Developing ribosome profiling for the marine model diatom *Thalassiosira pseudonana*

Monica Pichler

100225613/1

A thesis submitted for the degree of Doctor of Philosophy

University of East Anglia, Norwich, UK

School of Environmental Sciences

October 2023

Abstract

Diatoms are unicellular, eukaryotic microalgae which have evolved a vast number of regulatory mechanisms to adapt their protein synthesis in response to the changing environmental conditions they live in. Due to their key role in marine biochemical cycling and their high biotechnological potential, gene regulatory mechanism in diatoms have been extensively studied using transcriptomics and proteomics. However, regulation of protein synthesis on the translational level is largely unexplored. A ribosome profiling protocol has been developed for the model diatom *Thalassiosira pseudonana* which allows genome-wide analysis of translation in this globally important phytoplankton by deep sequencing of ribosome protected mRNA fragments. This method has been applied to high light stressed cells to better understand the role of translational regulation in response to changing environments. The generated dataset is the first ribosome profiling data for any marine microalgae and expands the molecular toolbox.

To study the impact of codon usage change on protein synthesis in diatoms, the codon usage of the light-harvesting complex associated Lhcx6 gene of *T. pseudonana* has been optimized via CRISPR/Cas9-mediated homologous recombination. Efficient gene targeting using this method has already been achieved in this species. However, this is the first time that a gene was replaced with a non-selective marker gene. Phenotyping of homozygous knock-in cell lines gives first insights into the role of codon usage in diatoms and provides preliminary data for potential future work.

Table of Contents

List of Figures

List of Tables

Preface

All contributions to this thesis of members of the Mock lab and other collaborators are outlined below.

Chapter 1 gives a brief introduction to diatoms and the field of diatom genomics. It also presents the model diatom *Thalassiosira pseudonana*, which is the diatom of interest throughout this thesis.

Chapter 2 and 3 describe the development and application of a ribosome profiling protocol for *Thalassiosira pseudonana.* Preliminary ribosome profiling data used to design rRNA depletion oligos was generated by Master student Annemarie Eckes and postdoctoral researcher Amanda Hopes. I prepared the ribosome profiling sequencing libraries together with PhD student Andreas Meindl and technician Markus Romberger who both work at the Medenbach lab at the University of Regensburg.

Chapter 4 describes the modification of codon usage for two genes in *T. pseudonana* via CRISPR/Cas mediated homologous recombination. Codon usage of genes RPL10a and Lhcx6 were previously modified by postdoctoral researcher Amanda Hopes. She also designed the Lhcx6 sgRNAs as well as sgRNA_RPL10a_AH. She further assembled both HR plasmids and the Lhcx6 CRISPR plasmid. I designed off-target site specific primers together with Master student Ji Nah. He performed PCR amplification and sequencing of potential off-target sites under my supervision.

Chapter 1: Introduction

1.1 Introduction to Diatoms

Diatoms are important microalgae found throughout the world's ocean and freshwater environments (Field et al., 1998). They are unicellular photosynthetic eukaryotes playing a critical role in regulating the global carbon cycle via the ocean's biological pump. Diatoms are responsible for about 40% of marine productivity, thus are strongly influencing atmospheric CO_2 levels (Nelson et al., 1995). Besides their key role in global carbon fixation, marine food webs and biochemical cycling of nutrients, diatoms also show high biotechnological potential (Dolatabadi and de la Guardia, 2011). They are characterised by their silicified cell walls, called frustules, displaying intricate morphologies which are widely applied in nanotechnology, including biosensing, drug and gene delivery (Dolatabadi and de la Guardia, 2011; Jeffryes et al., 2011). In addition, substances harvested from diatoms find wide applications in biofuel production, cosmetics and the food industry (Sharma et al., 2021).

Based on valve symmetry, diatoms are generally divided into two classes: the centrics, which are circular with a radial symmetry and the pennates, which are elongated and display a bilateral symmetry. In both groups, the frustule is composed of two valves called epivalve (upper valve) and hypovalve (lower valve), whereas the upper valve is slightly larger. The two valves of the diatom frustule are linked by multiple silica bands called girdle bands (Figure 1.1) (Fu et al., 2022).

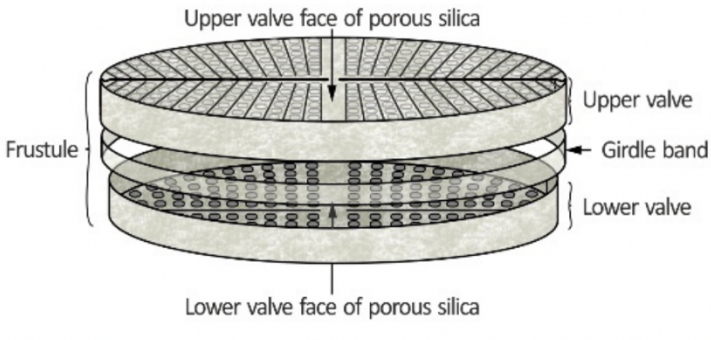


Figure 1.1: Overview of a diatom frustule. The frustule is composed of an upper and a lower valve which are linked by girdle bands (Allen et al., 2020).

Diatoms predominately undergo vegetative reproduction by mitotic cell division. Thereby, each daughter cell receives one valve and synthesizes a new one inside the existing valve. Every division thus results in an average cell size reduction of the diatom population (Bowler et al., 2010). Once a critical size threshold is reached, cell size is usually restored via auxospore formation resulting from sexual reproduction (Chepurnov et al., 2004). This further increases genetic variability and preventing clonal death (Moore et al., 2017). However, some diatom species seem to be able to avoid mitotic size reduction (Koester et al., 2018).

Diatoms form one of the most diverse groups of phytoplankton with an estimated species richness of ≥ 100.000 (Mann and Vanormelingen, 2013). Diatoms evolved around the Jurassic period with the first recorded fossil records dating back to ~190 million years ago (Behrenfeld et al., 2021). This was followed by extensive diversification with centrics evolving in the Jurassic and early Cretaceous, followed by pennates in the late Cretaceous (Litchman and Klausmeier, 2008).
Diatom cell size ranges from 2 µm in the smallest species to a maximum diameter of 2 mm in the largest ones (Halse and Syvertsen, 1996). The most characteristic feature of diatoms is their elaborately and beautifully decorated frustule, which can vary widely in ornamentation and has fascinated scientists for a long time (Figure 1.2).

Figure 1.2: The diversity of diatoms illustrated by Ernst Haeckel in 'Kunstformen der Natur' (1904).

1.2 Diatom genomics

The field of diatom genomics has emerged in 2004 with the publication of the genome of the centric diatom *Thalassiosira pseudonana* (Armbrust et al., 2004a). The second diatom genome, of the pennate species *Phaeodactylum tricornutum,* was published in 2008 (Bowler et al., 2008). These two quickly became the model organisms providing valuable insights into diatom evolution, ecology and diversity (Bowler et al., 2009). Following the publication of the first two reference genomes, several more diatom genomes have been released including the first genome of a polar diatom, *Fragilariopsis cylindrus* (Mock et al., 2017). The '100 Diatom Genomes Project' which aims to sequence 100 diatom species from across the tree of life (https://jgi.doe.gov/csp-2021-100-diatom-genomes/), will drastically increase the number of available diatom reference genomes and facilitate genomic studies.

The success of diatoms is largely based on their evolutionary history which gave rise to their mosaic genome consisting of genes from different organism – host, plastids and bacteria (Armbrust et al., 2004a; Bowler et al., 2008). While Chlorophytes, glaucophytes and rhodophytes are derived from a primary endosymbiotic event in which a eukaryote took up a plastid by engulfing a cyanobacteria, diatoms (belonging to the heterokontophytes) are derived from a secondary endosymbiosis (Figure 1.3) (Hopes and Mock, 2015). This secondary endosymbiosis event involves a heterotrophic eukaryote acquiring a plastid by engulfing a photosynthetic eukaryote, likely red algae (Armbrust et al., 2004a). A substantial number of genes derived from green algae which were discovered in diatom genomes provide evidence for another secondary endosymbiosis event having occurred, most likely predating the acquisition of a red alga (Moustafa et al., 2009). Due to this evolutionary history, plastids found in diatom are distinguished by having four membranes with the outermost membrane being continuous with the endoplasmic reticulum (Dorrell et al., 2022). Besides genes acquired via endosymbiotic gene transfer (EGT), diatom genomes also comprise of bacterial genes likely derived from horizontal gene transfer (HGT) (Bowler et al., 2008), however the extent of it is still an ongoing debate (Mock et al., 2022).

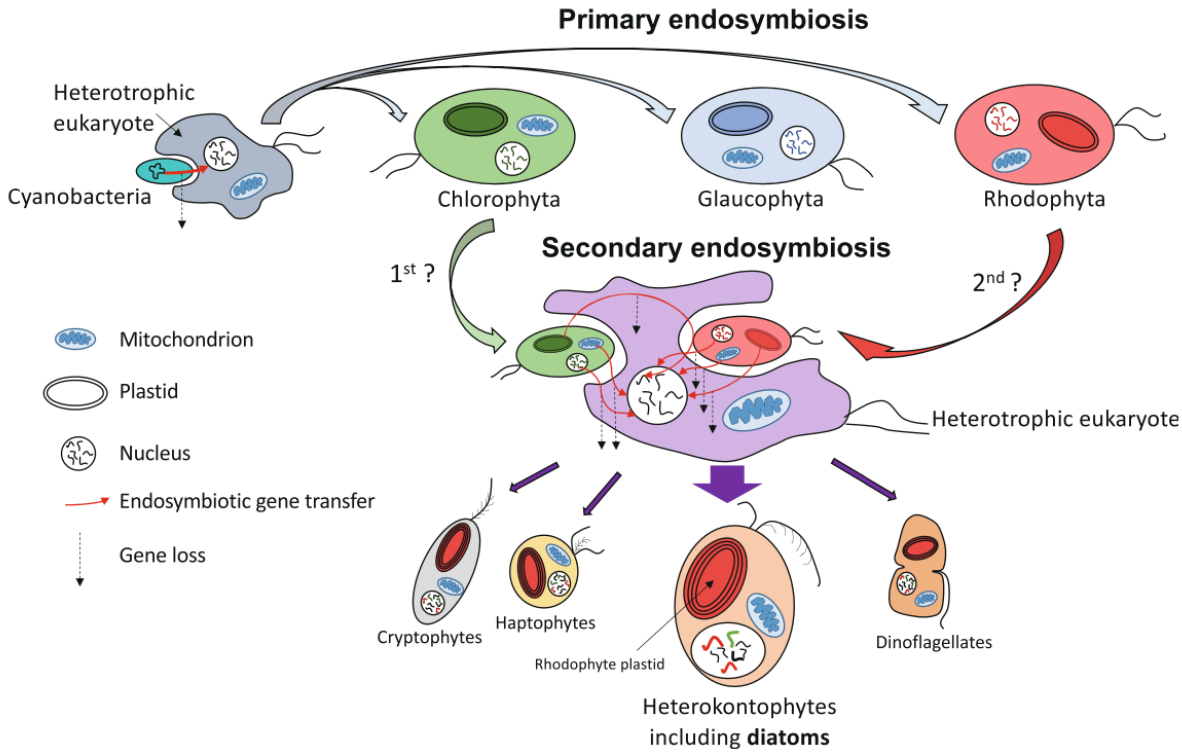


Figure 1.3: Evolution of diatoms according to primary and secondary endosymbiotic events. Primary endosymbiosis describes the process of a eukaryote engulfing a cyanobacteria and acquiring the plastid. A debate is still ongoing about secondary endosymbiotic events giving rise to heterokontophytes including diatoms. New evidence suggests that a heterotrophic eukaryote acquired a green alga prior of a red alga. Endosymbiotic gene transfer led to the loss and transfer of genes from the endosymbiont to the host genome and is visible in the mosaic genomes of diatoms, contributing to their metabolic plasticity (Mock et al., 2022).

More details of molecular analyses in diatoms as well as mechanisms and drivers of their genome evolution and adaptation leading to the success of this class can be found in the published book chapter 'Structure and Evolution of Diatom Nuclear Genes and Genomes' (Appendix A).

1.3 The model diatom Thalassiosira pseudonana

The diatom of interest throughout this thesis is *Thalassiosira pseudonana*. It is a centric diatom and belongs to the genus *Thalassiosira* which is globally distributed in temperate environments. *T. pseudonana* is heavily silicified and one of the smaller diatom species with a cell diameter of less than 10 µm (Figure 1.4) (Tirichine et al., 2017).

The species was selected for this thesis because it is a model diatom for physiology studies (Poulsen and Kröger, 2004) with the first sequenced genome of any eukaryotic marine phytoplankton (Armbrust et al., 2004a). *T. pseudonana* has a relatively small nuclear genome size of 34 Mb, consisting of 24 chromosomes and a predicted total of 11,242 protein-coding genes (Armbrust et al., 2004a). Novel insight of the sequenced genome included the identification of multiple transporters for the acquisition of inorganic nutrients and a range of metabolic pathways which contribute to the success of diatoms. The presence of a complete urea cycle was one of the most unexpected findings. This has never been described in any eukaryotic phototroph before and allows fast recovery from prolonged nitrogen limitation (Armbrust et al., 2004a; Allen et al., 2011).

The first genetic transformation system for *T. pseudonana* was published in 2006 (Poulsen et al., 2006), and the first successful editing via CRISPR/Cas was reported ten years later (Hopes et al., 2016). Extensive genetic research and the availability of several molecular tools established *T. pseudonana* as a model organism.

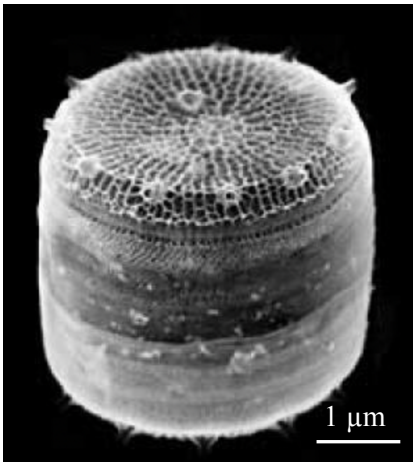


Figure 1.4: Scanning electron microscopy image of the model diatom Thalassiosira pseudonana. This centric diatom is characterized by a heavily silicified cell wall. Image credit Nils Kröger, Georgia Institute of Technology, Atlanta (USA).

1.4 Introduction to ribosome profiling and the study of translation

1.4.1 Eukaryotic translation of mRNA

Translation is one of the fundamental processes in biology and the most energetically costly, consuming approximately half the energy expended in rapidly growing cells (McGlincy and Ingolia, 2017). Across all domains of life, protein synthesis is performed by ribosomes which translate mRNA transcripts into functional proteins (Schuller and Green, 2018). Eukaryotic ribosomes (80S) are complex macromolecules composing of a small 40S (SSU) and a large 60S subunit (LSU). They consist of more than 5500 nucleotides of RNA (SSU, 18S rRNA; LSU, 5S, 5.8S, and 28S rRNA) and 80 proteins and comprise three tRNA binding sites: Aminoacyl-site (A-site), Peptidyl-site (P-site) and Exit site (E-site) (Wilson and Doudna Cate, 2012). The translation process occurs in four main stages: initiation, elongation, termination and ribosome recycling (Figure 1.5). During translation initiation, several eukaryotic initiation factors (eIFs) bind to the small subunit of the ribosome. This 43S pre-initiation complex together with an initiator methionyl-tRNA (Met-tRNA^{iMet}) then binds to the 5' untranslated region (5'UTR) and scans it for a start codon. Once the AUG start codon is recognized, the large ribosomal subunit joins to assemble the 80S ribosome with the initiator tRNA bound in its P-site (Clancy and Brown, 2008). During elongation, the ribosome moves along the mRNA in the 5'-3'direction, three nucleotides at a time, translating it into a protein through the actions of tRNAs and the involvement of eukaryotic elongation factors (eEFs). Aminoacyl-tRNAs (aa-tRNAs) are charged with their corresponding amino acid and can read the triplet code of the mRNA through complementary base-pairing in the ribosomal A-site (Clancy and Brown, 2008). The tRNA

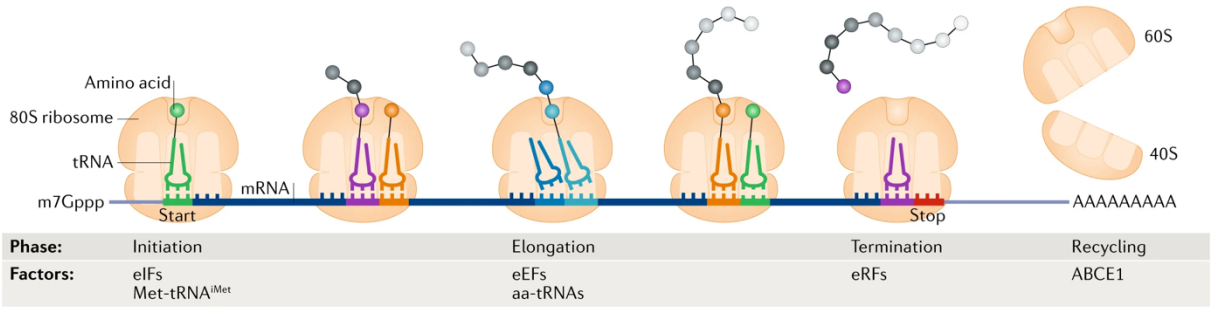


Figure 1.5: Overview of eukaryotic translation. Translation gets initiated at the start codon of an open reading frame by a coordination of many eukaryotic initiation factors (eIFs), initiator methionyl-tRNA (Met-tRNA^{iMet}) and the ribosomal subunits. During elongation, a polypeptide chain gets synthesized using eukaryotic elongation factors and aminoacyl-tRNAs (aa-tRNAs). Once the ribosome reaches a stop codon, the peptide gets released through the action of eukaryotic peptide chain release factors (eRFs) and the subunits get recycled by ATP-binding cassette sub-family E member 1 (ABCE1) (Schuller and Green, 2018).

molecule gets shifted to the P-site where its amino acid gets transferred to the growing polypeptide chain. After another translocation by the ribosome, the empty tRNA now occupies the E-site from where it gets released back to the cytoplasm allowing another aa-tRNA to bind at the A-site and repeating the process (Clancy and Brown, 2008). Termination occurs when a stop codon (UAA, UAG or UGA) is encountered, which, along with eukaryotic peptide chain release factors (eRFs), triggers the release of the peptide from the ribosome (Kapp and Lorsch, 2004; Schuller and Green, 2018). In the final recycling phase, the 80S ribosome complex gets dissociated into its subunits by ATP-binding cassette subfamily E member 1 (ABCE1) and a new translation cycle can start (Schuller and Green, 2018).

The standard genetic code is followed by all organisms to ensure correct translation of mRNA into protein. It describes how triplets of DNA nucleotides (codons) correspond to a certain amino acid which are then connected with peptide bonds to form proteins. As shown in Figure 1.6, the genetic code is degenerate, because a single amino acid can be encoded by more than one codon. Synonymous codons differ from each other in the third codon position and do not alter the amino acid sequence of a protein (Brule and Grayhack, 2017).

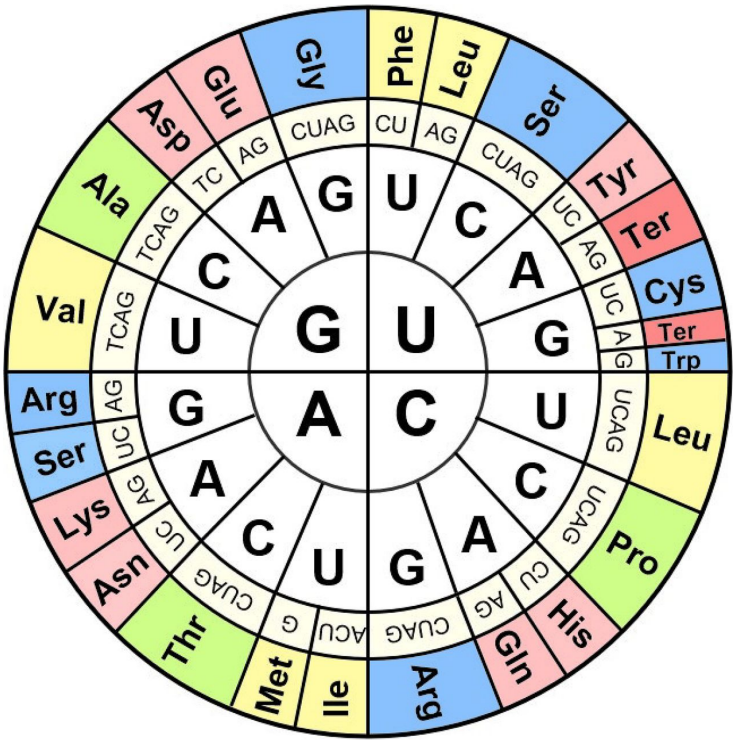


Figure 1.6: The codon wheel. Starting from the center, each amino acid is specified by three nucleotides (Saier, 2019).

1.4.2 Mechanisms of translational regulation

Gene expression is regulated at multiple levels, including the regulation at the transcriptional level. However, translational control of existing mRNA allows for more rapid changes of cellular protein levels (Sonenberg and Hinnebusch, 2009). Translational regulation in eukaryotes is a complex process which can generally be divided into global control, which regulates the translation of most mRNAs in the cell, and mRNA-specific control, which modulates the translation of a particular group of mRNAs without affecting general protein biosynthesis or the translational status of the overall cellular transcriptome (Gebauer and Hentze, 2004). Global control of translation is often achieved by phosphorylation or by modulating the availability of eIFs (King and Gerber, 2016). A well-known example is the phosphorylation of eIF2α, which reduces the levels of active initiation complexes and thus leading to a rapid reduction in mRNA translation (King and Gerber, 2016). Another example is the regulation of the availability of the cap-binding protein eIF4E, which is controlled by 4E-binding proteins (4E-BPs). These proteins compete with eIF4G for binding to eIF4E, leading to inhibition of the association between the 43S complex and the mRNA, which results in translational repression (Gebauer and Hentze, 2004). Translation of specific mRNAs can be regulated by RNA-binding proteins (RBP) which interact with sequences or structures located in the UTRs (Imig et al., 2012). The protein bound to the mRNA can compete directly with ribosome binding in some cases, while in others, it can promote the formation of an RNA structure that inhibits ribosome binding or trap the ribosome in a complex that prevents the initiation of translation (Babitzke et al., 2009). Non-coding RNAs, such as microRNAs (miRNAs) are important regulatory components. These ∼22-nucleotide-long RNA molecules are processed from larger precursors and finally incorporated into the RNA-induced silencing complex (RISC). They can repress translation or promote mRNA decay by guiding the RISC complex to target partially complementary sequences located in the 3'UTR (Imig et al., 2012). Internal ribosome entry sites (IRES) are another means of translation control. These structures are located in the 5'UTR and can mediate translational initiation independently of the cap structure by recruiting the ribosome directly to an internal position of the mRNA (Gebauer and Hentze, 2004). Upstream ORFs (uORFs), short protein-coding regions, which are present in the 5'UTR, can repress translation of the main open reading frame (ORF) via ribosome stalling, inhibition of translation reinitiation or uORF induced nonsense-mediated decay (NMD) (Ruiz-Orera and Alba, 2019). Figure 1.7 displays all of the above mentioned elements that influence translation of mRNA.

Although translation initiation is the primary target of regulation, it is possible to regulate the translation elongation phase as well. The presence of rare codons within the coding sequence of an mRNA can decrease the elongation rate to such an extent that initiation is no longer the rate-limiting factor. This can ultimately lead to frame-shifting at specific regions of some mRNAs leading to the production of a protein with a different sequence and length than the unshifted version (Hershey et al., 2012).

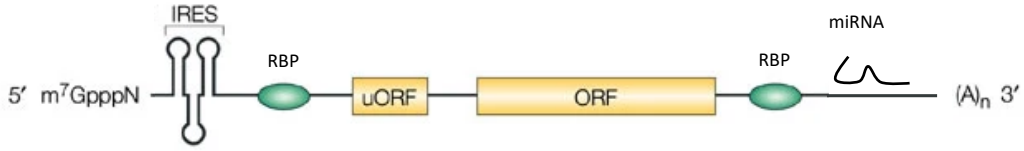


Figure 1.7: Structural features influencing translation of mRNA. RNA-binding proteins (RBPs), micro RNAs (miRNAs), upstream open reading frames (uORFs) and internal ribosome entry sequences (IRESs) all play regulatory roles during translation initiation. Adapted from Gebauer and Hentze, 2004.

1.4.3 Translational regulation in response to stress in plants and algae

Under stress conditions, the levels of translation of a large number of transcripts have to be regulated rapidly. In *Arabidopsis thaliana*, phosphorylation of eIF2α has been linked to the down-regulation of translation in response to oxidative stress, amino acid and purine starvation, UV radiation, cold stress, wounding, cadmium treatment, and various phytohormones such as jasmonate, ethylene, and salicylic acid (Merchante et al., 2017; Sesma et al., 2017). However, eIF2α phosphorylation does not appear to be involved in plant responses to heat or osmotic stresses (Merchante et al., 2017). In maize and wheat, the eIF4A helicase can undergo phosphorylation in response to hypoxia and heat shock (Webster et al., 1991; Le et al., 1998). The target of rapamycin (TOR) signaling pathway plays a major role in plant responses to various stress factors (Fu et al., 2020). Through regulation of the S6 kinase (S6K) activity, the TOR pathway is significantly involved in the phosphorylation of the ribosomal protein S6 (RPS6), which is critical for controlling of translation initiation (Muench et al., 2012). It was shown that *Arabidopsis* subjected to osmotic stress showed a reduction in S6K activity (Mahfouz et al., 2006). Moreover, strong down-regulation of the *Arabidopsis* TOR resulted in reduced translation rates, mimicking the effect of the plant hormone abscisic acid (Deprost et al., 2007). These findings suggest a close association between the TOR signaling pathway and environmental cues in plants. Figure 1.8 provides a summary of how developmental and

environmental cues regulate translation initiation through the TOR pathway, eIF4E activity, and eIF2a phosphorylation, as detailed by Sesma et al., (2017). Translation regulation in plants and algae is still elusive with many of the detailed molecular mechanisms involved often remaining unknown.

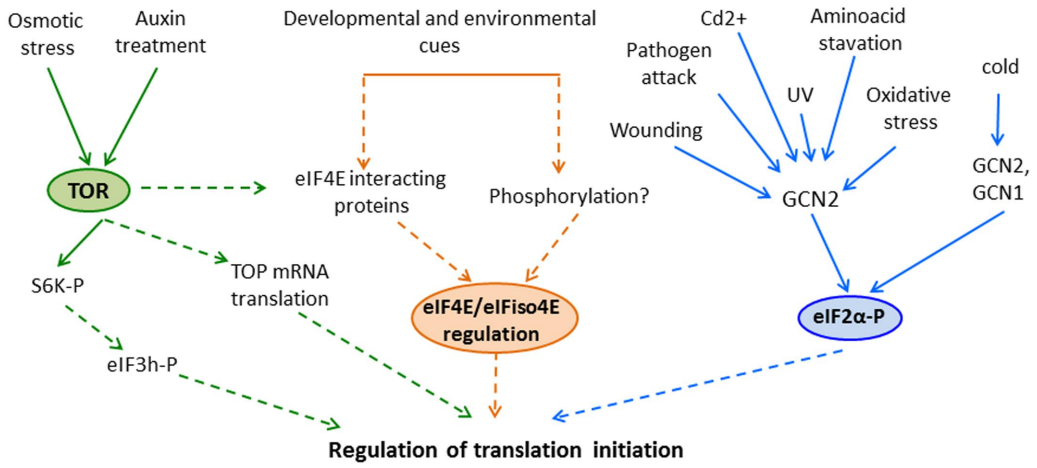


Figure 1.8: Regulation of translation by TOR, eIF4E and eIF2α in plants. Various treatments trigger the activation of TOR and the general control non-derepressible 2 (GCN2) in plants, which promotes S6K and eIF2α phosphorylation, respectively. The functioning of eIF4E and eIFiso4E might also be subject to regulation, but the precise mechanisms are not fully comprehended. In general, translational regulation in plants is still poorly understood. Solid lines represent confirmed connections between processes based on experimental evidence, while dashed lines indicate possible connections that are either missing or remain unresolved within the context of plants (Sesma et al., 2017).

Light is an important environmental signal with immense effects on the development and physiology of photosynthetic organisms, thus they regulate translation in response to this factor (Merchante et al., 2017). In response to fluctuating light conditions, translational regulation is adapted to minimize photodamage and to efficiently harvest light (Chotewutmontri and Barkan, 2018). To study translation in response to light, *Arabidopsis* seedlings were subjected to an unanticipated 1 h period of darkness which resulted in a 17% decrease of global translation levels consistent with inhibition of translation initiation. Among those over 1600 mRNAs which rapidly changed their translation status in response to light availability was the light-harvesting chlorophyll-protein complex A4 (LHCA4) and the ribosomal protein large subunit 12A (RPL12A). This rapid down-regulation of translation was reversible within 10 min of being re-illuminated (Juntawong and Bailey-Serres, 2012).

Ribosome profiling in *Arabidopsis* revealed genes which experience translational upregulation in response to light. These genes are primarily associated with processes related to the organization and function of chloroplasts (Liu et al., 2013).

McKim and Durnford (2006) investigated the role of translational control in response to high-light stress in the green alga *Chlamydomonas reinhardtii.* Their results show a down-regulation of the light harvesting capacity at Photosystem II after short-term exposure to excess light via translational repression of two light-harvesting complex genes, Lhcbm and Lhcb4.

The D1 reaction center protein of PSII, which is encoded by the chloroplast psbA gene, is a crucial target of photodamage. When damaged, D1 undergoes degradation and is subsequently replaced by newly synthesized D1 through an elaborate repair cycle (Järvi et al., 2015; Chotewutmontri and Barkan, 2018). Due to its high rate of turnover, the psbA mRNA is the most actively translated mRNA in response to high light stress (Sun and Zerges, 2015). A study investigated chloroplast translation in maize after a shift from dark to light and discovered that this transition is accompanied by a global increase in translation elongation rate in chloroplasts (Chotewutmontri and Barkan, 2018). Interestingly, the psbA mRNA exhibits large light-induced changes in ribosome occupancy in response to light-dark shifts, while ribosome recruitment on all other chloroplast mRNAs remained largely unchanged. These findings highlight the unique translational response of psbA to produce a rapid increase in protein levels under light stress (Chotewutmontri and Barkan, 2018).

1.4.4 Methods to study translation and translational regulation

Given the importance of translation and its regulation, various techniques have emerged to study translated mRNAs. Polysome profiling is based on sucrose-gradient separation of actively translated mRNAs bound by multiple ribosomes (polysomes) from untranslated mRNA. The distribution pattern of specific mRNAs in the gradient can then be determined by northern blotting, qRT-PCR or RNA sequencing (King and Gerber, 2016). Since the discovery that multiple ribosomes can be held together by a single mRNA (Warner et al., 1963), this method has been widely used to target specific mRNAs (del Prete et al., 2007; Chassé et al., 2016) as well as for genome-wide analyses of translation (Chen et al., 2011; Kronja et al., 2014). Combined with immunoblotting and/or proteomics, polysome profiling further allows for monitoring of proteins associated with ribosome and/or initiation complexes (Chassé et al., 2017). However, broad application of polysome profiling has been hampered by the technical

difficulty of fractionating mRNAs which harbour many ribosomes per transcript, as well as by the need to collect and analyse multiple polysome fractions per sample (Ingolia, 2014).

Translating ribosome affinity purification (TRAP) is another commonly used method to study translation. Organisms are genetically engineered to express affinity-tagged ribosomal proteins (RPs) *in vivo*, enabling the selection of tagged ribosome-mRNA complexes via affinity purification and further quantitative analysis of the RNA by deep sequencing, northern blotting or qPCR (King and Gerber, 2016). An advantage of TRAP is that the technique is able to study cell-type-specific gene expression in many organisms (Heiman et al., 2014). In *Drosophila melanogaster* this can be accomplished by controlling the expression of the tagged RPs via the GAL4-UAS system (Bertin et al., 2015) or in mice via the cell-type specific Cre-lox promoters (Salussolia et al., 2022). However, a limitation of this method is the requirement to generate transgenic cell lines (Heiman et al., 2014). Polysome profiling and TRAP do not provide any information about the distribution of ribosomes on the transcripts, thus those methods cannot distinguish between translating and non-translating ribosomes (Kage et al., 2020). Ribosome profiling (see below) is a technique that circumvents this problem by precisely determining the position of translating ribosomes on the mRNA (Ingolia et al., 2009).

To overcome limitations related to each approach, several papers have reported a combination of these methods. Incorporation of TRAP into ribosome profiling has been proven successful in revealing translational dynamics of *Arabidopsis thaliana* under hypoxic stress (Juntawong et al., 2014). Further, a method termed 'Poly-Ribo-Seq' was developed which subjects polysomal fractions to ribosome profiling to facilitate the detection of small ORFs (sORFs) in *D. melanogaster* (Aspden et al., 2014).

1.4.5 The ribosome profiling technique

Precise monitoring of translation has been challenging until the development of ribosome profiling or Ribo-Seq in 2009 (Ingolia et al., 2009). This technique is based on the analysis of ~30 nucleotide (nt) long mRNA fragments which are enclosed by translating ribosomes and protected from nuclease digestion (Steitz, 1969). Deep sequencing of these generated footprints or ribosome-protected fragments (RPFs) thus provides a genome-wide snapshot of translation (Ingolia et al., 2009). The main steps of the ribosome profiling method are shown in Figure 1.9. It starts with rapid harvesting of the samples to avoid ribosome run-off. This is usually achieved

by treating cells with elongation inhibitors such as cycloheximide (CHX) or by flash-freezing in liquid nitrogen (Ingolia et al., 2012; Eastman et al., 2018b). Nuclease digestion of the lysed cells is a critical step in the ribosome profiling protocol. Optimal conditions are needed to ensure both complete digestion of unprotected mRNA and preservation of the ribosomal integrity. In eukaryotic organism, ribonuclease (RNase) I, A, T1 and micrococcal S7 are commonly used but their cutting efficiencies are species-dependent. Studies in bacteria however are more restricted to the use of S7 (Gerashchenko and Gladyshev, 2017). The next step involves the isolation of mRNA-ribosome-complexes which is usually done by ultracentrifugation in sucrose cushions or gradients. RPFs are size selected from a denaturing poly-acrylamide gel electrophoresis (PAGE) using appropriate markers to excise the ~30 nt fragments as precisely as possible to minimize contamination with other RNA fragments. Even with stringent size selection conditions, contamination can represent up to 90% of ribosome profiling samples. Thus, most protocols implement a rRNA removal step using commercially available depletion kits or subtractive hybridization using biotinylated oligos which are antisense to the most common contaminants. The later requires preliminary sequencing results for the design of the depletion oligos but is typically the best or only option when working with non-model organism (Zinshteyn et al., 2020). The RPFs are then converted into a cDNA library and subjected to deep sequencing. Bioinformatic analysis typically involves pre-processing steps for quality control, mapping of the reads against a data base of rRNA to remove contamination, alignment of unmapped reads against a reference genome or transcriptome and statistical analysis (Eastman et al., 2018a). The main advantage of this method is that the RPFs reveal the exact position and number of ribosomes on transcripts. Combined with RNA sequencing generated from the same sample, it reveals the translation efficiency (the ratio of ribosome footprint density to mRNA density) of any gene and thus allows to estimate relative translation levels (Ingolia et al., 2009).

