub.2009.09.061

Auer, H., Mobley, J., Ayers, L., Bowen, J., Chuaqui, R., Johnson, L., ... Ramirez, N. (2014). The effects of frozen tissue storage conditions on the integrity of RNA and protein. *Biotechnic and Histochemistry*, *89*(7), 518–528. doi: 10.3109/10520295.2014.904927

Auffret, G. A. (1981). *Dynamique sédimentaire de la marge continentale celtique* (PhD dissertation). UNIVERSITE .DE· BORDEAUX 1.

Augusto, L., Dupouey, J. L., Picard, J. F., & Ranger, J. (2001). Potential contribution of the seed bank in coniferous plantations to the restoration of native deciduous forest vegetation. *Acta Oecologica*, *22*(2), 87–98. doi: 10.1016/S1146-609X(01)01104-3

Aylward, F. O., Eppley, J. M., Smith, J. M., Chavez, F. P., Scholin, C. A., & DeLong, E. F. (2015). Microbial community transcriptional networks are conserved in three domains at ocean basin scales [Journal Article]. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(17), 5443-5448. doi: 10.1073/pnas.1502883112

Bachvaroff, T. R., & Place, A. R. (2008). From stop to start: Tandem gene arrangement, copy number and Trans-splicing sites in the dinoflagellate Amphidinium carterae. *PLoS ONE*, *3*(8). doi: 10.1371/journal.pone.0002929

Bakker, F. T., Olsen, J. L., Stam, W. T., & van den Hoek, C. (1992). Nuclear ribosomal dna internal transcribed spacer regions (its1 and its2) define discrete biogeographic groups in cladophora albida (chlorophyta) 1. *Journal of Phycology*, *28*(6), 839–845.

Bargiela, R., Herbst, F. A., Martínez-Martínez, M., Seifert, J., Rojo, D., Cappello, S.,

… Golyshin, P. N. (2015). Metaproteomics and metabolomics analyses of chronically petroleum-polluted sites reveal the importance of general anaerobic processes uncoupled with degradation. *Proteomics*, *15*(20), 3508–3520. doi: 10.1002/pmic.201400614

Baum, A., García-Sastre, A., Baughman, B. M., Jake Slavish, P., Dubois, R. M., Boyd, V. a., … Banerjee, a. K. (1975). Novel initiation of RNA synthesis in vitro by vesicular stomatitis virus. *Journal of virology*, *255*(2), 37–40. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/165900{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/165428{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=113057{&}tool=pmcentrez{&}rendertype=abstract{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4277 doi: 10.1261/rna.2340405

Bechler, K. (1997). Influence of capping and polyadenylation on mRNA expression and on antisense RNA mediated inhibition of gene expression. *Biochem Biophys Res Commun*, *241*(1), 193–199. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{&}db=PubMed{&}dopt=Citation{&}list{_}uids=9405256 doi: S0006-291X(97)97789-5[pii]\r10.1006/bbrc.1997.7789

Behnke, A., Engel, M., Christen, R., Nebel, M., Klein, R. R., & Stoeck, T. (2011). Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environmental Microbiology*, *13*(2), 340–349. doi: 10.1111/j.1462-2920.2010.02332.x

Belin, C. (1993). Distribution of Dinophysis spp. and Alexandrium minutum along French coasts since 1984 and their DSP and PSP toxicity levels. In T. J. Smayda & Y. Shimizu (Eds.), *Toxic phytoplankton blooms in the sea* (pp. 469 – 474). Amsterdam, The Netherlands: Elsevier.

Berg, C., Dupont, C. L., Asplund-Samuelsson, J., Celepli, N. A., Eiler, A., Allen, A. E., … Ininbergs, K. (2018). Dissection of Microbial Community Functions during a Cyanobacterial Bloom in the Baltic Sea via Metatranscriptomics. *Frontiers in Marine Science*, *5*(February), 1–12. Retrieved from http://journal.frontiersin.org/article/10.3389/fmars.2018.00055/full doi: 10.3389/fmars.2018.00055

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., … Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology and Evolution*, *22*(3), 148–155. doi: 10.1016/j.tree.2006.11.004

Blanquart, F., Valero, M., Alves-De-Souza, C., Dia, A., Lepelletier, F., Bigeard, E., … Guillou, L. (2016). Evidence for parasite-mediated selection during short-lasting toxic algal blooms. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1841). doi: 10.1098/rspb.2016.1870

Bobbie, R. J., & White, D. C. (1980). Characterization of Benthic Microbial Community Structure by High-Resolution Gas Chromatography of Fatty Acid Methyl Esters. *Applied and Environmental Microbiology*, *39*(6), 1212–1222.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Bolnick, D. I., & Fitzpatrick, B. M. (2007). Sympatric Speciation: Models and Empirical Evidence. *Annual Review of Ecology, Evolution, and Systematics*, *38*(1), 459–487. Retrieved from http://www.annualreviews.org/doi/10.1146/annurev.ecolsys.38.091206.095804 doi: 10.1146/annurev.ecolsys.38.091206.095804

Bonen, L. (1993). Trans-splicing of pre-mRNA in plants, animals, and protists. *Federation of American Societies for Experimental Biology*, *7*(1), 40–46.

Bravo, I., Vila, M., Masó, M., Figueroa, R. I., & Ramilo, I. (2008). Alexandrium catenella and Alexandrium minutum blooms in the Mediterranean Sea: Toward the identification of ecological niches. *Harmful Algae*, *7*(4), 515–522. doi: 10.1016/j.hal.2007.11.005

Brosnahan, M. L., Velo-Suárez, L., Ralston, D. K., Fox, S. E., Sehein, T. R., Shalapyonok, A., ... Anderson, D. M. (2015). Rapid growth and concerted sexual transitions by a bloom of the harmful dinoflagellate Alexandrium fundyense (Dinophyceae). *Limnology and Oceanography*, *60*(6), 2059–2078. doi: 10.1002/lno.10155

Brown, T. A. (2002). Molecular Phylogenetics. Retrieved from `https://www.ncbi.nlm.nih.gov/books/NBK21122/`

Busby, M. A., Gray, J. M., Costa, A. M., Stewart, C., Stromberg, M. P., Barnett, D., ... Marth, G. T. (2011). Expression divergence measured by transcriptome sequencing of four yeast species. *BMC genomics*, *12*(1), 635.

Calbet, A., Vaqué, D., Felipe, J., Vila, M., Sala, M. M., Alcaraz, M., & Estrada, M. (2003). Relative grazing impact of microzooplankton and mesozooplankton on a bloom of the toxic dinoflagellate Alexandrium minutum. *Marine Ecology Progress Series*, *259*(August), 303–309. doi: 10.3354/meps259303

Caron, D. A., Alexander, H., Allen, A. E., Archibald, J. M., Armbrust, E. V., Bachy, C., ... Worden, A. Z. (2017). Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews Microbiology*, *15*(1), 6–20. Retrieved from `http://dx.doi.org/10.1038/nrmicro.2016.160` doi: 10.1038/nrmicro.2016.160

Casabianca, S., Penna, A., Pecchioli, E., Jordi, A., Basterretxea, G., & Vernesi, C. (2012). Population genetic structure and connectivity of the harmful dinoflagellate Alexandrium minutum in the Mediterranean Sea. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1726), 129–138. Retrieved from `http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2011.0708` doi: 10.1098/rspb.2011.0708

Cembella, A. D. (2003). Chemical ecology of eukaryotic microalgae in marine ecosystems. *Phycologia*, *42*(4), 420–447. Retrieved from `http://www.phycologia.org/doi/abs/10.2216/i0031-8884-42-4-420.1` doi: 10.2216/i0031-8884-42-4-420.1

Chambouvet, A., Morin, P., Marie, D., & Guillou, L. (2008). Control of toxic marine dinoflagellate blooms by serial parasitic killers. *Science*, *322*(5905), 1254–1257. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/19023082{%}0Ahttp://www.sciencemag.org/cgi/doi/10.1126/science.1164387` doi: 10.1126/science.1164387

Chan, Y. H., & Wong, J. T. Y. (2007). Concentration-dependent organization of DNA by the dinoflagellate histone-like protein HCc3. *Nucleic Acids Research*, *35*(8), 2573–2583. doi: 10.1093/nar/gkm165

Chang, F., Anderson, D. M., Kulis, D. M., & Till, D. G. (1997, mar). Toxin production of Alexandrium minutum (Dinophyceae) from the Bay of Plenty, New Zealand. *Toxicon*, *35*(3), 393–409. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0041010196001687?via{%}3Dihub` doi: 10.1016/S0041-0101(96)00168-7

Chang, F. H., Garthwaite, I., Anderson, D. M., Towers, N., Stewart, R., & MacKenzie, L. (1999, dec). Immunofluorescent detection of a PSPproducing dinoflagellate, <i>Alexandrium minutum</i> , from Bay of Plenty, New Zealand. *New Zealand Journal of Marine and Freshwater Research*, *33*(4), 533–543. Retrieved from `http://www.tandfonline.com/doi/abs/10.1080/`

00288330.1999.9516898 doi: 10.1080/00288330.1999.9516898

Chapelle, A., Le Bec, C., Le Gac, M., Labry, C., Amzil, Z., Guillou, L., ... Malestroit, P. (2014). Étude sur la prolifération de la micro algue Alexandrium minutum en rade de Brest. , 61.

Chapelle, A., Le Gac, M., Labry, C., Siano, R., Quere, J., Caradec, F., ... Gouriou, J. (2015a). The Bay of Brest (France), a new risky site for toxic Alexandrium minutum blooms and PSP shellfish contamination. *Harmful Algae News*, *51*(51), 4–5.

Chapelle, A., Le Gac, M., Labry, C., Siano, R., Quere, J., Caradec, F., ... Gouriou, J. (2015b). The Bay of Brest (France), a new risky site for toxic Alexandrium minutum blooms and PSP shellfish contamination. *Harmful Algae News*, *51*(51), 4–5.

Chauvaud, L., Jean, F., Ragueneau, O., & Thouzeau, G. (2000). Long-term variation of the Bay of Brest ecosystem: Benthic-pelagic coupling revisited. *Marine Ecology Progress Series*, *200*, 35–48. doi: 10.3354/meps200035

Chen, S., Yang, P., Jiang, F., Wei, Y., Ma, Z., & Kang, L. (2010). De Novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. *PLoS ONE*, *5*(12). doi: 10.1371/journal.pone.0015633

Chevaldonné, P., Jollivet, D., Vangriesheim, A., & Desbruye, D. (1997). Hydrothermal-vent alvinellid polychaete dispersal in the eastern Pacific. *Limnology and Oceanography*, *42*(1), 67–80.

Choi, C. J., Brosnahan, M. L., Sehein, T. R., Anderson, D. M., & Erdner, D. L. (2017). Insights into the loss factors of phytoplankton blooms: The role of cell mortality in the decline of two inshore alexandrium blooms. *Limnology and Oceanography*, *62*(4), 1742–1753.

Chomérat, N., Sellos, D. Y., Zentz, F., & Nézan, E. (2010). Morphology and molecular phylogeny of Prorocentrum consutum SP. Nov. (dinophyceae), a new benthic dinoflagellate from South Brittany (northwestern France). *Journal of Phycology*, *46*(1), 183–194. doi: 10.1111/j.1529-8817.2009.00774.x

Chust, G., Villarino, E., Chenuil, A., Irigoien, X., Bizsel, N., Bode, A., ... Borja, A. (2016). Dispersal similarly shapes both population genetics and community patterns in the marine realm. *Scientific Reports*, *6*(February). Retrieved from http://dx.doi.org/10.1038/srep28730 doi: 10.1038/srep28730

Cohen, N. R., Ellis, K. A., Lampe, R. H., McNair, H., Twining, B. S., Maldonado, M. T., ... Marchetti, A. (2017). Diatom Transcriptional and Physiological Responses to Changes in Iron Bioavailability across Ocean Provinces. *Frontiers in Marine Science*, *4*(November). Retrieved from http://journal.frontiersin.org/article/10.3389/fmars.2017.00360/full doi: 10.3389/fmars.2017.00360

Colborn, J., Crabtree, R. E., Shaklee, J. B., Pfeiler, E., & Bowen, B. W. (2007). the Evolutionary Enigma of Bonefishes (Albula Spp.): Cryptic Species and Ancient Separations in a Globally Distributed Shorefish. *Evolution*, *55*(4), 807–820. Retrieved from http://doi.wiley.com/10.1111/j.0014-3820.2001.tb00816.x doi: 10.1111/j.0014-3820.2001.tb00816.x

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 1–19. doi: 10.1186/s13059-016-0881-8

Connell, J. H., & Slatyer, R. O. (1977). Mechanisms of succession in natural communities and their role in community stability and organization. *The American Naturalist*, *111*(982), 1119–1144.

