Phaeodactylum tricornutum genome and epigenome : characterization of natural variants

Achal Rastogi

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

IBENS, Ecole Normale Supérieure (ENS)

Phaeodactylum tricornutum genome and epigenome: characterization of natural variants

Ecole doctorale ED515

Complexité du vivant

Spécialité Bio-informatique

COMPOSITION DU JURY:

Prof. MOCK Thomas
University of East Anglia, Rapporteur

Prof. VYVERMAN Wim
Ghent University, Rapporteur

Dr. IMMACOLATA FERRANTE Maria
SZN, Examinatrice

Dr. COCK Mark
Station Biologique, Examinateur

Dr. TIRICHINE DELACOUR Leila
ENS, Directrice de thèse

Prof. LE CROM Stéphane
UPMC-ENS, Co-directeur de thèse

Soutenue par Achal **RASTOGI**
le 27 Octobre 2016

Dirigée par **Catherine DE VITRY**

Thèse de Doctorat d'Ecole Normale Supérieure

Spécialité

Bio-informatique

École doctorale ED515: Complexité du vivant

Présentée par

Achal RASTOGI

Pour obtenir le grade de

DOCTEUR ès Sciences

Phaeodactylum tricornutum genome and epigenome: characterization of natural variants

Soutenue le **27 Octobre 2016**

Devant le jury composé de

Dr. Catherine de Vitry (IBPC, Paris)	Présidente
Prof. Thomas Mock (University of East Anglia, UK)	Rapporteur
Prof. Wim Vyverman (Ghent University, Belgium)	Rapporteur
Dr. Maria Immacolata Ferrante (SZN, Italy)	Examinatrice
Dr. Mark Cock (Station Biologique, Roscoff)	Examinateur
Dr. Leila Tirichine Delacour (ENS, Paris)	Directrice de thèse
Prof. Stéphane Le Crom (ENS, Paris)	Co-directeur de thèse

I dedicate my thesis to
all the beautiful women of my life.
You all are
the real architects
of my achievements and success.

Acknowledgements

"No one who achieves success does so without acknowledging the help of others. The wise and confident acknowledge this help with gratitude."

-Alfred North Whitehead

It is a heart warming and rewarding experience to pay tribute to the people whose invaluable contribution helped me through out my time as a PhD researcher at ENS, Paris. My sincere and profound gratitude goes to *Dr. Leila Tirichine Delacour* and *Prof. Stephane Le Crom* who gave me a chance to accomplish my PhD under their esteemed supervision. I thank *Prof. Chris Bowler* for giving me an opportunity to join his group and for the trust that he put into me. It was a unique experience to work with them as a team. I thank PSL and Labex-Memolife fellowship, which enabled me to pursue this PhD.

Dr. Leila Tirichine Delacour provided excellent scientific guidance that helped me expand my capabilities to address a biological problem like a *"Biologist"*. I would like to extend my special thanks to you for being so patient and understanding. Your tutelage helped me think and perform beyond my capacity. You gave me an opportunity to work on different projects that helped me develop my scientific aptitude. I can't forget the times when you just went out of the way to help me and found a way to sort out things on your own. My words are not enough to express my immense gratitude and respect towards you. I will always be grateful to you.

My special thanks go to my thesis committee members, *Dr. Lionel Navarro, Prof. Denis*

Thieffry and *Dr. Jean-Pierre Bouly* for assessing my progress and encouraging me to shape up my thesis. I thank my jury members for the kind acceptance to assess and help me to accomplish my thesis. I would like to appreciate and thanks *Prof. Vincent Colot*, *Dr. Lionel Navarro* and *Dr. Fredy Barneche* for discussions and boosting me during the sneak peaks and meet ups in the corridor and section meetings. I would like to express my thankfulness to *Prof. Martine Boccara* for being the youngest soul of the lab. Your presence kept me motivated. I also thank all my collaborators from round the globe for their massive support with the data and valuable discussions.

Oldies are the goldies: This goes for the "not-so-long-gone" mates of the lab - *Shruti, Omer, Heni, Javier, Amos, Alaguraj and Anne-Sophie,* whose presence is missed for sure. I appreciate each and every moment of rise and fall that we spent together during the first two years of my PhD. Not to forget to mention the ones who stood by and made the rest of the time easy going, deserve an imperative appreciation remark - *Anne-Flore, Zhanru, Catherine, Flora, Yann, Richard and Imen*. *Ana*, *Flora, Martin* being part of the same run always encouraged me to touch the finish line. I thank you all for being really supportive and maintaining the friendly environment, keeping the tensions out. I would also like to extend my gratitude to the new members of the Bowler's lab - *Fabio, Federico*, *Clara.*

Science can never be carried out as a one-man army and on this note I would like to thank *François, Daniel, Emmanuelle, Mohamed, Claudia, Barbara, Amira and Delase,* for engaging into long and short-term discussions and always keeping the spirits high. You all have been the best officemates I could ever ask for. I also appreciate the time spent in discussing and captivating advises about the future (not just the projects) with *Leandro, Valentina, Amanda, Angelique, Thierry* and *Odon".*

The Canteen team and my best mates *"Meenu, Anne Flore, Zhanru and Catherine"* always reminded me to take care of my lunch and I can't thank you enough for always looking out for me even if I was not there. Your support, encouragement and thoughtfulness can never be traded for anything. The little parties and get together were really fun and will always be the best memories. I cannot thank enough to *Yann* for making me witness some of my first and the best experiences - *MotoGP and grand cliff view at Etretat*. *Heni,* thank you for introducing me to Belgian fries (De Clercq, les Rois de la Frite) and Le Mayflower - *Dark times were not so dark around a pint and sauce Brazil*. Sweet thanks to *Magali* for sharing the delicious cookies and cupcakes from her kitchen, *Lionel* for the

Belgian beers, *Jerome* to join and finish those beers and last but not the least all the members of the fourth floor and the section for participating in the section meetings and sharing their scientific ideas. I would like to thank everybody who was important to the successful completion of my thesis, and would like to apologize that I could not mention everyone personally one by one.

I would like to appreciate the prompt help, I got from Sandrine, Julien, Beatrice, Pierre Vincens and Bilel, in all my administrative and technical endeavors. Thank you for your kind cooperation.

Someone said, " *Wherever you go, whatever you do, your Family will always be there to love and support you".* I never realized how far I have come in my life and its only because my family was always there to care about me. However hard the times were, they always gave me the strength to find a way out. I am sure that my *Amma (Grandmother)* and *Baba (Grandfather)* must be really happy to see me succeed. Thank you for shaping my childhood with your immense love and nurturing me with your values of life. I cannot express my thankfulness to my *Papa* and *Chacha* (Uncle) in words as I can see myself to be their reflection. They have given me the strength to deal with all sort of stumbling rocks, which comes in the path of success. On the other hand, my *Mumma* and *Chachi* (Aunt) have provided me with the softness in the heart and love for the people around me. I am grateful to them for being an eternal part of my life. They are my first teachers and the four building blocks of my life. My three sisters, *Aneesha Dee*, *Sonal*, *Komal* and my lady luck, *Khwaish* (my niece) are the most beautiful and idealistic inspirations and I would like to thank them all for being supportive as well as for being critically and practically helping me in taking decisions of my life. I also thank my brother-in-laws, *Ankit Jijz* and *Ashish Jijz*, for guiding me and encouraging me for being the true self. I want to thank my extended family, my *Parent-in-laws*, *Aman* and *Priya* for believing in my capabilities and trusting me.

I would like to express my not so formal thankfulness to my childhood friend, *Abhishek* for being there and laughing at my weird situations, making it worse but the best times of my life. I will for sure have to kill you to erase all the truths of my life. *Alok Sir*, you deserve special thanks and respect for guiding and helping me throughout, since we met. *Shruti Mam*, without you the journey would have not been possible. I will always be indebted to your cute little family. *Venkat "boy"*, sharing drinks and talks over the drinks were always helpful any which way. You deserve special thanks, as you were there to

share all the good and bad moments in Paris when I was alone and far away from my family. *Sai boy,* you are a gem too. We made a trio and I would like to thank you both deep down my heart for being there for me. I cannot end this stanza without a word of appreciation for my present and former Maison de L'inde floor mates *"Snig, Mithila, Suraya, Aditya, Soumya, Aman, Kamal, Nupur, Samta, Samarth, Chaitrali, Nishtha, Gaurav, Aayushi, Ashima, Bhavya, Aamir and Neerj*a". The moments we spent together were really an *Incredible India* times.

In the end, I would like to acknowledge my beautiful wife, *Meenu* who is there with me since always. She seems to be an eternal part of my past, present as well as the future. I don't want to thank her but congratulate her for accomplishing this venture with me, hand in hand. I cannot forget to mention the hard times when we were apart and still managed to match the timings to discuss about our work and our life over *Skype* (special thanks to you too, for making the long distances short). She not only discussed but also critically analyzed my work and provided me with her intriguing ideas. Being a PhD student herself, she understood the ups and downs I had to face and was supportive in her best possible way. She is the best mate I can ever ask for. Thank you for being there!!

Last but not the least, I would like to thank the almighty GOD for being there somewhere around me with all the positive energy and omnipresent grace!!

Thesis Content

***Note:** Supplementary files are online.

Download Link: https://www.dropbox.com/s/1o7bnqioyo0p98s/ThesisAR_SupplementaryFiles.zip?dl=0

Thesis Summary

"Reading after a certain age,
diverts the mind too much from its creative pursuits.
Any man who reads too much and
uses his own brain too little falls into lazy habits of thinking"

-Albert Einstein

Since the discovery of *Phaeodactylum tricornutum* by Bohlin in 1897, its classification within the tree of life has been controversial. It was in 1958 when Lewin, using oval and fusiform morphotypes, described multiple characteristic features of this species that resemble diatoms structure, the debate to whether classify *P. tricornutum* as a member of Bacillariophyceae was ended. To this point three morphotypes (oval, fusiform and triradiate) of *Phaeodactylum tricornutum* have been observed. Over the course of approximately 100 years, from 1908 till 2000, 10 strains of *Phaeodactylum tricornutum* (referred to as ecotypes) have been collected and stored axenically as cryopreserved stocks at various stock centers. Various cellular and molecular tools have been established to dissect and understand the physiology and evolution of *P. tricornutum*, and/or diatoms in general. It is because of decades of research and efforts by many laboratories that now *P. tricornutum* is considered to be a model diatom species.

My thesis majorly focuses in understanding the genetic and epigenetic makeup of *P. tricornutum* genome to decipher the underlying morphological and physiological diversity within different ecotype populations. To do so, I established the epigenetic landscape within *P. tricornutum* genome using various histone post-translational modification marks (chapter 1 and chapter 2) and also compared the natural variation in the distribution of some key histone PTMs between two ecotype populations (chapter 4). We also generated a genome-wide genetic diversity map across 10 ecotypes of *P. tricornutum* revealing the presence of a species-complex within the genus *Phaeodactylum* as a consequence of ancient hybridization (Chapter 3). Based on the evidences from many previous reports and similar observations within *P. tricornutum*, we propose natural hybridization as a strong and potential foundation for explaining unprecedented species diversity within the diatom clade. Moreover, we updated the functional and structural annotations of *P. tricornutum* genome (Phatr3, chapter 2) and developed a user-friendly software algorithm to fetch CRISPR/Cas9 targets, which is a basis to perform knockout studies using CRISPR/Cas9 genome editing protocol, in 13 phytoplankton genomes including *P. tricornutum* (chapter 5). To accomplish all this, I used various state-of-the-art technologies like Mass-Spectrometry, ChIP sequencing, Whole genome sequencing, RNA sequencing, CRISPR genome editing protocols and several computational softwares/pipelines. In brief, the thesis work provides a comprehensive platform for future epigenetic, genetic and functional molecular studies in diatoms using *Phaeodactylum tricornutum* as a model. The work is an add-on value to the current state of diatom research by answering questions that have never been asked before and opens a completely new horizon and demand of epigenetics research that underlie the ecological success of diatoms in modern-day ocean.

Summarily, this thesis is based on the following papers, which are either published, under revision, submitted or in preparation:

I. **Achal Rastogi**, Omer Murik, Chris Bowler, and Leila Tirichine. "PhytoCRISP-Ex: a web-based and stand-alone application to find specific target sequences for CRISPR/CAS editing." *BMC bioinformatics* 17, no. 1 (2016): 261.

II. **Achal Rastogi**, Xin Lin, Bérangère Lombard, Damarys Loew, and Leïla Tirichine. "Probing the evolutionary history of epigenetic mechanisms: what can we learn from marine diatoms." *AIMS Genetics* (2015).

III. Alaguraj Veluchamy, **Achal Rastogi**, Xin Lin, Bérangère Lombard, Omer Murik, Yann Thomas, Florent Dingli et al. "An integrative analysis of post-translational histone modifications in the marine diatom Phaeodactylum tricornutum." *Genome biology* 16, no. 1 (2015): 1.

IV. **Achal Rastogi**, Uma Maheswari, Richard G. Dorrell, Florian Maumus, Adam Kustka, James McCarthy, Andy E. Allen, Paul Kersey, Chris Bowler and Leila Tirichine. “Improved annotation of *Phaeodactylum tricornutum* genome reveals novel genes, alternative splicing and extended repertoire of transposable elements.” [In Revision].

V. **Achal Rastogi**, Anne-Flore Deton Cabanillas, Alaguraj Veluchamy, Catherine Cantrel, Gaohong Wang, Pieter Vanormelingen, Chris Bowler, Gwenael Piganeau, Leila Tirichine and Hanhua Hu. “Whole genome sequencing of *Phaeodactylum tricornutum* ecotypes reveals multiple sub-species as a consequence of ancient hybridization” [Submitted].

VI. **Achal Rastogi**, Xin Lin, Alaguraj Veluchamy, Anne Flore Deton Cabanillas, Catherine Cantrel, Javier Paz, Murik Omer, Cédric Gaillard, Chris Bowler and Leila Tirichine. “Insights into the role of H3K27me3 mediated regulation of morphogenesis in the marine diatom *Phaeodactylum tricornutum*” [Manuscript in preparation].

Introduction

Diatoms are highly diversified group of phytoplankton with estimates from few to thousands of species and recent evaluation of their diversity in the TARA Oceans Expedition reveals 4,748 operational taxonomic units (OTUs) distributed in all ocean provinces (Malviya et al. 2016). They are unicellular micro-eukaryotes mostly photosynthetic and found in almost all aquatic environments, from marine to fresh water. C.G. Ehrenberg first discovered them in the 19th century in dust samples collected by Charles Darwin in the Azores (Gorbushina et al. 2007). Diatoms belong to a big group of heterokonts, constituent of chromalveolate (SAR group) which are believed to be derived from serial endosymbiosis acquiring genes from green and red algae predecessors (Bowler et al. 2008; Moustafa et al. 2009). They further diversified via horizontal transfer of bacterial genes (Bowler et al. 2008). According to earliest well-preserved fossil record, diatoms are believed to be around since 190 million years (myr) (Armbrust 2009) and their closest sister group are the Bolidomonads. Based on the characteristic organization of the valve (the intricately sculptured unit of each end of the diatom frustule), diatoms are divided into two major classes: Centrics and Pennates (Round et al. 1990). The Pennates are believed to have evolved from the polar centric diatoms (Figure 1) (Theriot 2010) and are first found in the fossil record of about 70 million years ago (mya) (Montsant et al. 2005).

Phaeodactylum tricornutum: a model species for diatom research

Phaeodactylum tricornutum is a raphid pennate diatom species classified within the suborder Phaeodactylineae, to which *Phaeodactylum* is assigned as the only known genus, till date. Bohlin, in 1897, described *Phaeodactylum tricornutum* as a single known species within genus *Phaeodactylum*. Bohlin based on stellate, three narrow arm structure named the species as *P. tricornutum*. Because of *P. tricornutum* cells description as weakly silicified by Bohlin and its close morphological resemblance with *Nitzschia closterium,* described by Wilson in 1946, the taxonomic assignation of *P. tricornutum* remained controversial until Lewin in 1958 described other morphological forms which are silicified and possess characteristics like golden-brown chromatophores, store leucosin and oil. Precisely, Joyce C. Lewin in 1958 reported the pleomorphic nature of *P. tricornutum* cells and suggested that it can be found in three major morphotypes viz. triradiate, oval and fusiform. He also described some rare morphotypes like cruciform, lunate etc. Recently, cruciform cells are also observed in lab maintained samples isolated from China (He et al. 2014). Among the three major morphotypes, fusiform and triradiate cells are planktonic while oval are benthic. The genome sequence and annotation of *P. tricornutum* was released in 2008 (Bowler et al. 2008). For sequencing, DNA was extracted from monoclonal cultures grown from single cell isolated from an English Channel strain Pt1. The latter cultured strain was denoted as Pt1 8.6. The DNA from Pt1 8.6 cell population was sequenced using Sanger sequencing protocol and assembled into 27.4 mega base pair (Mbp) with 33 chromosomes [12 complete (from one telomeric end to another) and 21 partial (with either one or no telomeric ends)] and 55 scaffolds (designated as bottom drawer, bd). The last publically available functional annotation of *P. tricornutum* genome is Phatr2 that includes 10,402 genes with an average gene length of 1474 bp. These annotations are majorly based on the comprehensively created EST libraries in 16 different conditions (Maheswari et al. 2009; Maheswari et al. 2010). The genome release along with exhaustive EST data source and establishment of numerous genetic

transformation tools (Falciatore et al. 1999; Diner et al. 2016) boosted *P. tricornutum* as a model species for comparative genetics and functional characterization studies of other diatoms. Recent advancement in the field of reverse genetics is dominated by RNAi silencing and TALEN (De Riso et al. 2009; Daboussi et al. 2014; Kaur and Spillane 2015; Weyman et al. 2015). Simpler and more efficient techniques have been recently developed including CRISPR based genome-editing. Be it knocking any gene out off a system or to manipulate the epigenetic makeup of the genome (Vojta et al. 2016), CRISPR based editing tools are now widely used and accepted. Recently, for the first time, researchers from Norway, our lab and others are able to successfully optimize CRISPR/Cas9 technology to efficiently generate stable targeted gene mutations in *P. tricornutum* (Nymark et al. 2016), strengthening its molecular toolbox.

Since the discovery of *P. tricornutum*, many of its isolates from different locations of the world have been sampled (referred to as ecotypes) within a time frame of approximately 100 years, between 1908 and 2000 (Figure 2) (De Martino 2007). All of the strains have been maintained either axenically or with their native bacterial population in different stock centers and cryopreserved after isolation (De Martino 2007).

Although variable cell shapes are observed within individual ecotypes, some forms are dominant, hence can be said as selected, over others. Among the three major morphotypes, fusiform is observed to be highly abundant across most sampled isolates and with the fastest doubling time it was also proposed to be highly adaptable to changing environmental conditions (De Martino 2007; Bartual Ana 2011). *P. tricornutum* is reported as non-abundant and suggested to be found majorly in unsteady coastal locations like seashores, estuaries, rock pools, tidal creeks, etc. The physicochemical conditions in such habitats change rapidly suggesting that evolution and diversity within morphotypes of *P. tricornutum* might be an adaptation mechanism. Studies have reported many instances where an external environmental

stress can induce the morphotype to change from one to another, which in some cases are reversible when the external stress is removed (Figure 3) (Tesson 2009).

Furthermore, many reports have shown that each ecotype has remarkably developed a unique environment-specific functional and phenotypic behavior of ecotypes in response to various environmental cues (Stanley 2007; Bailleul et al. 2010; Abida et al. 2015). However, diversity within the ecotypes was overlooked and no attempts were made at the molecular level to dissect this diversity. Changes in the species diversity/composition within diatoms and other microalgae are frequently regular (Harnstrom 2009). Genetic studies on microalgae during the last decades have revealed strong cryptic diversity with some evident reports from protists (Harnstrom 2009). However, the ecological and evolutionary processes generating new species and their maintenance is yet to be understood. The concept of species within microalgae is debatable because of the use of sequence-based clades in defining cryptic species (Harnstrom 2009), while the underlying genetic difference is also described as intra-genomic variation within a species (Harnstrom 2009). Among dinoflagellates, many genotypes with small genetic variations have been described as clades, ribotypes or species complex, instead of species (Harnstrom 2009). Comparative genomics study using ITS2 region of the genome from 10 isolates of *P. tricornutum* was not able to provide any significant evidence for genetically driven morphological appearance of *P. tricornutum* cells (De Martino 2007). However, the analysis provided a firm opinion over the presence of various genotypes within *P. tricornutum* strains. Thus, in the absence of genome wide analysis and diversity map of *P. tricornutum* ecotypes, it is naïve to produce any hypothesis about, if or not, the genotype hosts any signatures for the evolution of morphogenesis in *P. tricornutum*. However because of an instantaneous nature of *P. tricornutum* morphotypes to switch from one to another under the influence of reported environmental inducers, the mechanism of morphogenesis and other environment induced physiological changes

seems to be driven by more complex non-genetic forces, e.g. epigenetics and not by genetics alone.

Epigenetics is the study of heritable non-genetic changes that are induced by the physical environment around us. These changes provide fitness and establish new characters or phenotypes in an organism to survive better in its changing ecological niche and/or in newly invaded environmental conditions. The epigenetic control of gene regulation within a cell is governed by three components: DNA methylation, Histone protein post-translational modifications and noncoding RNA mediated pathways (Figure 4). In a eukaryotic cell nucleus, DNA exists as a condensed form called chromosomes. These chromosomes are composed of long threads called chromatin and are made by tightly wrapping of DNA around histone protein complex forming compact spherical beads like structure called nucleosomes. The histone protein complex is an octamer unit composed of two copies each of the core histone proteins H3, H4, H2A and H2B. The chain of nucleosomes is then mold into a 30nm spiral called solenoid, where additional H1 linker histone proteins are associated with each nucleosome to maintain the chromosome structure (Ramakrishnan 1997).

Epigenetic mechanisms are well defined within plants and animals, however a lot has yet to be learnt from far diverged clades like stramenopiles. In the following review I have discussed how the study of emerging model organisms like diatoms helps understand the evolutionary history of epigenetic mechanisms with a particular focus on DNA methylation and histone modification landscape in *P. tricornutum*. The review article also provides, in light of studied epigenetic phenomenon in multicellular organisms, a holistic view of the epigenetic code underlying the ecological success of diatoms in the contemporary oceans.

References

Abida H, Dolch LJ, Mei C, Villanova V, Conte M, Block MA, Finazzi G, Bastien O, Tirichine L, Bowler C et al. 2015. Membrane glycerolipid remodeling triggered by nitrogen and phosphorus starvation in Phaeodactylum tricornutum. *Plant physiology* **167**: 118-136.

Armbrust EV. 2009. The life of diatoms in the world's oceans. *Nature* **459**: 185-192.

Bailleul B, Rogato A, de Martino A, Coesel S, Cardol P, Bowler C, Falciatore A, Finazzi G. 2010. An atypical member of the light-harvesting complex stress-related protein family modulates diatom responses to light. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 18214-18219.

Bartual Ana VB, G. Brun Feranando 2011. Monitoring the long-term stability of pelagic morphotypes in the model diatom Pheodactylum triconrutum. *Diatom Research* doi:10.1080/0269249X.2011.619365: 243-253.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP et al. 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456**: 239-244.

Daboussi F, Leduc S, Marechal A, Dubois G, Guyot V, Perez-Michaut C, Amato A, Falciatore A, Juillerat A, Beurdeley M et al. 2014. Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology. *Nat Commun* **5**: 3831.

De Martino AM, A. Juan Shi, K.P. Bowler, C. 2007. Genetic and phenotypic characterization of Phaeodactylum tricornutum (Bacillariophyceae) accessions. *J Phycol* **43**: 992–1009.

De Riso V, Raniello R, Maumus F, Rogato A, Bowler C, Falciatore A. 2009. Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic acids research* **37**: e96.

Diner RE, Bielinski VA, Dupont CL, Allen AE, Weyman PD. 2016. Refinement of the Diatom Episome Maintenance Sequence and Improvement of Conjugation-Based DNA Delivery Methods. *Front Bioeng Biotechnol* **4**: 65.

Falciatore A, Casotti R, Leblanc C, Abrescia C, Bowler C. 1999. Transformation of Nonselectable Reporter Genes in Marine Diatoms. *Mar Biotechnol (NY)* **1**: 239-251.

Gorbushina AA, Kort R, Schulte A, Lazarus D, Schnetger B, Brumsack HJ, Broughton WJ, Favet J. 2007. Life in Darwin's dust: intercontinental transport and survival of microbes in the nineteenth century. *Environ Microbiol* **9**: 2911-2922.

Harnstrom K. 2009. Bloom dynamics and population genetics of marine phytoplankton - Community, species and population aspects.

He L, Han X, Yu Z. 2014. A rare Phaeodactylum tricornutum cruciform morphotype: culture conditions, transformation and unique fatty acid characteristics. *PloS one* **9**: e93922.

Kaur S, Spillane C. 2015. Reduction in carotenoid levels in the marine diatom Phaeodactylum tricornutum by artificial microRNAs targeted against the endogenous phytoene synthase gene. *Mar Biotechnol (NY)* **17**: 1-7.

Maheswari U, Jabbari K, Petit JL, Porcel BM, Allen AE, Cadoret JP, De Martino A, Heijde M, Kaas R, La Roche J et al. 2010. Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome biology* **11**: R85.

Maheswari U, Mock T, Armbrust EV, Bowler C. 2009. Update of the Diatom EST Database: a new tool for digital transcriptomics. *Nucleic acids research* **37**: D1001-1005.

Malviya S, Scalco E, Audic S, Vincent F, Veluchamy A, Poulain J, Wincker P, Iudicone D, de Vargas C, Bittner L et al. 2016. Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences of the United States of America* doi:10.1073/pnas.1509523113.

Matouk CC, Marsden PA. 2008. Epigenetic regulation of vascular endothelial gene expression. *Circ Res* **102**: 873-887.

Montsant A, Jabbari K, Maheswari U, Bowler C. 2005. Comparative genomics of the pennate diatom Phaeodactylum tricornutum. *Plant physiology* **137**: 500-513.

Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D. 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324**: 1724-1726.

Nymark M, Sharma AK, Sparstad T, Bones AM, Winge P. 2016. A CRISPR/Cas9 system adapted for gene editing in marine algae. *Sci Rep* **6**: 24951.

Ramakrishnan V. 1997. Histone structure and the organization of the nucleosome. *Annu Rev Biophys Biomol Struct* **26**: 83-112.

Round FE, Crawford RM, Mann DG. 1990. *The Diatoms: Biology and Morphology of the Genera*. Cambridge University Press, London, UK.

Stanley JAC. 2007. Whole cell adhesion strength of morphotypes and isolates of Phaeodactylum tricornutum (Bacillariophyceae). *European Journal of Phycology* doi:10.1080/09670260701240863.

Tesson BG, C. ; Martin-Jézéquel, V. . 2009. Insights into the polymorphism of the diatom Phaeodactylum tricornutum Bohlin. *Botanica Marina* **52**.

Theriot ECA, M., Ruck, E., Nakov, T, Jensen, R.K. 2010. A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. *Plant Ecology and Evolution* **143**: 278-296.

Vojta A, Dobrinic P, Tadic V, Bockor L, Korac P, Julg B, Klasic M, Zoldos V. 2016. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic acids research* **44**: 5615-5628.

Weyman PD, Beeri K, Lefebvre SC, Rivera J, McCarthy JK, Heuberger AL, Peers G, Allen AE, Dupont CL. 2015. Inactivation of Phaeodactylum tricornutum urease gene using transcription activator-like effector nuclease-based targeted mutagenesis. *Plant Biotechnol J* **13**: 460-470.

Figures

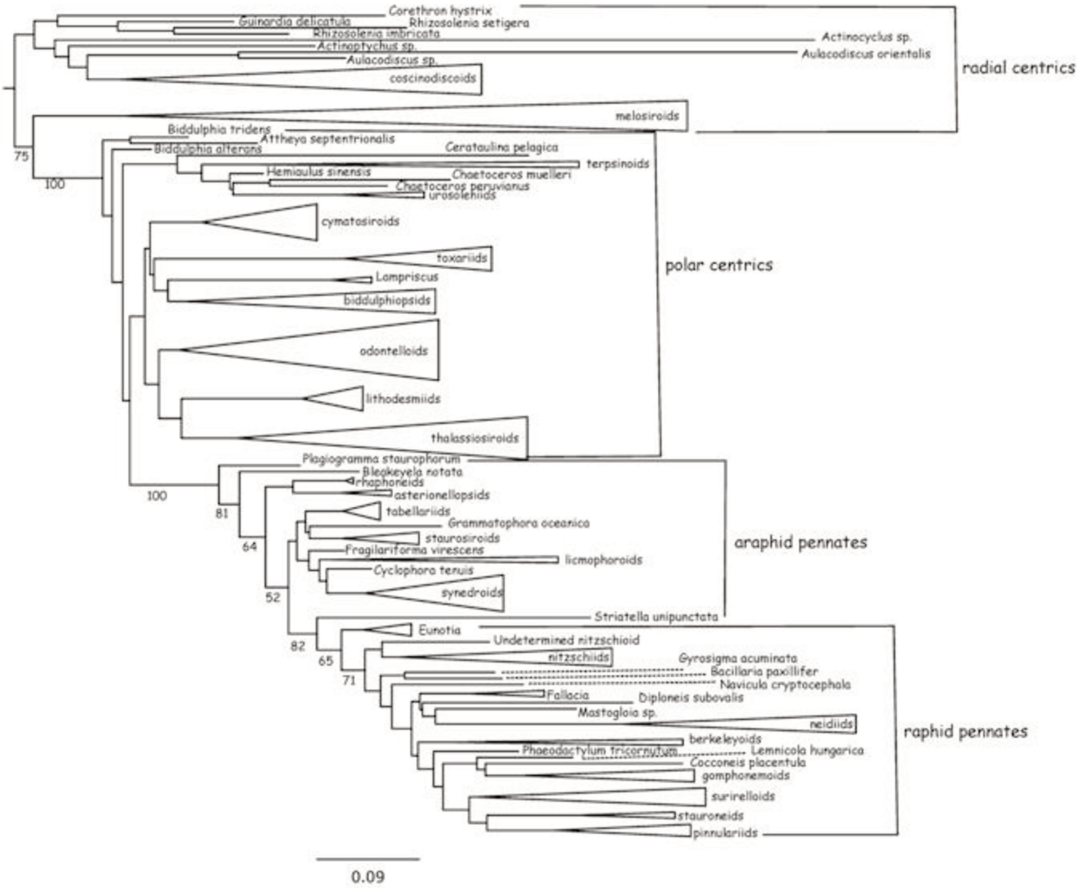


Figure 1. Maximum likelihood phylogeny inferred from SSU. Image is taken from Edward C. Theriot work in 2010 (Theriot 2010).

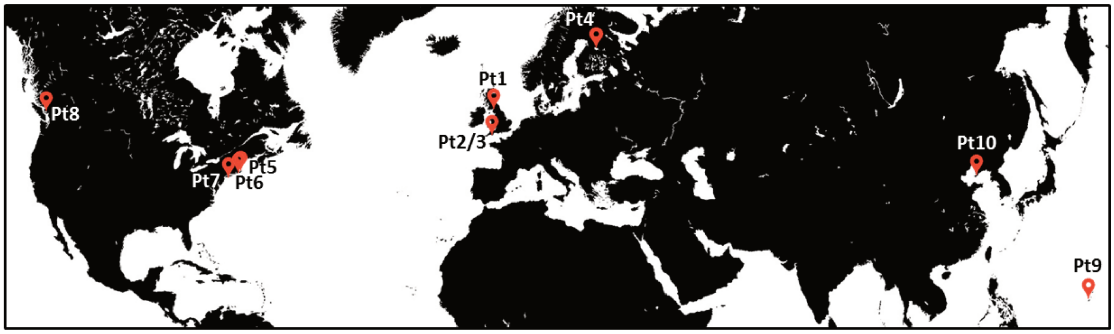


Figure 2. The ecotype world map represents sampling locations of *P. tricornutum*.

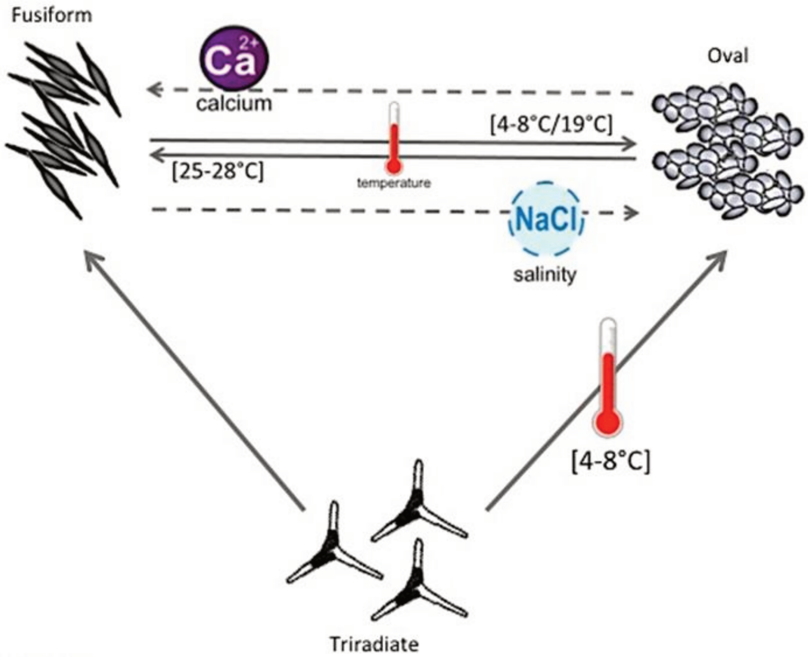


Figure 3. Cartoon representation of the effect of various abiotic stresses on the morphology of *P. tricornutum* cells. The data is taken from (Tesson 2009)

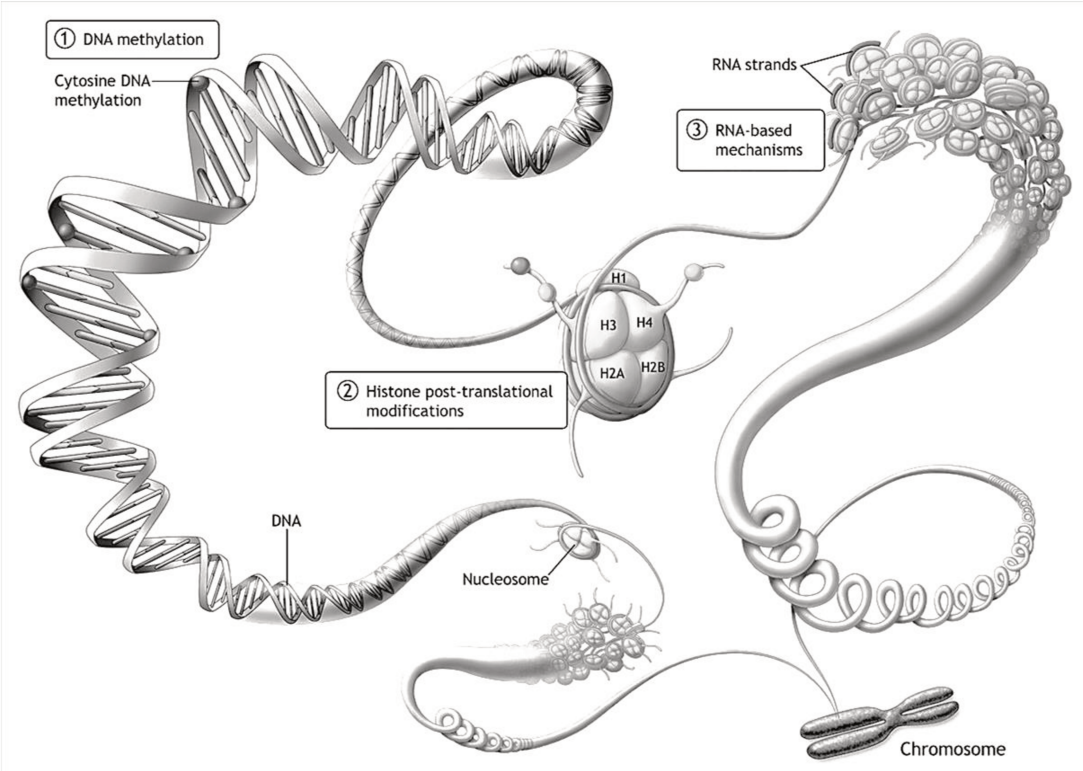


Figure 4. Components of epigenetic control. The image is adapted from (Matouk and Marsden 2008).

Genetics

Volume 2, Issue 3, 173-191.
DOI: 10.3934/genet.2015.3.173
Received date 19 May 2015,
Accepted date 13 July 2015,
Published date 20 July 2015

http://www.aimspress.com/

Review

Probing the evolutionary history of epigenetic mechanisms: what can we learn from marine diatoms

Achal Rastogi[1], Xin Lin[1,2], Bérangère Lombard[3], Damarys Loew[3] and Leïla Tirichine[1,*]

[1] Ecology and Evolutionary Biology Section, Institut de Biologie de l'École Normale Supérieure (IBENS), CNRS UMR8197 INSERM U1024, 46 rue d'Ulm 75005 Paris, France
[2] State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen 361005, China
[3] Institut Curie, PSL Research University, Centre de Recherche, Laboratoire de Spectrométrie de Masse Protéomique, 26 rue d'Ulm 75248 Cedex 05 Paris, France

* **Correspondence:** Email: tirichin@biologie.ens.fr; Tel: 33 1 44 32 35 34.

Abstract: Recent progress made on epigenetic studies revealed the conservation of epigenetic features in deep diverse branching species including Stramenopiles, plants and animals. This suggests their fundamental role in shaping species genomes across different evolutionary time scales. Diatoms are a highly successful and diverse group of phytoplankton with a fossil record of about 190 million years ago. They are distantly related from other super-groups of Eukaryotes and have retained some of the epigenetic features found in mammals and plants suggesting their ancient origin. *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*, pennate and centric diatoms, respectively, emerged as model species to address questions on the evolution of epigenetic phenomena such as what has been lost, retained or has evolved in contemporary species. In the present work, we will discuss how the study of non-model or emerging model organisms, such as diatoms, helps understand the evolutionary history of epigenetic mechanisms with a particular focus on DNA methylation and histone modifications.

Keywords: diatoms; *Phaeodactylum tricornutum*; *Thalassiosira pseudonana*; epigenetics; DNA methylation; histone modifications; non-coding RNA; comparative epigenetics; evolution

1. Introduction

Research in the field of epigenetics has taken off in the last decade as evidenced by the growing

number of published literature and scientific meetings. This is obviously due to numerous findings of its critical role in diseases such as cancer, development and responses to environmental cues in a wide range of species. Epigenetics means in addition to or above genetics implying changes in gene expression without altering the DNA sequence. These changes are inherited from cell to cell and trans-generationally from parent to offspring. Such changes involve chemical modifications of the DNA such as methylation, histone post-translational modifications leading to chromatin modifications, remodeling and attachment to the nuclear matrix, packaging of DNA around nucleosomes and RNA mediated gene silencing. Epigenetic mediated modifications are usually influenced by environmental cues, including diet, physical stresses such as temperature, or chemicals such as toxins and can also be stochastic due to random effects. A striking example is seen in Agouti mice exposed to bisphenol A, a ubiquitous chemical found in our environment. These are genetically identical twins but have a different size and fur color. In slim healthy brown mice, Agouti gene is prevented from transcription by DNA methylation while in yellow obese mice which are prone to diabetes and cancer, the same gene is not methylated resulting in its expression [1,2]. This is a fine example of the trans-generational inheritance of an epigenetic state where the Agouti locus escaped the usual resetting of epigenetic states during reproduction.

In the fruit fly *Drosophila melanogaster*, temperature treatment changes the eye color from white to red, and the treated individual flies pass on the change to their offspring over several generations without further requirement of temperature treatment [3]. The DNA sequence of the gene responsible for eye color remained the same for white eyed parents and red eyed offspring and the change was attributed to a specific histone modification [3]. Consistent with the work described above, a more recent study in *Drosophila* showed that the fission yeast homolog of activation transcription factor 2 (ATF2) that usually contributes to heterochromatin formation becomes phosphorylated leading to its release from heterochromatin upon heat shock or osmotic stress [4]. This new heterochromatin state that does not involve any DNA sequence change is transmitted over multiple generations [4].

In an ecological context, variation of DNA methylation was observed in a wild population of *Viola cazorlensis* which is a perennial plant [5]. Using a modeling approach on data collected over many years, the authors have observed that epigenetic variation is significantly correlated with long-term differences in herbivory, but only weakly with herbivory-related DNA sequence variation suggesting that besides habitat, substrate and genetic variation, epigenetic variation may be an additional, and at least partly independent, factor influencing plant-herbivore interactions in the field [5].

The above-discussed examples show a remarkable conservation of the function of epigenetic mechanisms in regulating gene expression among mammals, plants and invertebrates. This conservation goes beyond these species including early diverging single celled organisms such as microalgae. In this work, we will discuss how the study of non-model or emerging model organisms such as diatoms helps understand the evolutionary history of epigenetic mechanisms with a particular focus on DNA methylation and histone modifications.

2. Diatoms, what are they?

Diatoms are photosynthetic eukaryotic algae with cell sizes that usually range between 10 and 200 μm. They are found in all aquatic habitats including fresh and marine waters. These single celled species belong to Stramenopiles, which are part of the supergroup, Chromalveolates, containing also

the Alveolata, the Haptophyta and Cryptophyceae (Figure 1, [6,7]). Diatoms are one of the most diverse and widespread phytoplankton with more than 100,000 extant species which are divided into two orders: centric that are round with radial symmetry and pennate that are elongate with bilateral symmetry (Figure 2). Fossil evidence suggests that diatoms originated during or before the early Jurassic period (~ 210–144 Mya). They are hypothesized to be derived from successive endosymbiosis where a heterotrophic eukaryotic host engulfed cells, phylogenetically close to red and green alga [8], combining therefore features from both green and red algae predecessors [9]. The diversity of diatoms increased further via the horizontal transfer of bacterial genes [10]. Diatoms and bacteria have indeed co-occurred in common habitats throughout the oceans for more than 200 million years, fostering interactions between these two diverse groups over evolutionary time scales [11]. Diatoms are at the base of the food web contributing to one fifth of the planet's oxygen and representing 40% of primary marine productivity [12]. They therefore play a critical role sustaining life not only in the oceans but also on Earth as a whole through their role in the global carbon cycle. Diatoms are also important for human society, providing food through the aquatic food chain and high value compounds for cosmetic, pharmaceutical and industrial applications.

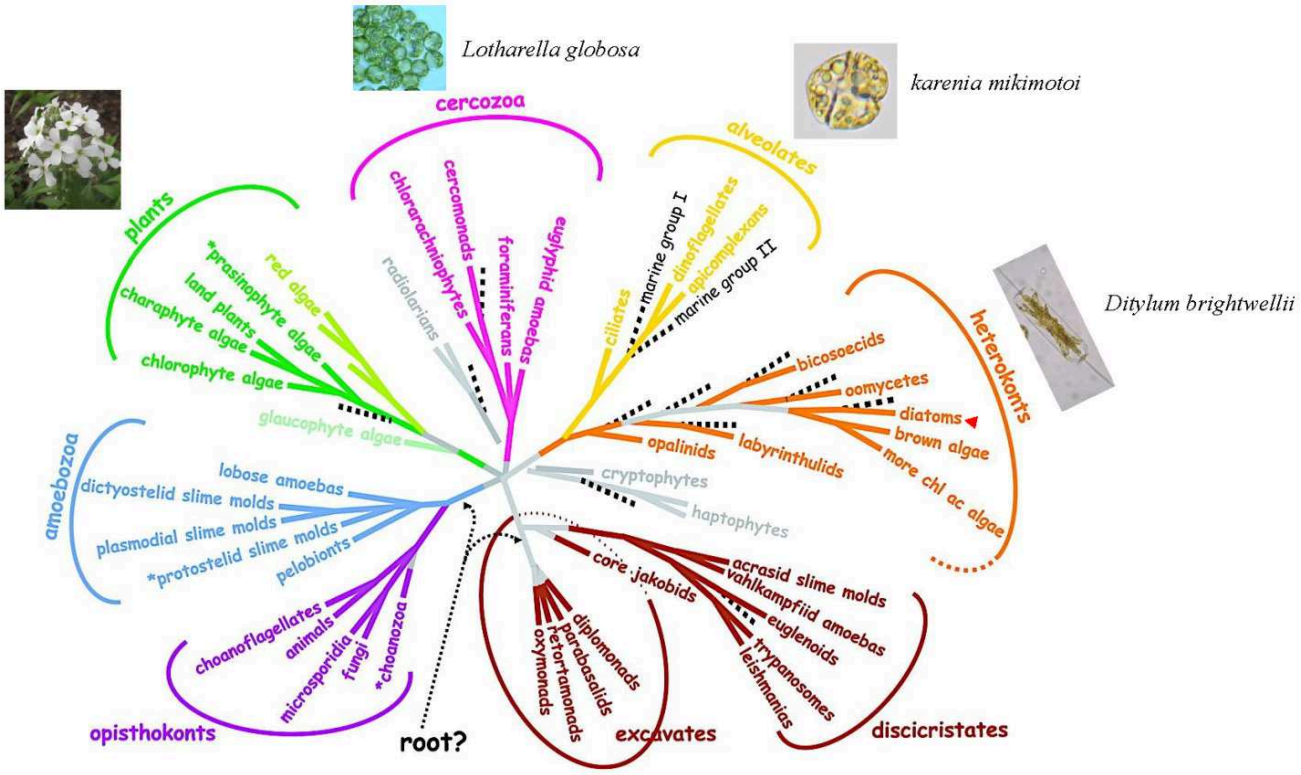


Figure 1. Eukaryote phylogenetic tree. The tree is derived from different molecular phylogenetic and ultrastructural studies (adapted from [13]). Images courtesy of NCMA, the Culture Collection of Marine Phytoplankton at Bigelow Laboratory for Ocean Sciences, and for dinoflagellates (image courtesy of Richard Dorrell). Red arrow head points to diatoms.

Several diatom genome sequences are now available including the two centrics, *Thalassiosira*

pseudonana (32 Mbp), (http://genome.jgi-psf.org/Thaps3/Thaps3.home.html) [14] and *Thalassiosira oceanica* (81.6 Mbp, [15]) *Phaeodactylum tricornutum* (27 Mbp) [10], (http://genome.jgi-psf.org/Phatr2/Phatr2.home.html) (Figure 2), the polar, cold-loving species *Fragilariopsis cylindrus* (80 Mbp; http://genome.jgi-psf.org/Fracy1/Fracy1.home.html), the toxigenic coastal species *Pseudo-nitzschia multiseries* (300 Mb; by the Joint Genome Institute) and the high lipid content diatom *Fistulifera* sp. strain JPCC DA058 [16]. The ecological success of diatoms suggests that they have developed sophisticated ways to cope with changing environments. Complete sequencing of *P. tricornutum* genome [17] showed that it has an unusual genetic composition, which arose through successive endosymbioses and horizontal gene transfers from bacteria. These events have provided diatoms with several unusual metabolic pathways, such as the urea cycle which was previously considered to exist only in animals [14,18]. The ability of diatoms to survive in rapidly changing environments with all the fluctuating conditions (UV radiations, temperature, salinity, toxins, nutrients, grazing pressure etc.) is also attributable to another layer of regulation known as epigenetics [19]. It was previously shown using McrBC, an enzyme sensitive to methylated DNA, that there is an induction of LTR-retrotransposon called Blackbeard (*Bkb*) with a decrease in cytosine methylation under nitrate limitation suggesting that nitrate depletion induces demethylation and upregulation of *Bkb* [20]. Although de novo insertion of *Bkb* was not shown in this study, its distribution with two other retrotransposons was analyzed in thirteen different accessions of *P. tricornutum.* The work showed clear differences in the distribution of the three retrotransposons among the tested accessions demonstrating their transposition in natural environments [20]. Besides this experimental clue of the occurrence of DNA methylation, our recent work [21-24] revealed an amazing conservation of the epigenetic machinery in this model diatom. *P. tricornutum* possesses histone modifying enzymes, small RNA [25,26] as well as DNA methylation which is absent in the multicellular brown algae *Ectocarpus siliculosus* which belongs to Stramenopiles [27].

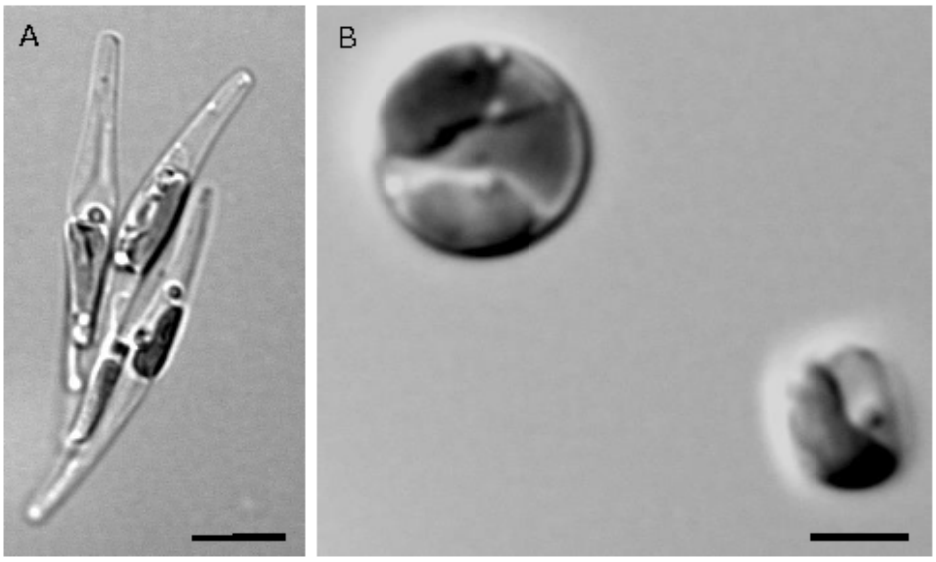


Figure 2. Light microscopy micrographs of representative model diatoms. A. The pennate diatom *Phaeodactylum tricornutum*. Scale bar 5 μm. B. The centric diatom *Thalassiosira pseudonana*. Scale bar 2 μm.

3. DNA methylation

Cytosine DNA methylation is so far the best characterized epigenetic mark. It is a biochemical process in which a methyl group is added to the cytosine pyrimidine ring at position five (5meC) common to all three super kingdoms. Cytosine methylation is a conserved epigenetic mechanism crucial for a number of developmental processes such as regulation of imprinted genes, X-chromosome inactivation, silencing of repetitive elements including viral DNA and transposons and regulation of gene expression [28,29]. DNA methylation is widespread among protists, plants, fungi and animals [30,31]. It is however absent or poor in some species such as the budding yeast *Saccharomyces cerevisiae*, the fruit fly *Drosophila melanogaster*, the nematode worm *Caenorhabditis elegans* and the brown algae *Ectocarpus siliculosus* [27,32].