Ribosome profiling data exhibits a strong triplet periodicity which is absent from RNA-Seq data. This distinct feature is due to the fact that ribosomes move along the mRNA 3 nt at a time resulting in a strong reading frame preference throughout the whole coding sequence (CDS) (Eastman et al., 2018a). The majority of ribosome profiling reads are expected to map to CDSs. However, a low number of reads can also map to 5'-untranslated regions (UTR) representing the translation of upstream ORFs (uORFs) or scanning ribosomes (Rodriguez et al., 2019) and to 3' UTR, which reveals translation of downstream ORFs (dORFs) (Bazzini et al., 2014).

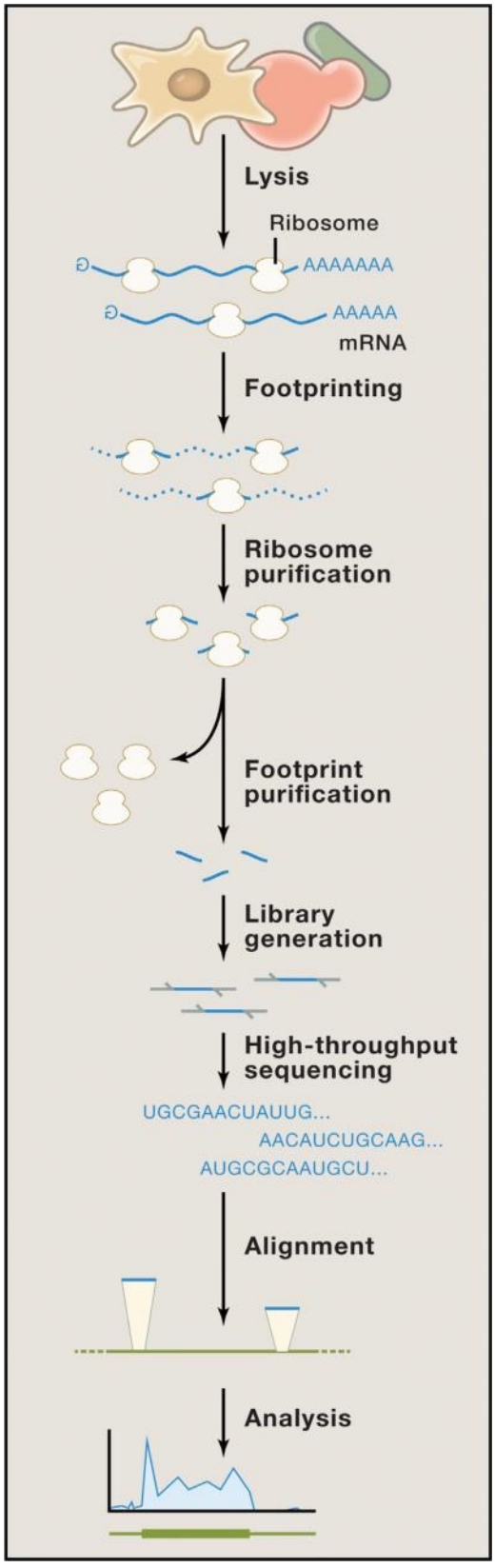


Figure 1.9: Overview of the ribosome profiling method. Cells are lysed, and RNA is subjected to nuclease digestion (footprinting). Ribosomes are recovered and purified and converted to cDNA. High-throughput sequencing is followed by bioinformatic analysis (Ingolia, 2016).

This positional information revealed by ribosome profiling data relies on the ability to determine the exact location of the ribosomal P-site, which is the site holding the tRNA associated with the growing polypeptide chain (Lauria et al., 2018). Several computational tools have been developed to calculate the so-called P-site offset (PO), which is the distance from both the 5' and 3' extremities of a read and the first nucleotide of the P-site within the RPF (Figure 1.10) (Dunn and Weissman, 2016; Popa et al., 2016; Lauria et al., 2018). Given the source of information provided by ribosome profiling, such as precise positional information, RPF abundance or length distribution, the applications of this method are broad, and some important ones are summarized in the next section.

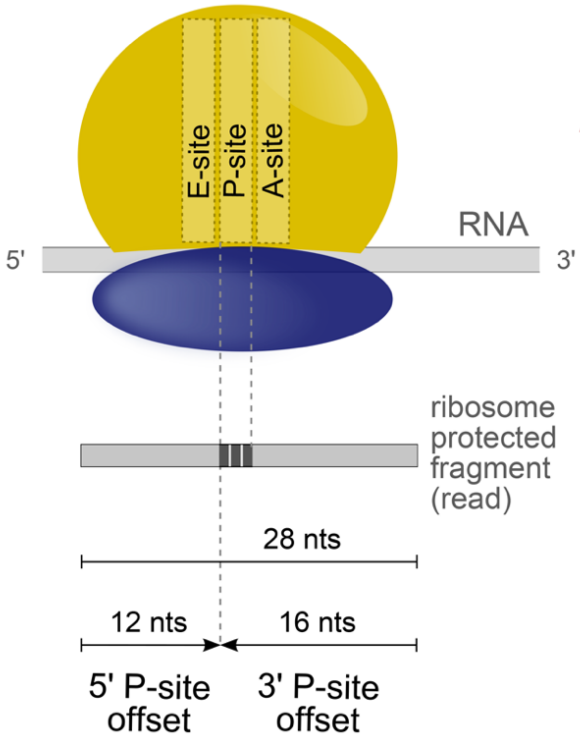


Figure 1.10: Illustration of the ribosomal P-site within the ribosome-protected fragment (RPF) and determination of its offset. Offsets can be defined both for the 5' end and the 3' end of the read (Lauria et al., 2018).

1.4.6 Applications of ribosome profiling

Ribosome profiling has first been developed in *Saccharomyces cerevisiae* (Ingolia et al., 2009) and has since been applied to study translation in a wide range of organisms such as bacteria (Li et al., 2012), fishes (Bazzini et al., 2012), mammals (Guo et al., 2010; Ingolia et al., 2011) and plants (Juntawong et al., 2014; Hsu et al., 2016). The most widespread application of ribosome profiling data is the analysis of differential gene expression at the level of translation. Many studies revealed that gene expression levels estimated from ribosome profiling data correlate better with protein abundance than transcriptome-derived estimations, highlighting the importance of translational regulation (Eastman et al., 2018a). Ribosome profiling has been applied in parallel with RNA-Seq for genome-wide analysis of translational regulation in a variety of organisms, revealing complex regulatory strategies in response to cellular stress (Ingolia et al., 2009; Ingolia et al., 2011; Gerashchenko et al., 2012; Zhang et al., 2017).

Ribosome profiling experiments have contributed to reveal the complexity of the translatome not just by re-annotating known ORFs but also by identifying previously unannotated ORFs leading to the discovery of novel peptides (Calviello et al., 2016; Hsu et al., 2016; Wu et al., 2019). Extensive translation of uORFs has been revealed by many eukaryotic ribosome profiling studies. Their translation seems to be upregulated upon exposure to environmental stress, suggesting important regulatory roles of these uORFs (Ingolia et al., 2009; Brar et al., 2012; Gerashchenko et al., 2012). Studies have further reported the translation of sORFs in long-non-coding RNAs (lncRNAs), many of which are likely to express functional proteins (Aspden et al., 2014; Bazzini et al., 2014; Ji et al., 2015).

Translation initiation sites (TIS) have been identified by using drugs such as harringtonine and lactimidomycin which lead to stalling of initiating ribosomes on the mRNA. This treatment thus results in an enrichment of RPFs at the start codon of an ORF. This approach has facilitated the discovery of novel non-AUG initiation sites in many uORFs (Ingolia et al., 2011; Fritsch et al., 2012; Lee et al., 2012). Such non-AUG start codons have been shown to play important roles in stress response (Starck et al., 2016) and drive tumour initiation (Sendoel et al., 2017). The combined use of harringtonine and cycloheximide to arrest both initiating and elongating ribosomes allows for monitoring of translational kinetics like codon-specific elongation rates (Ingolia et al., 2012; Li et al., 2012; Pop et al., 2014). Ribosome profiling data showed that codons matching less abundant tRNAs are more slowly translated, however, this relationship

between ribosome density and tRNA abundance was only present in data without CHX pre-treatment (Lareau et al., 2014; Weinberg et al., 2016).

Analysis of ribosome profiling data revealed the widespread presence of ribosomal pausing sites in both prokaryotes and eukaryotes. These pause sites are suggested to have regulatory effects on translation and appear as peaks of RPFs within ORFs (Ingolia et al., 2011; Shalgi et al., 2013; Brar and Weissman, 2015; Karlsen et al., 2018). In mammalian cells, heat shock has resulted in global ribosome pausing early on in the transcript, which is thought to be caused by ribosome-associated chaperons (Shalgi et al., 2013). Ribosome profiling has further led to the discovery of novel cases of ribosomal frameshifting in yeast under oxidative stress (Gerashchenko et al., 2012).

1.4.7 Limitations of ribosome profiling

Although ribosome profiling is a powerful technique to study genome-wide translation, it also faces some major difficulties and limitations. Digestion with nuclease results in contaminating RNA fragments resembling the size of 80S footprints, which co-migrate with the ribosome in a sucrose gradient or cushion (Brar and Weissman, 2015). These contaminants are mostly derived from non-coding RNAs and significantly reduce the amount of informative footprint sequencing data (Ingolia et al., 2009). rRNAs are typically removed during library preparation by commercial depletion kits or by subtractive hybridization using custom biotinylated depletion oligonucleotides as well as bioinformatically after sequencing (McGlincy and Ingolia, 2017). The small size of ribosome footprints makes correct alignment to the reference genome or transcriptome challenging, especially for reads derived from repetitive regions or alternative splicing (Brar and Weissman, 2015). However, several computational tools have been developed to resolve the issue of ambiguous read mapping (Calviello et al., 2016; Wang et al., 2016). The large amount of input material needed still poses the major limitation of ribosome profiling (Brar and Weissman, 2015). The requirement for large quantities of sample is due to the number of processing steps involved in the protocol to isolate ribosomes (McGlincy and Ingolia, 2017). Continuous improvements to this technology have already drastically reduced the RNA input requirements allowing for selective ribosome profiling of specific ribosomal subunits as well as small tissue samples such as clinical biopsies (Liu et al., 2019; Meindl et al., 2022). Further, ribosome profiling is a labour-intensive method, and requires specialised equipment (e.g. ultracentrifuge, gradient fractionation system) which is not

available in every laboratory (King and Gerber, 2016). Once ribosome profiling data has been generated, processing and analysis of the sequencing data can be challenging. In recent years, many ribosome profiling data specific tools have been developed ranging from analysis of nucleotide periodicity and translation efficiency to identification of ribosomal P-site (Michel et al., 2016; Lauria et al., 2018; Chothani et al., 2019; Francois et al., 2021). Ribosome profiling data however still lacks standardization of computational protocols which hampers the analysis of this specialized data and leads to increased data variability coming from the shear amount of available software packages (Berg et al., 2020).

1.5 Genome editing using the CRISPR/Cas9 system

Continuing advancements of genetic engineering tools have revolutionized the field of biology. Most prominently, the advent of the clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) technology enabled precise, facile, rapid, and cost-efficient editing of any target DNA in living cells (Doudna and Charpentier, 2014; Sander and Joung, 2014). CRISPR-Cas is an adaptive viral defence system used by bacteria and archaea (Jinek et al., 2012). The technology is comprised of a two-component system, including a Cas9 endonuclease and a single guide RNA (sgRNA) (Doudna and Charpentier, 2014). The chimeric sgRNA is created by fusing two noncoding RNAs, a CRISPR RNA (crRNA) and a trans activation RNA (tracr). The CRISPR-associated protein Cas9 is an endonuclease that, when paired with a sgRNA, induces a double-strand break (DSB) in a target sequence complementary to a 20 nt (guide) sequence in the crRNA (Jinek et al., 2012; Doudna and Charpentier, 2014; Sander and Joung, 2014). Earlier approaches for genome editing include the site-directed zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) (Doudna and Charpentier, 2014). However, CRISPR-Cas has the advantage that it is a cheap, efficient, and easily adaptable tool (Hopes et al., 2016). By simply changing the guide sequence of the sgRNA, it is possible to target any position in the genome (Doudna and Charpentier, 2014; Sander and Joung, 2014) provided that a protospacer adjacent motif (PAM) immediately follows the targeted regions, allowing Cas9 to bind. The Cas9 enzyme cleaves the DNA 3 nt upstream of the PAM sequence that matches a 20-nucleotide sequence of the sgRNA (Doudna and Charpentier, 2014) which triggers DNA repair by one of two major pathways: error-prone non-homologous end joining (NHEJ), or homologous recombination (HR) (Figure 1.11) (Sander and Joung, 2014). NHEJ simply re-joins the broken DNA ends, which can lead to insertions/deletions (indels) to be introduced, possibly resulting in frame shift mutations and premature stop codons (Hiom, 2000; Sander and Joung, 2014). HR uses an undamaged copy of

the DNA as a template to repair the DSB which makes this repair mechanism less susceptible to errors (Hiom, 2000). In the presence of a template DNA with flanking regions homologous to the DSB, any sequence of interest can be knocked-in by HR. However, the efficiency of NHEJ-mediated gene knock-out is much higher than knock-in by HR repair mechanisms (Zhang et al., 2021). Genome editing using HR has been employed in many marine microalgae including *T. pseudonana* (Belshaw et al., 2022), *P. tricornutum* (Daboussi et al., 2014), *C. reinhardtti* (Greiner et al., 2017) and *Nannochloropsis sp*. (Kilian et al., 2011).

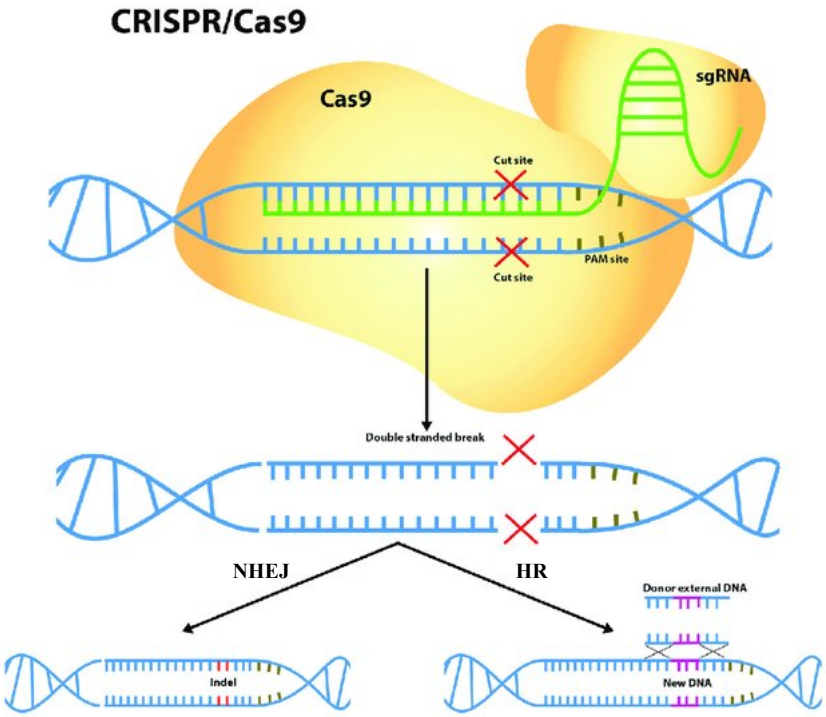


Figure 1.11: Schematic of the CRISPR/Cas9 genome editing system. A single guide RNA (sgRNA) targets a genomic region followed by a protospacer adjacent motif (PAM). This allows the Cas9 endonuclease to cleave the DNA and introduce a double-stranded break (DSB) which gets repaired by error-prone non-homologous end joining (NHEJ) that can lead to indels, or by homologous recombination (HR) if a donor DNA is present. Adapted from Cribbs and Perera (2017).

The CRISPR/Cas system has been delivered to diatom cells by using three different transformation methods: biolistic particle bombardment (Hopes et al., 2016; Nymark et al., 2016), electroporation (Niu et al., 2012; Zhang and Hu, 2014) and more recently bacterial conjugation (Karas et al., 2015; Diner et al., 2016; Sharma et al., 2018; Moosburner et al., 2020).

A major concern about CRISPR/Cas is unspecific binding of the guide RNA to identical or highly homologous regions, leading to unintended off-target activity (Kim et al., 2021). Mismatches between the target DNA and the guide RNA are tolerated by Cas9 but seem to be determined by their number and position (Hsu et al., 2013). A systematic evaluation of Cas9 specificity revealed that single-base mutations in the target sequence which are close to the PAM region are less tolerated than mutations further upstream (Hsu et al., 2013). Predominantly, off-target sites have less than four or five mismatches to the guide RNA (Haeussler et al., 2016). Accurate prediction and validation of Cas9 off-target activity is therefore crucial for any genome editing project. Alignment tools based on sequence homology allow for *in silico* prediction of potential off-target sites in any genome with up to several mismatches (Bae et al., 2014). Validation of predicted off-target sites can be done *in vitro* based on PCR amplification or sequencing (Kim et al., 2015; Tsai et al., 2015; Höijer et al., 2020) and *in vivo* based on CHip-seq or tagging (Wienert et al., 2019; Liang et al., 2022). Delivery of the CRISPR/Cas complex via plasmids poses a much higher risk of off-target activity due to long-term nuclease expression in the target cell (Kim and Kim, 2014). To reduce off-target effects, many researchers now directly deliver Cas9 and sgRNAs via ribonucleoprotein (RNP) complexes (Naduthodi et al., 2019; Kang et al., 2020). In 2018, *P. tricornutum* was the first diatom to be transformed without plasmid-based delivery of exogenous DNA (Serif et al., 2018). Their RNP genome editing approach via biolistics resulted in multiple gene knock-out strains, avoiding random DNA integration and prolonged Cas expression.

1.6 Aims of this thesis

The overarching goal of this thesis was to provide first insights into translational regulation of the model diatom *Thalasssiosira pseudonana*. Ribosome profiling has unveiled the complexity and regulation of translation in many prokaryotic and eukaryotic species. To date, the regulatory mechanisms of protein synthesis at the translational level remain largely unexplored in diatoms due to the lack of an adapted ribosome profiling protocol.

Thus, the first aim of this study was to develop a detailed ribosome profiling protocol for the diatom *T. pseudonana* to study genome-wide translation. The intent was to demonstrate the protocol's proficiency in generating high-quality ribosome profiling data, while also being readily adaptable for usage in other diatom species. This is the first application of this technique for any diatom and expands the available molecular toolbox for this globally important group of phytoplankton.

The second aim of the thesis was to study translational regulation under changing environmental conditions by applying ribosome profiling and RNA-seq in parallel. For this purpose, replicate cultures of *T. pseudonana* were exposed to high light for 4 hours and 24 hours. The generated dataset provides first insights into translational regulation of any diatom species in response to light stress. The questions that were explored revolved around identifying genes expressed at different levels of gene expression, determining whether specific genes are exclusively regulated translationally, and examining whether short-term and prolonged high-light stress result in distinct responses. This information is important for advancing our knowledge of how photosynthetic organisms adapt their cellular responses and energy allocation in the face of changing environmental conditions.

The third aim of this thesis tries to elucidate the role of codon usage in diatoms. For this, we tried to modify the codon usage for two genes in *T. pseudonana* via CRISPR/Cas mediated homologous recombination. Initially, we assessed the feasibility of this approach for both genes and aimed to find a method to reliably isolate codon modified cells. Phenotyping of cells with verified bi-allelic modifications provide preliminary insights into the role of codon usage in diatoms. These cell lines can be used for future ribosome profiling studies to investigate the impact of codon usage on translation.

Chapter 2: Development of a ribosome profiling protocol for *Thalassiosira pseudonana*

2.1 Introduction

In recent years, transcriptome data has been widely used to estimate gene expression levels, however, due to a series of post-transcriptional regulation factors, the correlation of mRNA levels with proteomic data is far from perfect (Eastman et al., 2018b). Ribosome profiling fills this gap by quantifying translation at a genome-wide level. This technique is based on the analysis of mRNA fragments (28-32 nt) which are enclosed by elongating ribosomes and shielded from nuclease digestion (Steitz, 1969). Ribosome profiling enables direct estimation of protein synthesis at the translational level by deep sequencing of these ribosome protected fragments (RPF) (Ingolia et al., 2009).

Ribosome profiling has been applied to study the translational regulation of gene expression in a number of photosynthetic organisms. In *Arabidopsis thaliana*, exposure to hypoxic conditions resulted in an ~100-fold decrease in translation efficiency of some mRNAs. The reduced presence of RPFs at AUG codons strongly suggests that hypoxia acts as a translation initiation inhibitor (Juntawong et al., 2014). A recent ribosome profiling study revealed the substantial role of uORFs and microRNAs in regulating translation in tomato roots. Their data suggests that e.g. protein phosphorylation/dephosporylation and signal transduction pathways are translationally regulated through uORFs (Wu et al., 2019). At the beginning of this project, ribosome profiling has only been performed on one algal species, the green algae *Chlamydomonas reinhardtii.* Chung et al. (2015) used ribosome profiling data to identify and correct misannotations in the *C. rheinhardtii* reference genome. In another study, ribosome profiling was applied in combination with other methods to quantify the effects of miRNA on protein expression in this freshwater microalgae (Chung et al., 2017). Trösch et al. (2018) used a targeted chloroplast ribosome profiling approach to study chloroplast gene expression levels in *C. reinhardtii* compared to land plants. The first ribosome profiling data for cyanobacteria reveals novel regulation strategies for *Synechocystis* under carbon starvation. Translation was decreased by 80% and ribosome pausing was increased at stop and start codons as well as in 5' and 3' UTRs, suggesting a sequestration mechanism to deactivate ribosomes upon carbon depletion (Karlsen et al., 2018).

To the best of our knowledge, this chapter provides the first application of genome-wide ribosome profiling for any marine algae. When starting ribosome profiling with a new class of

organisms, several experimental setups need to be adapted from published protocols to generate accurate footprint data. These include optimizing harvesting strategies and buffer conditions and the amount of nuclease for mRNA digestion. Here, we describe the detailed development of a ribosome profiling protocol to study translation in the marine diatom *T. pseudonana*. Extensive research has been carried out on this species as it was the first diatom with a sequenced genome (Armbrust et al., 2004b), serves as the model organism for studying silicification (Poulsen and Kröger, 2004) and had its genome successfully edited via CRISPR/Cas (Hopes et al., 2016). Ribosome profiling further expands the molecular toolbox for diatoms and will hopefully improve our understanding of translational regulation of this globally important phytoplankton.

2.2 Material and Methods

2.2.1 Monosome preparation

2.2.1.1 Cell culture conditions

Thalassiosira pseudonana (strain CCMP1335) was grown to mid-exponential phase (500.000-700.000 cells/ml) in Aquil media (pH 8) (Price et al., 1989) under constant light (75 µmol photons m^{-2} s^{-1}) at 20°C. Cells were either harvested immediately or exposed to pre-heated media (32°C) or high light (500 µmol photons m^{-2} s^{-1}) for 4 and 24 hours. All experiments were carried out in biological triplicates.

2.2.1.2 Cell harvest and lysis

Cells were harvested by vacuum filtration onto 1.2 µm Isopore membrane filters (Millipore) and immediately flash-frozen in liquid nitrogen. Per sample, 2 x 200 ml of culture was required. The cell paste was thawed on ice and washed off the filter with 400 µl pre-chilled polysome resuspension buffer (PRB) (50 mM Tris Hcl, pH 8; 100 mM NH_4Cl; 200 mM sucrose; 10.5 mM Mg acetate; 0.5 mM EDTA; 1 mM DTT; 0.01 % Cycloheximide; 0.1 % Triton X-100; 25U/ml Turbo DNAse). Sterile glass beads (450-600 µm) were added to vortex the sample in a Mini-beadbeater (BioSpec) for 2 min (max speed) with a break on ice after 1 min. The lysate was clarified by centrifugation for 10 min at 14000 rpm at 4°C. The RNA concentration of the extract was estimated by measuring UV absorption at 260nm with a NanoDrop

spectrophotometer. Lysate containing an estimated amount of 60 µg of total RNA (1.5 AU_{260nm}) was treated with 7U RNase I (100 U/µl) per µg and incubated for 45 min at room temperature with gentle shaking. Digestion conditions were previously assessed via sucrose gradient density ultracentrifugation and gradient fractionation (see section 2.2.2.3 Density gradient centrifugation). The digestion was placed on ice, followed by addition of 2 µl SUPERase-In RNase inhibitor (20 U/µl) to stop the nuclease digestion.

2.2.2 Isolation of monosomes

2.2.2.1 Sucrose cushion ultracentrifugation

The sample was transferred to a 11 x 34 mm thickwall polycarbonate ultracentrifuge tube (Beckman Coulter) and underlaid with 750 µl of 1M sucrose cushion prepared as outlined in McGlincy and Ingolia (2017) but with PRB. The sample was centrifuged for 2 h using a TLA120.2 rotor (Beckman Coulter) at 75.000 rpm at 4°C.

2.2.2.2 RNA extraction and size selection of ribosome protected fragments

After removal of the supernatant, the ribosomal pellet was resuspended in 350 µl of TRIzol and replicates were pooled at this point. The RNA was purified using the Direct-zol RNA MiniPrep kit (Zymo Research) following the manufacturer's instruction for purification of total RNA, eluted in 50 µl RNAse-free water and precipitated overnight. Around 10-15 µg of RNA was loaded onto a 15% denaturing polyacrylamide gel and RPFs were size selected (26 to 32 nt). RNA was eluted from the gel slices via overnight extraction, precipitated with ethanol, and quantified on a Qubit using the microRNA Assay kit.

2.2.2.3 Density gradient ultracentrifugation

The following steps were done at the Medenbach Lab (University of Regensburg, Germany) as described in Meindl et al., (2022). Briefly, about 200 µg of RNase I digested total RNA was loaded on top of a 10-50% sucrose gradient and centrifuged for 3 h using a SW41Ti rotor (Beckman Coulter) at 35000 rpm at 4˚C. A siFractor (siTOOLs Biotech) connected to an ÄKTApurifier fast protein liquid chromatography (FPLC) system (Cytiva) was used for semi-automated fractionation of the sample. The gradient was displaced with a dense chase solution stained with Ponceau S and 1 ml fractions were collected. A variety of digestion conditions

were tested, ranging from 0.2U to 7U of RNase I per µg of total RNA, to find the optimal one for *T. pseudonana*.

2.2.3 Library construction

Preparation of cDNA libraries was done as described in Meindl et al., (2022) with modifications to the rRNA depletion step.

2.2.3.1 Design of biotinylated depletion oligos for subtractive hybridization

The design of custom rRNA depletion probes was based on previously generated sequencing data from a small-scale experiment using an undepleted *T. pseudonana* wild-type sample. This data was generated by Annemarie Eckes and Amanda Hopes, and preparation of the sample differed from our final protocol as it includes CHX pre-treatment, digestion of 300 µl lysate with 7.5 µl RNase I, and library preparation using the NEBnext kit (New England BioLabs). XPRESSpipe (Berg et al., 2020) and its sub-module rrnaProbe was used to create a list of the most dominant sequences, according to which eight 5' biotinylated oligos (Table 2.1) were designed to subtract the most abundant rRNA fragments from the samples. Sequences were verified using BLAST (Altschul et al., 1990) prior to synthesis (Eurofins Genomics).

2.2.3.2 Implementation of rRNA removal step into protocol

A custom subtractive hybridization step using the 5' biotinylated depletion oligos described above was implemented into the protocol following Zinshteyn et al. (2020) with some modifications. Briefly, a depletion oligo mixture was made by diluting each oligo to 2.5 µM in 4x SSC buffer. 10 µl of the oligo mixture was added to 35µl of the 3'- adapter ligation product and incubated for 2 min at 80°C. 80 µl of MyOne Streptavidin C1 dynabeads were washed three times and finally resuspended in 45 µl of 2x Binding/Washing buffer. Beads and sample were mixed and incubated for 15 min at 25°C with shaking at 500 rpm before placing it on a magnetic rack for > 1 min and transferring the supernatant into a new tube.

2.2.4 Analysis of fractions from density gradient ultracentrifugation

Ribosome protected fragments from two fractionations, generated via density gradient centrifugation, were isolated and library preparation was done following Meindl et al., (2022) but omitting the rRNA removal step. Libraries were sequenced on an Illumina MiSeq (V3 kit, 150 cycles, 90 cycles single end). Sequencing data was processed and visualized using the RiboDoc pipeline (Francois et al., 2021) and the implemented riboWaltz package (Lauria et al.,

2018). The *T. pseudonana* genome (Thaps3_chromosomes_assembly_chromosomes_repeatmasked.fasta) and the annotations (Thaps3.filtered_proteins.FilteredModels2.gff3) were downloaded from the Joint Genome Institute's (JGI) website. A file containing a comprehensive list of *T. pseudonana* rRNA and other non-coding sequences was compiled using the RNAcentral database. The RSeQC package (version 2.6.4) was used to calculate the distribution of mapped reads over genome features.

2.2.5 Ribosome profiling of wildtype samples to verify the protocol

Ribosome profiling of *T. pseudonana* wild-type samples was done following our protocol (see detailed protocol in Appendix B). Three sample were processed omitting the rRNA removal step while three samples were subjected to subtractive hybridization. Sequencing of the cDNA libraries was done on an Illumina MiSeq (V3 kit, 150 cycles, 90 cycles single end).

The sequencing reads were de-multiplexed, and adapters were trimmed using Cutadapt (version 2.8, adapter=AGATCGGAAGAGCACACGTCT, overlap=10, minimum-length=10, discard-untrimmed). UMIs were extracted and appended to the read name using UMI tools (version 1.0.1, extract-method=regex). Reads were filtered by lengths and only reads between 16 and 40 nt were mapped with bowtie2 (version 2.3.5.1) against *T. pseudonana* rRNA and noncoding sequences compiled using the RNAcentral database. Unmapped reads were then mapped against the JGI *T. pseudonana* reference genome (Thaps3_chromosomes_assembly_chromosomes_repeatmasked.fasta) with genome annotations (Thaps3.filtered_proteins.FilteredModels2.gff3) using STAR (version, --quantMode TranscriptomeSAM GeneCounts, --outSAMattributes All, outFilterMismatchNmax 2). PCR duplicates were removed using UMI tools (extract-umi-method, read_id --method unique). Analyses and interpretation of ribosome profiling data was performed in R (version 3.5.1) (R Core Team, 2021) using the riboWaltz package (version 1.2.0) (Lauria et al, 2018). Statistical analysis was performed using the RiboDoc pipeline (Francois et al., 2021) and in R (version 3.5.1) (R Core Team, 2021).

2.3 Results and Discussions

2.2.1 Development of a ribosome profiling protocol to study translation in *T. pseudonana*

We describe a ribosome profiling protocol suitable for the analysis of *T. pseudonana* cells. The first part of the protocol, from cell harvest to size selection of the ribosome protected footprints, is largely based on previously published protocols by Ingolia et al., (2012) and McGlincy and Ingolia (2017). *T. pseudonana* cultures were rapidly harvested and flash frozen to retain the translational state of the cell. Cells were lysed using a mechanic bead beater and glass beads. The RNA lysate was digested with RNase I to remove mRNA unprotected by ribosomes. 80S ribosomes were isolated via sucrose cushion ultracentrifugation and size-selected on a polyacrylamide gel.

The recently published protocol by Meindl et al., (2022) was used for conversion of ribosome protected fragments into cDNA libraries. A subtractive hybridization step was added to deplete abundant rRNA fragments.

An overview of the adjustments and optimizations will be highlighted in the following sections.

2.2.1.1 Translation inhibitors

Translation inhibitors such as cycloheximide were frequently used prior to cell harvest in early ribosome profiling studies (e.g. Ingolia et al., 2009). However, more recent data suggests that they can introduce significant biases by enhancing ribosomal occupancy profiles at the initiation site in *S. cerevisiae* (Gerashchenko and Gladyshev, 2014). Similar artefacts were observed in *C. reinhardtii* when treated with translation inhibitors (Chung et al., 2017). Thus, cycloheximide pre-treatment was omitted in this protocol and instead a flash-freezing method using liquid nitrogen was selected. In accordance with other protocols, cycloheximide was still present in the buffer to prevent any post-lysis ribosomal movement during thawing of the cell lysate.

2.2.1.2 Starting volumes, harvesting strategy and cell opening

T. pseudonana is characterized by a heavily silicified cell wall, which hampers the extraction of the large amounts of total RNA needed for ribosome profiling. In the end, two replicates, each 200 ml of cells in mid-exponential phase (500.000-700.000 cell/ml), were sufficient for providing enough starting material. Our harvesting strategy requires replicates to keep the time

between filtration and flash-freezing of the sample under 60 seconds, which preserves the state of the cell as accurately as possible. Simply ramping up the filtration speed would have posed another stress factor to the cells and possibly introduced a bias in the translational landscape. For lysis, we used a polysome resuspension buffer which has previously been tested on marine dinoflagellates (Schröder-Lorenz and Rensing, 1987) and thus might closely reflect the cell properties of diatoms. Our selected starting culture volume, together with a bead beating cell opening approach, yielded about 60 µg of total RNA per replicate, which is twice the recommended total RNA amount of 30 µg used by McGlincy and Ingolia (2017) in *S. cerevisiae.* They, however, measured the concentration on a Qubit, while we used a NanoDrop which only provides a rough estimate due to the large number of metabolites and light harvesting complexes present in the crude diatom RNA lysate which absorb UV light of 260 nm. Cell opening via cryogenic mill would further improve isolation of total RNA from diatom cells and circumvents the issue of heat generation during vigorous beat beating which can lead to thawing of the cell extracts. Even though we likely overestimate the amount of RNA by UV absorption measurement, this starting amount generates the 10-15 µg of RNA needed for loading onto the PAGE gel.