Coolen, S., Proietti, S., Hickman, R., Davila Olivas, N. H., Huang, P. P., Van Verk, M. C., ... Van Wees, S. C. (2016). Transcriptome dynamics of Arabidopsis during sequential biotic and abiotic stresses. *The Plant journal : for cell and molecular biology*, *86*(3), 249–267. doi: 10.1111/tpj.13167

Cosgrove, S., Rathaille, A. N., & Raine, R. (2014). The influence of bloom intensity on the encystment rate and persistence of alexandrium minutum in cork harbor, ireland. *Harmful algae*, *31*, 114–124.

Cowen, K. R., & Sponaugle, S. (2009, 01). Larval dispersal and marine population connectivity. *Annual review of marine science*, *1*, 443-66.

Cowen, R. K., Gawarkiewicz, G., Pineda, J., Thorrold, S. R., & Werner, F. E. (2007). Population Connectivity in Marine Systems: An Overview. *Oceanography*, *20*(3), 14–21. doi: 10.5670/oceanog.2009.01

Coyne, J. A., & Orr, H. A. (2004). Speciation. *Sinauer.Com*. Retrieved from http://www.sinauer.com/media/wysiwyg/tocs/Speciation.pdf

Crovadore, J., Soljan, V., Calmin, G., Chablais, R., Cochard, B., & Lefort, F. (2017). Metatranscriptomic and metagenomic description of the bacterial nitrogen metabolism in waste water wet oxidation effluents. *Heliyon*, *3*(10), e00427. Retrieved from http://dx.doi.org/10.1016/j.heliyon.2017.e00427 doi: 10.1016/j.heliyon.2017.e00427

Cruz, D., Suárez, J. P., Kottke, I., & Piepenbring, M. (2014). Cryptic species revealed by molecular phylogenetic analysis of sequences obtained from basidiomata of <i>Tulasnella</i>. *Mycologia*, *106*(4), 708–722. Retrieved from https://www.tandfonline.com/doi/full/10.3852/12-386 doi: 10.3852/12-386

Cullen, J. J., & MacIntyre, J. (1998). Behavior, physiology and the niche of depth-regulating phytoplankton. *Nato Asi Series G Ecological Sciences*, *41*(1978), 559–580. Retrieved from http://cmore.ancl.hawaii.edu/agouron/2007/documents/Cullen-MacIntyre-NATO98.pdf

Cushman, S. A., Landguth, E. L., & Flather, C. H. (2013). Evaluating population connectivity for species of conservation concern in the American Great Plains. *Biodiversity and Conservation*, *22*(11), 2583–2605. doi: 10.1007/s10531-013-0541-1

Cusick, K. D., & Sayler, G. S. (2013). An overview on the marine neurotoxin, saxitoxin: Genetics, moleculartargets, methods of detection and ecological functions. *Marine Drugs*, *11*(4), 991–1018. doi: 10.3390/md11040991

Darwin, C. (1859). On the origin of species by means of natural selection. *Murray, London*.

De Clerck, O., Guiry, M. D., Leliaert, F., Samyn, Y., & Verbruggen, H. (2013). Algal Taxonomy: A Road to Nowhere? *Journal of Phycology*, *49*(2), 215–225. doi: 10.1111/jpy.12020

De Meester, N., Derycke, S., Rigaux, A., & Moens, T. (2015). Temperature and salinity induce differential responses in life histories of cryptic nematode species. *Journal of Experimental Marine Biology and Ecology*, *472*, 54–62. Retrieved from http://dx.doi.org/10.1016/j.jembe.2015.07.002 doi: 10.1016/j.jembe.2015.07.002

Delong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N.-u., ... Karl, D. M. (2006). Community Genomics among microbial assemblages in the Ocean' s Interior. *Science*, *311*(January), 496–503. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16439655 doi: 10.1126/science.1120250

Deng, J., Auchtung, J. M., Konstantinidis, K. T., Caro-Quintero, A., Brettar, I., Höfle, M., & Tiedje, J. M. (2018). Divergence in gene regulation contributes to sympatric speciation of Shewanella baltica strains. *Applied and Environmental Microbiology*, *84*(4). doi: 10.1128/AEM.02015-17 de Queiroz, K. (2005). Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences*, *102*(Supplement 1), 6600–6607. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.0502030102 doi: 10.1073/pnas.0502030102

Destombe, C., Valero, M., & Guillemin, M. L. (2010). Delineation of two sibling red algal species, gracilaria gracilis and gracilaria dura (gracilariales, rhodophyta), using multiple dna markers: Resurrection of the species g. dura previously described in the northern atlantic 200 years ago. *Journal of Phycology*, *46*(4), 720–727. doi: 10.1111/j.1529-8817.2010.00846.x

Dia, A., Guillou, L., Mauger, S., Bigeard, E., Marie, D., Valero, M., & Destombe, C. (2014). Spatiotemporal changes in the genetic diversity of harmful algal blooms caused by the toxic dinoflagellate Alexandrium minutum. *Molecular Ecology*, *23*(3), 549–560. doi: 10.1111/mec.12617

Dilmaghani, A., Gladieux, P., Gout, L., Giraud, T., Brunner, P. C., Stachowiak, A., ... Rouxel, T. (2012). Migration patterns and changes in population biology associated with the worldwide spread of the oilseed rape pathogen Leptosphaeria maculans. *Molecular Ecology*, *21*(10), 2519–2533. doi: 10.1111/j.1365-294X.2012.05535.x

Dobzhansky, T. (1937). Genetics and the orign of species. (Xi).

Dodge, J. D. (1965). Chromosome structure in the dinoflagellates and the problem of mesokaryotic cells. *Excerpta Medica, International Congress Series*, *97*, 339 – 345.

Dray, S., & Dufour, A.-B. (2007, sep). The <b>ade4</b> Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, *22*(4), 1–20. Retrieved from http://www.jstatsoft.org/v22/i04/ doi: 10.18637/jss.v022.i04

Duale, N., Lipkin, W. I., Briese, T., Aarem, J., Rønningen, K. S., Aas, K. K., ... Brunborg, G. (2014). Long-term storage of blood RNA collected in RNA stabilizing Tempus tubes in a large biobank - Evaluation of RNA quality and stability. *BMC Research Notes*, *7*(1), 1–10. doi: 10.1186/1756-0500-7-633

Ekelund, N. G. (1991, jul). The Effects of UV-B Radiation on Dinoflagellates. *Journal of Plant Physiology*, *138*(3), 274–278. Retrieved from https://www.sciencedirect.com/science/article/pii/S0176161711802877 doi: 10.1016/S0176-1617(11)80287-7

Elbrachter, M. (1998). Exotic flagellates of coastal North Sea waters. *Helgol. Wiss. Meeresunters.*, *52*(3-4), 235–242. doi: 10.1007/BF02908899

Erdner, D. L., & Anderson, D. M. (2006, apr). Global transcriptional profiling of the toxic dinoflagellate Alexandrium fundyense using Massively Parallel Signature Sequencing. *BMC Genomics*, *7*(1), 88. Retrieved from http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-7-88 doi: 10.1186/1471-2164-7-88

Erdner, D. L., Richlen, M., McCauley, L. A., & Anderson, D. M. (2011). Diversity and dynamics of a widespread bloom of the toxic dinoflagellate Alexandrium fundyense. *PLoS ONE*, *6*(7), 1–8. doi: 10.1371/journal.pone.0022965

Fabre, A. L., Colotte, M., Luis, A., Tuffet, S., & Bonnet, J. (2014). An efficient method for long-term room temperature storage of RNA. *European Journal of Human Genetics*, *22*(3), 379–385. doi: 10.1038/ejhg.2013.145

Falciatore, A., D'Alcalà, M. R., Croot, P., & Bowler, C. (2000, jun). Perception of environmental signals by a marine diatom. *Science (New York, N.Y.)*, *288*(5475), 2363–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10875921 doi: 10.1126/SCIENCE.288.5475.2363

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The Microbial Engines That

Drive Earth 's Biogeochemical Cycles. *Science*, *320*(5879), 1034–1039. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18497287 doi: 10.1126/science.1153213

Fay, J. C., & Wittkopp, P. J. (2008). Evaluating the role of natural selection in the evolution of gene regulation. *Heredity*, *100*(2), 191–199. doi: 10.1038/sj.hdy.6801000

Federle, T. W., Dobbins, D. C., Thornton-Manning, J. R., & Jones, D. D. (1986). Microbial biomass, activity, and community structure in subsurface soils. *Groundwater*, *24*(3), 365–374.

Figueroa, R. I., Dapena, C., Bravo, I., & Cuadrado, A. (2015). The hidden sexuality of Alexandrium minutum: An example of overlooked sex in dinoflagellates. *PLoS ONE*, *10*(11), 1–21. doi: 10.1371/journal.pone.0142667

Figueroa, R. I., Garcés, E., & Bravo, I. (2007). Comparative study of the life cycles of Alexandrium tamutum and Alexandrium minutum (Gonyaulacales, Dinophyceae) in culture. *Journal of Phycology*, *43*(5), 1039–1053. doi: 10.1111/j.1529-8817.2007.00393.x

Figueroa, R. I., Garcés, E., Massana, R., & Camp, J. (2008, oct). Description, Host-specificity, and Strain Selectivity of the Dinoflagellate Parasite Parvilucifera sinerae sp. nov. (Perkinsozoa). *Protist*, *159*(4), 563–578. Retrieved from https://www.sciencedirect.com/science/article/pii/S1434461008000400?via{%}3Dihub doi: 10.1016/J.PROTIS.2008.05.003

Figueroa, R. I., Vazquez, J. A., Massanet, A., Murado, M. A., & Bravo, I. (2011). Interactive effects of salinity and temperature on planozygote and cyst formation of alexandrium minutum (dinophyceae) in culture 1. *Journal of phycology*, *47*(1), 13–24.

Fišer, C., Robinson, C. T., & Malard, F. (2018). Cryptic species as a window into the paradigm shift of the species concept. *Molecular Ecology*, *27*(3), 613–635. doi: 10.1111/mec.14486

Forlani, M. C., Tonini, J. F., Cruz, C. A., Zaher, H., & de Sá, R. O. (2017). Molecular and morphological data reveal three new cryptic species of <i>Chiasmocleis</i> (Mehely 1904) (Anura, Microhylidae) endemic to the Atlantic Forest, Brazil. *PeerJ*, *5*(Mehely 1904), e3005. Retrieved from https://peerj.com/articles/3005 doi: 10.7717/peerj.3005

Francesconi, M., & Lehner, B. (2014). The effects of genetic variation on gene expression dynamics during development. *Nature*, *505*(7482), 208–211. Retrieved from http://dx.doi.org/10.1038/nature12772 doi: 10.1038/nature12772

Franco, J. M., Fernández, P., & Reguera, B. (1994, jun). Toxin profiles of natural populations and cultures ofAlexandrium minutum Halim from Galician (Spain) coastal waters. *Journal of Applied Phycology*, *6*(3), 275–279. Retrieved from http://link.springer.com/10.1007/BF02181938 doi: 10.1007/BF02181938

Frangópulos, M., Guisande, C., DeBlas, E., & Maneiro, I. (2004). Toxin production and competitive abilities under phosphorus limitation of Alexandrium species. *Harmful Algae*, *3*(2), 131–139. doi: 10.1016/S1568-9883(03)00061-1

Fraser, H. B. (2013). Gene expression drives local adaptation in humans Gene expression drives local adaptation in humans. , 1089–1096. doi: 10.1101/gr.152710.112

Fraser, H. B., Moses, A. M., & Schadt, E. E. (2010). Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proceedings of the National Academy of Sciences*, *107*(7), 2977–2982. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.0912245107 doi: 10.1073/pnas.0912245107

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W.,

& DeLong, E. F. (2008). Microbial community gene expression in ocean surface waters [Journal Article]. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(10), 3805-3810. Retrieved from `<GotoISI>://WOS:000253930600028` doi: 10.1073/pnas.0708897105

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. doi: 10.1093/bioinformatics/bts565

Gabaldón, C., Carmona, M. J., Montero-Pau, J., & Serra, M. (2015). Long-term competitive dynamics of two cryptic rotifer species: Diapause and fluctuating conditions. *PLoS ONE*, *10*(4), 1–13. doi: 10.1371/journal.pone.0124406

Garcés, E., Bravo, I., Vila, M., Figueroa, R. I., Masó, M., & Sampedro, N. (2004). Relationship between vegetative cells and cyst production during Alexandrium minutum bloom in Arenys de Mar harbour (NW Mediterranean). *Journal of Plankton Research*, *26*(6), 637–645. doi: 10.1093/plankt/fbh065

Geoffroy, A., Mauger, S., De Jode, A., Le Gall, L., & Destombe, C. (2015). Molecular evidence for the coexistence of two sibling species in Pylaiella littoralis (Ectocarpales, Phaeophyceae) along the Brittany coast. *Journal of Phycology*, *51*(3), 480–489. doi: 10.1111/jpy.12291

Ghalambor, C. K., Hoke, K. L., Ruell, E. W., Fischer, E. K., Reznick, D. N., & Hughes, K. A. (2015). Non-adaptive plasticity potentiates rapid adaptive evolution of gene expression in nature. *Nature*, *525*(7569), 372–375. doi: 10.1038/nature15256

Giacobbe, M., & Maimone, G. (1994). First report of Alexandrium minutum Halim in a Mediterranean lagoon. *Cryptogamie Algol*, *15*, 47–52.