With the advent of sequencing technologies and their increasing quality in terms of resolution and depth, our view and understanding of DNA methylation in the main supergroups of eukaryotes, plants and animals starts to emerge. The recently published methylome of *P. tricornutum* [23], which is phylogenetically distant from classic model organisms in the animal and green plant groups as well as diverse protists [31,33], drew a better picture and brought more insights into the evolutionary history of DNA methylation. With 27 Mb genome size, *P. tricornutum* shows a low level of DNA methylation compared to other eukaryotes such as human, *Arabidopsis* and the sea squirt *Ciona intestinalis* [31,33,34] (Figure 3). This is not correlated to the size of the genome as evidenced by the higher methylation occurrence of *Ostreoccocus* [33] that have much smaller genome and the low methylation in honey bee [31] whose genome is nearly ten times bigger than *P. tricornutum*. Although few species are compared in Figure 3, increase in cytosine DNA methylation seems to correlate with the average content of transposable elements, which presumably are kept silenced, and the complexity of the genome. Comparative epigenomics or methylomics provide some insights into the genes that might have impacted species evolutionary fate. A striking example are the differentially methylated genic regions (DMRs) found in human and its closely related primates such as chimpanzees, gorillas and orangutans which encode neurological functions suggesting species divergence correlated with developmental specialization [35,36]. In line with these observations, comparative epigenetic analysis of the two diatoms, the pennate *P. tricornutum* and the centric *T. pseudonana* [33], revealed no major differences in the fraction of the genome that is methylated or the context (Figure 4). However, out of 6199 shared genes, 408 are methylated only in *P. tricornutum* versus 461 only in *T. pseudonana*. DMRs between the two species are subsequently reflected in different GO categories enrichment [33] (Figure S1). Investigating further these genes might shed light on the history of their evolutionary divergence.

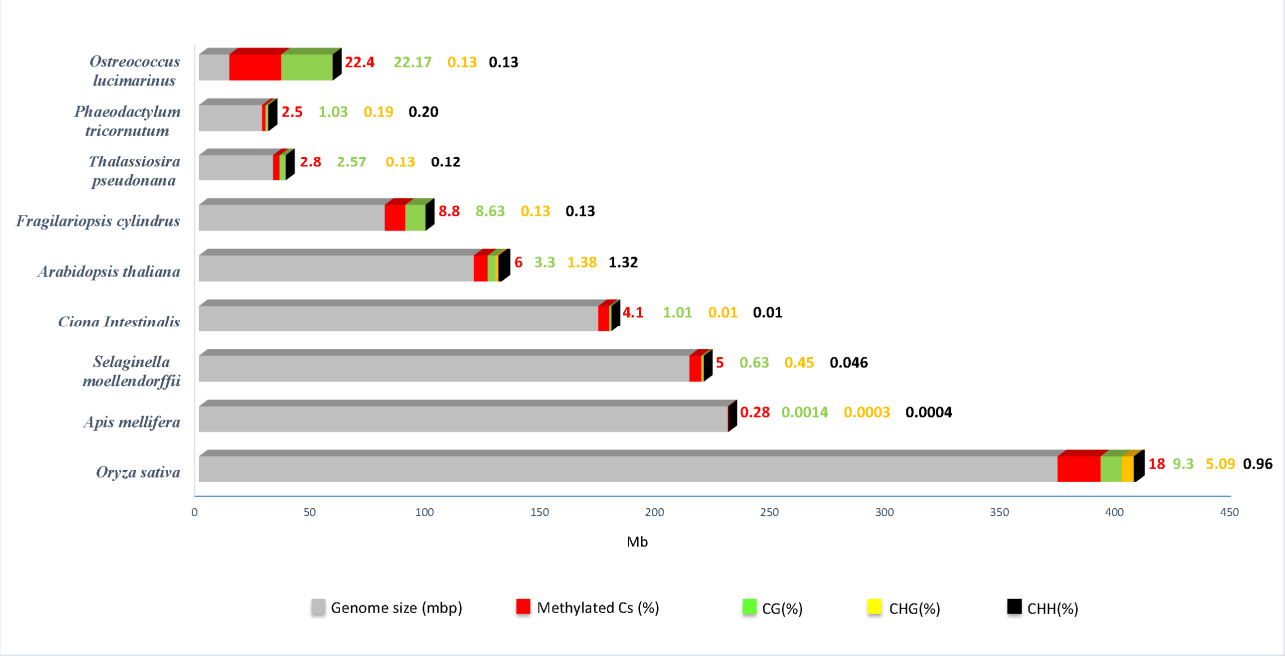


Figure 3. DNA methylation in diverse Eukaryotes. Graphical representation of genome-wide percentages of cytosine DNA methylation as well as in different contexts [C (red), CG (green), CHG (orange) and CHH (black)]. Species names are represented on the Y-axis. All the stated elements are represented as stacks over gray bar indicating the size of each genome measured as mega base pairs (Mbp). For comparison, the human genome methylation data is given: genome size (3381,94), methylated Cs (75%). Data was taken from [33,37,38], http://genome.jgi-psf.org/, http://phytozome.jgi.doe.gov/pz/portal.html.

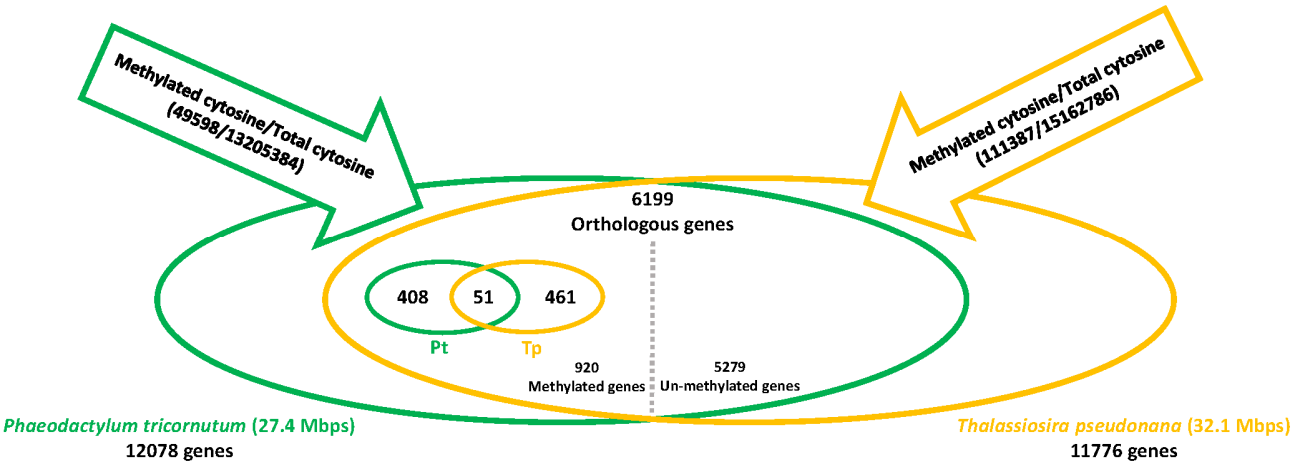


Figure 4. The orthologous gene body cytosine methylation analysis. The genes that are differentially methylated between *Phaeodactylum tricornutum* (Pt) and *Thalassiosira pseudonana* (Tp) are represented. Qualitative analysis of gene body cytosine methylation

on the orthologous genes between Pt and Tp genome. Using reciprocal best-hit BLAST approach, orthologous genes between Pt and Tp genomes are found. Out of 6199 orthologues, 459 genes are methylated in Pt whereas 512 genes are found methylated in Tp genome. The Venn comparison of these genes shows the conservation of gene body cytosine methylation over 51 genes while 408 and 461 genes are specifically methylated in Pt and Tp genomes, respectively. SRA accessions: Tp = GSM1134628; Pt = GSM1134626.

DNA methylation can occur in different contexts including CG, CHG and CHH where H can be any nucleotide except G. In *P. tricornutum,* DNA methylation was found in all contexts suggesting that CHG and CHH is not a plant innovation but existed already in a common ancestor and was lost from certain lineages. Indeed, Eukaryotes have evolved and/or retained different DNA methyltransferase complements responsible for the different context of methylation. Metazoans commonly encode DNMT1 and DNMT3 proteins, while higher plants additionally have plant-specific chromomethylase (CMT). On the other hand, fungi have DNMT1, Dim-2, DNMT4, and DNMT5 [39,40]. Previous phylogenetic analysis suggests that *P. tricornutum* genome encodes a peculiar set of DNMTs as compared to other eukaryotes [41]. DNMT1 appears to be absent in *P. tricornutum* as well as putative proteins coding for plant specific DNA methyltransferase CMT3 and DRM, which are responsible for non CG methylation. *P. tricornutum* encodes DNMT2 (Pt16674), which is an RNA methyltransferase that shows strong sequence similarities with DNA cytosine C5 methyltransferases. In addition to DNMT3 (Pt 46156), diatom genomes also encode DNMT5 (Pt45072) and DNMT6 (Pt36049) proteins as well as a bacterial-like DNMT (Pt47357) [41]. In bacteria, cytosine methylation acts in the restriction-modification system. Thus, the function of a bacterial-like DNMT in *P. tricornutum* is unclear. Interestingly, it is conserved in the centric diatom *T. pseudonana* (Tp 2094), from which pennate diatoms such as *P. tricornutum* diverged ~ 90 million years ago. This implies that a diatom common ancestor acquired DNMT from bacteria after a horizontal gene transfer prior to the centric/pennate diatom split [42]. Conservation of this gene in diatoms over this length of time suggests that it is functional. Because DNMT5 is also found in other algae and fungi, we postulate that it was present in a common ancestor. Furthermore, structural, functional, and phylogenetic data suggest that CMT, Dim-2 and DNMT1 are monophyletic [39,40]. Therefore, we propose that the common ancestor of plants, unikonts and stramenopiles possessed DNMT1 (subsequently lost in diatoms), DNMT3, and probably also DNMT5 (lost in metazoans and higher plants). This evolutionarily important loss is supported by the absence of DNA methyltransferases in the stramenopile *E. siliculosus* [27]. *P. tricornutum* encodes three putative DNA demethylases (Pt46865, Pt48620, Pt12645) with ENDO domain similar to the *Arabidopsis* DNA demethylases ROS1 domain suggesting similar mechanisms for DNA demethylation.

Dnmt5 was reported in a wide range of Eukaryotic single celled species that lack Dnmt1 but nevertheless retain CG methylation which was shown to be catalyzed by Dnmt5 [33]. In this work, the authors used *Cryptococcus neoformans* that has Dnmt5 as a unique DNA methylatransferase and showed that CG methylation is entirely lost when DNMT5 is deleted [33]. However, the authors did not exclude that another unknown methyltransferase catalyzes CG methylation and uses Dnmt5 as a required accessory or regulatory protein [33]. As mentioned above, typical Dnmt1 does not exist in *P. tricornutum* but our in-silico analysis revealed the presence of a gene which seems to be a Dnmt1 remnant protein which lacks the C5 methyltransferase catalytic domain but has retained two motifs

characteristic of Dnmt1, the Bromo-adjacent homology (BAH) domain and a cysteine rich region (ZF_CXXX) that binds zinc ions. In higher Eukaryotes, Dnmt1 is the enzyme that catalyzes CG methylation and the activity of its catalytic domain is regulated by the N terminal region of the protein. Indeed an isolated Dnmt1 catalytic domain was proven to be inactive [43,44]. Interestingly, both BAH and cysteine rich domains are found within the N terminal region of Dnmt1 in higher eukaryotes. A tempting hypothesis would be that *P. tricornutum* Dnmt1-like is the accessory protein that might interact with Dnmt5 to catalyze CG methylation. It is tempting to think that these two domains that are as independent proteins in *P. tricornutum* fused through evolutionary time in a single polypeptide protein in higher Eukaryotes and gave rise to the eukaryotic Dnmt1. We are currently using a reverse genetic approach to determine the function of Dnmts and the putative accessory protein in *P. tricornutum.* The work will help to better understand their role in processes such as maintenance and *de novo* DNA methylation as well as context specificities which will ultimately shed light on the function of DNMTs in an evolutionary context.

P. tricornutum methylome discussed in various studies [23,30,31] confirms the conservation of gene body methylation as an ancient feature and its methylation preference for exons over introns in all Eukaryotic genomes where it has been examined including *Arabidopsis*, *Ciona intestinalis*, honey-bee and human. Several hypotheses were made to explain this specific pattern and interestingly, in-silico analysis of *P. tricornutum* genome revealed few evidences that support them. *P. tricornutum* encodes ROS1 related glycolysases that were thought present only in *Arabidopsis* where they were shown to specifically remove DNA methylation from gene ends [45]. A more universal factor that might explain gene body methylation pattern is the histone mark H3K4me that antagonizes DNA methylation and is distributed around the transcription start site in the genomes where it has been examined. In *P. tricornutum*, H3K4me2 does not localize with DNA methylation and maps around the translation start site [24], which is in line with its potential contribution to DNA methylation pattern at gene bodies.

A conserved function for gene-body methylation at the whole-genome level has not yet been established. When examined, sets of body-methylated genes were found to be expressed constitutively at moderate levels such as in angiosperms and most invertebrates [34,46-48]. Nevertheless, in the silkworm, gene-body methylation correlates positively with gene expression levels [49]. In human, gene body methylation was shown to be involved in X chromosome activation [50] while it was recently reported that methylation of the first exon of autosomal genes correlates with transcriptional silencing [51]. It was also proposed that gene body methylation in human regulates the activity of intragenic alternative promoters [52]. In this line, a recent study [53] has established that body-methylated genes in *A. thaliana* are functionally more important, as measured by phenotypic effects of insertional mutants, than unmethylated genes. Using a probabilistic approach, the authors have reanalyzed single-base resolution bisulfite sequence data from *A. thaliana*. They demonstrated that body methylated genes are likely involved in either suppressing expression from cryptic promoters within coding regions and/or in enhancing accurate splicing of primary transcripts [53]. Interestingly, these functions were already proposed by previous studies [54-56], and the recent comparative study of honey-bee methylome has also established a link between gene-body methylation and splicing [57]. In our study, we found that gene-body methylation in *P. tricornutum* correlates positively with gene length and exon number. It is thus tempting to infer that intragenic methylation in *P. tricornutum* may play a role in avoiding aberrant transcription and/or mis-splicing. Furthermore, functional annotation of body-methylated genes reveals the presence of important

functional classes such as (1) transferases and catalytic enzymes that play important role in cell wall assembly and its rearrangement which is crucial for cell integrity, (2) hydrolase activity which is important in stress responses, and (3) transporter activity necessary for metabolites shuttling such as silicic acid. Considering previous studies and in light of our recent work in *P. tricornutum*, gene body methylation does not suppress expression but rather correlates with low to moderate transcriptional activity. This might have the putative function of preventing aberrant transcription from intragenic promoters and appears to be a common and ancestral eukaryotic feature as reported previously [31,54].

4. Histones and their modifications

Eukaryotic chromosomes are packaged in the nucleus by wrapping the DNA around an octamer of four core histone proteins H2A, H2B, H3 and H4 forming the basic unit of chromatin, the nucleosome. Further compaction is achieved by the interaction of the nucleosome to the linker histone H1. This phenomenon seems to be conserved among all Eukaryotes and even archaea, where the nucleosomes are formed of only a tetramer of two H3 and H4 histones found in the cell, as archaea do not have a nucleus. Furthermore, nucleosome occupancy was found similar in two species of Archaea with depletion over transcriptional start sites as well as a conservation of nucleosome positioning code [58,59]. This demonstration of similarities between Eukaryotes and Archaea chromatin, suggests that histones and chromatin architecture evolved before the divergence of Archaea and Eukarya. This also suggests that the initial function of nucleosomes and chromatin formation might have been for the regulation of gene expression rather than the packaging of DNA, which is an Eukaryotic invention [58].

Histones are subject to a variety of post-translational modifications (PTMs) that have an important role in several processes such as transcription, replication and DNA repair. Histone PTMs in particular at the N terminus include acetylation, methylation, phosphorylation and ubiquitination, which were extensively studied in diverse species, along with modifications like sumoylation, glycosylation, biotinylation, carbonylation, and ADP ribosylation for which little is known [60]. Histone PTMs function either by altering the accessibility of genes to the transcriptional machinery, or by binding to effector proteins via specialized chromatin domains that deposit or erase these histone modifications. PTMs function in a combinatorial pattern known as the histone code, which confers active or repressive chromatin states to specific chromosomal regions of the genome [60,61].

P. tricornutum possesses 14 histone genes encoding 9 histone proteins. They are dispersed throughout five chromosomes with most in clusters of two to six genes as seen for most Eukaryotes. *P. tricornutum* histones belong to the five known classes, histone H1, H3, H4, H2A and H2B. These histones are conserved among diatoms and eukaryotic species. With the exception of histones H4 and H2B, *P. tricornutum* encodes variants for each histone H1, H3 and H2A. Sequence alignment of histone H3 shows the presence of canonical and replacement histones similar to human, H3.2 and H3.3. Additionally, *P. tricornutum* expresses a centromere specific variant commonly called CenH3 that varies considerably from the rest of H3 histones especially in the N terminal tail. CenH3 is essential for recruitment of kinetochores components ensuring correct segregation of chromosomes during mitosis and meiosis [62].

H2A histone members constitute the most diverse group of histones with the greatest number of variants. *P. tricornutum* is no exception as it encodes two copies of the canonical H2A but also both

H2AZ (Pt28445) and H2AX variants while this latter is missing from *C. elegans* and protozoan parasites such as *Plasmodium* and *Trypanosomes*. The presence of the conserved motif SQE/D in the C terminal of *P. tricornutum* H2AX suggests a putative role of this histone in the maintenance of genome integrity via its contribution in the repair of double stranded DNA breaks. *P. tricornutum* encodes two histone H1 variants, which share nearly 50% identity. Interestingly, one of them (Pt44318), is expressed only in stress conditions such as high light which suggests its putative role in DNA repair as found previously in yeast and vertebrates [63,64]. The diversity of histone variants in *P. tricornutum* is interesting and suggests an adaptive evolution to the life history of diatoms via their chromatin interface to acquire new abilities to cope with the changing environment.

P. tricornutum and *T. pseudonana* genome sequencing revealed a long list of histone modifying and demodifying enzymes that are summarized in Table 1. This shows the great conservation of the writers and erasers of histone modification marks in diatoms and their ancient origin. Furthermore, Mass spectrometry analysis (MS) of PTMs in *P. tricornutum* showed similarities to that of plants and mammals including acetylation and/or methylation of several lysines on the N terminal tail of histones H2A, H2B, H3 and H4 and mono, di and tri-methylation of lysines 4, 9, 27 and 36 of histone H3 suggesting the early divergence of these PTMs and their important role in transcriptional regulation of many biological processes (Table 2). Interestingly, *P. tricornutum* combines histone PTMs found in both mammals and plants such as acetylation and mono-di methylation of lysine 79 of histone H3 found only in human and yeast [65] but not in *Arabidopsis* [66] underlying *P. tricornutum* genome diversity and the divergence of histone modifications among species throughout evolution. Another interesting example is the acetylation of lysine 20 of histone H4 which is shared with *Arabidopsis* but different from human where the residue is only methylated [66]. H4K20me which is known to be a repressive mark was detected neither by mass spectrometry nor by western blot using an antibody that recognizes this modification in *Arabidopsis* (data not shown). Furthermore, mono and dimethylation of lysine 79 of histone H4 are modifications that *P. tricornutum* shares only with *Toxoplasma gondii* which is an obligate intracellular parasitic protozoan belonging to Alveolates, a superphylum closely related to Stramenopiles [24]. A non-exhaustive mass spectrometry analysis of histones from an early diverging diatom *Thalassiosira pseudonana* shows the presence of similar histone PTMs (Figure 5), which points to the important role that histone PTMs might have had in shaping diatom genomes and ultimately in the diversification of eukaryotes.

Table 1. **Histone modifications enzymes in two diatom species**. Proteins encoding putative enzymes responsible for histone modification which are identified in *P. tricornutum* and *T. pseudonana.* New gene models are given for *P. tricornutum* (http://protists.ensembl.org/Phaeodactylum_tricornutum/Location/Genome).

Histone Modifiers	**Residues Modified**	**Homologs in *P. tricornutum* (Phatr2)**	**Homologs in *P. tricornutum* (Phatr3)**	**Homologs in *T. pseudonana***
Lysine Acetyltransferases (KATs)				
HAT1 (KAT1)	H4 (K5, K12)	54343	Phatr3_J54343	1397, 22580
GCN5 (KAT2)	H3 (K9, K14, K18, K23, K36)	46915	Phatr3_J2957	15161
Nejire (KAT3);	H3 (K14, K18,	45703, 45764,	Phatr3_J45703, Phatr3_J45764,	24331, 269496,

CBP/p300 (KAT3A/B)	K56) H4 (K5, K8); H2A (K5) H2B (K12, K15)	54505	Phatr3_J54505	263785
MYST1 (KAT8)	H4 (K16)	24733, 24393	Phatr3_J51406, Phatr3_J3062	37928, 36275
ELP3 (KAT9)	H3	50848	Phatr3_J50848	9040
Unknown				
RPD3 (Class I HDACS)	H2, H3, H4	51026, 49800	Phatr3_J51026, Phatr3_J49800	41025, 32098, 261393
HDA1 (Class II HDACS)	H2, H3, H4	45906, 50482, 35869	Phatr3_J45906, Phatr3_J50482, Phatr3_J35869	268655, 269060, 3235, 15819
NAD+ dependent (Class III HDACS)	H4 (K16)	52135, 45850, 24866, 45909, 52718, 21543, 39523	Phatr3_J52135, Phatr3_J45850, Phatr3_J8827, Phatr3_J12305, Phatr3_J16589, Phatr3_J21543, Phatr3_J39523	269475, 264809, 16405, 35693, 264494, 16384, 35956
Lysine Methyltransferases				
MLL	H3 (K4)	40183, 54436, 42693, 47328, 49473, 49476, 44935	Phatr3_EG00277, Phatr3_EG02316, Phatr3_J6915, Phatr3_J47328, Phatr3_EG00277, Phatr3_15913, Phatr3_J44935	35182, 35531, 22757
ASH1/WHSC1	H3 (K4)	43275	Phatr3_6093	264323
SETD1	H3 (K36), H4 (K20)	not found	not found	not found
SETD2	H3 (K36)	50375	Phatr3_EG02211	35510
SETDB1	H3 (K9)	not found	not found	not found
SETMAR	H3 (K4, K36)	not found	not found	not found
SMYD	H3 (K4)	bd1647, 43708	Phatr3_J1647, Phatr3_J43708	23831, 24988
TRX-related		not found	not found	not found
E(Z)	H3 (K9, K27)	32817	Phatr3_J6698	268872
EHMT2	H3 (K9, K27)	not found	not found	not found
SET+JmjC	Unknown	bd1647	Phatr3_J1647	not found
Lysine Demethylases (KDM)				
LSD1 (KDM1)	H3 (K4, K9)	51708, 44106, 48603	Phatr3_J51708, Phatr3_J44106, Phatr3_J48603	not found
FBXL (KDM2)	H3 (K36)	42595	Phatr3_J42595	not found
JMJD2 (KDM4)/JARID	H3 (K9, K36)	48747	Phatr3_J48747	2137
JMJ-MBT	Unknown	48109	Phatr3_J48109	22122
JMJ-CHROMO	Unknown	40322	Phatr3_J40322	1863

Table 2. Diversity of histone PTMs in *P. tricornutum*. Examples of PTMs of histones present in *P. tricornutum* but absent or not detected (ND) in representative of two major lineages, animals and plants. Data taken from [24,66,67].

Histone PTM	*P. tricornutum*	*H. sapiens*	*A. thaliana*
H4K31	present	ND	ND
H4K59Ac	present	ND	ND
H4K59me	present	ND	ND
H4K79me	present	ND	ND
H4K79me2	present	ND	ND
H4K20Ac	present	ND	present
H4K20me	present	present	ND
H3K79me	present	present	ND
H3K79me2	present	present	ND
H2BK107Ac	present	ND	ND

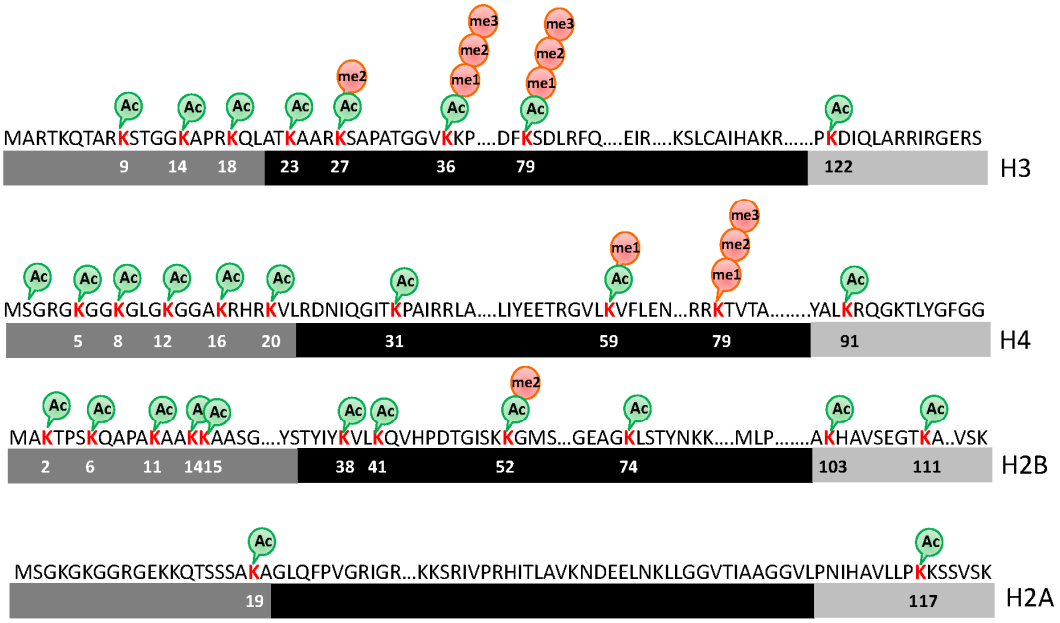


Figure 5. Histone PTMs in *T. pseudonana.* Diagram showing sites of PTMs of core and variant histones identified in *Thalassiosira pseudonana* by mass spectrometry. Amino acid residue number is indicated below the peptide sequence. Dark gray, black and light gray boxes indicate N-terminal, globular core and C-terminal domains, respectively. Acetylation and methylation are indicated in green and red respectively.

5. Non-coding RNA

Non-coding RNA is found in all kingdoms of life with fractions varying from 8% for bacteria to more than 98% for human genome (Figure 6). This non-coding fraction comprises functional non-coding RNAs such as transfer, ribosomal and regulatory RNAs as well as DNA that remains

untranscribed or gives rise to RNA molecules of unknown function. Genome size correlates positively with the amount of non-coding DNA and evolutionary age of the species suggesting that the smaller and early diverging the species are, the less non-coding fraction of their genome they have (Figure 6). This also suggests that non-coding RNAs arose with the complexity of species and the plethora of subsequent novel functions. Although initially argued to be spurious transcriptional noise or accumulated evolutionary debris arising from the early assembly of genes and/or the insertion of mobile genetic elements, we have now evidence suggesting that the previously named "junk DNA" may play a major biological role in cellular development, physiology and pathologies [68]. It is also argued that not all of it will be functional as the transcription machinery is not perfect and will generate non-coding RNA with no fitness advantage and simply tolerating them would be more feasible than evolving and maintaining more rigorous control mechanisms that could prevent their production [69]. Non-coding RNAs that appear to have an epigenetic function including heterochromatin formation, DNA methylation, histone modifications and transcriptional silencing can be divided into two main categories based on their length: short non-coding RNAs (< 30 nts) and long non-coding RNAs (> 200 nts). Short interfering RNAs (siRNA) of 21 nucleotides are produced by long double stranded RNA through a cleavage by the endonuclease Dicer and are bound by an Argonaute protein. They recognize and silence their target mRNAs by perfect sequence complementarity which is in contrast to micro RNAs (miRNAs, 20 to 23 nts) which silence their target sequences by incomplete homology and act primarily at the translational level. Long non-coding RNAs (lncRNAs) have been reported in several eukaryotic genomes including mouse [70], human [71], *Arabidopsis* [72] and Zebrafish [73].

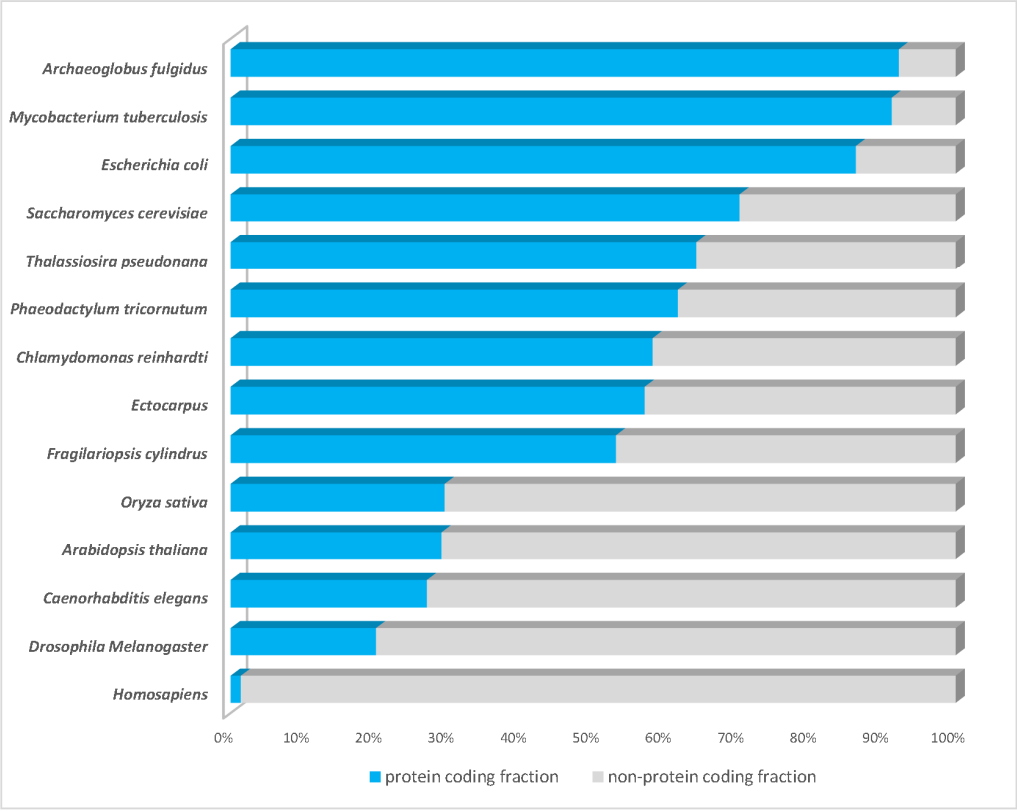


Figure 6. The percentage of coding fraction of several Eukaryotic and bacterial genomes (Adapted from [68]).

Non-coding RNAs are highly diverse and new classes are constantly being discovered. For an exhaustive list of known non-coding RNAs, refer to [74]. Non-coding RNA are known to occur in a wide range of species including human, insects, fish, plants, yeast, protists, even bacteria and archaea, underlying a conserved phenomenon. In *Chlamydomoans reinhardtii*, two studies reported the existence of miRNA that are reminiscent of the miRNAs of multicellular organisms as well as the phased transacting siRNAs (tasiRNAs) of plants. *Chlamydomonas* miRNA do not seem to have sequence homology to any known miRNAs in animals or plants, suggesting that miRNA genes may have evolved independently in the lineages leading to animals, plants and green algae [75,76]. The discovery of small RNA in diatoms and cocolithophores further confirmed the early divergence of such molecules [25,77,78].

6. Conclusions and future perspectives

Although epigenetics is recognized for its fundamental role in diseases such as cancer, there is still a long way to go before we appreciate its importance in shaping species genomes through evolutionary time scales. Epigenetics allows individuals and populations to cope with biotic and abiotic stresses and respond to environmental cues through its dynamic regulation of genes but also provides progenies with a better fitness when the parents experience a particular stress affecting therefore their evolutionary potential. This is exemplified by DNA methylation that acts as an inducer of mutations in DNA sequences via the deamination process impacting therefore genome nucleotide sequences. These mutations in chromosomal DNA might have an effect on the fitness and evolution of individuals and populations. Using model or non-model single celled eukaryotes such as diatoms which constitute an early diverging branch in the evolutionary tree will provide a solid complement to multicellular organisms to enhance our understanding of the impact and true contribution of epigenetics to biological processes and ultimately to their evolutionary history. It is becoming clear now that it is important to include epigenetics and its impact on the evolutionary biology of species in our way of thinking and designing of experiments in biology.

Acknowledgments

AR is a PhD student funded by the MEMO LIFE International PhD program.

Conflict of interest

All authors declare no conflicts of interest in this paper.

Supplementary

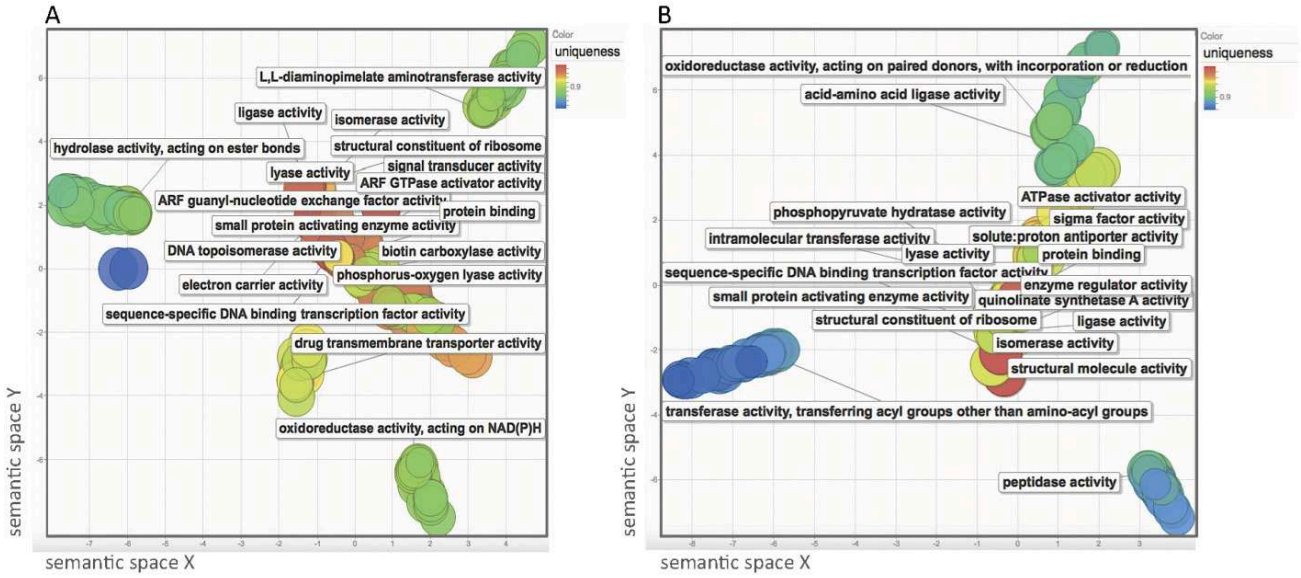


Figure S1. Gene Ontology (GO) enrichment analysis based on semantic clustering of molecular function (MF) associated to *P. tricornutum-T. pseudonana* orthologous genes which are (A) methylated only in *P. tricornutum* and (B) methylated in *T. pseudonana.* X and the Y axis represent the pairwise semantic similarity scores. Color in the sphere represents the uniqueness of each term when compared semantically to the whole list of molecular functions. More unique terms tends to be less dispensable. The graph was generated using Revigo [79].

References

1. Dolinoy DC (2008) The agouti mouse model: an epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome. *Nutr Rev* 66 Suppl 1: S7-11.
2. Dolinoy DC (2007) Epigenetic gene regulation: early environmental exposures. *Pharmacogenomics* 8: 5-10.
3. Tariq M, Nussbaumer U, Chen Y, et al. (2009) Trithorax requires Hsp90 for maintenance of active chromatin at sites of gene expression. *Proc Natl Acad Sci U S A* 106: 1157-1162.
4. Seong KH, Li D, Shimizu H, et al. (2011) Inheritance of stress-induced, ATF-2-dependent epigenetic change. *Cell* 145: 1049-1061.
5. Herrera CM, Bazaga P (2011) Untangling individual variation in natural populations: ecological, genetic and epigenetic correlates of long-term inequality in herbivory. *Mol Ecol* 20: 1675-1688.
6. Dorrell RG, Smith AG (2011) Do red and green make brown?: perspectives on plastid acquisitions within chromalveolates. *Eukaryot Cell* 10: 856-868.
7. Walker G, Dorrell RG, Schlacht A, et al. (2011) Eukaryotic systematics: a user's guide for cell biologists and parasitologists. *Parasitology* 138: 1638-1663.
8. Archibald JM (2009) The puzzle of plastid evolution. *Curr Biol* 19: R81-88.

9. Moustafa A, Beszteri B, Maier UG, et al. (2009) Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* 324: 1724-1726.
10. Bowler C, Allen AE, Badger JH, et al. (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* 456: 239-244.
11. Amin SA, Parker MS, Armbrust EV (2012) Interactions between diatoms and bacteria. *Microbiol Mol Biol Rev* 76: 667-684.
12. Falkowski PG, Barber RT, Smetacek VV (1998) Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* 281: 200-207.
13. Baldauf SL (2008) An overview of the phylogeny and diversity of eukaryotes. *J Syst Evol* 46: 263-273.
14. Armbrust EV, Berges JA, Bowler C, et al. (2004) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* 306: 79-86.
15. Lommer M, Specht M, Roy AS, et al. (2012) Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biol* 13: R66.
16. Tanaka T, Maeda Y, Veluchamy A, et al. (2015) Oil Accumulation by the Oleaginous Diatom Fistulifera solaris as Revealed by the Genome and Transcriptome. *Plant Cell* 27: 162-176.
17. Bowler C, De Martino A, Falciatore A (2010) Diatom cell division in an environmental context. *Curr Opin Plant Biol* 13: 623-630.
18. Allen AE, Dupont CL, Obornik M, et al. (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* 473: 203-207.
19. Tirichine L, Bowler C (2011) Decoding algal genomes: tracing back the history of photosynthetic life on Earth. *Plant J* 66: 45-57.
20. Maumus F, Allen AE, Mhiri C, et al. (2009) Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics* 10: 624.
21. Lin X, Tirichine L, Bowler C (2012) Protocol: Chromatin immunoprecipitation (ChIP) methodology to investigate histone modifications in two model diatom species. *Plant Methods* 8: 48.
22. Tirichine L, Lin X, Thomas Y, et al. (2014) Histone extraction protocol from the two model diatoms Phaeodactylum tricornutum and Thalassiosira pseudonana. *Mar Genomics* 13: 21-25.
23. Veluchamy A, Lin X, Maumus F, et al. (2013) Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum. *Nat Commun* 4: 2091.
24. Veluchamy A, Rastogi A, Lin X, et al. (2015) An integrative analysis of post-translational histone modifications in the marine diatom Phaeodactylum tricornutum. *Genome Biol* 16: 102.
25. Rogato A, Richard H, Sarazin A, et al. (2014) The diversity of small non-coding RNAs in the diatom Phaeodactylum tricornutum. *BMC Genomics* 15: 698.
26. Huang A, He L, Wang G (2011) Identification and characterization of microRNAs from Phaeodactylum tricornutum by high-throughput sequencing and bioinformatics analysis. *BMC Genomics* 12: 337.
27. Cock JM, Sterck L, Rouze P, et al. (2010) The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* 465: 617-621.
28. Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* 9: 465-476.
29. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11: 204-220.

30. Feng S, Cokus SJ, Zhang X, et al. (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* 107: 8689-8694.
31. Zemach A, McDaniel IE, Silva P, et al. (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328: 916-919.
32. Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* 74: 481-514.
33. Huff JT, Zilberman D (2014) Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell* 156: 1286-1297.
34. Zhang X, Yazaki J, Sundaresan A, et al. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* 126: 1189-1201.
35. Molaro A, Hodges E, Fang F, et al. (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* 146: 1029-1041.
36. Zeng J, Konopka G, Hunt BG, et al. (2012) Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am J Hum Genet* 91: 455-465.
37. Honeybee Genome Sequencing C (2006) Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* 443: 931-949.
38. Satou Y, Mineta K, Ogasawara M, et al. (2008) Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: new insight into intron and operon populations. *Genome Biol* 9: R152.
39. Goll MG, Kirpekar F, Maggert KA, et al. (2006) Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* 311: 395-398.
40. Ponger L, Li WH (2005) Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. *Mol Biol Evol* 22: 1119-1128.
41. Maumus F, Rabinowicz P, Bowler C, et al. (2011) Stemming Epigenetics in Marine Stramenopiles. *Current Genomics* 12: 357-370.
42. Bowler C, Vardi A, Allen AE (2010) Oceanographic and biogeochemical insights from diatom genomes. *Ann Rev Mar Sci* 2: 333-365.
43. Zimmermann C, Guhl E, Graessmann A (1997) Mouse DNA methyltransferase (MTase) deletion mutants that retain the catalytic domain display neither de novo nor maintenance methylation activity in vivo. *Biol Chem* 378: 393-405.
44. Fatemi M, Hermann A, Pradhan S, et al. (2001) The activity of the murine DNA methyltransferase Dnmt1 is controlled by interaction of the catalytic domain with the N-terminal part of the enzyme leading to an allosteric activation of the enzyme after binding to methylated DNA. *J Mol Biol* 309: 1189-1199.
45. Penterman J, Zilberman D, Huh JH, et al. (2007) DNA demethylation in the Arabidopsis genome. *Proc Natl Acad Sci U S A* 104: 6752-6757.
46. Cokus SJ, Feng S, Zhang X, et al. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452: 215-219.
47. Hunt BG, Brisson JA, Yi SV, et al. Functional conservation of DNA methylation in the pea aphid and the honeybee. *Genome Biol Evol* 2: 719-728.
48. Foret S, Kucharski R, Pittelkow Y, et al. (2009) Epigenetic regulation of the honey bee transcriptome: unravelling the nature of methylated genes. *BMC Genomics* 10: 472.

49. Xiang H, Zhu J, Chen Q, et al. (2010) Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nat Biotechnol* 28: 516-520.
50. Hellman A, Chess A (2007) Gene body-specific methylation on the active X chromosome. *Science* 315: 1141-1143.
51. Brenet F, Moh M, Funk P, et al. DNA methylation of the first exon is tightly linked to transcriptional silencing. *PLoS One* 6: e14524.
52. Maunakea AK, Nagarajan RP, Bilenky M, et al. (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466: 253-257.
53. Takuno S, Gaut BS (2012) Body-Methylated Genes in Arabidopsis thaliana Are Functionally Important and Evolve Slowly. *Mol Biol Evol* 29: 219-227.
54. Zilberman D, Gehring M, Tran RK, et al. (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* 39: 61-69.
55. Lorincz MC, Dickerson DR, Schmitt M, et al. (2004) Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol* 11: 1068-1075.
56. Luco RF, Pan Q, Tominaga K, et al. (2010) Regulation of alternative splicing by histone modifications. *Science* 327: 996-1000.
57. Lyko F, Foret S, Kucharski R, et al. (2010) The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol* 8: e1000506.
58. Ammar R, Torti D, Tsui K, et al. (2012) Chromatin is an ancient innovation conserved between Archaea and Eukarya. *Elife* 1: e00078.
59. Nalabothula N, Xi L, Bhattacharyya S, et al. (2013) Archaeal nucleosome positioning in vivo and in vitro is directed by primary sequence motifs. *BMC Genomics* 14: 391.
60. Mersfelder EL, Parthun MR (2006) The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. *Nucleic Acids Res* 34: 2653-2662.
61. Cosgrove MS, Boeke JD, Wolberger C (2004) Regulated nucleosome mobility and the histone code. *Nat Struct Mol Biol* 11: 1037-1043.
62. Lermontova I, Schubert V, Fuchs J, et al. (2006) Loading of Arabidopsis centromeric histone CENH3 occurs mainly during G2 and requires the presence of the histone fold domain. *Plant Cell* 18: 2443-2451.
63. Hashimoto H, Sonoda E, Takami Y, et al. (2007) Histone H1 variant, H1R is involved in DNA damage response. *DNA Repair (Amst)* 6: 1584-1595.
64. Maheswari U, Jabbari K, Petit JL, et al. (2010) Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome Biol* 11: R85.
65. Bheda P, Swatkoski S, Fiedler KL, et al. (2012) Biotinylation of lysine method identifies acetylated histone H3 lysine 79 in Saccharomyces cerevisiae as a substrate for Sir2. *Proc Natl Acad Sci U S A* 109: E916-925.
66. Zhang K, Sridhar VV, Zhu J, et al. (2007) Distinctive core histone post-translational modification patterns in Arabidopsis thaliana. *PLoS One* 2: e1210.
67. Tan M, Luo H, Lee S, et al. (2011) Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* 146: 1016-1028.
68. Sana J, Faltejskova P, Svoboda M, et al. (2012) Novel classes of non-coding RNAs and cancer. *J Transl Med* 10: 103.

69. Ulitsky I, Bartel DP (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell* 154: 26-46.
70. Okazaki Y, Furuno M, Kasukawa T, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420: 563-573.
71. Cabili MN, Trapnell C, Goff L, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25: 1915-1927.
72. Liu J, Jung C, Xu J, et al. (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell* 24: 4333-4345.
73. Pauli A, Valen E, Lin MF, et al. (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* 22: 577-591.
74. Cech TR, Steitz JA (2014) The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* 157: 77-94.
75. Molnar A, Schwach F, Studholme DJ, et al. (2007) miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature* 447: 1126-1129.
76. Zhao T, Li G, Mi S, et al. (2007) A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. *Genes Dev* 21: 1190-1203.
77. Lopez-Gomollon S, Beckers M, Rathjen T, et al. (2014) Global discovery and characterization of small non-coding RNAs in marine microalgae. *BMC Genomics* 15: 697.
78. Norden-Krichmar TM, Allen AE, Gaasterland T, et al. (2011) Characterization of the small RNA transcriptome of the diatom, Thalassiosira pseudonana. *PLoS One* 6: e22870.
79. Supek F, Bosnjak M, Skunca N, et al. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6: e21800.

AIMS Press

Chapter 1

The Histone Code

Post-translational modifications of histone proteins are one of the two components of epigenetic code that regulates and maintains the functional physiological behavior of a cell. The current chapter discusses the use of an integrative approach combining Mass Spectrometry (MS), Chromatin Immunoprecipitation (ChIP) and RNA-Seq to introduce the histone code in *Phaeodactylum tricornutum*, along with the functional characterization of five histone post-translational modifications. This study provides the first landscape of various histone modifications and their impact in shaping the genome using a stramenopile species. In the current chapter, we discuss thoroughly the role of different marks in maintaining active and repressive state of the genome of *P. tricornutum*. Furthermore, to decipher the dynamic role of histone marks in maintaining the cellular physiology of *P. tricornutum* in response to the environmental cues, we investigated the role and functional profile of some key histone marks under nitrate stress conditions. In conclusion, the combinatorial analysis of histone PTMs revealed different chromatin states and gene expression patterns, extending the histone code to Stramenopiles.

Veluchamy et al. Genome Biology (2015) 16:102
DOI 10.1186/s13059-015-0671-8

Genome Biology

RESEARCH Open Access

An integrative analysis of post-translational histone modifications in the marine diatom *Phaeodactylum tricornutum*

Alaguraj Veluchamy[1,6], Achal Rastogi[1], Xin Lin[1,7], Bérangère Lombard[2], Omer Murik[1], Yann Thomas[1], Florent Dingli[2], Maximo Rivarola[3,8], Sandra Ott[3], Xinyue Liu[3], Yezhou Sun[3], Pablo D. Rabinowicz[3], James McCarthy[4], Andrew E. Allen[4,5], Damarys Loew[2], Chris Bowler[1*] and Leïla Tirichine[1*]

Abstract

Background: Nucleosomes are the building blocks of chromatin where gene regulation takes place. Chromatin landscapes have been profiled for several species, providing insights into the fundamental mechanisms of chromatin-mediated transcriptional regulation of gene expression. However, knowledge is missing for several major and deep-branching eukaryotic groups, such as the Stramenopiles, which include the diatoms. Diatoms are highly diverse and ubiquitous species of phytoplankton that play a key role in global biogeochemical cycles. Dissecting chromatin-mediated regulation of genes in diatoms will help understand the ecological success of these organisms in contemporary oceans.

Results: Here, we use high resolution mass spectrometry to identify a full repertoire of post-translational modifications on histones of the marine diatom *Phaeodactylum tricornutum*, including eight novel modifications. We map five histone marks coupled with expression data and show that *P. tricornutum* displays both unique and broadly conserved chromatin features, reflecting the chimeric nature of its genome. Combinatorial analysis of histone marks and DNA methylation demonstrates the presence of an epigenetic code defining activating or repressive chromatin states. We further profile three specific histone marks under conditions of nitrate depletion and show that the histone code is dynamic and targets specific sets of genes.

Conclusions: This study is the first genome-wide characterization of the histone code from a stramenopile and a marine phytoplankton. The work represents an important initial step for understanding the evolutionary history of chromatin and how epigenetic modifications affect gene expression in response to environmental cues in marine environments.

Background

Eukaryotic histones are small proteins involved in the formation of nucleosomes, the basic repeating unit of chromatin comprising a histone core around which approximately 146 base pairs of DNA wrap, allowing it to be packaged into the nucleus [1]. The histone core consists of a histone octamer comprising two copies of histone H2A and H2B dimers and one copy of a histone H3-H4 tetramer, all linked to the next nucleosome by the histone linker H1, which appears to be an essential element for stabilizing the folding and condensation of chromatin [2]. Histones are substrates for a diverse range of post-translational modifications (PTMs). These PTMs can occur alone or in a combinatorial fashion (known as the 'histone code') and define dynamic transitions between active and silent chromatin states that co-regulate important biological processes [3]. PTMs occur primarily on the N-terminal tails of histones but also on their globular domains and their C-termini, and include methylation, acetylation, phosphorylation, ubiquitination, sumoylation, citrullination, ADP-ribosylation, hydroxylation, and crotonylation of specific residues [4]. While histone acetylation is generally associated with

* Correspondence: cbowler@biologie.ens.fr; leila.tirichine@ens.fr
[1]Ecology and Evolutionary Biology Section, Institut de Biologie de l'École Normale Supérieure (IBENS), CNRS UMR8197 INSERM U1024, 46 rue d'Ulm, 75005 Paris, France
Full list of author information is available at the end of the article

gene activation, methylation of specific lysine residues can be associated with either active or silent chromatin states depending on the residue that is modified, and whether it is mono-, di-, or tri-methylated. Furthermore, histone phosphorylation is involved in transcriptional regulation of a wide range of biological processes, such as mitosis, DNA replication and damage repair, stress responses, and activation of transcription.