2.2.1.3 Optimization of RNase digestion conditions

Nuclease treatment is a critical step during ribosome footprint preparation and impacts the quality of the sequencing data. Ideally, RNase I should only digest mRNA which is not protected by translating ribosomes and keep the ribosomes intact (Gerashchenko and Gladyshev, 2017). Digestion conditions need to be thoroughly tested for every species since ribosomes from different species vary widely in terms of resilience to nuclease digestion. Yeast ribosomes can tolerate vigorous RNase I treatment without compromising much of their

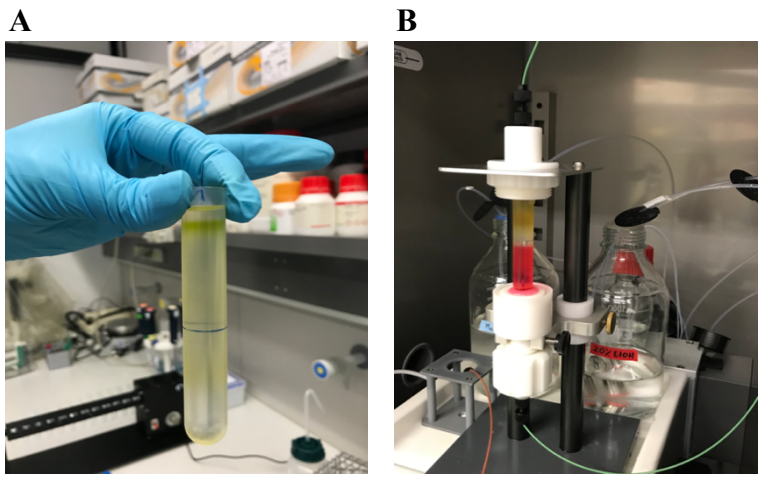


*Figure 2.1: Gradient fractionation. **A**. Gradient tube after density gradient ultracentrifugation. Sedimentation of green pigments from the photosystems can be seen towards the top of the gradient. **B**. Pierced gradient tube in fractionator unit connected to an ÄKTA FPLC system. Red chase solution is displacing the density gradient during the fractionation process.*

integrity while *Drosophila* ribosomes are easily degradable (Gerashchenko and Gladyshev, 2017). Nuclease tolerance of *T. pseudonana* ribosomes could not be deduced from model organisms but needed to be experimentally tested before continuing with the ribosome profiling protocol. Ribosome integrity cannot be monitored when using a sucrose cushion, thus a sucrose gradient ultracentrifugation step needed to be performed under a variety of nuclease digestion conditions (Figure 2.1). The efficacy of the different RNase I conditions was tested under 20˚C and 4˚C (Figure 2.2B-F). Undigested *T. pseudonana* cells showed several peaks corresponding to both nucleus and chloroplast-encoded ribosomes (Figure 2.2A). Diatoms are photosynthetic organisms, thus it was expected to see both cytoplasmic 80S ribosomes and chloroplast derived 70S ribosomes (Manuell et al., 2007). Smaller peaks towards the end of the gradient represent disomes and polysomes. A conversion from polysomal to monosomal peaks is clear indication for efficient nuclease treatment (Gerashchenko and Gladyshev, 2017). Light digestion conditions of 0.2U, 0.5U and 2U/µg did not result in a shift of the gradient profile towards an 80S peak. However, a strong 7U/µg digest performed well and led to the decrease of polysome peaks and what were possibly the chloroplast 70S ribosome and 60S and 40S subunits (Figure 2.2F). We did not see a clear monosomal peak. Instead the gradient profile showed two peaks which possibly contain 80S particles and a polysomal peak which was not fully resolved to monosomes yet. An increase in RNase I would have probably further decreased the height of any peaks containing 80S. Thus, a 7U digest at 20°C for 45 min was implemented in our protocol, as it seems to represent optimal conditions for efficiently trimming the mRNA while keeping the ribosomes intact and yielding sufficient 80S complexes. Ribosomes of microalgae seem to withstand strong RNase I treatment, as a similarly high tolerance is seen in the green algae *C. reinhardtii* (Chung et al., 2015). Preliminary gradient profile data (data not shown here) from the polar diatom *Fragilariopsis cylindrus* showed that our protocol is easily adaptable for other diatom species.

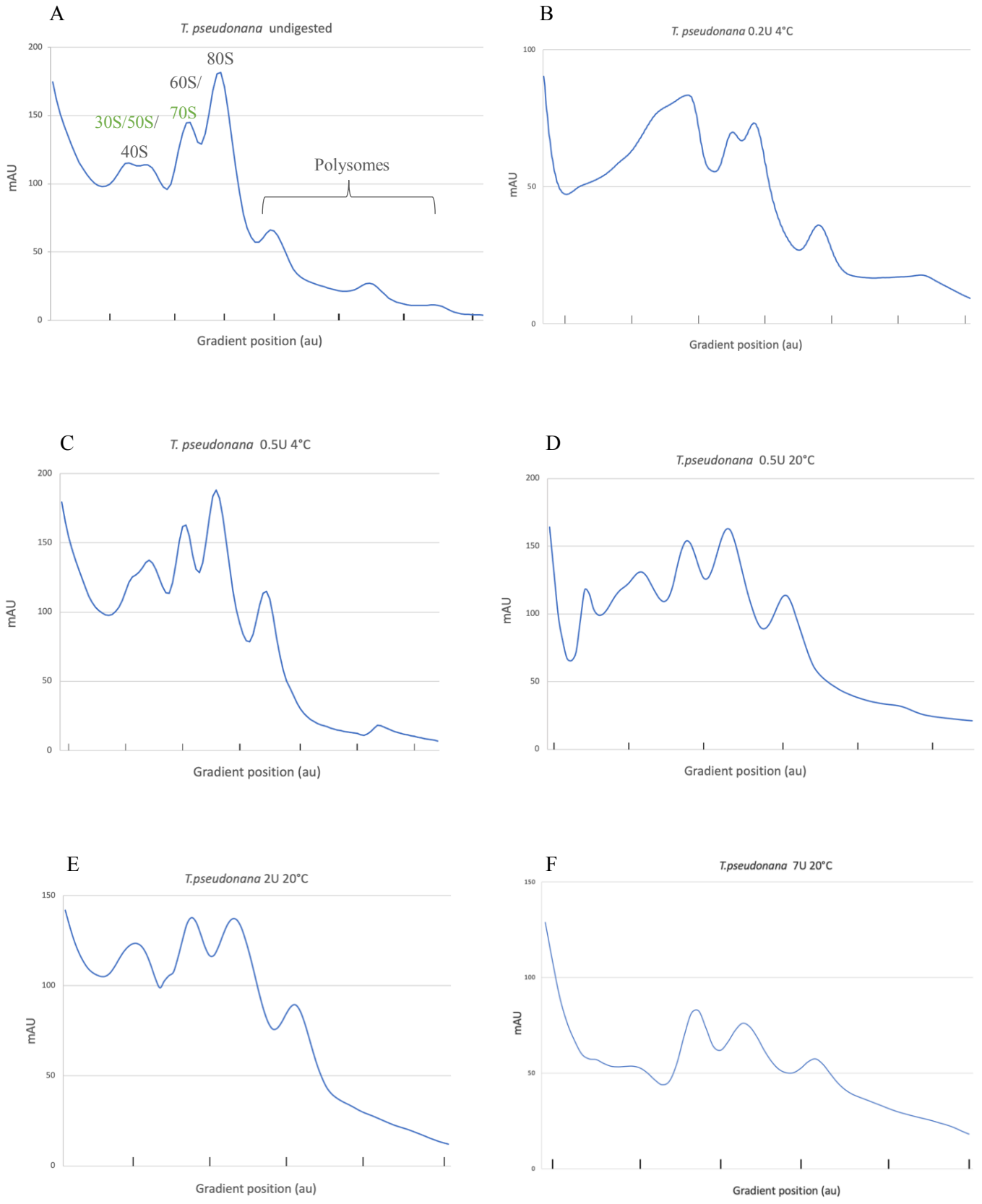


Figure 2.2: Density gradients. ***A****. Density gradient of undigested extracts from T. pseudonana cells. Peaks are derived from cytoplasmic ribosomes (black) and possibly chloroplast ribosomes (green).* ***B-F****. Density gradient of extracts treated with varying degrees of RNase I nuclease digestion. A strong 7U digest reduced the amount of polysome peaks but did not clearly result in a monosomal peak. milli-absorbance unit*

2.2.1.4 Isolation of ribosomes via ultracentrifugation

Different methods have been used in ribosome profiling studies to isolate digested monosomes, the most prevalent ones being ultracentrifugation via sucrose gradient and sucrose cushion (e.g. Meindl et al., 2022, Ingolia et al., 2012).

A density gradient centrifugation was performed to find the optimal nuclease digestion conditions for *T. pseudonana* which had not been established before. This method allows for evaluation of polysomal digestion and is the most reliable way to isolate monosomes without carrying over a large portion of messenger ribonucleoproteins (mRNPs) to the next steps in the protocol (McGlingy and Ingolia 2017). The disadvantage, however, is that this requires access to a FPLC system and a fractionation device, which was not available during protocol development at the University of East Anglia. Thus, once RNase I conditions have been established at the University of Regensburg, ultracentrifugation through a sucrose cushion was employed in our protocol.

For sedimentation of 80S ribosomes via sucrose cushion, centrifugation run times were adjusted for the use of a TLA120.2 rotor using the clearing factor or k factor, which represents the relative pelleting efficiency of a rotor (Heidcamp, 1995). McGlingy and Ingolia (2017) recommend pelleting the ribosomes by centrifugation in a TLA100.3 rotor at 100000 rpm, 4°C for 1 h (t_2). The clearing factor for TLA100.3 was k_2 = 14. To find an equivalent run on TLA120.2 and for a centrifuge time of t_1 = 2 h, the rotor speed was adjusted according to the following: first, the required adjusted clearing factor k was computed by the formula (Heidcamp, 1995):

$$\frac{t1}{k1} = \frac{t2}{k2} \rightarrow k_{adj} = 28$$

As the clearing factor of TLA120.2 was only k_1=8, rotor speed needed to be adjusted. This was computed using the following formula (Heidcamp, 1995):

$$k_{adj} = k_1 \left(\frac{maximum\ rotor\ speed}{desired\ \ rotor\ speed}\right)^2$$

Based on this equation, the desired rotor speed was 64142 rpm. To be confident the conditions are sufficient in pelleting 80S ribosomes, we centrifuged at 75000 rpm, 4°C for 2 h. These conditions resulted in the sedimentation of 80S particles visible at the bottom of the tube (Figure 2.3). Pelleted ribosomes are translucent, however the pellet also contains aggregates of the

photosystems which makes it easy to identify. The prominent green band represents the pigment containing photosystems of diatoms (Zill et al., 2019).

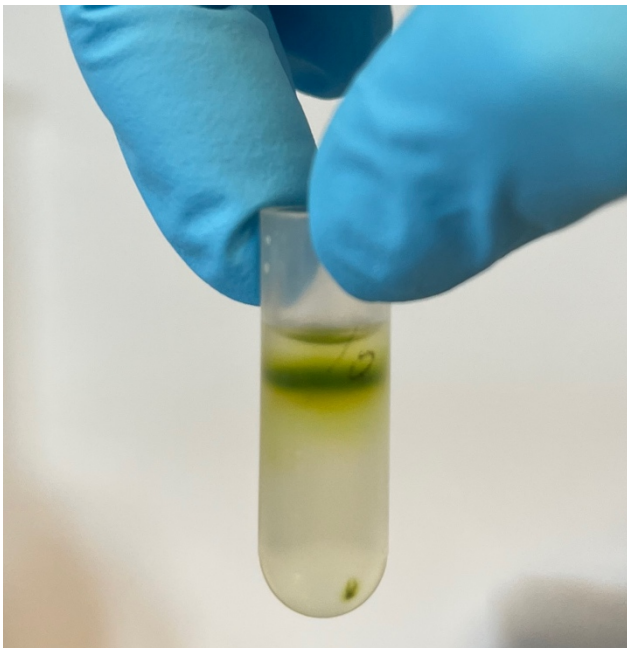


Figure 2.3: Ribosomal pellet is visible at the bottom of the tube after sucrose cushion centrifugation. The green band is representing the pigment containing photosystems.

2.2.1.5 Footprint fragment purification

Ribosome footprints are around 30 nt and need to be selected for by excision from a footprint fragment purification gel. To capture the footprints, two marker oligos (26 nt and 34 nt) were run alongside the samples. Figure 2.4 shows such a polyacrylamide size selection gel, with the red box highlighting the region to be excised. A negative control sample that was not treated with RNaseI (-RNaseI) was run alongside a nuclease treated sample (+RNaseI). The undigested sample was characterized by the presence of RNA in larger size ranges and the absence of significant material in the smaller size range compared to the digested sample. In 2014, Lareau et al. discovered another population of ribosome footprints with a length of ~20 nt, resulting from a conformational change of the translating ribosome. The authors proposed that the larger footprints originate from non-rotated ribosomes during the decoding stage of elongation, while the smaller ones originate from rotated ribosomes during the translocation process. The small footprints only accumulate when pre-treatment with the elongation inhibitor cycloheximide was omitted. This needs to be considered before starting any ribosome profiling experiment. Here,

we decided to only capture large footprints. However, by using markers with different size ranges, both large and small footprints can be captured.

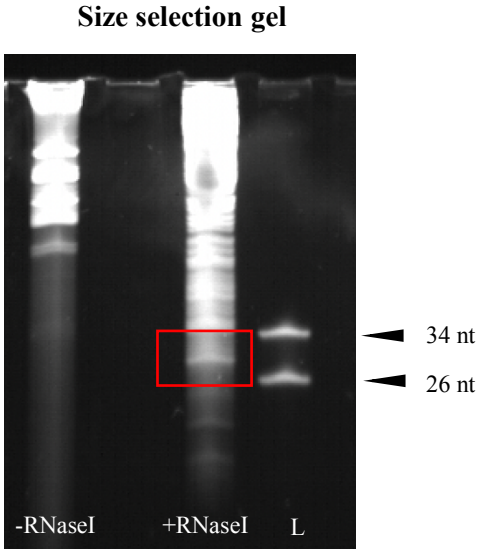


Figure 2.4: Image of a typical ribosome footprint size selection gel. -RNaseI: Undigested RNA from sucrose cushion ultracentrifugation (negative control). +RNaseI: Digested RNA from sucrose cushion ultracentrifugation. L: 26 nt and 34 nt size marker oligonucleotides. The red box highlights the region excised from the gel containing ribosome-protected fragments (RPFs).

Size selection via gel extraction typically results in low yield, thus an appropriate amount of input RNA needs to be determined. A minimum of 10-15 µg of RNA seems to generate enough purified footprints for further library preparation. For footprint fragment purification, we mainly followed the procedure outlined by McGlingy and Ingolia (2017). For gel extraction and purification, additional steps were implemented (Mohammad and Buskirk, 2019). These steps include shearing of the gel cut out and the use of spinX-columns, which consistently improved the RNA yield. We typically had a yield of 10 to 50 ng of RNA, corresponding to 1-5 pmol of ~30 nt fragments which was an optimal starting amount for library preparation.

2.2.1.6 Library construction

Construction of ribosome profiling libraries is a challenging and laborious process that can take up to several days. There are different strategies for ribosome footprint library preparation. The most commonly used is based on circularization of cDNA and a single adapter ligation step (Ingolia et al., 2012, Meindl et al., 2022). Dual ligation of both 3' and 5' adapters to the ribosome footprint has also been done (Weinberg et al., 2016). Different adapter ligation efficiencies can result in an uneven representation of reads, thus minimizing this ligation bias is crucial. A new approach for library construction is the ligation-free method based on template-switching activity of some reverse transcriptase (Hornstein et al., 2016).

The use of library preparation kits specifically made for small RNAs have found wide application in ribosome profiling studies due to their ease of application and reduced library preparation time of < 1 day (Reid et al., 2015; Simbriger et al., 2020). More recently, kits based on template-switching are starting to be employed in ribosome profiling studies (Tonn et al., 2021). However, ligation-free kits performed poorly in comparative studies due the formation of side products such as adapter dimers (Dard-Dascot et al., 2018).

Here, we decided to follow the recently published protocol by Meindl et al., (2022) for the following reasons. First, it uses a circularization approach and a 3' adapter which contains a degenerate sequence at its 5' end to reduce ligation bias. Secondly, it introduces unique molecular identifiers (UMIs) for bioinformatic identification of PCR duplicates. Further, this library preparation protocol is extremely sensitive and works with input material as low as 0.1 pmol of ribosome footprints and is thus ideal for the preparation of diatom derived ribosome footprints which sometimes yield low input amounts. The workflow is also considerably faster compared to similar protocols and can be completed in as little as 2 days. The protocol was adapted by implementing a subtractive hybridization step for removal of rRNA contaminants using custom depletion oligos. This will be discussed in detail in the next section.

2.2.1.7 rRNA depletion with biotinylated subtraction oligonucleotides

The majority of ribosome profiling sequencing reads comprise of contaminating fragments of rRNA. This is explained by the fact that rRNA accounts for up to 90% of total RNA in a cell. During ribosome profiling, ribosomal complexes are purified, further enriching rRNA content of the sample. These highly abundant transcripts are typically removed prior to sequencing to increase throughput of reads mapping to coding regions (Wilhelm and Landry, 2009).

Despite the implementation of rRNA removal steps in ribosome profiling protocols, the amount of rRNA fragments can still be above 80% (Gerashchenko et al., 2012; Fenton et al., 2022). The rRNA contamination is mainly derived from nuclease digestion, which causes nicks in rRNAs, generating fragments of similar size to ribosome footprints (Chung et al., 2015; Zinshteyn et al., 2020). Stringent gel purification of the expected size of the ribosome footprints reduces the number of rRNA fragments in the libraries, however, this approach is not possible when selecting footprints derived from different ribosomal conformations (Lareau et al., 2014). The fact that rRNA abundance varies between samples can be explained by slicing of ribosome

protected footprints during manual size selection from gels (Fenton et al., 2022). The composition and amount of contaminating rRNA fragments is specific to each organism and therefore the removal step needs to be customized. There are two approaches which are typically used to deplete rRNAs from ribosome profiling datasets – using commercially available depletion kits for RNA-seq and subtractive hybridization using custom biotinylated depletion oligonucleotides. A major limitation of rRNA depletion kits is that they are only available for use in a few model species. Zinshteyn et al., (2020) have tested several commercially available rRNA depletion kits and found them generally unsuitable for use with ribosome footprint libraries. Their data suggests that especially kits based on targeted nuclease cleavage can lead to ribosome footprint degradation, reduction of mappable reads, interference with global gene expression measurements and blurred nucleotide resolution.
The other routinely used approach is subtractive hybridization based on biotinylated antisense oligos which was first used in Ingolia et al., (2012). This approach requires initial sequencing data to design oligos complementary to the most abundant rRNA sequences. The step is species-specific and results in increased levels of mRNA with little to no bias in terms of fragment length or position and is thus the preferred method for rRNA depletion of ribosome profiling data (Zinshteyn et al., 2020).
An alternative approach to deplete rRNA using duplex-specific nucleases has been tested on *C. reinhardtii* libraries (Chung et al., 2015). This species-independent method increased the proportion of mRNA by up to fourfold.

For our protocol, we decided to implement a subtractive hybridization step. To identify the most abundant rRNA contaminants in a *T. pseudonana* library, sequencing reads from a small-scale ribosome profiling experiment, previously generated by Annemarie Eckes and Amanda Hopes, were analysed. The eight most abundant sequences comprised approximately 70% of all contaminants and were used to design biotinylated oligos for removal in subsequent large-scale experiments (Table 2.1). Two of the sequences mapped to the chloroplast 23S rRNA gene, which is an expected source of rRNA contamination in any photosynthetic organism. Out of those, sequence AAGTAGCATGGAGCACGTGGAATTC CGTGTGAAT represents the majority of contaminants, comprising almost 28% of all rRNAs. The other six prominent fragments were identified to be derived from 28S and 5.8S rRNA.

When depleting with magnetic beads, the ratio between beads and RNA needs to be carefully balanced as the addition of ‘empty’ beads (without hybridization probes in the reaction) to

footprint fragments can result in complete loss of the sample (personal communication Jan Medenbach). Despite the rRNA removal, our dataset still comprised a large number of fragments mapping to rRNA contaminants (> 70%), which is in accordance with other rRNA depleted datasets (Fenton et al. 2022). Mapping the reads to a comprehensive list of *T. pseudonana* rRNAs showed lower alignment rates for depleted samples (Table 2.3), however no significant difference to undepleted samples was detected. To assess the efficiency of the subtractive hybridization step, we analysed the most abundant rRNA sequences before and after depletion and found that almost 100% of all targeted rRNA fragments were removed. Due to the variation between samples and experimental conditions, depletion oligos would ideally be re-designed and/or more oligos included to cover a larger range of rRNA genes (Fenton et al., 2022). Implementation of efficient rRNA removal steps can be expensive, especially when working with non-model organisms like diatoms for which no commercial kits are available. siTOOLs Biotech offers the development of a riboPOOL, a mixture of several hundred DNA oligonucleotides covering any organisms entire rRNA. However, even this highly optimized

Table 2.1: 5' biotinylated antisense oligos, derived from small-scale experiment used for removal of targeted rRNA fragments via subtractive hybridization.

Oligos	DNA sequence	Antisense sequence	Target	% of contaminants
# 1	AAGTAGCATGGAGCACGTG GAATTCCGTGTGAAT	/5biotin-TEG/ATTCACACGGAATTCCACGTGCTCCAT GCTACTT	23S Chloroplast	27.7
# 2	AGTGCCTTCTGTTCATGTC CTGGGTCGGGCTGAGGT	/5biotin-TEG/ACCTCAGCCCGACCCAGGACATGAACA GAAGGCACT	28S	21.6
# 3	GTGCATCGAATTGTGGTCT GGAGAAGTA	/5biotin-TEG/TACTTCTCCAGACCACAATTCGATGCA C	5.8S	7.9
# 4	AAGGAACGTGGTACTGTAA GCATGAGAGTAGCC	/5biotin-TEG/GGCTACTCTCATGCTTACAGTACCACG TTCCTT	28S	5.0
# 5	TCGAGGGACGGAGAAGGCT AAGCTAGCC	/5biotin-TEG/GGCTAGCTTAGCCTTCTCCGTCCCTCG A	23S Chloroplast	4.1
# 6	GTGAATCATCAAACTTTTG AACGCACATTGCGCTTTT	/5biotin-TEG/AAAAGCGCAATGTGCGTTCAAAAGTTT GATGATTCAC	5.8S	3.2
# 7	TCGATGTCGGCTCTTCCTA TCATTGTGTC	/5biotin-TEG/GACACAATGATAGGAAGAGCCGACATC GA	28S	3.1
# 8	TGCCCTCGGCCTATTCTCA AACTTT	/5biotin-TEG/AAAGTTTGAGAATAGGCCGAGGGCA	28S	2.8

strategy only leads to a moderate decrease of rRNA fragments in ribosome profiling datasets, approximately doubling mRNA-derived reads. While we decided to implement a rRNA removal step in our library preparation protocol, omitting this step and instead increasing the sequencing depth, combined with bioinformatically removing the contaminants, is also an option.

2.2.1.8 Analysis of fractions to verify presence of footprints

Two fractions (F_04, F_05, Figure 2.5) derived from *T. pseudonana* extracts treated with 7U RNase I were sequenced to verify the presence of RPF and thus confirm whether the selected digestion conditions were ideal. Table 2.2 summarizes the results from bioinformatic analysis. For both fractions, approximately 70% of the pre-processed reads mapped to the *T. pseudonana* genome.

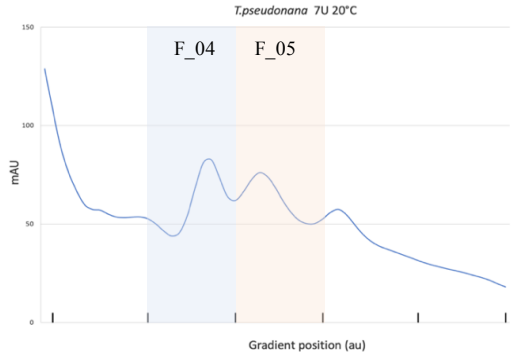


Figure 2.5: Gradient profile of 7U RNase I digested T. pseudonana extracts. The two sequenced fractions which are presumably containing 80S ribosomes are highlighted in blue (F_04) and red (F_05).

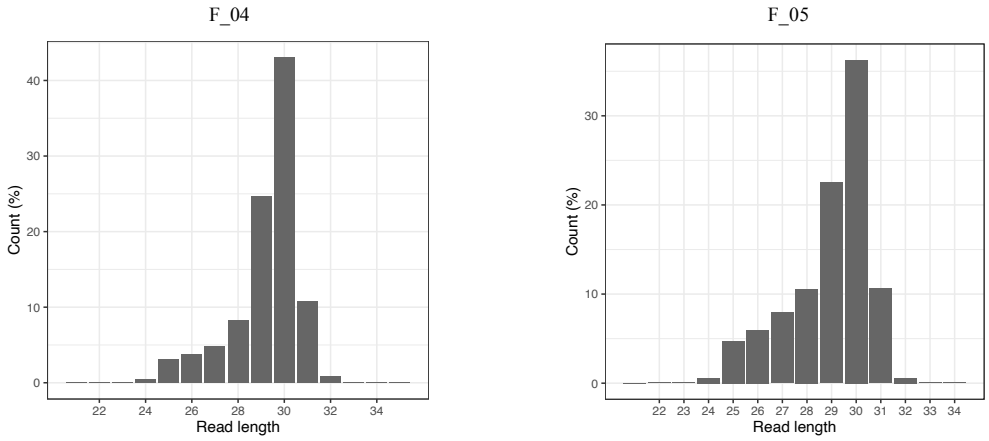
Read length distribution reveals a peak of 30 nt long reads across both fractions (Figure 2.6A). A characteristic feature of high-quality ribosome profiling data is a strong periodicity score, which arises from the translocation of a ribosome along the mRNA three nucleotides at a time. Both fractions show a strong nucleotide periodicity, which is evident by an enrichment of ribosomal P-sites (corresponding to start codon) in the first reading frame on the CDS but not on the 5' and 3' UTRs (Figure 2.6B). This sequencing data suggests that genuine footprints are present in both fractions, with slightly more reads mapping to CDS in F_04 (Figure 2.6C). This indicates that a 7U digest represents optimal nuclease digestion conditions resulting in reads displaying triplet periodicity which is indicative of high-quality ribosome footprint data.
Separation in the gradients is not perfect but it serves to enrich for meaningful reads and to reduce the number of contaminants. Our previous assumption that first peak is mainly derived from chloroplasts might be wrong, as the amount of reads mapping to the *T. pseudonana* nuclear

genome after rRNA removal is equally high in both fractions (Table 2.2). This small-scale analysis only served to verify the presence of footprints with our chosen digestion conditions. Gradient profiles generated during the preparation of the samples used in our protocol paper depicted a different profile with one 80S peak (Figure 2E in Appendix B).

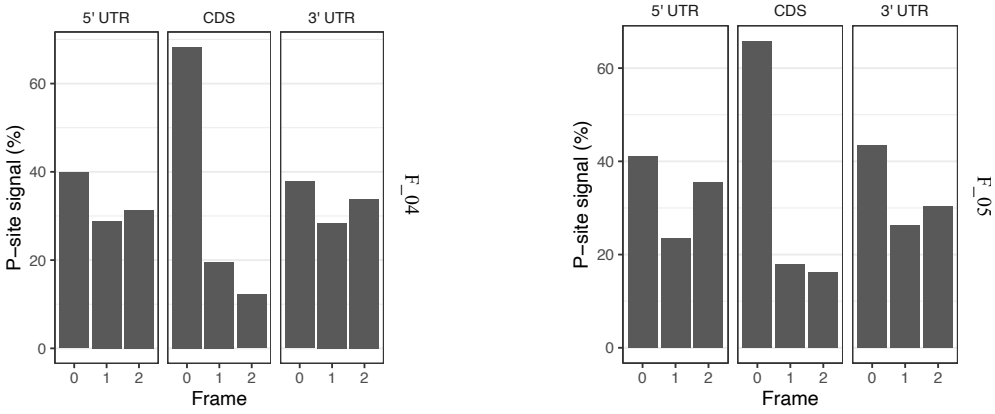
Table 2.2: Run metrics and analysis obtained from sequencing of fractions in an initial ribosome profiling experiment.

Fraction	# of raw reads	# of trimmed and filtered reads	% of reads mapping to rRNAs	% of unmapped reads mapping to genome
F_04	2696139	1667976	73.01	73.35
F_05	3217700	2479256	86.81	72.3

A



B



C

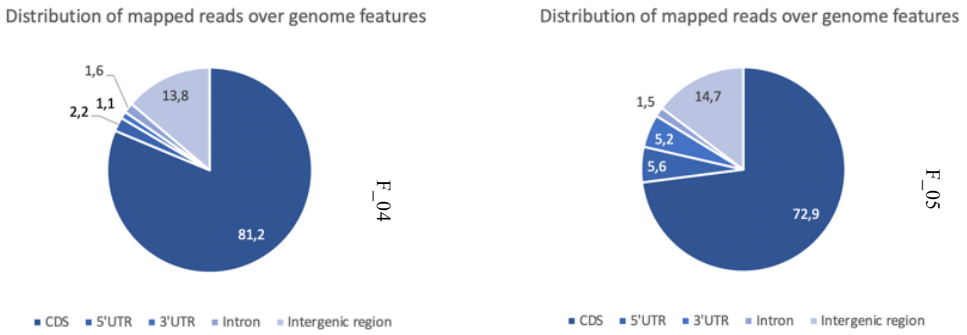


Figure 2.6: Analyses of two fractions from a 7U digest of T. pseudonana extracts. Fraction F_04 on the left, Fraction F_05 on the right. ***A*** *Read length distribution. The number of reads according to length (between 25 and 35 nucleotides).* ***B*** *Percentage of P-sites in the three reading frames (Periodicity score) along the 5'UTR, CDS and 3'UTR.* ***C*** *Distribution of reads across known gene features.*

2.2.2 Ribosome profiling of *T. pseudonana* cells to verify the protocol

To evaluate the performance of our ribosome profiling protocol and to assess the quality of the generated footprint data, we tested it on *T. pseudonana* wild-type and experimental samples. To compare the impact of the implemented rRNA removal step via subtractive hybridization, both depleted and undepleted samples were sequenced.

Total RNA derived from 2 x 200ml culture replicates, treated with nuclease and ultracentrifuged typically yielded a clearly visible 80S ribosomal pellet. Visualizing the purified RNA on a polyacrylamide gel resulted in a distinct gel pattern. Gel extraction of the RPF resulted in 10-40 ng which was sufficient input amount for library preparation.
PCR cycle optimization was performed with 9, 12 and 15 cycles to find the ideal number for each library (Figures 2.7 and 2.8). Different cycle numbers were chosen for each library depending on the concentration of PCR product and the occurrence of daisy chains, which are fragments shifting into higher size ranges and are an indication of over-amplification (Huppertz et al., 2014). Increasing the number of PCR cycles subsequently also increased the number of 'empty' amplicons. Those are derived from the rApp-L7 adapter failing to ligate to RNA and acting as a template for the extension of the P7 RT oligonucleotide during reverse transcription, resulting in PCR amplification of the partially extended RT oligos (Meindl et al., 2022). However, those 'empty' amplicons were efficiently removed in a final purification step by automated agarose gel electrophoresis and not subjected to sequencing. Depleted libraries are usually expected to require a higher number of PCR cycles. This was not the case for our libraries, possibly due to different amounts of input RNA used.

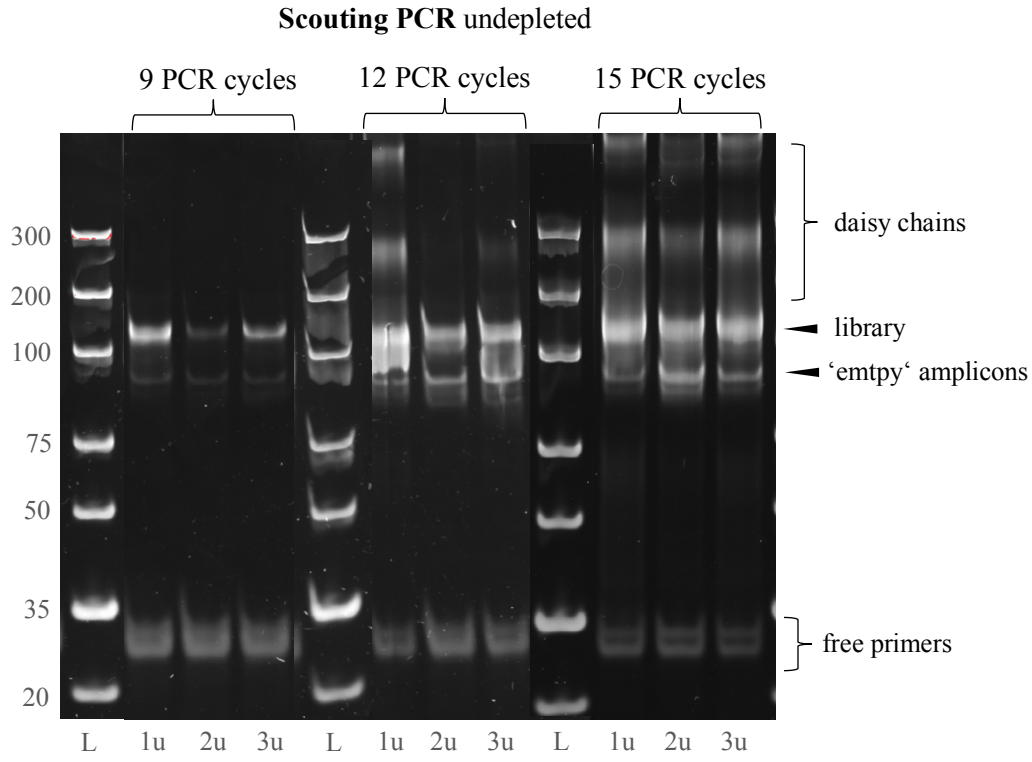


Figure 2.7: Gel after scouting PCR (undepleted). Libraries (1u-3u) were prepared from T. pseudonana RPFs without a subtractive hybridization step for rRNA removal. The size of the final amplicon (with a ~30nt insert) is approximately 170 bp. L: GeneRuler Ultra Low Range DNA Ladder.

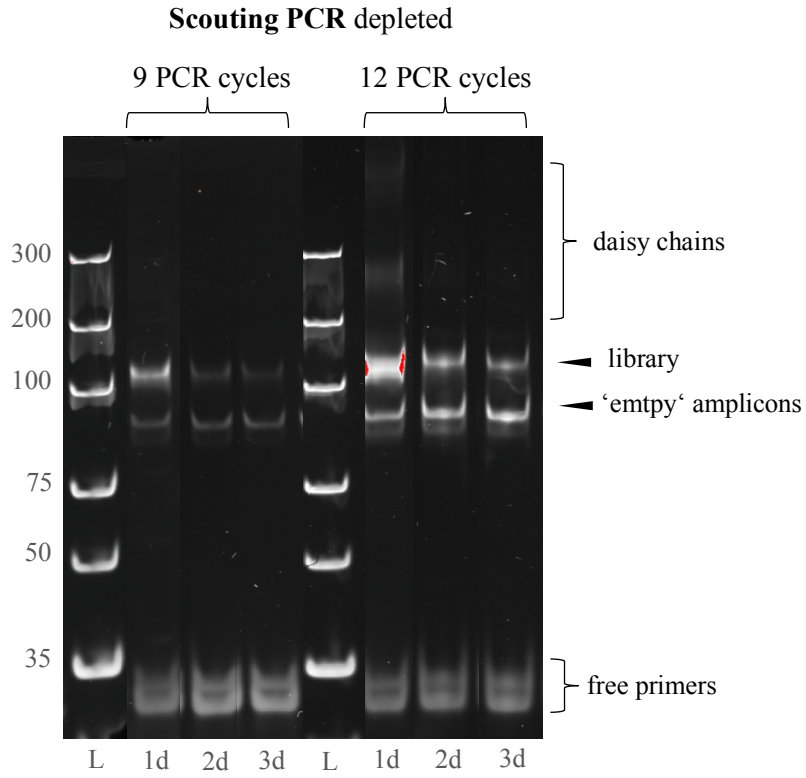


Figure 2.8: Gel after scouting PCR (depleted). Libraries (1d-3d) were prepared from T. pseudonana RPFs with a subtractive hybridizations step for rRNA removal. The size of the final amplicon (with a ~30nt insert) is approximately 170 bp. L: GeneRuler Ultra Low Range DNA Ladder

Several computational analyses were performed to validate the quality of the generated footprint data. We found that 60-90% of the pre-processed reads aligned to rRNA genes (Table 2.3). After bioinformatically removing the rRNA contaminants, 57-82% of the reads across all samples mapped to the *T. pseudonana* genome. Removal of UMIs revealed that the number of PCR duplicates in our libraries was between 19-58%. While this is still in an acceptable range (Buchbender et al., 2020), Meindl et al. (2022) report significantly fewer PCR duplicates. Even with low amount of input material, their libraries contained less than 2% PCR duplicates. To reach sufficient cDNA concentration, some of our libraries required 12-15 cycles of amplification, which is a relatively high and probably contributed to the number of PCR duplicates (Buchbender et al., 2020).