Giacobbe, M. G., Oliva, F. D., & Maimone, G. (1996). Environmental factors and seasonal occurrence of the dinoflagellate Alexandrium minutum, a PSP potential producer, in a Mediterranean lagoon. *Estuarine, Coastal and Shelf Science*, *42*(5), 539–549. doi: 10.1006/ecss.1996.0035

Gilad, Y., Oshlack, A., & Rifkin, S. A. (2006). Natural selection on gene expression. *Trends in Genetics*, *22*(8), 456–461. doi: 10.1016/j.tig.2006.06.002

Gilg, M. R., & Hilbish, T. J. (2003). The Geography of Marine Larval Dispersal : Coupling Genetics with Fine-Scale Physical Oceanography Published by : Ecological Society of America THE GEOGRAPHY OF MARINE LARVAL DISPERSAL : COUPLING GENETICS WITH FINE-SCALE PHYSICAL OCEANOGRAPHY. *Ecology*, *84*(11), 2989–2998.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, *11*(5), 759–769. doi: 10.1111/j.1755-0998.2011.03024.x

Gong, W. D., Paerl, H., & Marchetti, A. (2018). Eukaryotic phytoplankton community spatiotemporal dynamics as identified through gene expression within a eutrophic estuary [Journal Article]. *Environmental Microbiology*, *20*(3), 1095-1111. Retrieved from `<GotoISI>://WOS:000428393900014` doi: 10.1111/1462-2920.14049

Goodenough, U. W., Shames, B., Small, L., Saito, T., Crain, R. C., Sanders, M. A., & Salisbury, J. L. (1993). The role of calcium in the chlamydomonas-reinhardtii mating reaction. *Journal of Cell Biology*, *121*(2), 365–374.

Gorokhova, E. (2005). OCEANOGRAPHY : METHODS Effects of preservation and storage of microcrustaceans in RNA later on RNA and DNA degradation. *LIMNOLOGY and OCEANOGRAPHY: METHODS*, *3*, 143–148. doi: 10.4319/lom.2005.3.143

Gosalbes, M. J., Durbán, A., Pignatelli, M., Abellan, J. J., Jiménez-Hernández, N., Pérez-Cobas, A. E., . . . Moya, A. (2011). Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS ONE*, *6*(3), 1–9. doi: 10.1371/journal.pone.0017447

Gross, J. (1989). Re-occurrence of red tide in Cork Harbor, Ireland. *Red Tide News*, *2*(45).

Grzebyk, D., Béchemin, C., Ward, C. J., Vérité, C., Codd, G. A., & Maestrini, S. Y. (2003). Effects of salinity and two coastal waters on the growth and toxin content of the dinoflagellate Alexandrium minutum. *Journal of Plankton Research*, *25*(10), 1185–1199. doi: 10.1093/plankt/fbg088

Guallar, C., Bacher, C., & Chapelle, A. (2017). Global and local factors driving the phenology of Alexandrium minutum (Halim) blooms and its toxicity. *Harmful Algae*, *67*, 44–60. Retrieved from http://dx.doi.org/10.1016/j.hal.2017.05.005 doi: 10.1016/j.hal.2017.05.005

Guarner, F., & Malagelada, J. R. (2003). Gut flora in health and disease. *Lancet*, *361*(9356), 512–519. doi: 10.1016/S0140-6736(03)12489-0

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., … Christen, R. (2013). The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), 597–604. doi: 10.1093/nar/gks1160

Guiry, M.D and Guiry, G.M. (2014). *Algaebase. world-wide electronic publication, national university of ireland, galway.* (http://www.algaebase.org; searched on 29 June 2018)

Guisande, C., Frangópulos, M., Maneiro, I., Vergara, A. R., & Riveiro, I. (2002). Ecological advantages of toxin production by the dinoflagellate Alexandrium minutum under phosphorus limitation. *Marine Ecology Progress Series*, *225*, 169–176. doi: 10.3354/meps225169

Gygi, S. P., Rochon, Y., Franza, B. R., & Aebersold, R. (1999). Correlation between Protein and mRNA Abundance in Yeast. *Molecular and Cellular Biology*, *19*(3), 1720–1730. Retrieved from http://mcb.asm.org/lookup/doi/10.1128/MCB.19.3.1720 doi: 10.1128/MCB.19.3.1720

Hackett, J. D., Scheetz, T. E., Yoon, H. S., Soares, M. B., Bonaldo, M. F., Casavant, T. L., & Bhattacharya, D. (2005). Insights into a dinoflagellate genome through expressed sequence tag analysis. *BMC Genomics*, *6*, 1–13. doi: 10.1186/1471-2164-6-80

Hagino, K., Okada, H., & Matsuoka, H. (2005). Coccolithophore assemblages and morphotypes of Emiliania huxleyi in the boundary zone between the cold Oyashio and warm Kuroshio currents off the coast of Japan. *Marine Micropaleontology*, *55*(1-2), 19–47. doi: 10.1016/j.marmicro.2005.02.002

Halim, Y. (1960). Alexandrium minutum n. gen. n. sp., dinoagelle provocant des eaux rouges. *Vie et Milieu*, *11*, 102 – 105.

Hallegraeff, G. M., Bolch, C. J., Blackburn, S. I., & Oshima, Y. (1991). Species of the Toxigenic Dinoflagellate Genus Alexandrium in Southeastern Australian Waters. *Botanica Marina*, *34*(6), 575–588. doi: 10.1515/botm.1991.34.6.575

Hamm, J., & Mattaj, I. W. (1990). Monomethylated cap structures facilitate RNA export from the nucleus. *Cell*, *63*(1), 109–118. doi: 10.1016/0092-8674(90)90292-M

Han, Y., Wan, H., Cheng, T., Wang, J., Yang, W., Pan, H., & Zhang, Q. (2017). Comparative RNA-seq analysis of transcriptome dynamics during petal development in Rosa chinensis. *Scientific Reports*, *7*(January), 1–14. Retrieved from http://dx.doi.org/10.1038/srep43382 doi: 10.1038/srep43382

Hansen, G., Daugbjerg, N., & Franco, J. M. (2003). Morphology, toxin composition and LSU rDNA phylogeny of Alexandrium minutum (Dinophyceae) from

Denmark, with some morphological observations on other European strains. *Harmful Algae*, *2*(4), 317–335. doi: 10.1016/S1568-9883(03)00060-X

Harz, H., & Hegemann, P. (1991). *Rhodopsin-regulated calcium currents in Chlamydomonas* (Vol. 351) (No. 6326). Retrieved from http://untitled-6.pdf doi: 10.1038/351489a0

He, T., Lamont, B. B., Krauss, S. L., & Enright, N. J. (2010). Genetic connectivity and inter-population seed dispersal of Banksia hookeriana at the landscape scale. *Annals of Botany*, *106*(3), 457–466. doi: 10.1093/aob/mcq140

Heard, E., & Martienssen, R. A. (2014). Transgenerational epigenetic inheritance: myths and mechanisms. *Cell*, *157*(1), 95–109.

Hedgecock, D., Barber, P. H., & Edmands, S. (2007). Genetic Approaches to Measuring Connectivity. *Oceanography*, *20*(3), 70–79. doi: 10.1073/pnas.0401921101

Heise, T., Schug, M., Storm, D., Ellinger-Ziegelbauer, H., J. Ahr, H., Hellwig, B., ... G. Hengstler, J. (2012). In Vitro - In Vivo Correlation of Gene Expression Alterations Induced by Liver Carcinogens. *Current Medicinal Chemistry*, *19*(11), 1721–1730. Retrieved from http://www.eurekaselect.com/openurl/content.php?genre=article{&}issn=0929-8673{&}volume=19{&}issue=11{&}spage=1721 doi: 10.2174/092986712799945049

Henderiks, J., Winter, A., Elbrächter, M., Feistel, R., Van Der Plas, A., Nausch, G., & Barlow, R. (2012). Environmental controls on Emiliania huxleyi morphotypes in the Benguela coastal upwelling system (SE Atlantic). *Marine Ecology Progress Series*, *448*, 51–66. doi: 10.3354/meps09535

Hendry, A. P., Berg, O. K., & Quinn, T. P. (1999). Condition Dependence and Adaptation-by-Time: Breeding Date, Life History, and Energy Allocation within a Population of Salmon. *Oikos*, *85*(3), 499. Retrieved from http://www.jstor.org/stable/3546699?origin=crossref doi: 10.2307/3546699

Herrmann, M., Ravindran, S. P., Schwenk, K., & Cordellier, M. (2018). Population transcriptomics in Daphnia: The role of thermal selection. *Molecular Ecology*, *27*(2), 387–402. doi: 10.1111/mec.14450

Hewson, I., Eggleston, E. M., Doherty, M., Lee, D. Y., Owens, M., Shapleigh, J. P., ... Crump, B. C. (2014). Metatranscriptomic analyses of plankton communities inhabiting surface and subpycnocline waters of the chesapeake bay during oxic-anoxic-oxic transitions. *Applied and environmental microbiology*, *80*(1), 328–338.

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring genetic differentiation from pool-seq data. *bioRxiv*, 282400.

Holland, E., Braun, F., Nonnengässer, C., Harz, H., & Hegemann, P. (1996, feb). The nature of rhodopsin-triggered photocurrents in Chlamydomonas. I. Kinetics and influence of divalent ions. *Biophysical Journal*, *70*(2), 924–931. Retrieved from http://www.sciencedirect.com/science/article/pii/S0006349596796352?via{%}3Dihub doi: 10.1016/S0006-3495(96)79635-2

Honnay, O., Bossuyt, B., Jacquemyn, H., Shimono, A., & Uchiyama, K. (2008). Can a seed bank maintain the genetic variation in the above ground plant population? *Oikos*, *117*(1), 1–5. doi: 10.1111/j.2007.0030-1299.16188.x

Howe, C. J., Nisbet, R. E. R., & Barbrook, A. C. (2008). The remarkable chloroplast genome of dinoflagellates. *Journal of Experimental Botany*, *59*(5), 1035–1045. doi: 10.1093/jxb/erm292

Hu, S. K., Liu, Z., Alexander, H., Campbell, V., Connell, P. E., Dyhrman, S. T., ... Caron, D. A. (2018). Shifting metabolic priorities among key protistan taxa within and below the euphotic zone. *Environmental Microbiology*, *00*, 1–15. Retrieved from http://doi.wiley.com/10.1111/1462-2920.14259 doi:

10.1111/1462-2920.14259

Huang, J., & van der Ploeg, L. H. (1991). Maturation of polycistronic pre-mRNA in Trypanosoma brucei: analysis of trans splicing and poly(A) addition at nascent RNA transcripts from the hsp70 locus. *Molecular and cellular biology*, *11*(6), 3180–3190. doi: 10.1128/MCB.11.6.3180.Updated

Hutchinson, G. E. (1957). Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*(22), 415–427.

Hutchinson, G. E. (1961). The Paradox of the Plankton. *The American Naturalist*, *106*(948), 254–257.