Histones are one of the most highly conserved groups of proteins throughout evolution, highlighting their important role in living organisms. They have been found in almost all eukaryotes so far examined, and although they are not found in bacteria, they do occur in some Archaea [5], indicating their ancient origin. They have been extensively studied in several model organisms, including human, *Drosophila*, yeast and *Arabidopsis*. However, little is known about their role in genome organization in phylogenetically distant groups of eukaryotes beyond the Opisthokonta (including metazoans and fungi) and the Archaeplastida (higher plants, green and red algae).

The chromalveolate group is one of the most diverse groups of eukaryotes, and includes ciliates and dinoflagellates (members of the Alveolata), as well as oomycetes and diatoms (representatives of the Stramenopila, also known as Heterokonta) [6]. Very little is known about the genome structure of these organisms. Ciliates, for example, show a peculiar genome organization reminiscent of the germline-soma distinction in other eukaryotes, with a macronucleus, where transcription of protein coding genes takes place, and a germline micronucleus, which remains silenced [7]. This diversity in genome organization is also seen in dinoflagellates, whose chromosomes are attached to the nuclear membrane and lack canonical histones [8].

Diatoms (Bacillariophyta) are one of the major groups of chromalveolates. Although chronically understudied from a molecular perspective, they are a fundamental component of phytoplankton in most aquatic ecosystems, and are believed to contribute around 40 % of primary production in marine ecosystems [9]. Whole genome sequencing of two marine diatoms, *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*, has revealed their unusual genomic composition, proposed to be a result of endosymbiotic gene transfers involving green and red algae, as well as a significant amount of horizontal gene transfer from bacteria [10]. The combination of genes from different origins has attributed them with novel metabolic capacities for photosynthetic organisms, such as fatty acid oxidation pathways and a urea cycle centered in their mitochondria [11]. These pathways are central hubs of diatom primary metabolism and are also used for diatom-specific processes, such as the construction of their silicified cell walls, known as frustules [11, 12].

Diatoms are remarkably successful organisms with a broad distribution in contemporary oceans and with a well-known capacity to adapt rapidly and outcompete other phytoplankton when favorable conditions arise [13], suggesting that epigenetic regulation mechanisms might contribute to their success. We therefore used high accuracy mass spectrometry (MS) to draw a comprehensive landscape of PTMs in *P. tricornutum*. Using chromatin immunoprecipitation (ChIP), we generated whole genome maps of five PTMs and compared their distributions with a previously generated DNA methylation landscape [14]. Finally, we demonstrate the dynamic nature of the chromatin code by revealing changes in response to nutrient limitation.

Results

Identification of histone PTMs using mass spectrometry

The *P. tricornutum* genome encodes 14 histone genes dispersed on 5 of the 34 chromosome scaffolds characterized previously [15]. Most are found in clusters of two to six genes, as seen in other, albeit not all, eukaryotes such as the ciliates *Stylonychia lemnae*, *Tetrahymena thermophila* and related species [16, 17]. The phylogenetic clustering of *P. tricornutum* histones in doublets of H3-H4 and H2A-H2B reflects their similar evolutionary history, which involves the progressive diversification and differentiation of the four core histone families through a mechanism of recurrent gene duplication [18] (Additional file 1). While histones H4 and H2B are highly conserved, the *P. tricornutum* genome encodes one variant of histone H1, and two variants of each histone H3 and H2A. Further *in silico* analysis revealed that the 27 Mb genome contains a plethora of histone-modifying enzymes, including histone acetyl transferases and deacetylases, and methyl transferases and demethylases [19].

To identify histone PTMs in *P. tricornutum*, we used high-accuracy MS combined with different enzyme digests with purified histone preparations [20]. Subsequent manual inspection and validation of MS data resulted in a high level of confidence in discriminating between modified sites with the same nominal masses (see Materials and methods for details). In total we identified 62 PTMs on the core and variant histones, among which eight are novel or have been identified previously in only one species (Fig. 1a). As expected, most PTMs are on the protruding N-terminal tails, although a substantial number of modified sites were also detected on the globular domains (Fig. 1a; Additional file 2). A range of lysine residues exhibited multiple modifications comprising mono-, di- and tri-methylation, acetylation and mono-ubiquitination, many of which are shared with mammals and plants. However, the positions of some modified sites were not conserved, such as the

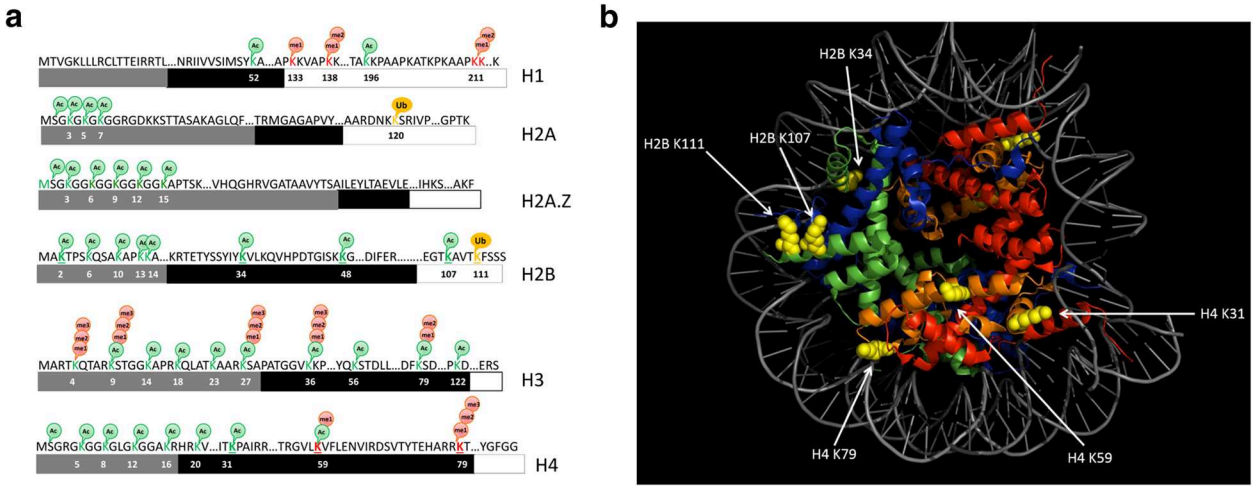


Fig. 1 Histone PTMs in *P. tricornutum*. **a** Sites of PTMs of linker, core and variant histones identified in this work. Amino acid residue number is indicated below the peptide sequence. Gray, black and white boxes indicate N-terminal, globular core and C-terminal domains, respectively. Because K211 and K212 are contiguous, their methylation state cannot be discriminated. Me2 can be located on K211 or K212 of histone H1. H3K9me2 and H3K9me3 were detected by western blot. Novel modifications are underlined. **b** Novel histone modifications are modeled on the crystal structure of the nucleosome (Protein Data Bank file 3A6N). The histone proteins are shown in ribbon diagram with histone H2A in red, H2B in orange, H3 in blue, and H4 in green. The DNA helix is shown in gray. Modified residues are visible as yellow spheres. The image was generated using the program Pymol [24]

ubiquitination of lysine 111 of histone H2B. N-terminal acetylation was observed on H2A.Z and H4, where the initial methionine was lost during protein processing and the subsequent serine residue was acetylated. We could not detect either di- or tri-methylation of lysine 9 of histone H3 despite the presence of the histone modifying enzyme of the SuVar family. However, we have shown the presence of both modifications by western blot in a previous work [21]. Of note, neither H3K9me2 nor H3K9me3 were detectable by MS in *Arabidopsis* despite their occurrence in vivo [22, 23]. Although several arginine methylases are encoded in the *P. tricornutum genome* [19], methylation of arginine was not detected, which might be due to its low abundance. Some modifications were shared only with metazoans and not plants, such as mono-, di- and tri-methylation of lysine 79 of histone H4. Besides these novel PTMs, we identified five additional unique modifications, acetylation of lysines 31 and 59 of histone H4 as well as acetylation of lysines 2, 34 and 107 of H2B (Fig. 1a, b).

Genome-wide distribution of H3K4me2, H3K9me2, H3K9me3, H3AcK9/K14 and H3K27me3

The identification of several histone PTMs by MS provides an opportunity to investigate their biological role and significance in *P. tricornutum*. We chose to focus our analyses on a few histone marks that have known functions in transcriptional activation or repression so as to draw inferences of possible regulatory roles in diverse biological processes. Genome-wide mapping of five histone marks (H3K4me2, H3K9me2, H3K9me3, H3K27me3 and H3AcK9/14) as well as nucleosome occupancy was generated by ChIP followed by deep sequencing (ChIP-Seq; see Materials and methods). Two biological replicates were sequenced for each mark and statistical tests showed a positive correlation between replicates (multiple hypothesis testing with 10 % false discovery rate, FDR p-value <0.005), thus validating the accuracy of the data. We generated high coverage maps with 4 to 38 million reads that uniquely mapped to the *P. tricornutum* genome (Additional file 3). Almost 40 % of the genome was marked by the five histone modifications. H3K4me2-, H3K9/14Ac-, H3K9me2-, H3K9me3- and H3K27me3-marked regions covered ~29 %, ~25 %, ~25 %, ~8 % and ~14 % of the genome, respectively (Additional file 4). The genome shows regions marked by at least one histone modification, and a total of 119,000 genomic regions — genes, intergenic regions, transposable elements (TEs) (1,228,620 bp) — are shared between all the marks (Fig. 2a).

We further investigated the distribution of histone modification peaks on genes, TEs and intergenic regions. Based on the number of modified domains, we found more enriched domains within genic regions than on TEs for all the marks except H3K9me3 (Additional file 5). Furthermore, a significant percentage of H3K4me2- and H3K9/14Ac-modified domains lay within intergenic regions.

A systematic analysis of the locations of H3K4me2- and H3K9/14Ac-marked regions revealed a highly significant correlation between their location and the

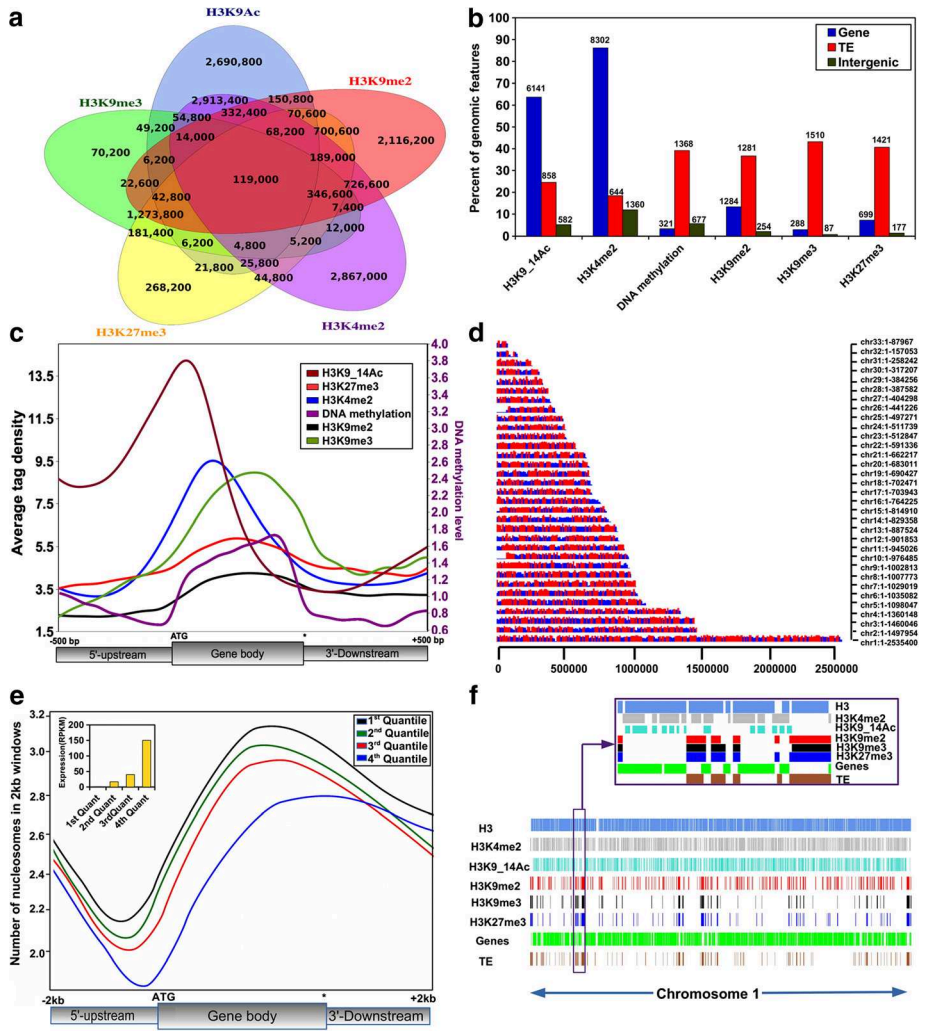


Fig. 2 Distribution of ChIP-Seq signal peaks for H3K4me2, H3K9_14Ac, H3K9me2, H3K9me3, and H3K27me3. **a** Venn diagram showing the overlap (in base pairs) of histone marks with each other. **b** Percentage of genomic features (genes, TEs, and intergenic regions) found enriched for each of the marks. The numbers above each bar refer to total number of peaks overlapping genes, TEs or intergenic regions. **c** Enrichment profile of H3K4me2, H3K9_14Ac, H3K9me2, H3K9me3, H3K27me3 along genes (upstream 500 bp, coding region, downstream 500 bp). Average tag density is the number of sequence reads per gene. Note that the small rise in enrichment seen in the flanking regions (both at 5′ and 3′ ends) is a result of the presence of nearby genes in the densely packed genome. **d** Genome-wide nucleosome distribution. The color code refers to the level of enrichment, blue being low and red high. **e** Nucleosome occupancy along genes and flanking regions and its correlation with gene expression quantiles (Reads Per Kilobase Per Million, *RPKM*). **f** A snapshot of chromosome 1, showing the six epigenetic modifications within genes and TEs

presence of annotated genes. Around 86 % and 63 % of annotated genes were found to be associated with H3K4me2 and H3K9/14Ac, respectively, while only 605 and 741 TEs, respectively, were marked (Fig. 2b). In stark contrast with these two marks, H3K9me2 and H3K9me3 were found associated principally with annotated TEs. A total of 1281 and 1510 TEs were found to be marked with H3K9me2 and H3K9me3, respectively, even though a significant number of genes were also marked by H3K9me2 (Fig. 2b). Compared with H3K9me2, H3K27me3 was even more highly enriched on TEs, which is surprising and unusual compared with other organisms for which this mark has been profiled [25, 26]. A total of 1421 out of 3493 TEs were associated with H3K27me3 while only 700 genes were marked with it (Fig. 2b).

Distribution patterns of H3K4me2, H3K9me2, H3K9me3, H3AcK9/K14 and H3K27me3 on genes

To investigate the enrichment patterns of each histone modification on genes, we examined the average enrichment of the five histone marks on predicted genes over their entire coding sequence (CDS), and within regions 500 bp upstream and downstream of the CDS. The enrichment of H3K9/14Ac and H3K4me2 peaks significantly close to the 5′ end of CDSs, with a sharper peak for H3K9/14Ac (Fig. 2c).

To assess whether the genes marked with specific histone modifications are enriched in specific functional categories, we performed a gene ontology (GO) classification. The genes marked by H3K4me2 (6047; 62.8 % of annotated genes) are enriched in the structural constituent of ribosome GO category compared with the rest of the genes in the genome while those marked by H3K27me3 (700; 7 % of annotated genes) are enriched in protein kinase activity, cAMP-dependent protein kinase, phosphotransferase, and diamine N-acetyl transferase GO categories. H3K9me2- and H3K9me3-marked genes (218; 2.3 % of annotated genes) were found to be enriched in hydrolase, ATPase, inorganic cation transmembrane transport, nucleoside tri-phosphatase activity, helicases, and structural constituent of cytoskeleton GO categories. H3K9/14ac marked genes were not enriched in any particular GO category (Additional file 6).

Genome-wide mapping of H3K4me2, H3K9me2, H3K9me3, H3AcK9/K14 and H3K27me3 on TEs

P. tricornutum TEs contain class I elements, including long terminal repeat retrotransposons (LTR-RTs; Copia), relics of non LTR-RT retrotransposon-like elements, and a few copies of class II transposons, including Piggybac, Tpase-like, and MuDR-like elements [27]. Among them, 1350 (38.7 %), 1510 (43 %) and 1421 TEs (41 %) were marked by H3K9me2, H3K9me3 and H3K27me3, respectively (Figure S5A in Additional file 7). Most of these marked TEs belong to Copia-type elements. A total of 1158, 1163 and 1281 Copia TEs were found to be marked by H3K9me2, H3K9me3 and H3K27me3, respectively, and most of them are not transcribed (Figure S5A, B in Additional file 7). As for H3K9/14Ac and H3K4me2, which have a transcription activating effect on TEs (Figure S5B in Additional file 7), only 858 and 644 TEs were marked, respectively, and most of these were found to be simple repeats (n = 657 and n = 337, respectively). Overall, a significant fraction of potentially active Copia TEs were found associated with H3K9me2, H3K9me3 and H3K27me3, which implies that these marks may regulate the transcriptional activation of TEs, especially Copia-type TEs, which appear likely to have amplified recently in the genome of *P. tricornutum* [27].

Nucleosome occupancy in the *P. tricornutum* genome

Nucleosome occupancy plays an important role in cellular processes, allowing selective access to the DNA by regulatory elements such as transcription factors [28]. To assess the relative size of nucleosomes, we performed micrococcal nuclease (MNase) digestion of isolated nuclei using increasing concentrations of MNase. Separation of the digested product in agarose gels showed a major band around 150 bp, which is a similar size to that found in plant and metazoan nucleosomes (Figure S6A in Additional file 8).

To evaluate the relative nucleosome occupancy over the *P. tricornutum* genome, we performed ChIP-Seq with an antibody against the unmodified carboxyl terminus of histone H3. Close to 60 % of the genome was found to be occupied by nucleosomes, with densely packed segments interspersed by nucleosome-depleted regions (Fig. 2d). Most of the nucleosomes fall within exons and a significant number cover TEs and intergenic regions (Figure S6B in Additional file 8). We further examined nucleosome distribution over CDSs, upstream of the transcription start site and downstream of the stop codon, and assessed how this correlated with the expression state of the genes. Our data show that nucleosome depletion occurs around 150 bp upstream of the transcription start site for genes that have high expression quantiles while nucleosome density increases over gene bodies and drops towards the 3′ end for all genes, regardless of their expression quantile, which is consistent with what has been reported in other species (Fig. 2e).

Previous work in other organisms identified nucleosome positions based on DNA sequence motifs [29]. We therefore tested whether DNA sequence-guided nucleosome positioning is of relevance in *P. tricornutum*. We found that GC and CG are dinucleotide sequences where nucleosomes are preferentially positioned whereas AA, TA and TT can be considered nucleosome-excluding sequences and rather tend to peak outside the nucleosomes (Figure S6C in Additional file 8), as observed in other species [29].

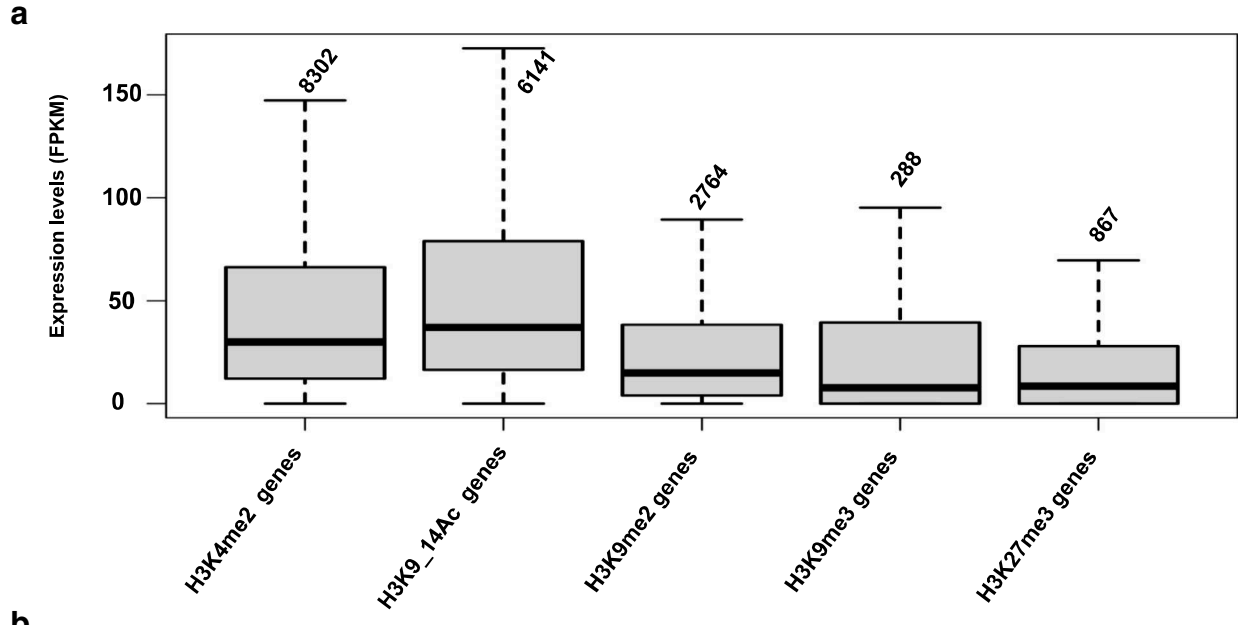
Correlation of chromatin marks with gene expression

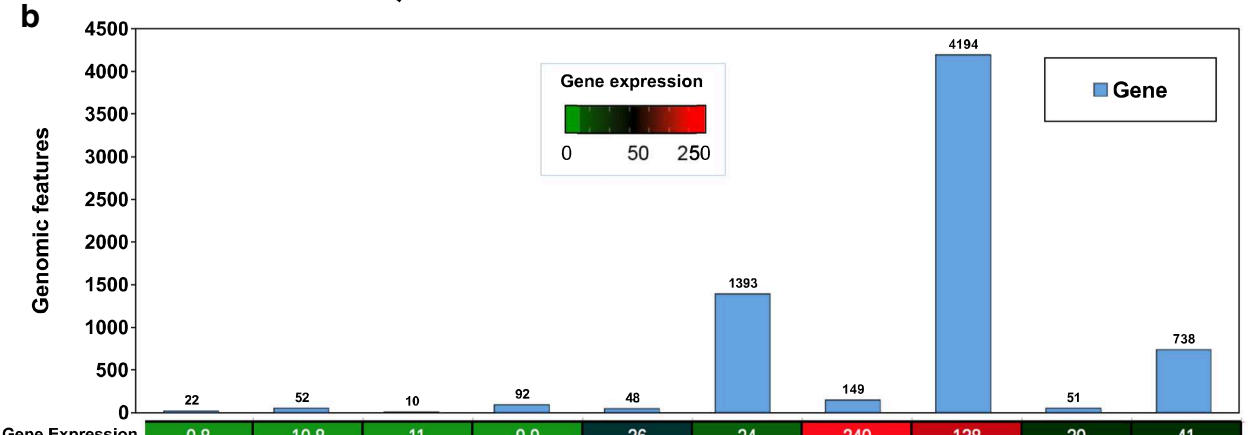
A representative genomic region of chromosome 1 is shown in Fig. 2f to demonstrate the general distribution of the five histone marks along with histone H3 on genes and TEs. It shows that, in general, genes are co-marked by H3K4me2 and acetylation while TEs tend to be marked by H3K9me2, H3K9me3 as well as H3K27me3.

To explore the relationship between gene expression and each of the five histone marks, the mRNA levels of genes were assessed genome-wide using RNA-Seq in the same growth conditions that were used for the cells used for chromatin analyses. The genes marked by both H3K4me2 and H3K9/14Ac showed the highest average expression levels, and the latter showed the largest variation in expression (Fig. 3a). By contrast, both H3K9me3- and H3K27me3-marked genes displayed the lowest gene expression levels, indicating their association with repressed genes in *P. tricornutum*, consistent with previous studies in different organisms [30–32]. H3K9me2 also displayed a moderate repressive effect on genes. Taken together, these observations suggest that H3K4me2 and H3K9/14Ac represent general marks for expressed genes, whereas H3K9me3, H3K27me3 and H3K9me2 appear to associate with repressed genes.

Combination of two or more histone modifications is known to have an impact on gene regulation beyond those of individual marks [33, 34]. We therefore examined whether particular combinations of histone PTMs may influence transcriptional regulation of genes in *P. tricornutum*. A large number of genes marked with both H3K4me2 and H3K9/14Ac were significantly correlated with high levels of expression, further supporting the role of these two marks in the activation of gene expression, while co-occurrence of H3K27me3 with either H3K9me2 or H3K9me3 correlated with a low level of gene expression, indicating the repressed state of these co-marked genes (Fig. 3b).

We further correlated DNA methylation from previously published work [14] with histone PTMs and examined gene expression patterns of marked genes. As noted above, genes co-marked with H3K4me2 and H3K9/14Ac were upregulated, and this was largely





Gene Expression	0.8	10.8	11	9.9	36	24	240	138	39	41
H3K4me2	-	-	-	-	+	+	+	+	-	-
H3K9_14Ac	-	-	+	+	-	-	+	+	-	-
H3K9me2	+	+	-	-	-	-	-	-	-	-
H3K9me3	+	+	+	+	-	-	-	-	-	-
H3K27me3	+	+	+	+	-	-	-	-	-	-
DNA met	+	-	+	-	+	-	+	-	+	-

Fig. 3 Comparison of the presence of histone marks with gene expression. **a** Boxplots showing the gene expression profiles correlating with each histone mark. Gene expression was quantified in standard growth conditions using RNA-Seq data. Around 85 % of genes are expressed and quantified as fragments per kilobase of exons per million reads mapped (*FPKM*). Number of genes is indicated above each bar. **b** Combinatorial effects of histone modifications and DNA methylation on gene expression. FPKM values are indicated below each bar

unaffected by the presence or absence of DNA methylation (Fig. 3b). DNA methylation had no major effect on expression patterns, except on 48 genes labeled with H3K4me2, which were significantly highly expressed compared with other H3K4me2 genes that were not methylated (Fig. 3b). We have already shown that DNA methylation has no significant effect on expression of genes except when they are extensively methylated [14]. Furthermore, the genes co-marked by H3K4me2 and DNA methylation might have additional activating histone marks that we did not investigate in this study, such as H3K4me3, H3K36me3 and H3AcK27, which could explain the significant increase in gene expression.

Combination of the five histone marks with DNA methylation defined three main chromatin states (CSs): CS1, which is activating and correlates with the presence on genes of H3 acetylation and H3K4me2; CS2, which is repressive and is defined predominantly by H3K27me3, H3K9me3 and H3K9me2; and CS3, which combines activating and repressive marks with an intermediate expression level of genes. It should also be noted that a significant number of genes are not marked and may contain histone marks that were not investigated in this study (Fig. 3b).

H3K27me3 is characterized by a broad distribution pattern over several kilobases in animals, while plants such as *Arabidopsis* display shorter H3K27me3-marked domains restricted mainly to transcribed regions [32, 36]. Although a photosynthetic organism, *P. tricornutum* shows an animal-like distribution pattern of H3K27me3, perhaps suggesting similar mechanisms of deposition and transcriptional regulation of genes. Polycomb repressive complex 2, containing four proteins, methylates H3K27me via the SET domain of its subunit enhancer of zeste [37]. The PRC2 complex is widely distributed among plants, metazoans and algae but appears to be absent from the yeast species *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* [38]. Considering the presence of different clusters within the wide H3K27me3 domains reported in different species [26, 31, 37] and the lack of knowledge about the distribution pattern of this mark in single-celled organisms, we assessed in more detail the pattern of H3K27me3 enrichment over genes. Interestingly, we could distinguish six different profiles. The first cluster of genes shows a distinct enrichment over the region 500 bp downstream of the stop codon (C1), while the other clusters target the gene body and 500 bp downstream (C2), the gene body only (C3), the entire gene length (C4), only the region 500 bp upstream of the TSS (C5), and 500 bp upstream and the gene body (C6) (Fig. 4a). When correlated with expression data, only four clusters (C2, C3, C4 and C5) correlate clearly with repressed genes while clusters C1 and C6 show positive correlations with gene expression compared with unmarked genes, suggesting a different or a diversified role of H3K27me3 in *P. tricornutum* which is known to be repressive (Fig. 4b). We performed a homology estimation analysis (see Materials and methods) and found out that out of 700 H3K27me3-marked genes, 39 % have no homologs with known function and nearly 20 % are found only in *P. tricornutum* (Additional file 9). Of note, none of the other investigated marks showed such a clustering pattern.

Changes in chromatin marks in response to nutrient-limiting conditions

To gain insights into the dynamic nature of the *P. tricornutum* epigenome in response to an environmental cue, we analyzed the impact of nitrate depletion. Nitrate is an important nutrient in marine ecosystems and its appearance in surface waters, e.g., following upwelling events, is often associated with diatom proliferation [13]. We specifically examined three histone modifications (H3K4me2, H3K9/14Ac and H3K9me3) using Chip-seq, as well as DNA methylation by bisulfite deep sequencing. We also assessed gene expression changes by RNA-Seq.

In parallel with the reduced growth rate and chlorotic phenotype observed during nitrate limitation (Additional file 10), the number of genes that lost or gained histone marks and/or DNA methylation was noteworthy, in particular H3K9/14 acetylation and H3K4me2 (Fig. 5a). These changes were more prominent on genes than on TEs, except for H3K9me3 and DNA methylation, which showed an opposite profile, indicating that TEs are probably tightly regulated by these two marks, which show repressive effects in response to stress (Fig. 5b). Almost 20 % of H3K4me2-marked genes lost this mark under nitrate depletion while ~16 % of H3K9/14Ac-free genes gained this mark. The loss of both H3K9me3 and DNA methylation was even more significant (31 % and 35 %, respectively). As expected, the chromatin profiles of most genes and TEs remained the same, suggesting that only certain sets of genes and TEs were affected by nitrate limitation. Very few intergenic regions were differentially marked between both conditions, suggesting they have a minor role in gene regulation in response to nitrate limitation (data not shown).

To determine whether these patterns of histone modifications were correlated with changes in transcriptional regulation, we used a global quantification approach to examine the link between differentially marked genes and the fold change in expression under nitrate replete and limiting conditions. The overall effect of differential marking by both acetylation and H3K4me2 had a positive effect on gene expression, while the differential marking of H3K9me3 showed a rather repressive effect on gene expression (Fig. 5c, d). Many genes gained acetylation and/or H3K4me2 under low nitrate and

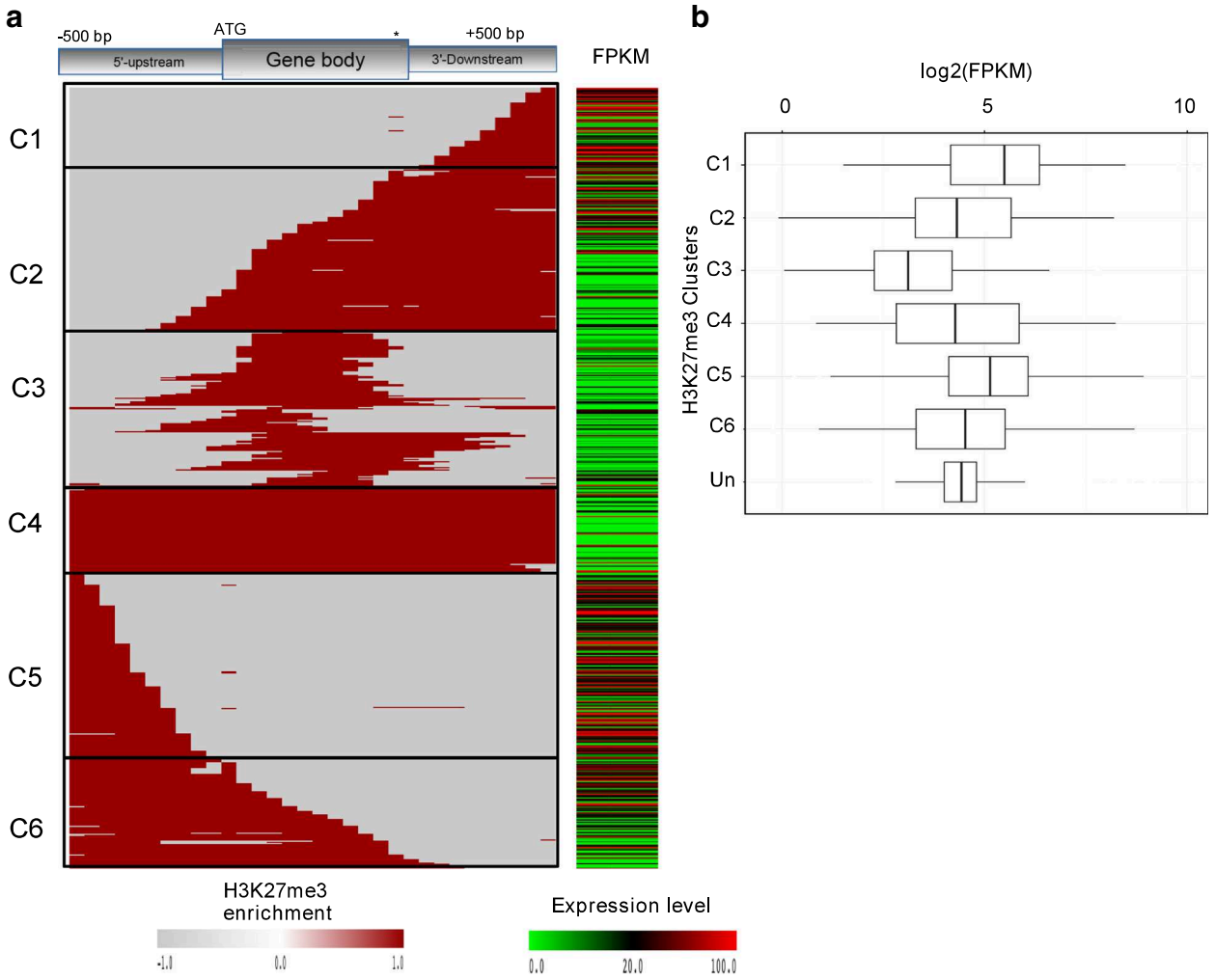


Fig. 4 Clustering analysis of H3K27me3 over genes and correlation with gene expression. **a** H3K27me3 enrichment is shown in red while absence is represented in gray. Gene expression levels are shown as a heat map with red reflecting high (expression level FPKM [fragments per kilobase of exons per million reads mapped] greater than 20) and green representing low expression (expression level FPKM between 0 and 20). **b** Expression boxplots for each H3K27me3 enrichment cluster. *Un* refers to genes unmarked by H3K27me3

therefore became upregulated, as did those that lost H3K9me3. This analysis pinpointed genes involved in nitrate metabolism, e.g., ferredoxin-dependent nitrite reductase (Pt12902), and nitrite (Pt13076) and nitrate (Pt26029) transporters, which were all acetylated under nitrate starvation, which correlated with their transcriptional upregulation (Fig. 6; Additional file 11).

To gain further insights into the functional categories of genes that were differentially marked and regulated under low nitrate, we performed a GO classification as well as a Mapman analysis and found that, as expected, there is an enrichment in genes encoding proteins involved in nitrate metabolic pathways (nitrate transport, reduction and assimilation) as well as genes involved in lipid transport and metabolism, and stress response in conditions of nitrate limitation (Additional files 11, 12, 13, 14, and 15). Other genes involved in different pathways related to nitrate availability were also found to be marked and expressed differentially. For example, genes encoding phytoene desaturase-like3 (Pt15806), known to be involved in carotenoid biosynthesis, coproporphyrinogen III oxidase (Pt10640), involved in heme and chlorophyll synthesis, as well as several proteins involved in light harvesting (Pt14442, Pt25168, Pt22956, Pt22395, Pt47697) gain H3K4me2 under low nitrate, while genes encoding an ATP binding protein (Pt46431) or encoding secondary metabolite biosynthesis components (Pt16295) gained H3K9me3. Stress response genes were also found to be differentially regulated. Most of the genes that gain or lose DNA methylation encode proteins involved in catalytic activities and metabolism. A few other genes indirectly related to nitrate depletion showed differential regulation

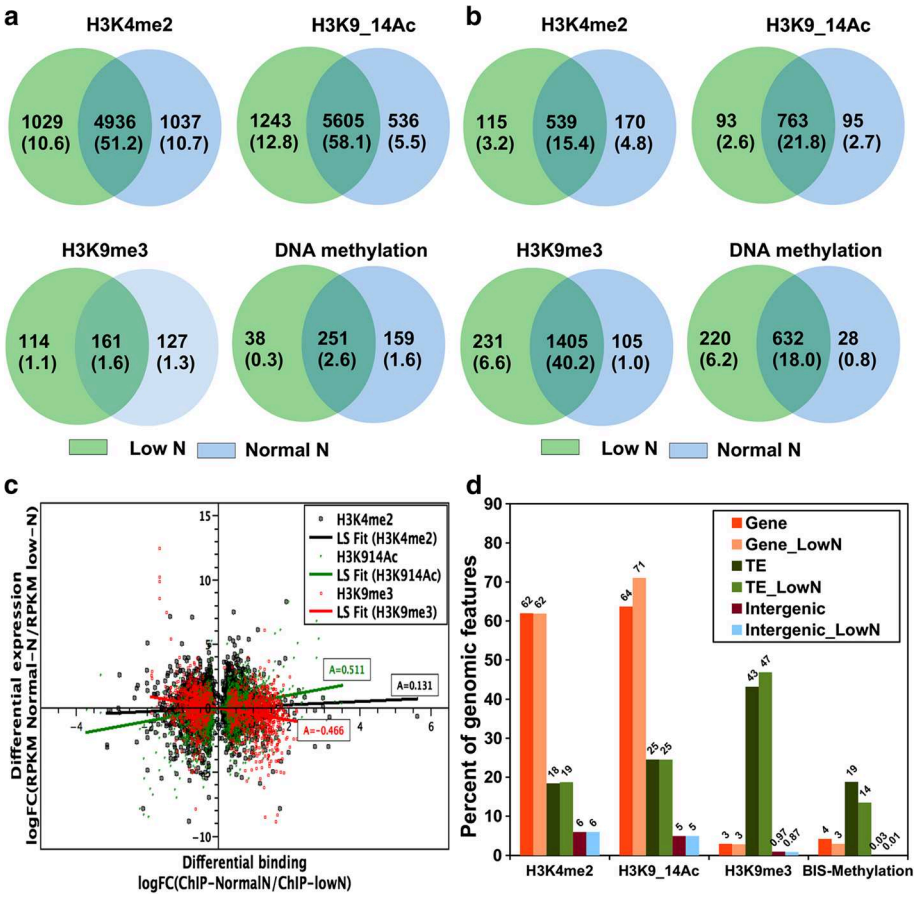


Fig. 5 Dynamic changes in chromatin marks in nitrate replete and limiting conditions. **a** Venn diagrams showing overlap of different histone marks on genes in nitrate replete (*Normal N*) and nitrate-limiting (*Low N*) conditions. **b** Venn diagrams showing overlap of different histone marks on TEs in nitrate replete and nitrate-limiting conditions. Percentages of differentially marked genes and TEs are indicated between parentheses. **c** Comparison of levels of ChIP binding and expression levels. Fold change of ChIP binding (differential binding) is calculated from the number of reads in the peak region using diffReps. Fold change in expression level is calculated from the levels of RPKM using the Cuffdiff [78]. Pearson correlation values between differentially marked and expressed genes under the two conditions (normal and low nitrate growth conditions) are marked along the trend line. **d** Number of genome features (genes and transposable elements) marked by each of the three histone marks, as well as DNA methylation. Percentage was derived using the known genome feature annotations. Number of genes, TEs and intergenic regions under normal or low nitrate are indicated above each bar

in response to nitrate limitation, including Pt15815, which encodes an ortholog of a pyrophosphatase-energized proton pump (involved in auxin transport in plants), whose enhanced activity was shown to improve nitrogen uptake in roman lettuce [39], and Pt51183, which encodes a CREG1 ortholog, involved in cellular repression of transcription. Interestingly, many genes of unknown function lost one of the marks, in particular acetylation under low nitrate conditions. Among these genes, many were found to be diatom specific — 30 % of H3K9me3-, 36 % of H3AcK9/14- and 40 % of H3K4me2-marked genes; these represent 9 %, 9.6 % and 3.6 %, respectively, of the *Phaeodactylum* genome, suggesting particular pathways have been recruited by diatoms to survive nitrate depletion in the oceans.

Discussion

We report here a comprehensive analysis of histone PTMs in the model diatom *P. tricornutum* using MS and ChIP-Seq. MS analysis revealed a large conservation of histone modifications but also new ones, thus expanding the list of histone PTMs in eukaryotes. Most of the histone modifications showed similarities to those of plants and mammals, including acetylation of several lysines on the N-terminal tails of histones H2A, H2B, H3 and H4 and mono-, di- and tri-methylation of lysines 4, 9, 27 and 36 of histone H3, suggesting their role in transcriptional regulation of many biological processes.

MS analysis revealed eight less-characterized modifications, namely acetylation of lysine 59 of histone H4, which was instead reported to be methylated in bovine

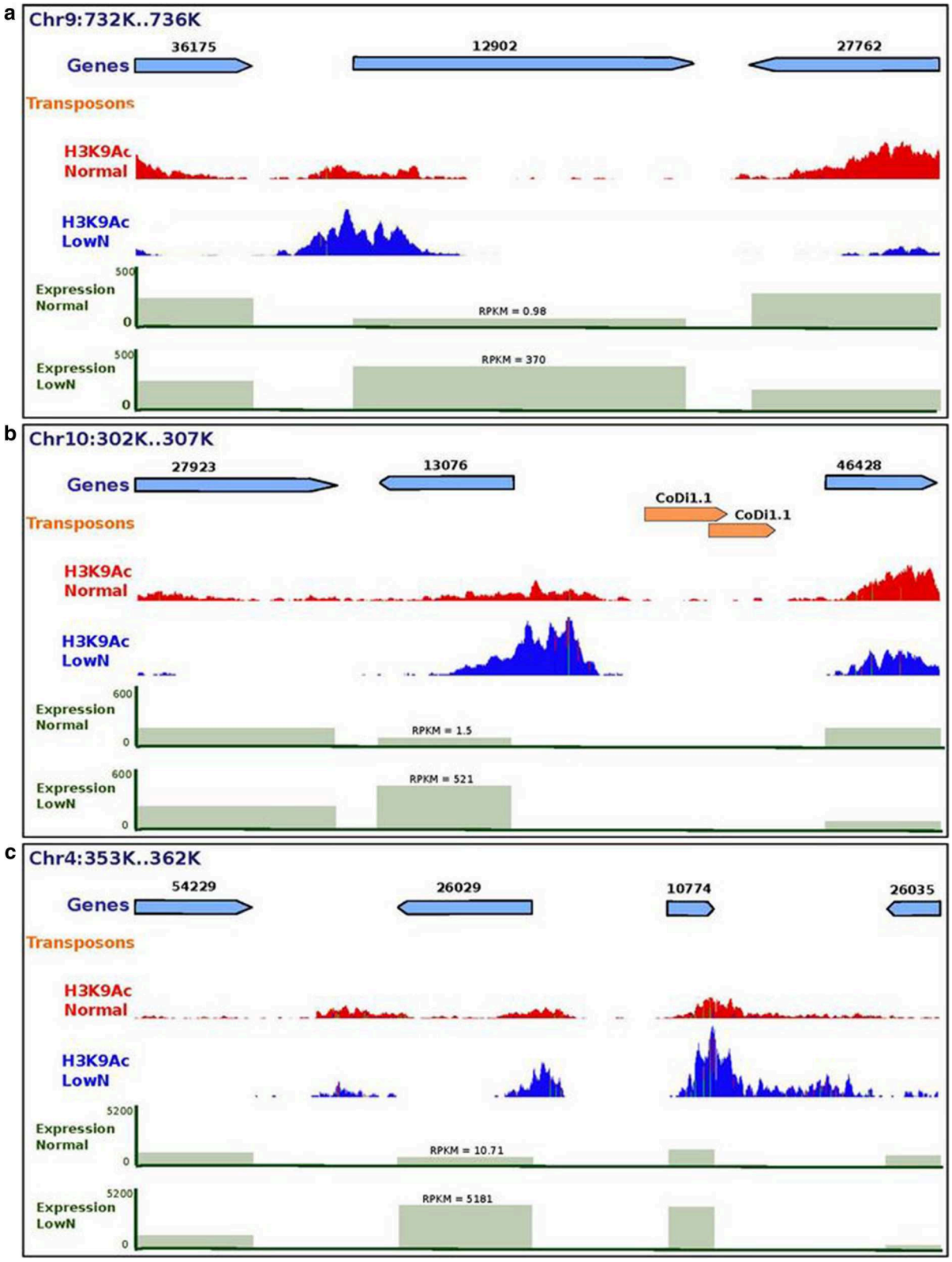


Fig. 6 Snapshots of genes differentially marked and regulated under nitrate depletion. Ferredoxin nitrite reductase (12902) (**a**), a nitrite transporter (13076) (**b**) and a nitrate transporter (26029) (**c**) as well as chloroplast ribosomal proteins are shown with higher levels of acetylation and expression under low nitrate conditions

calf thymus histones [40], and acetylation of lysine 31 of histone H4 (reported only in *Toxoplasma* [41], although not confirmed in an independent study [42]). In addition, we also detected acetylation of lysines 2, 34 and 107 of H2B, which are reported for the first time in this study, as well as the ubiquitination of lysine 111 of the same histone. Due to their accessibility, histone tail modifications have been widely studied and have been shown to act primarily through altering the ability of non-histone proteins to interact with chromatin. On the other hand, less-studied histone modifications in the core domain, such as the novel ones identified in this work, are likely to exert their function through mechanisms that are distinct from those reported for histone tail modifications. H2BAcK34, H4AcK31, H4AcK59, H4K79me1, H4K79me2 and H4K79me3 are located on the lateral surface of the nucleosome (Fig. 1b), suggesting a primary function in the regulation of histone-DNA interactions. To explain how these modifications might alter chromatin structure, a model has been proposed whereby a chromatin remodeling activity acts on the nucleosomes to alter histone-DNA interactions, thereby exposing sites on the lateral surface which in turn become modified and alter the mobility of nucleosomes. This altered mobility can either lead to changes in the accessibility of specific DNA sequences or changes in higher order chromatin structure [39, 42].

Interestingly, *P. tricornutum* combines histone PTMs found in both mammals and plants, such as acetylation and mono- and di-methylation of lysine 79 of histone H3 found only in human and yeast [44] but not in *Arabidopsis* [23], underlying the mosaic nature of the *P. tricornutum* genome. Another interesting example is the acetylation of lysine 20 of histone H4, which is shared with *Arabidopsis* but different from human where this residue is only methylated [23]. H4K20me, which is known to be a repressive mark, was not detected by either MS or western blotting using an antibody that recognizes this modification in *Arabidopsis* (data not shown). Furthermore, mono- and di-methylation of lysine 79 of histone H4 are modifications that *P. tricornutum* shares only with *Toxoplasma gondii*, which is an obligate intracellular parasitic protozoan belonging to the Alveolata, a superphylum closely related to Stramenopiles. Novel histone core domain modifications identified in *P. tricornutum* are probably ancient modifications that likely were lost from the divergent lineages (or have not yet been detected). It will be interesting to investigate the presence of these novel PTMs in closely related species to trace back their evolutionary history more precisely.

We further generated an integrated epigenomic map of five histone marks known to be activating or repressive using ChIP-Seq. Combined with previously published genome-wide DNA methylation data [14], comprehensive and combinatorial analyses revealed some conserved and specific epigenetic features in *P. tricornutum*, thereby extending the existence of the histone code to Stramenopiles. As expected, both acetylation of lysines 9 and 14 and di-methylation of lysine 4 of histone H3 map predominantly to genes, followed by intergenic regions and TEs. This is in contrast to H3K9me3, H3K9me2 and H3K27me3, which we found mainly within TEs. As in yeast, mammals and plants, H3K4me2 in *P. tricornutum* does not appear to index genes in relation to their expression level and may not be directly implicated in transcriptional activation [45, 46]. By contrast, acetylation correlates with transcription activation and is enriched in gene promoters, which is in line with genome-wide studies in yeast, human and *Arabidopsis* [47–49]. H3K9me2 marking was mainly found on TEs and a substantial number of transcriptionally repressed genes, consistent with what has been observed in plants and animals in which this mark has been profiled [33, 49–50]. The H3K9me3 mark mapped mainly on TEs and is repressive, which is similar to mammals but different from *Arabidopsis*, where it is exclusively found on euchromatic regions where it has a positive effect on gene regulation [32].

H3K27me3 covered 3.84 Mb (14 %) of the 27.4 Mb *P. tricornutum* genome, which is more than *Neurospora*, *Arabidopsis*, *Drosophila* and mammals, where it covers around 6 % of each genome [25, 32]. This high percentage compared with other species can be explained by the large coverage of H3K27me3 over TEs and is in line with this mark being an ancient histone modification with a primary role in TE silencing, as previously suggested [25]. H3K27me3 is known to be repressive and to mark mainly genes in *Drosophila*, mammals, *Arabidopsis* and *Neurospora* [25, 32]. However, its distribution is different and unusual in *P. tricornutum*, being predominantly on TEs, and it has a repressive effect, implying that the functions and mechanisms of H3K27me3 in single-celled eukaryotes may be different from in their multicellular counterparts.

The H3K27me3 mark is established by the Polycomb repressive complex PRC2 and its absence in the model unicellular fungi *S. pombe* and *S. cerevisiae* initially suggested that it arose to regulate developmental processes in multicellular organisms [53]. This hypothesis has recently been questioned because PRC2 has been found in several single-celled species [37]. Our results showing genome-wide mapping of H3K27me3 in a unicellular organism confirm its early evolution prior to the last common ancestor of animals and plants.

Unlike in *Arabidopsis*, where H3K27me3 marks short regions, typically <1 kb, which tends to be restricted to the coding regions of single genes, H3K27me3-modified regions show blanket-type coverage over large domains

in *P. tricornutum* (≥2 kb), which resembles the enriched profiles of H3K27me3 in animals [35, 53]. Our observation of H3K27me3 enrichment over promoter regions and its correlation with highly transcribed genes is also surprising and contrasts with what has been previously reported [37, 55]. The blanket-like coverage of H3K27me3 in animals has been overlooked and such correlations might have been missed from *Drosophila*, human and mouse cells because the wide enrichment of H3K27me3 does not appear to have been analyzed in detail. However, a more detailed study reported recently in mouse embryonic stem cells revealed a similar profile, where H3K27me3 mapping on promoters correlated with high expression, suggesting that these regions might serve as bivalent domains harboring additional activation marks such as H3K4me3 and H3K36me3 [30]. The observed enrichment over the entire gene, gene body alone, or together with either 500 bp upstream or downstream regions represents the majority of H3K27me3-marked genes in *P. tricornutum* and correlates with low expression, thus corresponding to the canonical view of H3K27me3 as being inhibitory to transcription [25, 31, 35]. This suggests that a repressive role of H3K27me3 in *P. tricornutum* might be mediated by gene body marking that occurs in all four clusters identified in Fig. 4, and may compromise transcription elongation. This does not exclude repression by transcription initiation for the upstream marked regions. A novel and intriguing pattern is the presence of H3K27me3 downstream of gene bodies that correlates with activation, suggesting that H3K27me3 does not interfere with transcription termination and that other unknown additional factors allow the transcription of these genes to take place. Overall, H3K27me3-marked genes belong to many functional categories. However, there is a tendency for H3K27me3 to mark genes that have no known function or to be poorly conserved, among which a large proportion have no orthologs. For the rest of the genes that are functionally annotated, a significant number encode 'developmental' genes, as seen in mammals and *Arabidopsis*. Cluster-wise, most of the genes marked at their promoter by H3K27me3 are co-marked by acetylation, which might explain their transcriptional activity, while the others, in particular those marked over their entire length, tend to encode 'developmental' genes as well as defense response genes.