Table 2.3: Overview of sequencing metrics for our ribosome profiling libraries. Only reads with lengths between 16 and 40 nt were considered for mapping to rRNAs using Bowtie2. All reads which did not align to rRNAs were further mapped to the T. pseudonana genome using STAR.

Sample	# of raw reads	% of reads mapping to rRNAs	% of unmapped reads mapping to genome	% PCR duplicates
Ribo_undepleted.1	43782636	80.74	69.32	19.21
Ribo_undepleted.2	41691610	78.62	75.3	33.5
Ribo_undepleted.3	42626206	89.33	68.88	22.12
Ribo_Ctrl.1	57958849	74.85	74.07	19.84
Ribo_Ctrl.2	52538883	76.53	78.31	31.45
Ribo_Ctrl.3	40049661	70.89	71.1	37.72
Ribo_HT4.1	42804034	73.33	74.95	31.52
Ribo_HT4.2	39673611	78.82	69.26	24.84
Ribo_HT4.3	45366929	83.48	56.88	24.73
Ribo_HT24.1	44679566	66.14	75.65	45.41
Ribo_HT24.2	39073057	63.67	78.64	40.31
Ribo_HT24.3	47975947	82.36	68.9	20.74
Ribo_HL4.1	48081630	78.08	70.86	58.68
Ribo_HL4.2	38246722	75.11	74.23	39.3
Ribo_HL4.3	41364340	81.47	80.36	23.87
Ribo_HL24.1	42332929	80.09	75.75	24.55
Ribo_HL24.2	42928262	79.09	81.82	24.49
Ribo_HL24.3	44708557	89.11	76.46	23.95

The distribution of read lengths is a quality metric for ribosome profiling data and is expected to be around 30 nt. Our results showed a peak of ribosome-protected fragments at 31 nt for all samples (Figure 2.9A). As expected for ribosome profiling data, which gives a snapshot of active translation, the majority of reads across all samples mapped to CDS. This was shown by analysing the percentage of ribosomal P-sites (corresponding to the start codon) which are located in the 5' UTR, CDS and 3'UTR of mRNAs (Figure 2.9C shows Ctrl.1, for all other samples see Appendix C). Triplet periodicity of ribosome footprints along CDS is another distinct characteristic of ribosome profiling data. This is due to the ribosome moving along the mRNA 3 nt at a time during elongation. We observed an enrichment of ribosomal P-sites in the first reading frame on the CDS but not on the 5' and 3' UTRs (Figure 2.9B and 2.9D for Ctrl.1, also see Appendix C). The heatmap (2.9D) analyses all read lengths separately for sample Ctrl.1 and reveals that fragments of 30-32 nt have a strong frame preference for one of the three subcodon positions. The triplet periodicity is further visualized using metaprofile plots which display the distance of the P-sites of aligned reads to the start and stop codon of annotated coding sequences. The peaks around the start and the end of the CDS are attributed to longer ribosome dwell times at these positions due to initiation and termination taking more time than elongation at a codon. This phenomenon is typically seen in ribosome profiling datasets.

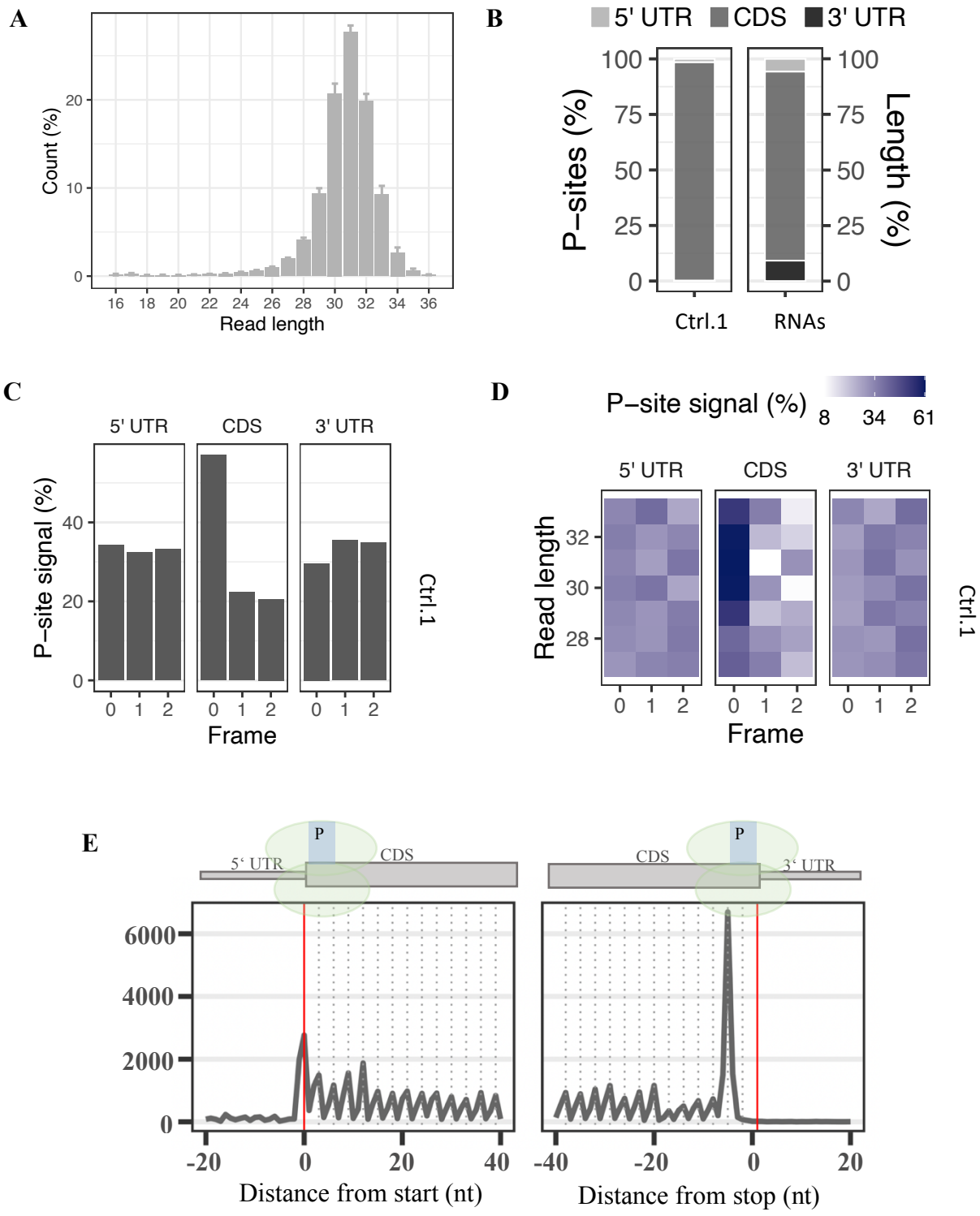


*Figure 2.9: Quality control of ribosome profiling data. **A** Average read length distribution. **B** Left, percentage of P-sites in the 5'UTR, CDS and 3'UTR. Right, percentage of region lengths in mRNAs sequences (Sample Ctrl.1). **C** Triplet periodicity of the footprint data is shown by the percentage of P-sites falling into one of the three reading frames for 5'UTR, CDS and 3'UTR. **D** Percentage of P-sites in the three frames along the 5' UTR, CDS and 3' UTR, stratified for read length (Sample Ctrl.1). **E** Metaprofile showing the periodicity of ribosomes along the coding sequence based on P-sites mapping around the start and stop codon (Sample Ctrl.1).*

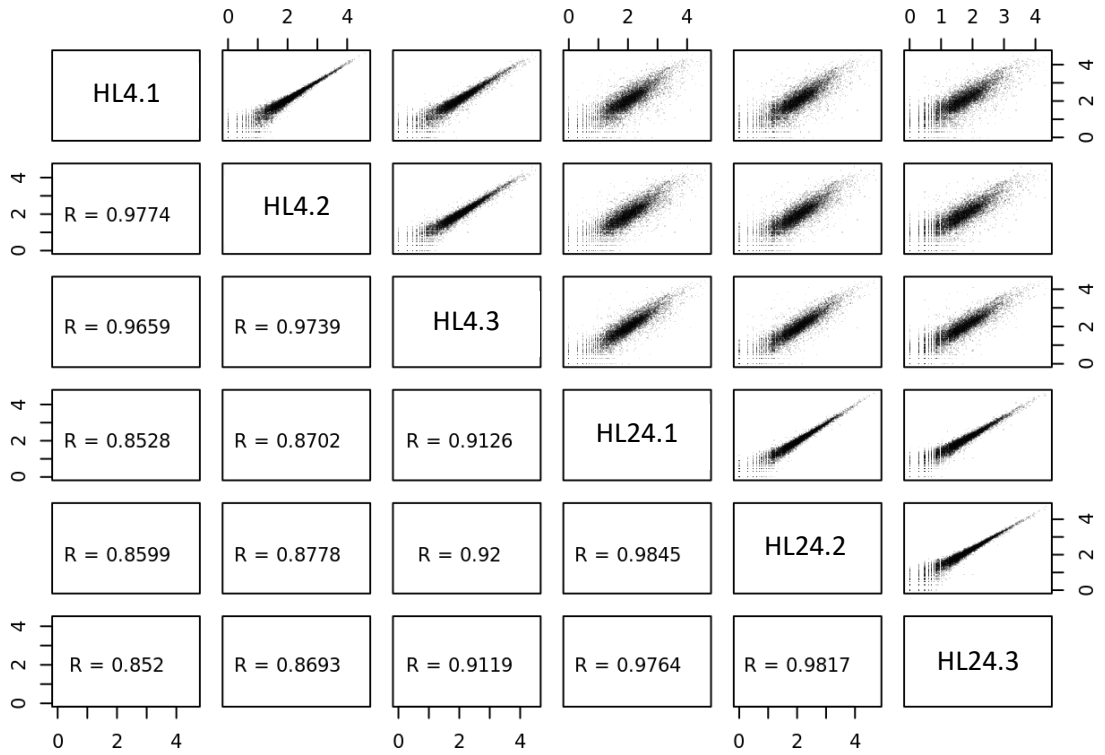


Figure 2.10: Statistical validation. Scatter plots for T. pseudonana wild-type cells (HL= High light treatment for 4 h and 24 h) show high levels of reproducibility between biological replicates indicated by a Spearman's correlation coefficient >0.96. The scatter plots show the log_{10} read counts per gene in a sample relative to another sample.

Once quality control of the ribosome profiling data was done, we performed statistical analysis to assess the reproducibility of biological replicates. We found a high Spearman correlation of > 0.8 for replicates from the same biological condition. Figure 2.10 shows the correlation between and within biological replicates of the cell treated with high light (see also Appendix C). Further, principal component analysis (PCA) of the Ribo-Seq data as well as of RNA-Seq data which was sequenced in parallel was done to assess inter- and intragroup variability (see Figure 7.6 Appendix C). Samples are displayed along the first component (PC1) and the second component (PC2), with PCA1 accounting for 43% of the variance in the Ribo-Seq data and for 47% of the variance in the RNA-Seq data. The biological replicates generally exhibit a strong grouping pattern, with the exception of HT24.3 in the Ribo-Seq data. HT24.3 stands out with a noticeably lower correlation coefficient when compared to the other samples. However, this correlation still falls within an acceptable range. Results of a Spearman's rank correlation analysis of the corresponding RNA-Seq data can be found in Table 7.1 in Appendix C.

These results demonstrate that we have developed a robust protocol capable of generating high-quality ribosome profiling data. This data is suitable for use in further gene expression analysis as done in the next chapter of this thesis.

2.4 Conclusion

We have developed a ribosome profiling protocol to study genome-wide translation in *T. pseudonana*. Our protocol is based on previously published protocols with several adaptions including sample harvesting strategy, RNase I digestion conditions and rRNA depletion. We demonstrated that the protocol is capable of generating high-quality sequencing data of actively translating ribosomes. We detected a strong enrichment of ribosome profiling reads in the canonical coding sequences. Our data further displays triplet periodicity, a bias generated from an elongating ribosome moving along the mRNA one codon at a time. This protocol can easily be adapted for use with other diatom species. Now that a robust ribosome profiling protocol is in place it can be used for future gene expression analysis. This ribosome profiling protocol will be a valuable resource adding to the diatom molecular toolbox.

Chapter 3: Investigating the response of *Thalassiosira pseudonana* to high light stress by ribosome profiling and RNA-seq

3.1 Introduction

Diatoms live in a highly dynamic marine environment and need to cope with abrupt and unpredictable changes in irradiance. As photosynthetic organisms they need to efficiently acquire photons (light-harvesting) under low light conditions but also protect their photosynthetic apparatus from light damage (photoprotection) when the light energy exceeds the uptake capacity (Nymark et al., 2009). Light absorption and transfer of energy takes place in two protein-pigment containing complexes, photosystem II (PSII) and photosystem I (PSI), which are located in the thylakoid membranes inside the chloroplasts. Photodamage occurs mainly in the PSII complex, where light-driven water oxidation takes place. Photosynthetic organisms possess light-harvesting complex proteins (LHCs) which are bound to the PS cores and play a major role in light-harvesting and photoprotection. Diatoms possess a huge number of members of the nucleus-encoded LHC superfamily, known as fucoxanthin (Fx) chlorophyll (Chl) a/c-binding proteins (FCPs) which have the capacity to harvest light and dissipate excess energy (Wang et al., 2020). These FCPs are divided into three groups: the major fucoxanthin Chl a/c proteins (Lhcf), the red algal-like proteins (Lhcr) and the green algal-like proteins (Lhcx) (Dong et al., 2016).

The effect of high light stress is associated with the production of harmful reactive oxygen species (ROS) which inactivate PSII mainly by targeting the primary electron-accepting protein D1 (Vass et al., 2007). Phytoplankton have evolved various cellular mechanisms to regulate the rate of photosynthesis when exposed to fluctuating light regimes. Some important mechanisms include PSII and PSI electron cycles, fast repair of the D1 protein of the PSII reaction centre and non-photochemical fluorescence quenching (NPQ) (Müller et al., 2001; Brunet and Lavaud, 2010). The latter is one of the most important short-term photoprotective mechanisms in chloroplasts of plants and algae and is attributed to the xanthophyll cycle (XC). In diatoms, the main XC comprises the de-epoxidation of the pigment diadinoxanthin (Ddx) to diatoxanthin (Dtx) under high light, triggered by an increase in pH of the thylakoid lumen. The accumulation of Dtx is a prerequisite for NPQ, which leads to dissipation of excess excitation energy as heat (Goss and Jakob, 2010). Photoprotective responses have been extensively studied on the transcriptome and the proteome levels providing new insights into metabolic pathways and cellular mechanisms. In *T. pseudonana*, the Lhcx6 protein is suggested to bind to Dtx and plays

a role in heat dissipation via NPQ under high light stress (Zhu and Green, 2010). In *Phaeodactylum tricornutum*, D1 protein degradation and re-synthesis rates were increased under high light stress (Domingues et al., 2012). Nymark et al. (2009) performed global transcriptional profiling and revealed that photoprotective processes in *P.tricornutum* can be divided in three distinct response phases: an initial response phase (0-0.5 h), an intermediate acclimation phase (3-12 h) and a late acclimation phase (12-48 h). A proteomic analysis of light protection mechanisms in *T. pseudonana* revealed 143 differentially expressed genes under high light (Dong et al., 2016). However, translational regulation of gene expression upon light stress is still largely unexplored in diatoms. Ribosome profiling is a method based on deep sequencing of ribosome-protected fragments (RPF) and quantifies ribosome density of transcripts on a genome-wide level (Ingolia et al., 2009). Combined with RNA sequencing, the translation efficiency (TE) of genes can be evaluated to reveal translational regulation. TE is basically the number of ribosomes per gene, normalized to transcript abundance. A gene is considered to be translationally regulated if the TE changes between conditions and the changes in the number of RPFs are not accompanied by changes in mRNA counts (Chothani et al., 2019).

Here, we apply our newly developed ribosome profiling protocol to identify translational regulation in *T. pseudonana* in response to high light stress. This is the first ribosome profiling study to investigate photoprotective responses in a marine diatom providing insights into regulatory mechanisms of this important phytoplankton.

3.2 Materials and Methods

3.2.1 Cultivation and experimental design

Thalassiosira pseudonana (strain CCMP1335) was incubated at 20˚C in ½ salinity Aquil media (pH~8) with constant irradiance of approximately 40 µmol photons $m^{-2}s^{-1}$ and was kept in exponential growth phase for 3 weeks to ensure acclimation of cells. Untreated samples (low light, LL) were harvested (T0) in mid-exponential phase (500.000-700.000 cells/ml) before the rest of the cells were cultured at 500 µmol photons $m^{-2}s^{-1}$ (high light, HL). Treated samples were harvested at time points +4 h and +24 h after transfer to HL. Experiments were done in triplicates for each time point.

3.2.2 Ribosome profiling

Ribosome profiling data was generated following our protocol described in Chapter 2 (for detailed protocol see Appendix B). Libraries were generated together with Andreas Meindl and Markus Romberger and sequenced on an Illumina MiSeq (V3 kit, 150 cycles, 90 cycles SE).

3.2.3 RNA sequencing

50 ml of culture per sample was harvested by vacuum filtration onto 47 mm, 1.2 µm Isopore filters (Millipore) and immediately flash-frozen in liquid nitrogen. RNA was isolated using the Direct-zol RNA Miniprep Kit (Zymo Research) with some modifications. Cell lysis was done by adding 1ml of 65˚C preheated TRI reagent and sterile glass beads (425-600 µm, Sigma-Aldrich) to the frozen filter and bead beating for 2 min at maximum speed (BioSpec 3110BX Mini-BeadBeater-1). Samples were centrifuged for 2 min at 15000 g at 4˚C and the supernatant was transferred into a nuclease free tube. RNA was purified according to the manufacturer's protocol, with an in-column DNase I treatment. For all samples, RNA was eluted in 30 µl nuclease free water and quantified on a NanoDrop (ND-2000 spectrophotometer; Thermo Scientific). RNA quality was assessed by running it on a 1% TAE agarose gel at 60 volts for 1 hour. Libraries were prepared using the NEBNext Ultra II Directional RNA library prep kit with poly(A) enrichment and sequenced on an Illumina NovaSeq 6000 (SP, v1.5 kit, 100 PE) by the Earlham Institute (Norwich, UK).

3.2.4 Data analysis

Ribosome profiling sequencing data was processed as described in Chapter 2. RNA-seq data was de-multiplexed and adapter trimming was done with Trim Galore! (version 0.6.5). Reads were mapped against the *T. pseudonana* reference genome (Thaps3_chromosomes _assembly_chromosomes_repeatmasked.fasta) with HISAT2 (version 2.1.0) using standard parameters. SAMtools (version 1.10) was used to generate sorted and indexed bam files. mRNA and RPF reads for each gene were counted using the featureCounts command in the Rsubread package (version 2.8.2). Detection of genes with different translation efficiencies (TE) was done in R (version 4.2.1) (R Core Team, 2021) following the ΔTE approach (Chothani et al. 2019). Differentially translated genes were determined based on significant change in TE and an associated false discovery rate (FDR) < 0.05. Functional analysis of genes with up- or downregulation of TE between normal and light stress conditions was done using Gene Ontology (GO) enrichment analysis and EuKaryotic Orthologous Groups (KOG) analysis.

3.3 Results and Discussion

3.3.1 Detection of differentially transcribed genes and differential translation efficiency genes under 4 h of high light stress

To investigate how *T. pseudonana* deals with high light stress, ribosome profiling and RNA-seq data was integrated to calculate translation efficiency (TE). Genes are grouped as differential translation efficiency genes (DTEGs) if their TE changes between conditions and the change in RPFs cannot be explained by a change in mRNA count (Chothani et al., 2019). In other words, differential translation happens if the change between RPF and mRNA levels is significantly different between two conditions, using a false discovery rate of < 0.05. We identified 461 DTEGs in *T. pseudonana* under 4 h of light stress. Genes falling into the class of differentially transcribed genes (DTGs) are transcriptionally regulated with changes in mRNA counts matching the changes in RPFs. 2822 DTGs were detected in the dataset, indicating that

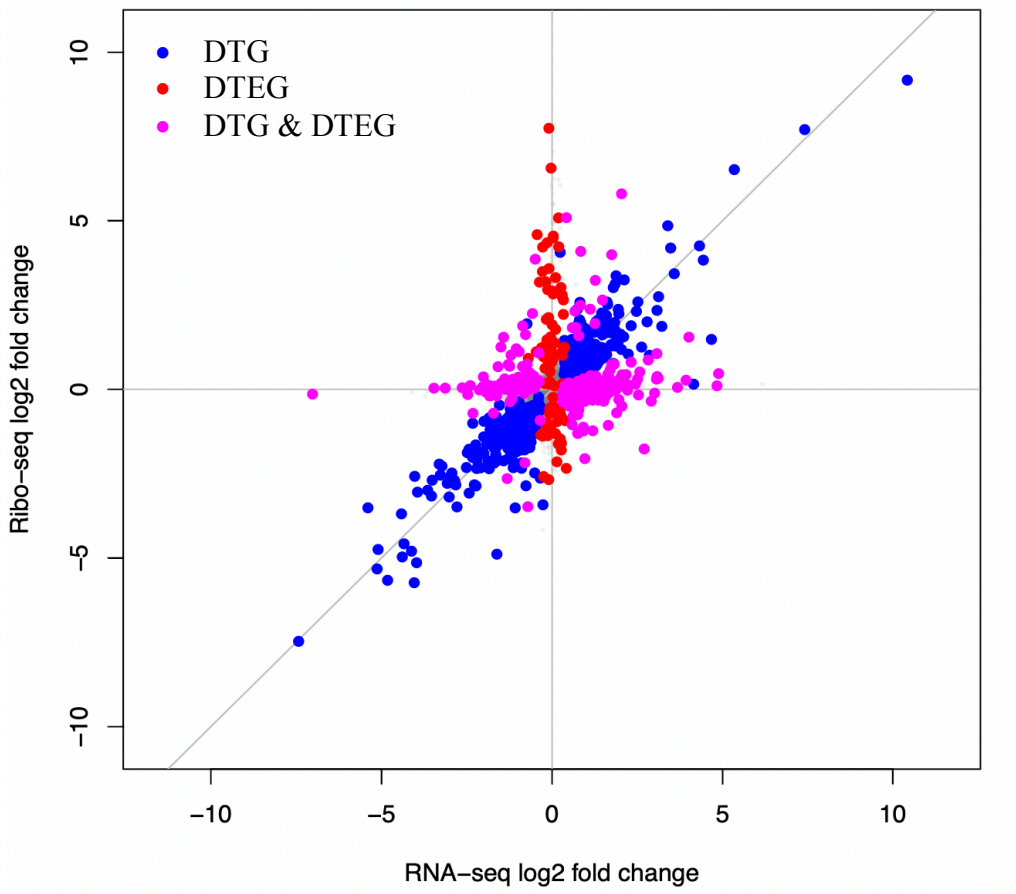


Figure 3.1: Scatter plot of log fold changes for each gene in the ribosome profiling and the RNA-seq data under 4 h of HL vs the control group. Differentially transcribed genes (DTGs, in blue), differentially translation efficiency genes (DTEGs, in red, exclusive) and genes that are both DTG and DTEG (in pink) (intensified and buffered).

for most genes transcriptional regulation is the major component of short-term light stress response. Genes can be DTG and/or DTEG, thus they are further divided into regulatory classes: exclusive, intensified, or buffered, depending on changes in RPF, RNA and TE. We detected 93 exclusive genes which exhibited a change in TE but not in transcription, which suggests that they are translationally driven. Buffered and intensified genes show changes in translational efficiency (ΔTE) and transcription (ΔRNA), thus are regulated both transcriptionally and translationally. We discovered 248 buffered genes which exhibited a change in TE which is counteracting the change in RNA. Twenty genes were considered intensified with their translational change leading in the same direction as their transcriptional change (Chothani et al., 2019). Figure 3.1 plots all genes falling into distinct groups according to changes in expression levels between control conditions and 4 h HL stress. Our results suggest that HL stress is leading to altered levels of gene expression, regulated both transcriptionally and translationally, which is in accordance with studies in other organisms focusing on revealing response mechanisms to various kinds of abiotic stress (Lei et al., 2015; Zhang et al., 2017). *T. pseudonana* has 11242 predicted protein-coding genes (Armbrust et al., 2004b), of those, we have categorized 25.1% as DTGs, 4.1% as DTEGs and 0.8% as exclusive genes upon 4 h of HL stress (Table 3.1).

Table 3.1: Percentage of genes expressed upon 4 h and 24 h of high light stress in relation to the total number of genes in the genome of T. pseudonana. Genes are grouped as differentially transcribed genes (DTGs), differential translation efficiency genes (DTEGs) and genes which are exclusively translationally regulated (exclusive genes).

Time point	Category	% of total genes in genome
+ 4 h	**DTGs**	25.1%
	DTEGs	4.1%
	Exclusive genes	0.8%
+ 24 h	**DTGs**	58.2%
	DTEGs	13%
	Exclusive genes	2.1%

3.3.2 Functional analysis of genes with differential TE upon 4 h of light stress

Short-term high light stress resulted in a differential expression of genes, including 247 genes with upregulated TE and 214 genes with downregulated TE (Figure 3.2). The Gene Ontology (GO) was used for specifying molecular function, biological processes and cellular components associated with the identified DTEGs. They were involved in 42 biological processes (BP) and 139 molecular functions (MF) and 13 cellular components (CC) (Tables 7.2-7.4 of Appendix D). However, none of these were found significantly enriched ($p_{adj} < 0.05$).

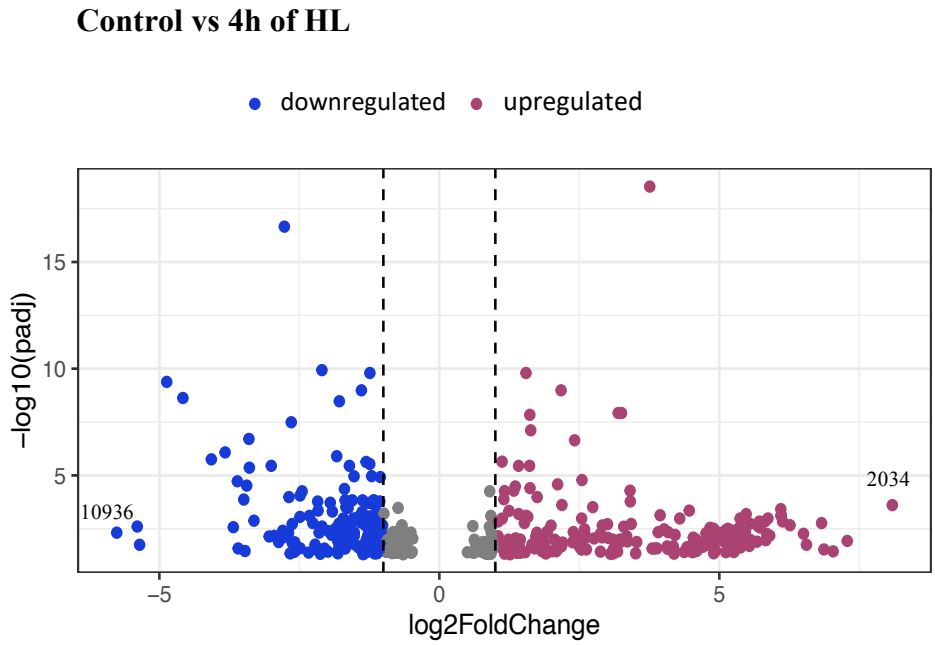


Figure 3.2: Volcano plot of upregulated and downregulated DTEGs of cells under 4 h of HL vs the control group. Blue dots represent downregulated genes with log2FC < 1. Red dots represent upregulated genes with log2FC > 1. Dash lines indicate log2FC values -1 or +1. The –log10 (adjusted p values) represents the level of significance of each DTEG. Genes with an adjusted p value of < 0.05 were assigned as DTEGs. Genes with the highest up- and downregulation were annotated with ProteinID numbers from JGI.

In general, changes of light-harvesting complex and xanthophyll cycle genes in RPFs were matching the changes in mRNA counts. Thus, those genes were not found in the ΔTE group, with the exception of two FCP genes. The Lhcr5 gene showed a 3.25-fold-change increase in TE and classified as intensified, meaning it is both transcriptionally and translationally regulated. FCPs are involved in photoprotection via NPQ in *T. pseudonana* during HL stress (Zhu and Green, 2010; Dong et al., 2016). These results agree with a proteomic study which

showed that the protein expression levels of Lhcr5 increased by 3.68-fold after 10 h of HL treatment (800 µmol photons $m^{-2}s^{-1}$) (Dong et al., 2016).

3.3.3 Exclusively translationally regulated genes under 4 h of light stress

Upon 4 h of high light stress, 93 exclusively translationally regulated genes, with a false discovery rate of < 0.05, were detected. Of those, 56 were upregulated and 37 were downregulated, suggesting a significantly higher number of upregulated genes as demonstrated through binominal distribution testing (p = 0.03). Figure 3.3 shows all exclusively downregulated and upregulated genes with log2 fold change of -1 or +1 under 4h of HL vs control conditions. A detailed list of all genes with differential translational regulation (p_{adj} < 0.05) under 4 h of HL is provided in Table 7.9 of Appendix D. GO term analysis of exclusively translationally regulated genes under 4 h of high light stress revealed that 11 GO terms fell in the BP category, 15 in the MF category and 4 in the CC category (Figure 3.4 A-C), however, none of these results were significant (p_{adj} < 0.05).

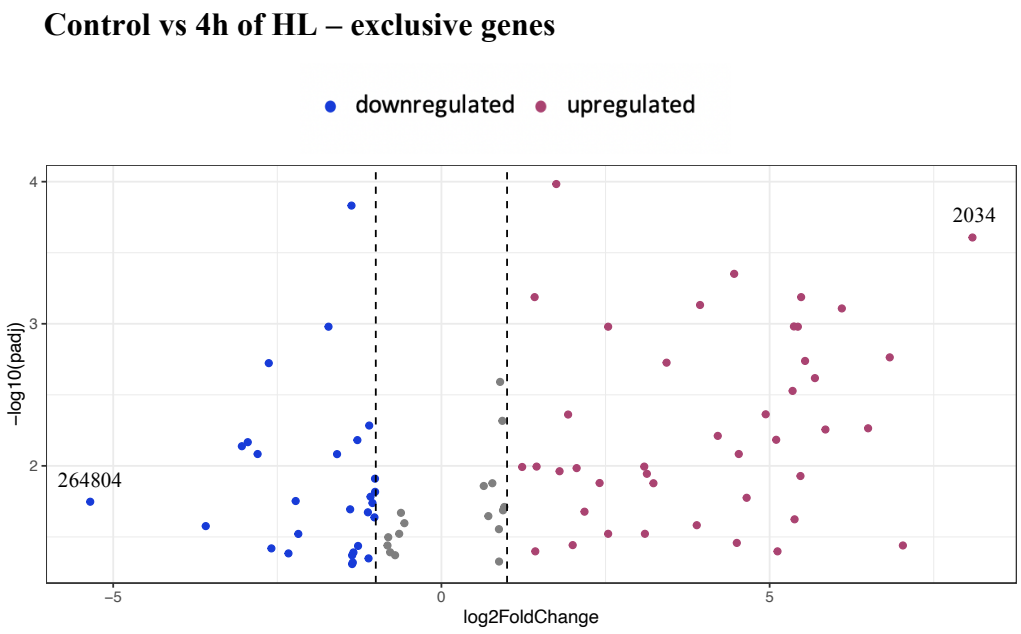


Figure 3.3: Volcano plot of exclusive genes which are upregulated or downregulated under 4 h of HL vs the control condition. Blue dots represent downregulated genes with log2FC < 1. Red dots represent upregulated genes with log2FC > 1. Dash lines indicate log2FC values -1 or +1. The –log10 (adjusted p values) represents the level of significance of each exclusively translationally regulated gene. Genes with an adjusted p value of < 0.05 were assigned as exclusive genes. Genes with the highest up- and downregulation were annotated with ProteinID numbers from JGI.

Interestingly, the majority of genes with the highest ΔTE are poorly annotated in our gene model, predicting proteins of unknow function (Table 3.2). Those results emphasize how little is known about translational regulation upon light stress in *T. pseudonana*. EuKaryotic Orthologous Groups (KOG) analysis was performed to further investigate the potential functions of those genes using orthologous gene products. Hits in the KOG database revealed that 10 out of the 20 genes with the largest upregulated TE are clustered with the functional group *Cellular Processes and Signalling* (not significant). Under light stress, signalling proteins may allow light signals from photoreceptors to activate gene expression which facilitate cellular response mechanisms (Dong et al., 2016).

An exonuclease was found upregulated by 6.1-fold after 4 h of HL treatment (Table 3.2). It is suggested to have similar activity to Fen1, a nuclease involved in the base excision repair (BER) pathway which facilitates DNA repair (Robertson et al., 2009). High UV exposure leads to an increase in reactive oxygen species (ROS) which are by-products of photosynthetic processes (Pospíšil, 2016). The gene may be activated under HL to quickly remove DNA damage that arises due to oxidative stress.