Hwang, D. F., & Lu, Y. H. (2000, nov). Influence of environmental and nutritional factors on growth, toxicity, and toxin profile of dinoflagellate Alexandrium minutum. *Toxicon*, *38*(11), 1491–1503. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0041010100000805?via{%}3Dihub` doi: 10.1016/S0041-0101(00)00080-5

Jackson, C. J., Norman, J. E., Schnare, M. N., Gray, M. W., Keeling, P. J., & Waller, R. F. (2007). Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. *BMC Biology*, *5*, 1–17. doi: 10.1186/1741-7007-5-41

Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J., & Kimmel, M. (2015). Microarray experiments and factors which affect their reliability. *Biology Direct*, *10*(1), 1–14. Retrieved from `http://dx.doi.org/10.1186/s13062-015-0077-2` doi: 10.1186/s13062-015-0077-2

Jiang, Y., Xiong, X., Danska, J., & Parkinson, J. (2016). Metatranscriptomic analysis of diverse microbial communities reveals core metabolic pathways and microbiomespecific functionality. *Microbiome*, *4*, 1–18. Retrieved from `http://dx.doi.org/10.1186/s40168-015-0146-x` doi: 10.1186/s40168-015-0146-x

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., & D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of Sciences*, *109*(40), 16213–16216. Retrieved from `http://www.pnas.org/cgi/doi/10.1073/pnas.1203849109` doi: 10.1073/pnas.1203849109

Kawecki, T. J., & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters*, *7*(12), 1225–1241. doi: 10.1111/j.1461-0248.2004.00684.x

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., … Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology*, *12*(6). doi: 10.1371/journal.pbio.1001889

Keeling, P. J., & del Campo, J. (2017). Marine Protists Are Not Just Big Bacteria. *Current Biology*, *27*(11), R541–R549. Retrieved from `http://dx.doi.org/10.1016/j.cub.2017.03.075` doi: 10.1016/j.cub.2017.03.075

Kelly, S. A., Panhuis, T. M., & Stoehr, A. M. (2012). Phenotypic plasticity: molecular mechanisms and adaptive significance. *Compr Physiol*, *2*(2), 1417–39.

Klouch, Z. K., Caradec, F., Plus, M., Hernández-Fariñas, T., Pineau-Guillou, L., Chapelle, A., … Siano, R. (2016). Heterogeneous distribution in sediments and dispersal in waters of Alexandrium minutum in a semi-enclosed coastal ecosystem. *Harmful Algae*, *60*, 81–91. Retrieved from `http://dx.doi.org/10.1016/j.hal.2016.11.001` doi: 10.1016/j.hal.2016.11.001

Knijn, H. M., Wrenzycki, C., Hendriksen, P. J. M., Vos, P. L. a. M., Zeinstra, E. C., van der Weijden, G. C., … Dieleman, S. J. (2005). In vitro and in vivo culture effects on mRNA expression of genes involved in metabolism and apoptosis in bovine embryos. *Reproduction, fertility, and development*, *17*, 775–784. doi: 10.1071/RD05038

Kohli, G. S., John, U., Van Dolah, F. M., & Murray, S. A. (2016). Evolutionary distinctiveness of fatty acid and polyketide synthesis in eukaryotes. *ISME Journal*, *10*(8), 1877–1890. Retrieved from `http://dx.doi.org/10.1038/ismej.2015.263` doi: 10.1038/ismej.2015.263

Koumandou, V. L., Nisbet, R. E. R., Barbrook, A. C., & Howe, C. J. (2004). Dinoflagellate chloroplasts - Where have all the genes gone? *Trends in Genetics*, *20*(5), 261–267. doi: 10.1016/j.tig.2004.03.008

Kuylenstierna, M., & Karlson, B. (2000). *Checklist of phytoplank- ton in the Skagerrak-Kattegat.* Retrieved from `http://www.marbot.gu.se/sss/ssshome.html`

Laanaia, N., Vaquer, A., Fiandrino, A., Genovesi, B., Pastoureaud, A., Cecchi, P., & Collos, Y. (2013). Wind and temperature controls on Alexandrium blooms (2000-2007) in Thau lagoon (Western Mediterranean). *Harmful Algae*, *28*, 31–36. doi: 10.1016/j.hal.2013.05.016

Labry, C., Erard-Le Denn, E., Chapelle, A., Fauchot, J., Youenou, A., Crassous, M. P., ... Lorgeoux, B. (2008). Competition for phosphorus between two dinoflagellates: A toxic Alexandrium minutum and a non-toxic Heterocapsa triquetra. *Journal of Experimental Marine Biology and Ecology*, *358*(2), 124–135. doi: 10.1016/j.jembe.2008.01.025

Laird, P. W. (1989). Trans splicing in trypanosomes - archaism or adaptation? *Trends in Genetics*, *5*(C), 204–208. doi: 10.1016/0168-9525(89)90082-6

LaJeunesse, T. C., Lambert, G., Andersen, R. A., Coffroth, M. A., & Galbraith, D. W. (2005). Symbiodinium (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates. *Journal of Phycology*, *41*(4), 880–886. doi: 10.1111/j.0022-3646.2005.04231.x

Lajus, D., Sukhikh, N., & Alekseev, V. (2015). Cryptic or pseudocryptic: Can morphological methods inform copepod taxonomy? An analysis of publications and a case study of the Eurytemora affinis species complex. *Ecology and Evolution*, *5*(12), 2374–2385. doi: 10.1002/ece3.1521

Lam, C. M. C., New, D. C., & Wong, J. T. Y. (2001). cAMP in the cell cycle of the dinoflagellate <i>Crypthecodinium cohnii</i> (Dinophyta). *Journal of Phycology*, *37*, 79–85.

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*. doi: 10.1186/1471-2105-9-559

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, *9*(4), 357.

Lassus, P., & Bardouil, M. (1988). Présence d'un protogonyaulax sp. sur le littoral atlantique français pendant l'hiver 1987. *Cryptogamie Algologie*, *9*(4), 273–278.

Laza-Martinez, A., Orive, E., & Miguel, I. (2011). Morphological and genetic characterization of benthic dinoflagellates of the genera coolia, ostreopsis and prorocentrum from the south-eastern bay of biscay. *European Journal of Phycology*, *46*(1), 45–65. doi: 10.1080/09670262.2010.550387

Le Bescot, N., Mahé, F., Audic, S., Dimier, C., Garet, M. J., Poulain, J., ... Siano, R. (2016). Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. *Environmental Microbiology*, *18*(2), 609–626. doi: 10.1111/1462-2920.13039

Le Calvez, T., Burgaud, G., Mahé, S., Barbier, G., & Vandenkoornhuyse, P. (2009). Fungal diversity in deep-sea hydrothermal ecosystems. *Applied and Environmental Microbiology*, *75*(20), 6415–6421. doi: 10.1128/AEM.00653-09

Le Gac, M., Metegnier, G., Chomérat, N., Malestroit, P., Quéré, J., Bouchez, O., ...

Chapelle, A. (2016). Evolutionary processes and cellular functions underlying divergence in Alexandrium minutum. *Molecular ecology*, *25*(20). doi: 10.1111/mec.13815

Lebret, K., Kritzberg, E. S., Figueroa, R., & Rengefors, K. (2012). Genetic diversity within and genetic differentiation between blooms of a microalgal species. *Environmental Microbiology*, *14*(9), 2395–2404. doi: 10.1111/j.1462-2920.2012.02769.x

Ledoux, M., Bardouil, M., Fremy, J. M., Lassus, P., Murail, I., & Bohec, M. (1993). Use of HPLC for toxin analysis of shellfish contaminated by an Alexandrium minutum strain. In T. Smayda & Y. Shimizu (Eds.), *Toxic phytoplankton blooms in the sea* (Elsevier ed., pp. 413–418). Amsterdam, The Netherlands.

Lee, Y. S. (2006, oct). Factors affecting outbreaks of high-density Cochlodinium polykrikoides red tides in the coastal seawaters around Yeosu and Tongyeong, Korea. *Marine Pollution Bulletin*, *52*(10), 1249–1259. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0025326X06000889` doi: 10.1016/J.MARPOLBUL.2006.02.024

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012, mar). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)*, *28*(6), 882–3. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/22257669http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3307112` doi: 10.1093/bioinformatics/bts034

Legendre, L., & Rassoulzadegan, F. (1995, feb). Plankton and nutrient dynamics in marine waters. *Ophelia*, *41*(1), 153–172. Retrieved from `http://www.tandfonline.com/doi/abs/10.1080/00785236.1995.10422042` doi: 10.1080/00785236.1995.10422042

Leggat, W., Seneca, F., Wasmund, K., Ukani, L., Yellowlees, D., & Ainsworth, T. D. (2011, oct). Differential Responses of the Coral Host and Their Algal Symbiont to Thermal Stress. *PLoS ONE*, *6*(10), e26687. Retrieved from `http://dx.plos.org/10.1371/journal.pone.0026687` doi: 10.1371/journal.pone.0026687

Lehahn, Y., Koren, I., Sharoni, S., d'Ovidio, F., Vardi, A., & Boss, E. (2017). Dispersion/dilution enhances phytoplankton blooms in low-nutrient waters. *Nature Communications*, *8*, 14868.

Lejzerowicz, F., Voltsky, I., & Pawlowski, J. (2013). Identifying active foraminifera in the Sea of Japan using metatranscriptomic approach. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *86-87*, 214–220. Retrieved from `http://dx.doi.org/10.1016/j.dsr2.2012.08.008` doi: 10.1016/j.dsr2.2012.08.008

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. (2014). DNA-based species delimitation in algae. *European Journal of Phycology*, *49*(2), 179–196. Retrieved from `http://dx.doi.org/10.1080/09670262.2014.904524` doi: 10.1080/09670262.2014.904524

Lema, K., Metegnier, G., Quéré, J., Latimier, M., Youenou, A., Lambert, C., . . . Le Gac, M. (2018 *Submitted*). Linking *in vitro* vs *in situ* gene expression variations to genetic background : investigating species divergence under the molecular physiology scope. *FEMS Microbiology Ecology*.

Lett, C., Verley, P., Mullon, C., Parada, C., Brochier, T., Penven, P., & Blanke, B. (2008). A Lagrangian tool for modelling ichthyoplankton dynamics. *Environmental Modelling and Software*, *23*(9), 1210–1214. doi: 10.1016/j.envsoft.2008.02.005

Levin, L. A., Huggett, D., Myers, P., Bridges, T., & Weaver, J. (1993). Rare earth tagging methods for the study of larval dispersal by marine invertebrates. *Limnology and Oceanography*, *38*(2), 346–360. doi: 10.4319/lo.1993.38.2.0346

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009, aug). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. Retrieved from `https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352` doi: 10.1093/bioinformatics/btp352

Li, M. H., & Merilä, J. (2010). Genetic evidence for male-biased dispersal in the Siberian jay (Perisoreus infaustus) based on autosomal and Z-chromosomal markers. *Molecular Ecology*, *19*(23), 5281–5295. doi: 10.1111/j.1365-294X.2010.04870.x

Lidie, K. B., & Van Dolah, F. M. (2007). Spliced leader RNA-mediated trans-splicing in a dinoflagellate, Karenia brevis. *Journal of Eukaryotic Microbiology*, *54*(5), 427–435. doi: 10.1111/j.1550-7408.2007.00282.x

Liebeke, M., Bruford, M. W., Donnelly, R. K., Ebbels, T. M. D., Hao, J., Kille, P., ... Bundy, J. G. (2014). Identifying biochemical phenotypic differences between cryptic species. *Biology Letters*, *10*(9), 20140615–20140615. Retrieved from `http://rsbl.royalsocietypublishing.org/cgi/doi/10.1098/rsbl.2014.0615` doi: 10.1098/rsbl.2014.0615

Lilly, E. L., Halanych, K. M., & Anderson, D. M. (2005). Phylogeny, biogeography, and species boundaries within the Alexandrium minutum group. *Harmful Algae*, *4*(6), 1004–1020. doi: 10.1016/j.hal.2005.02.001

Lilly, E. L., Halanych, K. M., & Anderson, D. M. (2007). Species boundaries and global biogeography of the alexandrium tamarense complex (dinophyceae) 1. *Journal of Phycology*, *43*(6), 1329–1338.