Our work has also shown the importance of chromatin level regulation in diatoms in response to nitrate starvation, as the changes in the examined histone marks had a considerable impact on gene expression. Epigenetic profiling of nitrate-starved cells revealed a set of genes involved in nitrate assimilation, transport and metabolism which either gain activating marks or lose repressive marks and become upregulated. As expected, many diatom-specific genes of unknown function show up in this analysis, suggesting a key role in surviving nitrate starvation. These uncharacterized genes might help diatoms to cope with a scarcity of nitrate until better conditions become available and allow them to bloom and out-compete other plankton. Functional characterization of these genes will shed light on the pathways that diatoms recruit to survive nutrient depletion and will ultimately contribute to better understanding of diatom ecological success in contemporary oceans.

In line with previous studies in *Drosophila* and *Arabidopsis*, where different chromatin states have been identified, the combinatorial analysis of histone marks with DNA methylation allowed us to define three chromatin states — active, repressive or intermediate — supporting the existence of an epigenetic code in addition to the histone code in *P. tricornutum*. Mapping of additional marks will undoubtedly refine this analysis and provide new insights into the role of chromatin modifications in marine diatoms.

Conclusions

To gain insights into the evolution of chromatin-mediated regulation of genes, we used an integrative approach combining MS, ChIP and RNA-Seq to analyze post-translational modifications of histones in a stramenopile, the model diatom *P. tricornutum*, which is phylogenetically distant from well-known model organisms from other lineages of life such as plants and animals. MS analysis revealed the strong conservation of histone modifications across distantly related species but also new ones, thus expanding the list of histone PTMs in eukaryotes. Remarkably, *Phaeodactylum* combines histone PTMs found in plants and/or mammals, underscoring the chimeric nature of its genome and suggesting a different evolution of histone PTMs in plants and animals. Genome-wide mapping of some key PTMs revealed shared features with plants and animals, such as the distribution of acetylation, and di-methylation of lysine 4 of histone H3, which map mainly on genes and have an activating effect. Our work shows also some divergence from green lineages exemplified by the H3K9me3 profile, which is found exclusively on genes and is activating in *Arabidopsis* while it is distributed mainly on TEs and is repressive in *P. tricornutum* and animals. Interestingly, the pioneering genome-wide mapping of H3K27me3 has revealed an unorthodox distribution as it maps mainly on TEs and has a repressive effect, while this mark is known to repress mostly genes in euchromatic regions in *Arabidopsis*. The H3K27me3 profile in *P. tricornutum* suggests this mark has an evolutionarily ancient function in transcriptional repression of TEs. The presence of H3K27me3 in *P. tricornutum* and several other algae suggests an ancient origin of Polycomb repressive complex proteins and raises the question of its role in single-celled species. Combinatorial analysis of histone

PTMs revealed different chromatin states and gene expression patterns, extending the histone code to Stramenopiles. Investigation of histone modifications under nitrate-limiting conditions revealed the dynamic role of chromatin modifications in regulating some key target genes, indicating their importance for adaptation of diatoms to changing environments.

Materials and methods

Materials and growth conditions

Phaeodactylum tricornutum Bohlin Clone Pt1 8.6 (CCMP2561) cells were grown as described previously [54]. Under low nitrate, cells were grown as described in [14].

Extraction of histones

Histones from *P. tricornutum* were extracted as described previously [20].

MNase digest assay

MNase digest was performed as described previously [54] with a few modifications. Nuclei were washed three times with MNase digestion buffer. The nuclei suspension was aliquoted into 100 μl to which 0.5, 12 and 16 units of MNase were added. After 1 h of incubation with the stop buffer, 1 μl of RNAse was added to each sample and further incubated as described previously [55].

MS assay

Protein in-gel digestion using multiple proteases

Comprehensive localization of PTMs on histones requires observation of each amino acid. Efforts to increase histone coverage have been achieved by use of a multiple protease strategy and chemical derivatization. Enzymatic digestion with trypsin results in small peptides that are difficult to retain on nano-high-performance liquid chromatography (HPLC) columns for analysis by MS. As an alternative, lysine amino groups can first be chemically modified by reaction with propionic anhydride to further generate propionylated residues that would be resistant to trypsin proteolysis. Under these conditions, reproducible and MS-compatible Arg-C-type peptides can be obtained [56].

Proteins were separated by 14 % SDS-PAGE gels and stained with colloidal Coomassie blue (LabSafe Gel Blue™, AGRO-BIO) reagent, which does not contain methanol or acetic acid. Histone bands were excised and washed and proteins were reduced with 10 mM dithiothreitol prior to alkylation with 55 mM iodoacetamide or chloroacetamide for ubiquitylation studies. After washing and shrinking of the gel pieces with 100 % acetonitrile, propionylation or in-gel digestion was performed. All digestions were performed overnight in 25 mM ammonium bicarbonate at 30 °C, by adding 10–20 μl endoproteinase (12.5 ng/μl) trypsin (Promega) or 12.5 ng/μl chymotrypsin (Promega) or 12.5 ng/μl ArgC (Promega) or 20 ng/μl elastase (Sigma-Aldrich). The shrunken gel bands were chemically derivatized by treatment with propionic anhydride before and after trypsin digestion. Briefly, this reaction mixture was created using 3/4 propionyl anhydride (Sigma-Aldrich) and 1/4 methanol. Propionylation reagent (20 μl) and 100 μl of 25 mM ammonium bicarbonate were added to each band, adjusted to pH 8.0, and allowed to react at 51 °C for 20 minutes and reduced to dryness using a SpeedVac concentrator for removal of reaction remnants before trypsin digestion. A second round of propionylation was performed to propionylate the newly created peptide N-termini. Ultrasound-assisted extraction was used to extract peptides with 60 % acetonitrile/5 % formic acid extraction solution. The extract was dried in a vacuum concentrator at room temperature and re-dissolved in solvent A (2 % acetonitrile, 0.1 % formic acid). Peptides were then subjected to MS analysis.

MS and data analysis

Samples were analyzed by nano-HPLC/MS/MS using an Ultimate3000 system (Dionex S.A.) coupled to an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Samples were loaded on a C18 pre-column (300 μm inner diameter × 5 mm; Dionex) at 20 μl/minute in 2 % acetonitrile, 0.1 % trifluoroacetic acid. After 3 minutes of desalting, the pre-column was switched on line with the analytical C18 column (75 μm inner diameter × 50 cm; C18 PepMap™, Dionex) equilibrated in 100 % solvent A. Bound peptides were eluted using a 0 to 30 % gradient of solvent B (80 % acetonitrile, 0.085 % formic acid) during 157 minutes, then a 30 to 50 % gradient of solvent B during 20 minutes at a 150 nl/minute flow rate (40 °C). Data-dependent acquisition was performed on the LTQ-Orbitrap mass spectrometer in the positive ion mode. Survey MS scans were acquired on the Orbitrap in the 400–1200 m/z range with resolution set to a value of 100,000. Each scan was recalibrated in real time by co-injecting an internal standard from ambient air into the C-trap ('lock mass option'). The five most intense ions per survey scan were selected for collision-induced dissociation fragmentation and the resulting fragments were analyzed in the linear trap (LTQ). Target ions already selected for MS/MS were dynamically excluded for 20 s.

Data were acquired using the Xcalibur software (version 2.0.7) and the resulting spectra were then analyzed via the Mascot™ Software created with Proteome Discoverer (version 1.4, Thermo Scientific) using an in-house database containing the sequences of histone proteins from *P. tricornutum* (PtH3_50695, PtH3_21239, PtH4_26896, PtH2A_34798, PtH2A_28445, PtH2B_11823, PtH1_54381)

or the UniProtKB Phaeodactylum tricornutum database (15,832 proteins) with a Mascot score of 1 % FDR (or <5 %; shown in bold in Additional file 16). Carbamidomethylation of cysteine, oxidation of methionine, acetylation of lysine and protein N-termini, methylation, dimethylation of lysine, arginine and trimethylation of lysine, methylation of aspartic and glutamic acid, di-glycine of lysine, propionylation of lysine and N-termini of peptides, phosphorylated histidine, serine, threonine and tyrosine were set as variable modifications for Mascot searches. The mass tolerances in MS and MS/MS were set to 5 ppm and 0.5 Da, respectively. The resulting Mascot files were further processed using myProMS [57]. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium [58] via the PRIDE partner repository [59] with the dataset identifier PXD002148.

Isolation and immunoprecipitation of chromatin

Chromatin isolation and immunoprecipitation were performed as described previously [21]. The following antibodies were used for immunoprecipitation: H4K9/14Ac (005–044) from Diagenode; H3K4m2 (32356) and H3K9me3 (8898) from Abcam; H3 (07–690), H3K4me2 (07–030), H3K9me2 (17–681) and H3K27me3 (07–449) from Millipore. Peptide competition assays were performed for H3K4me2, H3K9me2 and H3K27me3 as described previously [21].

Chlorophyll analysis

Cells (1 ml) were harvested by centrifugation at 1400 *g* for 10 minutes, resuspended in 100 % ethanol, and incubated for 30 minutes in the dark. The crude extract was cleared by 1-minute centrifugation at 12,000 *g* and the supernatant was used for chlorophyll quantification according to [60] with a spectrophotometer Biossacte 3 from Thermo Spectronic reading at 629 nm and 665 nm wavelengths. The calculated chlorophyll contents were normalized to 10^6 cells.

Oxygen evolution

Photosynthesis was measured as O_2 exchange rates using a Clark-type oxygen electrode at 25 °C (Oxy-Lab, Hansatech Instruments, King's Lynn, UK). The actinic light was provided by light-emitting diodes with an emission maximum around 650 nm. For each measurement cells were concentrated by 10-minute centrifugation at 1400 *g* and resuspended in Artificial Sea Water (ASW) to a final concentration of 10^7 to 3×10^7 cells/ml. Net O_2 evolution V_{max} was measured at 800 μE and is presented as nmol O_2 evolved per minute per 10^6 cells.

Data analysis

For mapping and analysis we used *P. tricornutum* genome v.2.0 available at the Joint Genome Institute [61]. Reads obtained were quality controlled with a standardized procedure using FASTQC [62]. Trimmomatic [63] was used for quality trimming. GO-based functional analysis on ChIP-marked genes and methylated genes were performed using BLAST2GO [64] with a significant FDR cutoff of 0.05 % probability level. R [65] and Biopython [66] were extensively used for data analysis. For pattern-based analysis on genes and flanking regions, genes were normalized to equal size, and flanking 2-kb regions were selected as the average intergenic size of ~1500 bp. Data processing, analysis, and plotting were performed using Python, R/Bioconductor and Hyperbrowser [67]. Results of the analysis have been made available on the Gbrowse-based genome browser at [68].

Computational analysis of histone modifications in *P. tricornutum* by ChIP-Seq

Single-end sequencing of the five ChIP samples was performed using an Illumina GAIIx with a read length between 36 and 50 bp. This yielded an average of approximately 37 million reads each (Additional file 3). Data for all ChIP samples and input were of good quality with mean quality scores of 30, with 50 % mean GC content. The reads were mapped onto the *P. tricornutum* genome v.2.0 using Bowtie [69] with mismatch permission of 2 bp. Unique mapping of reads was adopted. To identify regions that were significantly enriched, we used MACS [70] and SICER [71] with parameters of W:200 (window length), G:200 (gap size) for H3K4me2 and H3K9_14Ac; W:200, G:600 for H3K27me3, H3K9me2 and H3K9me3; and a FDR <1E-2. Enriched regions were detected against the islands of background control with the same parameter.

MACS and SICER both generated peaks with similar peak ranges and comparable overlapping genomic regions. But for H3K9me2 and H3K27me3, MACS showed much fewer peaks and in these two cases SICER showed significant diffused peaks which overlapped on repeat regions. Visualization and analysis of genome-wide enrichment profiles were done with IGV [72]. Peak annotations such as proximity to genes and overlap on genomic features such as transposons and genes were performed using Peak Analyzer [73]. The high-throughput Chip sequencing data have been deposited in NCBI's Gene Expression Omnibus under accession number GSE68513.

H3K27me3 clustering

A 700 × 30 matrix was created based on position of the H3K27me3 mark along the 500 bp upstream, gene body, and 500 bp downstream regions. Hierarchical clustering was done on this matrix using the complete-linkage method [74]. This resulted in six clusters and each cluster was then correlated to expression level. Both

expression levels and ChIP seq peaks were plotted as heat maps with red reflecting high expression (expression level RPKM greater than 20), and green low expression (expression level RPKM between 0 and 20).

Homology estimation analysis

We downloaded 57,224,756 protein sequences from 344,297 species from UniProt [75] and MMETSP [76] databases. The sequences were then categorized into various groups of life, Archaea, Bacteria and as described in the Eukaryote tree of life [77]. Finally, for the homology estimation analysis, 662,593, 20,176,346, 7,250,814, 1,741,270 and 3,065,341 protein sequences from Archaea (~267 species), Bacteria (~4435 species), Opisthokonta (~853 species), Archaeplastida (~144 species) and diatoms (~70 species) were considered, respectively. The homology was assigned to a pair by BLASTP [78] using an expected cutoff value of 1e-5.

Detection of nucleosome occupancy sites

Nucleosome prediction with NuPoP [79] (HMM order, 3; Markov model, Linker-Nucleosome; Markov model species, NULL) on the Pt1.86 were compared with nucleosome mapping using NucHunter [80]. The following parameters are used: chunk size, 1 Mb; *p*-value threshold, 1E-6; Z-score, 3.0; interval length, 146 bp. NuPoP predicts 161,875 nucleosomes. Genome-wide di-nucleotide preference over nucleosomes and nucleosome-depleted regions was estimated by fetching their corresponding nucleotide sequences using GFF-Ex [81] and calculating the frequency of AT, TA, TT, GC and CG occurrence within the sequences using compseq [82].

Bisulfite-Seq methylation analysis

We mapped Bisulfite-Seq reads from an Illumina GAII from DNA extracted from both nitrate replete and depleted conditions after filtering through FASTQC to the Pt1.86 reference genome available using Bismark [83]. Five million reads for the replete nitrogen condition and 3.3 million reads for the low nitrogen condition were uniquely mapped and de-duplicated. Average fold coverage was 17. We extracted the methylation calls for each base and for calling a CpG/CHH/CHG site as methylated, we used a cutoff of at least three reads and a minimum of 20 % reads being methylated.

RNA-Seq data analysis for gene expression quantification

TopHat v.1.1.3 [84] and Cufflinks [85] were used to map and estimate the transcripts from the RNA-Seq data. Relative abundances of transcripts were measured as fragments per kilobase of exon per million fragments mapped (FPKM).

Additional files

Additional file 1: Figure S1. Phylogeny of different classes of histone proteins. **A** Relationship between different classes of histone proteins in *P. triconutum*. Boot strap numbers are indicated. **B** Protein sequences from different species belonging to respective classes of histone proteins were obtained from NCBI, aligned using Gonnet [87], and used to construct a neighbor-joining tree. The numeric values on the tree indicate the bootstrap confidence between two branches. Only core histones are considered in this tree. NCBI accession numbers used are as follows: [H3: *Drosophila melanogaster* (NP_724345), *Homo sapiens* (NP_003520), *Caenorhabditis elegans* (NP_496899), *Arabidopsis thaliana* (NP_195713), *Chlamydomonas reinhardtii* (XP_001690671), *Phaeodactylum tricornutum* (XP_002177514), *Thalassiosira pseudonana* (XP_002288694), *Paramecium tetraurelia* (BAF03646), *Saccharomyces cerevisiae* (AAS64349), *Dictyostelium discoideum* (XP_647577), *Leishmania infantum* (XP_001463740), *Giardia intestinalis* (ESU45648)]; [H4: *Arabidopsis thaliana* (NP_180441), *Chlamydomonas reinhardtii* (XP_001690685), *Caenorhabditis elegans* (NP_492641), *Drosophila melanogaster* (NP_524352), *Paramecium tetraurelia* (CAD97571), *Dictyostelium discoideum* (XP_642712), *Leishmania infantum* (XP_001464339), *Giardia intestinalis* (ADW95184), *Saccharomyces cerevisiae* (EDV12280), *Homo sapiens* (ESW55528), *Phaeodactylum tricornutum* (XP_002179286), *Thalassiosira pseudonana* (XP_002288196)]; [H2A: *Phaeodactylum tricornutum* (XP_002181345), *Thalassiosira pseudonana* (XP_002286413), *Paramecium tetraurelia* (XP_001433287), *Drosophila melanogaster* (NP_524519), *Caenorhabditis elegans* (NP_001263788), *Homo sapiens* (NP_003507), *Arabidopsis thaliana* (Q90681), *Chlamydomonas reinhardtii* (EDO96006), *Saccharomyces cerevisiae* (CAA81267), *Leishmania infantum* (CAD11891), *Dictyostelium discoideum* (XP_636327), *Giardia intestinalis* (ESU37209)]; [H2B: *Phaeodactylum tricornutum* (XP_002179210), *Thalassiosira pseudonana* (XP_002290856), *Saccharomyces cerevisiae* (CAA81268), *Arabidopsis thaliana* (NP_180440), *Chlamydomonas reinhardtii* (XP_001700194), *Drosophila melanogaster* (NP_724342), *Caenorhabditis elegans* (NP_505464), *Homo sapiens* (NP_003510), *Giardia intestinalis* (ESU39253), *Paramecium tetraurelia* (XP_001428087), *Leishmania infantum* (XP_001470056), *Dictyostelium discoideum* (XP_628972)]; [H1: *Drosophila melanogaster* (NP_724341), *Homo sapiens* (NP_722575), *Chlamydomonas reinhardtii* (XP_001693443), *Arabidopsis thaliana* (AAD20121), *Caenorhabditis elegans* (NP_491678), *Saccharomyces cerevisiae* (NP_015198), *Phaeodactylum tricornutum* (XP_002179285), *Thalassiosira pseudonana* (XP_002294007)].

Additional file 2: Figure S2. Identification of histone modification sites. MS/MS spectrum of novel PTMs detected in histones H3, H4, H2A and H2B, as well as histone PTMs that were mapped.

Additional file 3: Table S1. Sequencing data from each chromatin immunoprecipitation experiment followed by Illumina HiSeq 2000. The number of sequencing reads analyzed in the ChIP-Seq and RNA-Seq data are shown.

Additional file 4: Figure S3. Coverage (in base pairs) of the peaks of each histone mark on the genomic features (genes, TEs and intergenice regions) of *P. tricornutum*.

Additional file 5: Figure S4. Distributions of histone modifications on genes, TEs, Genes and TEs, and intergenic regions.

Additional file 6: Table S2. Functions of encoded proteins of genes marked with a range of histone modifications. KOG functional categories and GO enriched categories are shown. For functional enrichment, a hypergeometric test was performed with KOG classes and a GO enrichment test with GO classes. GO enrichment was performed with a Fisher exact test using a reference set of whole genome annotation (4772 genes) and an FDR value of 0.05. The comparison was performed against the unmarked genes.

Additional file 7: Figure S5. Distribution of histone marks on TEs and their correlation with expression. **A** Proportion of different classes of TEs marked by six modifications, including DNA methylation. The numbers of the different classes of TEs are indicated above each bar. **B** Expression of different classes of TEs marked by the five histone modifications.

Additional file 8: Figure S6. Nucleosome features. **A** Micrococcal nuclease digest. The three right most lanes indicate digestions of nuclei with increasing concentrations of MNase resulting in a major band of

mononucleosomes around 150 bp. Lane C indicates undigested control. The two left-most lanes indicate low and high molecular weight DNA ladders. **B** Pie chart showing genome-wide nucleosome distributions (H3) along genes, TEs, and intergenic regions. The number of each genomic feature is indicated. **C** Dinucleotide frequencies around nucleosome occupancy sites and over linker regions.

Additional file 9: Table S3. GO categories and orthology groups of H3K27me3 clusters of genes.

Additional file 10: Figure S7. Characterization of *P. tricornutum* cells grown under nitrate limiting conditions. **A** Chlorotic phenotype of cells grown under low nitrate at 75 μM (left) versus replete control at 880 μM (right). **B** Growth curves of cells grown under low (75 μM) and normal (880 μM) nitrate for 3 weeks. **C** Chlorophyll *a* and *c* contents in each culture condition measured using ethanol extraction. **D** Maximal oxygen evolution rates in each culture condition. Measurements in C and D were made after seven days of culture.

Additional file 11: Table S4. List of differentially marked and expressed genes under nitrate depleted condition.

Additional file 12: Figure S8. GO categories of genes differentially regulated and marked by H3K4me2 under low nitrate.

Additional file 13: Figure S9. GO categories of genes differentially regulated and marked by H3AcK9/K14 under low nitrate.

Additional file 14: Figure S10. GO categories of genes differentially regulated and marked by H3K9me3 under low nitrate.

Additional file 15: Figure S11. MapMan display of genes differentially marked and regulated under low nitrate showing assignment to different metabolic compartments, including light harvesting, photorespiration, amino acid biosynthesis, and lipid metabolism. MapMan of the model plant *Arabidopsis thaliana* was used.

Additional file 16: Table S5. The sites of post-tanslational modifications (PTMs) on histones include amino acid (residue) and peptide sequences which are acetylated (*Acetyl*), methylated (*Methyl*, *Dimethyl* and *Trimethyl*) or ubiquitinated.

Abbreviations

bp: base pair; CDS: coding sequence; ChIP: chromatin immunoprecipitation; CS: chromatin state; FDR: false discovery rate; GO: gene ontology; HPLC: high-performance liquid chromatography; MNase: micrococcal nuclease; MS: mass spectrometry; PTM: post-translational modification; TE: transposable element.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LT and CB conceived and designed the study. LT designed and coordinated the study. Xin Lin, BL, FD and LT performed the experiments. OM and YT characterized the cells under low N. AV and AR performed the bioinformatics analysis. MR, SO, XL, YS, JM, PR and AA performed ChIP-Seq and RNA-Seq experiments under low nitrate. AV, AR, CB, DL and LT analyzed and interpreted the data. AV, AR, DL and LT made the figures. LT wrote the manuscript with input from AV, CB, DL and AR. All authors read and approved the final manuscript.

Acknowledgments

Funding is acknowledged from CNRS (Défi Transition énergétique: ressources, société, environnement - ENRS) to LT, the ERC "Diatomite" and EU MicroB3 to CB and Cancéropôle Ile de France to DL and LT. CB additionally thanks PSL Research University (ANR-11-IDEX-0001-02). AR is a PhD student funded by the MEMO LIFE International PhD program (ANR-10-LABX-54).

Author details

[1]Ecology and Evolutionary Biology Section, Institut de Biologie de l'École Normale Supérieure (IBENS), CNRS UMR8197 INSERM U1024, 46 rue d'Ulm, 75005 Paris, France. [2]Institut Curie, PSL Research University, Centre de Recherche, Laboratoire de Spectrométrie de Masse Protéomique, 26 rue d'Ulm, 75248 Cedex 05 Paris, France. [3]Institute for Genome Sciences (IGS), University of Maryland School of Medicine, Baltimore, MD 21201, USA. [4]J. Craig Venter Institute, 10355 Science Center Drive, San Diego, CA 92121, USA. [5]Scripps Institution of Oceanography, Integrative Oceanography Division, University of California, San Diego, CA 92093, USA. [6]Present address: BESE Division, Center for Desert Agriculture, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. [7]Present address: State key lab of Marine Environmental Science, Xiamen University, Xiamen 361005, China. [8]Present address: Instituto de Biotecnología, CICVyA, Instituto Nacional de Tecnología Agropecuaria (INTA Castelar), CC 25, Castelar B1712WAA, Argentina.

Received: 16 January 2015 Accepted: 11 May 2015
Published online: 20 May 2015

References

1. Bednar J, Horowitz RA, Grigoryev SA, Carruthers LM, Hansen JC, Koster AJ, et al. Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. Proc Natl Acad Sci U S A. 1998;95:14173–8.
2. Caterino TL, Hayes JJ. Structure of the H1 C-terminal domain and function in chromatin condensation. Biochem Cell Biol. 2011;89:35–44.
3. Wolffe AP. Transcriptional regulation in the context of chromatin structure. Essays Biochem. 2001;37:45–57.
4. Tan M, Luo H, Lee S, Jin F, Yang JS, Montellier E, et al. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell. 2011;146:1016–28.
5. Bailey KA, Pereira SL, Widom J, Reeve JN. Archaeal histone selection of nucleosome positioning sequences and the procaryotic origin of histone-dependent genome evolution. J Mol Biol. 2000;303:25–34.
6. Cavalier-Smith T. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. Biol Lett. 2009;6:342–5.
7. Caron F, Meyer E. Does Paramecium primaurelia use a different genetic code in its macronucleus? Nature. 1985;314:185–8.
8. Gao XP, Li JY. Nuclear division in the marine dinoflagellate Oxyrrhis marina. J Cell Sci. 1986;85:161–75.
9. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: integrating terrestrial and oceanic components. Science. 1998;281:237–40.
10. Bowler C, Vardi A, Allen AE. Oceanographic and biogeochemical insights from diatom genomes. Annu Rev Mar Sci. 2010;2:333–65.
11. Allen AE, Dupont CL, Obornik M, Horak A, Nunes-Nesi A, McCrow JP, et al. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. Nature. 2011;473:203–7.
12. Shrestha RP, Tesson B, Norden-Krichmar T, Federowicz S, Hildebrand M, Allen AE. Whole transcriptome analysis of the silicon response of the diatom Thalassiosira pseudonana. BMC Genomics. 2012;13:499.
13. Armbrust EV. The life of diatoms in the world's oceans. Nature. 2009;459:185–92.
14. Veluchamy A, Lin X, Maumus F, Rivarola M, Bhavsar J, Creasy T, et al. Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum. Nat Commun. 2013;4:2091.
15. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 2008;456:239–44.
16. Allis CD, Glover CV, Gorovsky MA. Micronuclei of Tetrahymena contain two types of histone H3. Proc Natl Acad Sci U S A. 1979;76:4857–61.
17. Prescott DM. The DNA, of ciliated protozoa. Microbiol Rev. 1994;58:233–67.
18. Pontarotti P. Evolutionary biology: concept, modeling and application. Springer; 2009.
19. Maumus F, Rabinowicz P, Bowler C, Rivarola M. Stemming epigenetics in marine stramenopiles. Curr Genomics. 2011;12:357–70.
20. Tirichine L, Lin X, Thomas Y, Lombard B, Loew D, Bowler C. Histone extraction protocol from the two model diatoms Phaeodactylum tricornutum and Thalassiosira pseudonana. Mar Genomics. 2014;13:21–5.
21. Lin X, Tirichine L, Bowler C. Protocol: Chromatin immunoprecipitation (ChIP) methodology to investigate histone modifications in two model diatom species. Plant Methods. 2012;8:48.
22. Johnson L, Mollah S, Garcia BA, Muratore TL, Shabanowitz J, Hunt DF, et al. Mass spectrometry analysis of Arabidopsis histone H3 reveals distinct combinations of post-translational modifications. Nucleic Acids Res. 2004;32:6511–8.
23. Zhang K, Sridhar VV, Zhu J, Kapoor A, Zhu JK. Distinctive core histone post-translational modification patterns in Arabidopsis thaliana. PLoS One. 2007;2:e1210.

24. Roudier F, Teixeira FK, Colot V. Chromatin indexing in Arabidopsis: an epigenomic tale of tails and more. Trends Genet. 2009;25:511–7.
25. Jamieson K, Rountree MR, Lewis ZA, Stajich JE, Selker EU. Regional control of histone H3 lysine 27 methylation in Neurospora. Proc Natl Acad Sci U S A. 2013;110:6027–32.
26. Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, Vardi A, et al. Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. BMC Genomics. 2009;10:624.
27. Clark DJ. Nucleosome positioning, nucleosome spacing and the nucleosome code. J Biomol Struct Dyn. 2010;27:781–93.
28. Bai L, Morozov AV. Gene regulation by nucleosome positioning. Trends Genet. 2010;26:476–83.
29. Lafos M, Kroll P, Hohenstatt ML, Thorpe FL, Clarenz O, Schubert D. Dynamic regulation of H3K27 trimethylation during Arabidopsis differentiation. PLoS Genet. 2011;7:e1002040.
30. Young MD, Willson TA, Wakefield MJ, Trounson E, Hilton DJ, Blewitt ME, et al. ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. Nucleic Acids Res. 2011;39:7415–27.
31. Zhang X, Clarenz O, Cokus S, Bernatavichute YV, Pellegrini M, Goodrich J, et al. Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. PLoS Biol. 2007;5:e129.
32. Roudier F, Ahmed I, Berard C, Sarazin A, Mary-Huard T, Cortijo S, et al. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J. 2011;30:1928–38.
33. Widiez T, Symeonidi A, Luo C, Lam E, Lawton M, Rensing SA. The chromatin landscape of the moss Physcomitrella patens and its dynamics during development and drought stress. Plant J. 2014;79:67–81.
34. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell. 2010;143:212–24.
35. Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, et al. H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. Genome Res. 2009;19:221–33.
36. Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. Nature. 2011;469:343–9.
37. Shaver S, Casas-Mollano JA, Cerny RL, Cerutti H. Origin of the polycomb repressive complex 2 and gene silencing by an E(z) homolog in the unicellular alga Chlamydomonas. Epigenetics. 2010;5:301–12.
38. Paez-Valencia J, Sanchez-Lares J, Marsh E, Dorneles LT, Santos MP, Sanchez D, et al. Enhanced proton translocating pyrophosphatase activity improves nitrogen use efficiency in Romaine lettuce. Plant Physiol. 2013;161:1557–69.
39. Mersfelder EL, Parthun MR. The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. Nucleic Acids Res. 2006;34:2653–62.
40. Jeffers V, Sullivan Jr WJ. Lysine acetylation is widespread on proteins of diverse function and localization in the protozoan parasite Toxoplasma gondii. Eukaryot Cell. 2012;11:735–42.
41. Nardelli SC, Che FY, de Monerri NCS, Xiao H, Nieves E, Madrid-Aliste C, et al. The histone code of Toxoplasma gondii comprises conserved and unique posttranslational modifications. MBio. 2013;4:e00922–00913.
42. Cosgrove MS, Boeke JD, Wolberger C. Regulated nucleosome mobility and the histone code. Nat Struct Mol Biol. 2004;11:1037–43.
43. Bheda P, Swatkoski S, Fiedler KL, Boeke JD, Cotter RJ, Wolberger C. Biotinylation of lysine method identifies acetylated histone H3 lysine 79 in Saccharomyces cerevisiae as a substrate for Sir2. Proc Natl Acad Sci U S A. 2012;109:E916–25.
44. Zhang X, Bernatavichute YV, Cokus S, Pellegrini M, Jacobsen SE. Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. Genome Biol. 2009;10:R62.
45. Millar CB, Grunstein M. Genome-wide patterns of histone modifications in yeast. Nat Rev Mol Cell Biol. 2006;7:657–66.
46. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, et al. Genome-wide map of nucleosome acetylation and methylation in yeast. Cell. 2005;122:517–27.
47. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet. 2008;40:897–903.
48. Zhou J, Wang X, He K, Charron JB, Elling AA, Deng XW. Genome-wide profiling of histone H3 lysine 9 acetylation and dimethylation in Arabidopsis reveals correlation between multiple histone marks and gene expression. Plant Mol Biol. 2010;72:585–95.
49. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.
50. Filion GJ, van Steensel B. Reassessing the abundance of H3K9me2 chromatin domains in embryonic stem cells. Nat Genet. 2010;42:4. author reply 5–6.
51. Kohler C, Villar CB. Programming of gene expression by Polycomb group proteins. Trends Cell Biol. 2008;18:236–43.
52. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, et al. A cis-regulatory map of the Drosophila genome. Nature. 2011;471:527–31.
53. Hosogane M, Funayama R, Nishida Y, Nagashima T, Nakayama K. Ras-induced changes in H3K27me3 occur after those in transcriptional activity. PLoS Genet. 2013;9:e1003698.
54. Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, et al. Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. Gene. 2007;406:23–35.
55. Cui K, Zhao K. Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq. Methods Mol Biol. 2012;833:413–9.
56. Garcia BA, Mollah S, Ueberheide BM, Busby SA, Muratore TL, Shabanowitz J, et al. Chemical derivatization of histones for facilitated analysis by mass spectrometry. Nat Protoc. 2007;2:933–8.
57. Poullet P, Carpentier S, Barillot E. myProMS, a web server for management and validation of mass spectrometry-based proteomic data. Proteomics. 2007;7:2553–6.
58. ProteomExchange. http://proteomecentral.proteomexchange.org/cgi/GetDataset.
59. Vizcaino JA, Cote RG, Csordas A, Dianes JA, Fabregat A, Foster JM, et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. 2013;41:D1063–9.
60. Ritchie RJ. Consistent sets of spectrophotometric chlorophyll equations for acetone, methanol and ethanol solvents. Photosynth Res. 2006;89:27–41.
61. Joint Genome Institute. http://genome.jgi-psf.org/Phatr2/Phatr2.home.html.
62. FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
63. Bolger AM, Lohse M, Usadel B. "Trimmomatic: a flexible trimmer for Illumina sequence data". Bioinformatics. 2014: btu170.
64. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.". Bioinformatics. 2005;21:3674–6.
65. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat. 1996;5:299–314.
66. Cock PJA et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.
67. Sandve GK, Gundersen S, Rydbeck H, Glad IK, Holden L, Holden M, et al. The Genomic HyperBrowser: inferential genomics at the sequence level. Genome Biol. 2010;11:R121.
68. EnsemblProtists: Phaeodactylum tricornutum. http://protists.ensembl.org/Phaeodactylum_tricornutum/Location/Genome.
69. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
70. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.
71. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics. 2009;25:1952–8.
72. Thorvaldsdóttir H, Robinson JT, Mesirov JP. "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration". Brief Bioinform. 2012: bbs017.
73. Salmon-Divon M, Dvinge H, Tammoja K, Bertone P. "PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci". BMC bioinformatics. 2010;11:415.
74. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: a free, open-source system for microarray data management and analysis. Biotechniques. 2003;34:374–8.
75. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2005;33:D154–9.
76. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project

(MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 2014;12:e1001889.
77. Baldauf SL. An overview of the phylogeny and diversity of eukaryotes. J Syst Evol. 2008;46:263–73.
78. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
79. Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J, Wang JP. "Predicting nucleosome positioning using a duration Hidden Markov Model". BMC bioinformatics. 2010;11:346.
80. Mammana A, Martin V, Ho-Ryun C. "Inferring nucleosome positions with their histone mark annotation from ChIP data". Bioinformatics. 2013;29:2547–54.
81. Rastogi A, Gupta D. GFF-Ex: a genome feature extraction package. BMC Res Notes. 2014;7:315.
82. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16:276–7.
83. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2.
84. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.
85. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
86. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. Science. 1992;256:1443–5.
87. Shen L, Shao NY, Liu X, Maze I, Feng J, Nestler EJ. diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. PLoS One. 2013;8:e65598.

Chapter 2

Phatr3: The Functional Genome

The first genome assembly of *Phaeodactylum tricornutum*, Phatr1, was released in October 2005 with 588 scaffolds and 10681 gene models, constituting 31 MB genome size. The next release Phatr2 was made in 2008 with 33 main chromosomes and 55 scaffolds composing the genome size of 27.4 MB. Phatr2 is composed of 10402 gene models with an average gene length of 1617 bp(s). This version is currently in use and was majorly established based on a comprehensive set of EST(s) generated using numerous *P. tricornutum* cell lines maintained under different conditions. In the current chapter, I discuss the release of new genome annotation, Phatr3, for *P. tricornutum* genome. In summary, I used 90 RNA sequencing libraries, done using Illumina NGS platform, that were maintained in different conditions, to discover new and improved gene models compared to the last release Phatr2. In light of new gene models the current chapter also provides valuable insights on the occurrence of alternative splicing, which has never been studied in *P. tricornutum*. Furthermore, using an updated set of diatom transcriptomes and state-of-the-art computational tools, I attempted to understand the phylogenetic origin of *P. tricornutum* and/or diatom genomes in general.

(Manuscript in revision in NAR)

Improved annotation of *Phaeodactylum tricornutum* genome reveals novel genes, alternative splicing and extended repertoire of transposable elements

Achal Rastogi[1], Uma Maheswari[2], Richard G. Dorrell[1], Florian Maumus[3], Adam Kustka[4], James McCarthy[5], Andy E. Allen[5, 6], Paul Kersey[2], Chris Bowler[1] and Leila Tirichine[1]

[1]Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France

[2]EMBL-EBI, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, United Kingdom

[3]INRA, UR1164 URGI—Research Unit in Genomics-Info, INRA de Versailles-Grignon, Route de Saint-Cyr, Versailles 78026, France

[4]Earth and Environmental Sciences, Rutgers University, 101 Warren Street, 07102 Newark, New Jersey, USA

[5]J. Craig Venter Institute, 10355 Science Center Drive, 92121 San Diego, California, USA

[6]Integrative Oceanography Division, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA

Abstract

Recent decades have seen a remarkable increase in the development of molecular tools and genetic resources to understand the biology of diatoms, a divergent group of photosynthetic eukaryotes of secondary endosymbiotic origin. The model pennate diatom *Phaeodactylum tricornutum* is one of the first diatoms with a completed genome sequence, which has been very useful for dissecting key genes in diatom metabolism and responses to environmental cues. However, the previous genome annotations have missed several genes and transposable elements, and contain numerous incomplete or erroneous gene models. We used a large RNA-Seq dataset, combined with published expressed sequence tags, protein sequences, and chromatin landscapes to improve annotation quality across the *P. tricornutum* genome. The new annotation, denoted Phatr3, introduces 1,489 new genes, which together with the unchanged gene models, provide valuable insights into alternative splicing and updated knowledge of epigenetic control of gene regulation in diatoms. In addition, we have used up-to-date reference sequence libraries to improve our understanding of the phylogenetic origins of diatom genomes. Finally, we have performed a comprehensive *de novo* annotation of repetitive elements showing that repeats contribute over 10% of the *P. tricornutum* genome, with a high predominance of Copia-type transposable elements. These improvements will facilitate studies of diatom gene function and evolution.

Introduction

Diatoms are one of the most important and abundant photosynthetic micro-eukaryotes, and are believed to contribute annually about 40% of marine primary productivity and 20% of global carbon fixation (1). Marine diatoms are highly diverse and span a wide range of latitudes, from tropical to Polar Regions. The diversity of planktonic diatoms was recently estimated using metabarcoding to be around 4,748 operational taxonomic units (OTUs) (2,3). In addition to performing key biogeochemical functions (4), marine diatoms are also important for human society, being the anchor of marine food webs, and providing high value compounds for pharmaceutical, cosmetic and industrial applications (5,6). A deeper understanding of their genomes can therefore provide key insights into their ecology and biology.

In the last decades, several diatom species have been sequenced, including the pennate diatom *Phaeodactylum tricornutum* (7) and the centric diatom *Thalassiosira pseudonana* (8). Analysis of the *P. tricornutum* genome revealed an evolutionarily chimeric signal with genes apparently derived from red and green algal sources as well as the endosymbiotic host, and from a range of bacteria by lateral gene transfers (7,9), although the exact contributions of different donors to the *P. tricornutum* genome remains debated (10,11), alongside vertically inherited and lineage-specific genes. Such a diversity of genes has likely provided *P. tricornutum* and diatoms in general with a high degree of metabolic flexibility that has played a major role in determining their success in contemporary oceans.

The availability of a sequenced genome for *P. tricornutum* has also opened the gate for functional genomics studies, using Gateway, RNAi, CRISPR (12,13), TALEN and conjugation and have revealed novel proteins and metabolic capabilities such as the urea cycle (14), proteins important for iron acquisition (15,16), cell cycle progression (17), lipid metabolism for biofuel production (18,19), as well as for red and far/red light sensing (20). Deeper understanding of the ecology and success of diatoms as well as a thorough dissection of the gene repertoire of *P. tricornutum* will however

require better information regarding genome composition and gene structure. The first draft of the *P. tricornutum* genome (Phatr1), based on Sanger sequencing, was released in 2005, and contained 588 genome scaffolds totalling 31 Mb. The draft genome was further re-annotated and released as Phatr2 in 2008 with an improved assembly condensed into 33 scaffolds and 55 unmapped sequences which could not be assigned to any of the mapped chromosomes (denoted Phatr2 bottom drawer) (7). Subsequently, sequencing technologies have evolved and RNA-seq has been established as a gold standard for transcriptome investigation, and knowledge has advanced rapidly concerning the roles of small RNAs and histone modifications on transcriptional regulation. Therefore, we exploited a large set of RNA-Seq reads derived from cells grown in different conditions, in combination with expressed sequence tags (ESTs) (21) and protein sequences (UniProt), as well as histone post-translational modifications (22) to re-annotate the *P. tricornutum* genome. This allowed identification of a significant number of novel transcripts as well as improved annotation of gene models including splice variants and a refined structure of exon intron junctions and gene ends. This resource was released as *Phaeodactylum tricornutum* annotation 3 (Phatr3) and is available to the community on the Ensembl portal (http://protists.ensembl.org/Phaeodactylum_tricornutum/Info/Index).

Results and Discussion

Building Phatr3 gene models

To generate a new annotation of the *P. tricornutum* nuclear genome, our approach combined high-throughput RNA sequencing data along with ESTs and protein sequences. 14,505 gene models were predicted from RNA-Seq mapping and aligning the EST data-set using est2genome. Additionally, we used SNAP and Augustus to predict 10,743 and 11,611 gene models, respectively. MAKER2 was used to merge the three sets and predicted 12,055 gene models with an average gene length of 1,624 bp and ~1.7 exons per gene. The predicted putative genes were then compared to the Phatr2 gene models (http://genome.jgi.doe.gov/Phatr2/Phatr2.home.html) and categorized as: 1) New, 2) Unchanged or Modified at 5' and/or 3', 3) Split (former Phatr2 genes that were split into two or more gene models), or 4) Merged gene models (former two or more Phatr2 genes that were merged as one Phatr3 gene model) (Table 1). The list of genes in each category is shown in Supplementary Table S3. Comparison between Phatr3 and Phatr2 revealed a total of 1,489 novel genes. We found no notable evidence from the Phatr3 annotation for 142 Phatr2 genes. A total of 122 of these were however considered as true genes based on the presence of epigenetic features known to be hallmarks of coding sequences, like acetylation which peaks over the 5' end of genes and H3K4me2 which lies over gene bodies. The remaining 20 genes were not considered in Phatr3 and can be accessed via the Phatr2 portal hosted by the Joint Genome Institute (JGI). The basic statistics of the structural comparison between Phatr3 and Phatr2 gene models are presented in Table 1.

Phatr3 extends the repertoire of *bona fide* genes

To provide further evidence for the 1,489 novel transcripts, we examined the potential influence of chromatin-based regulatory processes based on DNA methylation and histone (H3) post-translational modification (PTMs) profiles, in light

of our previously published work (22,23). We have shown that genes are preferentially labelled by active marks such as acetylation and H3K4me2 with a particular pattern of mapping but can also carry marks known to be repressive in other systems including H3K27me3, H3K9me3 and H3K9me2. Among the novel transcripts, 99 (~7%) genes show DNA methylation among which 91 (~92%), 17 (~17%) and 9 (~9%) genes are found to be methylated in CG, CHH and CHG contexts, respectively (Fig 1A). Compared to the non-methylated novel transcripts, the expression of methylated genes is low (Fig 1B) and the expression of genes marked specifically by CHG is the lowest. We then looked at five histone H3 post-translational modifications associated with these genes. In general, from 1,489 novel transcripts, 1,481 (~97%) were marked either by DNA methylation or the studied H3 PTMs. Of the 1,489 candidate genes, 1,364 (~91%) genes are marked by at least one epigenetic mark among which 57 (~4%) are marked by all five modifications (Fig 1C). While most of the new genes (1,227; ~90%) are co-marked by at least two epigenetic marks, 101 (~7%), 17 (~1%), 13 (~1%) and 6 (~1%) genes are marked specifically by H3K4me2, H3K9me2, H3K9_14Ac and H3K27me3, respectively (Fig 1B). Among the genes specifically labeled by one or other studied PTMs, those marked by H3K27me3 have the lowest expression, confirming the role of this mark in maintaining a repressive state of the targeted gene , while acetylation as well as H3K4me2 increases the expression of target genes.

Functional annotation of novel transcripts

A total of 586 (~39%) of the new gene models in Phatr3 can be functionally annotated using the UniProt-GOA (UniProt release 2015_03) database. Specifically, 579 and 492 genes out of 1,489 new gene models have protein domain information from InterPro and PFAM databases, respectively. Only 66 genes have known cellular component (CC) information, whereas biological process (BP) and molecular function (MF) can be inferred for 190 and 344 genes, respectively (Fig 2A). Consequently, we managed to evaluate the enriched (chi-squared test; P-value < 0.05) biological processes within novel genes (1489 genes) using only 190 (~13%) genes. The novel

genes perform a variety of different biological processes, with significant enrichments in 7-methylguanosine mRNA capping, chromatin remodeling menaquinone biosynthesis, and potassium ion transport (refer to Table S3 for all enriched processes). Most of the biological processes, like ones mentioned latter, are exclusively enriched in novel genes relative to unchanged gene models (Table S3), indicating the value of Phatr3 annotations in elevating the functional knowledge-base of *P. tricornutum* genome.

In addition to the functional annotations, we used ASAFind and HECTAR to predict the sub-cellular targeting of the new gene models (Fig 2B and 2C, Table S4). The proteins encoded by the new gene models were predicted to localize to all three organelles considered by each targeting programme, namely the plastid (196 proteins predicted with ASAFind, 63 with HECTAR), the endomembrane system (188 with ASAFind, 370 with HECTAR) and the mitochondria (92 with HECTAR, mitochondrial targeting not considered with ASAFind) (Table S4). This is broadly analogous to the targeting predictions made for proteins encoded by Phatr3 genes with unchanged Phatr2 models, although both programs suggested a slightly higher proportion of the novel gene models localized to the cytoplasm or lacked clear targeting information (Fig 2B and 2C). Together, the presence of domains performing known biological processes, and of clear subcellular localization predictions, confirms that many of the novel genes identified within the *P. tricornutum* genome possess specific biological functions.

Homology estimation of *P. tricornutum* proteome

The initial publication of the *P. tricornutum* genome in 2008 (7) suggested that nearly 40% of the genes in the *P. tricornutum* genome were species-specific, i.e. were not shared with the closest relative compared (the diatom *Thalassiosira pseudonana (8)*). Since then, the amount of sequence information available for other lineages of the tree of life has increased dramatically, including multiple genome sequences for previously under sampled taxa (e.g., red algae), and large-scale

transcriptome sequence data for a variety of eukaryotic species (e.g., the Marine Microbial Eukaryote Transcriptome Sequencing Project). We wished to use this data, alongside improved gene models within Phatr3, to quantify the lineage specific versus conserved genes in the *P. tricornutum* genome, and to determine whether genes with different levels of conservation have different functional properties.

We assembled a large sequence library, consisting of the complete UniRef library, supplemented with multiple genome and transcriptome sequences, currently not integrated into UniRef (Table S2). This library was split into 73 phylogenetic sub-categories, based on recent published phylogenies (See Methods), which were grouped into nine lineages (diatoms, other ochrophytes, other algae with complex plastids, other heterotrophic SAR clade members, green algae and plants, red algae, other eukaryotes, prokaryotes, and viruses; Table S2). Two complementary techniques were employed to estimate the conservation status of each gene in the *P. tricornutum* genome. Firstly, to minimize the identification of false positives due to random associations, Reciprocal Best-Hit BLAST (RBH-BLAST) was performed across the entire *P. tricornutum* genome. Genes were only deemed to be "conserved" with a particular lineage if orthologues were detected using RBH-BLAST with an expect value of <1E -05 (Tables S3, S5). Secondly, to minimize the identification of false positives due to contamination in sequence libraries and species-specific gene transfer, the conservation of genes within the different sub-categories that comprise a particular lineage were considered. For this analysis, genes were deemed to be conserved if a BLAST top hit with an expect value of <1E -05 were found in more than half of the sub-categories that comprise a particular lineage (Table S1, S4). We repeated the latter analysis with three different threshold expect values (1E-05, 1E-10, 1E-50) and three threshold levels of conservation (presence in >1/2, >1/3, and >2/3 sub-categories within a particular lineage (Table S5).

The conservation analyses revealed that a significantly greater proportion (83-86%) of the Phatr3 genes are of the *Phaeodactylum* genes are shared among one or more

groups within the tree of life (10,110 genes, RBH analysis; 10, 417 genes, majority conservation; Table S3) than was previously found (7). 14 different patterns of conservation were found consistently, in all analyses, to occur more frequently than would be expected by random association (chi-squared test, P < E-02; Table S3). 3-9% of the Phat3 genes were conserved across all other lineages studied (333 genes, RBH analysis; 1007 genes, majority conservation); 9-14% across all lineages except viruses (1749 genes, RBH analysis; 1111 genes, majority conservation); 1-3% across all lineages except prokaryotes (55 genes, RBH analysis; 364 genes, majority conservation); and 11-15% across all eukaryotes (1325 genes, RBH analysis; 1871 genes, majority conservation). Smaller but significant numbers of genes were found to have more complex patterns of conservation (Table S3), with the largest of these categories being all eukaryotes except red algae (which may reflect ancient reduction events in red algal genomes), and all eukaryotic algal lineages (inclusive or exclusive of red algae) (Table S3).

We still found, with the expanded dataset, that many genes within the *Phaeodactylum* genome have limited evolutionary conservation. 14-17% of the genes (2,067 genes, RBH analysis; 1,784 genes, majority conservation) were found to be specific to *P. tricornutum*, while 13-19% were only shared between *P. tricornutum* and other diatoms (1616 genes, RBH analysis; 2300 genes, majority conservation), 3-4% were only shared with diatoms and other ochrophytes (424 genes, RBH analysis; 469 genes, majority conservation), 1-3% with ochrophytes and other algae with complex plastids (337 genes, RBH analysis; 153 genes, majority conservation), and approximately 1% with ochrophytes and other SAR clade members (145 genes, RBH analysis; 154 genes, majority conservation). We could not find significant evidence for genes that were shared between *P. tricornutum* and non-diatom lineages, but absent in other diatoms, or between *P. tricornutum* and non-ochrophyte lineages, but absent in other ochrophytes (Tables S3, S5). Thus, the *Phaeodactylum* genome contains many lineage-specific genes, and there has been largely vertical inheritance of genes in the recent evolutionary history of *P. tricornutum*, at least since the radiation of ochrophytes.