Table 3.2: Genes with exclusive differential translational regulation upon 4 h of high light stress with high TE changes. Adjusted p-value < 0.05. FC, fold change, TE, translation efficiency

Category	Protein ID	Description	KOG group	Log_2FC TE
Upregulated TE	269871	Exonuclease 1		6.10
	9674	Unknown protein	Cellular processes and signalling	6.82
	7145	Unknown protein	Cellular processes and signalling	5.54
	24722	Unknown protein	Cellular processes and signalling	5.37
Downregulated TE	264804	Unknown protein		-5.35
	12179	Unknown protein		-3.59

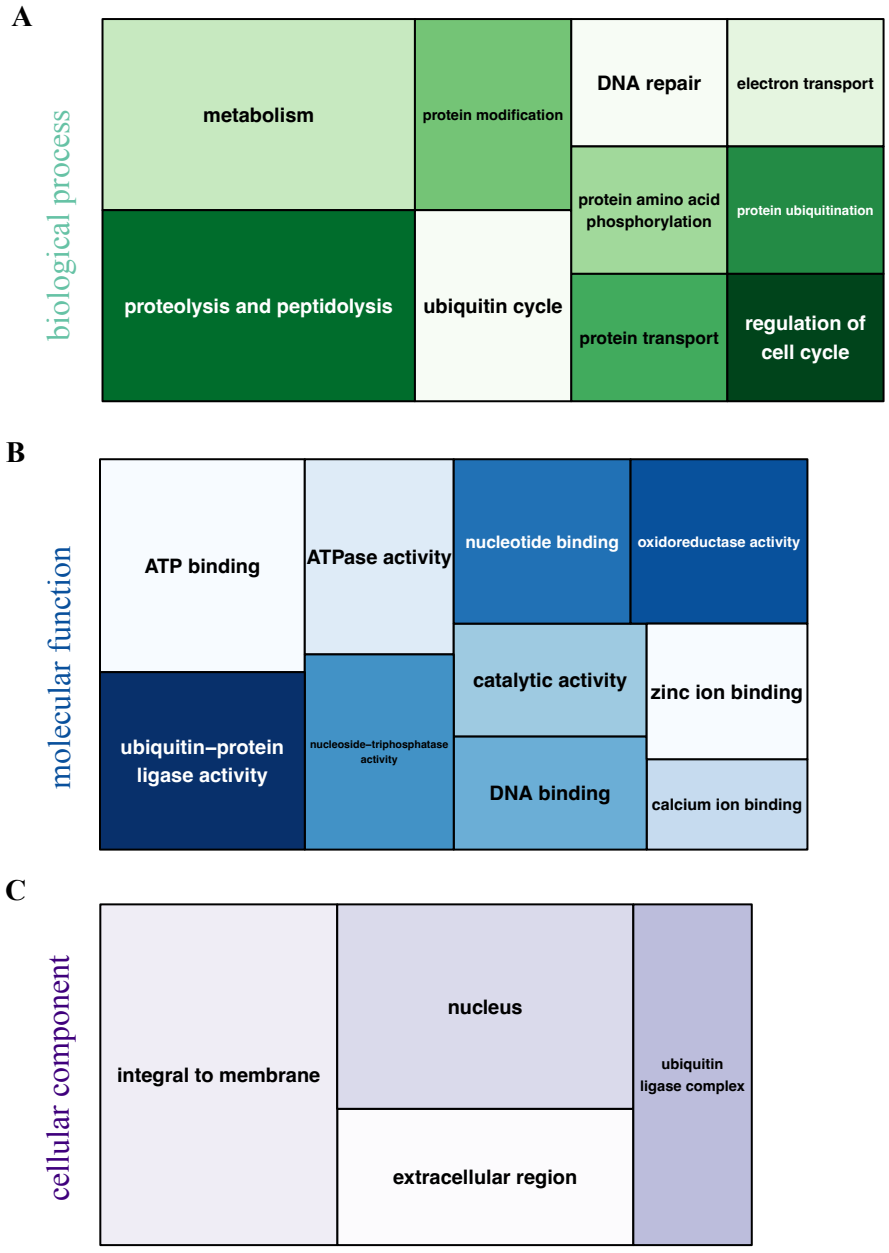


Figure 3.4: Gene Ontology (GO) treemap for exclusively translationally regulated genes after 4 h light stress. GO classification of identified genes in terms of biological process ***(A)*** *molecular function* ***(B)*** *and cellular component* ***(C)****. The top 10 GO terms are shown for each category with at least two counts.*

3.3.4 Detection of differentially transcribed genes and differential translation efficiency genes under 24 h of high light stress

To compare how gene expression patterns of *T. pseudonana* under high light change over time, DTGs and DTEGs were again analysed after 24 h of high light stress. Prolonged light stress led to an increase in differentially expressed genes compared to treatment for 4 h. In total, 1469 DTEGs and 6546 DTGs were detected. Of those, 246 were identified as exclusive (translationally regulated) and 981 and 153 as buffered and intensified (both transcriptionally and translationally regulated), respectively (Figure 3.5). Under prolonged HL treatment, transcriptional regulation is still the major stress response mechanism. Table 3.1 highlights the percentage of DTGs, DTEGs and exclusive genes after 24 h HL stress in relation to the total number of genes in the genome of *T. pseudonana*.

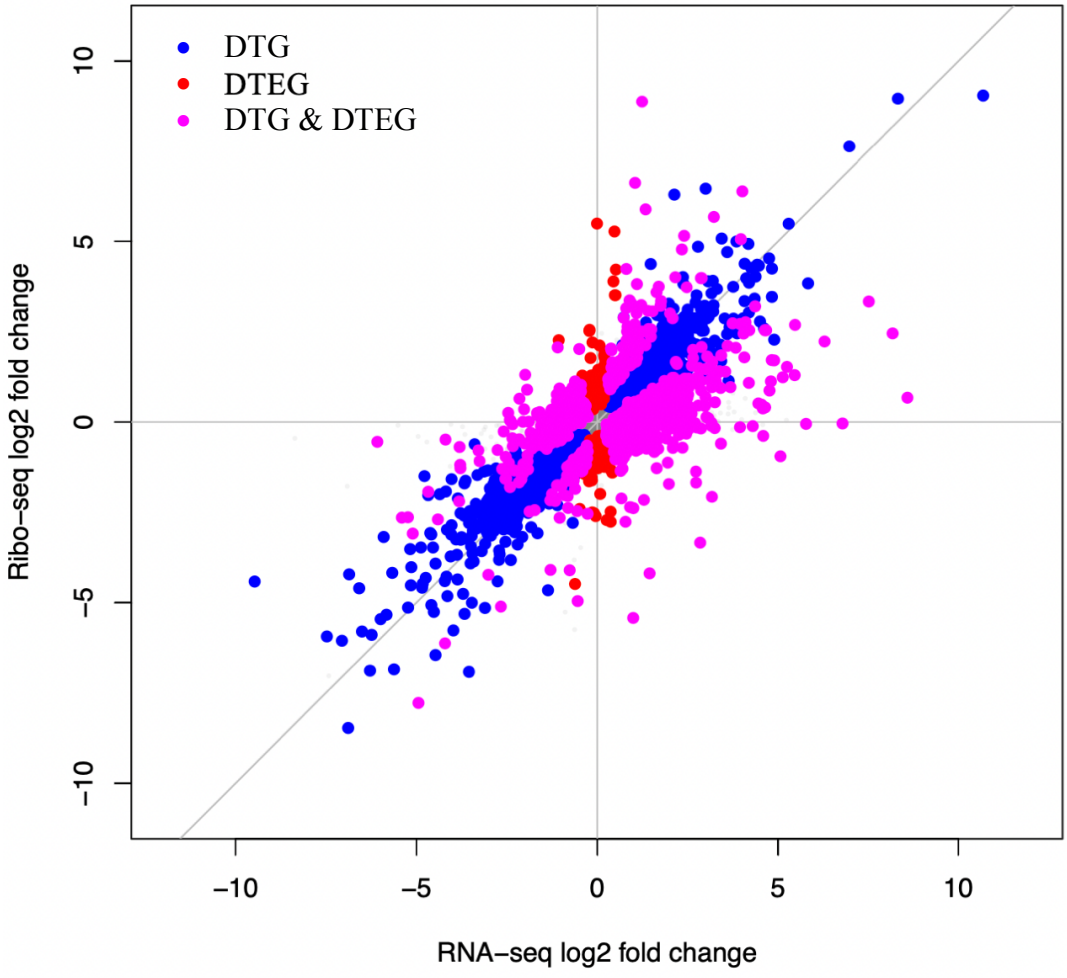


Figure 3.5: Scatter plot of log fold changes for each gene in the ribosome profiling and the RNA-seq data (NL/HL24). Differentially transcribed genes (DTGs, in blue), differentially translation efficiency genes (DTEGs, in red, exclusive) and genes that are both DTG and DTEG (in pink) (intensified and buffered).

3.3.5 Functional analysis of DTEGs under 24 h of light stress

Prolonged high light stress resulted in an extensive reprogramming of genes, including 537 genes with upregulated TE and 932 genes with downregulated TE (Figure 3.6). This finding suggests a significantly higher number of downregulated genes compared to upregulated genes, as determined through binomial distribution testing (p = 2.26E-25).

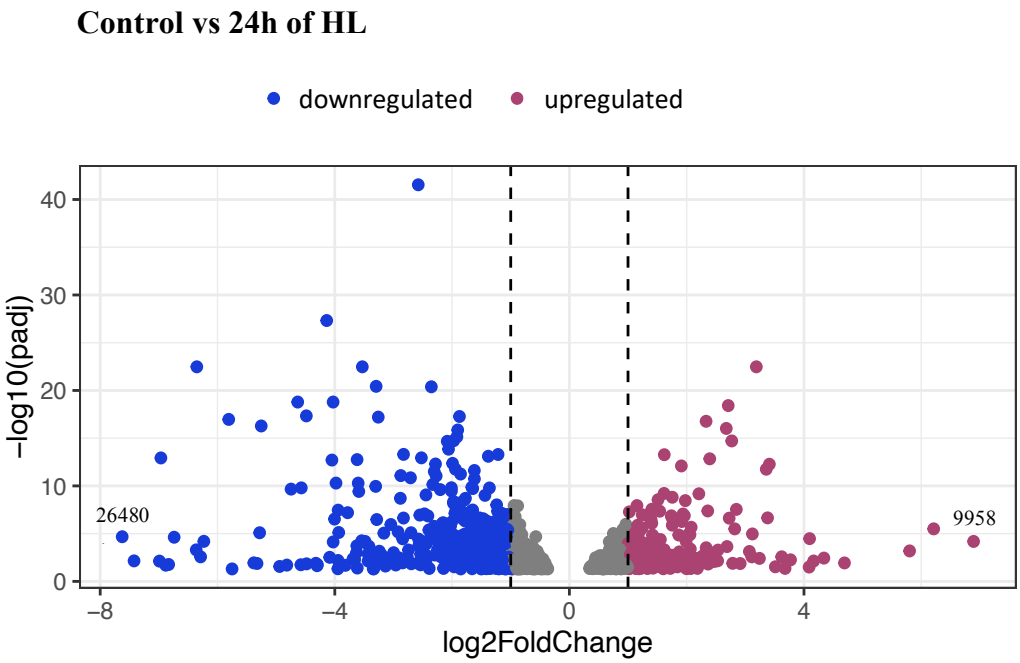


Figure 3.6: Volcano plot of upregulated and downregulated DTEGs of cells under 24 h of HL vs the control group. Blue dots represent downregulated genes with $log_2FC < 1$. Red dots represent upregulated genes with $log_2FC > 1$. Dash lines indicate log2FC values -1 or +1. The –log10 (adjusted p values) represents the level of significance of each DTEG. Genes with an adjusted p value of < 0.05 were assigned as DTEGs. Genes with the highest up- and downregulation were annotated with ProteinID numbers from JGI.

GO term analysis of DTEGs under 24 h of HL stress compared to the control revealed involvement in 87 biological processes, 351 molecular functions and 25 cellular components (Tables 7.5-7.7 of Appendix D). Two of the molecular functions, oxidoreductase activity (GO:0016491) and ligase activity (GO:0016874), were significantly overrepresented ($p_{adj} < 0.05$). Table 3.3 summarizes the genes involved in those two overrepresented GO terms.

Table 3.3: Differential translation efficiency genes identified in T. pseudonana under 24 h of high light stress which are involved in significantly overrepresented GO terms with fold change (FC) < -1.4 and > 1.5. Dashes indicate that gene names or descriptions were missing from the annotation from JGI.

Name	Protein ID	FC
Oxidoreductase activity		
ferric reductase	260785	-6,83
nitrate reductase	25299	-4,48
-	264757	-3,35
putative pyruvate formate-lyase activating enzyme	263834	-3,34
-	7359	-3,29
-	264753	-2,54
putative pyruvate formate-lyase activating enzyme	263830	-2,50
-	23654	-2,27
protoporphyrinogen IX oxidase (Ppx1)	264901	-1,60
-	20953	-1,55
putative 3-oxoacyl-[acyl-carrier protein] reductase (KAR1)	268493	2,18
-	34283	2,06
-	2066	1,72
-	20622	1,53
Ligase activity		
Pyruvate carboxylase	11076	-1,95
Acetyl-CoA carboxylase	12234	-1,93
Carbamyl phosphate synthetase III	24248	-1,91
Pyruvate carboxylase	11075	-1,48
biotin carboxylase (PCB1)	269328	-1,40

After 24 h of HL stress, the expression of 5 FCPs were upregulated by 1.2- to 2.6-fold suggesting their role in photoprotection in *T. pseudonana.* Among them, the Lhcr5 gene is still transcriptionally and translationally upregulated, with a 1.91-fold change in TE, however the change in TE has decreased after prolonged stress. This suggests that the gene is mainly involved in initial and intermediate photoprotective response. This is in accordance with other studies showing transcript of FCP genes Lhcx1, Lhcx4 and Lhcx6 achieved the highest levels after 1-3 h of exposure to HL and decreased again afterwards (Zhu and Green, 2010; Dong et al., 2016).

3.3.6 Exclusively translationally regulated genes under 24 h of light stress

Significant translational response for 246 genes was detected, of which 125 and 121 were upregulated and downregulated, respectively, after 24 h of high light treatment. Figure 3.7 shows all exclusively downregulated and upregulated genes with log2 fold change of -1 or +1 under 24h of HL vs control conditions. GO term and KOG analysis was performed to explore the potential functions of genes which are translationally driven. GO analysis identified gene involvement in several categories (Figure 3.8A-C), however none of them were significant.

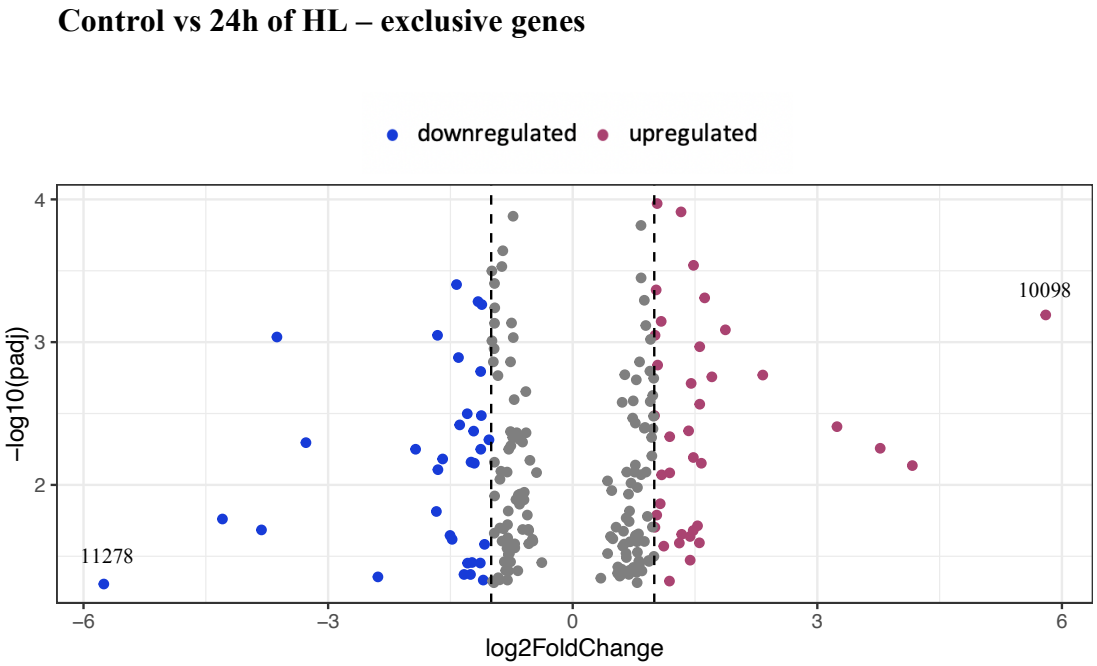


Figure 3.7: Volcano plot of exclusive genes which are upregulated or downregulated under 24 h of HL vs the control condition. Blue dots represent downregulated genes with log2FC < 1. Red dots represent upregulated genes with log2FC > 1. Dash lines indicate log2FC values -1 or +1. The –log10 (adjusted p values) represents the level of significance of each exclusively translationally regulated gene. Genes with an adjusted p value of < 0.05 were assigned as exclusive genes. Genes with the highest up- and downregulation were annotated with ProteinID numbers from JGI.

Ten genes with downregulated TE were found in the functional KOG class of *translation, ribosomal structure and biogenesis* (not significant) (Table 3.4). Interestingly, three of these genes encode ribosomal proteins, RPS7, RPS13 and RPL30, which were downregulated by 0.82 to 1.12-fold. Proteomic analysis showed that 21 of ribosomal proteins of *T. pseudonana* were significantly upregulated after 10 h of HL stress (Dong et al., 2016). In our study, none of those ribosomal genes showed significant ΔTE. RPS7, RPS13 and RPL30 were not significantly expressed in the study by Dong after 10 h of HL. This agrees with our results

which did not show significant expression at 4 h, suggesting those genes might only be involved in translational regulation upon prolonged HL stress.

Among the genes with the largest changes in TE between control and 24 h high light condition are several cyclin genes (Table 3.4). Studies have shown that elevated transcription levels of cyclins increase the ability of plants to defend against stress by enhancing immune response, leading to prolonged cell cycle progression or programmed cell death (Qi and Zhang, 2019). Diatoms possess a large number of cyclin genes. In addition, diatom-specific cyclins were discovered in *T. pseudonana* and *P. tricornutum,* which are predominantly expressed at the G1-to-S transition and are hinted to play regulatory roles in fluctuating environmental conditions (Huysman et al., 2010)

Ppx1 is a protoporphyrinogen IX oxidase involved in chlorophyll biosynthesis, which is downregulated by 1.59-fold under prolonged HL stress (Table 3.4). Downregulation of the gene under abiotic stress conditions leads to impaired Chl biosynthesis in plants (Dalal and Tripathy, 2012).

The TE of a gene belonging to the photolyase family was upregulated by 1.47-fold (Table 3.4). Photolyases are a class of flavoproteins that catalyse the repair of UV-induced DNA damage (Sancar, 2003). Photolyase enzymes are present in twelve species of diatoms from Antarctica, demonstrating their involvement in response mechanisms to minimize UV damage in high light environments (Karentz et al., 1991).

Table 3.4: Genes with exclusive differential translational regulation upon 24 h of high light stress discussed in this chapter. Adjusted p-value < 0.05. FC, fold change, TE, translation efficiency

Category	Protein ID	Description	Log_2FC TE
Upregulated TE	10098	G1/S-specific cyclin E	5,80
	11267	G1/S-specific cyclin D	2,84
	35005	Photolyase	1.47
Downregulated TE	39550	RPS7	-1.12
	29217	RPL30	-0.96
	26221	RPS13	-0.82
	264901	Ppx1	-1.59

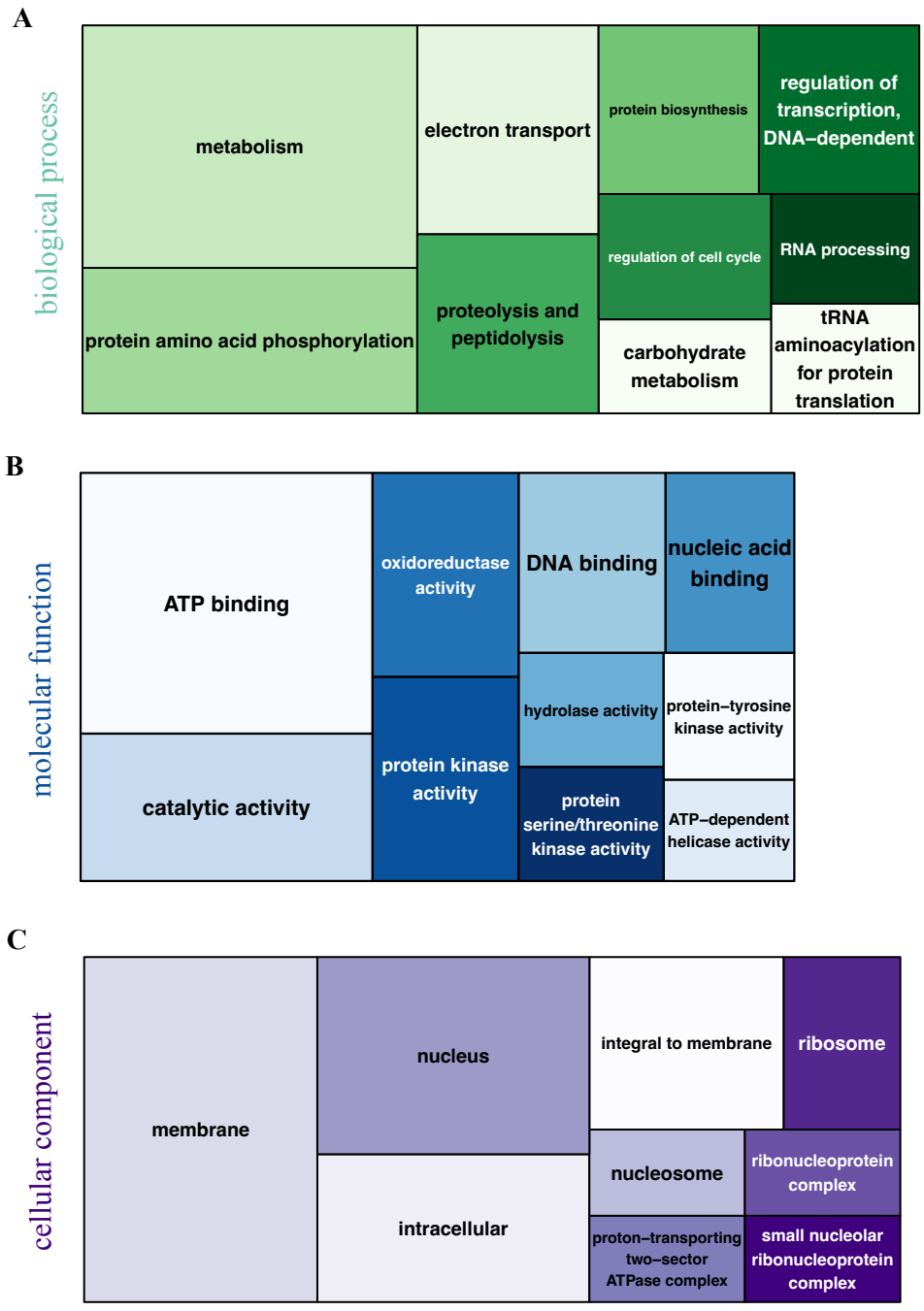


Figure 3.8: Gene Ontology (GO) treemap for exclusively translationally regulated genes after 24 h light stress. GO classification of identified genes in terms of biological process ***(A)*** *molecular function* ***(B)*** *and cellular component* ***(C)****. The top 10 GO terms are shown for each category with at least two counts.*

3.3.7 Short-term and prolonged high light stress trigger distinct translatomic responses

To investigate if similar sets of genes are involved in translationally regulated response mechanisms after 4 h and 24 h of high light treatment, we compared all genes with changes in TE. Among all DTEGs, 230 genes (13.5%) are differentially regulated at both short-term and prolonged HL stress (Figure 3.9A). After 24 h of HL, we found 36 DTEGs involved in the biological process of *electron transport* (GO:0006118), which is an increase compared to the 12 which were involved at 4 h of HL treatment. However, only three of those genes are shared between the two conditions.

Translational regulation of LHC genes in response to light has been shown in several plants (Frigerio et al., 2007; Floris et al., 2013) and the green algae *C. reinhardtii* (Mussgnug et al., 2005; McKim and Durnford, 2006). Both FCPs which were translationally regulated in our dataset after 4 h of HL were still differentially expressed at 24 h but showed a decrease in TE. This indicates a photoprotective function for these gene products which is most pronounced after short-term HL treatment. In total, five FCPs were upregulated upon 24 h HL treatment, suggesting that a different set of light harvesting proteins is involved in response to prolonged stress.

Prolonged HL stress resulted in an increase of exclusive genes involved in *protein amino acid phosphorylation* (GO:0006468) from two detected after 4 h exposure to nine after 24 h. The ΔTE of the one gene (Protein ID 38513) which was expressed at both time points stayed constant (1.41/1.63-fold). Interestingly, the highest upregulated TE (5.69-fold) was detected for a gene (Protein ID 3330) which was only expressed after short-term treatment. Phosphorylation is a rapid and transient mechanism and is thus commonly used by cells to post-translationally regulate protein function upon stress (Withers and Dong, 2017; Soma et al., 2021).

Differentially expressed genes related to oxidoreductase activity are associated with stress tolerance in photosynthetic organisms (Rezayian et al., 2019; Zhang et al., 2020). We identified 4 exclusively regulated genes involved in *oxidoreductase activity* (GO:0016491) after 4 h of HL and 9 genes after 24 h.

Comparison of all exclusively translationally regulated genes between the two time points revealed that only 13 genes (4%) were shared (Figure 3.9B). Of those, 10 were upregulated and 3 were downregulated. Interpretation of this set of genes is difficult as most are annotated as

unknown proteins. Nevertheless, the low number of genes shared indicate two distinct translationally regulated responses upon short-term and prolonged HL stress in *T. pseudonana*.

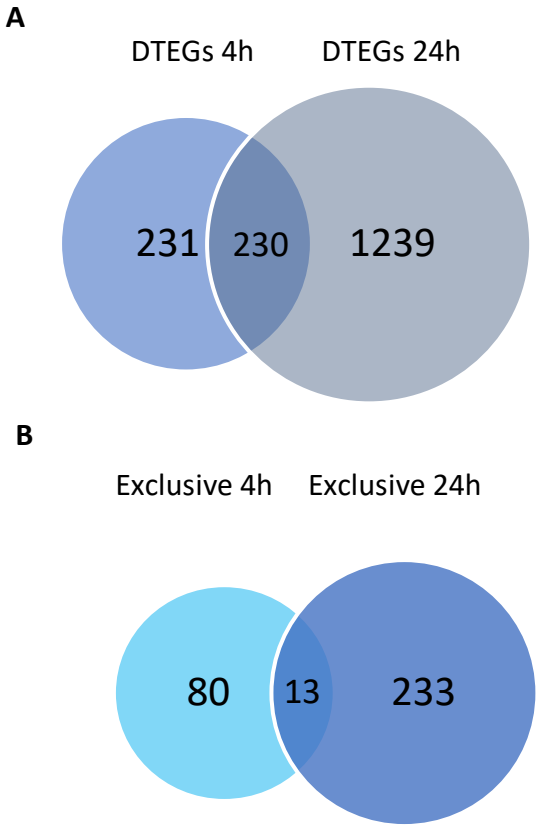


*Figure 3.9: Venn diagram of genes with differential TE upon high light stress. **A** The overlapping circles show the number of shared DTEGs between 4 h and 24 h of HL stress. **B** The overlapping circles show the number of exclusively translationally regulated genes which are shared after 4 h and 24 h of HL stress.*

3.4 Conclusion

This is the first study to reveal the landscape of translational regulation in response to high light stress in diatoms. By combining ribosome profiling and RNA-seq data, we could determine genes exhibiting a change in TE, which means they are translationally regulated. For the majority of genes, changes in RPFs and mRNA levels match, indicating that light stress induced gene expression is primarily regulated transcriptionally. However, our data suggest several mechanisms at the translational level in *T. pseudonana* which regulate the balance between cell growth and photoprotection during HL stress. These include an increase in FCP expression upon 24 h of light exposure. We were able to detect several exclusively translationally regulated genes upon short and prolonged HL stress whose roles in diatom light stress response still need to be analysed. Our data gives a first overview of genes involved in regulation at the translational level upon light stress on a genome-wide level. However, a more in-depth analysis of their molecular functions and involved pathways is needed to decipher the translatomic response in *T. pseudonana*.

Chapter 4: Modifying the codon usage of *Thalassiosira pseudonana* via CRISPR/Cas9-mediated homologous recombination

4.1 Introduction

Considering the ecological relevance and biotechnological potential of diatoms, rapid developments in gene editing tools are of global interest. These tools have already improved our understanding of key cellular processes in diatoms and have enabled the production of desired compounds via redesigning of metabolic pathways (Huang and Daboussi, 2017). CRISPR/Cas methods have dramatically increased our understanding of gene functions and have led to new paradigms in gene expression (Akinci et al., 2021). However, regulation of gene expression through codon usage bias (CUB) has not been studied well in diatoms. Due to the degeneracy of the genetic code, the 20 amino acids that occur in proteins are encoded by 61 different codons. Besides methionine and tryptophan, all amino acids are encoded by two to six synonymous codons, some of which are used more frequently than others. This phenomenon is termed CUB, and it is shown to affect gene expression and cellular function by influencing translation machinery processes such as RNA processing, protein translation and folding, as well as mRNA stability and translation elongation (Plotkin and Kudla, 2011; Hanson and Coller, 2018). In contrast to the central dogma in molecular biology, which suggests that synonymous mutations are assumed to be neutral because they do not have an effect on protein synthesis (Plotkin and Kudla, 2011), synonymous mutations can have phenotypic consequences (Peng et al., 2018). Whether a codon is defined as optimal or sub-optimal depends on how efficiently the cognate tRNA can be selected from the tRNA pool (Hanson and Coller, 2018). Codon bias has been correlated with tRNA abundance in several prokaryotic and eukaryotic organisms (Ikemura, 1981; Duret, 2000; dos Reis et al., 2004) suggesting that tRNA abundance is the selective force determining synonymous codon usage (Ikemura, 1981, 1982; Lynn et al., 2002). Synonymous codon usage can directly modulate the efficiency and accuracy of protein synthesis (Gingold and Pilpel, 2011). Optimal codons are thought to be translated both faster and more accurately, while sub-optimal codons can slow down elongation rate, leading to ribosome stalling and reduced protein synthesis (Pechmann and Frydman, 2013; Schuller and Green, 2018). Studies have shown that CUB towards preferred codons is more pronounced in highly expressed genes than in genes with lower expression rates (Ikemura, 1981; Shields and Sharp, 1987). This suggests that the codon usage of genes with elevated expression levels has been optimized through evolution to enhance translation efficiency. A sub-optimal codon usage

in genes with lower expression levels might however be a strategy to avoid competition for efficient translation with highly expressed genes (Hiraoka et al., 2009). Despite extensive research on codon-mediated regulation, its exact impact on regulating translation remains elusive.

Each species shows a distinct codon usage and the extent of this CUB varies (Hershberg and Petrov, 2008; Krasovec and Filatov, 2019). A metagenomic study revealed that microbial organisms living in the same ecological niche share a common preference for synonymous codons (Roller et al., 2013). The similarity of tRNA pools of organisms living in these communities also increases the chance of successful horizontal gene transfer (Tuller et al., 2011). Further, microbes found in a broad range of habitats tend to have lower CUB than organisms living in specialized environments, indicating the role of codon usage on the ability to adapt to different environments (Botzman and Margalit, 2011). A study analysing the relationship between CUB, phenotypic traits (lifestyle) and habitats in 615 microbial organisms confirmed these results and showed that CUB and tRNA pools are under weaker selective pressure in species inhabiting multiple environments (Arella et al., 2021).

The first genome-wide analysis of CUB in diatoms showed that most species have modest CUB, which is relatively surprising given their large population sizes (Krasovec and Filatov, 2019). It is argued that their effective population size is large enough for selection for optimal codons to overpower genetic drift. Therefore, Krasovec and Filatov (2019) rather suggest that frequent changes of preferred codons lead to the low CUB in diatoms. Shifts in the set of preferred codons, such as from GC-rich to AT-rich codons seen in *Chaetoceros* species, constantly change the direction of selection for codon usage. This never allows codon usage to catch up with the optimal set of codons, thus never resulting in a strong CUB.

The aim of this study was to modify the codon usage of two genes (Lhcx6, RPL10a) in the diatom *T. pseudonana* via CRISPR/Cas-mediated homologous recombination. The Lhcx6 gene (JGI protein ID 12097) is part of the light harvesting complex (LHC) superfamily and plays a role in light harvesting and photoprotection via non-photochemical quenching (NPQ) (Zhu and Green, 2010). Lhcx6 is only moderately expressed in *T. pseudonana* and shows a slightly non-optimal synonymous codon usage. The Lhcx6 WT gene will be replaced with a codon-optimized gene. The second gene of interest is RPL10a (JGI protein ID 23025), a highly expressed gene in *T. pseudonana* which encodes a ribosomal protein that is part of the large

ribosomal subunit (60S). It has a key role in ribosome assembly and is essential for the translation process (Uniprot.org). RPL10a is under strong translational selection showing a codon usage which has been optimized through evolution to enhance translation efficiency. We tried to target the RPL10a WT gene and replace it with a codon sub-optimized gene. In this chapter, the construction of plasmids used for biolistic transformations of *T. pseudonana* and subsequent genotyping and phenotyping of HR cell lines is described. Our study provides preliminary insights for future study of gene regulation via codon usage in diatoms.

4.2 Material and methods

4.2.1 Diatom strain and growth conditions

T. pseudonana (strain CCMP1335) used for biolistic transformation was grown to exponential phase (~1 x 10^6) in ½ salinity Aquil media (pH~8) under 24 hours light (ca. 80 µmol photons m^{-2} s^{-1}) at 20°C. Cells used for protein extraction and growth rate measurement were kept under low light conditions (50 µmol photons m^{-2} s^{-1}) prior to high light exposure (500 µmol photons m^{-2} s^{-1}).

4.2.2 Synthesis of codon modified genes

The following part was previously performed by Amanda Hopes. Two genes with different expression levels were selected by according to a RNA expression dataset published by Lopez-Gomollon et al. (2014): the ribosomal RPL10a gene displaying high expression levels and the light-harvesting complex associated Lhcx6 gene which is moderately expressed. Codon usage for *T. pseudonana* was analysed using the general codon usage analyser (GCUA) software (http://mcinerneylab.com) (McInerney, 1998). Relative synonymous codon usage (RSCU) values were calculated for the *T. pseudonana* reference dataset. RSCU values indicate how often a particular codon is used relative to the expected number of times that codon would be used in the absence of any codon usage bias (McInerney, 1998). The highly expressed RPL10a gene showed an optimized codon usage while the codon usage of the moderately expressed Lhcx6 gene was sub-optimal. The web application OPTIMIZER (Puigbo et al., 2007) used the mean codon usage of *T. pseudonana* as a reference set to either sub-optimize or optimize the codon usage of genes RPL10a and Lhcx6, respectively (see Figures 7.7-7.8 of Appendix E). The 'one amino acid – one codon' method was applied, which replaces all codons encoding for the same amino acid with the synonymous codon used most frequently in the reference table or

uses an inverse reference table to generate a less optimized sequence (Puigbo et al., 2007). Codon-modified genes for integration via homologous recombination were *de novo* synthesised using GENEWIZ services (http://genewiz.com).