Lim, P. T., Leaw, C. P., Sato, S., van Thuoc, C., Kobiyama, A., & Ogata, T. (2011). Effect of salinity on growth and toxin production of Alexandrium minutum isolated from a shrimp culture pond in northern Vietnam. *Journal of Applied Phycology*, *23*(5), 857–864. doi: 10.1007/s10811-010-9593-8

Lim, P. T., Leaw, C. P., Usup, G., Kobiyama, A., Koike, K., & Ogata, T. (2006). Effects of light and temperature on growth, nitrate uptake, and toxin production of two tropical dinoflagellates: Alexandrium tamiyavanichii and Alexandrium minutum (Dinophyceae). *Journal of Phycology*, *42*(4), 786–799. doi: 10.1111/j.1529-8817.2006.00249.x

Lin, S. (2011). Genomic understanding of dinoflagellates. *Research in Microbiology*, *162*(6), 551–569. doi: 10.1016/j.resmic.2011.04.006

Lin, S., Zhang, H., Hou, Y., Zhuang, Y., & Miranda, L. (2009). High-level diversity of dinoflagellates in the natural environment, revealed by assessment of mitochondrial cox1 and cob genes for dinoflagellate DNA barcoding. *Applied and Environmental Microbiology*, *75*(5), 1279–1290. doi: 10.1128/AEM.01578-08

Liu, M. G., Li, H., Xu, X., Barnstable, C. J., & Zhang, S. S. (2008). Comparison of gene expression during in vivo and in vitro postnatal retina development. *J Ocul Biol Dis Infor*, *1*(2-4), 59–72. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/20072636` doi: 10.1007/s12177-008-9009-z

Liu, Y., Beyer, A., & Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, *165*(3), 535–550. Retrieved from `http://dx.doi.org/10.1016/j.cell.2016.03.014` doi: 10.1016/j.cell.2016.03.014

Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, *113*(21), 5970–5975. Retrieved from `http://www.pnas.org/lookup/doi/10.1073/pnas.1521291113` doi: 10

.1073/pnas.1521291113

Lopez-Maestre, H., Brinza, L., Marchet, C., Kielbassa, J., Bastien, S., Boutigny, M., … Lacroix, V. (2016). SNP calling from RNA-seq data without a reference genome: Identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Research*, *44*(19), 1–13. doi: 10.1093/nar/gkw655

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. doi: 10.1186/s13059-014-0550-8

Lundholm, N., Bates, S. S., Baugh, K. A., Bill, B. D., Connell, L. B., Léger, C., & Trainer, V. L. (2012). Cryptic and pseudo-cryptic diversity in diatoms-with descriptions of pseudo-nitzschia hasleana SP. NOV. AND P. Fryxelliana SP. NOV. *Journal of Phycology*, *48*(2), 436–454. doi: 10.1111/j.1529-8817.2012.01132.x

MacKenzie, L., & Berkett, N. (1997). Cell morphology and PSP-toxin profiles of Alexandrium minutum in the Marlborough Sounds, New Zealand. *New Zealand Journal of Marine and Freshwater Research*, *31*(3), 403–409. doi: 10.1080/00288330.1997.9516773

Maes, G. E., Pujolar, J. M., Hellemans, B., & Volckaert, F. A. (2006). Evidence for isolation by time in the European eel (Anguilla anguilla L.). *Molecular Ecology*, *15*(8), 2095–2107. doi: 10.1111/j.1365-294X.2006.02925.x

Maguer, J.-F., Wafar, M., Madec, C., Morin, P., & Erard-Le Denn, E. (2004). Nitrogen and phosphorus requirements of an Alexandrium minutum bloom in the Penzé estuary, France. *Limnology and Oceanography*, *49*(4), 1108–1114. doi: 10.4319/lo.2004.49.4.1108

Mahé, F., De Vargas, C., Bass, D., Czech, L., Stamatakis, A., Lara, E., … Dunthorn, M. (2017). Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nature Ecology and Evolution*, *1*(4), 1–8. Retrieved from `http://dx.doi.org/10.1038/s41559-017-0091` doi: 10.1038/s41559-017-0091

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, *3*, e1420. Retrieved from `https://peerj.com/articles/1420` doi: 10.7717/peerj.1420

Marchant, A., Mougel, F., Almeida, C., Jacquin-Joly, E., Costa, J., & Harry, M. (2015). De novo transcriptome assembly for a non-model species, the blood-sucking bug Triatoma brasiliensis, a vector of Chagas disease. *Genetica*, *143*(2), 225–239. doi: 10.1007/s10709-014-9790-5

Marchetti, A., Schruth, D. M., Durkin, C. A., Parker, M. S., Kodner, R. B., Berthiaume, C. T., … Armbrust, E. V. (2012). Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability [Journal Article]. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(6), E317-E325. doi: 10.1073/pnas.1118408109

Marie, A. D., Lejeusne, C., Karapatsiou, E., Cuesta, J. A., Drake, P., Macpherson, E., … Rico, C. (2016). Implications for management and conservation of the population genetic structure of the wedge clam Donax trunculus across two biogeographic boundaries. *Scientific Reports*, *6*(May), 1–10. Retrieved from `http://dx.doi.org/10.1038/srep39152` doi: 10.1038/srep39152

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*.

Mayali, X., Franks, P. J., & Azam, F. (2007). Bacterial induction of temporary cyst formation by the dinoflagellate Lingulodinium polyedrum. *Aquatic Microbial Ecology*, *50*(1), 51–62. doi: 10.3354/ame01143

Mayden, R. L. (1997). A hierarchy of species concepts: the denouement in the saga of the species problem. *Species. The units of biodiversity.*, 381–423. doi: http://dx.doi.org/10.1016/S0169-5347(97)85758-8

Mayr, E. (1942). *Systematics and the Origin of Species*. Columbia University Press, New York.

Mayr, E. (1963). *Animal Species and Evolution*. Harvard University Press, Cambridge, MA.

McManus, M. A., & Woodson, C. B. (2012). Plankton distribution and ocean dispersal. *Journal of Experimental Biology*, *215*(6), 1008–1016. Retrieved from http://jeb.biologists.org/cgi/doi/10.1242/jeb.059014 doi: 10.1242/jeb.059014

Medlin, L., & Cembella, A. (2013). *Biodiversity of harmful marine algae*.

Mehta, S., Tsai, P., Lasham, A., Campbell, H., Reddel, R., Braithwaite, A., & Print, C. (2016). A study of TP53 RNA splicing illustrates pitfalls of RNA-seq methodology. *Cancer Research*, *76*(24), 7151–7159. doi: 10.1158/0008-5472.CAN-16-1624

Metegnier, G., Destombe, C., Quéré, J., & Le Gac, M. (2018a *Submitted*). Challenging microorganisms surveys *in situ* : new highlights into the gene expression dynamics of a marine phytoplankton. *Molecular Ecology*.

Metegnier, G., Destombe, C., Quéré, J., & Le Gac, M. (2018b *Submitted*). Inter and intra specific transcriptional and phenotypic responses of *Pseudo-nitzchia* under different nutrient conditions. *Molecular Ecology*.

Meyer, C. P., Geller, J. B., & Paulay, G. (2005). Fine scale ednemism o coral reefs: Archipelagic differentiation in turbid gastropods . *Evolution*, *59*(1), 113–125.

Mims, M. C., Hauser, L., Goldberg, C. S., & Olden, J. D. (2016). Genetic differentiation, isolation-by-distance, and metapopulation dynamics of the Arizona treefrog (Hyla wrightorum) in an isolated portion of its range. *PLoS ONE*, *11*(8), 1–23. doi: 10.1371/journal.pone.0160655

Mitarai, S., Siegel, D. A., Watson, J. R., Dong, C., & McWilliams, J. C. (2009). Quantifying connectivity in the coastal ocean with application to the Southern California Bight. *Journal of Geophysical Research: Oceans*, *114*(10), 1–21. doi: 10.1029/2008JC005166

Monks, S. A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., ... Schadt, E. E. (2004). Genetic inheritance of gene expression in human cell lines. *American journal of human genetics*, *75*(6), 1094–1105. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed{&}id=15514893{&}retmode=ref{&}cmd=prlinks{%}5Cnpapers2://publication/doi/10.1086/426461 doi: 10.1086/426461

Montecinos, A. E., Couceiro, L., Peters, A. F., Desrut, A., Valero, M., & Guillemin, M. L. (2017). Species delimitation and phylogeographic analyses in the Ectocarpus subgroup siliculosi (Ectocarpales, Phaeophyceae). *Journal of Phycology*, *53*(1), 17–31. doi: 10.1111/jpy.12452

Montresor, M., Marino, D., Zingone, A., & Dafnis G. (1990). Three Alexandrium species from coastal Tyrrhenian waters. In E. Graneli, B. Sundstrom, L. Edler, & D. Anderson (Eds.), *Toxic marine phytoplankton* (Elsevier ed., pp. 82–87). New York.

Moran, M. A., Satinsky, B., Gifford, S. M., Luo, H., Rivers, A., Chan, L. K., ... Hopkinson, B. M. (2013). Sizing up metatranscriptomics. *ISME Journal*, *7*(2), 237–243. Retrieved from http://dx.doi.org/10.1038/ismej.2012.94 doi: 10.1038/ismej.2012.94

Morard, R., Garet-Delmas, M. J., Mahé, F., Romac, S., Poulain, J., Kucera, M., & De Vargas, C. (2018). Surface ocean metabarcoding confirms limited diversity in planktonic foraminifera but reveals unknown hyper-abundant lineages. *Scientific Reports*, *8*(1), 1–10. doi: 10.1038/s41598-018-20833-z

Moreno Díaz de la Espina, S., Alverca, E., Cuadrado, A., & Franca, S. (2005). Organization of the genome and gene expression in a nuclear environment lacking histones and nucleosomes: The amazing dinoflagellates. *European Journal of Cell Biology*, *84*(2-3), 137–149. doi: 10.1016/j.ejcb.2005.01.002

Moustafa, A., Evans, A. N., Kulis, D. M., Hackett, J. D., Erdner, D. L., Anderson, D. M., & Bhattacharya, D. (2010). Transcriptome profiling of a toxic dinoflagellate reveals a gene-rich protist and a potential impact on gene expression due to bacterial presence. *PLoS ONE*, *5*(3). doi: 10.1371/journal.pone.0009688

Mukherjee, M., Banik, S. K., Pradhan, S. K., Sharma, A. P., Raghavan, S. V., Manna, R. K., … Mandal, S. (2015). Diversity and distribution of tintinnids in chilika lagoon with description of new records. *Indian Journal of Fisheries*, *62*(1).

Muller, E. E., Glaab, E., May, P., Vlassis, N., & Wilmes, P. (2013). Condensing the omics fog of microbial communities. *Trends in Microbiology*, *21*(7), 325–333. Retrieved from http://dx.doi.org/10.1016/j.tim.2013.04.009 doi: 10.1016/j.tim.2013.04.009

Murphy, W. J., Watkins, K. P., & Agabian, N. (1986). Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: Evidence for Trans splicing. *Cell*, *47*(4), 517–525. doi: 10.1016/0092-8674(86)90616-1

Murray, S. A., Garby, T., Hoppenrath, M., & Neilan, B. A. (2012). Genetic diversity, morphological uniformity and polyketide production in dinoflagellates (amphidinium, dinoflagellata). *PLoS ONE*, *7*(6), 1–14. doi: 10.1371/journal.pone.0038253

Muth, T., Benndorf, D., Reichl, U., Rapp, E., & Martens, L. (2013). Searching for a needle in a stack of needles: Challenges in metaproteomics data analysis. *Molecular BioSystems*, *9*(4), 578–585. doi: 10.1039/c2mb25415h

Mutshinda, C. M., Finkel, Z. V., Widdicombe, C. E., & Irwin, A. J. (2016). Ecological equivalence of species within phytoplankton functional groups [Journal Article]. *Functional Ecology*, *30*(10), 1714-1722. Retrieved from <GotoISI>://WOS:000385511500011 doi: 10.1111/1365-2435.12641

Nagai, S., Lian, C., Yamaguchi, S., Hamaguchi, M., Matsuyama, Y., Itakura, S., … others (2007). Microsatellite markers reveal population genetic structure of the toxic dinoflagellate alexandrium tamarense (dinophyceae) in japanese coastal waters 1. *Journal of Phycology*, *43*(1), 43–54.