We wished to determine whether genes with different levels of conservation have different functional properties in the *Phaeodactylum* genome. Firstly, we tested to see whether the 1,489 novel genes uncovered by Phatr3 differ in terms of evolutionary conservation to those previously identified. While many of the novel genes are specific to *P. tricornutum* (33-49%: 728 genes, RBH analysis; 502 genes, majority conservation; Table S3) or are limited to diatoms (13-28%: 188 genes, RBH analysis; 409 genes, majority conservation; Table S1), as with the full dataset, significant numbers of novel genes were found to be conserved across the tree of life (2-4%: 66 genes, RBH analysis; 35 genes, majority conservation; Table S3), or across all eukaryotes (5-8%: 74 genes, RBH analysis; 121 genes, majority conservation; Table S3), confirming that many of these genes are likely to have important biological functions. Next, we performed GO enrichment analysis (using assigned biological process (BP), chi-squared test; P-value < 0.05) to evaluate the functional nature of genes, as determined by RBH analysis, that are specific to diatoms (uniquely shared with diatoms, 1,619 genes), those that are uniquely shared with eukaryotes (1325 genes) and are broadly conserved across the tree of life, except viruses (1749 genes). Overall, ~12%, ~47% and ~76% genes are found to be have at least one biological process associated to them, from diatom specific, eukaryotic specific and broadly conserved genes, respectively. Among the very few genes that have predicted BP assignations within diatom specific gene-set, the top five enriched BP includes: cell division, lipid catabolism and metabolism, protein phosphorylation and DNA-templated regulation of transcription. While processes corresponding to intracellular protein transport, mRNA splicing via spliceosomes, protein and vesicle mediated transport are highly enriched in eukaryotic specific genes, processes like protein phosphorylation, DNA templated regulation of transcription, tRNA aminoacylation for protein translation and vesicle-mediated transport are enriched in *P. tricornutum* genes that are broadly conserved across the tree of life, except viruses. By analyzing all enriched processes within different sets of specific genes, it appears that *P. tricornutum* or in general most diatoms have attained numerous genes involved in signaling, exchanging and acclimatizing genetic

and nutritional elements from their ecological niches. All enriched processes are available in Table S3.

Characterization of ancient gene transfers in the *P. tricornutum* genome

We wished to use the Phatr3 genome annotation, and our expanded reference sequence dataset, to reassess the different gene transfer events that have occurred in the evolutionary history of *P. tricornutum*. In particular, we wished to validate the presence of genes derived from green algae and from prokaryotes, which have previously been controversial. To do this, the phylogenetic distribution of BLAST top hits across the Phatr3 genome was tabulated using the full reference sequence dataset previously generated (Fig. 3; Table S6). A phylogenetic origin was only considered to be real if the first three top hits originated from the same lineage (i.e. was phylogenetically consistent). Given the evidence from our conservation analysis pipeline for largely vertical gene inheritance in the recent history of *P. tricornutum* (Table S1), we also profiled the presence of more ancient gene transfers. To do this, the BLAST top hit analysis was repeated using versions of the reference library from which five different lineages (all pennate diatoms, all diatoms, all ochrophytes, all algae with complex plastids, and all remaining SAR clade members) were iteratively removed (Fig. 3).

Only 40 of the 12177 Phatr3 genes producing phylogenetically unambiguous BLAST tophits against non-diatom (i.e. non-vertically inherited) lineages, when searched against the full reference dataset (Fig. 3). Of these, the greatest number (13) yielded top hits against prokaryotes, and five of yielded top hits with one specific prokaryotic sub-category (the Deinococcus-Thermus clade), which was significantly more than might be expected through random assortment (chi-squared test, P= 0.02; Table S6). Removal of all pennate diatom sequences from the reference library did not have any major effects on the BLAST top hits recovered, confirming that *P. tricornutum* has received few genes from other sources in its recent evolutionary history.

Substantial changes in BLAST top hits were found by searches using libraries from which all remaining diatom, ochrophyte, complex algal sequences, or SAR clade sequences had been removed (Fig. 4). 504 probable gene transfer events were identified between prokaryotes and different ancestors of the diatom lineage (Table S6. These top hits were found to be distributed across multiple prokaryotic lineages, consistent with multiple gene transfer events, with significant enrichments (chi-squared test, P< 1E-03) found for the Deinococcus-Thermus clade, Chlamydiae, Chloroflexi, cyanobacteria, and Planctomycetes (Fig. S1; Table S6). 386 genes produced top hits against members of the red algae, consistent with the red algal ancestry of the diatom plastid (Fig. 3). Only 38 of these red algal genes were identified by removing heterotrophic members of the SAR clade (i.e. oomycetes, labyrinthulomycetes, slopalinids, Rhizaria, and ciliates) from the dataset (Fig. 3; Table S6). The limited number of genes of red algal origin identified within these groups is consistent with them never having possessed red algal plastids.

Finally, 1,383 genes generated top hits from members of the green lineage (green algae and plants), consistent with large-scale gene transfer between diatom ancestors and green algae (Fig. 3; Table S6). Some of these genes may be misidentified genes of red algal origin, as has been discussed elsewhere; however, we believe that many are genuinely of green origin, for three reasons. Firstly, compared to previous phylogenomic studies of diatom genomes, our reference library contains a much larger amount of red algal sequence information, including five complete genomes, and large-scale transcriptomes for a further twelve red algal species (Table S2). Secondly, green gene transfers appear to have occurred at a different time point to the red algal gene transfers, as the largest number (624/ 1,383) were only identified following the removal of all heterotrophic SAR clade members from the reference dataset (Fig. 3), in contrast to the relatively limited number of red genes identified with this dataset. Finally, the green top hits were biased towards specific sub-categories, whereas misidentified red genes should be distributed randomly across the entire green lineage. Only one-fifth of the green lineage top hits were found to come from streptophytes, despite these organisms

comprising three-quarters of the green lineage sequences (Fig. S1), while strong signal enrichments (chi-squared test, P=0) were found for five chlorophyte/ prasinophyte sub-categories (chlorodendrophytes, Pyramimonadales, the *Micromonas/ Mantoniella* clade, *Ostreococcus*, and *Nephroselmis/ Prasinoderma*) (Fig. S1, Table S6).

Examination of repetitive sequence content

In the context of the Phatr3 gene prediction, we also revisited the annotation of repetitive elements in the genome assembly. For the first time, we applied a robust and *de novo* approach for the whole genome annotation of repeat sequences. Collectively, repeats were found to contribute 3.1 Mb (11%) of the assembly, including transposable elements (TEs), unclassified and tandem repeats, as well as fragments of host genes (Table 2). TEs are the dominant repetitive elements in *P. tricornutum* and represent 75% of the repeat set, i.e., 2.3 Mb as compared to 1.7 Mb in the previous TE annotation. In line with previous analyses, Copia-type LTR retrotransposons (LTR-RTs) are the most abundant type of TEs, contributing over 55% of the repeat annotation, while Gypsy-type LTR-RTs remain undetected. In agreement with other studies(24), this new TE annotation also reveals the presence of traces of Crypton-type transposons in *P. tricornutum,* suggesting an ancient origin of this type of transposons which were also found in fungi and multiple invertebrate lineages . By comparing the current TEs with the previous TE annotations, 1,993 (~54%) TEs are found to be novel.

We further checked the distribution of epigenetic marks (H3K27me3, H3K9me3, H3K9me2, H3K4me2, and H3K9-14Ac) and DNA methylation on these new TEs. Of the latter, 1,236 (~62%) TEs are found to be marked by at least one of the epigenetic marks, from which only 75 (~6%), 25 (~2%), 111 (~9%), 58 (~5%) and 62 (~5%) are marked specifically by H3K27me3, H3K9me3, H3K9me2, H3K9-14Ac and H3K4me2, respectively (Fig 4A). TEs specifically marked by H3K27me3 and H3K9me3 exhibit least expression and are consistent with their role in maintaining repressive state of

the genome (Fig 4B). Although the expression of novel genes that are specifically marked by H3K9me2 is slightly higher than those specifically marked by H3K9_14Ac (Fig 4B), it is reported to be repressive genome wide. We further investigated the localization pattern of DNA methylation over the newly annotated TEs. From 458 (~23%) of the 1,993 new TEs that we found to be methylated, most of them (368; ~80%) are methylated in CG context only (Fig 4C). On the other hand, 19 (~4%) TEs are specifically methylated in CHH context, and 34 (~7%) are found to be methylated by at least two sub-contexts of DNA methylation (CG, CHH and CHG) (Fig 4C). TEs those are methylated specifically in CG context reveals least expression compared to those specifically methylated in CHG or CHH contexts (Fig 4D).

By classifying the entire repertoire of TEs based on their method of transposition, Class I and II TEs show distinct patterns of DNA methylation and epigenetic marking. While most of the class I TEs are enriched with CG methylation, co-localizing with or without CHH methylation, class II TEs are predominantly marked by DNA methylation in CHH context and CGs co-localize with CHG methylation (Fig S2). A similar pattern was reported in soybean where a high abundance of CHH methylation over class II elements was correlated with the presence of small RNAs(25). This specific pattern might also be relevant to the nature of replication of each class of TEs; class I relies on a replicative mechanism for transposition while class II elements move by a cut and paste mechanism. As class I TE copy number can rapidly increase, their higher methylation in both CG and CHG contexts might be a means to keep their expression under tight control. An interesting pattern of co-occurrence of epigenetic marks over TEs emerges from this analysis that shows a systematic repression of TEs when associated with CHG methylation (Figure S2). Although associated with active marks such as H3K9/14 Ac and H3K4me2, both classes of TEs show a significant decrease in their expression, suggesting the importance of maintenance of a heterochromatic environment. When checked, most of these TEs are found to be inserted into or overlapping with genes, reflecting the importance of keeping these TEs silent. A similar phenomenon has been

observed for Arabidopsis TEs which are inserted into genes and whose repression is required to avoid the deleterious effects of TE insertion into host genes (26).

Alternative splicing is widespread in *P. tricornutum*

Finally, we profiled the distribution and dynamics of introns in the *P. tricornutum* genome. 4,014 (~33%) Phatr3 genes were predicted to contain only one intron, while 1,730 (~14%) genes contain more than one intron and 6,434 (~53%) are predicted to not contain any introns at all. The low density of introns observed in *P. tricornutum* (on average 0.7 introns per gene) is similar to what has been observed in the unicellular yeast *Saccharomyces cerevisiae* and other fungi(27).. The average intron size in *P. tricornutum* (142 bp) is small compared to other eukaryotes which might simply mirror genome size. However, small introns have also been shown to correlate with species experiencing metabolically demanding behaviors, for which small introns can increase transcriptional efficiency or splicing accuracy. Unlike most eukaryotes (28), in *P. tricornutum* we found no significant difference between the average length of the first intron with respect to the others. This is similar to some species of the genus *Phytophthora* (Fig S3) and to what has been reported in *Schizosaccharomyces pombe* and *Aspergillus nidulans* (28). However, the observation is not consistent with what has been observed in *Ostreococcus tauri,* and in multiple species of the genus *Plasmodium*, where first introns are found to be significantly longer than non-first introns (Fig S3).

It is well known that introns can have regulatory functions because they are the basis of alternative splicing, generating multiple proteins from a single gene, and they are also known to produce non-coding RNA molecules that can regulate transcription(29). Exon-skipping (ES) and intron-retention (IR) are two major constituents of the alternative splicing code in higher eukaryotes and plants, respectively(30). Because AS may have an important impact on gene expression in *P. tricornutum*, we evaluated the type of AS that can be found in Phatr3 with RNA-Seq data generated in different growth and stress conditions (Table S1). From the 12,177

Phatr3 gene models, 2,924 (~24%) genes are found to have introns that can be retained in more than 20% of the total experimental samples studied, while 2,444 (~20%) genes are observed to skip one or more exons in various samples. A total of 1,335 (~11%) genes are found to undergo both ES and IR, hence can perform alternative splicing (Fig 5A). Like most unicellular eukaryotes and unlike metazoans, *P. tricornutum* shows a higher rate of IR than ES, supporting the hypothesis that ES has become more prevalent over the course of metazoan evolution (23).

Alternative splicing was investigated in a subset of samples that have experienced different forms of nitrate stress. This revealed significant differences in the number of genes alternatively spliced between normal and stress conditions. 730 and 646 genes that are not found to be alternatively spliced in normal growth conditions (WT), undergo IR and ES under N-stress conditions, respectively (Fig 5B and 5C). GO enrichment analysis of the latter genes revealed a top enrichment of biological processes like DNA replication, mitochondrial translational elongation, ATP synthesis coupled proton transport, and glutamine catabolism (Table S3 for further classes). Thus, AS may have important roles in regulating the physiology of *P. tricornutum* under fluctuating environmental conditions.

Conclusions

The use of high coverage RNA-Seq transcriptomics in *P. tricornutum* grown in various conditions, coupled with information from chromatin landscaping and improved annotation tools, has increased the repertoire of genes and transposable element classes, and improved predictions about the structure of proteins previously identified in Phatr2. In-depth analysis of RNA-Seq reads from different libraries identified alternatively-spliced forms including intron retention and exon skipping, the former being more prevalent, a common feature in single celled eukaryotes. Comparison of the extended *P. tricornutum* genome to expanded reference libraries has illuminated ongoing questions regarding diatom evolution, such as the timing and extent of gene transfer from bacteria, red algae and green algae into diatom genomes. This work provides a more accurate annotation of a widely used model diatom, which will undoubtedly be useful for functional as well as comparative and evolutionary genomics studies. Further manual curation of gene models are ongoing and will be provided to the community members through the Web Apollo annotation tool.

Methods

Data

Phaeodactylum tricornutum genome re-annotation (Phatr3) was done on the Phatr2 assembly (ASM15095v2). The Phatr2 assembly was generated by the Joint Genome Institute (JGI), which resulted in 10,402 gene models from 33 assembled scaffolds (12 complete and 21 partial chromosomes) and 55 unassembled scaffolds (7). Apart from the previous assembly information, the species-specific data used in this re-annotation included 13,828 non-redundant ESTs (21,31), 42 libraries of RNA-Seq generated using Illumina technology and 49 libraries of RNA-Seq data generated under various conditions of iron availability using SoLiD sequencing technology. Other data used included 93,206 Bacillariophyta ESTs from dbEST (32), 22,502 Bacillariophyta, 118,041 Stramenopile protein sequences from UniProt (33) and, histone (H3) post translational modification data (H3K27me3; H3K9me2; H3K9me3; H3K4me2 and H3K9_14Ac) from our previous study (22).

Mapping and gene discovery: Phatr3

The 42 RNA-Seq Illumina dataset was mapped to the genome using Genomic Short-read Nucleotide Alignment Program, GSNAP (34), whereas the 49 SoLiD sequence datasets were aligned to the genome in color space (35). The mapping percentage was estimated using Samtools and varied between 70 – 96% (Table S1). Transcripts from the mapped reads were generated using Cufflinks (36). These transcripts along with EST libraries and protein sequences were used to train SNAP and Augustus(37) gene prediction programs using the MAKER2 annotation pipeline (38).

Annotation

All predicted gene models were annotated for protein function using InterProScan(39) and hardware-accelerated double-affine Smith-Waterman alignments (www.timelogic.com) against UniProt and other specialized databases such as KEGG(40). Finally, KEGG matches were used to map EC numbers (http://www.expasy.org/enzyme/) and UniProt hits were used to map GO terms. The results are available at http://protists.ensembl.org/

Phaeodactylum_tricornutum/Info/Index/. We investigated the presence of organelle signaling signatures within the entire Phatr3 gene repertoire using ASAFind and HECTAR, under the default conditions as specified in the original publications for each programme (41,42). ASAFind v 2.0 was run remotely across the entire dataset, using SignalP v 3.0 (43) to predict the presence of signal peptide sequences. HECTAR was run remotely, using the Galaxy integrated server provided by the Roscoff Culture Collection (http://webtools.sb-roscoff.fr/).

Distribution of epigenetic marks

Data corresponding to epigenetic marks, H3K27me3; H3K9me2/3; H3K14_Ac; H3K4me2 and DNA methylation (CG, CHH, CHG), were taken from (22) and (23,44), respectively. RNA-Seq data for Pt18.6 (normal growth condition) was also downloaded from the same resource. Distribution of all marks along with the expression were then checked over the new genes and new TE models by adapting the methodology applied in (22) and (23).

Evolutionary analysis

The evolution of the *P. tricornutum* genome was examined using gene homology searches. Amino acid identities were calculated from Smith–Waterman alignments. A composite reference library, consisting of 75001602 non-redundant sequences was compiled from UniPROT (http://www.uniprot.org/help/uniref) (downloaded February 2016), alongside additional genomic and transcriptomic resources from JGI, MMETSP, and the 1kp project currently not located on UniPROT (http://genome.jgi.doe.gov;http://marinemicroeukaryotes.org/;https://sites.google.com/a/ualberta.ca/onekp/) (Table S2). To minimize phylogenetic artifacts arising from contamination, each MMETSP library was pre-cleaned using a BLAST pipeline kindly provided by Daniel Richter (Roscoff Culture Collection), which identifies and removes sequences with 100% pairwise identity that are present in two different species libraries, and a further twelve MMETSP libraries found to contain high levels of residual contamination were excluded (see Table S2). The reference sequence library was then split into twenty-six prokaryotic sub-categories (using UniPROT

taxonomy as a guideline), and forty-seven eukaryotic sub-categories, defined using up-to-date phylogenetic information (45-52), and these categories were finally binned into nine distinct lineages (diatoms, ochrophytes, algae with complex plastids, heterotrophic SAR clade members, red algae, green algae and plants).

Conceptual translations of the Phatr3 genome were then searched against the composite reference library using RBH-BLAST (Reciprocal best hit BLAST), under default alignment conditions, with a threshold expect value of 1E-05. Genes were identified as likely to be conserved between *P. tricornutum* and a given lineage if at least one orthologue were identified within the lineage. A secondary series of conservation analysis was performed using the initial primary BLAST output for each gene. In this case, genes were deemed to be conserved if BLAST tophits below a specified threshold expect value (1E-05, 1E-10, or 1E-50) could be identified in multiple sub-categories within a given lineage (either > 1/3, >1/2, or >2/3 all sub-categories within a lineage). The values obtained from each analysis were compared, and only forms of conservation that in every analysis were found to occur significantly more frequently (chi-squared test, 1E-05) than would be expected through uniform distribution of the data were deemed to be genuine.

To identify the probable closest evolutionary relative of each gene, the BLAST top hits obtained for each gene against each phylogenetic sub-category were compiled to form a localized reference directory. As this directory only contains the single BLAST best hits for each category, it is de-enriched for phylogenetically redundant and duplicated genes. BLAST top hit analysis was then performed for each gene against this directory, with an evolutionary affinity only recorded if a phylogenetically consistent signal was obtained from the first three BLAST hits (for example, a gene for which the first three BLAST hits were of diatom origin would be recorded as being a diatom-derived gene). To distinguish between BLAST results that represent a genuine phylogenetic affinity between two lineages, versus noise and/ or random phylogenetic assignation within the dataset (53,54), the number of top hits observed for a particular phylogenetic lineage was compared to what would be

expected given random assortment, and the size of the reference dataset for that lineage (chi-squared test).

Annotation of repetitive elements

We used the REPET v2.2 package (55) to detect the repetitive fraction of the *P. tricornutum* genome. The TEdenovo pipeline (https://urgi.versailles.inra.fr/Tools/REPET) was launched including the RepeatScout approach (56) to build a library of consensus sequences representatives of the repeated elements in the genome assembly. The classification comes from decision rules applied to the evidence collected from the consensus sequences. These include: search for structural features, search for tandem repeats, comparison to PFAM, comparison to Repbase (57) and to a local library of known TEs. The library of manually curated TEs that has been established in previous work was appended to the TEdenovo library and redundancy was removed from the combined library. The TEannot pipeline was then run with default settings using the sequences from the filtered combined library as probes.

Alternative splicing

To explore the set of genes undergoing alternative splicing, exon skipping or intron retention, we used 17 RNA-Seq samples prepared under different conditions of nutrient limitation (Table S1). Only genes that were annotated as having two or more exons, and containing introns with a minimum length of 50 bp, were considered for the analysis. RNA-Seq reads were mapped on the reference genome using Bowtie(58) with parameters: -n 2 –k 2 --best. To filter the significant candidate features, we considered horizontal (along the gene) and vertical coverage (depth of reads) to be more than 80% and 4x, respectively. Measuring the rate of consensus observation within multiple samples studied provided theoretical support to the candidate features showing exon-skipping or intron-retention. For exons anticipated to show exon-skipping, the observation had a consensus from more than 20% and less than 80% samples. On the other hand, introns having a consensus observation of their retention from more than 20% samples were considered as true events.

Functional association studies were further performed on genes showing evidence for exon-skipping or intron-retention. Genes were clustered based on conditions used to prepare the RNA-Seq samples and gene ontology (GO) terms were assigned to the genes (wherever possible) using UniProt-GOA (http://www.ebi.ac.uk/GOA). Significance of these terms was interpreted by calculating the observed to expected ratio of their percent occurring enrichment.

Acknowledgements

The authors acknowledge Daniel Richter (Roscoff Culture Collection) for providing the pipeline to clean the MMETSP sequence libraries, and to Andrew Alverson (University of Arkansas), Neal Clarke (Yale University), Michael Melkonian (University of Koln), Gane Ka-Shu Wong (University of Alberta), Jun Yu (Beijing Institute of Genomics), Ramon Massana (ICM) and Patrick Wincker (Genoscope) for early access to additional transcriptome data used for compilation of the reference sequence library. Funding is acknowledged from the ERC Advanced Award ''Diatomite'' and from the Louis D Foundation to C.B., as well as the French Government ''Investissements d'Avenir'' programmes MEMO LIFE (ANR- 10-LABX-54) and PSL* Research University (ANR-11-IDEX-0001-02). C.B and A.E.A. acknowledge funding from the Gordon and Betty Moore Foundation. AR acknowledge MEMO-LIFE International PhD fellowship program. RD is funded by EMBO early career fellowship (ALTF 1124-2014).

Competing financial interests

The authors declare no competing interests.

References

1. Field, C.B., Behrenfeld, M.J., Randerson, J.T. and Falkowski, P. (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, **281**, 237-240.
2. Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., de Vargas, C., Bittner, L. *et al.* (2016) Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences of the United States of America*.
3. de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I. *et al.* (2015) Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**, 1261605.
4. Bowler, C., Vardi, A. and Allen, A.E. (2010) Oceanographic and biogeochemical insights from diatom genomes. *Ann Rev Mar Sci*, **2**, 333-365.
5. Kroger, N. and Poulsen, N. (2008) Diatoms-from cell wall biogenesis to nanotechnology. *Annu Rev Genet*, **42**, 83-107.

6. Mata, T.M., Martins, A.A. and Caetano, N.S. (2010) Microalgae for biodiesel production and other applications: A review. *Renew Sust Energ Rev*, **14**, 217-232.
7. Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P. *et al.* (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, **456**, 239-244.
8. Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M. *et al.* (2004) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science*, **306**, 79-86.
9. Moustafa, A., Beszteri, B., Maier, U.G., Bowler, C., Valentin, K. and Bhattacharya, D. (2009) Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science*, **324**, 1724-1726.
10. Ku, C., Nelson-Sathi, S., Roettger, M., Sousa, F.L., Lockhart, P.J., Bryant, D., Hazkani-Covo, E., McInerney, J.O., Landan, G. and Martin, W.F. (2015) Endosymbiotic origin and differential loss of eukaryotic genes. *Nature*, **524**, 427-432.
11. Deschamps, P. and Moreira, D. (2012) Reevaluating the green contribution to diatom genomes. *Genome biology and evolution*, **4**, 683-688.
12. Rastogi, A., Murik, O., Bowler, C. and Tirichine, L. (2016) PhytoCRISP-Ex: a web-based and stand-alone application to find specific target sequences for CRISPR/CAS editing. *BMC bioinformatics*, **17**, 261.
13. Nymark, M., Sharma, A.K., Sparstad, T., Bones, A.M. and Winge, P. (2016) A CRISPR/Cas9 system adapted for gene editing in marine algae. *Sci Rep*, **6**, 24951.
14. Allen, A.E., Dupont, C.L., Obornik, M., Horak, A., Nunes-Nesi, A., McCrow, J.P., Zheng, H., Johnson, D.A., Hu, H., Fernie, A.R. *et al.* (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature*, **473**, 203-207.
15. Morrissey, J., Sutak, R., Paz-Yepes, J., Tanaka, A., Moustafa, A., Veluchamy, A., Thomas, Y., Botebol, H., Bouget, F.Y., McQuaid, J.B. *et al.* (2015) A novel protein, ubiquitous in marine phytoplankton, concentrates iron at the cell surface and facilitates uptake. *Current biology : CB*, **25**, 364-371.
16. Allen, A.E., Laroche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P.J., Finazzi, G., Fernie, A.R. and Bowler, C. (2008) Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 10438-10443.
17. Huysman, M.J., Martens, C., Vandepoele, K., Gillard, J., Rayko, E., Heijde, M., Bowler, C., Inze, D., Van de Peer, Y., De Veylder, L. *et al.* (2010) Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome biology*, **11**, R17.
18. Xue, J., Niu, Y.F., Huang, T., Yang, W.D., Liu, J.S. and Li, H.Y. (2015) Genetic improvement of the microalga Phaeodactylum tricornutum for boosting neutral lipid accumulation. *Metab Eng*, **27**, 1-9.

19. Kaur, S. and Spillane, C. (2015) Reduction in carotenoid levels in the marine diatom Phaeodactylum tricornutum by artificial microRNAs targeted against the endogenous phytoene synthase gene. *Mar Biotechnol (NY)*, **17**, 1-7.
20. Fortunato, A.E., Jaubert, M., Enomoto, G., Bouly, J.P., Raniello, R., Thaler, M., Malviya, S., Bernardes, J.S., Rappaport, F., Gentili, B. *et al.* (2016) Diatom Phytochromes Reveal the Existence of Far-Red-Light-Based Sensing in the Ocean. *The Plant cell*, **28**, 616-628.
21. Maheswari, U., Jabbari, K., Petit, J.L., Porcel, B.M., Allen, A.E., Cadoret, J.P., De Martino, A., Heijde, M., Kaas, R., La Roche, J. *et al.* (2010) Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome biology*, **11**, R85.
22. Veluchamy, A., Rastogi, A., Lin, X., Lombard, B., Murik, O., Thomas, Y., Dingli, F., Rivarola, M., Ott, S., Liu, X. *et al.* (2015) An integrative analysis of post-translational histone modifications in the marine diatom Phaeodactylum tricornutum. *Genome biology*, **16**, 102.
23. Veluchamy, A., Lin, X., Maumus, F., Rivarola, M., Bhavsar, J., Creasy, T., O'Brien, K., Sengamalay, N.A., Tallon, L.J., Smith, A.D. *et al.* (2013) Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum. *Nat Commun*, **4**.
24. Kojima, K.K. and Jurka, J. (2011) Crypton transposons: identification of new diverse families and ancient domestication events. *Mobile DNA*, **2**, 12.
25. Song, Q.X., Lu, X., Li, Q.T., Chen, H., Hu, X.Y., Ma, B., Zhang, W.K., Chen, S.Y. and Zhang, J.S. (2013) Genome-wide analysis of DNA methylation in soybean. *Mol Plant*, **6**, 1961-1974.
26. Ito, H. and Kakutani, T. (2014) Control of transposable elements in Arabidopsis thaliana. *Chromosome Res*, **22**, 217-223.
27. Kim, E., Magen, A. and Ast, G. (2007) Different levels of alternative splicing among eukaryotes. *Nucleic acids research*, **35**, 125-131.
28. Bradnam, K.R. and Korf, I. (2008) Longer first introns are a general property of eukaryotic gene structure. *PloS one*, **3**, e3093.
29. Zhang, Q. and Edwards, S.V. (2012) The evolution of intron size in amniotes: a role for powered flight? *Genome biology and evolution*, **4**, 1033-1043.
30. Keren, H., Lev-Maor, G. and Ast, G. (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nature reviews. Genetics*, **11**, 345-355.
31. Maheswari, U., Mock, T., Armbrust, E.V. and Bowler, C. (2009) Update of the Diatom EST Database: a new tool for digital transcriptomics. *Nucleic acids research*, **37**, D1001-1005.
32. Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993) dbEST--database for "expressed sequence tags". *Nature genetics*, **4**, 332-333.
33. Dimmer, E.C., Huntley, R.P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M.J., Bely, B., Browne, P., Mun Chan, W., Eberhardt, R. *et al.* (2012) The UniProt-GO Annotation database in 2011. *Nucleic acids research*, **40**, D565-570.
34. Wu, T.D. and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873-881.

35. Ondov, B.D., Varadarajan, A., Passalacqua, K.D. and Bergman, N.H. (2008) Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. *Bioinformatics*, **24**, 2776-2777.
36. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, **7**, 562-578.
37. Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19 Suppl 2**, ii215-225.
38. Holt, C. and Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics*, **12**, 491.
39. Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic acids research*, **37**, D211-215.
40. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**, 27-30.
41. Gruber, A., Rocap, G., Kroth, P.G., Armbrust, E.V. and Mock, T. (2015) Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *Plant J*, **81**, 519-528.
42. Gschloessl, B., Guermeur, Y. and Cock, J.M. (2008) HECTAR: a method to predict subcellular targeting in heterokonts. *BMC bioinformatics*, **9**, 393.
43. Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340**, 783-795.
44. Huff, J.T. and Zilberman, D. (2014) Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell*, **156**, 1286-1297.
45. Yang, E.C., Boo, G.H., Kim, H.J., Cho, S.M., Boo, S.M., Andersen, R.A. and Yoon, H.S. (2012) Supermatrix data highlight the phylogenetic relationships of photosynthetic stramenopiles. *Protist*, **163**, 217-231.
46. Theriot, E.C., Ashworth, M., Ruck, E., Nakov, T. and Jansen, R.K. (2010) A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. *Plant Ecology and Evolution*, **143**, 278-296.
47. Adl, S.M., Simpson, A.G., Lane, C.E., Lukes, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V. *et al.* (2012) The revised classification of eukaryotes. *J Eukaryot Microbiol*, **59**, 429-493.
48. Yoon, H.S., Muller, K.M., Sheath, R.G., Ott, F.D. and Bhattacharya, D. (2006) Defining the major lineages of red algae (Rhodophyta). *Journal of Phycology*, **42**, 482-492.
49. Leliaert, F., Smith, D.R., Moreau, H., Herron, M.D., Verbruggen, H., Delwiche, C.F. and De Clerck, O. (2012) Phylogeny and Molecular Evolution of the Green Algae. *Critical Reviews in Plant Sciences*, **31**, 1-46.
50. Shalchian-Tabrizi, K., Brate, J., Logares, R., Klaveness, D., Berney, C. and Jakobsen, K.S. (2008) Diversification of unicellular eukaryotes: cryptomonad colonizations of marine and fresh waters inferred from revised 18S rRNA phylogeny. *Environmental Microbiology*, **10**, 2635-2644.

51. Simon, M., Lopez-Garcia, P., Moreira, D. and Jardillier, L. (2013) New haptophyte lineages and multiple independent colonizations of freshwater ecosystems. *Environ Microbiol Rep*, **5**, 322-332.
52. Bachvaroff, T.R., Gornik, S.G., Concepcion, G.T., Waller, R.F., Mendez, G.S., Lippmeier, J.C. and Delwiche, C.F. (2014) Dinoflagellate phylogeny revisited: using ribosomal proteins to resolve deep branching dinoflagellate clades. *Mol Phylogenet Evol*, **70**, 314-322.
53. Stiller, J.W., Huang, J., Ding, Q., Tian, J. and Goodwillie, C. (2009) Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? *BMC genomics*, **10**, 484.
54. Burki, F., Imanian, B., Hehenberger, E., Hirakawa, Y., Maruyama, S. and Keeling, P.J. (2014) Endosymbiotic gene transfer in tertiary plastid-containing dinoflagellates. *Eukaryot Cell*, **13**, 246-255.
55. Flutre, T., Duprat, E., Feuillet, C. and Quesneville, H. (2011) Considering transposable element diversification in de novo annotation approaches. *PloS one*, **6**, e16526.
56. Price, A.L., Jones, N.C. and Pevzner, P.A. (2005) De novo identification of repeat families in large genomes. *Bioinformatics*, **21 Suppl 1**, i351-358.
57. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, **110**, 462-467.
58. Langmead, B. (2010) Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, **Chapter 11**, Unit 11 17.

Figures

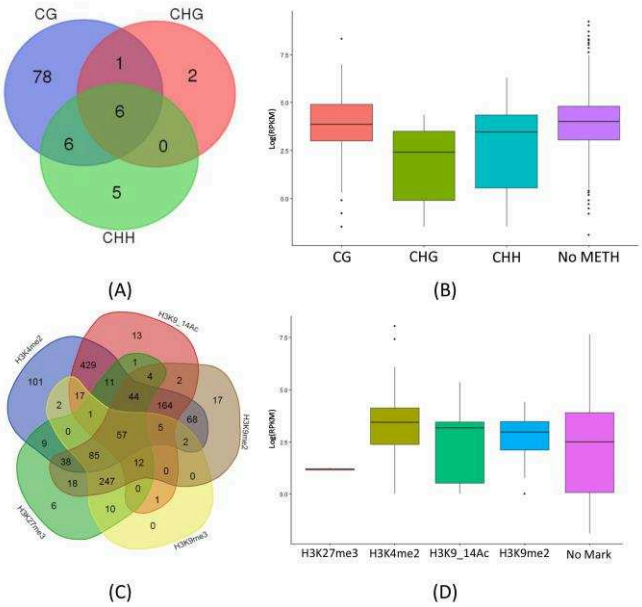


Figure 1. Distribution of epigenetic modifications. The Venn diagram depicts (A) the distribution of DNA methylation in different contexts (CG, CHG and CHH) along with expression of DNA methylated genes (B). Panel C shows the distribution of histone H3 post-translational modifications (PTM) over the set of new genes from Phatr3. The expression profiles of genes marked uniquely by the studied PTMs are represented as a box plot in D.

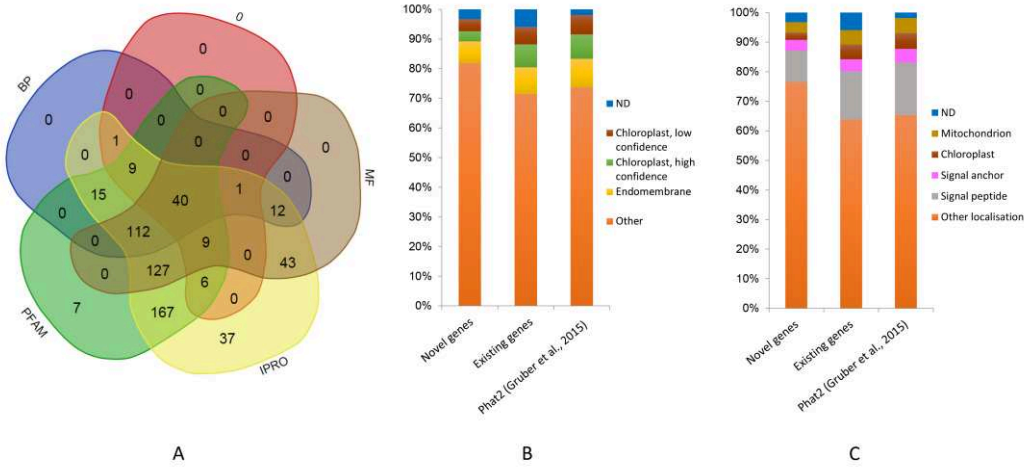


Figure 2. Dissecting the functional characteristics of *P. tricornutum* novel proteome (A) In-silico functional characterization of novel transcripts from Phatr3 was performed using the UniProt-GOA database. For each gene we mapped the associated biological process (BP), cellular component (CC) and molecular function (MF), along with the presence of functional protein domains using PFAM and

InterPro (IPRO) databases. The Venn diagram shows the number of genes found to have one or more of the functional variables. Figure (B) and (C) shows the targeting predictions for proteins encoded within the *P. tricornutum* genome as assessed using the diatom targeting predictor softwares, ASAFind (Gruber et al., 2015) and HECTAR (Gschoessl et al., 2008), respectively. Targeting predictions are shown (left to right) for: all new genes identified within the Phatr3 annotation, all Phatr3 genes for which a Phatr2 equivalent previously existed, and the modified version of Phatr2 (using all gene models, clustered to remove redundant sequences, and modified to be N-terminally complete) generated by Gruber et al. (2015) for exemplar profiling of the ASAFind predictor. “ND” implies a targeting prediction could not be obtained either because the protein in question was N-incomplete (i.e. did not begin with a methionine) or was too short to screen using ASAFind, or the Predotar module of HECTAR.

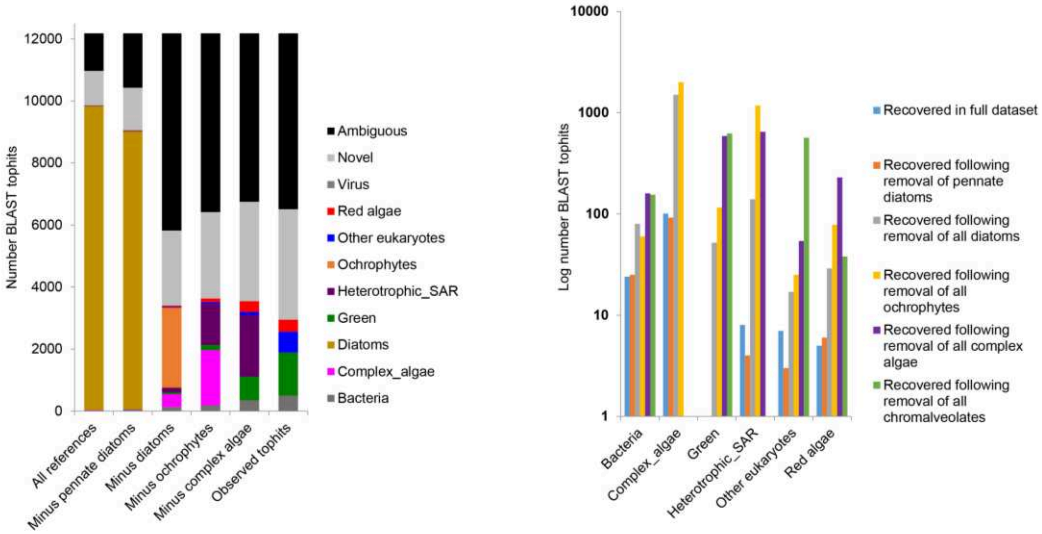


Figure 3. BLAST top hit analysis of *P. tricornutum* genes. The bar chart on the left shows the results of BLAST top hit analysis of the entire Phatr3 genome using the entire reference dataset, and five reference datasets generated by the progressive removal of five lineages (pennate diatoms, all diatoms, all ochrophytes, all complex algae, and all remaining SAR clade lineages). A phylogenetic affinity for each gene was only recorded if the top three hits obtained in the BLAST analysis originated from the same lineage. Genes were labelled as "ambiguous" if the top three BLAST hits contained sequences with multiple phylogenetic affinity, and "novel" if no hits were found with an expect value below 1E -05. The bar chart on the right compares the number of genes assigned to six different phylogenetic affinities (bacteria, complex algae, green algae and plants, heterotrophic SAR clade members red algae, and other eukaryotes) by BLAST analysis with each of the six different reference datasets used. The largest value recorded for each gene category corresponds to the most probable evolutionary time point at which genes were acquired; for example, the largest number of genes of red algal affinity were recovered following the removal of all complex algae from the reference dataset, indicating a large-scale donation of red algal genes into algae with complex plastids.

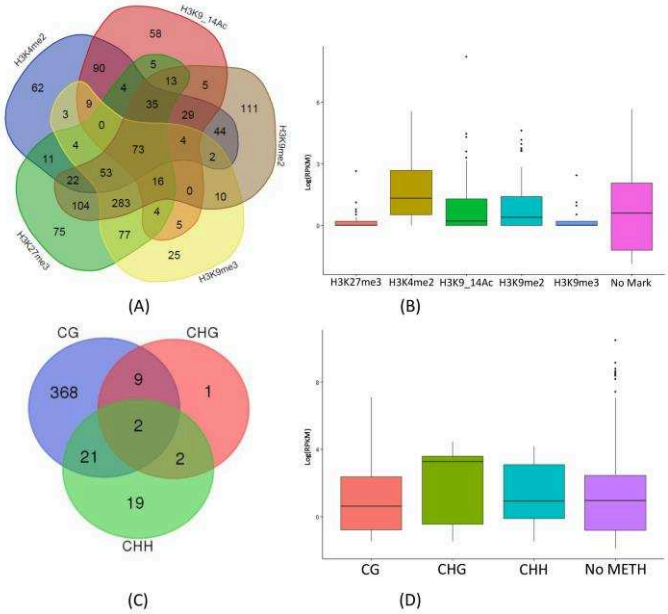


Figure 4. Distribution of epigenetic modifications over transposable elements. The Venn diagram presents (A) the number of new TEs found to be localized by one or more histone H3 PTMs, and (B) the number of new TEs found to be methylated in CG, CHH, and CHG contexts.

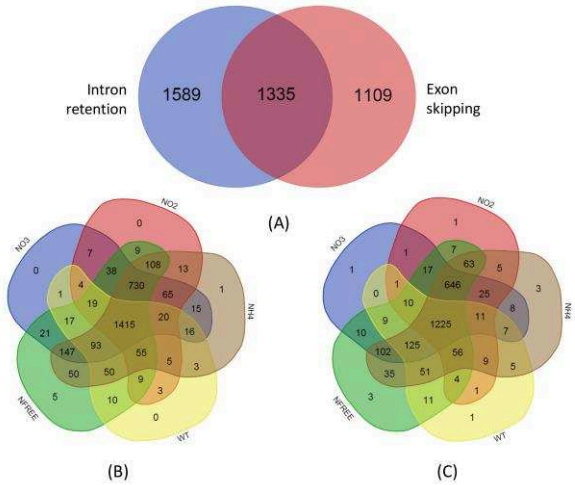


Figure 5. Identification of splice variants under different nutrient conditions. (A) The Venn diagram represents genes observed to either retain introns or to skip exons in multiple biological samples. (B) & (C) represent a Venn diagram analysis depicting genes that show intron-retention and exon-skipping under normal (WT) and various N-stress conditions (NFREE, NO3, NO2 and NH4), respectively.

Tables

Features	Phatr2	Phatr3
Number of genes	**10,402**	**12,177**
New gene models	-	1,489
Unchanged gene models	-	9,192
Merged gene models	-	236
Split gene models	-	1,260
Average gene length	1,474 (bp)	1,624 (bp)
Number of exons	18,552	20,885
Average exon length	770 (bp)	886 (bp)
Number of introns	8,058	10,932
Average intron length	963 (bp)	142 (bp)
Number of intergenic regions	5,157	5,542
Average length of intergenic region	1159 (bp)	307 (bp)
Completeness	83.4%	99.1%
Number of known protein domains	65,988	74,171
Number of genes with known protein domain	9,152	9,910

Table 1. Comparison of Phatr3 and Phatr2 annotations. The table presents a summary of Phatr3 and Phatr2 gene comparison statistics. Only filtered gene models have been considered in each case. The structural differences between Phatr2 and Phatr3 gene models have been classified into 4 major classes: Merged, New, Split, and Unchanged gene models. The number of genes in each category is given. Average length, number of exons, and average exon length is based on evaluation of all 12,177 genes. Completeness here refers to the percentage of gene models found to contain both start and stop codons.

Type	Class	Order	Superfamily	Coverage (bp)
Transposable Elements (TEs)	Class I	LTR retrotransposon	Copia	1, 764,927
			Gypsy	0
			DIRS	16,618
			Putative TRIM/LARD	7,517
		Non-LTR retrotransposon	LINE	0
			Putative SINE	1,467
	Class II	Subclass I	MuDR	56,587
			PiggyBac	75,644
			Other transposase	12,077
			Putative non-autonomous	189,491
		Subclass II	Crypton	16,569
	Putative TEs		Confused TE	232,238
Other	Undetermined		Unclassified repeats	400,050
	Tandem repeats		Tandem repeats	46,660
	Host genes		Putative host genes	244,861
			Total	**3, 064,706**

Table 2. Composition of repetitive sequence content.

Chapter 3

Ecotype diversity

During the time frame of approximately a century, 10 strains (referred to as ecotypes) of *Phaeodactylum tricornutum* have been sampled from different parts of the world. Although the sampling locations are widespread around the globe, their microenvironment is suggested to be highly unsteady where physiochemical conditions change very rapidly. Studies in the past have also suggested that these strains have distinct functional and morphological characteristics. However, so far no distinct evidence for their genetic diversity has been reported. In the current chapter, we used whole genome sequencing data of all the 10 strains to elucidate the genetic diversity that underlies distinct functional behavior of the ecotypes. Total diversity between the ecotypes is estimated via small polymorphisms and large structural variants and their specificity to individual ecotypes. Furthermore, using comparative and population genomics we suggest multiple sub-species within the 10 ecotypes of *P. tricornutum*. Finally based on multiple reports of hybridization in diatoms and other marine micro-eukaryotes like dinoflagellates, we propose natural hybridization as a prominent phenomenon in creating high estimated diversity within Bacillariophyceae.

Whole genome sequencing of *Phaeodactylum tricornutum* ecotypes reveals multiple sub-species as a consequence of ancient hybridization

Achal Rastogi[1], Anne-Flore Deton[1], Alaguraj Veluchamy[1,$], Catherine Cantrel[1], Gaohong Wang[2], Pieter Vanormelingen[3], Chris Bowler[1], Gwenael Piganeau[4,5], Leila Tirichine[1] and Hanhua Hu[2]

[1]Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France

[2]Key Laboratory of Algal Biology, Institute of Hydrobiology, Donghu south road, Wuchang district, Wuhan, Hubei Province, China

[3]Ghent University, Department of Biology, Research Group Protistology and Aquatic Ecology Krijgslaan 281/S8 9000 Gent, Belgium

[4]CNRS, UMR 7232, Observatoire Océanologique

[5]UPMC University Paris 06, Observatoire Océanologique, Sorbonne Universités

[$]Current address: Biological and Environmental Sciences and Engineering Division, Center for Desert Agriculture, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

Abstract

Diatoms are a highly diversified group of eukaryotic phytoplankton with estimates of up to 200,000 species. Multiple factors have been proposed to account for their current diversity patterns in the world's ocean. Natural hybridization has also been proposed to play a vital role in generating diversity in both terrestrial and aquatic organisms, including marine micro-eukaryotes such as diatoms and dinoflagellates. From decades, *Phaeodactylum tricornutum* is used as a model diatom species to characterize the functional physiology and evolution of diatoms in general. Since its discovery in 1897 by Bohlin, multiple *P. tricornutum* strains have been isolated from different geographic locations (referred to as ecotypes). These ecotypes have been shown to persist high diversity and explicit functional behavior in response to various environmental cues and suggest a distinctive genetic-makeup of the ecotypes. In the current study we have used whole genome sequencing data from ten ecotypes of the model diatom *Phaeodactylum tricornutum* to unveil the presence of a species-complex within the genus *Phaeodactylum* as a consequence of natural hybridization. In light of previous observations reporting natural hybrids within diatom species together with our current findings, we propose natural hybridization as a persistent phenomenon in shaping diatom evolution and diversity. We also attempted to understand the underlying functional consequences of the genetic diversity that persist between different sub-species of genus *Phaeodactylum*. Lastly, our work provides a firm genomic platform for the classification of proposed species-complex and/or, in general the genus *Phaeodactylum*, which was earlier shown to be contentious.

Introduction

Diatoms are unicellular, obligate photosynthetic eukaryotes. Ehrenberg first discovered them in the 19th century in dust samples collected by Charles Darwin in the Azores. They belong to a big group of heterokonts, constituent of chromalveolate (SAR group) which are believed to be derived from serial endosymbiosis combining genes from green and red algae predecessors (Bowler et al. 2008; Moustafa et al. 2009). According to earliest well-preserved fossil record, diatoms are believed to be in existence since 190 myr (Armbrust 2009) and their closest sister group are the Bolidomonads. Based on the characteristic symmetry and organization of their silicified valves, diatoms are divided into three major classes: Coscinodiscophyceae (referred to as centric diatoms; further subdivided into radial and polar centrics), Fragilariophyceae (araphid pennate diatoms) and Bacillariophyceae (raphid pennate diatoms) (Round et al. 1990). The pennates are believed to have evolved from the polar centric diatoms (Theriot et al. 2010) and are first found in the fossil record about 70 mya (Montsant et al. 2005).

Phaeodactylum tricornutum is a marine diploid raphid pennate diatom species. In nature, it is usually found under highly unstable coastal environments such as estuaries and rock pools (De Martino 2007). Since its discovery by Bohlin in 1897 and the characterization of different morphologies or morphotypes, denoted fusiform, triradiate and oval by Wilson in 1946, and recently, round and cruciform by (De Martino et al. 2011) and (He et al. 2014), 10 strains from 9 different geographic locations (sea shores, estuaries, rock pools, tidal creeks, etc.) around the world, from sub-polar to tropical latitudes, have been isolated (well described in (De Martino 2007)). These ecotypes have been collected within the time frame of approximately one century, from 1908 (Plymouth strain, Pt2/3) to 2000 (Chinese strain, Pt10) (De Martino 2007). All of the strains have been maintained either axenically or with their native bacterial population in different stock centers and cryopreserved after isolation (De Martino 2007). Based on sequence similarity of the ITS2 region within the 28S rDNA repeat sequence, previous reports have suggested the presence of four major ribotypes (Ribotype A: Pt1, Pt2, Pt3 and Pt9; Ribotype B: Pt4; Ribotype C: Pt5 and Pt10; Ribotype D: Pt6, Pt7 and Pt8) across the 10 ecotypes, with ribotype B and C being the most distant genetically (De Martino 2007). In 2008, Chris Bowler and colleagues (Bowler et al. 2008)

reported the sequencing and assembly of the genome of *P. tricornutum,* from a monoclonal culture of a single cell isolated from Pt1 (denoted Pt1 8.6) population. Using Sanger sequencing, de-novo assembly resulted in ~27.4 Mb genome with 33 chromosome scaffolds (12 of which were telomeric from end-to-end) and 55 scaffolds [designated bottom drawer (bd)]. This new resource further boosted *P. tricornutum* as a model diatom, and a range of methods are now available to manipulate and functionally characterize diatom genes, such as an extensively curated EST database, transformation methods, generic vectors for transgene expression, CRISPR, TALEN and RNAi (Falciatore et al. 1999; Siaut et al. 2007; De Riso et al. 2009; Maheswari et al. 2009; Huysman et al. 2010; Kaur and Spillane 2015; Nymark et al. 2016; Rastogi et al. 2016). These resources have also been used to explore the gene repertoire of *P. tricornutum* and other diatoms in an evolutionary context. We currently consider that diatom gene content is a chimera derived from endosymbiotic gene transfer from green and red algae as well as from the host cell, and from wholesale horizontal gene transfer from a wide range of prokaryotes (Bowler et al. 2008; Moustafa et al. 2009).