4.2.3 Designing sgRNAs

sgRNAs were designed according to Hopes et al. (2017). This method uses two sgRNAs per construct to increase the probability of target site cutting. The protocol uses the chimeric sgRNA with the following sequence: NNNNNNNNNNNNNNNNNNGTTTTAGAGCTAG AAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCG AGTCGGTGCTTTTTT, where the underlined sequence represents the 20 nt target region. For both genes of interest, sgRNA targets were designed using the RGEN Cas-Designer and the Broad Institute sgRNA designer (Table 4.1). Lhcx6 sgRNAs and sgRNA_RPL10a_AH were previously designed by Amanda Hopes.

Table 4.1: Overview of sgRNAs designed and tested for cleavage efficiencies. Each sgRNA is followed by a specific PAM sequence. High out-of-frame scores (0-100) are desired to avoid unwanted in frame deletions (RGEN Cas-Designer).

sgRNA ID	Target (5' to 3')	Position	Out-of-frame-score	PAM sequence
sgRNA_Lhcx6_1	GAAACGTGCTAATATTGGAT	453	74.0	GGG
sgRNA_Lhcx6_2	GCAGGTCCGGTAATTCCAAA	358	50.8	TGG
sgRNA_RPL10a_AH	GACCGGGTCCCAAAAGACGG	396	80.7	GGG
sgRNA_RPL10a_1	GAGAATCGCTTGTCACGCTG	163	66.2	AGG
sgRNA_RPL10a_2	CGTAGCTCACATGAGTACCG	279	62.4	AGG
sgRNA_RPL10a_3	TAAGACCGGGTCCCAAAAGA	399	68.5	CGG
sgRNA_RPL10a_4	ATGTTGGGACGGGGGATGGC	202	67.9	AGG
sgRNA_RPL10a_5	AACATCAAGTGCTGCATGCT	220	78.8	CGG
sgRNA_RPL10a_6	CTCCGCCAGCAAGAAGGGTG	438	60.2	GGG

4.2.4 Testing cleavage efficiency

A Cas9 *in vitro* cleavage assay developed by OmicronCr (www.omicroncr.co.uk) was used to evaluate the efficiency of all designed sgRNAs. Nine sgRNAs to target genes RPL10a and Lhcx6 were *in vitro* transcribed. A detailed description of the OmicronCr method cannot be provided here. ImageJ (Schneider et al., 2012) was used to calculate cutting efficiency for each sgRNA.

4.2.5 Design of plasmids for co-transformations

Our co-transformation approach uses two plasmids to introduce all required components for CRISPR-mediated HR as outlined in Belshaw et al. (2022). The CRISPR plasmid expresses a cassette for nourseothricin resistance, Cas9 and two sgRNAs to target the gene of interest and to introduce double-strand breaks (DSBs). The homologous recombination donor (HR) plasmid contains the codon modified RPL10a or Lhcx6 gene, flanked by two regions which are homologous to the non-coding 5' and 3' ends of the wild-type genes. The constructs were assembled using the Golden Gate cloning method (Weber et al., 2011; Belhaj et al., 2013). For the assembly of the individual modules, the protocol from Hopes et al., (2017) was followed, which also describes available modules from the Addgene database.

4.2.6 CRISPR plasmid assembly

sgRNAs were assembled directly into L1 vectors as a PCR product: the forward primer 1-5 (Table 4.2) includes the target region and amplifies together with the reverse primer 6 (Table 4.2) the scaffold from pICH86966::AtU6p::sgRNA_PDS. PCR for each sgRNA was done with Phusion DNA polymerase and 55°C annealing temperature. PCR products were purified using the Monarch PCR&DNA clean-up kit (NEB) and run on a 1% agarose gel. The L0 module containing the U6 promoter and the purified PCR product (sgRNA1 or sgRNA2) was assembled into the L1 destination vector pICH47751 and pICH47761, respectively. For L1 assembly, 40 fmol of each component was added to a total reaction volume of 20 µl with 10U of BsaI (10.000U/ml) and 10U of T4 DNA ligase (Promega, 10U/µl). Reactions were incubated for 5 hours at 37°C, 5 min at 50°C and 10 min at 80°C, and five µl of each reaction was transformed into 50 µl of high efficiency NEB 5-alpha chemically competent *E. coli* following the NEB protocol. L1 modules contain the cassette for carbenicillin resistance for selection in *E. coli*. Correct insertion removed the LacZ gene and blue/white colony screening was done. Colonies from each L1 assembly were picked and DNA was extracted following the QIAprep Spin Miniprep Kit (Qiagen). 500 ng of plasmid was digested with 20 units of XbaI for 15 min at 37°C, followed by 20 min of heat inactivation at 65°C. Linearized products were run on a 1% agarose gel. To confirm the correct insertion of the U6 promoter and the sgRNA into the L1 vector, plasmids were sequenced using reverse primer 7 (Table 4.2). L1 modules pICH47732_TpFCP:NAT, pICH47742_TpFCP:Cas9:YFP, pICH47751_TpU6:sgRNA1, pICH47761_TpU6:sgRNA2 and the L4E linker pICH41780 were assembled into the L2 backbone pAGM4723. Ligation was performed as described above for L1 assembly but with 10U of BpiI restriction enzyme. Five µl of the reaction was used for transformation in high

efficiency NEB 5-alpha chemically competent *E. coli.* L2 modules contain the cassette for kanamycin resistance for selection in *E. coli*. Correct insertion removed the gene for canthaxanthin and allowed for pink/white screening. Colonies were picked from selective agar plates and DNA was extracted following the QIAprep Spin Miniprep Kit (Qiagen). 500 ng of plasmid was used in a digest with 20 units of double cutter EcoRV-HF for 15 min at 37°C, followed by 20 min of heat inactivation at 65°C. Products were run on a 1% agarose gel. Constructs displaying the correct band patters were further sequenced (Eurofins Genomics) using primers 7, 8 and 9 (Table 4.2) to confirm the correct insertion of all modules. Prior to biolistic transformations, plasmids were sodium-acetate ethanol precipitated and eluted in nuclease free water. Assembly of the Lhcx6 CRISPR plasmid was previously done by Amanda Hopes.

Table 4.2: Oligonucleotides used for cloning and screening of CRISPR and HR plasmids for biolistic co-transformations. BsaI/BpiI restriction sites are underlined, overhangs are in bold and sgRNA targets in italic.

Primer	Sequence	No.
sgRNA_RPL10a_AH_F	aggtctca***ttgt****GACCGGGTCCCAAAAGACGG*GTTTTAGAGCTAGAAATAGCAAG	1
sgRNA_RPL10a_1_F	aggtctca***ttgt****GAGAATCGCTTGTCACGCTG*GTTTTAGAGCTAGAAATAGCAAG	2
sgRNA_RPLL10a_2_F	aggtctca***ttgtG****CGTAGCTCACATGAGTACCG*GTTTTAGAGCTAGAAATAGCAAG	3
sgRNA_Lhcx6_1_F	aggtctca***ttgt****GAAACGTGCTAATATTGGAT*GTTTTAGAGCTAGAAATAGCAAG	4
sgRNA_Lhcx6_2_F	aggtctca***ttgt****GCAGGTCCGGTAATTCCAAA*GTTTTAGAGCTAGAAATAGCAAG	5
sgRNA_R	tggtctca***agcg***TAATGCCAACTTTGTACAAG	6
piCH_L1_R	GCCAATATATCCTGTCAAACAC	7
GG_Cas9: YFP_F	ATGGACAAGAAGTACTCCATTGG	8
NAT_F	ATGACCACTCTTGACGACAC	9
RPL10a_UFlank_F	AAAGAAGACCA**tgcc**AGTCAGAAGTGCGTGAGTTTC	10
RPL10a_coding_F	ATTGAAGACCAATGTCGAATAAATTAAATTCGG	11
Lhcx6_UFlank_F	AAAGAAGACCA**tgcc**ACCAGGCTGATCGTAAGAAG	12
Lhcx6_coding_F	ATTGAAGACCAATGAAGTTTACCCTTTTGTCC	13

4.2.7 HR plasmid assembly

The 5' and 3' flanking regions for RPL10a (785bp and 639bp) and Lhcx6 (813bp and 743bp) were amplified from genomic DNA using primers 10 and 12 (Table 4.2) to introduce overhang sequences and BpiI restriction sites for directional cloning following the method outlined in Hopes et al., (2017). PCR products were assembled into the L2 backbone pAGM4723 as described above for CRISPR plasmids. This part was previously carried out by Amanda Hopes. I then digested the RPL10a plasmid DNA with BsaI (10 units) and BamH-HF (20 units) restriction enzymes and Lhcx6 plasmid DNA with EcoRV (20 units) for 1 h at 37°C. Banding patterns were analysed on an agarose gel. To verify the correct insertion, plasmids were sequenced using primers 7 and 10-13 (Table 4.2).

4.2.8 Biolistic co-transformations

Co-transformations by biolistic bombardment, using both a CRISPR plasmid and a HR donor plasmid, were carried out according to Hopes et al., (2017). For each shot, 5 x 10^7 cells were filtered onto 47 mm, 1.2 µm Isopore filters (Millipore) using vacuum filtration (<150 mbar Hg) and placed onto agar plate with 1.5% agar and ½ salinity Aquil media. Five µl of each plasmid (CRISPR and HR) were coated with 30 µl of prepared tungsten particles, 30 µl of 2.5 M $CaCl_2$ and 12 µl of 0.1 M spermidine to prepare three replicates per construct. A flight distance of 7 cm, 1.350 psi rupture discs and a vacuum of 25 hg were used. Following transformations, filters were placed directly into 25 ml of ½ salinity Aquil and incubated under standard growth conditions for 24 h. Then, 5 x 10^6 cells were spread onto each selective agar plate (0.8% agar, ½ salinity Aquil, 100 µg/ml nourseothricin). Colonies appeared after approximately 14 days and were re-streaked onto fresh selective plates. These secondary clones were resuspended in selective liquid media and cells were used for colony PCR and to grow up liquid cultures in 96-well plates.

4.2.9 Screening of transformants

Colony lysates were used as template in PCR reactions using MyTaq HS MM Red (Bioline) unless stated otherwise and several primer pairs (Table 4.3 and Figure 4.1) as described in Belshaw et al., (2022). First, integration of the Cas9 gene was screened using primers 38 and 39. In a nested PCR approach using LongAmp Taq 2x Master Mix (NEB), primers 14-15 and 26-27 were binding to a sequence outside the flanking regions to amplify the entire region of interest. To confirm the presence of both the 5' and the 3' flanking regions of the modified

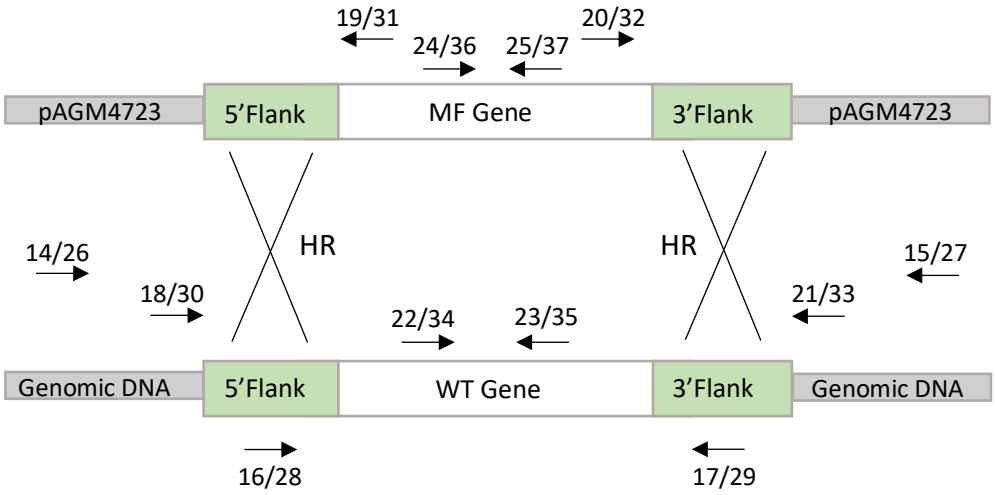


Figure 4.1: Overview of HR repair mechanisms after Cas9-induced double-strand break between donor plasmid containing the modified gene and the genomic wild-type gene. Arrows depict primers used for screening of transformants with numbers corresponding to table 4.4. Adapted from Belshaw et al., (2022).

genes, primers 18-21 and 30-33 targeted the codon modified gene and the 5' and 3' flanking region, respectively using products of the previous PCR as template. Clones, which showed bands were further screened for the presence/absence of the wild-type gene by using primers 22-23 and 34-35. If amplification resulted in bands, possible mono-allelic transformants were identified. Clones showing no WT amplification were further used to amplify a section of the codon modified gene using primers 24-25 and 36-37. Bands from the final amplification are a strong indication for HR having occurred in both alleles. Genomic DNA of potential bi-allelic knock-ins was extracted using MagAttract HMW DNA kit (Qiagen) following manufactures instructions but doubling all reagents. Primer pairs 14-15 and 26-27 binding to regions outside of the 5' and 3' regions were used to amplify the RPLL10a and Lhcx6 locus using LongAmp Taq 2x Master Mix (NEB). The products were sent off for Sanger sequencing (Eurofins) to confirm bi-allelic replacement of the genes by HR using sequencing primers (16-17 and 28-29).

Table 4.3: Oligonucleotides used for screening of transformants.

Name	Sequence 5'-3'	No.
Outer_Primer_Lhcx6_F	GGCTACAGGGTGATACCAAATG	14
Outer_Primer_Lhcx6_R	CGCCCCTACTCCGTGTTTAG	15
Inner_Primer_Lhcx6_F	ATCACACCGCCAGCACC	16
Inner_Primer_Lhcx6_R	TCGTTTGGGGCAGTACACAT	17
5'_Lhcx6_F	CGTTCTGAAGCTCCTTGGTAATC	18
5'_Lhcx6_R	GCGGTAGATGCGATAGTAGAAGGG	19
3'_Lhcx6_F	CAGACAAGAGAGCTACAGAAC	20
3'_Lhcx6_R	CACCGTGAGAGATGAGAATCTG	21
WT_Lhcx6_F	CTGCCTTCGTCGCTCCTTCA	22
WT_Lhcx6_R	CCGCCCATTCTGCAATTCACG	23
MF_Lhcx6_F	GAGCCCTTCTACTATCGCATCT	24
MF_Lhcx6_R	ACGACCGTTCTGTAGCTCTCT	25
Outer_Primer_RPL10a_F	CCATCAAGAGGTTTCGGCTAAAG	26
Outer_Primer_RPL10a_R	TCCCTCTCTTTCCGTTGATTG	27
Inner_Primer_RPL10a_F	TTGACCCCCTTCTAACCCGA	28
Inner_Primer_RPL10a_R	TCTTCAATCCCAAGCCTGCC	29
5'_RPL10a_F	TGGGGGAGACTGTGAAAACG	30
5'_RPL10a_R	GTCCTCTACCGCTTTGTCTAATAAC	31
3'_RPL10a_F	TGTGCTTAAATGTAGCGATAGGG	32
3'_RPL10a_R	ACCCATCCAACCCTCATTTG	33
WT_RPL10a_F	GTCTTTTGGGACCCGGTCTT	34
WT_RPL10a_R	AGCAGACAACTGAGTGTTGAC	35
MF_RPL10a_F	GGCGTTATTAGACAAAGCGGTA	36
MF_RPL10a_R	TAACCCCGGCCCTAATAAC	37
Cas9_F	CCGAGACAAGCAGAGTGGAAAG	38
Cas9_R	AGAGCCGATTGATGTCCAGTTC	39

4.2.10 Single cell sorting

To get a monoclonal culture, single cells were isolated from sample Lhcx6_MF_85 using the BD FACSMelody cell sorter at the John Innes Centre (Norwich, UK). This automated instrument sorted single cells directly into 96-well plates containing 100 µl ½ salinity Aquil media. Growth was visible after three weeks and cells were grown up in larger volumes for PCR screening and Sanger sequencing as described above.

4.2.11 Verify genome editing via Oxford Nanopore sequencing and Illumina sequencing

High molecular weight genomic DNA of the Lhcx6_MF_85 mutant cell lines was extracted using the MagAttract kit as described above. Up to 400 ng of gDNA was prepared with the Rapid Sequencing Library kit (Oxford Nanopore SQK-RAD004) and sequenced with a FLO-MIN106 flow cell on the MinION (Oxford Nanopore). Base calling was performed using Guppy version 3.1.5. Sequencing data was analysed on the NanoGalaxy web platform (de Koning et al., 2020) using Porechop (version 0.2.4) to trim adapters. Reads were mapped to the *T. pseudonana* reference genome using minimap2 (version 2.24) and visualized using the integrative genomics viewer (IGV) (Robinson et al., 2011).

For Illumina sequencing, extracted gDNA (samples Lhcx6_MF85s_opt, _1.1, _1.3) was diluted to 20-50 ng/µl in 20µl low TE buffer (10 mM Tris, pH 8, 0.1mM EDTA) and library construction was done by the Earlham Institute (Norwich, UK) using the LITE (Low Input, Transposase Enabled) pipeline (Perez-Sepulveda et al. 2021). Pooled libraries were sequenced on one lane (SP v1.5) of Illumina NovaSeq 6000, with a 2 x 250 paired end read metrics. DNA data was processed using the Galaxy web platform (Afgan et al., 2018). Adapter and read trimming were done using Trim Galore (version 0.6.5) and quality control was performed using FastQC (version 0.11.9). Reads were mapped against the *T. pseudonana* reference genome (Thaps3_chromosomes_assembly_chromosomes_ repeatmasked.fasta) using botwie2 (version 2.4.5) and visualized with IGV (Robinson et al., 2011).

4.2.12 *In silico* prediction and *in vitro* validation of off-target activity

Cas-OFFinder (Bae et al., 2014) was used to computationally identify all possible Cas9 cleavage sites with up to four mismatches in the genome of *T. pseudonana* for Lhcx6_sgRNA_1 GAAACGTGCTAATATTGGAT and Lhcx6_sgRNA_2 GCAGGTCCGGTAATTCCAAA. Off-target site specific primers were designed using Primer-Blast (Ye et al., 2012) and regions of interest were amplified using MyTaq HS Red Mix (Bioline) in a volume of 20 µl with primers 40-51 listed in Table 4.4 and DNA from cell line Lhcx6_MF_85 as template. PCR program was run as follows: initial denaturation at 95°C for 1 min, followed by 35 cycles of denaturation at 95°C for 15 s, annealing at 57°C for 15 s, elongation at 72°C for 10 s and a final elongation step at 72°C for 5 min. 5µl of PCR reaction was run on a 1% agarose gel to check the product size before purifying the product using the Monarch PCR&DNA clean-up kit (NEB) and sending it to Eurofins Genomics for sequencing. All sequences were aligned to the *T. pseudonana* reference genome using Geneious Prime version 2022.2.2

(https://www.geneious.com). PCR amplification and sequencing were performed by Ji Nah under my supervision.

Table 4.4: List of primers to target all potential off-target sites for both Lhcx6 sgRNAs.

Name	Sequence 5'-3'	Ref. No.
Lhcx6_sgRNA1_1_F	ACCTTGAACTGCACACCCAA	40
Lhcx6_sgRNA1_1_R	GCCTGAGAAGTTCCCTGCTA	41
Lhcx6_sgRNA1_2_F	ACACCCAGTCAGTCGAAACG	42
Lhcx6_sgRNA1_2_R	CAGCAACCGCAGCATTACAG	43
Lhcx6_sgRNA2_1_F	GCCATGGAGTTACGCTGTCT	44
Lhcx6_sgRNA2_1_R	ACCAAAGCAGTCCTACACCG	45
Lhcx6_sgRNA2_2_F	TGGATTGATGTGCCTCCCAC	46
Lhcx6_sgRNA2_2_R	ACTTGCTGGTCGTTCATCGT	47
Lhcx6_sgRNA2_3_F	TCTCGGCCTTTCGTGTCTTC	48
Lhcx6_sgRNA2_3_R	TGTGCCTTGAGCTGTAGACG	49
Lhcx6_sgRNA2_4_F	TACCGATGCAACTCTGAGCC	50
Lhcx6_sgRNA2_4_R	AACTTGGAGTGCTTGAGGGG	51

To predict off-target activity of both Lhcx6 sgRNAs *in vitro*, the Nano-OTS protocol (Höijer et al., 2020) was followed with some modifications. High molecular weight genomic DNA of wild-type *T. pseudonana* was extracted using the MagAttract kit as described above. 5-10 µg of gDNA was sheared via needle shearing by passing the DNA 4-5 times through a 26G needle. Size selection was done using the QIAEX II Gel Extraction Kit (Qiagen). RNPs were assembled by directly adding 2.5 µl of each Lhcx6 IVT sgRNA (10µM) to the mix. Sequencing was performed on a Nanopore MinION (ONT) with a FLO-MIN106 flow cell.

4.2.13 Protein extraction, SDS-Page and immunoblotting

For protein extraction of *T. pseudonana* wild-type and mutant cell line Lhcx6_MF_85s_opt, approximately 10^8 cells (100-150 ml of culture in mid-exponential phase) were harvested by centrifugation prior and 24 h after HL stress. Cells were resuspended in 100 µl lysis buffer (50mM Tris-HCl pH 6.8, 2% SDS). Samples were briefly vortexed, incubated for 30 min at RT, and centrifuged for 30 min at 13000 rpm at 4˚C to remove cell debris. The supernatant was transferred into a nuclease-free 1.5 ml Eppendorf tube and kept on ice. Protein concentration was determined using the Pierce BCA Protein Assay Kit (Thermo Scientific). Sample and

standard preparation was done in a 96-well plate following the protocol's microplate procedure. Absorbance was measured at 562 nm using a SpectraMax iD3 plate reader (Molecular Devices). Denaturing protein gel electrophoresis was done using NuPAGE Bis-Tris Mini Gels (Thermo Scientific) following the manufacturer's protocol for reduced samples and NuPAGE MES running buffer. 24 µg of protein per sample was loaded onto the gel along with a dual color standard (10-250kD, Bio-Rad) and separated on a XCell SureLock Mini-Cell system for 40 min at 200V. Proteins were transferred onto a PVDF membrane (Amersham, 0.2 µm) at 30V for 1 h using 1x transfer buffer (25mM Tris, 192 mM glycine, 10% methanol) with 0.1% NuPAGE Antioxidant. The membrane was blocked overnight at 4˚C using 5% powdered milk in 1x Tris-Buffered Saline, 0.1% Tween (TBST) followed by overnight incubation with a primary antibody raised against a C-terminal sequence of the Lhcx6 protein (Agrisera AB) at a dilution of 1:1000 at 4˚C with gentle agitation. The secondary antibody used was a horseradish peroxidase (HRP) conjugated goat-anti rabbit (Agrisera AB) with a dilution of 1:50000 in 1xTBST containing 5% milk. Blots were incubated with TMB substrate solution (Thermo Scientific) for 3 min and imaged with an iPhone13 camera. The intensities of the bands were quantified using ImageJ (Schneider et al., 2012).

4.2.14 Growth rate measurement

Cell counts were generated using the Multisizer 3 Coulter Counter (Beckman Coulter) with a 100 µm aperture tube. For each measurement, cells were diluted to a 1/10 ratio with 0.2 µm filtered 1% NaCl solution. Specific growth rate (µ) was calculated from the linear regression of the natural log of cell numbers during the exponential growth phase following the equation:

$$\mu = \frac{\ln(N_2) - \ln(N_1)}{t_2 - t_1}$$

where N_2 represents the number of cells at time t_2, N_1 the number of cells at time t_1 and t_2-t_1 the time difference between sampling points. 50.0000 cells/ml of both the *T. pseudonana* wild-type cell line and the codon modified cell line incubated at low light (LL, 50 µmol photons m^{-2} s^{-1}) were transferred to fresh Aquil media and were either kept at LL or exposed to high light (HL, 500 µmol photons m^{-2} s^{-1}) for an entire growth cycle. Growth rates were calculated for each of the three biological replicates per condition.

4.3 Results and Discussion

4.3.1 Testing of cleavage efficiencies

Bioinformatically predicted sgRNAs for two genes in *T. pseudonana* were screened for their *in vitro* cleavage efficiency using a Cas9 cleavage assay.

Two Lhcx6 sgRNAs were synthesized through IVT and tested against the target gene. Both sgRNAs-Cas9 complexes achieved similar *in vitro* cleavage efficiencies of > 90% (Figure 4.2A).

A previously used RPL10a CRISPR construct included only a single sgRNA (sgRNA_AH) and resulted in no transformed cell lines. The low cleavage efficiency of ~40% of the sgRNA could be one explanation for the unsuccessful transformation (Figure 4.2B). Therefore, six new RPL10a sgRNAs were synthesized through IVT and tested against the gene of interest. The results displayed in figure 4.2C suggest that five out of six are suitable for using *in vivo* as they are showing a high cleavage efficiency (> 90%). sgRNA_RPL10a_1 and sgRNA_RPL10a_2 were picked for subsequent usage in *in vivo* experiments as they are efficiently cutting DNA at a specific genomic locus *in vitro*.

The cleavage efficiency of Cas9 differs vastly between these sgRNAs, ranging from 5-99%. Identifying the most suitable sgRNAs for introducing a DSB at the target site increases the success rate of any genome editing project and avoids laborious repetition of cell transformation steps. Similar screening assays have shown a clear correlation between *in vitro* and *in vivo* sgRNA cleavage efficiencies (Grainger et al., 2017; Mehravar et al., 2019).

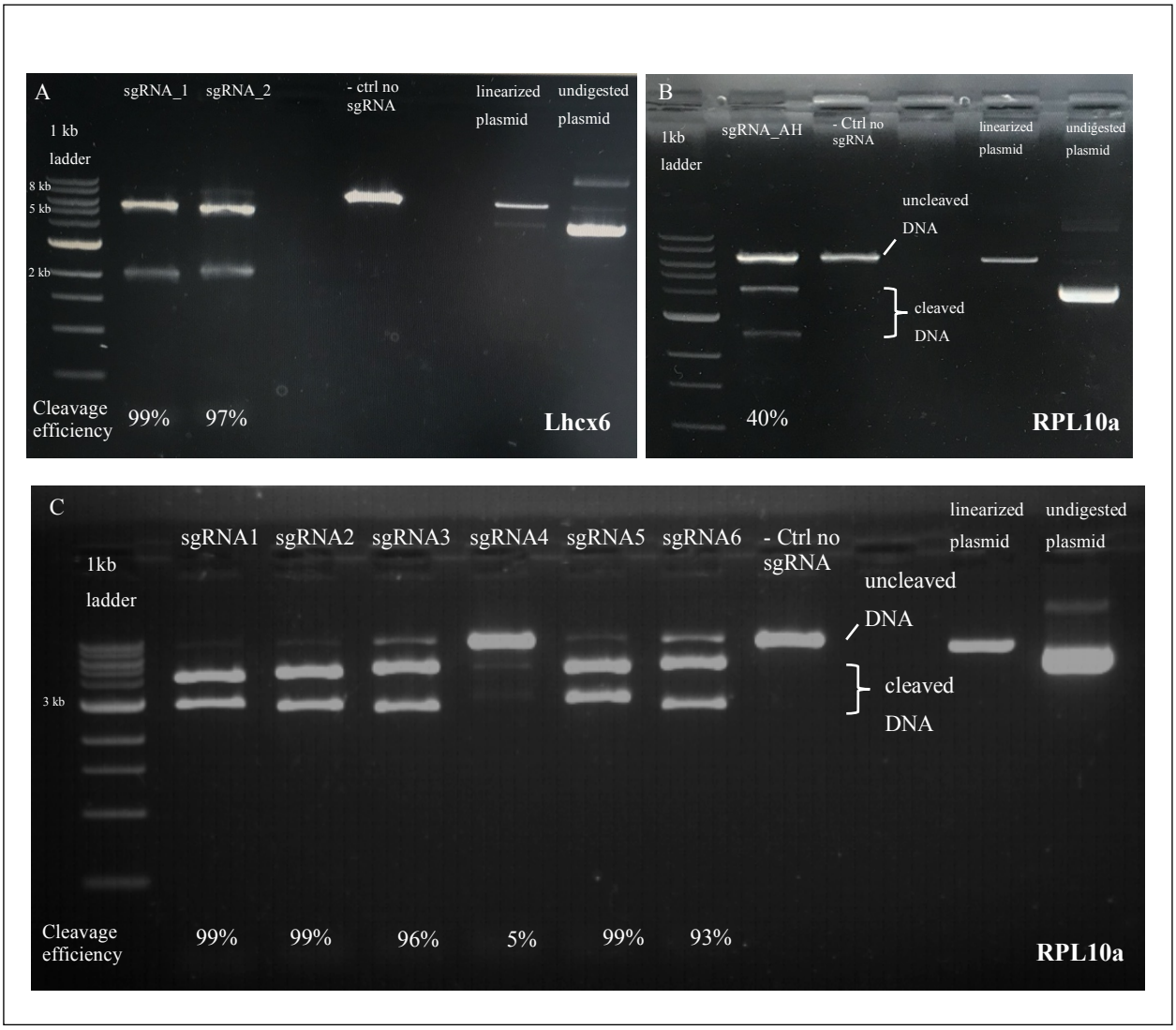


Figure 4.2: in vitro cleavage efficiencies of sgRNAs. ***A*** *Both sgRNAs used for Lhcx6 CRISPR plasmid construction showed high cleavage efficiencies > 97%.* ***B*** *RPL10a sgRNA_AH performed poorly with a low efficiency of only 40%.* ***C*** *Six new RPL10a sgRNAs were designed and tested for cleavage efficiencies, out of which five showed promising results for future in vivo cutting. Linearized (digested) and undigested plasmids as well as a negative control without sgRNAs were run for comparison.*

4.3.2 CRISPR plasmid construction

A CRISPR plasmid containing sgRNAs_RPL10a_1 and sgRNA_RPL10a_2, both showing high *in vitro* cleavage efficiency, was successfully assembled using Golden Gate cloning. PCR products of both sgRNAs (Figure 4.3A) were directly assembled into separate L1 backbones along with the U6 promoter. Restriction digest further revealed the correct insertion of the U6:sgRNA cassettes for all colonies which were screened for each Ll module (Figure 4.3B). The correct assembly of L1 modules was further confirmed by sequencing.

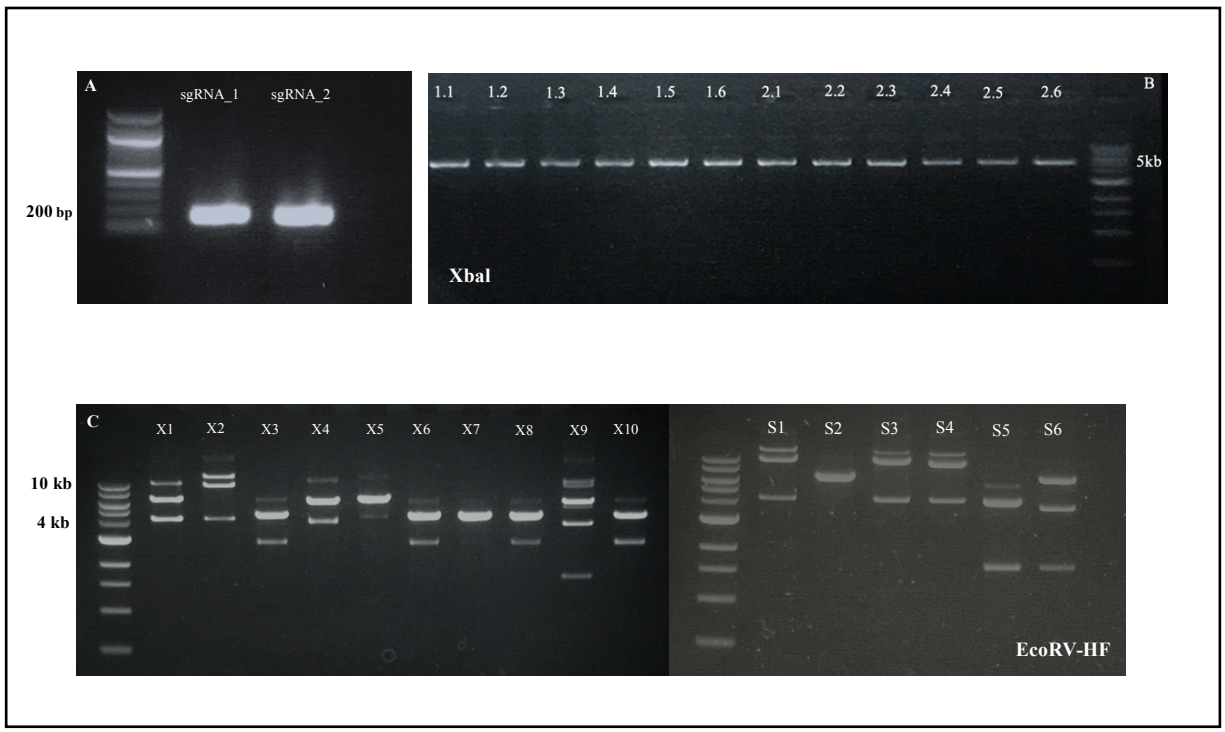


Figure 4.3: Screening of Golden Gate cloning levels via restriction digest for correct assembly of the RPL10a CRISPR plasmid. ***A*** *Bands at around 160 bp indicate the successful assembly of sgRNAs into U6:sgRNA1 and U6:sgRNA2. 100 bp DNA ladder (NEB).* ***B*** *L1 levels were screened for correct insertion of the U6:sgRNA1 (1.1-1.6) or U6:sgRNA2 cassette (2.1-2.6), respectively, via restriction digest with XbaI. All linearized plasmids displayed the expected size of 5 kb.* ***C*** *L2 constructs were screened by digestion with EcoRV-HF. Correct assembly gives a distinct band pattern at 14.6, 10.3 and 4.3 kb.*

The final CRISPR L2 level (Figure 4.4) encodes Cas9, nourseothricin resistance gene NAT, both U6:sgRNA cassettes and a linker, and is therefore rather large (14.6kb). Digestion of the plasmid with EcoRV-HF gives a very distinct band pattern (bands at 14.6kb, 10.3kb and 4.3kb), found in four out of sixteen colonies (Figure 4.3C, X2, S1, S3, S4). The large size of the CRISPR plasmids often leads to low transformation efficiencies and viability of cells to be transformed (Lesueur et al., 2016). Correct assembly of those four constructs was confirmed by sequencing.

The final Lhcx6 CRISPR plasmid containing both Lhcx6_sgRNA_1 and Lhcx6_sgRNA_2 is 14.6kb in size (Figure 4.5). Correct assembly was also confirmed via restriction digest and sequencing.