Nash, E. A., Barbrook, A. C., Edwards-Stuart, R. K., Bernhardt, K., Howe, C. J., & Nisbet, R. E. R. (2007). Organization of the mitochondrial genome in the dinoflagellate Amphidinium carterae. *Molecular Biology and Evolution*, *24*(7), 1528–1536. doi: 10.1093/molbev/msm074

Nawar, T., & Sibaoka, T. (1987). Local Ion Currents Controlling the Localized Cytoplasmic Movement Associated with Feeding Initiation of Noctiluca. *Protoplasma*, *137*, 125–133. Retrieved from https://link.springer.com/content/pdf/10.1007/BF01281147.pdf

Nehring, S. (1998). Non-indigenous phytoplankton species in the North Sea: supposed region of origin and possible transport vector. *Archive of Fishery and Marine Research*, *46*(3), 181–194. Retrieved from http://www.stefannehring.de/downloads/076a{_}Nehring-1998{_}AFMR-46{_}non-indi-plankton.pdf

Ní Rathaille, A., & Raine, R. (2011). Seasonality in the excystment of Alexandrium minutum and Alexandrium tamarense in Irish coastal waters. *Harmful Algae*, *10*(6), 629–635. Retrieved from http://dx.doi.org/10.1016/j.hal.2011.04

.015 doi: 10.1016/j.hal.2011.04.015

Nickrent, D. L., & Sargent, M. L. (1991). An overview of the secondary structure of the V4 region of eukaryotic small-subunit ribosomal RNA. *Nucleic Acids Research*, *19*(2), 227–235. doi: 10.1093/nar/19.2.227

Nourmohammad, A., Rambeau, J., Held, T., Kovacova, V., Berg, J., & Lässig, M. (2017). Adaptive Evolution of Gene Expression in Drosophila. *Cell Reports*, *20*(6), 1385–1395. doi: 10.1016/j.celrep.2017.07.033

Oami, K., Naitoh, Y., & Sibaoka, T. (1995). Modification of voltage-sensitive inactivation of Na+ current by external Ca2+ in the marine dinoflagellate Noctiluca miliaris. *Journal of Comparative Physiology A*, *176*(5), 635–640. doi: 10.1007/BF00192492

Okamoto, O. K., & Hastings, J. W. (2003). Novel dinoflagellate clock-related genes identified through microarray analysis. *Journal of Phycology*, *39*(3), 519–526. doi: 10.1046/j.1529-8817.2003.02170.x

Oliver, P. M., Adams, M., Lee, M. S., Hutchinson, M. N., & Doughty, P. (2009). Cryptic diversity in vertebrates: Molecular data double estimates of species diversity in a radiation of Australian lizards (Diplodactylus, Gekkota). *Proceedings of the Royal Society B: Biological Sciences*, *276*(1664), 2001–2007. doi: 10.1098/rspb.2008.1881

Olivieri, E. H. R., de Andrade Franco, L., Pereira, R. G., Carvalho Mota, L. D., Campos, A. H. J. F. M., & Carraro, D. M. (2014). Biobanking Practice: RNA Storage at Low Concentration Affects Integrity. *Biopreservation and Biobanking*, *12*(1), 46–52. Retrieved from http://online.liebertpub.com/doi/abs/10.1089/bio.2013.0056 doi: 10.1089/bio.2013.0056

Oshlack, A., & Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, *4*, 1–10. doi: 10.1186/1745-6150-4-14

Ottesen, E. A., Marin, R., Preston, C. M., Young, C. R., Ryan, J. P., Scholin, C. A., & DeLong, E. F. (2011). Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton [Journal Article]. *Isme Journal*, *5*(12), 1881-1895. doi: 10.1038/ismej.2011.70

Ottesen, E. A., Young, C. R., Eppley, J. M., Ryan, J. P., Chavez, F. P., Scholin, C. A., & DeLong, E. F. (2013). Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proceedings of the National Academy of Sciences*, *110*(6), E488–E497. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.1222099110 doi: 10.1073/pnas.1222099110

Ottesen, E. A., Young, C. R., Gifford, S. M., Eppley, J. M., Marin, R., Schuster, S. C., . . . DeLong, E. F. (2014). Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science*, *345*(6193), 207–212. doi: 10.1126/science.1252476

Pai, A. A., Pritchard, J. K., & Gilad, Y. (2015). The Genetic and Mechanistic Basis for Variation in Gene Regulation. *PLoS Genetics*, *11*(1). doi: 10.1371/journal.pgen.1004857

Parchman, T. L., Geist, K. S., Grahnen, J. A., Benkman, C. W., & Buerkle, C. A. (2010). Transcriptome sequencing in an ecologically important tree species: Assembly, annotation, and marker discovery. *BMC Genomics*, *11*(1). doi: 10.1186/1471-2164-11-180

Parkinson, J. E., Baumgarten, S., Michell, C. T., Baums, I. B., LaJeunesse, T. C., & Voolstra, C. R. (2016). Gene Expression Variation Resolves Species and Individual Strains among Coral-Associated Dinoflagellates within the Genus Symbiodinium. *Genome biology and evolution*, *8*(3), 665–680. doi: 10.1093/gbe/evw019

Payo, D. A., Leliaert, F., Verbruggen, H., D'hondt, S., Calumpong, H. P., & De Clerck, O. (2012). Extensive cryptic species diversity and fine-scale endemism in the marine red alga <i>Portieria</i> in the Philippines. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1753). doi: 10.1098/rspb.2012.2660

Peck, L. S. (2011). Organisms and responses to environmental change. *Marine Genomics*, *4*(4), 237–243. Retrieved from `http://dx.doi.org/10.1016/j.margen.2011.07.001` doi: 10.1016/j.margen.2011.07.001

Percopo, I., Ruggiero, M. V., Balzano, S., Gourvil, P., Lundholm, N., Siano, R., ... Sarno, D. (2016). Pseudo-nitzschia arctica sp. nov., a new cold-water cryptic Pseudo-nitzschia species within the P. pseudodelicatissima complex. *Journal of Phycology*, *52*(2), 184–199. doi: 10.1111/jpy.12395

Pernice, M. C., Giner, C. R., Logares, R., Perera-Bel, J., Acinas, S. G., Duarte, C. M., ... Massana, R. (2016). Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. *ISME Journal*, *10*(4), 945–958. doi: 10.1038/ismej.2015.170

Persich, G. R., Kulis, D. M., Lilly, E. L., Anderson, D. M., & Garcia, V. M. (2006). Probable origin and toxin profile of Alexandrium tamarense (Lebour) Balech from southern Brazil. *Harmful Algae*, *5*(1), 36–44. doi: 10.1016/j.hal.2005.04.002

Persson, A., Godhe, A., & Karlson, B. (2000). Dinoflagellate Cysts in Recent Sediments from the West Coast of Sweden. *Offprint Botanica Marina*, *43*, 69–79. Retrieved from `https://www.degruyter.com/downloadpdf/j/botm.2000.43.issue-1/bot.2000.006/bot.2000.006.pdf`

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., ... others (2018). Genome evolution across 1,011 saccharomyces cerevisiae isolates. *Nature*, *556*(7701), 339.

Pigliucci, M., & Murren, C. J. (2003). Perspective: Genetic Assimilation and a Possible Evolutionary Paradox: Can Macroevolution Sometimes Be So Fast As To Pass Us By? *Evolution*, *57*(7), 1455. Retrieved from `http://www.bioone.org/perlserv/?request=get-abstract{&}doi=10.1554{%}2F02-381` doi: 10.1554/02-381

Pinceel, J., Jordaens, K., van Houtte, N., de Winter, A., & Backeljau, T. (2004). Molecular and morphological data reveal cryptic taxonomic diversity in the terrestrial slug complex. *Biological Journal of the Linnean Society*(April), 23–38.

Pineda, J., Hare, J., & Sponaugle, S. (2007). Larval Transport and Dispersal in the Coastal Ocean and Consequences for Population Connectivity. *Oceanography*, *20*(3), 22–39. Retrieved from `https://tos.org/oceanography/article/larval-transport-and-dispersal-in-the-coastal-ocean-and-consequences-for-po` doi: 10.5670/oceanog.2007.27

Poretsky, R. S., Bano, N., Buchan, A., LeCleir, G., Kleikemper, J., Pickering, M., ... Hollibaugh, J. T. (2005). Analysis of microbial gene transcripts in environmental samples. *Applied and Environmental Microbiology*, *71*(7), 4121–4126. Retrieved from `isi:000230445700092{%}5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC1168992/pdf/1747-04.pdf` doi: 10.1128/AEM.71.7.4121

Poretsky, R. S., Hewson, I., Sun, S., Allen, A. E., Zehr, J. P., & Moran, M. A. (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environmental Microbiology*, *11*(6), 1358–1375. doi: 10.1111/j.1462-2920.2008.01863.x

Preußer, C., Jaé, N., & Bindereif, A. (2012). MRNA splicing in trypanosomes. *International Journal of Medical Microbiology*, *302*(4-5), 221–224. Retrieved from `http://dx.doi.org/10.1016/j.ijmm.2012.07.004` doi: 10.1016/j.ijmm.2012.07.004

Probert, I. P. (1999). *Sexual reproduction and ecophysiology of the marine dinoflagellate Alexandrium minutum Halim* (Unpublished doctoral dissertation).

Prosser, J. I. (2012). Ecosystem processes and interactions in a morass of diversity. *FEMS Microbiology Ecology*, *81*(3), 507–519. doi: 10.1111/j.1574-6941.2012.01435.x

Pusadee, T., Jamjod, S., Chiang, Y.-C., Rerkasem, B., & Schaal, B. A. (2009). Genetic structure and isolation by distance in a landrace of thai rice. *Proceedings of the National Academy of Sciences*, *106*(33), 13880–13885.

Quarmby, L. M. (1994). Signal transduction in the sexual life of Chlamydomonas. *Plant Molecular Biology*, *26*(5), 1271–1287. doi: 10.1007/BF00016474

Radax, R., Rattei, T., Lanzen, A., Bayer, C., Rapp, H. T., Urich, T., & Schleper, C. (2012). Metatranscriptomics of the marine sponge Geodia barretti: Tackling phylogeny and function of its microbial community. *Environmental Microbiology*, *14*(5), 1308–1324. doi: 10.1111/j.1462-2920.2012.02714.x

Raine, R. (2014). A review of the biophysical interactions relevant to the promotion of HABs in stratified systems: The case study of Ireland. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *101*, 21–31. Retrieved from http://dx.doi.org/10.1016/j.dsr2.2013.06.021 doi: 10.1016/j.dsr2.2013.06.021

Ram, R. J., VerBerkmoes, N. C., Thelen, M. P., Tyson, G. W., Baker, B. J., Blake, R. C., ... Banfield, J. F. (2005). Community proteomics of a natural microbial biofilm (Supporting Online Material). *Science (New York, NY)*, *308*(5730), 1915–1920. Retrieved from (null) doi: 10.1126/science.

Ramond, P., Sourisseau, M., Simon, N., Romac, S., Schmitt, S., Rigaut-Jalabert, F., ... Siano, R. (2018 *Submitted*). Coupling between taxonomic and functional diversity in protistan coastal communities. *Environmental Microbiology*.

Read, B. A., Kegel, J., Klute, M. J., Kuo, A., Lefebvre, S. C., Maumus, F., ... Wurch, L. L. (2013). Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature*, *499*(7457), 209–213. doi: 10.1038/nature12221

Ribolli, J., Hoeinghaus, D. J., Johnson, J. A., Zaniboni-Filho, E., de Freitas, P. D., & Galetti, P. M. (2017). Isolation-by-time population structure in potamodromous Dourado Salminus brasiliensis in southern Brazil. *Conservation Genetics*, *18*(1), 67–76. doi: 10.1007/s10592-016-0882-x

Richlen, M. L., Erdner, D. L., McCauley, L. A. R., Liberal, K., & Anderson, D. M. (2012). Extensive genetic diversity and rapid population differentiation during blooms of alexandrium fundyense (dinophyceae) in an isolated salt pond on cape cod, MA, USA. *Ecology and Evolution*, *2*(10), 2588–2599. doi: 10.1002/ece3.373

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4402510/pdf/gkv007.pdf doi: 10.1093/nar/gkv007

RIZZO, P. J. (1991). The Enigma of the Dinoflagellate Chromosome. *The Journal of Protozoology*, *38*(3), 246–252. doi: 10.1111/j.1550-7408.1991.tb04437.x

RIZZO, P. J. (2003). Those amazing dinoflagellate chromosomes. *Cell Research*, *13*(4), 215–217. Retrieved from http://www.nature.com/doifinder/10.1038/sj.cr.7290166 doi: 10.1038/sj.cr.7290166

Robinson, M., Mccarthy, D., Chen, Y., & Smyth, G. K. (2011). edgeR : differential expression analysis of digital gene expression data. *October*, *23*(October), 2887–2887.