Multiple studies in past have reported distinct functional behavior of different ecotype strains as an adaptive response to various environmental cues, suggesting a distinct genetic makeup of different ecotypes a (Stanley 2007; Bailleul et al. 2010; Abida et al. 2015). The accumulated effect of diverse evolutionary and environmental forces such as recombination, mutation, selection, admixtures and introgression has been found to dictate the structure and diversity of genomes in a wide range of species (Liti et al. 2009; Cao et al. 2011; Flowers et al. 2015). Deciphering the existence and role of such processes in shaping the genetic diversity of *P. tricornutum* across different ecotype populations is an important first step to assess their role in shaping the evolution and adaptive capacities of diatoms. In order to understand the underlying genomic diversity within different ecotypes and to establish the functional implications of such diversity, we performed deep whole genome sequencing of all the ecotypes. Here we present the first genome wide diversity map and population genetic structure of 10 *P. tricornutum* ecotypes.

Results

Heterozygous alleles account for most of the genetic diversity within *P. tricornutum* ecotypes

We sequenced the genomes of 10 isolates of *P. tricornutum* and performed a reference-based assembly using the genome sequence of Pt1 8.6 (Bowler et al. 2008). Across all the ecotypes, the alignment depth ranged from 26X to 162X covering 92% to 98% of the genome (Table 1). We discovered 462,514 (depth >= 4x) single nucleotide variants (SNVs), including ~25% singleton SNVs, 573 insertions (of length 1 bp to 312 bp) and 1,801 deletions (of length 1 bp to 400 bp). Most of the SNVs are heterozygous and are shared between different ecotypes (Fig 1A & 1B). Most INDELS are also shared between different ecotypes, except for Pt4, which possesses the highest proportion of specific INDELS and SNVS (Fig 1B). The total number of heterozygous SNVS, along the whole genome, is estimated to range between 45% (Pt10) and 99% (Pt3). On average, 1 out of every 219 bp exhibits a heterozygous polymorphism. On the other hand, homozygous mutations were observed much less frequently (on average 1 every 8,314 bp) within ecotypes having high levels of heterozygosity (>90%; Pt1, Pt2, Pt3, Pt9) and more frequently (on average 1 out of 319 bp) in ecotypes where heterozygosity was comparatively low (<65%; Pt4, Pt5, Pt6, Pt7, Pt8, Pt10). With an average transition to transversion ratio of ~1.6, the spectrum of mutations across all the ecotypes reveals a higher rate of transitions over transversions. In total, six possible types of single nucleotide changes could be distinguished, among which, G:C -> A:T and A:T -> G:C, while maintaining balance between each other, accounted for more than ~60% of the observed mutations (Fig 1C).

Along with small variations, we also discovered large structural variants within different ecotypes. Using a normalized measure of read depth, we found that 259 and 588 genes, representing ~2% and 5% of the total gene content, had been lost or displayed copy number variation (CNV), respectively, across all the 10 ecotypes (Fig 1D). Using PCR, we experimentally validated some randomly chosen loci which are lost in different ecotypes (Figure S4). Approximately 70% of the genes that were either lost within ecotypes or present

in many copies are shared among multiple ecotypes. This indicates close association and functional behavior between different populations that could be because of their shared coastal macro-environment. Additionally, each ecotype has a specific molecular make-up, possibly linked to the explicit functional behavior of some ecotypes in response to various environmental cues as reported previously (Stanley 2007; Bailleul et al. 2010; Abida et al. 2015). Among all the enriched biological processes (chi-square test, $P<0.01$) that are associated to genes exhibiting ecotype-specific CNVs, a gene (Phatr3_EG02286) associated to nitrate assimilation, is observed to have high copy number (estimated read-depth is 4 folds higher compared to other genes and in other ecotypes) specifically in Pt4 population. Nitrate-assimilation is shown to regulate extensively under low light or dark conditions, to overcome nitrate limitation of growth in *Thalassiosira weissflogii* (Darren R. Clark 2002). Pt4 is well adapted to its low light ambient environment which can affect the nitrate assimilation capacity (Ivanikova Natalia V. 2005; Weiguo Li 2011) and so the dependent growth of the strain. Although it is not evident that low light corresponds to low nitrate assimilation and limited growth in Pt4 cultures, it seems Pt4 cells have attained a specific genetic makeup to acquire more nitrate and compensate low assimilation of nitrate due to low light environment.

Monoallelic gene expression contributes excessively to the genetic diversity in *P. tricornutum*

In the absence of clear characterization of sexual reproduction, the observation depicting high levels of heterozygosity within *P. tricornutum* can be explained as a phenomenon of introgression via cryptic speciation or sub-speciation within a given cell population. We examined the latter hypothesis by comparing the rate and specificity of heterozygous alleles between laboratory cultured natural sampled populations and monoclonal populations derived from a single cell. This would give us a fair idea if the heterozygosity is because of sequencing of two different very closely associated species together or if other evolutionary pressures are dictating the effect. We performed WGS of monoclonal populations derived from single cells of Pt1, Pt3 and Pt8 ecotypes, named as Pt1 8.6, Pt3Ov and Pt8Tc, respectively. We further compared the nature of all SNVs between Pt1 vs Pt1 8.6; Pt3 vs

Pt3Ov and Pt8 vs Pt8Tc (data not provided). On comparing the natural vs monoclonal cell populations, we found consensus nature of most polymorphic sites, across all the 3 comparisons (Pt1 vs Pt1 8.6; Pt3 vs Pt3Ov and Pt8 vs Pt8Tc). Most of the polymorphic sites (>90%) were observed to be heterozygous in both backgrounds (natural and monoclonal population, data not provided). The little difference in the nature of polymorphisms between natural and monoclonal population data can be because of the clonally grown cells. Clearly, the results reject our hypothesis of relating high levels of heterozygosity with introgression via speciation. However introgression can still persist via migration and ecological mixing of different natural strains. We further questioned the existence and role of high levels of heterozygosity within *P. tricornutum* genome. In order to do so, we performed RNA sequencing from Pt1 8.6, Pt3Ov and Pt8Tc and compared it to the corresponding WGS data. On comparing the SNP sites over the genes, we observed that the reasonable numbers of heterozygous alleles from the WGS data were not supported by SNP data from RNA sequencing (data not included), suggesting a strong influence of mono-allelic gene expression. Interestingly, the number of genes exhibiting monoallelic expression (MAE) is highly variable between the three ecotypes studied. With 3546 genes (~29% of the total genes) exhibiting MAE, which is higher than Humans (~19%) (Savova et al. 2016), MAE is more prevalent within Pt1 8.6 population, and reveals significant (chi-squared test; P-value < 0.02) top enrichment of biological processes like DNA metabolic process; DNA replication; nitrate assimilation; obsolete ATP catabolic process; purine nucleobase biosynthetic process. Whereas only 574 (~5%) and 1539 (~13%) genes exhibits MAE in Pt3Ov and Pt8Tc, respectively. Genes that exhibits MAE in Pt3Ov are enriched in processes like acetyl-CoA biosynthetic process; DNA integration; nitrate assimilation; organic acid metabolic process; oxidation-reduction process. While, genes exhibiting MAE in Pt8Tc are enriched in processes like, 'de novo' IMP biosynthetic process; DNA metabolic process; obsolete ATP catabolic process; purine nucleobase biosynthetic process; rRNA processing. Further, among the genes, which shows MAE in all the studied ecotypes (185 genes), biological processes like acetyl-CoA biosynthetic process; ammonium transmembrane transport; intracellular signal transduction; organic acid metabolic process and tRNA dihydrouridine synthesis, are among the top enriched processes. Although there is a noticeable difference between the numbers of genes showing MAE in different ecotype populations, genes with MAE are least expressed

in all the three ecotypes (Fig 2A, 2B and 2C). Moreover we also confirmed that even if the genes specifically demonstrate MAE in either of the three ecotypes, their expression remains low in all the ecotype populations (Fig 2D), which suggest that the signals of evolution within a population can only be achieved from genes that are moderately or lowly expressed (Pal et al. 2001; Dotsch et al. 2010) or can be said that the evolutionary forces are more prominent on moderate/low expressing genes. Our analyses of 3546 genes exhibiting MAE are significantly enriched with diatom or *P. tricornutum* specific genes. Apart from the genes that are monoallelicly regulated within Pt1 8.6, we interestingly found the isoforms (alternative spliced forms) of many genes to exhibit both monoallelic and biallelic expression (Figure 2E). This unexpected and interesting observation further adds to our knowledgebase the complex nature of expression control of genes within single cell organisms. Based on these preliminary results, the ongoing work focus on characterizing the role of mono vs biallelic gene expression in *P. tricornutum*. We are also interested in understanding the epigenetic control, if any, in the regulation of monoallelism including DNA methylation (Meaburn et al. 2010; Schalkwyk et al. 2010), and post translational modifications of histones as reported in human lymphoid cells (Nag et al. 2013) and murine cells (Nag et al. 2015).

Ribotype analysis reveals species-complex within genus *Phaeodactylum*

Localization of the polymorphic sites over genomic features (genes, transposable elements, and intergenic regions) revealed the highest densities of polymorphisms over genes (Fig 3E), specifically on exons, and was consistent across all the studied ecotype populations. An average non-synonymous to synonymous mutation ratio (N/S) was estimated to be ~0.87, which is higher than *Chlamydomonas reinhardtii*, N/S = 0.58 (Flowers et al. 2015). These observations, coupled with the high levels of genetic variation, and fast doubling time, indicate that *P. tricornutum* populations are likely able to adapt quickly to survive in a dynamically changing environment.

Although nothing is known about sexual reproduction in *P. tricornutum*, stress due to metabolically demanding and challenging niches have been shown to induce and promote

sexual reproduction (genetic exchange) in asexually reproducing microorganisms (Taylor et al. 1999; Grishkan et al. 2003; Schoustra et al. 2010). The phenomenon leads to the establishment of new and fit genotypes, as a consequence of recombination of advantageous mutations. Hybridization events leading to the existence of multiple ribotypes have been reported recently in three pennate diatom species, which are *Pseudo-nitzschia pungens (Casteleyn et al. 2009)*, *Pseudo-nitzschia multistriata* (D'Alelio et al. 2009), *and Fistulifera solaris* (Tanaka et al. 2015). Likewise, inter-specific hybridization based on rDNA markers has also been reported in dinoflagellates (Edvardsen 2003; Hart 2007). These findings have extended our perception of natural hybridization to persist not only in macro-eukaryotes (Casteleyn et al. 2009) but also in micro-eukaryotes. By comparing the sequences of the 18S and internal transcribed spacer 2 (ITS2) region of rDNA within the *P. tricornutum* genome, the ecotypes can be clustered into 4 ribotypes (Fig 3A and 3B): ribotype A (Pt1, Pt2, Pt3 and Pt9), ribotype B (Pt4), ribotype C (Pt5 and Pt10), and ribotype D (Pt6, Pt7 and Pt8). Inspection of the 18S rDNA gene sequence across ribogroups (ecotypes possessing the same ribotypes) indicated the presence of two different alleles in populations of ribogroup B, C and D (Fig 3B). Specifically, out of the two alleles observed among populations within ribogroups B, C and D, one of the alleles was specific to ribogroup A (Fig 3B), supporting the idea of ecological mixing via migration leading to interbreeding or genetic exchange between populations of this ribogroup with other ecotypes, consistent with three past hybridization events within the genus *Phaeodactylum*. However, it is non-evident that any such process has happened or if so, how? Moreover, *P. tricornutum* is a coastal species and apparently there are no evidences of its migration from one location to another. Though, dispersal of diatoms between localities has been previously shown to be limited to human activities (Vanormelingen 2007), and can account for the dispersal of *Phaeodactylum* species. Further, we confirmed the presence and expression of both the alleles using whole genome and total-RNA sequencing monoclonal population, grown using a single cell from the Pt8 population (referred to as Pt8Tc) (Fig 3B). Furthermore, topology of the whole genome tree generated by maximum likelihood algorithm using all polymorphic sites (SNVS and INDELS), across all the ecotypes, reflects the close association of ecotypes clustered within a ribogroup (Fig 3E).

Population genetics structure exhibits close association between different species of genus *Phaeodactylum*

Pairwise nucleotide diversity (π) estimated in non-overlapping 1 kb windows across all the ecotypes is quite low and ranges between 0.000005 and 0.0097 per site. A mean π of 0.002 ± 0.001 per site indicates that two homologous sequences taken at random across different populations will on average differ at ~0.2%, which is slightly lower than dimorphic fungi *Candida albicans* (Hirakawa et al. 2015) and much lower than green alga *Chlamydomonas reinhardtii (Flowers et al. 2015)*. High level of heterozygosity and low nucleotide diversity has been shown to be associated with high growth rate in *Candida albicans* (Hirakawa et al. 2015) and seems adequate in explaining high estimates of heterozygosity within fast reproducing (Fig S1) *Phaeodactylum* species. Moreover, heterozygosity may be selected under balancing selection (Sellis et al. 2011; Sellis et al. 2016). Not surprisingly, we observed low levels of heterozygosity and less alleles, genome-wide, that are deviated from Hardy-Weinberg equilibrium within populations that support the occurrence of hybridization (ribogroup B, C and D), compared to ribogroup A populations where no hybridization is observed. The linkage disequilibrium analysis using only homozygous SNP sites, revealed, on average, high linkage disequilibrium (LD > 0.7) over pairs of polymorphism, which is consistent with low levels of recombination. Linkage disequilibrium moderately declines with the increase in pairwise distance between associated alleles (Fig 3C). The Fixation Index (F_{ST}), studied among all possible pairs of populations ranges between 0.08 and 0.4 with a mean of 0.28 ± 0.1 (Fig 3D). These observations indicate a strong signal of convergent evolution between *Phaeodactylum* species/ribogroup populations. The latter can be conceptualized based on low nucleotide diversity, consequently low genetic differentiation (calculated as F_{ST}) (Fig 3D), and strong linkage disequilibrium between the populations belonging to different ribogroups. The closely associated behaviors of populations, at both inter and intra-ribogroup level, suggests hybridization to be an ancient event, happened in the life history of *Phaeodactylum* strains. It furthermore appears that the species complex is under strong convergent evolution where asexuality and heterozygosity may provide *Phaeodactylum* species better fitness to adapt rapidly to changing environments.

Functional characterization of the polymorphisms suggest change of selection pressures under laboratory conditions

Considering Darwinian selection theory, species are under continuous pressure to get fitter with time, in order to better survive. To understand the functional consequences of the polymorphisms driving selection of certain loci for better fitness, we identified genes within different ribogroups experiencing strong selection based on their high Ka/Ks (dN/dS) ratios. Across all the ecotypes, 100 such genes could be detected (Fig 4A), among which 47% are specific to one or other ribogroup. Furthermore, many genes (902) are found to have loss-of-function (LoF) alleles (Fig 4A), including frame-shift mutations and mutations leading to theoretical start/stop codon loss or gain of premature start/stop codon.

Based on the presence of functional domains, all *P. tricornutum* annotated genes (Phatr3, Chapter 2) were grouped into 3,020 gene families. These families can be as large as the reverse transcriptase gene family, which is also highly abundant in marine plankton (Lescot et al. 2016), representing 149 candidate genes having reverse transcriptase domain or as small as families that constitute single gene candidates. Across all the ecotypes we observed that the majority of genes experiencing LoF mutations belong to large gene families (Fig 4B). This is consistent with a previous observation of the existence of functional redundancy in gene families as a balancing mechanism for null mutations in yeast (Gu et al. 2003). To estimate an unbiased effect of any evolutionary pressure (LoF allele or Positive selection mutations) on different gene families, we calculated a ratio, named the effect ratio (EfR, see Methods), which normalizes the fact that if any gene family has enough candidates to buffer the effect on some genes influencing evolutionary pressure, it will be considered as being less affected compared to those where all or most of the constituents are under selection pressure. From the analysis each ribogroup revealed a specific set of gene families to be under selection pressure. Significantly enriched biological processes (chi-squared test; P-value<0.05) that are associated to ribogroup A specific gene families, which are under selection, revealed the top enrichment of chlorophyll biosynthetic process, fatty acid biosynthetic process, fructose 6-phosphate metabolic process and photosynthesis. Similarly, ribogroup B specific gene families that are under strong adaptive selection exhibits top

enrichment of processes like posttranslational protein targeting to membrane, translocation, protein complex assembly, RNA 3'end processing involving polyadenylation and terpenoid biosynthetic process. Ribogroup C gene families that are under selection exhibits top enrichment of processes like DNA repair, intracellular protein transport, nucleotide-excision repair, DNA-templated transcription and vesicle-mediated transport. Likewise, ribogroup D specific set of gene families that are under selection shows top enrichment of biological processes like arginyl-tRNA aminoacylation, intracellular signal transduction, lipid catabolic process and N-glycan processing. Apart from the ribogroup specific families that are under selection pressure, interestingly across all the ribogroups, a group of gene families associated to methionine biosynthesis is also observed as experiencing strong adaptive selection (Fig 4C).

In *P. tricornutum*, MetE (cobalamin-independent methionine synthase) and MetH (cobalamin-dependent methionine synthase) genes are known to metabolize cobalamin in the presence of symbiotic bacteria and vitamin B12, respectively. Previous reports have suggested that growing axenic cultures in conditions of high cobalamin availability results in repression, leading to the loss of MetE function and high expression of the *MetH* gene in *P. tricornutum* and *C. reinhardtii* (Helliwell et al. 2011; Bertrand et al. 2012; Helliwell et al. 2015). In accordance with these results, we observed a positive correlation between expression pattern of MetH and axenically grown cultures (Fig 4D) and thus a strong selection signal over the *MetH* gene (Fig 4C), which we speculate to be because of high availability of cobalamin in the laboratory growth media used to maintain all the ecotype strains over the last decades, suggesting that evolution can be triggered in laboratory culture conditions. However, we were not able to trace any significant signature for loss of the *MetE* gene although its expression is significantly lower in axenic cobalamin containing cultures, suggesting that its loss might require further generations. Similar observations were obtained, as previously reported (Bertrand et al. 2012), for CBA1 and SHMT genes, which under cobalamin scarcity enhances cobalamin acquisition and manage, reduced methionine synthase activity, respectively (Fig S2).

Considering all pairwise correlated gene families exhibiting similar selection signals among the 10 ecotypes, we used hierarchical clustering to examine the functional closeness of ecotype populations with one another. Consistent with the population and genetic structure, ecotypes within a ribogroup are more closely related than the ecotypes belonging to other ribogroups (Fig S3), suggesting variation in functional relatedness between different proposed groups of sub-species within the genus *Phaeodactylum*.

Discussion

Phaeodactylum tricornutum is a diploid diatom known to reproduce asexually although sexual reproduction is not excluded. Since the discovery of the species by Bohlin in 1897, many strains around the world have been sampled. These strains were isolated over approximately a time period of 100 years (De Martino 2007). Since the release of its genome in 2008, the species is used as a model for deciphering various metabolic pathways and unveiling the evolutionary history of diatoms (Bowler et al. 2008, (Huysman et al. 2013); Tanaka et al. 2015)(Allen et al. 2011; Morrissey et al. 2015; Fortunato et al. 2016). Along with sequencing of the genome, for most of the studies, Pt1 8.6, monoclonal population of a single cell derived from Pt1 population, is used as a reference. However, some studies have involved multiple ecotypes of *P. tricornutum* to compare its genetic and functional behavior in response to dynamically changing niches. From these studies, it appears that some ecotypes have different and unique features in response to various environmental cues. Like Pt3 population, which was derived originally from Pt2 population, shows the dominance of oval cells in the cultures (De Martino 2007); Pt4 population is known to be uniquely adapted for low light ambient environment displaying a reduced non photochemical quenching capacity (Bailleul et al. 2010); Oval cells from Pt5 population has maximum adherence to culturing flasks (Stanley 2007); Pt6 population has shown to accumulate high lipid content (specifically TAG) when grown in nitrate limiting conditions (Abida et al. 2015), and Pt8 population exhibits cells that are mostly triradiate in shape (De Martino 2007). Although the genetic basis of such diversity is unknown, attempts have been made to decipher the underlying genotype leading to different morphologies observed within different ecotypes. Sequencing of ITS2 region of the genome from all the ecotypes shows no correlation

between ecotypes and their genotypes (De Martino 2007). However, four ribotypes: A (Pt1, 2, 3 and 9), B (Pt4), C (Pt5 and Pt10) and D (Pt6, 7 and 8), were proposed based on the ITS2 sequences from all the ecotypes (De Martino 2007). In the current study, we sequenced the whole genome of all the 10 ecotypes to understand and compare their population genetics structure and genomic diversity. Genome wide diversity highlights high levels of heterozygosity within all ecotypes. Confirming the observations of the previous study, Pt4 population is the most genetically distinct strain with the highest total and specific numbers of SNVS and INDELS. Based on the presence of multiple variant positions within 18S and ITS2 molecular marker genes, we can cluster all ecotypes within 4 ribogroups, which is in consensus with previous report (De Martino 2007). The topology of association between the ribogroups was further confirmed to exhibit coherency at the whole genome scale with multiple genotypes. These levels of heterozygosity suggest frequent hybridization events within the evolution of *P. tricornutum*, and we propose the existence of a species-complex within the genus *Phaeodactylum*. In the absence of clear morphological evidences, a strong genetic elucidation supports the idea of speciation. Furthermore, in light of other reported hybrids within diatom species (Casteleyn et al. 2009; Casteleyn et al. 2010), current analysis strongly promotes the idea of possible genetic exchange between known non-sexually reproducing organisms. Although the mechanism and conditions that favor such genetic exchanges are not clear, population genetics structure of the ecotypes and the fact that there are no evidences of *P. tricornutum* presence in the open ocean indicates an ancient dispersal of the ancestors of ribogroup A populations to localities specific to ancestor of other ribogroup populations, most likely by human activities, rafting (Thiel 2005; Nikula et al. 2013), birds and water masses (Schlichting 1960; Proctor 1966; Foissner 2006). Further, with no clear characterization of gametes, phenomenon like introgression seems more compelling in describing genetic exchange between the migrated and already inhabited strains. Additionally, there is little evidence of recombination in the dataset suggesting that the *Phaeodactylum* species-complex is best fit for reproducing asexually making it more flexible to adapt in its dynamically changing niche by maintaining a heterozygous genetic makeup.

These species complexes is further supported by functional specialization of individual groups well illustrated with Pt4 in ribotype B which shows low non photochemical quenching

capacity (NPQ) (Bailleul et al. 2010) suggested as adaptive trait to low light condition at this latitude and most likely subsequent upregulation of a peculiar light harvesting protein LHCX4 in dark conditions (Bailleul et al. 2010; Taddei et al. 2016). In line with these observations, a gene involved in nitrate assimilation (Phatr3_EG02286) in Ribotype B shows high copy number suggesting a different mode of nutrient acquisition within this ribotype to back up autotrophy via photosynthesis in prolonged dark or low light conditions. In this case, osmotrophy or a different nutrient uptake mechanism can be a way in Pt4 (Ribotype B) to use dissolved amino acids or other sources of nitrogen. Similarly, another gene (Phatr3_J50146), an amino acid transporter is positively selected suggesting a role in nutrient uptake when mixotrophy is advantageous, most likely via a diffusion mode rather than pinocytosis, which is more energy consuming. Interestingly, a predicted seven transmembrane receptor belonging to rhodopsin gene family characterizes ribotype B. It is tempting to speculate on its role in light perception and photo-sensing in low light environments such as Scandinavian fjords. No distinction of specific genes is obvious in ribotype A. Ribotype C (Pt5 and Pt10) contains a protein with a role in adhesion among other functions, which is in line with high adherence reported in Pt5 (Stanley 2007). Additional functions emerge from this ribotype including vacuolar sorting and vesicle mediated transport which could be an indication of some intracellular motility (Pickett-Heaps and Forer 2001) useful for yet an unknown phenomenon or some vesicle transport outside the cell for cell-cell communication process. Ribotype D is largely enriched with proteins with reverse transcriptase domain suggesting a dominant role of retro elements in adaptation and evolution of this ribotype.

Methods

Sample preparation, sequencing and mapping

Ten different accessions of *P. tricornutum* were obtained from the culture collections of the Provasoli-Guillard National Center for Culture of Marine Phytoplankton (CCMP, Pt1=CCMP632, Pt5=CCMP630, Pt6=CCMP631, Pt7=CCMP1327, Pt9=CCMP633), the Culture Collection of Algae and Protozoa (CCAP, Pt2=CCAP 1052/1A, Pt3= CCAP 1052/1B, Pt4= CCAP 1052/6), the Canadian Center for the Culture of Microorganisms (CCCM, Pt8=NEPCC 640), and the Microalgae Culture Collection of Qingdao University (MACC, Pt10=MACC B228). All of the accessions were grown axenically in batch in flask cultures with a photon fluency rate of 75 µmol photons m-2 s-1 provided by cool-white fluorescent tubes in a 12:12 light: dark (L:D) photoperiod at 20 °C. Exponentially growing cells were harvested and total DNA was extracted with the cetyltrimethylammonium bromide (CTAB) method. At least 6 µg of genomic DNA from each accession was used to construct a sequencing library following the manufacturer's instructions (Illumina Inc.). Paired-end sequencing libraries with a read size of 100 bp and an insert size of approximately 400 bp were sequenced on an Illumina HiSeq 2000 sequencer at Berry Genomics Company (China). Low quality read-pairs were discarded using FASTQC with a read quality (Phred score) cutoff of 30. Using the genome assembly published in 2008 as reference (Bowler et al. 2008), we performed reference-assisted assembly of all the ecotypes. We used BOWTIE (-n 2 –X 400) for mapping the high quality NGS reads to the reference genome followed by the processing and filtering of the alignments using SAMTOOLS and BEDTOOLS.

Discovery of small polymorphisms and large structural variants

GATK (McKenna et al. 2010), configured for diploid genomes, was used for variant calling, which included single nucleotide polymorphisms (SNVs), small insertions and

deletions ranging between 1 and 300 base pairs (bp). The genotyping mode was kept default (genotyping mode = DISCOVERY), Emission confidence threshold (-stand_emit_conf) was kept 10 and calling confidence threshold (-stand_call_conf) was kept at 30. The minimum number of reads per base to be called as a high quality SNP was kept at 4 (i.e., read-depth >=4x).

Considering Z-score as a normalized measure of read-depth, gene candidates showing multiple copies (representing CNV) or apparently being lost (representing gene loss) were determined. The read-depth data was normalized to eliminate 1) the effect of variable size of the sequence libraries, 2) irregularity in sequencing regions with similar depth across the entire genome, and 3) variable coverage breadth (horizontal coverage) of a gene across different ecotypes. The horizontal coverage cut-off of each gene to be called as showing copy number variation, across all the ecotypes, was set to >=95%. Following the normalization step we calculated the fold-change between the average of normalized read depth (average Z-score) and normalized read depth per gene (Z-score per gene), per sample.

Population genetics structure

Population genetic parameters were estimated using Vcftools (Danecek et al. 2011) using SNP data. In the absence of data from individuals of each ecotype/sample, we assumed the behavior of each individual in a population to be coherent. Conclusively, instead of estimating the population genetic structure within an ecotype, we compared it across all the ecotypes.

For haplotype analysis, ITS2 gene (chr13: 42150-43145) and 18S gene (chr13: 43553-45338) were used. Polymorphic sites across all the ecotypes within ITS2 and 18S were called and the assembled sequences were aligned using CLUSTALW. The same approach was employed to perform diversity analysis at the whole genome scale. Later, a maximum likelihood algorithm was used to generate the whole genome tree with bootstrap values of 1,000. Further, we measured various population genetic functions to estimate the effect of evolutionary pressure in shaping the diversity and resemblance between different ecotype populations. We evaluated average R2 as a

function to measure the linkage disequilibrium with increasing distance (1 kb, 5 kb, 10 kb, 20 kb, 30 kb, 40 kb and 50 kb) between any given pair of mutant alleles across all the ecotypes using expectation-maximization (EM) algorithm. Nucleotide diversity (π) was estimated in a 1 kb non-overlapping window along the whole genome across all the ecotypes, using the method described by (Nei and Li 1979). Genetic differentiation or variability between the ecotypes was assessed using the mathematical value of Fixation index (FST), as described by Wright in 1931, as also stated in (Whitlock 1999; Rottenstreich et al. 2007). We estimated FST as a function to measure, mathematically, the similarity between different pairs of ecotypes sharing multiple SNP positions using the following formula, $FST = \frac{Hp-He}{Hp}$, where Hp and He represent the total number of polymorphic positions between any given pair of ecotypes and number of total polymorphic sites within an individual ecotype, respectively.

Functional characterization of polymorphisms

SNPEFF and KaKs calculator were used to annotate the functional nature of the polymorphisms. Along with the non-synonymous, synonymous, loss-of-function (LOF) alleles, transition to transversion ratio and mutational spectrum of the single nucleotide polymorphisms were also measured. Genes with Ka/Ks aka dN/dS ratio more than 1 with a p-value less than 0.05 are considered as undergoing natural or Darwinian selection. Various in-house scripts were also used at different levels for analysis and for plotting graphs. Data visualization and graphical analysis were performed principally using CIRCOS, CYTOSCAPE, IGV and R.

Association of the ecotypes

Based on the presence of functional domains all the genes were grouped into 3,020 gene families. Subsequently, the constituents of each gene family was checked for being affected by either loss-of-function mutations or experiencing natural selection. To estimate an unbiased effect of such phenomena over the gene families, a

normalized ratio named as effect ratio (EfR), was calculated. The ratio was estimated as shown below and gene families with EfR larger than 1 were considered as being significantly affected.

$$Effect\ Ratio\ (EfR) = \frac{\frac{Number\ of\ genes\ affected\ within\ the\ given\ gene\ family}{Total\ number\ of\ genes\ in\ the\ given\ gene\ family}}{\frac{Total\ number\ of\ genes\ affected\ in\ all\ the\ gene\ families}{Total\ number\ of\ genes\ in\ all\ the\ gene\ families}}$$

Later, considering gene family EfR as a function to measure the association rate, we deduced Pearson pairwise correlations between different ecotypes. The correlation matrix describes that if many equally affected gene families are shared between any given pair of ecotypes, they will have higher correlation compared to others. Finally, hierarchical clustering using Pearson pairwise correlation matrix assessed the association between the ecotypes.

Validation of gene loss and quantitative PCR analysis

In order to validate gene loss, DNA was extracted from all the ecotypes as described previously (Falciatore et al. 1999) and PCR was performed with the primers listed in Table S1. PCR products were loaded in 1% agarose gel and after migration, later, UV light and photographs taken using a gel documentation apparatus visualize presence and absence of amplified fragment. To assess gene expression, RNA was extracted as described in (Siaut et al. 2007) from ecotypes grown axenically in Artificial Sea Water (ASW) (Vartanian et al. 2009) supplemented with vitamins as well as in the presence of their endemic bacteria in ASW without vitamins. QPCR was performed as described previously (Siaut et al. 2007).

Conflict of interest

The authors declare no conflicts of interest

References

Abida H, Dolch LJ, Mei C, Villanova V, Conte M, Block MA, Finazzi G, Bastien O, Tirichine L, Bowler C et al. 2015. Membrane glycerolipid remodeling triggered by nitrogen and phosphorus starvation in Phaeodactylum tricornutum. *Plant physiology* **167**: 118-136.

Allen AE, Dupont CL, Obornik M, Horak A, Nunes-Nesi A, McCrow JP, Zheng H, Johnson DA, Hu H, Fernie AR et al. 2011. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**: 203-207.

Armbrust EV. 2009. The life of diatoms in the world's oceans. *Nature* **459**: 185-192.

Bailleul B, Rogato A, de Martino A, Coesel S, Cardol P, Bowler C, Falciatore A, Finazzi G. 2010. An atypical member of the light-harvesting complex stress-related protein family modulates diatom responses to light. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 18214-18219.

Bertrand EM, Allen AE, Dupont CL, Norden-Krichmar TM, Bai J, Valas RE, Saito MA. 2012. Influence of cobalamin scarcity on diatom molecular physiology and identification of a cobalamin acquisition protein. *Proceedings of the National Academy of Sciences of the United States of America* **109**: E1762-1771.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP et al. 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456**: 239-244.

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C et al. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature genetics* **43**: 956-963.

Casteleyn G, Adams NG, Vanormelingen P, Debeer AE, Sabbe K, Vyverman W. 2009. Natural hybrids in the marine diatom Pseudo-nitzschia pungens (Bacillariophyceae): genetic and morphological evidence. *Protist* **160**: 343-354.

Casteleyn G, Leliaert F, Backeljau T, Debeer AE, Kotaki Y, Rhodes L, Lundholm N, Sabbe K, Vyverman W. 2010. Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 12952-12957.

D'Alelio D, Amato A, Kooistra WH, Procaccini G, Casotti R, Montresor M. 2009. Internal transcribed spacer polymorphism in Pseudo-nitzschia multistriata (Bacillariophyceae) in the Gulf of Naples: recent divergence or intraspecific hybridization? *Protist* **160**: 9-20.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158.

Darren R. Clark KJF, Nicholas J.P. Owens. 2002. The large capacity for dark nitrate-assimilation in diatoms may overcome nitrate limitation of growth. *New Phytologist* doi:10.1046/j.1469-8137.2002.00435.x.

De Martino A, Bartual A, Willis A, Meichenin A, Villazan B, Maheswari U, Bowler C. 2011. Physiological and Molecular Evidence that Environmental Changes Elicit Morphological Interconversion in the Model Diatom Phaeodactylum tricornutum. *Protist* **162**: 462-481.

De Martino AM, A. Juan Shi, K.P. Bowler, C. 2007. Genetic and phenotypic characterization of Phaeodactylum tricornutum (Bacillariophyceae) accessions. *J Phycol* **43**: 992–1009.

De Riso V, Raniello R, Maumus F, Rogato A, Bowler C, Falciatore A. 2009. Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic acids research* **37**: e96.

Dotsch A, Klawonn F, Jarek M, Scharfe M, Blocker H, Haussler S. 2010. Evolutionary conservation of essential and highly expressed genes in Pseudomonas aeruginosa. *BMC genomics* **11**: 234.

Edvardsen BS-T, Kamran; Jakobsen, Kjetill S.; Medlin, Linda K.; Dahl, Einar; Brubak, Sissel; Paasche, Eystein. 2003. GENETIC VARIABILITY AND MOLECULAR PHYLOGENY OF DINOPHYSIS SPECIES (DINOPHYCEAE) FROM NORWEGIAN WATERS INFERRED FROM SINGLE CELL ANALYSES OF rDNA. *Journal of Phycology* doi:10.1046/j.1529-8817.2003.01252.x.

Falciatore A, Casotti R, Leblanc C, Abrescia C, Bowler C. 1999. Transformation of Nonselectable Reporter Genes in Marine Diatoms. *Mar Biotechnol (NY)* **1**: 239-251.

Flowers JM, Hazzouri KM, Pham GM, Rosas U, Bahmani T, Khraiwesh B, Nelson DR, Jijakli K, Abdrabu R, Harris EH et al. 2015. Whole-Genome Resequencing Reveals Extensive Natural Variation in the Model Green Alga Chlamydomonas reinhardtii. *The Plant cell* **27**: 2353-2369.

Foissner W. 2006. Biogeography and Dispersal of Micro-organisms: A Review Emphasizing Protists. *Acta Protozool* **45**: 111-136.

Fortunato AE, Jaubert M, Enomoto G, Bouly JP, Raniello R, Thaler M, Malviya S, Bernardes JS, Rappaport F, Gentili B et al. 2016. Diatom Phytochromes Reveal the Existence of Far-Red-Light-Based Sensing in the Ocean. *The Plant cell* **28**: 616-628.

Grishkan I, Korol AB, Nevo E, Wasser SP. 2003. Ecological stress and sex evolution in soil microfungi. *Proc Biol Sci* **270**: 13-18.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**: 63-66.

Hart MCG, David H.; Bresnan, Eileen; Bolch, Christopher J. 2007. Large subunit ribosomal RNA gene variation and sequence heterogeneity of Dinophysis (Dinophyceae) species from Scottish coastal waters. *Harmful Algae* doi:10.1016/j.hal.2006.10.001.

He L, Han X, Yu Z. 2014. A rare Phaeodactylum tricornutum cruciform morphotype: culture conditions, transformation and unique fatty acid characteristics. *PloS one* **9**: e93922.

Helliwell KE, Collins S, Kazamia E, Purton S, Wheeler GL, Smith AG. 2015. Fundamental shift in vitamin B12 eco-physiology of a model alga demonstrated by experimental evolution. *ISME J* **9**: 1446-1455.

Helliwell KE, Wheeler GL, Leptos KC, Goldstein RE, Smith AG. 2011. Insights into the evolution of vitamin B12 auxotrophy from sequenced algal genomes. *Mol Biol Evol* **28**: 2921-2933.

Hirakawa MP, Martinez DA, Sakthikumar S, Anderson MZ, Berlin A, Gujja S, Zeng Q, Zisson E, Wang JM, Greenberg JM et al. 2015. Genetic and phenotypic intra-species variation in Candida albicans. *Genome Res* **25**: 413-425.

Huysman MJ, Fortunato AE, Matthijs M, Costa BS, Vanderhaeghen R, Van den Daele H, Sachse M, Inze D, Bowler C, Kroth PG et al. 2013. AUREOCHROME1a-mediated induction of the diatom-specific cyclin dsCYC2 controls the onset of cell division in diatoms (Phaeodactylum tricornutum). *The Plant cell* **25**: 215-228.

Huysman MJ, Martens C, Vandepoele K, Gillard J, Rayko E, Heijde M, Bowler C, Inze D, Van de Peer Y, De Veylder L et al. 2010. Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome biology* **11**: R17.

Ivanikova Natalia V. RMLM, and George S. Bullerjahn. 2005. Construction and characterization of a cyanobacterial bioreporter

capable of assessing nitrate assimilatory capacity in freshwaters. *Limnology and Oceanography* **3**: 86-93.

Kaur S, Spillane C. 2015. Reduction in carotenoid levels in the marine diatom Phaeodactylum tricornutum by artificial microRNAs targeted against the endogenous phytoene synthase gene. *Mar Biotechnol (NY)* **17**: 1-7.

Lescot M, Hingamp P, Kojima KK, Villar E, Romac S, Veluchamy A, Boccara M, Jaillon O, Iudicone D, Bowler C et al. 2016. Reverse transcriptase genes are highly abundant and transcriptionally active in marine plankton assemblages. *ISME J* **10**: 1134-1146.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458**: 337-341.

Maheswari U, Mock T, Armbrust EV, Bowler C. 2009. Update of the Diatom EST Database: a new tool for digital transcriptomics. *Nucleic acids research* **37**: D1001-1005.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Meaburn EL, Schalkwyk LC, Mill J. 2010. Allele-specific methylation in the human genome: implications for genetic studies of complex disease. *Epigenetics* **5**: 578-582.

Montsant A, Jabbari K, Maheswari U, Bowler C. 2005. Comparative genomics of the pennate diatom Phaeodactylum tricornutum. *Plant physiology* **137**: 500-513.

Morrissey J, Sutak R, Paz-Yepes J, Tanaka A, Moustafa A, Veluchamy A, Thomas Y, Botebol H, Bouget FY, McQuaid JB et al. 2015. A novel protein, ubiquitous in marine phytoplankton, concentrates iron at the cell surface and facilitates uptake. *Current biology : CB* **25**: 364-371.

Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D. 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324**: 1724-1726.

Nag A, Savova V, Fung HL, Miron A, Yuan GC, Zhang K, Gimelbrant AA. 2013. Chromatin signature of widespread monoallelic expression. *Elife* **2**: e01256.

Nag A, Vigneau S, Savova V, Zwemer LM, Gimelbrant AA. 2015. Chromatin Signature Identifies Monoallelic Gene Expression Across Mammalian Cell Types. *G3 (Bethesda)* **5**: 1713-1720.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* **76**: 5269-5273.

Nikula R, Spencer HG, Waters JM. 2013. Passive rafting is a powerful driver of transoceanic gene flow. *Biol Lett* **9**: 20120821.

Nymark M, Sharma AK, Sparstad T, Bones AM, Winge P. 2016. A CRISPR/Cas9 system adapted for gene editing in marine algae. *Sci Rep* **6**: 24951.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* **158**: 927-931.

Pickett-Heaps JD, Forer A. 2001. Pac-Man does not resolve the enduring problem of anaphase chromosome movement. *Protoplasma* **215**: 16-20.

Proctor VW. 1966. Dispersal of Desmids by birds. *Phycologia* **5**: 227-232.

Rastogi A, Murik O, Bowler C, Tirichine L. 2016. PhytoCRISP-Ex: a web-based and stand-alone application to find specific target sequences for CRISPR/CAS editing. *BMC bioinformatics* **17**: 261.

Rottenstreich S, Hamilton MB, Miller JR. 2007. Dynamics of Fst for the island model. *Theor Popul Biol* **72**: 485-503.

Round FE, Crawford RM, Mann DG. 1990. *The Diatoms: Biology and Morphology of the Genera*. Cambridge University Press, London, UK.

Savova V, Chun S, Sohail M, McCole RB, Witwicki R, Gai L, Lenz TL, Wu CT, Sunyaev SR, Gimelbrant AA. 2016. Genes with monoallelic expression contribute disproportionately to genetic diversity in humans. *Nature genetics* **48**: 231-237.

Schalkwyk LC, Meaburn EL, Smith R, Dempster EL, Jeffries AR, Davies MN, Plomin R, Mill J. 2010. Allelic skewing of DNA methylation is widespread across the genome. *American journal of human genetics* **86**: 196-212.

Schlichting HE. 1960. The rôle of waterfowl in the dispersal of algae. *Trans Am Microsc*

Soc **79**: 160-166.

Schoustra S, Rundle HD, Dali R, Kassen R. 2010. Fitness-associated sexual reproduction in a filamentous fungus. *Current biology : CB* **20**: 1350-1355.

Sellis D, Callahan BJ, Petrov DA, Messer PW. 2011. Heterozygote advantage as a natural consequence of adaptation in diploids. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 20666-20671.

Sellis D, Kvitek DJ, Dunn B, Sherlock G, Petrov DA. 2016. Heterozygote Advantage Is a Common Outcome of Adaptation in Saccharomyces cerevisiae. *Genetics* **203**: 1401-1413.

Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falciatore A, Bowler C. 2007. Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene* **406**: 23-35.

Stanley JAC. 2007. Whole cell adhesion strength of morphotypes and isolates of Phaeodactylum tricornutum (Bacillariophyceae). *European Journal of Phycology* doi:10.1080/09670260701240863.

Taddei L, Stella GR, Rogato A, Bailleul B, Fortunato AE, Annunziata R, Sanges R, Thaler M, Lepetit B, Lavaud J et al. 2016. Multisignal control of expression of the LHCX protein family in the marine diatom Phaeodactylum tricornutum. *J Exp Bot* **67**: 3939-3951.

Tanaka T, Maeda Y, Veluchamy A, Tanaka M, Abida H, Marechal E, Bowler C, Muto M, Sunaga Y, Tanaka M et al. 2015. Oil accumulation by the oleaginous diatom Fistulifera solaris as revealed by the genome and transcriptome. *The Plant cell* **27**: 162-176.

Taylor J, Jacobson D, Fisher M. 1999. THE EVOLUTION OF ASEXUAL FUNGI: Reproduction, Speciation and Classification. *Annu Rev Phytopathol* **37**: 197-246.

Theriot EC, Ashworth M, Ruck E, Nakov T, Jansen RK. 2010. A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. *Plant Ecology and Evolution* **143**: 278-296.

Thiel MaG, L. 2005. The ecology of rafting in the marine environment. II. The rafting organisms and community. *Oceanography and Marine Biology: An Annual Review* **43**: 279-418.

Vanormelingen PV, Elie; Vyverman, Wim 2007. The diversity and distribution of diatoms: from cosmopolitanism to narrow endemism. *Biodiversity and Conservation* **17**: 393-405.

Vartanian M, Descles J, Quinet M, Douady S, Lopez PJ. 2009. Plasticity and robustness of pattern formation in the model diatom Phaeodactylum tricornutum. *The New phytologist* **182**: 429-442.

Weiguo Li JW. 2011. Influence of light and nitrate assimilation on the growth strategy in clonal weed Eichhornia crassipes. *Aquatic Ecology* doi:10.1007/s10452-010-9318-8.

Whitlock MC. 1999. Neutral additive genetic variance in a metapopulation. *Genet Res* **74**: 215-221.

Figures and Tables

Library Name	Origin	Year of Isolation	Mapped Read-Pairs	Alignment Depth (X)	Genome Coverage (%)
Pt1	Blackpool, UK	1956	3, 642,044	26.5	98.0
Pt2	Plymouth, UK	Prior to 1910	6, 016,241	43.8	98.0
Pt3	Plymouth, UK	1930s	6, 373,591	46.4	98.3
Pt3Ov	Plymouth, UK	2012	17,694,466	131.2	98.4
Pt4	Finland	1951	15, 583,665	113.5	94.0
Pt5	West Dennis, MA, USA	1972	5, 346,009	38.9	93.2
Pt6	MA, USA	1956	3, 922,830	28.5	94.1
Pt7	Long Island, NY, USA	1952	4, 937,516	35.9	94.9
Pt8	Vancouver, Canada	1987	22, 235,170	162.1	94.4
Pt8Tc	Vancouver, Canada	2012	9,478,820	73.2	94.5
Pt9	Guam, Micronesia	1981	7, 551,099	55.2	97.5
Pt10	Dalian, China	2000	5, 436,057	39.6	92.1

Table 1. Reference-assisted mapping statistics. The table summarizes the origin and year of sampling of each isolate of P. tricornutum along with the number of total reads mapped on the reference. Average depth (X=average number of reads aligned on each base covered across the entire genome) was estimated using the number of mapped read pairs and the horizontal coverage (aka. coverage breadth) across the whole genome.

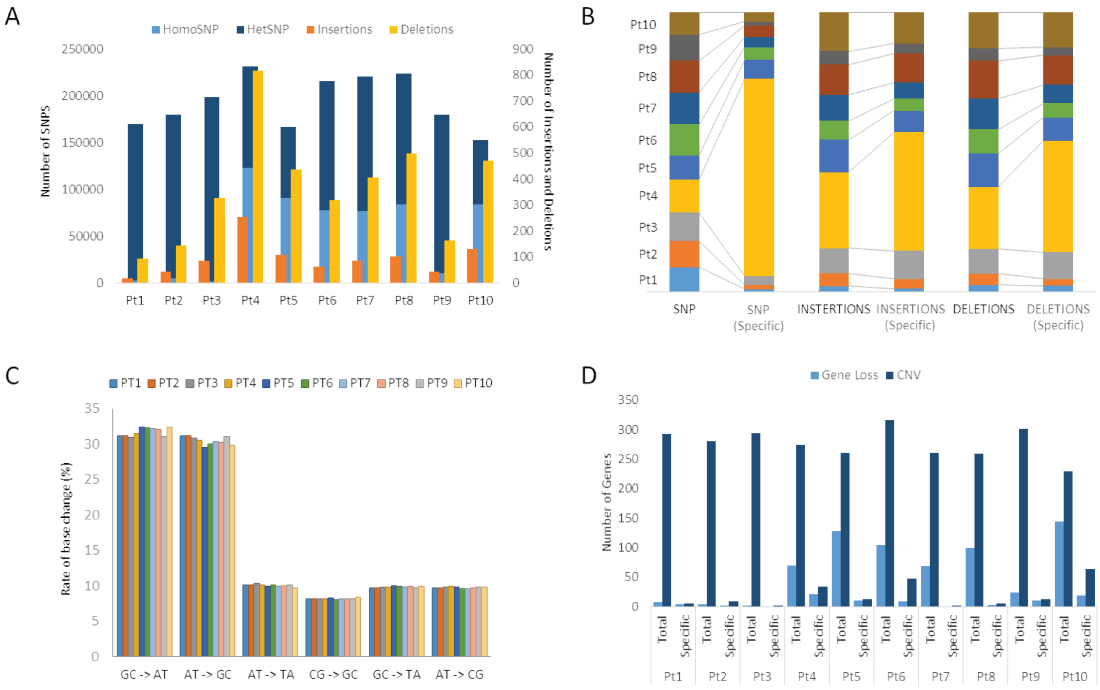


Figure 1. Global diversity patterns of 10 sequenced strains of P. tricornutum. (A) The bar plot represents the number of polymorphic sites discovered across all the ecotypes. The left and right Y-axis denotes the number of SNP sites and number of INDELS, respectively. (B) The stack bar plot represents the proportion of total (total number of variations in corresponding ecotype) vs specific polymorphic sites (number of variations specific to corresponding ecotype out of the total variations and not shared among other ecotypes), including SNPs, insertions and deletions (from Left to Right, respectively) across all the ecotypes. Colors representing the ecotypes are chosen randomly and have no biological significance. (C) The bar plots represents the mutational spectrum of all the SNVS discovered across the ecotypes. Y-axis denotes the total percentage of individual mutations observed as denoted on X-axis. (D) The bar plots represent the total and specific number of genes, denoted on Y-axis that exhibits a loss or multiple copies (signifying CNV) within individual ecotype strains. A Gene is deemed to exhibit CNV if the normalized read depth (z-score) covering atleast 95% (horizontal coverage) of the gene is equal or more than 4 folds higher than the median depth of all the genes. Total and specific on the X-axis corresponding to each ecotype refers to the total number of genes showing CNV or loss in the corresponding ecotype and specific number of genes out of total which are specific to that ecotype and not shared across other ecotypes.

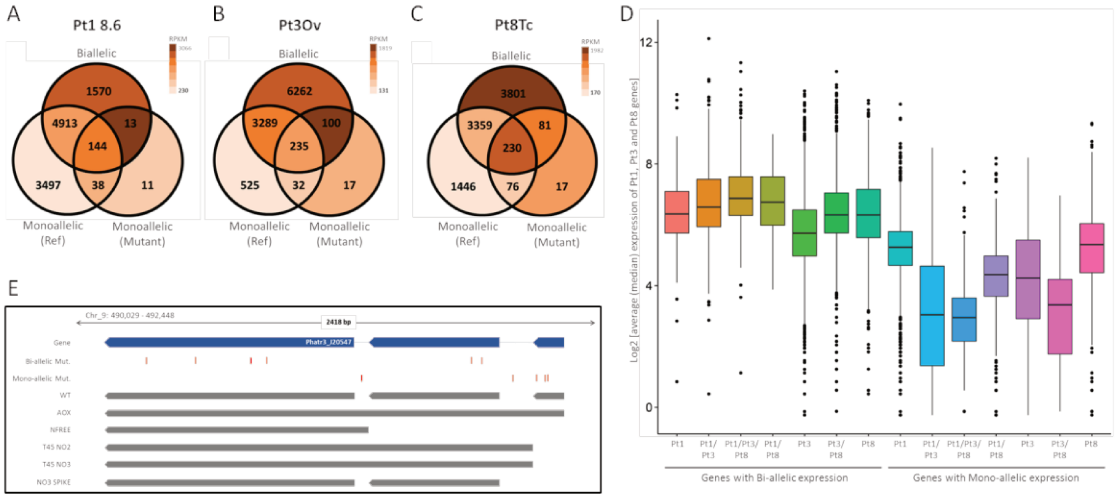


Figure 2. Characteristics of genes exhibiting mono-allelic expression. Venn analysis performed with genes exhibiting mono-allelic expression (RPKM) in all the three studied ecotypes or uniquely in Pt1 8.6, Pt8Tc and Pt3Ov (A, B and C). (D) The boxplot represents the expression of genes, within each ecotype or in all the ecotype, showing bi or mono-allelic expression. Values on the Y-axis are the average (median) expression of genes in all the three studied ecotypes. Monoallelic (Ref) indicates the expression of reference allele while Monoallelic (Mut) indicates the expression of allele with mutations. (E) A snapshot of a gene displays the mode of gene-expression i.e. mono-allelic or bi-allelic along with the structure of various isoforms expressed in different conditions as mentioned on Y-axis. WT refers to wild type or normal condition, AOX refers to the alternative oxidase mutant line, NFREE refers to cell cultures maintained under nitrogen free condition. T45 NO2 and T45 NO3 refer to cells taken after 45 min. from cultures maintained under NO2 and NO3 presence. NO3 SPIKE refers to cells taken from cultures with a spike treatment of NO3.