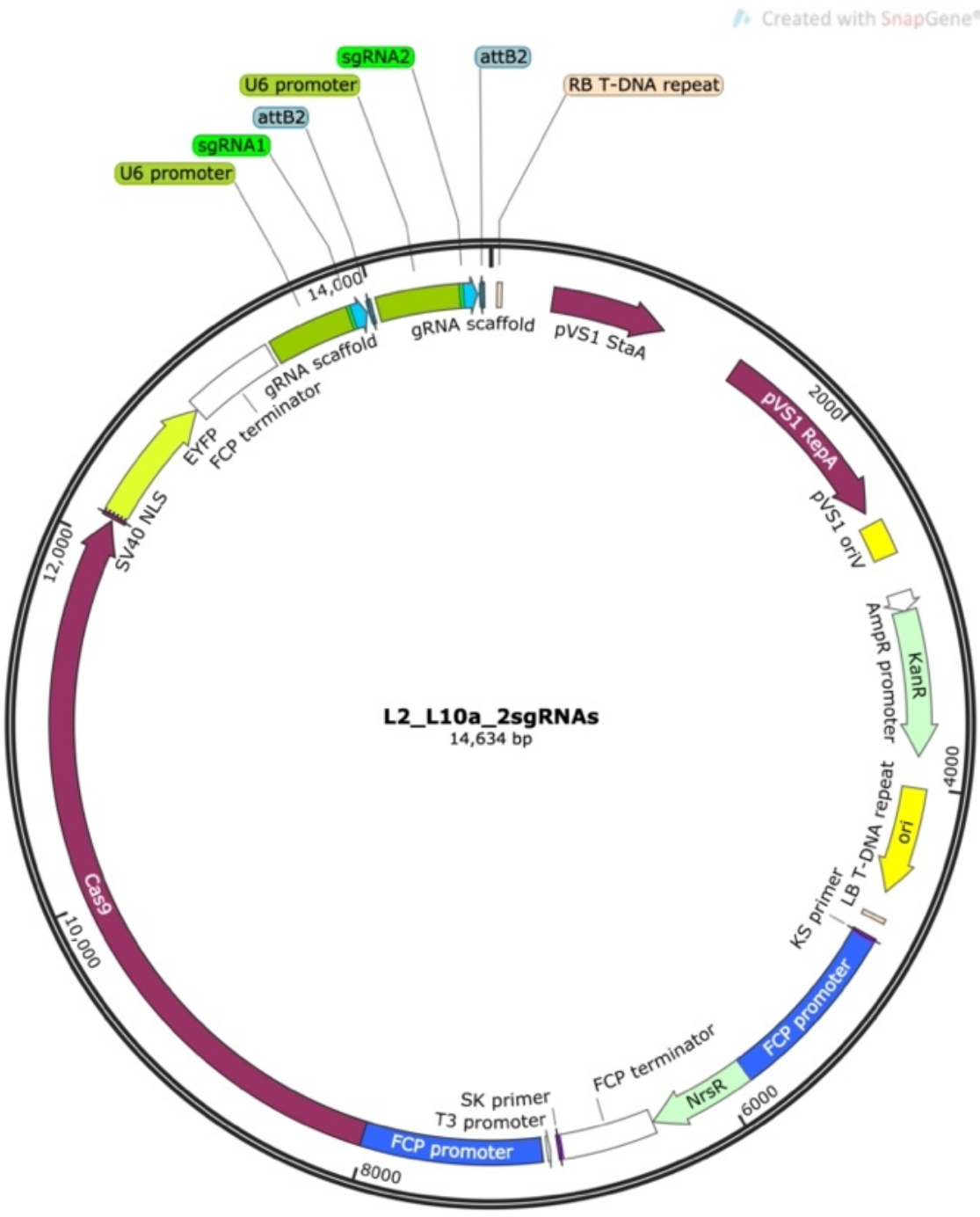


Figure 4.4: Plasmid map of the final RPL10a CRISPR level 2 construct containing Cas9 and two sgRNAs. Map created with SnapGene.

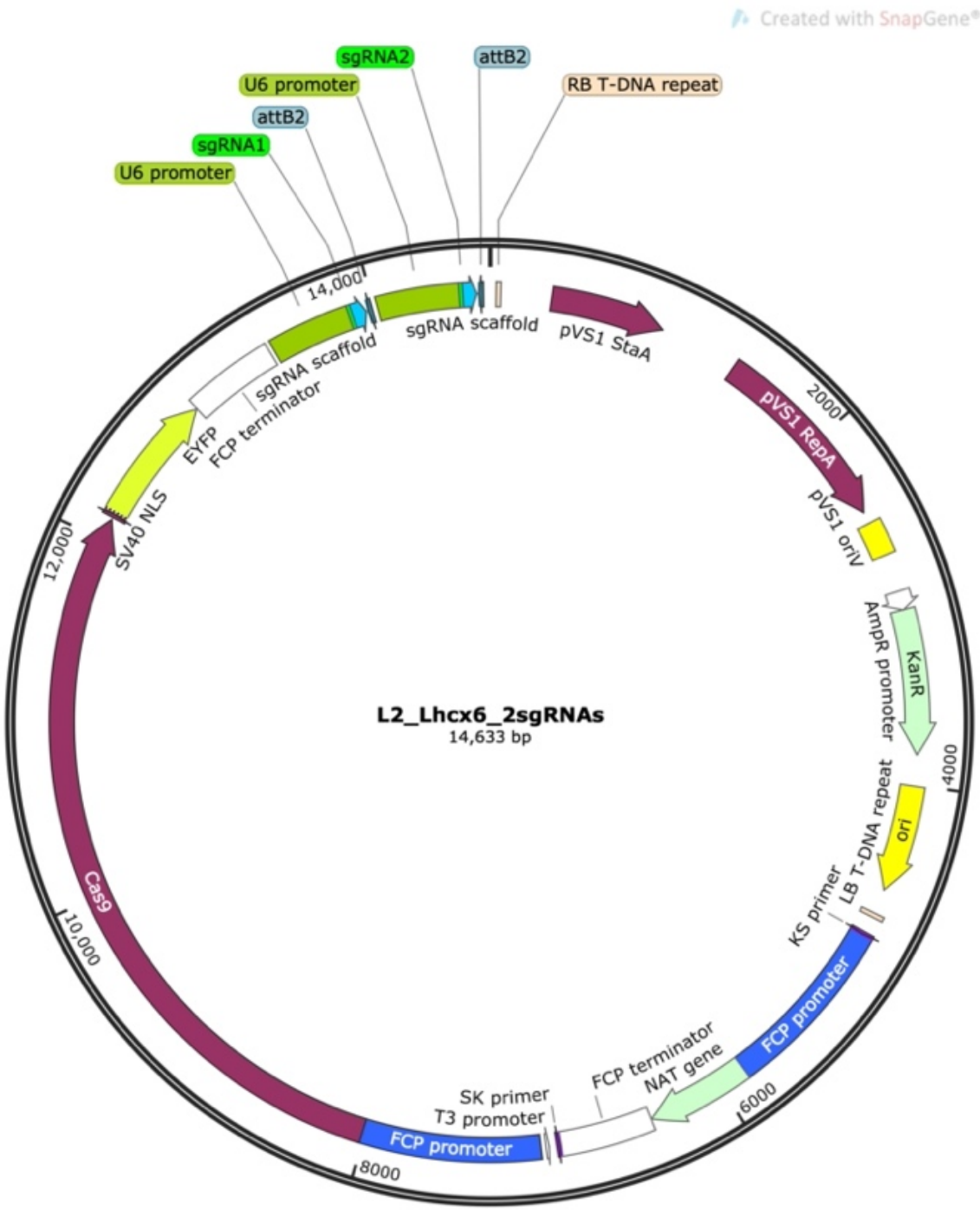


Figure 4.5: Plasmid map of the final Lhcx6 CRISPR level 2 construct containing Cas9 and two sgRNAs. Map created with SnapGene by Amanda Hopes.

4.3.3 HR plasmid construction

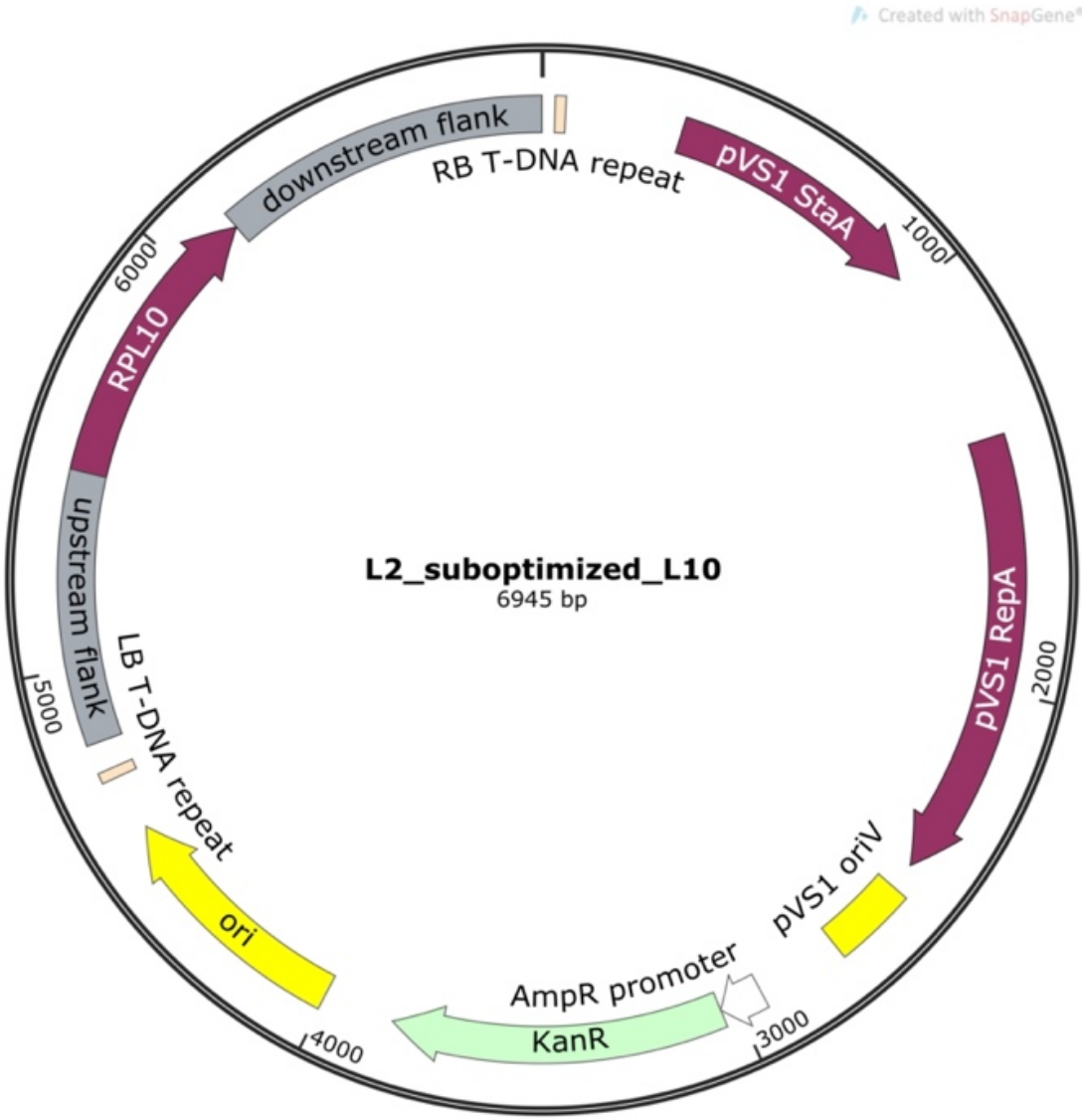


Figure 4.6: Plasmid map of the final RPL10a HR level 2 construct containing the codon modified gene flanked by two regions homologous to the non-coding 5' and 3' ends of the wild-type gene. Map created with SnapGene by Amanda Hopes.

An HR plasmid containing the codon modified RPL10a donor gene flanked by two regions homologous to the non-coding 5' and 3' ends of the RPL10a wild-type gene was successfully constructed (Figure 4.6). With a size of 6.9kb, it is substantially smaller than the CRISPR plasmid.

Assembly of the Lhcx6 plasmid resulted in a vector of 7.1kb in size and includes the codon optimized Lhcx6 gene with homologous flanking regions (Figure 4.7). Correct assembly of both HR plasmids was confirmed by sequencing of the individual construct fragments.

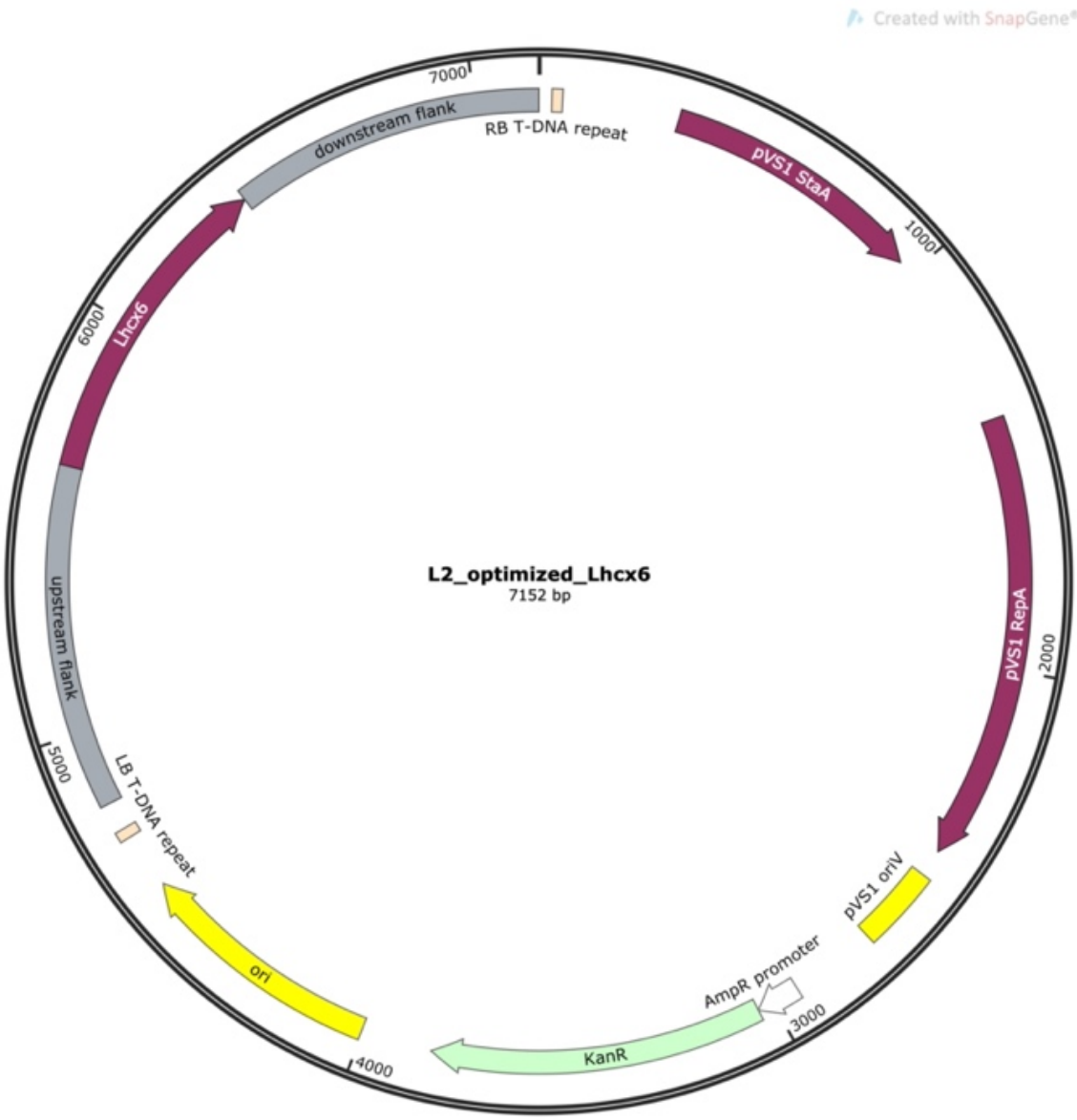


Figure 4.7: Plasmid map of the final Lhcx6 HR level 2 construct containing the codon modified gene flanked by two regions homologous to the non-coding 5' and 3' ends of the wild-type gene. Map created with SnapGene by Amanda Hopes.

4.3.4 Screening and verification of transformants

176 Lhcx6 secondary colonies were screened by colony PCR for CRISPR-Cas mediated HR. About 70% of the colonies screened positive for Cas9, which is a good indication for potential mutation, as Cas9 activity is required for HR to occur (Hopes et al., 2017; Belshaw et al., 2022). The entire Lhcx6 gene was amplified and the PCR product was used as template for nested PCRs targeting the 3' and 5' ends, respectively, of the transition between the genomic region to the HR insert. Screening for the Lhcx6 wild-type (WT) gene revealed that the WT was absent in 6% of clones which screened positive so far. In addition, the Lhcx6 codon modified (MF) gene was detected in all clones screening negative for the WT gene, which is a strong indication for bi-allelic HR to have taken place. Results of agarose gel electrophoresis indicated that clones Lhcx6_MF_85, _1.1 and _1.3 were bi-allelic (both WT alleles got replaced) and clones Lhcx6_MF_38, _53, _83, _90 were mono-allelic (only one WT allele got replaced) (Figure 4.8).

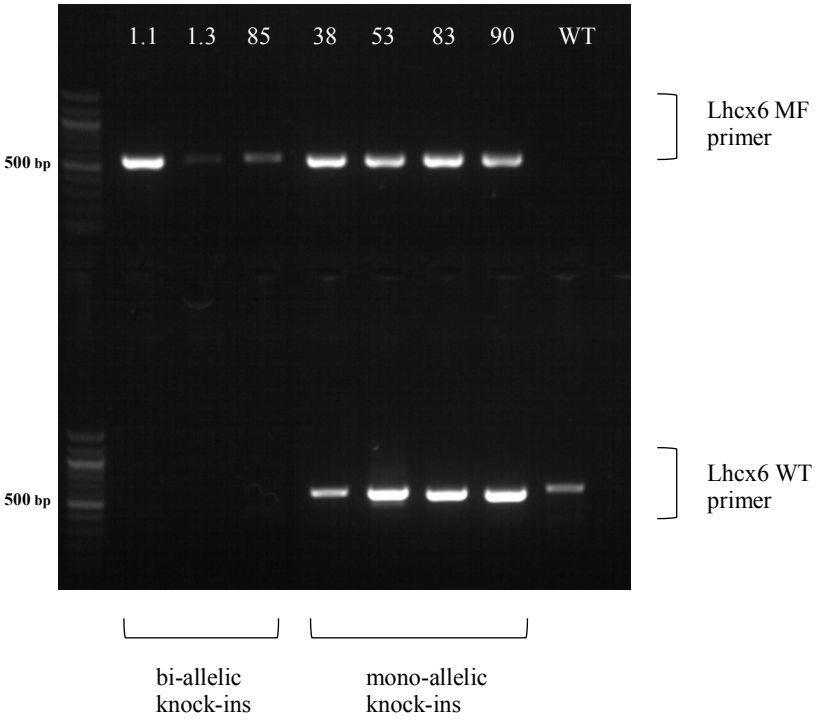


Figure 4.8: Genotype determined by PCR and agarose gel electrophoresis. Upper part: Primers targeting the modified Lhcx6 gene amplified a 571 bp fragment in bi-allelic and mono-allelic candidates. Lower part: Primers specific for the Lhcx6 wild-type gene only resulted in amplification of a 587 bp fragment in the mono-allelic clones. DNA derived from a wild-type sample was run as control. 100 bp DNA ladder (NEB).

Sanger sequencing data confirmed the replacement of Lhcx6 WT with Lhcx6 MF by HR in three potential bi-allelic clones. Sequencing chromatogram for clone Lhcx6_MF_85 showed single peaks corresponding to the codon optimized version of the Lhcx6 gene and indicating a bi-allelic replacement of the WT gene (Figure 4.9A). For comparison, monoallelic clone Lhcx6_MF_90 displays overlapping peaks at the position of changed nucleotides (Figure 4.9B).

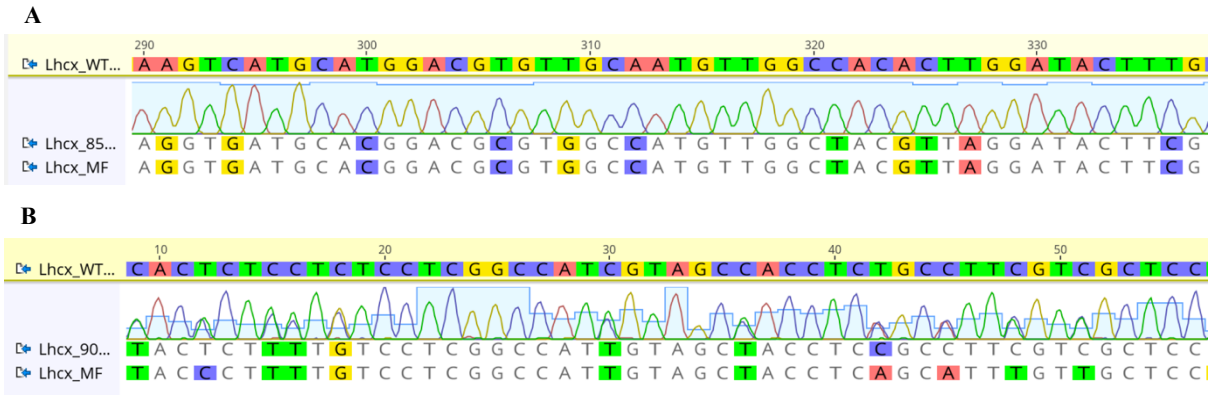


*Figure 4.9: Genotyping via Sanger sequencing of potential bi-allelic and mono-allelic Lhcx6 cell lines. **A** Single chromatogram peaks corresponding to the codon modified sequence of Lhcx6_MF_85 clone are suggesting that a bi-allelic HR event has occurred. **B** Mono-allelic HR has taken place in cell line Lhcx_MF_90 which shows two traces with overlapping peaks at the location of changed nucleotides. Top sequence: wild-type Lhcx6 sequence, middle sequence: modified clone, bottom sequence: modified Lhcx6 sequence*

Despite four attempts to replace the RPL10a gene with a codon sub-optimized version, no successful gene targeting via HR was discovered. A total of 459 secondary clones were screened, of which only 40% screened positive for integration of Cas9. The high cleavage efficiency of 99% for both RPL10a sgRNAs gives a good indication for DSBs to occur under perfect conditions, however this does not imply that it works in the living cell. A DSB might still have occurred at the target location, but the replacement of the ribosomal gene with an inferior codon usage might have been lethal to the cell and colonies could not be detected. The absence of heterozygous genotypes could indicate that the WT allele on the sister chromosome was not able to compensate for the non-functioning MF allele.

Knockdown of the rpl10a gene in zebrafish led to severe morphological abnormalities and resulted in the death of most knockdown embryos 3-7 days post-fertilization (Palasin et al., 2019). Such complex phenotypes can arise from the haploinsufficiency for a ribosomal protein which can result in inefficient translation and cell cycle arrest (Warner and McIntosh, 2009). In

yeast, mutations in the RPl10 prevents proper assembly of the small and large ribosomal subunits into a functional 80S ribosome (Bussiere et al., 2012).

It is possible, however, that the reason for not detecting codon modified RPL10a clones was due to insufficient screening, which may have required examining a significantly greater number of clones than the 459 that were screened.

Microparticle bombardment can lead to the integration of multiple copies of transgenes (between 1-10), which not only occurs at the target site but also at random positions in the host genome (Falciatore et al., 1999; Muto et al., 2013). In the case of RPL10a, HR did not occur, however, the codon-modified version of the gene was randomly integrated into the genome. This was suggested by the presence of bands using primers targeting the MF and WT gene and confirmed by sequencing of the HR loci, which indicated the complete absence of any gene replacement events. These random integration events will presumably not have had any effect on the cells as the donor DNA does not have its own promoter and will only be expressed when integrated at the correct location via HR. Although unlikely, the integration of the transgene into coding regions might unintentionally interrupt cellular pathways. Southern blotting or whole-genome sequencing could be done to detect these random integration events in the genome.

For final confirmation of HR having occurred in the potential bi-allelic mutants, Nanopore and Illumina sequencing was performed. First, long-read Nanopore sequencing of sample Lhcx6_MF_85 resulted in 4.54 million raw reads with an average length of 9 kb. In total, 25 GB of sequencing data was generated in a 72-hour run. Surprisingly, analysis of the mapped reads did not confirm bi-allelic HR but rather revealed a mosaic culture. *T. pseudonana* is a diploid organism and HR should only lead to two different mutations in a cell – either one allele was replaced or both. Divergent results indicate the presence of mosaic clones, which showed a mixed cell population with various genotypes as described in Belshaw et al. (2022). Besides WT and MF Lhcx6, truncated versions of Lhcx6 were also discovered. Those truncated versions are suggested to be the result of DSBs which were repaired by NHEJ as has been seen in previous studies in diatoms (Hopes et al., 2016; Belshaw et al., 2022). These results indicate that the secondary colony was not monoclonal despite suggested by PCR and Sanger sequencing. It is now thought that *T. pseudonana* does not grow monoclonal on plates and each colony is not arising from a single cell. A recent study just overturned the clonality myth in

E.coli and revealed that transformation of DNA plasmids into bacteria does not always produces clonal colonies (Tomoiaga et al., 2022).

For this reason, single cell sorting of the Lhcx6_MF_85 strain was done next. Nine out of twenty-two monoclonal cultures screened positive for bi-allelic HR using PCR and Sanger sequencing. One of those single-cell sorted cultures was randomly selected and it will be referred to as Lhcx6_MF_85s from now on.

Subsequent Illumina whole-genome sequencing of the Lhcx6_MF_85s and two more potential bi-allelic candidates (Lhcx6_MF_1.1 and _1.3) generated ~20 million reads. Bi-allelic knock-in of the codon optimized Lhcx6 at the homologous locus could be confirmed for sample Lhcx6_MF_85s. Visualization of mapped sequencing reads reveals a clear picture of successful gene replacement over the entire gene loci (Figure 4.10). Bi-allelic knock-in for the other two, not single-cell sorted samples, could not be confirmed. Instead, results resembled previous Nanopore data, indicating mosaic cultures. Figure 4.11 (A) shows a close up view of Lhcx6_MF_1.1 alignments to the *T. pseudonana* reference genome (the first ~60 nt downstream of the start codon), while (B) displays the alignments over the entire Lhcx6 gene loci. Coverage was low for the middle part of the gene, however, alignments mapped to the start and the end of the gene clearly showed the presence of both codon-modified and wild-type reads. Again, these results highlight the importance of single-cell sorting when performing transformations in diatom cells. Further, our data revealed that screening via PCR and Sanger sequencing was not sufficient for detecting bi-allelic HR. Our nested PCR approach might have introduced a bias (Yu et al., 2015b), resulting in detection of random gene insertions. Sanger sequencing results clearly indicate the presence of the codon modified Lhcx6 gene in the genome, however not at the targeted region.

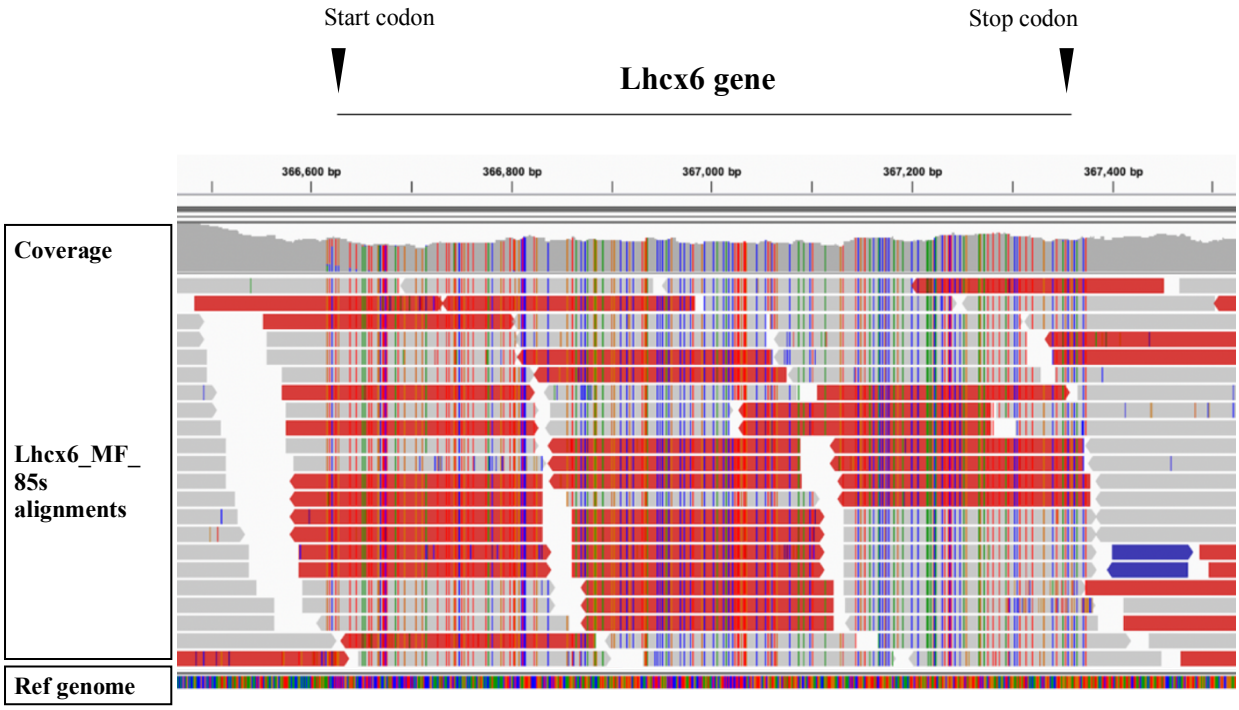


Figure 4.11: Illumina sequencing reads of Lhcx6_MF_85s cell line mapped to the T. pseudonana reference genome. Replacement of the gene with a codon optimized version is clearly visible by the mismatches highlighted between the sequencing reads and the reference genome. DNA base mismatches are located at the correct location (see Figure 7.7, Appendix E) between the start and stop codon of the Lhcx6 loci. Image created with IGV.

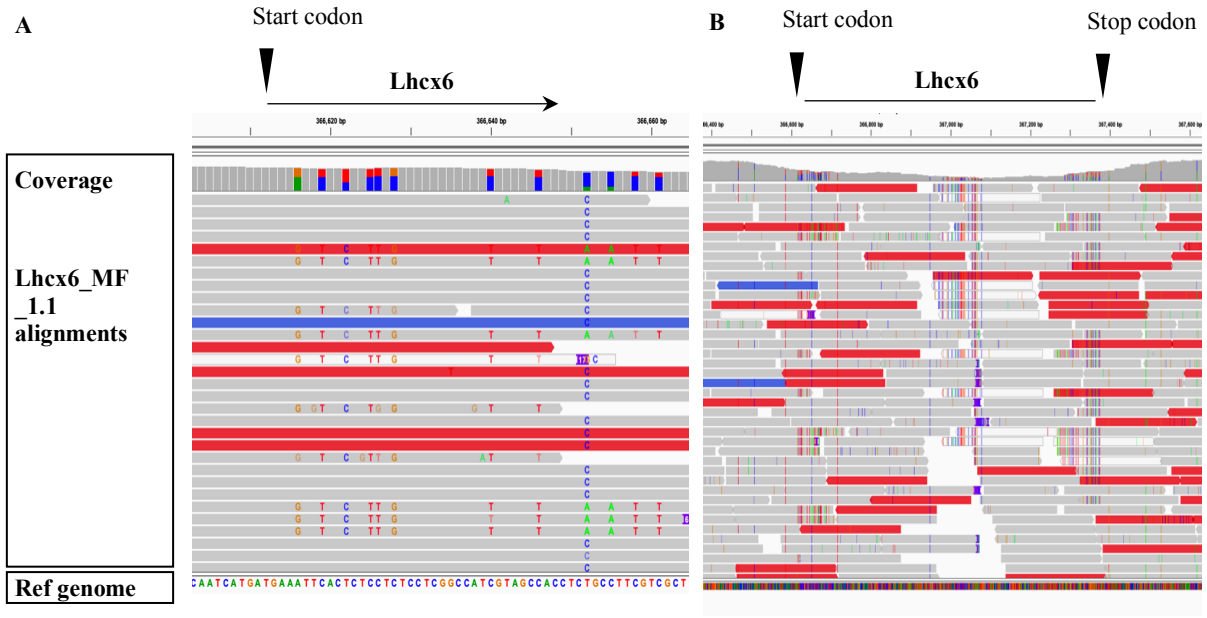


*Figure 4.10: Illumina sequencing reads of Lhcx6_MF_1.1 cell line mapped to the T. pseudonana reference genome. This cell line has not been subjected to single-cell sorting prior to sequencing and seems to be derived from a mosaic culture. **A** A close up view of the first ~60 nts upstream of the start codon showed that the replacement of the gene with a codon optimized version has happened, however, some sequencing reads are still wild-type. **B** Comparison of the mismatches over the entire Lhcx6 gene with Figure 4.10 clearly shows that no successful bi-allelic gene replacement took place. Image created with IGV.*

Highly efficient gene targeting via CRISPR/Cas-mediated HR has been achieved in *T. pseudonana* (Belshaw et al., 2022). The attempt to replace the silacidin gene with an antibiotics resistance cassette (FCP:NAT) resulted in approximately 85% of transformed colonies which screened positive for HR. However, in our study, we were not screening for a selective marker gene, which made the entire process to identify HR mutants more challenging and time-consuming. With a selectable Cas9 episome design, Moosburner et al. (2020) proposed an elegant approach to tackle this problem and to improve identification of mutant cell lines in *P. tricornutum*. They used the P2A self-cleaving peptide to transcriptionally fuse Cas9 to the selective gene *sh ble*, which enabled co-transcription under the same promoter and cleavage after translation. This allows for the selection of Cas9 via treatment with the antibiotic phleomycin. Future work could adapt this approach by fusing any HR donor gene or Cas9 with a selective marker gene to increase mutagenesis efficiency.

The choice of endonuclease used in transformations might also play a role in enhancing HR. Application of CRISPR/Cas12a has reported increased efficiency of gene targeting in a small number of studies (Begemann et al., 2017; Ferenczi et al., 2017; Li et al., 2020). Cas12a seems to have two advantages over Cas9 for genome editing via HR. First, Cas9 leaves blunt ends, while Cas12a produces a 5'overhang favouring HR over NHEJ (Volpi e Silva et al., 2021). Further, Cas12a cuts more distal to the PAM region, reducing the chances of mutagenesis in the PAM and binding region by NHEJ and therefore enhancing repeated repair until HR can occur (Huang and Puchta 2019). In *C. reinhardtti*, the use of Cas12a RNPs resulted in a gene targeting efficiency of 10% (Ferenczi et al., 2017). Disruption of the NHEJ pathway has been reported to be another way to increase HR. In *P. tricornutum*, upregulated HR rates were achieved by knockdown of a DNA ligase IV homologue (Angstenberger et al., 2019).

4.3.5 Off-target prediction and screening

Three different approaches were used to predict and validate potential off-target sites in the transformed TP cell line Lhcx6_MF_85. PCR amplification of potential off-target sites predicted by an alignment tool and subsequent sequencing was performed to detect indels (Atkins et al. 2021). Another *in vitro* approach to validate off-target sites using long-read Nanopore sequencing data was not successful.

The CasOFFinder algorithm identified six potential off-target sites for both Lhcx6 sgRNAs combined. One of those had three mismatches to the *T. pseudonana* reference genome, while

the rest differed in four nucleotides (Table 4.5). Alignment of the sequenced amplicons revealed no indels in proximity to the potential Cas9 cleavage sites, suggesting no off-target activity has occurred. Once Illumina sequencing data became available for the bi-allelic knock-in strain Lhcx6_MF_85s, all potential off-target sites were re-evaluated. No off-target activity was detected after this either.