Rocha-Olivares, A., Fleeger, J. W., & Foltz, D. W. (2004). Differential tolerance among cryptic species: A potential cause of pollutant-related reductions in genetic diversity. *Environmental Toxicology and Chemistry*, *23*(9), 2132–2137. doi: 10.1897/03-512

Rolshausen, G., Hobson, K. A., & Schaefer, H. M. (2010). Spring arrival along a migratory divide of sympatric blackcaps (Sylvia atricapilla). *Oecologia*, *162*(1), 175–183. doi: 10.1007/s00442-009-1445-3

Romero, E., Garnier, J., Lassaletta, L., Billen, G., Le Gendre, R., Riou, P., & Cugier, P. (2013). Large-scale patterns of river inputs in southwestern Europe: Seasonal and interannual variations and potential eutrophication effects at the coastal zone. *Biogeochemistry*, *113*(1-3), 481–505. doi: 10.1007/s10533-012-9778-0

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., ... Galtier, N. (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, *515*(7526), 261–263. doi: 10.1038/nature13685

Rose, N. H., Bay, R. A., Morikawa, M. K., & Palumbi, S. R. (2018). Polygenic evolution drives species divergence and climate adaptation in corals. *Evolution*, *72*(1), 82–94. doi: 10.1111/evo.13385

Roy, J., Gray, M., Stoinski, T., Robbins, M. M., Vigilant, L., Clutton-Brock, T., ... Hundertmark, K. (2014). Fine-scale genetic structure analyses suggest further male than female dispersal in mountain gorillas. *BMC Ecology 2014 14:1*, *14*(1), 70–72. doi: 10.1186/1472-6785-14-21

Rydgren, K., Hestmark, G., & Okland, R. H. (1998). Revegetation following experimental disturbance in a boreal old-growth Picea abies forest. *Journal of Vegetation Science*, *9*(3), 763–776. doi: 10.2307/3237042

Rynearson, T., Newton, J., & Armbrust, E. (2006). Spring bloom development, genetic variation, and population succession in the planktonic diatom ditylum brightwellii. *Limnology and Oceanography*, *51*(3), 1249–1261.

Sala-Rovira, M., Geraud, M. L., Caput, D., Jacques, F., Soyer-Gobillard, M. O., Vernet, G., & Herzog, M. (1991). Molecular cloning and immunolocalization of two variants of the major basic nuclear protein (HCc) from the histone-less eukaryote Crypthecodinium cohnii (Pyrrhophyta). *Chromosoma*, *100*(8), 510–518. doi: 10.1007/BF00352201

Salazar-Jaramillo, L., Jalvingh, K. M., de Haan, A., Kraaijeveld, K., Buermans, H., & Wertheim, B. (2017). Inter- and intra-species variation in genome-wide gene expression of Drosophila in response to parasitoid wasp attack. *BMC Genomics*, *18*(1), 1–14. doi: 10.1186/s12864-017-3697-3

San Diego-McGlone, M. L., Azanza, R. V., Villanoy, C. L., & Jacinto, G. S. (2008, jan). Eutrophic waters, algal bloom and fish kill in fish farming areas in Bolinao, Pangasinan, Philippines. *Marine Pollution Bulletin*, *57*(6-12), 295–301. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0025326X08001811?via{%}3Dihub` doi: 10.1016/J.MARPOLBUL.2008.03.028

Sarjeant, W. A. S. (1974). *Fossil and living dinoflagellates*. Academic Press.

Schlichting, C. D., Pigliucci, M., et al. (1998). *Phenotypic evolution: a reaction norm perspective.* Sinauer Associates Incorporated.

Schonrogge, K., Barr, B., Wardlaw, J., Napper, E., Gardner, M., Breen, J., ... Thomas, J. A. (2002). When rare species become endangered: Cryptic speciation in myrmecolphilous hoverflies. *Journal of the Linnean Society*, *75*, 291–300. Retrieved from `papers2://publication/uuid/10777671-65D9-4D06-933A-9581101B3A2F` doi: 10.1046/j.1095-8312.2002.00019.x

Sedillot, C. (1972). Comptes rendus hebdomadaires des séances de l ′ Académie des sciences . Série B. *Comptes rendus hebdomadaires des séances de l ' Académie des sciences, Série B*, *275*, 319–321.

Sefbom, J., Sassenhagen, I., Rengefors, K., & Godhe, A. (2015). Priority effects in a planktonic bloom-forming marine diatom. *Biology letters*, *11*(5), 20150184. doi: 10.1098/rsbl.2015.0184

Sendler, E., Johnson, G. D., & Krawetz, S. A. (2011). Local and global factors affecting RNA sequencing analysis. *Analytical Biochemistry*, *419*(2), 317–322. Retrieved from http://dx.doi.org/10.1016/j.ab.2011.08.013 doi: 10.1016/j.ab.2011.08.013

Seyhan, A. A., & Burke, J. M. (2000). Mg2+-independent hairpin ribozyme catalysis in hydrated RNA films. *Rna*, *6*(2), 189–198. doi: 10.1017/S1355838200991441

Shapiro, A. M. (1976). Seasonal Polyphenism. In *Evolutionary biology* (pp. 259–333). Boston, MA: Springer US. Retrieved from http://link.springer.com/10.1007/978-1-4615-6950-3{_}6 doi: 10.1007/978-1-4615-6950-3_6

Sharma, A. K., Becker, J. W., Ottesen, E. A., Bryant, J. A., Duhamel, S., Karl, D. M., . . . Delong, E. F. (2014). Distinct dissolved organic matter sources induce rapid transcriptional responses in coexisting populations of Prochlorococcus, Pelagibacter and the OM60 clade. *Environmental Microbiology*, *16*(9), 2815–2830. doi: 10.1111/1462-2920.12254

Sharma, R., & Sharma, P. K. (2018). Metatranscriptome sequencing and analysis of agriculture soil provided significant insights about the microbial community structure and function. *Ecological Genetics and Genomics*, *6*(September 2017), 9–15. Retrieved from http://dx.doi.org/10.1016/j.egg.2017.10.001 doi: 10.1016/j.egg.2017.10.001

Shatkin, A. J. (1976). Capping of eucaryotic mRNAs. *Cell*, *9*(4 PART 2), 645–653. doi: 10.1016/0092-8674(76)90128-8

Shi, Y. M., Tyson, G. W., Eppley, J. M., & DeLong, E. F. (2011). Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean [Journal Article]. *Isme Journal*, *5*(6), 999-1013. Retrieved from <GotoISI>://WOS:000295688200006 doi: 10.1038/ismej.2010.189

Sigee, D. C. (1986). The Dinoflagellate Chromosome. *Advances in Botanical Research*, *12*(C), 205–264. doi: 10.1016/S0065-2296(08)60195-0

Sjöqvist, C. O., & Kremp, A. (2016). Genetic diversity affects ecological performance and stress response of marine diatom populations. *The ISME Journal*, *10*(11), 2755–2766. Retrieved from http://www.nature.com/doifinder/10.1038/ismej.2016.44 doi: 10.1038/ismej.2016.44

Smith, B., & Oliver, J. D. (2006). In situ gene expression by Vibrio vulnificus. *Applied and Environmental Microbiology*, *72*(3), 2244–2246. doi: 10.1128/AEM.72.3.2244-2246.2006

Sorokin, Y. I., Sorokin, P. Y., & Ravagnan, G. (1996). On an extremely dense bloom of the dinoflagellate Alexandrium tamarense in lagoons of the po river delta: Impact on the environment. *Journal of Sea Research*, *35*(4), 251–255. doi: 10.1016/S1385-1101(96)90752-2

Sourisseau, M., Le Guennec, V., Le Gland, G., Plus, M., & Chapelle, A. (2017). Resource Competition Affects Plankton Community Structure; Evidence from Trait-Based Modeling. *Frontiers in Marine Science*, *4*(April), 1–14. Retrieved from http://journal.frontiersin.org/article/10.3389/fmars.2017.00052/full doi: 10.3389/fmars.2017.00052

Spector, D. L. (1984). *Dinoflagellate Nuclei* (Second Edition ed.). ACADEMIC PRESS, INC. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/B9780126565201500080 doi: 10.1016/B978-0-12-656520-1.50008-0

Stern, R. F., Horak, A., Andrew, R. L., Coffroth, M. A., Andersen, R. A., Küpper, F. C., . . . Keeling, P. J. (2010). Environmental barcoding reveals massive dinoflagellate diversity in marine environments. *PLoS ONE*, *5*(11). doi: 10.1371/journal.pone.0013991

Stewart, F. J., Sharma, A. K., Bryant, J. A., Eppley, J. M., & DeLong, E. F. (2011). Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities [Journal Article]. *Genome Biology*, *12*(3). Retrieved from `<GotoISI>://WOS:000291309200006` doi: 10.1186/gb-2011-12-3-r26

Stock, C., Ludwig, F. T., Hanley, P. J., & Schwab, A. (2013). Roles of ion transport in control of cell motility. *Comprehensive Physiology*, *3*(1), 59–119. doi: 10.1002/cphy.c110056

Stocklin, J., & Fischer, M. (1999). Plants with longer-lived seeds have lower local extinction rates in grassland remnants 1950-1985. *Oecologia*, *120*(4), 539–543. doi: 10.1007/s004420050888

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D., Breiner, H. W., & Richards, T. A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, *19*(SUPPL. 1), 21–31. doi: 10.1111/j.1365-294X.2009.04480.x

Su, H., & Chiang, Y. (1991). Dinoflagellates collected from aquaculture ponds in southern Taiwan. *Jpn. J. Phycol*, *39*, 227–238.

Sutton, R. E., & Boothroyd, J. C. (1986). Evidence for Trans splicing in trypanosomes. *Cell*, *47*(4), 527–535. doi: 10.1016/0092-8674(86)90617-3

Suzuki, Y., & Nijhout, H. F. (2006). Evolution of a Polyphenism by Genetic Accommodation. *Science*, *311*(5761), 650–652. Retrieved from `http://www.sciencemag.org/cgi/doi/10.1126/science.1118888` doi: 10.1126/science.1118888

Sword, G. A. (2002). A role for phenotypic plasticity in the evolution of aposematism. *Proceedings of the Royal Society B: Biological Sciences*, *269*(1501), 1639–1644. doi: 10.1098/rspb.2002.2060

Takahagi, K., Uehara-Yamaguchi, Y., Yoshida, T., Sakurai, T., Shinozaki, K., Mochida, K., & Saisho, D. (2016). Analysis of single nucleotide polymorphisms based on RNA sequencing data of diverse bio-geographical accessions in barley. *Scientific Reports*, *6*(October 2015), 1–7. Retrieved from `http://dx.doi.org/10.1038/srep33199` doi: 10.1038/srep33199

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in rna-seq: a matter of depth. *Genome research*, gr–124321.

Taylor, D. J., & Bruenn, J. (2009). The evolution of novel fungal genes from nonretroviral RNA viruses. *BMC Biology*, *7*(1), 1–11. doi: 10.1186/1741-7007-7-88

Temple, H. J., Hoffman, J. I., & Amos, W. (2006). Dispersal, philopatry and intergroup relatedness: Fine-scale genetic structure in the white-breasted thrasher, Ramphocinclus brachyurus. *Molecular Ecology*, *15*(11), 3449–3458. doi: 10.1111/j.1365-294X.2006.03006.x

Tirlapur, U., Scheuerlein, R., & Hader, D. P. (1993). Motility and orientation of a dinoflagellate, Gymnodium, impaired by solar and ultraviolet radiation . *FEMS Microbiol Ecol*, *102*(3/4), 167–174.

Toulza, E., Shin, M. S., Blanc, G., Audic, S., Laabir, M., Collos, Y., . . . Grzebyk, D. (2010). Gene expression in proliferating cells of the dinoflagellate alexandrium catenella (Dinophyceae). *Applied and Environmental Microbiology*, *76*(13), 4521–4529. doi: 10.1128/AEM.02345-09

Trainer, V. L., Bates, S. S., Lundholm, N., Thessen, A. E., Cochlan, W. P., Adams,

N. G., & Trick, C. G. (2012). Pseudo-nitzschia physiological ecology, phylogeny, toxicity, monitoring and impacts on ecosystem health. *Harmful Algae*, *14*, 271–300. Retrieved from `http://dx.doi.org/10.1016/j.hal.2011.10.025` doi: 10.1016/j.hal.2011.10.025

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, *31*(1), 46–53. doi: 10.1038/nbt.2450

Treml, E. A., & Halpin, P. N. (2012). Marine population connectivity identifies ecological neighbors for conservation planning in the Coral Triangle. *Conservation Letters*, *5*(6), 441–449. doi: 10.1111/j.1755-263X.2012.00260.x

Trerotola, M., Relli, V., Simeone, P., & Alberti, S. (2015). Epigenetic inheritance and the missing heritability. *Human genomics*, *9*, 17. Retrieved from `http://dx.doi.org/10.1186/s40246-015-0041-3` doi: 10.1186/s40246-015-0041-3

Turk, R., 't Hoen, P. A., Sterrenburg, E., de Menezes, R. X., de Meijer, E. J., Boer, J. M., ... den Dunnen, J. T. (2004). Gene expression variation between mouse inbred strains. *BMC Genomics*, *5*, 1–8. doi: 10.1186/1471-2164-5-57

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., ... Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43. doi: 10.1038/nature02340

Ullu, E., Matthews, K. R., & Tschudi, C. (1993). Temporal order of RNA-processing reactions in trypanosomes: rapid trans splicing precedes polyadenylation of newly synthesized tubulin transcripts. *Molecular and Cellular Biology*, *13*(1), 720–725. Retrieved from `http://mcb.asm.org/lookup/doi/10.1128/MCB.13.1.720` doi: 10.1128/MCB.13.1.720

Upton, A., Nedwell, D., & Wynn-Williams, D. (1990). The selection of microbial communities by constant or fluctuating temperatures. *FEMS Microbiology Letters*, *74*(4), 243–252.