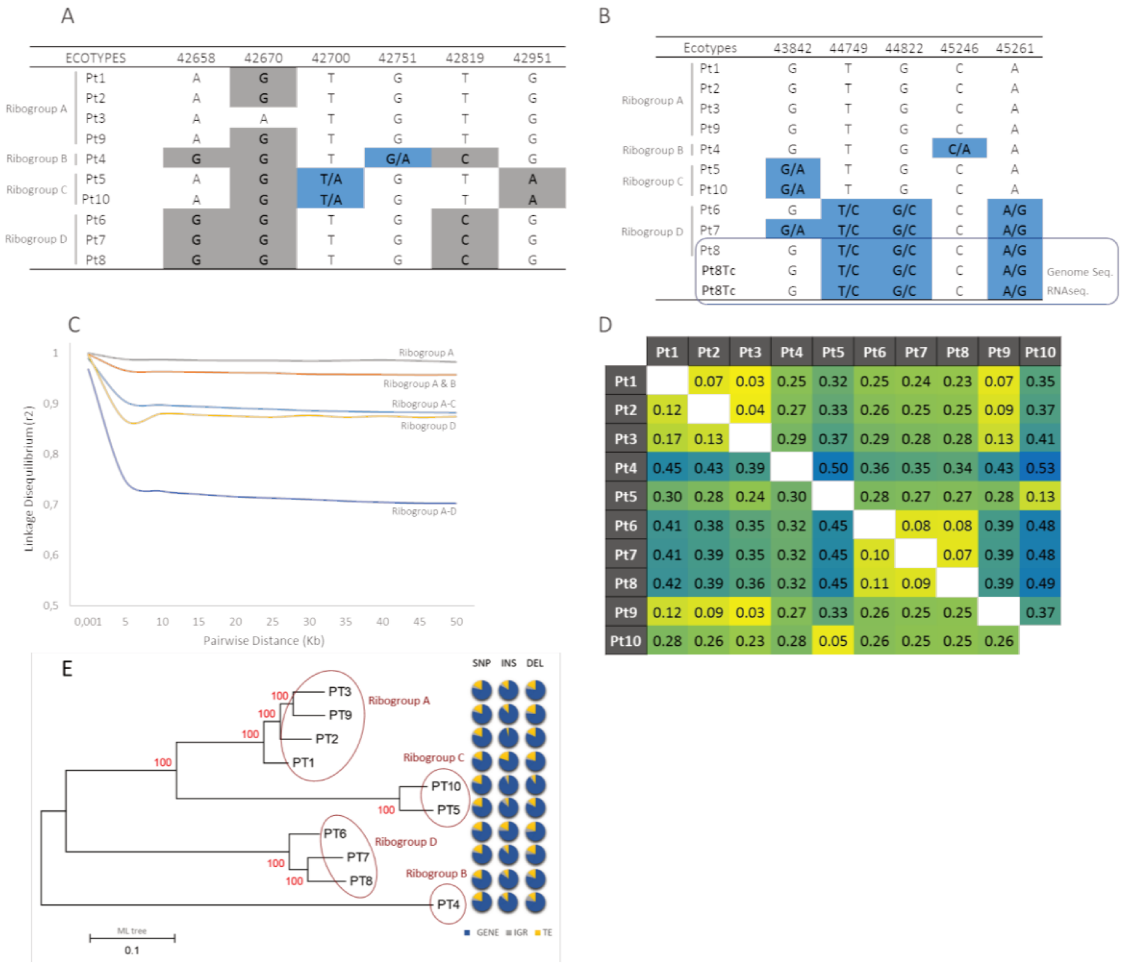


Figure 3. Population genomics reveals species-complex within genus Phaeodactylum. Haplotype analysis using ITS2 and 18S gene locus revealed the presence of 4 ribotypes as indicated in (A) and (B) respectively. Homozygous mutations are shown in grey, while heterozygous mutations are shown in blue with all possible alleles. (B) Last two lines indicate the confirmation of the presence of all the heterozygous alleles (both the alleles at a given position within 18S gene, indicated on the top Y-axis) within monoclonal population referred to as Pt8Tc. We verified the presence at both genomic and transcript level. (C) The line plot indicates the linkage disequilibrium (LD) decay (r2) across all the ecotypes with the increase of pairwise distance between any given pair of mutated allele. (D) The matrix measures the genetic differentiation or association between all possible pairs of ecotypes. The numbers in the matrix indicate the Fst values which ranges from 0 to 1 with a color gradient from yellow to blue, respectively. Values closer to 0 signify high genetic exchange and 1 indicates no exchange between the populations. (E) Phylogenetic association of the ecotypes based on 468,188 polymorphic sites (including SNP and INDELS) genome-wide, using maximum likelihood approach. Numbers on the branches indicate the bootstrap values. Pie chart adjacent to each node corresponds to proportion of SNVs and INDELs over all functional features of the genome; GENEs

(blue), TEs (Transposable elements, represented in yellow), IGRs (Intergenic regions, represented in grey).

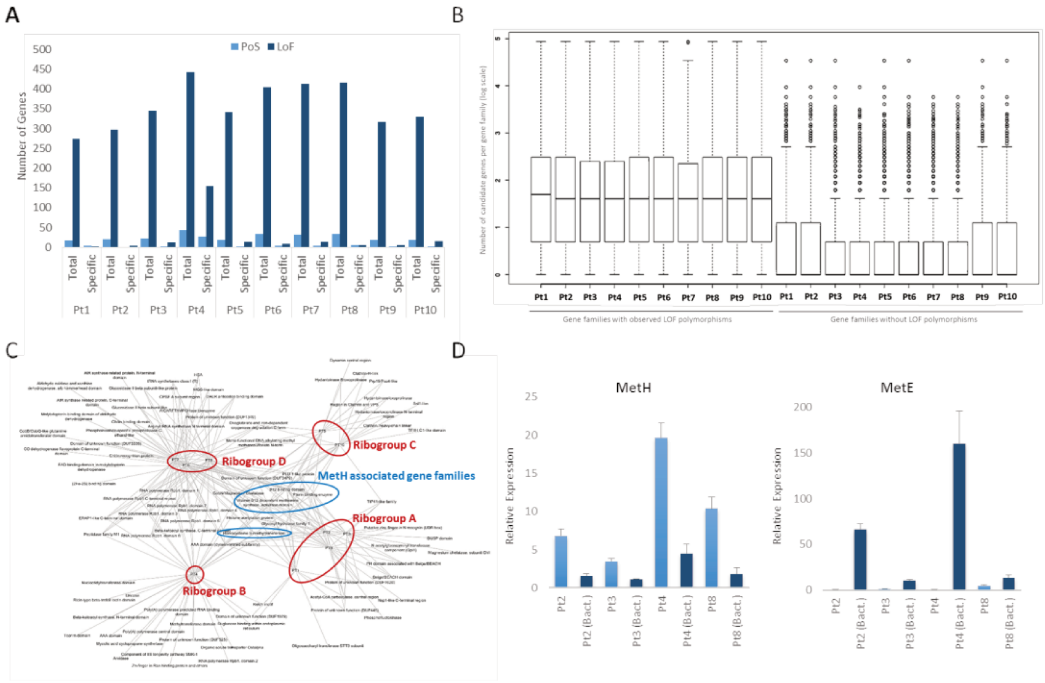


Figure 4. Evolutionary and functional consequences of polymorphisms. (A) The bar plot represents total vs specific number of genes that are subject to positive selection, or experiencing loss-of-function (LoF) mutations. For each category the ecotypes are plotted with total and specific number of genes. Y-axis denote the number of genes. (B) The box plot represents the number of gene families being affected by loss-of-function (LOF) mutations and suggests a bias of such mutations on the genes belonging to large gene families. Y-axis represent, as log scale, the number of genes in the gene families vs those that are not affected by LOF mutations. (C) Based on the EfR, the network displays highly affected gene families because of their constituents experiencing natural or Darwinian selection. Gene families associated to MetH genes under positive selection in all the ecotypes are indicated within blue circle (D) The bar plot represents relative expression of MetH and MetE gene in four (Pt2, Pt3, Pt4 and Pt8) of the ten ecotypes with the presence of vitamin B12 (light blue bars) and bacteria in the growing media (dark blue bars).

Chapter 4

Chromatin and Morphogenesis

Epigenetic modifications like tri-methylation at lysine 27 of histone H3 protein (H3K27me3) is well characterized as an essential component in development and in dictating cell fate decisions. In-silico identification of enhancer of zeste (E(z)) gene, a methyltransferase that transfer methyl group to lysine 27 of histone H3 and a central component of polycomb repressive complex proteins (PRC2), in *P. tricornutum* along in several other algae indicates an ancient origin of PRC2 and raises the question of its role in single-celled species. In the current chapter, we analyzed the level of natural variation in H3K27me3 profile of two populations, PT1 8.6 and PT8Tc representing two different morphotypes, fusiform and triradiate, respectively. Using chromatin Immunoprecipitation (ChIP) followed by sequencing from fusiform (PT1 8.6) and triradiate (PT8Tc) single cell monoclonal populations, we have established different H3K27me3 enrichment profiles of fusiform and triradiate populations, identifying putative targets of H3K27me3 involved in morphogenesis in *Phaeodactylum tricornutum*. Furthermore, using E(z)-knockout lines, RNA-seq and whole genome sequencing (WGS) data, we investigate the crosstalk between genetics and epigenetics in controlling morphogenesis in *P. tricornutum*.

(Manuscript in preparation)

Insights into the role of H3K27me3 mediated regulation of morphogenesis in the marine diatom *Phaeodactylum tricornutum*

Achal Rastogi[1], Xin Lin[2], Anne Flore Deton Cabanillas[1], Catherine Cantrel[1], Javier Paz[1], Murik Omer[3], Cédric Gaillard[4], Chris Bowler[1] and Leila Tirichine[1]

[1]Ecology and Evolutionary Biology Section, Institut de Biologie de l'École Normale Supérieure (IBENS), CNRS UMR8197 INSERM U1024, 46 rue d'Ulm 75005 Paris, France

Abstract

Tri-methylation of lysine 27 of histone H3 (H3K27me3) protein is well characterized in higher organisms like plants, insects and mammals where it plays a vital role at various stages of species development, organogenesis and in regulating cell fate decisions. Recently the functional profile of H3K27me3 has been established in the single cell marine model micro-eukaryote *Phaeodactylum tricornutum*. *P. tricornutum* is a coastal species and appears to have multiple morphological shapes (referred as morphotypes) that are fusiform, triradiate, oval, round and cruciform. These morphotypes have been shown to switch from one form to another under certain laboratory conditions and some transformations are reversible if the natural conditions are replenished. While genetic changes are not sufficient in explaining these dynamic and rapid switches, epigenetics holds a capacity to regulate these changes and identify the genes involved in giving the cell a specific morphotype. In order to understand the epigenetic mediated regulation of morphogenesis in *P. tricornutum*, we compared H3K27me3 profile of triradiate and fusiform cells sampled from two geographical isolated locations of the world. While doing so, we discovered many cell wall related genes being specifically marked and, regulated by H3K27me3 in triradiate cells as determined by knockout experiment of H3K27me3 methyltransferase gene, enhancer of zeste (Ez), in triradiate and fusiform backgrounds. Furthermore, identification of differentially marked genes between triradiate and fusiform cells revealed the presence of a genetic signature that might have a role in the presence of H3K27me3 over these genes.

Introduction

Diatoms are photosynthetic unicellular micro-eukaryotes found in almost all aquatic habitats. They belong to a large group called the heterokonts or stramenopiles, which is distantly related to the main eukaryotic lineages. Diatoms are cosmopolitan protists and highly diverse (de Vargas et al. 2015; Malviya et al. 2016). They are believed to originate from a serial secondary endosymbiosis where a heterotrophic cell engulfed red and green algae, and further diversified through horizontal transfer of bacterial genes (Bowler et al. 2008; Moustafa et al. 2009). Diatoms show a wide diversity of sizes and shapes and can be divided into two main groups, based on the organization of their silicified frustule, pennate and centric. *Phaeodactylum tricornutum* is a pennate diatom known to have different interconverting morphotypes, fusiform, triradiate, oval and cruciform (Borowitzka 1978; De Martino 2007; De Martino et al. 2011; He et al. 2014). Unlike most diatoms, *P.tricornutum* does not show the typical silicified frustule. Only the oval morphotype possesses a single silica valve while the valves in the other morphotypes consist entirely of organic material, although siliceous bands were observed in the girdle region of the fusiform cell wall (Borowitzka 1978). *P. tricornutum* is mainly found in coastal areas with high fluctuations of environmental conditions such as salinity, pH, light and nutrients, which might be in favor of the plasticity in the morphology inherent to this species. This is in line with several studies, which report a change in cell shape that can be reversible in response to an environmental trigger (De Martino 2007) . Ten natural variants have been isolated from different locations in the world oceans with a majority of fusiform morphotype, which was found to have the highest growth rate and adaptability to changing environments.

Comparative study using ITS2 region of the genome from ten natural isolates failed to designate any positive correlation between genotypes and morphotypes (De Martino 2007). Absence of such a correlation along with the reversible and rapid morphotype switch in *P. tricornutum* suggest an epigenetic mediated regulation, namely post translational modifications of histones (PTM) such as tri-methylation of lysine 27 of histone H3

(H3K27me3), a repressive mark well recognized for its role in governing morphogenesis and cell fate. H3K27me3 is deposited by polycomb repressive complex 2 (PRC2), a chromatin-remodeling complex that mediates silencing of gene expression during differentiation and development in both animals and plants (Surface et al. 2010; Aldiri and Vetter 2012; Fragola et al. 2013). PRC2 was first identified in *Drosophila melanogaster* and consists of four core proteins highly conserved among organisms: the histone methyltransferase Enhancer of zeste (E(z)), the WD40 domain containing polypeptide Extra Sex Comb (Esc), the C2H2 type zinc finger protein Suppressor of zeste 12 (Su(z)12) and the Nucleosome remodeling factor 55 kDa subunit (Nurf-55) (Martinez-Balbas et al. 1998; Schwartz and Pirrotta 2013) (Figure S1-S3). The absence of PRC2 in model unicellular fungi *S. pombe* and *S. cerevisiae* initially suggested that it arose to regulate developmental processes in multicellular organisms (Kohler and Villar 2008). This hypothesis has recently been questioned because PRC2 can be found in several single celled species (Shaver et al. 2010). This marked the beginning of attempts to understand the role of H3K27me3 in unicellular organisms. In-silico identification of polycomb complex in the pleomorphic model species, *P. tricornutum* (Figure S1-S3) presents a unique opportunity to decipher its role in single celled organisms. Moreover, based on our recent findings of the homologs of E(z) gene in multiple marine micro-eukaryotes establishes its ubiquitous role in unicellular organisms (Figure S4).

We report here for the first time comparative genome wide mapping of H3K27me3 in a unicellular species with two natural morphotype variants triradiate monoclonal population (PT8TC, TMP hereafter) and fusiform monoclonal population (PT1 8.6, FMP hereafter), which allowed the discrimination of putative targets with a role in cell fate determination. To gain insights into the role of polycomb proteins in unicellular species, we used CRISPR cas9 gene editing to knock out the catalytic unit of the PRC2 complex, which is E(z) and led to the loss of its expression, a total depletion of H3K27me3 from the genome and subsequent distortion in the morphotype. Our study brings new insights into the role of H3K27me3 in unicellular species and points to the role of the interplay between genetics and chromatin in cell fate determination.

Results and Discussions

Natural variation of H3K27me3 distribution between TMP and FMP

To investigate the function of PRC2 complex and uncover its targets, we carried out a genomic approach and performed Chromatin Immuno-Precipitation (ChIP) followed by deep sequencing using an antibody against H3K27me3 in TMP and FMP backgrounds. Additional marks (H3K4me2, H3K9me2) as well as RNA seq were also performed. CHIPseq data analysis revealed similar H3K27me3 enrichment profile between TMP and FMP that localizes majorly on transposable elements (TEs hereafter), specifically 58% and 60% within FMP and TMP, respectively (Figure 1A). However, the genome-wide H3K27me3 target coverage within TMP is slightly higher (2%) than the FMP (Figure 1B), which is due to more number of specific genes being targeted in the triradiate cells (Figure 1A). The mark was found to occupy, on average, ~11.55 % of the genome within FMP, targeting approximately 15% of genes (consistent with (Veluchamy et al. 2015)) and ~13.2% of the genome within TMP, targeting 19% of genes. This is slightly higher than H3K27me3 coverage in mammals (Ernst et al. 2011), *Arabidopsis* (Charron et al. 2009), Drosophila (Kharchenko et al. 2011) and *Neurospora* (Jamieson et al. 2013). However, the proportion of H3K27me3 targeted genes in *P. tricornutum* is less than in *Arabidopsis* where H3K27me3 majorly targets genes (~59%) (Charron et al. 2009; Ernst et al. 2011; Kharchenko et al. 2011; Jamieson et al. 2013) and remains higher in Neurospora *crassa* where a total of ~6.8% genome coverage by H3K27me3 targets only 8% of the genes. Further, most of the H3K27me3 target genes are shared between both ecotypes (Figure 1A, 1C) and exhibit consistent broad pattern of localization (Figure 1D) as described previously (Veluchamy et al. 2015). Among the target genes, 71% and 57% genes are flanked by transposable elements (referred as TGENES) within the vicinity of 500 bp upstream to TSS or 500 bp downstream to stop codon, in FMP and TMP respectively (Figure 1E, 1F). From 12177, 1836 (~15%) and 2377 (~19%) genes are marked by H3K27me3 in FMP and TMP, respectively. Most of the genes, from the latter categories, 1282/1836 (~70%) within FMP and 1558/2377 (~66%) in TMP, are marked at least in the gene-body (genic region between TSS and stop codon) and

are majorly shared between TMP (~66%) and FMP (~80%). Only 34% (853 genes/2377 genes marked by H3K27me3) and 20% (312 genes/1836 genes marked by H3K27me3) genes are found to be specific H3K27me3 targets within TMP and FMP respectively. Among the specifically marked genes, majority, 531 (~62%) in TMP and 251 (~80%) within FMP, are at least marked in the gene body. These observations indicate similar H3K27me3 gene target profile between the two backgrounds and also define the characteristic nature of H3K27me3 enrichment over gene-body in *P. tricornutum*. Consistent to genes and intergenic regions, most of the TE targets of H3K27me3 are shared between FMP and TMP (Figure 1A, 1C). The analysis of targeted TEs within classes I and II revealed strong H3M27me3 targeting of copy and paste transposition, mechanism that is specific to class I TEs, specifically Copia-type transposable elements (Figure 1G), in both the ecotype populations. This could be to avoid the free movement of the latter elements in the genome in order to keep them under tight control.

H3K27me3 majorly target genes specific to *P. tricornutum* and regulate cell wall related genes

Analysis of H3K27me3 distribution between the two morphotypes of *P. tricornutum* (fusiform and triradiate) revealed similar profiles of enrichment and localization, with some genomic features being targeted specifically in TMP and FMP. To investigate the functional regulation of H3K27me3 dynamic enrichment over various target loci, we performed high throughput RNA sequencing using total RNA from FMP and TMP. As reported previously (Veluchamy et al. 2015), we were able to distinguish six different enrichment profiles of H3K27me3 localization over the genes in both FMP and TMP (Figure 2A, 2C). They are; C1 (represented as “Fu” or “Tg” depending on cluster found in fusiform (F) population or triradiate (T) population) with genes marked only in the 500 bp upstream to TSS; C2 (represented as “Fg-Fu” or “Tg-Tu”) with genes marked within 500 bp upstream to TSS and gene body; C3 (represented as “Fg” or “Tg”) with genes marked only in gene body; C4 (represented as “Fg-Fd” or “Tg-Td”) with genes marked within gene body and 500 bp downstream to stop codon; C5 (represented as “Fd” or “Td”) with genes marked only in 500 bp downstream to stop codon; and C6 (represented as

"Fd-Fg-Fu" or "Td-Tg-Tu") with genes marked within 500 bp upstream to the TSS, gene body and 500 bp downstream to the stop codon (referred to as broadly marked). In addition, we found 14 and 18 genes marked only in the vicinity of 500 bp upstream to TSS and downstream to stop codon in FMP and TMP respectively (Figure 2A, 2C). Further evaluation of these genes revealed their close proximity to either TEs or genes, which are strongly marked by H3K27me3. Considering the broad nature of H3K27me3 marking over its targets that are often flanked to proximal features, we did not considered the genes, which are marked only in upstream and downstream with close proximity to H3K27me3 targets, as individual clusters.

With most of the H3K27me3 target genes being common between both morphotypes, their regulation corresponding to different enriched clusters is also coherent (Figure 2B, 2D). Among all the clusters, cluster with genes marked broadly (500bpUPstream-genebody-500bpDownstream) contributes 54% and 46% within FMP and TMP cells, respectively (Figure 2A, 2C). Compared to the unmarked genes, both FMP and TMP exhibit lowest expression in the cluster where genes are broadly marked (Fu-Fg-Fd/Tu-Tg-Td). In FMP, genes that are marked specifically either across upstream to TSS or downstream to stop-codon show higher expression, consistent to our previous observations showing the co-occurrence of active marks (like H3K9_14Ac) over these clusters (Veluchamy et al. 2015). This might also explain similar observation within TMP cells where genes that are marked only across the 500 bp upstream to TSS or 500 bp downstream to stop-codon have higher expression.

TMP possesses more number of H3K27me3 gene targets (2377 genes) compared to FMP (1836 genes) (Figure 1A), among which 1814 genes (either marked in both or marked in any one of them) are marked at least in the gene body. Further, total-RNA expression analysis exhibited 1097 (~9% of total 12177 Phatr3 gene annotations) differentially expressed genes between TMP and FMP. Among the latter, 554 and 543 genes are significantly upregulated ($\log_2$ (Fold-change) > 2; Padj. value < 0.05) and downregulated ($\log_2$ (Fold-change) < -2; Padj value < 0.05), respectively, in TMP cells compared to FMP. GO enrichment analysis of the upregulated genes revealed a significant (chi-squared test; P-value < 0.05) top enrichment of biological processes like gluconeogenesis, isopentenyl diphosphate biosynthetic process, mevalonate pathway, metal ion transport, proton transport and oxidation-reduction process. Whereas

downregulated genes show top enrichment of processes like leucyl-tRNA aminoacylation, mitochondrial alanyl-tRNA aminoacylation, nitrate assimilation, propyl-tRNA aminoacylation and S-adenosylmethionine biosynthetic process.

The expression of 711 (39%) from 1814 H3K27me3 target genes, localized at least in the gene-body, across either TMP or FMP is low expressed to repressed. Of the latter 711 genes, 308 (43%) and 102 (14%) genes are specifically marked in TMP and FMP, respectively. Although most of the genes are of unknown biological processes, we were able to trace most enriched processes among genes specifically marked in TMP or FMP using ~24% and ~27% genes known to have biological processes in TMP and FMP, respectively. Conclusively, genes being marked by H3K27me3 specifically in FMP exhibit enrichment of ATP biosynthetic process, glycerol-3-phosphate catabolic process, phosphatidylinositol biosynthetic process, phosphatidylinositol-mediated signaling and viral process. Whereas, genes that are specifically marked within TMP cells display top enrichment of cellular aldehyde metabolic process, lipid biosynthetic process, protein farnesylation, regulation of ARF protein signal transduction and regulation of cell proliferation. Apart from the latter, we found many putative cell wall associated genes to be specifically marked and regulated by H3K27me3 in TMP cells. We validated some of them using CHIP-qPCR (Figure 3A), which included genes similar to extensin (Phatr3_J44695), dentin (Phatr3_J47347), titin (Phatr3_J46523) proteins and sec14 gene (Phatr3_49062), which is known to be essential for vegetative growth in yeast (Bankaitis et al. 1989).

To gain insights into the function of Ez and its associated H3K27me3 mark in *P. tricornutum*, we performed CRISPR cas9 knockout of the gene in both morphotypes and performed genome wide expression (RNA-Seq) analysis of two mutants. With low or no expression of E(z) gene and an overall depletion of H3K27me3 genome-wide, observed in multiple mutant lines (Figure 3B), comparison of Ez mutant expression profiles with their respective wild type revealed that FMP and TMP exhibited 2795 (23%) and 1774 (15%) differentially expressed genes, respectively. Among the latter, 22/38 (58%) FMP and 24/44 (55%) TMP specific target genes, including some cell wall related genes specifically marked by H3K27me3 within TMP, show upregulation of gene expression, suggesting a loss of H3K27me3. For validation purposes, Chip-qPCR was performed for few genes confirming partial or total depletion of

H3K27me3 (Figure 3C, 3D). Furthermore, most of the genes (>25%) that are specifically marked in TMP and FMP populations does not have significant homology with other species in the tree of life and are specific to *P. tricornutum*. Approximately 14% of the genes are specific to diatoms. While ~8% of the genes shows conservation in almost all the lineages of Tree of life except viruses, more than 4% genes shows conservation across wide range of eukaryotic lineages.

The interplay between DNA variants and H3K27me3 in TMP and FMP

To evaluate the genetic diversity between TMP and FMP, contributing to morphotype determination, we sequenced the genome of TMP and performed a reference-based assembly using the genome of FMP (Pt1 8.6) (Bowler et al. 2008) as a reference. With an average depth of 110x and 94% of genome coverage, we discovered 232,323 (depth >= 4x, which means minimum of 4 reads per base) single nucleotide polymorphic (SNP) sites across the whole genome of TMP cells. We also discovered 133 small insertions and 642 deletions (INDELs) within TMP with highest insertion and deletion length to be 19 base pairs (bp) and 212 bp, respectively. Localization of the polymorphic sites over the functional features (genes, transposable elements, and intergenic regions) of the genome revealed high densities of the polymorphisms over genes specifically on exons, and is consistent across all the previously reported ecotype populations (Chapter 3 of the current thesis). With an average transition to transversion ratio of 1.6 and 1 mutation per 117 bases, the spectrum of mutations across all the ecotypes reveals a higher rate of transitions over transversions. In total, six possible types of single nucleotide changes can be distinguished, among which, G:C -> A:T and A:T -> G:C, while maintaining balance between each other, accounts for more than 72% of the observed mutations. An average non-synonymous to synonymous mutation ratio (N/S) is estimated to be 0.88 and indicates the fast evolving nature of TMP, which could be because of their adaptation to dynamically changing niches.

In order to understand the functional consequences of the polymorphisms, we clustered all the genes into two groups: Positive selection (PoS) and Loss of function (LoF) mutations. Both

groups reflect the characteristic functional nature of polymorphisms over the genes. Genes with high (more than 1) Ka/Ks (dN/dS) ratios are clustered together as highly evolving genes considered to be under positive selection. These polymorphisms are less likely to be deleterious and assumed to provide some kind of fitness to the species. Such genes display significant high ratio of non-synonymous (N) mutations/N sites over synonymous (S) mutations/S sites. Across all the ecotypes, 40 genes are found to be under natural selection. Among the latter, 8 (20%) genes are marked by H3K27me3 in both TMP and FMP, whereas 3 (5%) genes are marked specifically within TMP. Similarly, 915 genes are found to harbor frame-shift mutations, start/stop codon loss or premature start/stop codon mutations and are categorized under loss of function (LoF) alleles. Among the latter, 242 (26%) genes are marked by H3K27me3 in both TMP and FMP, whereas, only 50 (0.5%) and 22 (0.9%) genes are marked specifically within TMP and FMP, respectively. 98% (238 genes) of the genes that are marked by H3K27me3 and exhibiting LoF mutations in both TMP and FMP also contain numerous missense (mutations causing theoretical amino-acid substitutions) mutations. In addition, 6 out of 40 (15%) genes that are categorized as under PoS also exhibit LoF mutations. Among them, 2 (33%) and 1 (16%) are marked by H3K27me3 in both the populations and specifically in TMP, respectively. Interestingly, we found a significant (chi-square test; P-value < 0.01) number of genes with missense and LoF mutations, which are specifically marked and regulated by H3K27me3 in TMP, compared to the unmarked genes. Similar effects are observed when the genes are marked and regulated either specifically in FMP or both TMP and FMP cells. Mutational spectrum within genes that are specifically marked and regulated in TMP exhibits an abundance (~64%) of A:T -> G:C and C:G -> T:A mutations. These observations suggest a putative role of sequence variation in the direct or indirect recruitment of H3K27me3 in *P. tricornutum*.

Co-localization of H3K27me3 and H3K9me2 defines repressive chromatin state in both TMP and FMP

Along with H3K27me3, we also profiled two marks, H3K9me2 and H3K4me2 to study their distribution in the two morphotypes and analyze their co-occurrence with H3K27me3 and

effect over the marked genes. Lysine 9 of histone H3 can be mono-, di-, or tri- methylated both in plants and animals. Like H3K27me3, H3K9me2 is also associated with repression and targets mainly TE(s) in both TMP and FMP. Unlike H3K27me3 and H3K9me2, H3K4me2 is associated to constitutive expression of genes, which are its major targets. To identify different chromatin states (CS) we compared the co-localization profiles of H3K9me2, H3K27me3 and H3K4me2 over genes using FMP and TMP (Figure 4A, 4B). Both TMP and FMP possess approximately similar number of genes that are marked by H3K4me2. Whereas, ~1600 genes are uniquely marked by H3K9me2 within TMP cells (Figure 4B). Interestingly the co-localization profile of the three marks across FMP and TMP is coherent, in terms of maintaining similar chromatin states across the genome (Figure 4A, 4B). H3K27me3 and H3K9me2 often co-localize on their targets and the pattern is consistent in both the ecotypes (Figure 4A, 4B). Although, H3K9me2 and H3K27me3 peaks are reported to be mutually exclusive in humans, they are majorly found to be close neighbors (Lienert et al. 2011). Genes that are co-marked by the repressive marks, H3K27me3 and H3K9me2, shows very low or no expression within FMP and TMP cells (Figure 4A, 4B). Similarly, genes that are specifically marked by H3K9me2 show moderate to low expression compared to unmarked genes (Figure 4A, 4B). On the contrary, genes that are specifically marked by H3K4me2 are highly expressed. Interestingly, on an average, genes that are co-marked by H3K27me3 and H3K9me2 are flanked by more TEs, compared to those that are either marked specifically by H3K9me2 or H3K27me3, within the vicinity of 500 bp upstream till 500 bp downstream of a gene. Conclusively, co-localization of the three marks suggests four chromatin states that are consistent within both FMP and TMP cells (Figure 4A, 4B). These states are, **CS1**: Highly expressed, defined by genes that are uniquely marked by H3K4me2; **CS2**: Moderate to high expression, defined by genes that are marked by either all the three marks or by H3K4me2 with H3K9me2 and/or H3K27me3; **CS3**: Moderate to low expression, defined by genes that are marked specifically by either H3K9me2 or H3K27me3; and, **CS4**: Low or no expression, defined by genes that are marked by both the repressive marks, H3K27me3 and H3K9me2 suggesting an additive effect of both repressive marks. It is also noteworthy, that quantitatively, the enrichment of H3K9me2 and H3K27me3 within the co-localized target genes is higher compared to the enrichment of H3K9me2 and H3K27me3

within genes marked specifically by H3K9me2 and H3K27me3, respectively (Figure 5A) and is correlated to the expression (Figure 5B). This suggest the establishment of facultative heterochromatin, characterized as high TE(s) density and low or no expression state (CS4) of the genome, by the co-localization of H3K27me3 and H3K9me2

Conclusions

We demonstrate natural variation of the repressive chromatin mark H3K27me3 between two *P. tricornutum* ecotype populations, Pt1 8.6 and Pt8Tc, possessing fusiform (FMP) and triradiate morphotypes (TMP), respectively. The distribution of H3K27me3 in FMP and TMP cells is highly similar with TE(s) being the major targets suggesting an ancestral role of H3K27me3 in suppression of TEs. Although the global distribution of H3K27me3 is coherent, there are many genes, involved in wide range of biological processes including cell wall related genes, specifically marked within TMP and FMP. Along with the distribution of H3K27me3, its co-localization profile with H3K4me2 and H3K9me2, and the respective chromatin states are consistently maintained within both TMP and FMP. Furthermore, the function of H3K27me3 is for the first time demonstrated in a single celled species using CRISPR cas9 editing where the knockout of enhancer of zeste, a polycomb protein responsible of the deposition of H3K27me3 results in its depletion from the genome and a distortion of the morphology of the cell. Coupling the comparative distribution of H3K27me3 profiles with the genetic diversity between the two ecotypes revealed a significant role of sequence variations and putative cross talk between DNA sequence and epigenetic repressive mark H3K27me3 in order to attribute a specific fate to the cell via the repression of specific target genes most of which encode genes of unknown functions which interestingly are diatom or *P. tricornutum* specific. Among the genes that are differentially targeted by H3K27me3 and with functional domains, we identified and validated few with cell wall related functions.

Methods

Strains and growth conditions

Phaeodactylum tricornutum Bohlin Clone Pt1 8.6 (CCMP2561) and Clone Pt8Tc cells were grown as described previously (Siaut et al. 2007).

Isolation and immunoprecipitation of chromatin

Chromatin isolation and immunoprecipitation were performed as described previously (Lin et al. 2012). The following antibodies were used for immunoprecipitation: H3K27me3 (07-449) from Millipore and H3K27me3 from cell signaling technology. QPCR on recovered DNA was performed as described previously (Lin et al. 2012)

CRISPR-CAS plasmid construction

hCAS9n (CAS9 from *Streptococcus* pyogenes, adapted to human codon usage, fussed to SV40 nuclear localization sequence, and contains a D10A mutation) was amplified from pcDNA3.3-TOPO-hCAS9n (kindly received from Dr. Yonatan B. Tzur), using the primers 5'-CAC CAT GGA CAA GAA GTA CTC-3' and 5'- TCA CAC CTT CCT CTT CTT CTT-3'. The PCR product was first cloned into pENTR using pENTR/D-TOPO cloning kit (ThermoFisher Scientific), and then sub-cloned into a PT pDest, containing an N-terminal HA-tag (Siaut et al. 2007), following the manufacturer protocol, which was named pDest-HA-hCAS9n.

For the sgRNA vector we first cloned the snRNA U6 promoter(Rogato et al. 2014) from *P. tricornutum* genomic DNA using the primers 5'- AAA CGA CGG CCA GTG AAT TCT CGT TTC TGC TGT CAT CAC C-3' and 5'- TCT TTA ATT TCA GAA AAT TCC GAC TTT GAA GGT GTT TTT TG-3'. PU6::unc-119_sgRNA (kindly received from Dr. Yonatan B. Tzur) backbone was amplified using the primers 5'-CAA AAA ACA CCT TCA AAG TCG GAA TTT TCT GAA ATT AAA GA-3' and 5'-GGT GAT GAC AGC AGA AAC GAG AAT TCA CTG GCC GTC GTT T-3'. The two PCR products were used

as template for a second round fusion PCR reaction as described in (Hobert 2002). We further transformed the resulting product into *E. coli*, and extracted the ligated plasmid. The terminator sequence of the P. tricornutum U6 was amplified using the primers 5'-CATTCTAGAAGAACCGCTCACCCATGC-3' and 5'-GTTAAGCTTGAAAAGTTCGTCGAGACCATG-3', digested by XbaI/HindIII and ligated into XbaI/HindIII digested pU6::unc-119. The resulting vector, ptU6::unc-119-sgRNA, was used as template to replace the target sequence to E(Z) target by PCR using primers 32817TS12fwd GTG TCG GAG CCC GCC ATA CCG TTT TAG AGC TAG AAA TAG C and 32817TS12rev GGT ATG GCG GGC TCC GAC ACC GAC TTT GAA GGT GTT TTT TG. Target sequences were picked using PhytoCRISP-Ex (Rastogi et al. 2016).

Transformation of *P. tricornutum* cells and screening for mutants

Wild type cells of the reference strain Pt1 8.6 and the triradiate morphotype Pt8Tc were transformed with three plasmids (pPhat1, Cas9 and guide RNA with the target sequence) as described previously (Falciatore et al. 1999). Positive transformants were validated by triple PCR screen for pPhaT1 shble primers (ACT GCG TGCACTTCGTGGC/TCGGTCAGTCCTGCTCCTC), sgRNA (GAGCTGGAAATTGGTTGTC/GACTCGGTGCCACTTTTTCAAGTT) and CAS9n (GGGAGCAGGCAGAAAACATT/TCACACCTTCCTCTTCTTCTT). For each colony, a rapid DNA preparation was performed as described previously and fragment of 400 bp was amplified with primers flanking the target sequence in the enhancer of zeste gene (E(z)). The forward primer used is 5'-TAAGATGGAGTATGCCGAAATTC-3' and reverse primer is 5'-AGGCATTTATTCGTGTCTGTTCG-3' PCR product was run in 1% agarose gel and a single band was extracted using Machery Nagle kit and according to the manual manufacturer. PCR product was sequenced using the primer 5'-AGCCACCCTGCGTTAACTGAAAAT-3'.

Sequencing and Computational data analysis

Chromatin Immunoprecipitation (ChIP), Total RNA and DNA extraction was done with monoclonal cell cultures grown using single triradiate cell from Pt8 population (referred as

Pt8Tc). CHIPseq and RNA-Seq was performed as described previously (Veluchamy et al. 2013; Veluchamy et al. 2015). Whole genome sequencing was performed using genomic DNA from Pt8Tc to construct a sequencing library by following the manufacturer's instructions (Illumina Inc.). Paired-end sequencing libraries with an insert size of approximately 400 bp were sequenced using Illumina HiSeq 2000 sequencer at Fasteris (https://www.fasteris.com. Raw reads from each experiment were filtered and low quality read-pairs were discarded using FASTQC with a read quality (Phred score) cutoff of 30. Using the genome assembly published in 2008 as reference (Pt1 8.6), we performed reference- assisted assembly of all the filtered reads. We used BOWTIE for mapping the CHIP sequencing and whole genome sequencing reads to the reference genome followed by the processing and filtering of the alignments using SAMTOOLS and BEDTOOLS. SICER (Zang et al. 2009) was then used to identify significant enriched H3K27me3, H3K9me2 and H3K4me2 peaks by comparing it with the INPUT. Differential H3K27me3 and H3K9me2 peak enrichment analysis between Pt1 8.6 and Pt8Tc backgrounds was also done using SICER-df plugin. Peaks with atleast four folds with Padj < 0.05 differential enrichment or depletion were considered significant. Functional inferences were obtained by overlapping the differentially enriched peaks over structural annotations from Phatr3 genome annotation (Phatr3, Chapter 2). For estimating the genetic diversity between Pt1 8.6 and Pt8Tc genome, GATK (McKenna et al. 2010) configured for diploid genomes, was used for variant calling, which included single nucleotide polymorphisms (SNPs), small insertions and deletions ranging between 1 and 300 base pairs (bp). The genotyping mode was kept default (genotyping mode = DISCOVERY), Emission confidence threshold (-stand_emit_conf) was kept 10 and calling confidence threshold (-stand_call_conf) was kept at 30. The minimum number of reads per base to be called as a high quality SNP was kept at 4 (i.e., read-depth >=4x). SNPEFF and KaKs calculator were used to annotate the functional nature of the polymorphisms. Along with the non-synonymous, synonymous, loss-of-function (LOF) alleles, transition to transversion ratio and mutational spectrum of the single nucleotide polymorphisms were also measured. Genes with Ka/Ks aka dN/dS ratio more than 1 with a p-value less than 0.05 are considered as undergoing natural or Darwinian selection. Various in-house scripts were also used at different levels for analysis and for plotting graphs.

Basic RNA expression and differential gene expression analysis was performed using Eoulsan version 1.2.2 with default parameters (Jourdren et al. 2012). Data visualization and graphical analysis were performed principally using CIRCOS, CYTOSCAPE, IGV and R. GO enrichment analysis was performed by comparing the observed to expected frequency ratio of a biological process to occur in any given study relevant geneset. Chi-squared test with a P-value < 0.05 was considered significant to reject the null hypothesis.

Validation of enrichment and expression of target genes

QPCR: Total RNA was extracted from TMP and FMP cells as described previously (Siaut et al. 2007) and cDNA was synthesized with cDNA high yield synthesis kit according to the manufacturer user manual. Quantitative PCR was performed as described previously (Siaut et al. 2007) using the primer list in table S1 **Western blot analysis:** Chromatin was extracted from wild type as well as mutants of both TMP and FMP cells and western blot performed as described previously (Lin et al. 2012)

References

Aldiri I, Vetter ML. 2012. PRC2 during vertebrate organogenesis: a complex in transition. *Dev Biol* **367**: 91-99.

Bankaitis VA, Malehorn DE, Emr SD, Greene R. 1989. The Saccharomyces cerevisiae SEC14 gene encodes a cytosolic factor that is required for transport of secretory proteins from the yeast Golgi complex. *J Cell Biol* **108**: 1271-1281.

Borowitzka MA, Volcani, B.E. 1978. The polymorphic diatom Phaeodactylum tricornutum: Ultrastructure of its morphotypes. *J Phycolo* **14**: 10-21.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP et al. 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456**: 239-244.

Charron JB, He H, Elling AA, Deng XW. 2009. Dynamic landscapes of four histone modifications during deetiolation in Arabidopsis. *The Plant cell* **21**: 3732-3748.

De Martino A, Bartual A, Willis A, Meichenin A, Villazan B, Maheswari U, Bowler C. 2011. Physiological and Molecular Evidence that Environmental Changes Elicit Morphological Interconversion in the Model Diatom Phaeodactylum tricornutum. *Protist* **162**: 462-481.

De Martino AM, A. Juan Shi, K.P. Bowler, C. 2007. Genetic and phenotypic characterization of Phaeodactylum tricornutum (Bacillariophyceae) accessions. *J Phycol* **43**: 992–1009.

de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, Lara E, Berney C, Le Bescot N, Probert I et al. 2015. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43-49.

Falciatore A, Casotti R, Leblanc C, Abrescia C, Bowler C. 1999. Transformation of Nonselectable Reporter Genes in Marine Diatoms. *Mar Biotechnol (NY)* **1**: 239-251.

Fragola G, Germain PL, Laise P, Cuomo A, Blasimme A, Gross F, Signaroldi E, Bucci G, Sommer C, Pruneri G et al. 2013. Cell reprogramming requires silencing of a core subset of polycomb targets. *PLoS Genet* **9**: e1003292.

He L, Han X, Yu Z. 2014. A rare Phaeodactylum tricornutum cruciform morphotype: culture conditions, transformation and unique fatty acid characteristics. *PloS one* **9**: e93922.

Hobert O. 2002. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic C. elegans. *Biotechniques* **32**: 728-730.

Jamieson K, Rountree MR, Lewis ZA, Stajich JE, Selker EU. 2013. Regional control of histone H3 lysine 27 methylation in Neurospora. *Proceedings of the National Academy of Sciences of the United States of America* **110**: 6027-6032.

Jourdren L, Bernard M, Dillies MA, Le Crom S. 2012. Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. *Bioinformatics* **28**: 1542-1543.

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T et al. 2011. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**: 480-485.

Kohler C, Villar CB. 2008. Programming of gene expression by Polycomb group proteins. *Trends Cell Biol* **18**: 236-243.

Lienert F, Mohn F, Tiwari VK, Baubec T, Roloff TC, Gaidatzis D, Stadler MB, Schubeler D. 2011. Genomic prevalence of heterochromatic H3K9me2 and transcription do not discriminate pluripotent from terminally differentiated cells. *PLoS Genet* **7**: e1002090.

Lin X, Tirichine L, Bowler C. 2012. Protocol: Chromatin immunoprecipitation (ChIP) methodology to investigate histone modifications in two model diatom species. *Plant methods* **8**: 48.

Malviya S, Scalco E, Audic S, Vincent F, Veluchamy A, Poulain J, Wincker P, Iudicone D, de Vargas C, Bittner L et al. 2016. Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences of the United States of America* doi:10.1073/pnas.1509523113.

Martinez-Balbas MA, Tsukiyama T, Gdula D, Wu C. 1998. Drosophila NURF-55, a WD repeat protein involved in histone metabolism. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 132-137.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D. 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324**: 1724-1726.

Rastogi A, Murik O, Bowler C, Tirichine L. 2016. PhytoCRISP-Ex: a web-based and stand-alone application to find specific target sequences for CRISPR/CAS editing. *BMC bioinformatics* **17**: 261.

Rogato A, Richard H, Sarazin A, Voss B, Cheminant Navarro S, Champeimont R, Navarro L, Carbone A, Hess WR, Falciatore A. 2014. The diversity of small non-coding RNAs in the diatom Phaeodactylum tricornutum. *BMC genomics* **15**: 698.

Schwartz YB, Pirrotta V. 2013. A new world of Polycombs: unexpected partnerships and emerging functions. *Nature reviews Genetics* **14**: 853-864.

Shaver S, Casas-Mollano JA, Cerny RL, Cerutti H. 2010. Origin of the polycomb repressive complex 2 and gene silencing by an E(z) homolog in the unicellular alga Chlamydomonas. *Epigenetics* **5**: 301-312.

Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falciatore A, Bowler C. 2007. Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene* **406**: 23-35.

Surface LE, Thornton SR, Boyer LA. 2010. Polycomb group proteins set the stage for early lineage commitment. *Cell Stem Cell* **7**: 288-298.

Veluchamy A, Lin X, Maumus F, Rivarola M, Bhavsar J, Creasy T, O'Brien K, Sengamalay NA, Tallon LJ, Smith AD et al. 2013. Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum. *Nat Commun* **4**.

Veluchamy A, Rastogi A, Lin X, Lombard B, Murik O, Thomas Y, Dingli F, Rivarola M, Ott S, Liu X et al. 2015. An integrative analysis of post-translational histone modifications in the marine diatom Phaeodactylum tricornutum. *Genome biology* **16**: 102.

Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**: 1952-1958.

Figures

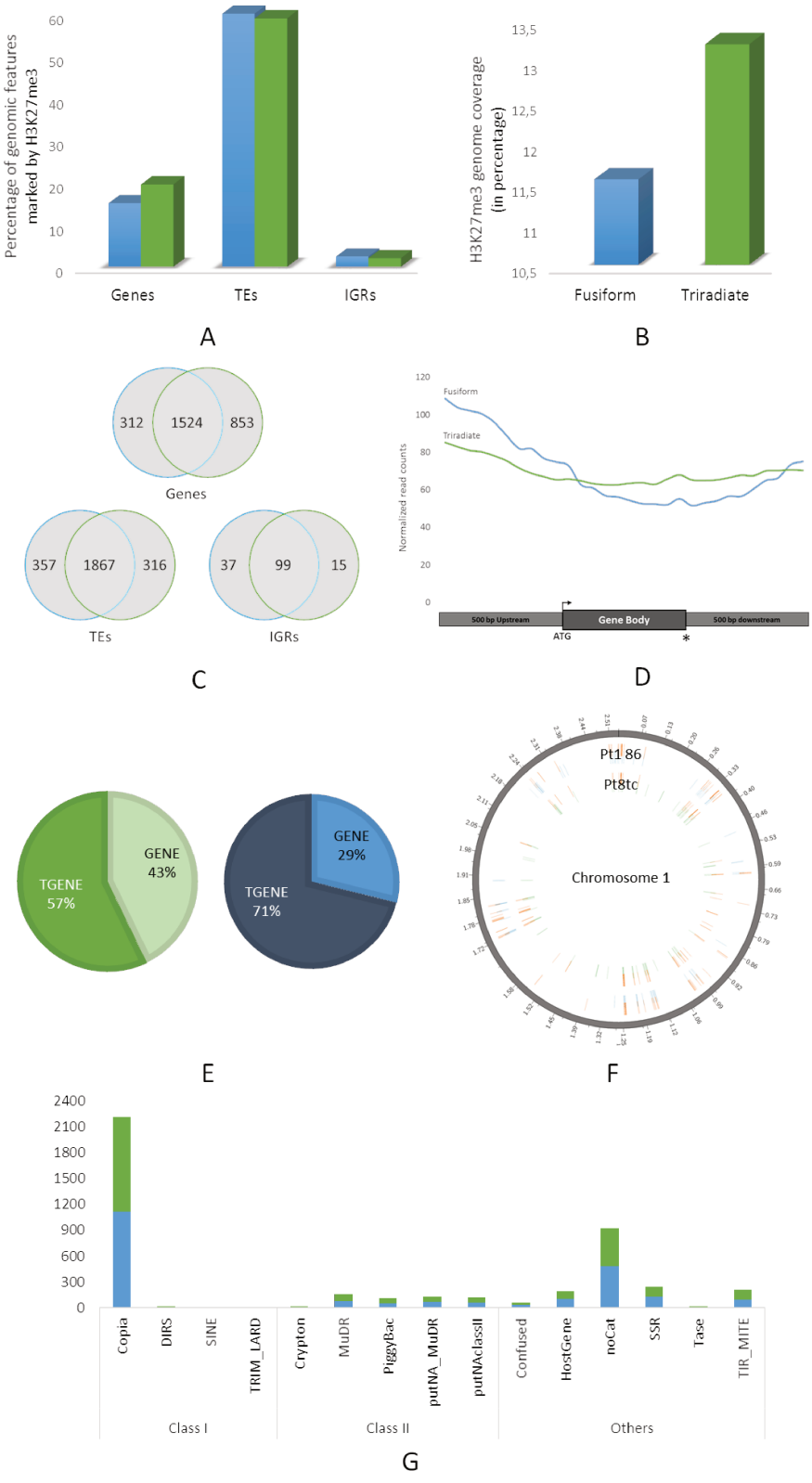

Percentage of genomic features marked by H3K27me3
Genes
TEs
IGRs
A
H3K27me3 genome coverage (in percentage)
Fusiform
Triradiate
B
312
1524
853
Genes
357
1867
316
TEs
37
99
15
IGRs
C
Fusiform
Triradiate
Normalized read counts
500 bp Upstream
Gene Body
500 bp downstream
ATG
D
GENE 43%
TGENE 57%
GENE 29%
TGENE 71%
E
Pt1 86
Pt8tc
Chromosome 1
F
Copia
DIRS
SINE
TRIM_LARD
Crypton
MuDR
Piggybac
putNA_MuDR
putNAclassII
Confused
HostGene
noCat
SSR
Tase
TIR_MITE
Class I
Class II
Others
G

Figure 1. Comparative genome-wide distribution of H3K27me3 within fusiform (PT186, FMP) and (PT8TC, TMP) triradiate monoclonal populations. Bar plots in (A) and (B) represents the percentage of different genomic features [Genes, transposable elements (TEs) and Intergenic regions (IGRs)] that are targeted by H3K27me3 in TMP (green bar) and FMP (blue bar). (B) The bar plot represent total genome coverage of H3K27me3 within TMP (green bar) and FMP (blue bar). (C) Venn diagram compares the number of common and specific genomic features targeted by H3K27me3 in TMP (green circles) and FMP (blue circles). (D) The line plot represents an average distribution profile of H3K27me3 over the 500 bp upstream, genebody and 500 bp downstream region of all the gene targets in FMP and TMP. (E) The pie chart depicts the portion of H3K27me3 target genes, in both TMP (green) and FMP (blue), are being flanked by TEs within the vicinity of 500 bp upstream till 500 bp downstream. (F) The circus plot, using chromosome 1 of the *P. tricornutum* genome, represents examples of TEs being in close proximity to the H3K27me3 target genes in both TMP (denoted by Pt8Tc) and FMP (denoted by Pt1 8.6). (G) The stack bar plot represents the proportion of different classes of transposable elements (TEs) that are targeted by H3K27me3 within TMP (green bars) and FMP (blue bars).