Urich, T., Lanzén, A., Qi, J., Huson, D. H., Schleper, C., & Schuster, S. C. (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE*, *3*(6). doi: 10.1371/journal.pone.0002527

Van Tienderen, P. H. (1990). Morphological variation in Plantago lanceolata : limits of plasticity. *Evol. Trends Plants*, *4*, 35 – 43.

van Belleghem, S. M., Roelofs, D., van Houdt, J., & Hendrickx, F. (2012). De novo transcriptome assembly and SNP discovery in the wing polymorphic Salt Marsh beetle pogonus chalceus (Coleoptera, Carabidae). *PLoS ONE*, *7*(8). doi: 10.1371/journal.pone.0042605

Van Dolah, F. M. (2000). Diversity of marine and freshwater algal toxins. In *Seafood and freshwater toxins* (pp. 43–71). CRC Press.

Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I., & Marden, J. H. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, *17*(7), 1636–1647. doi: 10.1111/j.1365-294X.2008.03666.x

Verbruggen, H., Vlaeminck, C., Sauvage, T., Sherwood, A. R., Leliaert, F., & De Clerck, O. (2009). Phylogenetic analysis of pseudochlorodesmis strains reveals cryptic diversity above the family level in the siphonous green algae (bryopsidales, chlorophyta). *Journal of Phycology*, *45*(3), 726–731. doi: 10.1111/j.1529-8817.2009.00690.x

Vila, M., Giacobbe, M. G., Masó, M., Gangemi, E., Penna, A., Sampedro, N., ... Galluzzi, L. (2005). A comparative study on recurrent blooms of Alexandrium minutum in two Mediterranean coastal areas. *Harmful Algae*, *4*(4), 673–695. doi: 10.1016/j.hal.2004.07.006

Waddington, C. (1953). Genetic assimilation of an acquired character. *Evolution*, *7*(2), 118–126. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed{&}cmd=Retrieve{&}dopt=AbstractPlus{&}list{_}uids=2405747{%}5Cnpapers2://publication/uuid/CA15A0FC-6827-4C61-BEB3-2DBB6F1FF4F7 doi: 10.2307/2405747

WADDINGTON, C. H. (1961). Genetic assimilation. *Advances in genetics*, *10*, 257–93. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14004267

Waller, R. F., & Jackson, C. J. (2009). Dinoflagellate mitochondrial genomes: Stretching the rules of molecular biology. *BioEssays*, *31*(2), 237–245. doi: 10.1002/bies.200800164

Wang, D. Z. (2008). Neurotoxins from marine dinoflagellates: A brief review. *Marine Drugs*, *6*(2), 349–371. doi: 10.3390/md6020349

Wang, Z., Gerstein, M., & Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57.

Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., … Leadbetter, J. R. (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, *450*(7169), 560–565. doi: 10.1038/nature06269

Wheeler, Q., & Meier, R. (2000). *Species concepts and phylogenetic theory: a debate*. Columbia University Press.

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, *95*(12), 6578–6583. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.95.12.6578 doi: 10.1073/pnas.95.12.6578

Whittaker, R. J., Araújo, M. B., Jepson, P., Ladle, R. J., Watson, J. E., & Willis, K. J. (2005). Conservation biogeography: Assessment and prospect. *Diversity and Distributions*, *11*(1), 3–23. doi: 10.1111/j.1366-9516.2005.00143.x

Williams, G. (1966). *Adaptation and Natural Selection* (PrincetonU ed.). Princeton.

Wirth, T., & Bernatchez, L. (2001). Genetic evidence against panmixia in the European eel. *Nature*, *409*(6823), 1037–1040. doi: 10.1038/35059079

Wisecaver, J. H., & Hackett, J. D. (2011). Dinoflagellate Genome Evolution. *Annual Review of Microbiology*, *65*(1), 369–387. Retrieved from http://www.annualreviews.org/doi/10.1146/annurev-micro-090110-102841 doi: 10.1146/annurev-micro-090110-102841

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, *74*(11), 5088–5090. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.74.11.5088 doi: 10.1073/pnas.74.11.5088

Wohlrab, S., Tillmann, U., Cembella, A., & John, U. (2016). Trait changes induced by species interactions in two phenotypically distinct strains of a marine dinoflagellate. *ISME Journal*, *10*(11), 2658–2668. Retrieved from http://dx.doi.org/10.1038/ismej.2016.57 doi: 10.1038/ismej.2016.57

Wollebæk, J., Heggenes, J., & Røed, K. H. (2011). Population connectivity: dam migration mitigations and contemporary site fidelity in arctic char. *BMC Evolutionary Biology*, *11*(1), 207. Retrieved from http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-11-207 doi: 10.1186/1471-2148-11-207

Wong, J. T. Y., New, D. C., & Hung, V. K. L. (2003). Histone-Like Proteins of the Dino agellate. *Society*, *2*(3), 646–650. doi: 10.1128/EC.2.3.646

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V., & Romano, L. A. (2003). The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, *20*(9), 1377–1419. doi: 10.1093/molbev/msg140

Wright, S. (1943). Isolation by Distance. *Genetics*, *28*(2), 114–138. doi: Article

Wyatt, T., & Jenkinson, I. R. (1997). Notes on Alexandrium population dynamics. *Journal of Plankton Research*, *19*(5), 551–575. Retrieved from http://plankt.oxfordjournals.org/cgi/doi/10.1093/plankt/19.5.551 doi: 10.1093/plankt/19.5.551

Xie, W., Wang, F., Guo, L., Chen, Z., Sievert, S. M., Meng, J., … Xu, A. (2011). Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. *ISME Journal*, *5*(3), 414–426. Retrieved from http://dx.doi.org/10.1038/ismej.2010.144 doi: 10.1038/ismej.2010.144

Xu, H. X., Hong, Y., Zhang, M. Z., Wang, Y. L., Liu, S. S., & Wang, X. W. (2015). Transcriptional responses of invasive and indigenous whiteflies to different host plants reveal their disparate capacity of adaptation. *Scientific Reports*, *5*(September 2014), 1–14. Retrieved from http://dx.doi.org/10.1038/srep10774 doi: 10.1038/srep10774

Xu, Y., Cahill, B., Wilkin, J., & Schofield, O. (2012). Role of wind in regulating phytoplankton blooms on the Mid-Atlantic Bight. (Mld). doi: 10.1016/j.csr.2012.09.011

Yang, I., John, U., Beszteri, S., Glöckner, G., Krock, B., Goesmann, A., & Cembella, A. D. (2010). Comparative gene expression in toxic versus non-toxic strains of the marine dinoflagellate Alexandrium minutum. *BMC Genomics*, *11*(1). doi: 10.1186/1471-2164-11-248

Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics*, *13*(5), 303–314. Retrieved from http://dx.doi.org/10.1038/nrg3186 doi: 10.1038/nrg3186

Yeung, P. K. K., Lam, C. M. C., Ma, Z. Y., Wong, Y. H., & Wong, J. T. (2006). Involvement of calcium mobilization from caffeine-sensitive stores in mechanically induced cell cycle arrest in the dinoflagellate Crypthecodinium cohnii. *Cell Calcium*, *39*(3), 259–274. doi: 10.1016/j.ceca.2005.11.001

Yoshida, M., Ogata, T., Van Thuoc, C., Matsuoka, K., Fukuyo, Y., Hoi, N., & Kodama, M. (2000). The first finding of toxic dinoflagellate Alexandrium minutum in Vietnam. *Fisheries Science*, *66*(1), 177–179.

Zaitseva, M., Vollenhoven, B. J., & Rogers, P. A. (2006). In vitro culture significantly alters gene expression profiles and reduces differences between myometrial and fibroid smooth muscle cells. *Molecular Human Reproduction*, *12*(3), 187–207. doi: 10.1093/molehr/gal018

Zampieri, E., Chiapello, M., Daghino, S., Bonfante, P., & Mello, A. (2016). Soil metaproteomics reveals an inter-kingdom stress response to the presence of black truffles. *Scientific Reports*, *6*(May), 1–11. Retrieved from http://dx.doi.org/10.1038/srep25773 doi: 10.1038/srep25773

Zhang, H., Hou, Y., Miranda, L., Campbell, D. A., Sturm, N. R., Gaasterland, T., & Lin, S. (2007). Spliced leader RNA trans-splicing in dinoflagellates. *Proceedings of the National Academy of Sciences*, *104*(11), 4618–4623. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.0700258104 doi: 10.1073/pnas.0700258104

Zhang, Z., Green, B. R., & Cavalier-Smith, T. (1999). Single gene circles in dinoflagellate chloroplast genomes. *Nature*, *400*(6740), 155–159. doi: 10.1038/22099

Zhou, M.-j., Shen, Z.-l., & Yu, R.-c. (2008, jul). Responses of a coastal phytoplankton community to increased nutrient input from the Changjiang (Yangtze) River. *Continental Shelf Research*, *28*(12), 1483–1489. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0278434308000794` doi: 10.1016/J.CSR.2007.02.009

Zonneveld, K. A., Marret, F., Versteegh, G. J., Bogus, K., Bonnet, S., Bouimetarhan, I., ... Young, M. (2013). Atlas of modern dinoflagellate cyst distribution based on 2405 data points. *Review of Palaeobotany and Palynology*, *191*, 1–197. doi: 10.1016/j.revpalbo.2012.08.003

SORBONNE UNIVERSITÉ

Résumé

From gene expression to genetic adaptation : insights into the spatio-temporal dynamics of *Alexandrium minutum* species complex

by Gabriel METEGNIER

Les populations naturelles sont constamment confrontées à des changements environnementaux (biotiques ou abiotiques). Pour faire face à ces perturbations, différentes réponses ont été sélectionnées au cours de l'évolution. Parmi elles se trouvent la plasticité phénotypique (qui est la capacité pour un même génotype d'exprimer différents phénotypes) et l'adaptation génétique (qui est l'augmentation en fréquence de mutations avantageuses au sein des populations). Toutes deux font parties d'un continuum et leurs apports respectifs à la réponse des populations est difficile à appréhender. Cependant, étudier les liens entre plasticité phénotypique et adaptation génétique est une manière de comprendre les dynamiques des populations et de prévoir leurs réponses à un environnement changeant. Dans la présente étude, je me suis attaché à étudier ces liens à plusieurs échelles (intra- et interspécifique), chez le complexe d'espèces cryptiques de la micro-algue *Alexandrium minutum*, et ce à la fois *in vitro* et *in situ*. En ce qui concerne la plasticité phénotypique, ces deux espèces, quoique génétiquement proche, montrent de profondes différences, soulignant les liens entre divergence génétique et écologique. Au niveau intraspécifique, il apparaît que face à des variations de facteurs abiotiques, les populations ajustent les niveaux d'expression de certains gènes. Notamment, des gènes impliqués dans des fonctions de motilité et d'interactions intercellulaires sont plus exprimés dans des environnements froids à faible salinité. En ce qui concerne la structuration génétique, les populations montrent de la différentiation génétique à la fois à faible échelle spatiale, au cours du temps, mais aussi lorsque la communauté d'espèces environnante change. Pour conclure il existe, aux niveaux intrapopulationnel comme interspécifique, une interaction directe entre divergence génétique et changements d'expression de gènes. En plus de poser de nombreuses questions quant aux capacités de réponse des populations, ces résultats soulignent comment plasticité phénotypique et changements génétique sont liés et interagissent. Ils offrent une perspective nouvelle sur les mécanismes qui sous-tendent les réponses des populations à leur environnement.

MOTS CLEFS : Expression de gènes; différentiation génétique; divergence; plasticité phénotypique; métatranscriptomique; micro-organisme