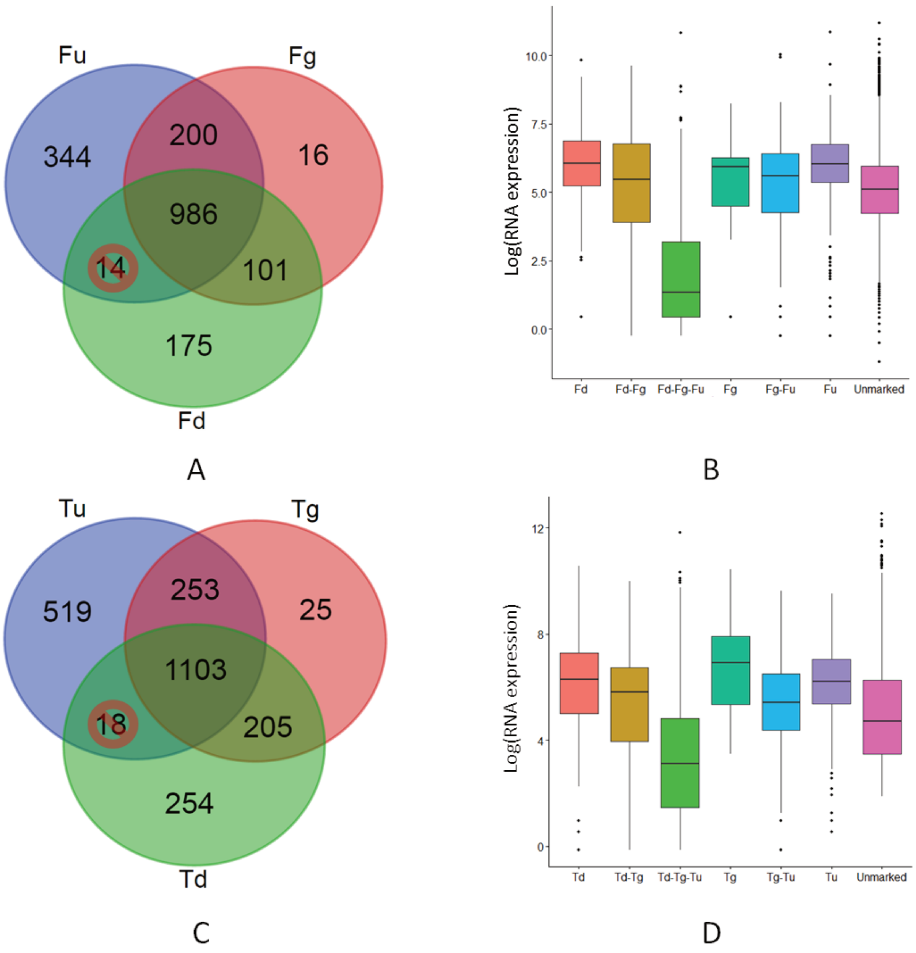


Figure 2. Enrichment profile of H3K27me3 over genes within TMP and FMP. Fu, Fg and Fd represent genes marked within fusiform monoclonal population on, anywhere within 500 bp upstream to TSS, gene-body and anywhere 500 bp downstream to stop codon, respectively. Similarly Tu, Tg and Td represent genes marked within triradiate monoclonal population. (A) The Venn diagram depicts the distribution profile of H3K27me3 target genes in FMP along with their average expression as represented by box plot (B). Similarly (C) represents the distribution of H3K27me3 in different regions of target genes in TMP, along with the average expression grouped into individual category (D).

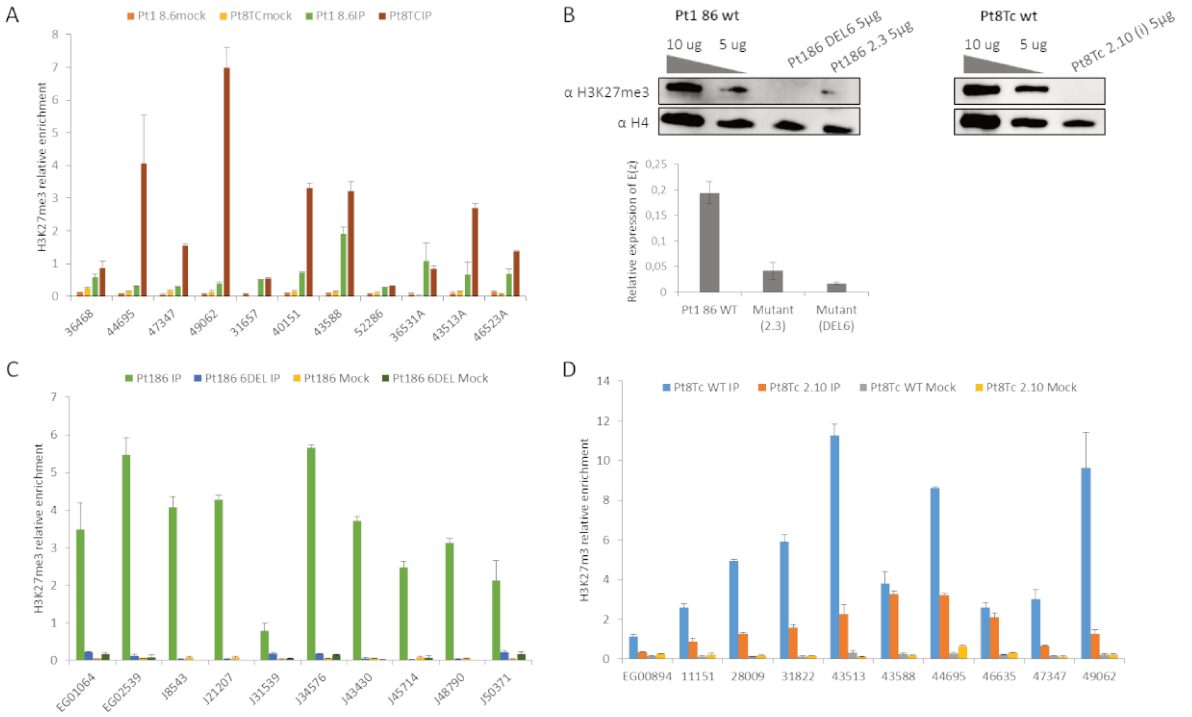


Figure 3. ChIP qPCR validation for the enrichment of H3K27me3 over target genes in both Pt 8Tc and Pt 1 8.6 (A, C, D), along with their corresponding knockout mutants (C, D). (B) Western blot analysis of chromatin extracts from wild type cells of *P. tricornutum* and enhancer of zeste knock out mutants using H3K27me3 monoclonal antibody. 1: Pt 1 8.6 DEL6 (6 nucleotides deletion inducing a premature stop codon), 2: Pt 1 86_2-3 (3 nucleotides insertion leading to an amino acid substitution), 3: Pt8Tc_2-10(i): a nucleotide insertion leading to an amino acid substitution). H4 is the loading control.

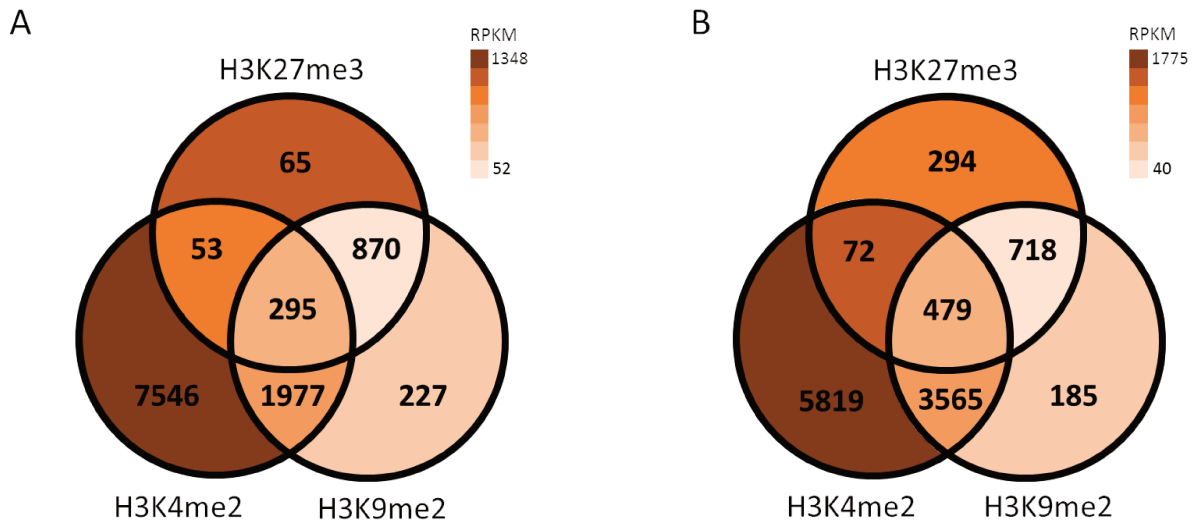


Figure 4. Co-localization profile of H3K27me3, H3K9me2 and H3K4me2 over gene body. Color gradient reflects an average expression of genes, and hence define various chromatin states with respect to the co-localization profile of different chromatin marks. (A) Co-localization analysis within FMP and (B) co-localization analysis within TMP.

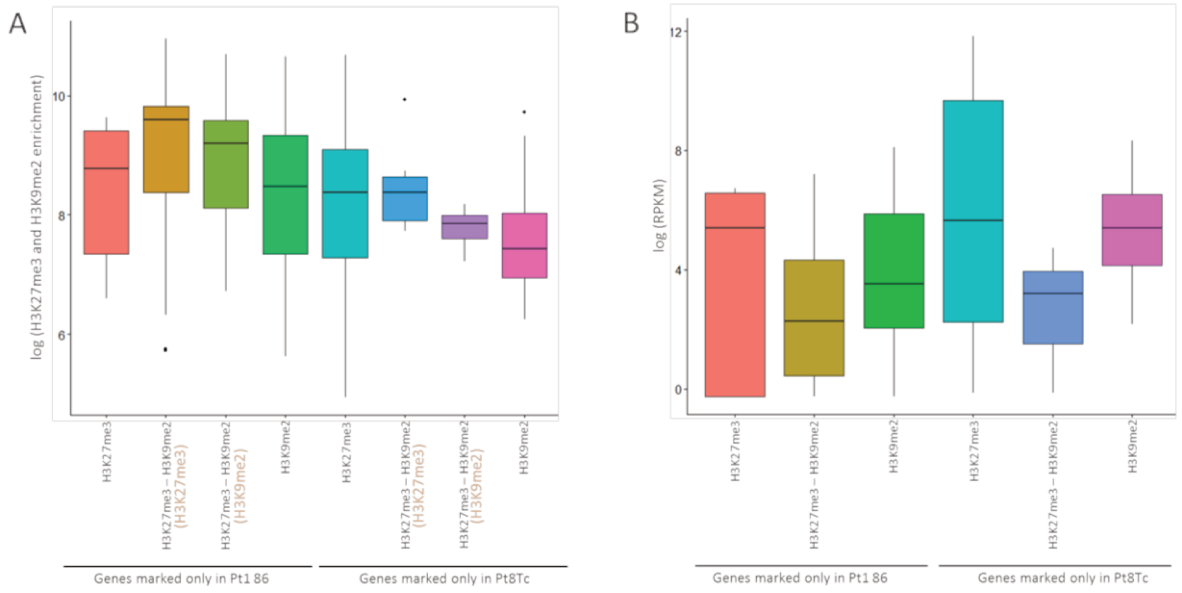


Figure 5. (A) The boxplot represents the enrichment of H3K27me3 and H3K9me2 (y-axis) within genes that are specifically marked either by H3K27me3 or H3K9me2, and genes that re co-localized by H3K27me3 and H3K9me2 (enrichment of individual marks is highlighted in orange, x-axis), in both TMP (Pt8Tc) and FMP (Pt1 8.6). (B) The boxplot represents the expression of genes within the categories described in (A).

Chapter 5

PhytoCRISP-Ex

Since past many years, gene editing has been widely employed to functionally characterize gene(s) and various metabolic pathways. Among numerous methods, CRISPR/Cas9 editing is a novel technique that allows researchers to edit genome with high precision, efficiency and flexibility. CRISPR was first discovered as a defense mechanism in wide range of bacteria(s) and got his name as "clustered regularly interspaced short palindromic repeats (CRISPR)" from its characteristic pattern of repeated DNA fragment interspaced by a unique sequence. These unique sequences were then distinguished as viral DNA fragments kept in the genome by bacteria as immune response inducers against virus containing these sequences. Later this system is engineered to work as a gene editing machinery in eukaryotes and requires a guide RNA (also called CRISPR targets) coupled with Cas9 (CRISPR associated proteins) protein to cleave the region of interest. Numerous softwares are now available which identifies CRISPR targets with no or low off target activity. Although these softwares have high sensitivity and specificity in finding CRISPR target sequences, they are of limited use in many ways and among many research communities. In the current chapter, I present PhytoCRISP-Ex as web-based and stand-alone software to predict CRISPR targets in a wide range of genomes including *Phaeodactylum tricornutum*. We also used the software to predict potential targets within E(z) gene, which led to a successful knockout of the gene (Chapter 4). PhytoCRISP-Ex is an adequately designed tool and out-pars most of the existing softwares by adopting new and essential utilities.

Rastogi et al. BMC Bioinformatics (2016) 17:261
DOI 10.1186/s12859-016-1143-1

BMC Bioinformatics

SOFTWARE Open Access

PhytoCRISP-Ex: a web-based and stand-alone application to find specific target sequences for CRISPR/CAS editing

Achal Rastogi, Omer Murik, Chris Bowler and Leila Tirichine*

Abstract

Background: With the emerging interest in phytoplankton research, the need to establish genetic tools for the functional characterization of genes is indispensable. The CRISPR/Cas9 system is now well recognized as an efficient and accurate reverse genetic tool for genome editing. Several computational tools have been published allowing researchers to find candidate target sequences for the engineering of the CRISPR vectors, while searching possible off-targets for the predicted candidates. These tools provide built-in genome databases of common model organisms that are used for CRISPR target prediction. Although their predictions are highly sensitive, the applicability to non-model genomes, most notably protists, makes their design inadequate. This motivated us to design a new CRISPR target finding tool, PhytoCRISP-Ex. Our software offers CRIPSR target predictions using an extended list of phytoplankton genomes and also delivers a user-friendly standalone application that can be used for any genome.

Results: The software attempts to integrate, for the first time, most available phytoplankton genomes information and provide a web-based platform for Cas9 target prediction within them with high sensitivity. By offering a standalone version, PhytoCRISP-Ex maintains an independence to be used with any organism and widens its applicability in high throughput pipelines. PhytoCRISP-Ex out pars all the existing tools by computing the availability of restriction sites over the most probable Cas9 cleavage sites, which can be ideal for mutant screens.

Conclusions: PhytoCRISP-Ex is a simple, fast and accurate web interface with 13 pre-indexed and presently updating phytoplankton genomes. The software was also designed as a UNIX-based standalone application that allows the user to search for target sequences in the genomes of a variety of other species.

Keywords: CRISPR, Cas9, Protists, Genome editing, Eukaryotes

Background

Phytoplankton are microalgae that form an essential constituent of the marine food chain. Though microscopic and mostly uncharacterized, these minute organisms have tremendously showcased themselves as potential research models [22, 20]. Recent large-scale sampling to understand the morphological and genetic diversity of this hidden community [3, 12, 21] has already established the foundation for further molecular studies. The successful achievement of this exploration also reflects the interest of research communities towards the functional characterization of phytoplankton in the near future.

Clustered regularly interspaced short palindromic repeats CRISPR/CAS systems have recently emerged as a simple and accurate tool for genome editing [13], and show facile editing in numerous organisms including bacteria [10], yeast [4], plants [6], human [14] and other animals [2, 5, 7, 9, 17, 23, 24]. Common designed CRISPR systems consist of expression of a Cas9 nuclease or nickase and a single guide RNA (sgRNA). The latter includes a 20-bp target sequence used to target the Cas9/RNA complex to the desired chromosomal location. By modifying only these 20-bp, the targeting of the whole CRISPR/Cas9 complex will change and thus the DNA cleavage site. A well designed target sequence must not only optimally bind to its desired target, but

* Correspondence: leila.tirichine@ens.fr
Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, PSL Research University, CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France

must also not target any other site in the genome being edited, to avoid undesired off-target mutations [11]. Several parameters were shown to contribute to a better targeting of the target sequence, and the general form of G-N19-NGG is widely accepted, although others were also suggested [1].

Because target site choice is a key point for promoting successful editing, several computational tools have been designed aiming to automate this procedure, all offering candidate target sites for a given input sequence/gene and potential off-target sites for a given background genome. This search is usually restricted for genomes of selected model organisms such as human, mice, fruit-fly, *C. elegans*, and yeast, making them irrelevant for researchers aiming to use CRISPR on other, less common, organisms. We designed PhytoCRISP-Ex, a user friendly web interface, as well as a stand-alone software, which predicts potential target sites for CRSIPR/CAS projects. The off-target analysis can be performed against indexed genomes from 13 algae (diatoms, green algae, haptophytes, etc.), or any user defined genome or transcriptome, assembled or not, making it useful for designing CRISPR projects for many communities.

Algorithm

PhytoCRISP-Ex is a rational and flexible tool for finding Cas9 target sites with low/no off-target potential. Given a DNA query (e.g., gene) sequence, the pipeline first fetches all the possible regions of length 23, structured as [5′-G/N(19 bps)PAM-3′]/[3′-PAM(19 bps)C/N-5′]. These regions are then evaluated and filtered to show low/no off-target activity across the whole genome (see Fig. 1a for general workflow). The choice of PAM sequence and CRISPR target start base is kept flexible. PAM sequence can be selected from NGG or NAG and CRISPR start base can be chosen as G or any (N) base. These options are also implemented in the standalone version of the software and can be used as arguments.

There are two levels of filters and passing both designates the region as a potential sgRNA for Cas9 activity. The first level accepts the target sequence and tags it "PASS" if it has less than or equal to 2 base-pair mismatch with the closest off-target in the genome. The second filter, accepts the target if the seed region (last 15 bases including the PAM sequence) has 0 mismatch with off-targets anywhere in the genome. The pipeline reports the instances which are accepted by both or either one of the filters, but the ones which passes both the filters are designated as potential Cas9 targets.

Once the potential sgRNAs are filtered, they are then checked for the presence of none, one or more restriction sites corresponding to pre-selected and most common restriction enzymes (Additional file 1: Figure S1). Cas9 enzyme cleaves both DNA strands ~3–4 bases upstream of the PAM [19, 25]. Therefore along with the presence of restriction sites on the entire target sequence, PhytoCRISP-Ex reports specifically, if any, the restriction sites overlapping three and four bases upstream to the PAM sequence. Embedding a restriction site in the target sequence will help screening for mutants using PCR after a digest with the appropriate restriction enzyme. Presence and absence of the restriction sites along with its position on the sgRNA are reported in the output file (Fig. 1b). The list of restriction enzymes (Additional file 2: Table S1) can be updated and used when using the standalone version of the software. The PhytoCRISP-Ex pipeline is mounted with the index of 13 genomes and is accessible via web-server. An off-line standalone software is also developed which gives liberty to its users to use it for any genome, assembled or un-assembled. The standalone package also includes a few example files and a README file to help users install and execute PhytoCRISP-Ex. The standalone version can be found under "Download" tab at http://www.phytocrispex.biologie.ens.fr/CRISP-Ex/.

Results

Several computational tools have been published aiming to identify candidates for CRISPR/CAS targeting. In order to validate the novelty and sustainability of PhytoCRISP-Ex, we compared it with 8 other previously published tools. Our criteria for comparison included first, the possibility to use the tool as a browser-based or stand-alone application, so to meet the need of a wide range of users. Second, whether it provides the flexibility to perform off-target analysis on any or a restricted list of genomes. The third criteria compares the possibility to perform restriction site mutant screening analysis. The summary of this comparison is presented in Table 1 and shows PhytoCRISP-Ex as a robust and adequately designed application. Further, we evaluated our tool against the whole exome of the model diatom *Phaeodactylum tricornutum* (http://protists.ensembl.org/Phaeodactylum_tricornutum/Info/Index/) and *Thalassiosira pseudonana* (http://genome.jgi.doe.gov/Thaps3/Thaps3.home.html). The statistics revealed that ~94 % and ~95 % of the genes in *P. tricornutum* and *T. pseudonana*, respectively, have atleast 1 potential guide RNA to be used as a Cas9 target against these genes. Among the latter, most of the genes have high percent efficiency in terms of constituting mostly potential targets among total predicted Cas9 targets (Additional file 1: Figure S1). These findings suggest that potentially almost any gene in these two species can be targeted with high probability for generating a single specific mutation. Out of the total, ~89 % genes in both the species (*P. tricornutum* and *T. pseudonana*) possess potential targets with restriction sites over the probable Cas9 cleavage site. Therefore, choosing such targets might help in fast screening of the mutants. PhytoCRISP-Ex is a

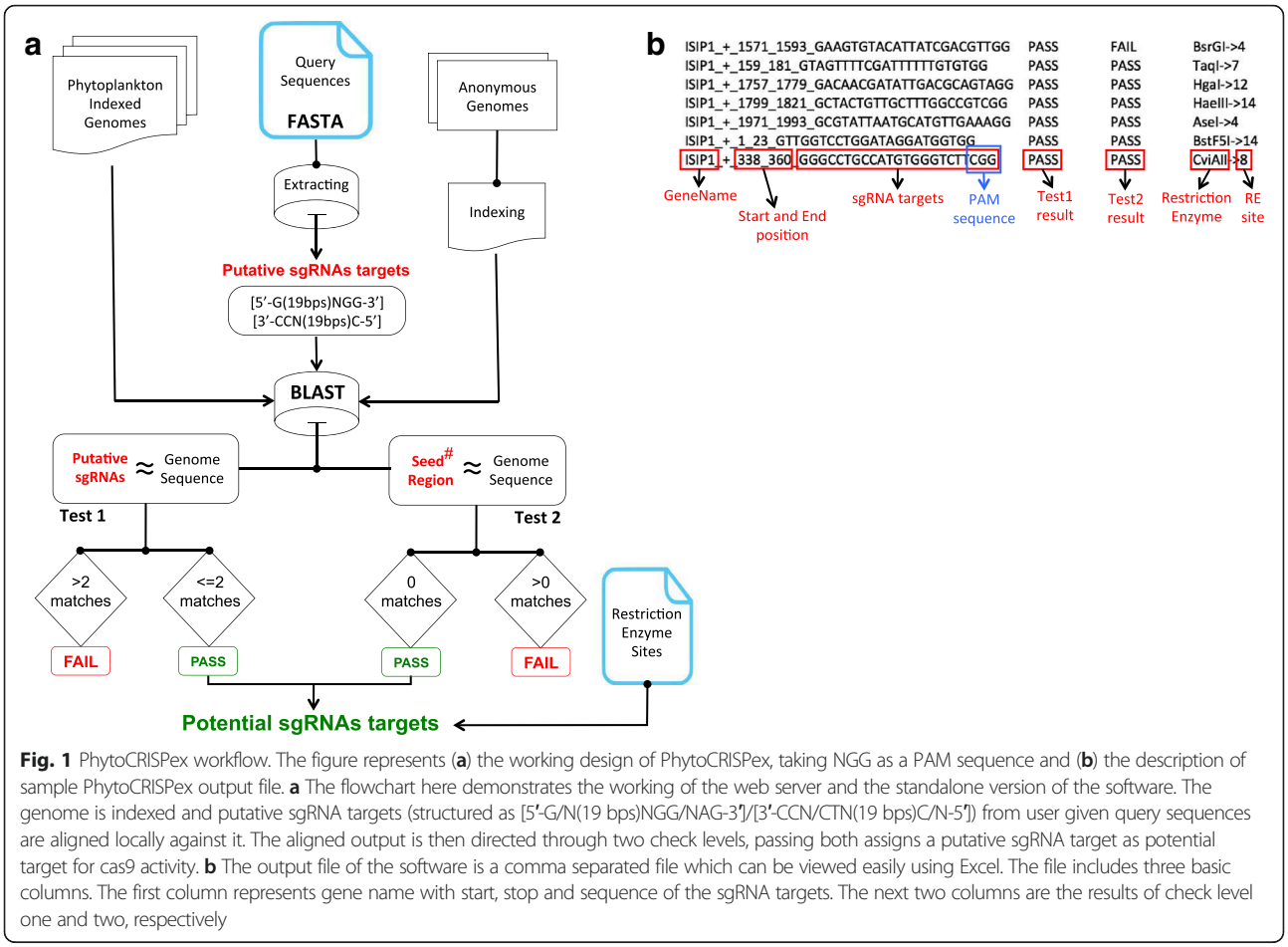


Fig. 1 PhytoCRISPex workflow. The figure represents (**a**) the working design of PhytoCRISPex, taking NGG as a PAM sequence and (**b**) the description of sample PhytoCRISPex output file. **a** The flowchart here demonstrates the working of the web server and the standalone version of the software. The genome is indexed and putative sgRNA targets (structured as [5′-G/N(19 bps)NGG/NAG-3′]/[3′-CCN/CTN(19 bps)C/N-5′]) from user given query sequences are aligned locally against it. The aligned output is then directed through two check levels, passing both assigns a putative sgRNA target as potential target for cas9 activity. **b** The output file of the software is a comma separated file which can be viewed easily using Excel. The file includes three basic columns. The first column represents gene name with start, stop and sequence of the sgRNA targets. The next two columns are the results of check level one and two, respectively

CRISPR/Cas9 target extraction web-package, built to extract targets using 13 model phytoplankton genomes. The algorithm has also been designed as a UNIX based standalone package which provides flexibility to its users to use it on other non-model genomes. The simple design of PhytoCRISP-Ex allows its use by end-users with moderate or no software programming background.

Conclusions

With the persuasive interest of scientific community towards phytoplankton research, the need of establishing genetic transformation tools for plankton species is thriving. PhytoCRISP-Ex provides a reliable and first ever application for predicting CRISPR/Cas9 targets within various plankton genomes. PhytoCRISP-Ex is also equipped with an easy to

Table 1 PhytoCRISP-Ex vs others

Softwares	Browser-based application	Stand-alone application	Restriction screening	Background Genome flexibility	Reference
PhytoCRISP-Ex	✓	✓	✓	✓	Current study
CRISPR MultiTargeter	✓	X	X	X	[18]
CasFinder	X	✓	X	X	[1]
CHOPCHOP	✓	X	X	X	[15]
CRISPRdirect	✓	X	X	X	[16]
E-CRISP	✓	X	X	X	[8]
sgRNAcas9	X	✓	X	✓	[27]
CasOT	X	✓	X	✓	[26]
CRISPRseek	X	✓	✓	✓	[28]

Comparison between PhytoCRISP-Ex and several previously published CRISPR target analysis tools

use, yet powerful, standalone version which gives its user the flexibility to use it on any genome. Many such and other unique features make the software more advance and appropriate for the use by a broad research community.

Additional files

Additional file 1: Figure S1. PhytoCRISP-Ex efficacy. The scatter plot depicts that most of the predicted Cas9 targets per gene in *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*, respectively, are potential candidates (passing both PhytoCRISP-Ex filter). X-axis represents the percent efficiency of each gene in terms of having high number of potential Cas9 targets compared to the total number of targets. Y-axis represents the number of all potential targets per gene. (PDF 182 kb)

Additional file 2: Table S1. The list of restriction enzymes being used by PhytoCRISP-Ex. (XLSX 13 kb)

Abbreviations

CAS, CRISPR associated protein; CRISPR, Clustered regularly-interspaced short palindromic repeats; DNA, Deoxyribo nucleic acid; PAM, Protospacer adjacent motif; PCR, Polymerase chain reaction; RNA, Ribo nucleic acid.

Acknowledgements

We thank Pierre Vincens for his help in mounting PhytoCRISP-Ex over the web-server. AR was supported by the Labex Memolife International doctoral program.

Availability of data and materials

Project name: PhytoCRISP-Ex
Project home page: http://www.phytocrispex.biologie.ens.fr/CRISP-Ex/
Compatible browsers: Mozilla, Internet Explorer, Chrome, etc.
Operating system(s): UNIX
Programming language: Shell, Perl, HTML
License: NA
Any restriction to use by non-academics: None

Authors' contributions

AR, OM and LT conceived and designed the structure of the software. AR wrote the software. AR, OM CB and LT wrote the manuscript. LT coordinated the study. All authors read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Received: 15 January 2016 Accepted: 22 June 2016
Published online: 01 July 2016

References

1. Aach JEA. CasFinder: Flexible algorithm for identifying specific Cas9 targets in genomes. bioRxiv. (2014);005074. doi:http://dx.doi.org/10.1101/005074.
2. Chari R, Mali P, Moosburner M, Church GM. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nat Methods. 2015;12:823–6.
3. de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, et al. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science. 2015;348:1261605.
4. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res. 2013;41:4336–43.
5. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016;34:184–91.
6. Feng Z, Zhang B, Ding W, Liu X, Yang DL, Wei P, Cao F, Zhu S, Zhang F, Mao Y, et al. Efficient genome editing in plants using a CRISPR/Cas system. Cell Res. 2013;23:1229–32.
7. Friedland AE, Tzur YB, Esvelt KM, Colaiacovo MP, Church GM, Calarco JA. Heritable genome editing in C. elegans via a CRISPR-Cas9 system. Nat Methods. 2013;10:741–3.
8. Heigwer F, Kerr G, Boutros M. E-CRISP: fast CRISPR target site identification. Nat Methods. 2014;11:122–3.
9. Hwang WY, Fu Y, Reyon D, Maeder ML, Kaini P, Sander JD, Joung JK, Peterson RT, and Yeh JR. Heritable and precise zebrafish genome editing using a CRISPR-Cas system. PLoS One. 2013;8, e68708.
10. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol. 2013;31:233–9.
11. Kuscu C, Arslan S, Singh R, Thorpe J, Adli M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. Nat Biotechnol. 2014;32:677–83.
12. Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, Chaffron S, Ignacio-Espinosa JC, Roux S, Vincent F, et al. Ocean plankton. Determinants of community structure in the global plankton interactome. Science. 2015;348:1262073.
13. Mali P, Esvelt KM, Church GM. Cas9 as a versatile tool for engineering biology. Nat Methods. 2013;10:957–63.
14. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, and Church GM. RNA-guided human genome engineering via Cas9. Science. 2013;339:823–6.
15. Montague TG, Cruz JM, Gagnon JA, Church GM, Valen E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. Nucleic Acids Res. 2014;42:W401–7.
16. Naito Y, Hino K, Bono H, Ui-Tei K. CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target sites. Bioinformatics. 2015;31:1120–3.
17. Niu Y, Shen B, Cui Y, Chen Y, Wang J, Wang L, Kang Y, Zhao X, Si W, Li W, et al. Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. Cell. 2014;156:836–43.
18. Prykhozhij SV, Rajan V, Gaston D, Berman JN. CRISPR multitargeter: a web tool to find common and unique CRISPR single guide RNA targets in a set of similar sequences. PLoS One. 2015;10, e0119372.
19. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. Nat Protoc. 2013;8:2281–308.
20. Rastogi A, Lin X, Lombard B, Loew D, Tirichine L. Probing the evolutionary history of epigenetic mechanisms: What can we learn from marine diatoms. AIMS Genet. 2015;2:173–91.
21. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science. 2015;348:1261359.
22. Tirichine L, Bowler C. Decoding algal genomes: tracing back the history of photosynthetic life on Earth. Plant J. 2011;66:45–57.
23. Tzur YB, Friedland AE, Nadarajan S, Church GM, Calarco JA, Colaiacovo MP. Heritable custom genomic modifications in Caenorhabditis elegans via a CRISPR-Cas9 system. Genetics. 2013;195:1181–5.
24. Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, and Jaenisch R. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell. 2013;153:910–8.
25. Wu X, Kriz AJ, Sharp PA. Target specificity of the CRISPR-Cas9 system. Quant Biol. 2014;2:59–70.
26. Xiao A, Cheng Z, Kong L, Zhu Z, Lin S, Gao G, Zhang B: CasOT: a genome-wide Cas9/gRNA off-target searching tool. Bioinformatics 2014.
27. Xie S, Shen B, Zhang C, Huang X, Zhang Y. sgRNAcas9: a software package for designing CRISPR sgRNA and evaluating potential off-target cleavage sites. PLoS One. 2014;9:e100448.
28. Zhu LJ, Holmes BR, Aronin N, Brodsky MH. CRISPRseek: a bioconductor package to identify target-specific guide RNAs for CRISPR-Cas9 genome-editing systems. PLoS One. 2014;9, e108424.

Conclusions and perspectives

"It is not the strongest of the species that survives,
nor the most intelligent,
but the one most responsive to change"

-Charles Darwin

Since the beginning of life on earth, phenomenon like speciation and selection has governed the establishment of new species and diversity within the tree of life. With time and ever-changing environment, the need of adaptation and the role of evolution became indispensable. Diatoms came into existence long after the life on earth was originated. It is suggested that diatoms emerged in the Jurassic era, approximately 190 million years (Myr) ago, long before homosapiens were evolved, and are now estimated to be one of the most diverse kingdom in the tree of life. They are found on all form of water bodies such as open oceans, polar waters, tropical waters, fresh water areas, snow and even glacial ice. Evolving as per their habitat, diatoms attained numerous shapes and sizes, which are grouped into centric and pennates and is based on their characteristic shape of cell frustule.

Phaeodactylum tricornutum is a raphid pennate diatom species and is classified under genus Phaeodactylum, where it is the only known species till date. *P. tricornutum* majorly resides in unsteady and/or nutrient limiting coastal areas like seashores, estuaries, rock pools, tidal creeks, etc. With its discovery in the late 18th century and efforts from numerous researchers, *P. tricornutum* is now extensively exploited as a model to understand, in general, diatom physiology, evolution and genome structure. To add on the existing experimental and theoretical resources that researchers have developed since a century, making *P. tricornutum* as a robust experimental model, a part of my thesis focused in advancing these resources (Chapter 1, 2 and 5), based on the current state-of-the-art technologies, and make the community access them with ease. Although the species is non-abundant in nature, it have been found and sampled from various parts of the world. Recent estimates of *P. tricornutum* abundance (Figure 1), around the world, have been estimated (personal communication by Shruti Malviya) from numerous locations which were the part of open sampling day (OSD) 2014 campaign (http://www.microb3.eu/osd). Among the 10 isolates (Pt1 – Pt10) of *P. tricornutum* that were sampled since 1902 till 2000 and are referred to as ecotypes, some have been frequently used in the past to understand the functional diversity of this species in response to various environmental cues. Using advanced Illumina sequencing technology, we sequenced 10 latter mentioned ecotype populations and established a genome-wide genetic diversity map (Chapter 3). Along with elucidating the population structure, this work unveils, for the first-time, the existence of multiple sub-species within the genus *Phaeodactylum* as a consequence of natural hybridization (Chapter 3). We were further interested in understanding the evolution of *P. tricornutum*, and the mystery behind the existence of its pleomorphic nature. Considering the dynamic nature of *P. tricornutum* cells in maintaining and transforming morphology from one form to another, we explored and advanced the epigenetic code (Chapter 1 and Chapter 5) underlying functional and genetic diversity between different ecotypes and morphotypes. We compared natural variation of H3K27me3 profile across two ecotypes, possessing two different morphotypes, of *P.*

tricornutum (Chapter 4). Using integrative-OMICS approach (Figure 2) we established the role of chromatin, precisely H3K27me3, in shaping the genome structure and governing the existence of adaptive phenotypes in *P. tricornutum*.

Based on the above findings, I propose a working evolutionary model/theory of morphogenesis in *P. tricornutum* where epigenetics provide a spontaneous and flexible control for acquiring various adaptive features, including morphological appearances, by the species under pressure (Figure 3). This flexibility allows the acquired feature to have reversible properties in the changing or fluctuating environmental conditions. Once these conditions get stable and the attained phenotype is selected, the underlying genetic makeup driven by epigenetics gets fixed in the population. The role of environment is well known in inducing various epigenetic changes in an organism resulting in the regulation of various chromatin states and adaptive acquisitions. We are familiar with the role of DNA methylation in regulating the expression of various loci, providing multiple elements for the evolutionary forces to act upon, causing high rates to C to T mutations genome-wide. In the current work we attempted to understand the role of H3K27me3 in providing a firm platform for the evolutionary forces to act upon. We also report the role of H3K27me3 in maintaining different expression profiles of genes involved in cell wall maintenance and morphogenesis in *P. tricornutum*.

The thesis work did not aim to explore specific environmental variables that induce morphogenesis in *P. tricornutum* and do not provide direct evidences that epigenetic governed mechanisms, like gene expression control, are adaptable to the changing environmental forces. Therefore, in light of the described model, an immediate perspective of the work lies in understanding the molecular control of gene expression within the population of *P. tricornutum* leading to the selection of certain loci or alleles, in its dynamically changing niche. Based on the recent findings (ongoing project) we discovered multiple genes with mono-allelic expression (Chapter 3),

advancing our knowledge to another horizon of molecular complexity that lies within single cell micro-eukaryotes like *P. tricornutum*. In a collaborative work environment, our future plan is to obtain experimental evidences to confirm H3K27me3 role in regulating genes involved in cell morphology and to establish the role and regulation of bi vs mono-allelism in *P. tricornutum* using newly sampled cell population from China which is characterized to have majorly the cruciform morphology. I vision to expand my knowledge of the molecular complexity within single cell organisms, which we deciphered using *P. tricornutum* as a model species, to understand the unprecedented diversity within diatoms using meta-data from TARA sampling project.

Conclusively, using integrative-OMICS approach my thesis provides new insights into the biology of extensively studied model diatom species *Phaeodactylum tricornutum*. Along with updating functional annotations (Phatr3) and identifying new gene models using current state-of-the-art technologies, I developed PhytoCRISP-Ex which will assist experimental biologist to perform reverse genetic studies in *P. tricornutum*. I discovered the presence of multiple sub-species within the genus *Phaeodactylum*, where *P. tricornutum* is classified as the only species. To do so, I established a genetic variant map across 10 most studied strains of *P. tricornutum*, providing a platform for future functional studies, including genome wide association analysis. I also discovered the molecular complexity underlying the expression control in single cell micro-eukaryotes, by identifying genes exhibiting bi and mon-allelic expression in the model diatom species *P. tricornutum*. After deciphering the histone code within *P. tricornutum*, I also studied the natural variation of H3K27me3 distribution across two ecotypes, possessing two different morphotypes. The study strengthens our knowledge over the role of H3K27me3 in regulating the genes associated with cell morphology and governing the genome structure. All these studies put together have advanced our knowledge about the molecular complexity that underlies the success of *P. tricornutum* survival and adaptation.

Figures

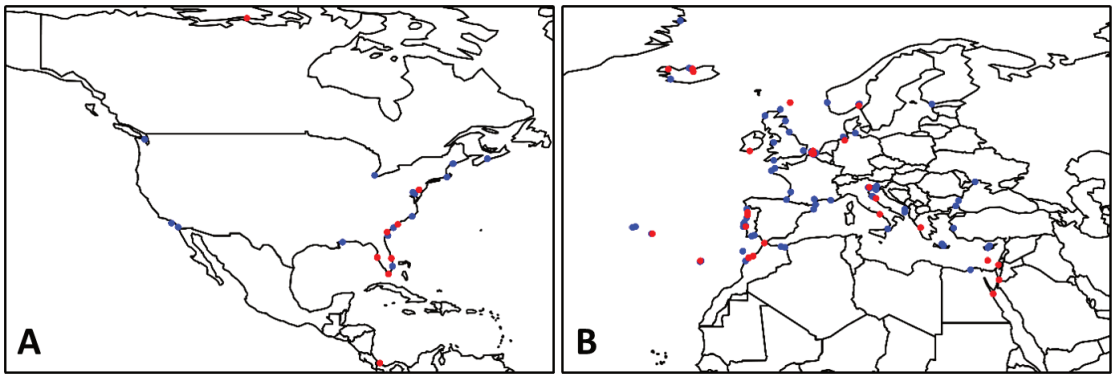


Figure 1. OSD 2014 sampling locations from geographic area in and around, United States of America (A) and Europe (B). Red dots indicate locations where the signals of *P. tricornutum* presence is achieved. Red dot indicate all other sampling locations.

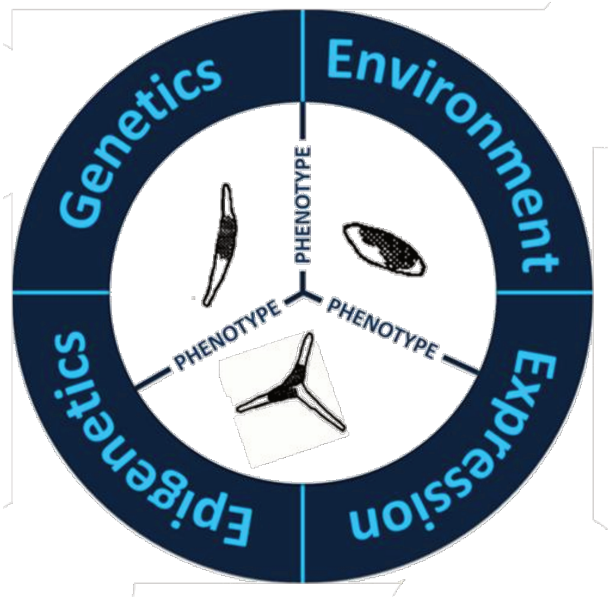


Figure 2. The components of integrative-OMICS approach.

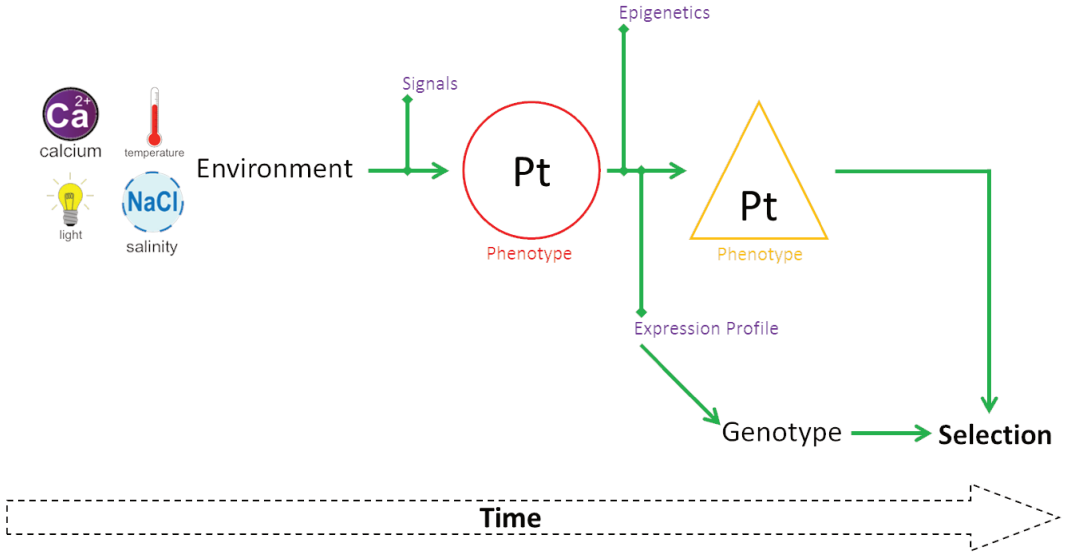


Figure 3. Evolutionary model of morphogenesis in *Phaeodactylum tricornutum*

Résumé

Depuis la découverte de *Phaeodactylum tricornutum* par Bohlin en 1897, sa classification au sein de l'arbre de la vie a été controversée. En utilisant des morphotypes ovales et fusiformes Lewin a décrit en 1958 plusieurs traits caractéristiques de cette espèce rappelant la structure des diatomées mettant ainsi fin à la controverse sur la classification de *P. tricornutum* au sein des Bacillariophycées. Pour se faire, trois morphotypes (ovale, fusiforme et triradié) de *Phaeodactylum tricornutum* ont été observés. Au cours d'une centaine d'années environ, de 1908 à 2000, 10 souches de *Phaeodactylum tricornutum* (appelées écotypes) ont été collectées et stockées soit de manière axénique ou en l'état avec leur populations naturelles de bactéries dans les centres des ressources génétiques pour algues, cryo-préservées quand cela est possible. Divers outils cellulaires et moléculaires ont été établis pour disséquer et comprendre la physiologie et l'évolution de *P. tricornutum*, et/ou les diatomées en général. Grâce à des décennies de recherche et les efforts déployés par de nombreux laboratoires que *P. tricornutum* est aujourd'hui considérée comme une espèce modèle des diatomées.

Le sujet de ma thèse traite majoritairement de la composition génétique et épigénétique du génome de *P. tricornutum* ainsi que de la diversité morphologique et physiologique sous-jacente au sein des populations naturelles prospectées à différents endroits du globe. Pour se faire, j'ai généré les profils chromatiniens en utilisant différentes marques des modifications post-traductionnelles des histones (chapitres 1 et 2) et a également comparé la variation naturelle dans la distribution de certaines marques clés entre deux populations d'écotypes (chapitre 4). Nous avons également généré une carte de la diversité génétique à l'échelle du génome chez 10 écotypes de *P. tricornutum* révélant ainsi la présence d'un complexe d'espèces dans le genre *Phaeodactylum* comme la conséquence d'une hybridation ancienne (chapitre 3). Sur la base de nombreux rapports antérieurs et des observations similaires au sein de *P. tricornutum*, nous proposons l'hybridation naturelle comme une base solide et une possibilité plausible pour expliquer la diversité des espèces chez lest diatomées. De plus, nous avons mis à jour les annotations fonctionnelles et structurelles du génome de *P. tricornutum* (Phatr3, chapitre 2) et mis au point un algorithme de logiciel convivial pour aller chercher les cibles CRISPR du système d'édition du génome CRISPR / cas9 chez 13 génomes de phytoplancton incluant *P. tricornutum* (chapitre 5). Pour accomplir tout cela, j'ai utilisé diverses méthodes à la pointe de l'état de l'art comme la spectrométrie de masse, l'immuno-précipitation de la chromatine suivie de séquençage à haut débit ainsi que les séquençages du génome entier, de l'ARN et des protocoles d'édition du génome CRISPR et plusieurs logiciels / pipelines de calcul. Ainsi, le travail de thèse fournit une plate-forme complète qui pourra être utilisée à l'avenir pour des études épigénétiques, de génétiques moléculaires et fonctionnelles chez les diatomées en utilisant comme espèce modèle *Phaeodactylum tricornutum*. Ce travail est pionnier et représente une valeur ajoutée importante dans le domaine de la recherche sur les diatomées en répondant à des questions nouvelles ouvrant ainsi de nouveaux horizons à la recherche en particulier en épigénétique qui joue un rôle important mais pas encore assez apprécié dans le succès écologique des diatomées dans les océans actuels.

Mots clés

Phaeodactylum tricornutum, Diatomées, Epigénétique, Génomique des Populations, Morphogenèse, Logiciels, Annotations fonctionnelles, H3K27me3, CRISPR/cas9.

Abstract

Since the discovery of *Phaeodactylum tricornutum* by Bohlin in 1897, its classification within the tree of life has been controversial. It was in 1958 when Lewin, using oval and fusiform morphotypes, described multiple characteristic features of this specie that resemble diatoms structure, the debate to whether classify *P. tricornutum* as a member of Bacillariophyceae was ended. To this point three morphotypes (oval, fusiform and triradiate) of *Phaeodactylum tricornutum* have been observed. Over the course of approximately 100 years, from 1908 till 2000, 10 strains of *Phaeodactylum tricornutum* (referred to as ecotypes) have been collected and stored axenically as cryopreserved stocks at various stock centers. Various cellular and molecular tools have been established to dissect and understand the physiology and evolution of *P. tricornutum*, and/or diatoms in general. It is because of decades of research and efforts by many laboratories that now *P. tricornutum* is considered to be a model diatom species.

My thesis majorly focuses in understanding the genetic and epigenetic makeup of *P. tricornutum* genome to decipher the underlying morphological and physiological diversity within different ecotype populations. To do so, I established the epigenetic landscape within *P. tricornutum* genome using various histone post-translational modification marks (chapter 1 and chapter 2) and also compared the natural variation in the distribution of some key histone PTMs between two ecotype populations (chapter 4). We also generated a genome-wide genetic diversity map across 10 ecotypes of *P. tricornutum* revealing the presence of a species-complex within the genus *Phaeodactylum* as a consequence of ancient hybridization (Chapter 3). Based on the evidences from many previous reports and similar observations within *P. tricornutum*, we propose natural hybridization as a strong and potential foundation for explaining unprecedented species diversity within the diatom clade. Moreover, we updated the functional and structural annotations of *P. tricornutum* genome (Phatr3, chapter 2) and developed a user-friendly software algorithm to fetch CRISPR/Cas9 targets, which is a basis to perform knockout studies using CRISPR/Cas9 genome editing protocol, in 13 phytoplankton genomes including *P. tricornutum* (chapter 5). To accomplish all this, I used various state-of-the-art technologies like Mass-Spectrometry, ChIP sequencing, Whole genome sequencing, RNA sequencing, CRISPR genome editing protocols and several computational softwares/pipelines. In brief, the thesis work provides a comprehensive platform for future epigenetic, genetic and functional molecular studies in diatoms using *Phaeodactylum tricornutum* as a model. The work is an add-on value to the current state of diatom research by answering questions that have never been asked before and opens a completely new horizon and demand of epigenetics research that underlie the ecological success of diatoms in modern-day ocean.

Keywords

Phaeodactylum tricornutum, Diatoms, Epigenetics, Population Genomics, Morphogenesis, Software, Functional annotations, H3K27me3, CRISPR/Cas9.