Table 4.5: Potential off-target sites for Lhcx6 sgRNA 1 and 2 predicted by CasOFFinder. Two off-target sites were found for sgRNA 1 and four for sgRNA 2. Mismatches to the T. pseudonana genome are in lowercase letters.

Name	Sequence	Mismatches	Chromosome	Position	Direction
Lhcx6_sgRNA1_1	tgtACGTGCTAATATTGGcTCGG	4	Chr_2	2,366,114	+
Lhcx6_sgRNA1_2	GAAAtGTGCacATATTGGAaAGG	4	Chr_9	1,017,647	-
Lhcx6_sgRNA2_1	GatGGTCCGGTAtTTCCAAATGG	3	Chr_2	1,829,610	+
Lhcx6_sgRNA2_2	GgAaGgCtGGTAATTCCAAAAGG	4	Chr_2	437,405	-
Lhcx6_sgRNA2_3	GCAGGTCtGGTcATTaCcAACGG	4	Chr_5	20,965	-
Lhcx6_sgRNA2_4	cCAGGTCCGtgAcTTCCAAATGG	4	Chr_18	466,717	-

In addition, an amplification-free protocol based on Nanopore long-read sequencing, Nano-OTS was used to predict off-target activity *in vitro*. Briefly, fragmented DNA gets digested by Cas9, which comes in the form of RNPs combined with the sgRNAs of interest. The sequenced reads are aligned to a reference genome and a specifically developed software tool detects and filters Cas9 cleavage sites which are characterized by multiple reads starting at the same loci (Höijer et al., 2020). Two attempts of following the Nano-OTS protocol only yielded 289.36 Mb and 86.92 Mb, respectively, on 72-hour MinION runs. Both runs did not generate sufficient amounts of data to proceed with bioinformatic analysis. Possible reasons for the failed runs include the use of expired flow cells and reagents as well as modifications made to the protocol, which have not been tested vigorously. Fragmentation of the input DNA was done manually via needle shearing which does not produce optimal or reproducible fragment lengths. The original protocol performs shearing with the Megaruptor 2 (PacBio) which leads to superior reproducible results compared to needle shearing. According to the original protocol, 3 µg of sheared and size-selected DNA is needed for the dephosphorylation step. However, the modified protocol only yielded between 1.2 and 1.7 µg of DNA prior to library preparation. Due to the lack of access to an automated BluePippin instrument (Sage Science), size selection was performed via gel extraction, which usually leads to a decrease in DNA yield. One suggestion is to entirely omit the size selection step via gel extraction to avoid major DNA loss.

However, the possible side effects of performing subsequent library preparation with samples which are contaminated with DNA fragments < 10 kb are unclear.

In the human cell line HEK293, the Nano-OTS method has identified multiple Cas9 cleavage sites, which were not detected by computational off-target prediction tools (Höijer et al., 2020). It would be interesting to compare results from our *in silico* predictions to Nano-OTS and validate them in the modified cell line through PCR. Future work is needed to optimize this protocol for use with *T. pseudonana*.

4.3.6 Phenotyping of codon modified cell lines

4.3.6.1 Protein expression of wild-type and modified cell lines under low light and high light

To test whether the modified Lhcx6 gene showed different protein expression levels compared to the WT gene, a specific antibody which was derived from the C-terminal peptide of Lhcx6 from *T. pseudonana* was used in the western blot procedure. WT and MF cells (from clone Lhcx6_MF_85s) cultured under LL and HL conditions were blotted in duplicates or triplicates. For both WT and MF samples, the Lhcx6 protein was undetectable under LL, however the protein was strongly induced at HL, indicated by a correctly sized band at around 25 kDa (Figure 4.12). Lhcx6 is strongly induced by HL (700 µmol photons m^{-2} s^{-1}) as has been reported by Zhu and Green (2010) who saw a 20-fold increase in Lhcx6 protein levels after a 10-hour HL treatment. In the same study, the Lhcx6 protein was detectable at LL (40 µmol photons m^{-2} s^{-1}), even though at very low levels. The lack of signal under LL (50 µmol photons m^{-2} s^{-1}) in our study could result from applying a less sensitive dye compared to the highly sensitive ECL substrate used in the study by Zhu and Green (2010). These preliminary immunoblotting results showed no differences of Lhcx6 protein abundance between WT and MF cell lines under HL stress, indicating that the codon optimization of the Lhcx6 gene does not lead to increased expression levels of the protein. It must be noted here that due to time constraints, no normalization of the western blot was done. Thus, experimental errors caused e.g. by unequal loading of the samples cannot be ruled out. Here, only raw density values of bands produced by WT and MF were compared and used for final conclusion. The mean of the WT samples (calculated via the area corresponding to each peak in a profile plot) was used as control to calculate the density values (mean of WT samples = 1, mean of MF samples = 1,01).

To get an absolute measure of Lhcx6 quantities, protein extracts of both WT and MF cell lines exposed to various light regimes should be sent for mass spectrometry analysis. Recent studies

have demonstrated that optimal synonymous codon usage speeds up translation elongation rates (Yu et al., 2015a; Zhao et al., 2017). However, at this point, semi-quantitative immunoblotting showed no evidence of increased Lhcx6 protein expression due to codon optimization.

However, different banding patterns appeared between 50 and 75 kDa. Those bands were present in both WT and MF cell lines under all light conditions with bands of MF shifting into higher size ranges. Lhcx6 is associated with the photosystem II-light-harvesting antenna super complex which consists of a large number of protein subunits (Arshad et al., 2021; Nagao et al., 2022). It is unclear if those bands represent any of those subunits and how the modified Lhcx6 gene could have impacted their expression. Yet, these banding patterns are the first evidence that modification of the codon usage did result in some change of the protein landscape.

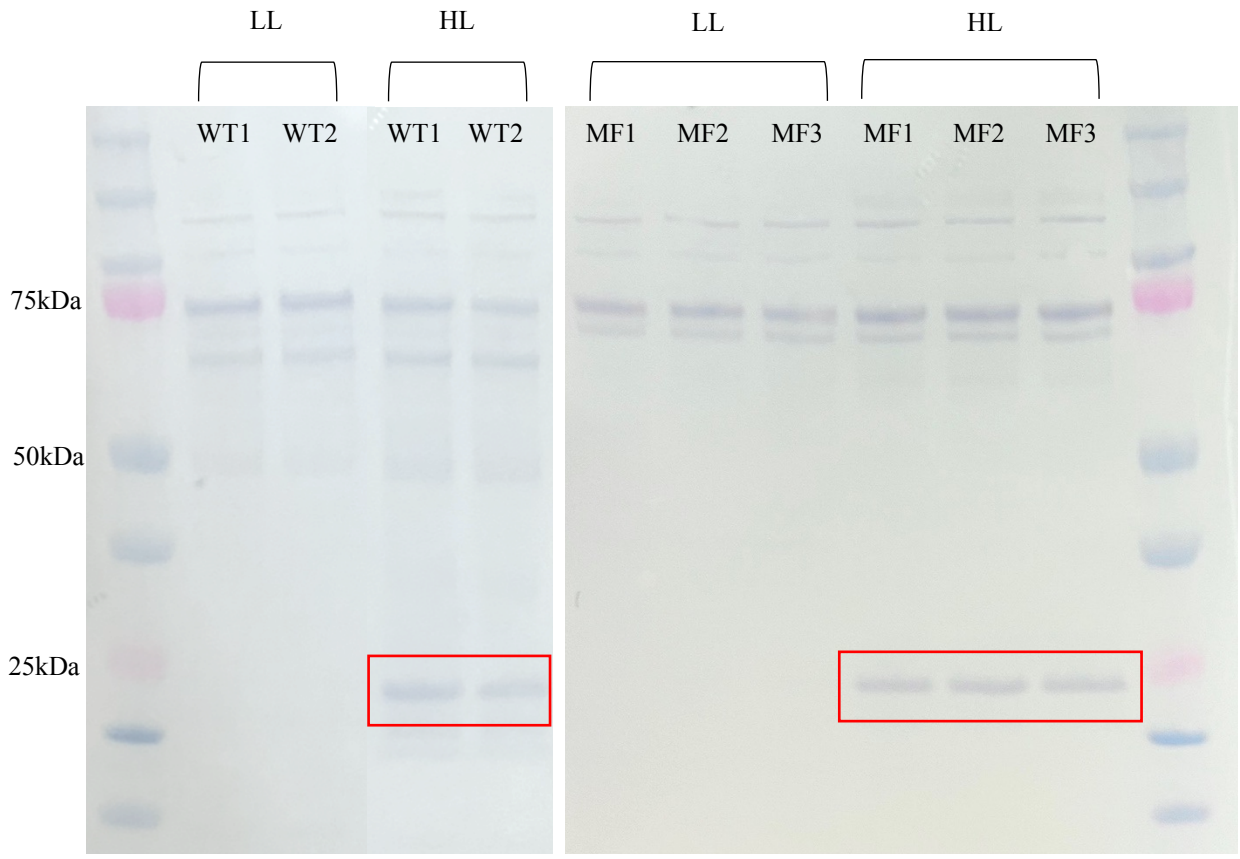


Figure 4.12: Immunoblot analysis of the Lhcx6 protein from wild-type (WT) and modified (MF) cultures incubated at low light (LL, 50 µmol photons m^{-2} s^{-1}) and after a shift to high light (HL, 500 µmol photons m^{-2} s^{-1}) for 24 hours. Red boxes highlight the bands representing the Lhcx6 protein. Two biological replicates of each WT treatment and three biological replicates of each MF treatment are shown. Each lane was loaded with 24 µg of protein lysate. Ladder: Precision plus protein dual color ladder, 10-250 kDa (Bio-Rad).

4.3.6.2 Growth rates of wild-type and modified cell lines under normal light and high light

Specific growth rates were calculated for wild-type (WT) and codon modified (MF) cell lines from the linear regression of the logarithmic growth during low light (LL) and high light (HL) exposure. Under LL, the growth rate of WT and MF cell lines did not show any significant difference (Figure 4.13). Furthermore, no significant difference was seen between the growth rates of WT and MF under HL stress. Increased growth rates were seen for both WT and MF strains under HL compared to LL. However, the codon optimization of the LHCX6 gene did not result in a significant increase of growth rate under HL stress.

Photosynthetic organisms generally increase their growth rates when exposed to higher light conditions, however this strongly depends on the balance between the capacity to absorb light and the photoprotective mechanisms to repair damage to the PSII caused by an increased flow of electrons. Under high light stress, a cell's specific growth rate only increases as long as the photodamage does not exceed the repair mechanisms of the PSII (Straka and Rittmann, 2018). Diatoms possess the ability to dissipate excess light energy as heat via non-photochemical fluorescence quenching (NPQ). This short-term photoprotective mechanisms involves the de-epoxidation of the pigment diadinoxanthin (Ddx) to diatoxanthin (Dtx) under high light (Goss and Jakob, 2010). The Lhcx6 gene is suggested to play a role in NPQ via binding to Dtx (Zhu and Green, 2010). Thus, we hypothesised that an optimized codon usage of the Lhcx6 gene, increases the capacity for NPQ and therefore leading to reduced photodamage and increased growth rate.

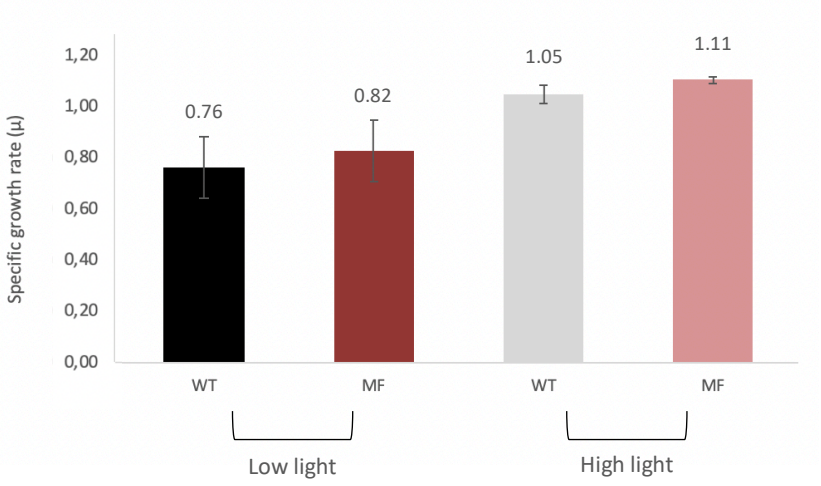


Figure 4.13: Growth rate under low light (50 µmol photons m^{-2} s^{-1}) and high light (500 µmol photons m^{-2} s^{-1}). Bars represent growth rate of three biological replicates of wild-type (WT) cells and codon modified (MF) cells. Error bars indicate standard deviation. N=3

In nature, diatoms are frequently exposed to high light irradiances that can be harmful to photosynthesis and growth. Light intensity at the ocean surface can reach up to 2000 µmol photons $m^{-2}s^{-1}$ (Long et al., 1994). In the diatom *Nitzschia aff. Pellucida*, photoinhibition and metabolic changes have been observed at irradiances ≥ 450 µmol m^{-2} s^{-1} (Lisondro et al., 2022). In *T. pseudonana*, light intensities above 110 µmol photons m^{-2} s^{-1} showed significant increase in NPQ (Zhu and Green, 2010), indicating that HL at 500 µmol photons m^{-2} s^{-1} which was used in this study, is high enough to impose a significant light stress response. NPQ measurements should be done next to get a better idea of how codon modification may impact photoprotective mechanisms in diatoms.

4.4 Conclusion

The codon usage of the Lhcx6 gene of *T. pseudonana* has been optimized via CRISPR/Cas9-mediated homologous recombination. The bi-allelic integration of the transgene was screened by PCR and confirmed by Illumina sequencing. In addition, no off-target activity of the Cas9 endonuclease was detected. Phenotyping of the Lhcx6 codon optimized cell line revealed no significant increase in growth rate under high light stress compared to the wild-type cell line. Further, immunoblotting analysis did not suggest that Lhcx6 protein abundance was elevated due to change in codon usage. Despite several transformation attempts and screening of sgRNA cleavage efficiencies *in vitro*, sub-optimization of the ribosomal RPL10a gene was not successful. More research is needed to increase the mutagenesis efficiency for these co-transformations as screening for a non-selectable gene poses extreme difficulty. The choice of synonymous codons is still elusive in diatoms. Future work should focus on ribosome profiling of the codon modified Lhcx6 cell line to investigate the role of codon usage on translation efficiency.

Chapter 5: Discussion and future work

5.1 Establishing ribosome profiling for *T. pseudonana*

The regulation of mRNA translation is a fundamental step in posttranscriptional gene regulation. Ribosome profiling is a powerful technique that provides a comprehensive understanding of this cellular translation. It involves sequencing mRNA fragments that are protected by ribosomes during nucleolytic digestion (Ingolia et al., 2009). These ribosome-protected fragments (RPFs) offer valuable information about the position of ribosomes on mRNA, enabling the determination of protein synthesis rates. So far, ribosome profiling has never been done on any marine microalgae before. Thus, in this thesis, the first ribosome profiling protocol for the diatom *T. pseudonana* has been developed. The only other microalgae for which a ribosome profiling method exists is the green algae *C. reinhardtii (Chung et al., 2015)*. Our protocol was mainly adapted from ribosome profiling protocols developed for yeast and human cell lines (McGlincy and Ingolia, 2017; Meindl et al., 2022). Key changes include optimization of cell harvesting and lysis conditions, the testing of several nuclease digestion conditions for use with *T. pseudonana* via density gradient centrifugation and the design of species-specific rRNA depletion oligos to be implemented into the library preparation protocol. Ribosome profiling requires large quantities of sample due to the processing steps involved to isolate ribosomes (McGlincy and Ingolia, 2017). *T. pseudonana* is characterized by a heavily silicified cell wall, which hampers the extraction of the large amounts of total RNA. To circumvent this problem, we have optimized the starting volumes and suggest using two replicates of each 200 ml of cells in mid-exponential phase (500.000-700.000 cell/ml). In terms of reducing work load and hands-on time, it may be worth trying to lower the starting culture volume even further, e.g. by using a cryogenic mill to extract larger amounts of total RNA without risking translocation of ribosomes (Fenton et al., 2022). A crucial step in developing ribosome footprints is determining the optimal nuclease digestion conditions since ribosomes from different species vary widely in terms of resilience to nuclease digestion (Gerashchenko and Gladyshev, 2017). Incorrect experimental conditions can lead to two potential issues in ribosome profiling: complete digestion of ribosomes or the absence of periodicity in the ribosome-protected fragments (RPFs) due to incomplete trimming.

Analysis of density gradient profiles of cells treated with varying degrees of nuclease shows the efficacy of digestion conditions. Correct digestion conditions convert polysomes to monosomes without compromising much of the ribosomal integrity (Figure 2.2) (Gerashchenko

and Gladyshev, 2017). Triplet periodicity plots, which provide a graphical representation of the periodic pattern observed in the ribosome-protected fragments (RPFs) along the mRNA, can also be used to evaluate the quality of the ribonucleic digestion (Power, 2022). Ribosome profiling data from diatoms digested with 7U/µg showed a strong enrichment of RPFs in one of the three translational reading frames, which is indicative of optimal digestion conditions. Chung et al. (2015) applied a similarly high RNase I concentration of > 9U/µg RNA for *C. reinhardtii*. Our findings, combined with their results, suggest that microalgal ribosomes may exhibit a higher nuclease tolerance compared to other organisms. Interestingly, a recent paper by Gotsmann et al. (2023) performed ribosome profiling in *C. reinhardtii* using as little as 1U/µg RNase I, similar to the concentrations used in yeast and Arabidopsis (Chartron et al., 2016; Hsu et al., 2016). While it is advantageous to reduce the amount of RNase I used, our results do not suggest a dissociation of polysomes to monosomes when applying RNase I concentration of 0.5U and 2U/µg. However, future research could explore the possibility of further reducing the nuclease concentration for *T. pseudonana* and assess whether a concentration below 7U/µg would already yield a strong triplet periodicity.

Reducing RNase I concentration is essential to minimize rRNA contamination in sequencing libraries, thereby ensuring a higher proportion of usable sequencing reads. Our results demonstrate that over-digestion did not occur, as the amount of rRNA fragments in our undepleted libraries closely matched findings from other studies (Fenton et al., 2022; Gotsmann et al., 2023). Our small-scale sequencing run revealed that a substantial amount of contamination originated from the cytosolic 5.8S rRNA and the chloroplast 23S rRNA, which is consistent with observations in Chlamydomonas (Gotsmann et al., 2023). To reduce the rRNA contamination, we developed custom depletion oligos that allows for specific depletion of the most prevailing rRNA fragments found in our *T. pseudonana* libraries. This targeted depletion successfully eliminated almost all of the specified rRNA species. However, due to variation between samples experimental conditions in subsequent sequencing runs, rRNA contamination levels remained relatively high, around 80%. Development of custom depletion oligos for rRNA removal is a labour-intensive process, as it requires preliminary ribosome profiling sequencing data to identify the most abundant contaminants. To avoid re-designing oligos for each diatom species and to deplete a larger range of rRNA genes, the development of a Pan-Diatom pool (by companies such as siTOOLsBiotech), which depletes all cytoplasmic, plastid and mitochondrial rRNA of diatoms with a sequenced genome may be worth considering once ribosome profiling has been established in the diatom community. In the meantime, using the already available Pan-Plant riboPool, designed for Angiosperms, could be

a cost-effective alternative, as it has demonstrated over 80% rRNA depletion efficiency when tested with *C. reinhardtii* (siTOOLsBiotech).

Our primary goal was to develop a ribosome profiling protocol which enables genome-wide study of translation in the model diatom *T. pseudonana.* Our optimized strategy yields high-quality footprint data with codon resolution comparable to other well-established protocols (e.g. Ingolia et al., 2009; Hsu et al., 2016; McGlincy and Ingolia, 2017). Moreover, our aim was to create a protocol which can easily be adapted for studying translation in other diatom species by primarily optimizing the nuclease digestion conditions. Performing ribosome profiling on various diatom species will greatly enhance our understanding of translation in these important microalgae. Preliminary data from a gradient profile of the polar diatom *F. cylindrus* suggests that a similar nuclease digest, as used in *T. pseudonana,* can be applied, albeit with the requirement of significantly smaller culture volumes. Thus, the development of this ribosome profiling protocol represents a substantial expansion of the molecular toolbox available for diatoms.

5.2 Ribosome Profiling of environmentally stressed cells

As photosynthetic organisms, diatoms are frequently exposed to fluctuating light levels in the ocean surface. This requires rapid regulation of gene expression to adapt to the environmental conditions. Transcriptional and proteomic studies have investigated the responses in diatoms following light stress (Nymark et al., 2009; Dong et al., 2016). Transcriptional responses include a rapid regulation of genes which encode proteins that are involved in photosynthesis, pigment metabolism and reactive oxygen species (ROS) scavenging systems (Nymark et al., 2009). Ribosome profiling fills this gap by allowing to study translation regulation. Ribosome profiling quantifies ribosome occupancy of transcripts on a genome-wide scale. By parallel sequencing of all expressed transcripts using RNA-seq, the translation efficiency (TE) of genes was calculated revealing genome-wide translational regulation upon high light stress in *T. pseudonana*.

Transcriptional regulation seems to be the dominant regulatory mechanism in response to light stress. This was indicated by the majority of genes displaying fold changes in mRNA and RPF levels which were concordant (Figures 3.1 and 3.5). However, a number of genes showed changes in translation efficiency (TE) between control and high light conditions, suggesting that they undergo translational regulation. After 4 h of HL stress, we identified 2822 differentially transcribed genes (DTGs) and 461 differential translation efficiency genes

(DTEGs). Among the DTEGs, we found 93 genes which were exclusively regulated on the translational level. Of these, 56 were upregulated and 37 were downregulated.
Prolonged HL stress (24 h) led to an increase in differentially expressed genes compared to short-term exposure. We found a total of 6546 DTGs and 1469 DTEGs and identified 246 exclusively translationally regulated genes, of which 125 and 121 were upregulated and downregulated, respectively.
Functional analysis revealed that several chlorophyll a/c-binding proteins (FCPs) were among genes with differential TE after both short-term and prolonged HL stress. They are members of the light-harvesting complex proteins (LHC) superfamily and are suggested to play important roles in photoprotection (Zhu and Green, 2010; Dong et al., 2016). Translational regulation of LHC genes in response to light have been shown in several photosynthetic organisms (McKim and Durnford, 2006; Floris et al., 2013). However, our results are the first evidence of their regulation at the level of translation in diatoms.
Comparison of differentially expressed genes after 4 h and after 24 h of HL stress revealed two distinct translationally regulated responses upon short-term and prolonged HL stress in *T. pseudonana.* These results indicate dynamic and fast responding regulatory mechanisms which are in line with results from a transcriptional profiling study (Nymark et al., 2009).

Even though our data suggest widespread translational regulation in diatoms under light stress conditions, the large number of genes which lack proper annotation are hampering meaningful functional analysis of genes involved in translationally regulated response mechanisms. Hopefully, ribosome profiling can contribute to improving gene characterization in the *T. pseudonana* gene model in the future. Considering the wide range of possible applications of ribosome profiling, our analyses of the change in TE upon high light stress does not unlock the full potential of the dataset yet. Processing of ribosome profiling data is challenging and requires computational expertise and often extensive knowledge of command line usage. In addition, the application of many tools is still limited to a set of model organisms and/or organisms with an available transcriptome of high quality, or they require annotations from a specific database (Perkins et al., 2019; Verbruggen et al., 2019; Liu et al., 2020). Some of these issues have been addressed by toolkits which automate and simplify the analysis procedure of ribosome profiling data (Michel et al., 2016; Berg et al., 2020; Francois et al., 2021). Protocol development of more non-model organisms, such as diatoms, will hopefully further increase the applicability of user-friendly tools to process and visualize ribosome profiling data.

Future work could focus on detection of novel translated ORFs in *T. pseudonana*. This can be a challenging task due to the high heterogeneity and noise levels in ribosome profiling data, especially when looking for sORFs (Choudhary et al., 2020). A number of programs have been developed in recent years to detect translated ORFs (e.g. Calviello et al., 2016; Ndah et al., 2017; Xiao et al., 2018; Choudhary et al., 2020), most of which are demanding in terms of bioinformatic expertise and computational power. The development of the Trips-viz browser allows visualization of footprint data at the level of individual RNAs and facilitates the detection of ORFs (Kiniry et al., 2021). To reduce users' computational workload, Trips-viz has been incorporated into the RiboGalaxy platform (Michel et al., 2016). Until recently, the use of Trips-viz has been limited to a small range of organisms. However, the integration of a new feature which allows uploading of custom transcriptomes to the browser finally enables the analysis of data obtained from any species (Kiniry et al., 2021). However, setting up a new species on Trips-viz is still not a straightforward task and requires considerable computational expertise (Personal communication Pasha Baranov and Jack Tierney). To perform this analysis on the browser, re-mapping of our sequencing reads to the transcriptome is needed.

Furthermore, ribosomal pausing is considered a way to regulate translation in response to environmental stress (Zhang et al., 2017; Karlsen et al., 2018). It may be worth looking into pause detection to see if this is a widespread regulatory mechanism in *T. pseudonana* upon HL stress.

5.3 CRISPR/Cas-mediated HR to change codon usage

Codon usage affects translation efficiency by regulating elongation rates during mRNA translation. Optimal codons lead to increased elongation speed while rare codons slow down elongation (Yu et al., 2015a; Liu et al., 2021). To investigate the role of codon usage in diatoms, we created a *T. pseudonana* cell line with a codon optimized Lhcx6 gene. Efficient gene targeting via CRISPR/Cas-mediated homologous recombination (HR) has been achieved in *T. pseudonana* before (Belshaw et al., 2022). Our study, however was the first to modify the codon usage of a gene via HR which did not allow screening for a selective marker gene. This meant that screening for HR mutants was challenging and laborious. Any future HR work needs to find a way to increase screening efficiency. It may be possible to use the P2A self-cleaving peptide (Moosburner et al., 2020) to transcriptionally fuse any HR gene with a selective marker. PCR screening of codon modified genes is difficult and requires a number of highly specific primer sets. Here, it resulted in the misleading assumption that bi-allelic replacement occurred, when the colonies were actually mosaic. Any future work with diatom transformations should

sidestep from laborious re-streaking of clones onto fresh agar plates as this approach does not necessarily lead to monoclonal colonies (Tomoiaga et al., 2022). Instead, single-cell sorting should become the standard method to isolate cells before screening for mutants via PCR. However, this specialised equipment might not be available in every laboratory.

Sub-optimization of the RPL10a gene was not successful despite several transformation attempts and selection of sgRNAs with high *in vitro* cleavage efficiencies. Knockdown of the gene in zebrafish resulted in high mortality rates (Palasin et al., 2019). It would be interesting to test if codon modification of the RPL10a gene was critical for lethality of the cells. This may be done by modifying a smaller number of codons and generating a gradient of codon modification to be able to calculate the threshold for lethality.

Phenotyping via immunoblotting and growth rate measurements of codon optimized cells did not show evidence of increased Lhcx6 protein expression or growth rate under high light stress (Figures 4.12. and 4.13). Further phenotyping is needed to determine the potential effects of codon modification in the Lhcx6 gene. The Cas9-edited cell line is currently maintained in the lab by sub-culturing but is also cryopreserved. Thus, it is readily available for further analysis. This could include studying NPQ development via chlorophyll fluorescence measurements and using mass spectrometry analysis to quantify absolute protein levels. Finally, since this thesis evolves around ribosome profiling, applying our newly developed protocol to the edited cell line should definitely be proposed for future work. This can provide new insights into the effect of codon change on elongation rate and translational pausing. Ribosome profiling studies have revealed a negative correlation between codon usage and ribosome density. This suggests that rare codons exhibit a higher ribosome density, implying a slower decoding process (Weinberg et al., 2016; Mohammad et al., 2019). Performing ribosome profiling on our cell line with modified codon usage would provide first insights into how codon usage affects translational regulation in diatoms.

5.4 Broader context of results

The development of ribosome profiling for *T. pseudonana* marks a significant expansion of the molecular tools available for diatom research. This advancement brings the field of diatom research in line with the progress achieved in the study of the green algae *Chlamydomonas reinhardtii*, a well-established model organism in plant biology and biotechnology (Mock et

al., 2022). Ribosome profiling for *C. reinhardtii* has already been established in 2015 (Chung et al., 2015) leading to a wide range of fascinating discoveries in this species (Chung et al., 2017; Trösch et al., 2018). Our protocol holds the promise of providing similar insights into the field of diatom genetics with focus on translational regulation. Diatoms are responsible for about 20% of the global primary production and play a key role in biochemical cycles and the marine food web (Field et al., 1998; Benoiston et al., 2017). Therefore, understanding how diatoms adapt to changing environmental conditions and stress factors through translational control can have implications for ecology, biochemistry and climate change.

Modeling studies show that rising ocean temperatures result in diatom communities shifting pole-wards. (Barton, 2016; Seinacher, 2010). This shift in geographical distribution also alters the light conditions to which microalgae are exposed. Ribosome profiling data detailing how diatoms adjust their translation in response to changes in light and other environmental factors provides crucial insights into the complex molecular responses and dynamics within these organisms. Our research has demonstrated that under high light stress, a substantial number of genes are subject to exclusive translational regulation, underscoring the critical significance of this aspect of gene expression. Using ribosome profiling on diatoms exposed to a variety of environmental stressors can help unlock the role of translation in response to climate change.

Further, diatoms find widespread application in various biotechnological processes due to their ability to produce valuable compounds, such as lipids, pigments or polysaccharides (Dolatabadi and de la Guardia, 2011; Sharma et al., 2021). Ribosome profiling allows researchers to identify and manipulate genes and pathways involved in the biosynthesis and regulation of these, as well as to optimize their production under different environmental conditions. Factors like codon usage influence the rate of elongation (Weinberg et al., 2016; Mohammad et al., 2019), potentially resulting in increased protein synthesis rates—a development with significant relevance for biotechnology. Thus, the knowledge gained from studying translational regulation in diatoms has the potential to yield valuable implications for further biotechnological applications.

List of abbreviations

A-site	aminoacyl-site
aa-tRNA	aminoacyl-tRNA
cDNA	complenetary DNA
CDS	coding sequence
Chl	chlorophyll
CHX	cycloheximide
CRISPR	clustered regularly interspaced short palindromic repeats
CUB	codon usage bias
Ddx	diadinoxanthin
DSB	double-strand break
DTEGs	differential translation efficiency genes
DTGs	differentially transcribed genes
Dtx	diatoxanthin
E-site	exit site
FCP	fucoxanthin chlorophyll a/c-binding protein
FPLC	fast protein liquid chromatography
GO	gene ontology
HL	high light
HR	homologous recombination
IGV	integrative genomics viewer
IRES	internal ribosome entry site
IVT	*in vitro* transcription
JGI	Joint Genome Institute
KOG	EuKaryotic Orthologous Groups
LHC	light-harvesting complex protein
LL	low light
LSU	large subunit
MF	modified
miRNA	microRNA
mRNA	messenger RNA
NAT	Nourseothricin N-acetyl transferase
NHEJ	non-homologous end joining

NPQ	non-photochemical fluorescence quenching
nt	nucleotides
ORF	open reading frame
P-site	peptidyl-site
PAGE	poly-acrylamide gel electrophoresis
PAM	protospacer adjacent motif
PRB	polysome resuspension buffer
PS	photosystem
RBP	RNA-binding protein
RISC	RNA-induced silencing complex
Rnase	ribonuclease
RNP	ribonucleoprotein
ROS	reactive oxygen species
RP	ribosomal protein
RPF	ribosome-protected fragment
rRNA	ribosomal RNA
RSCU	relative synonymous codon usage
sgRNA	single guide RNA
sORF	small open reading frame
SSU	small subunit
TALEN	transcription activator-like effector nuclease
TE	translation efficiency
TOR	target of rapamycin
TRAP	translating ribosome affinity purification
tRNA	transfer RNA
UMI	unique molecular identifier
uORF	upstream ORF
UTR	untranslated region
WT	wild-type
XC	xanthophyll cycle
ZFN	zinc finger nuclease

References

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M. et al. (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* **46**: W537-w544.

Akinci, E., Hamilton, M.C., Khowpinitchai, B., and Sherwood, R.I. (2021) Using CRISPR to understand and manipulate gene regulation. *Development* **148**.

Allen, A.E., Dupont, C.L., Obornik, M., Horak, A., Nunes-Nesi, A., McCrow, J.P. et al. (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. Nature 473: 203-+.

Allen, C.S., Thomas, E.R., Blagbrough, H., Tetzner, D.R., Warren, R.A., Ludlow, E.C., and Bracegirdle, T.J. (2020). Preliminary Evidence for the Role Played by South Westerly Wind Strength on the Marine Diatom Content of an Antarctic Peninsula Ice Core (1980–2010) [WWW document].

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Angstenberger, M., Krischer, J., Aktaş, O., and Büchel, C. (2019) Knock-Down of a ligIV Homologue Enables DNA Integration via Homologous Recombination in the Marine Diatom Phaeodactylum tricornutum. *ACS Synth Biol* **8**: 57-69.

Arella, D., Dilucca, M., and Giansanti, A. (2021) Codon usage bias and environmental adaptation in microbial organisms. *Mol Genet Genomics* **296**: 751-762.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H. et al. (2004a) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* **306**: 79-86.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H. et al. (2004b) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* **306**: 79-86.

Arshad, R., Calvaruso, C., Boekema, E.J., Büchel, C., and Kouřil, R. (2021) Revealing the architecture of the photosynthetic apparatus in the diatom Thalassiosira pseudonana. *Plant Physiol* **186**: 2124-2136.

Aspden, J.L., Eyre-Walker, Y.C., Phillips, R.J., Amin, U., Mumtaz, M.A., Brocard, M., and Couso, J.P. (2014) Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *Elife* **3**: e03528.

Babitzke, P., Baker, C.S., and Romeo, T. (2009) Regulation of translation initiation by RNA binding proteins. *Annu Rev Microbiol* **63**: 27-44.

Bae, S., Park, J., and Kim, J.S. (2014) Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**: 1473-1475.

Bazzini, A.A., Lee, M.T., and Giraldez, A.J. (2012) Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* **336**: 233-237.

Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S. et al. (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**: 981-993.

Begemann, M.B., Gray, B.N., January, E., Gordon, G.C., He, Y., Liu, H. et al. (2017) Precise insertion and guided editing of higher plant genomes using Cpf1 CRISPR nucleases. *Sci Rep* **7**: 11606.

Behrenfeld, M.J., Halsey, K.H., Boss, E., Karp-Boss, L., Milligan, A.J., and Peers, G. (2021) Thoughts on the evolution and ecological niche of diatoms. *Ecological Monographs* **91**: e01457.

Belhaj, K., Chaparro-Garcia, A., Kamoun, S., and Nekrasov, V. (2013) Plant genome editing made easy: targeted mutagenesis in model and crop plants using the CRISPR:Cas system. *Plant Methods* **9**.

Belshaw, N., Grouneva, I., Aram, L., Gal, A., Hopes, A., and Mock, T. (2022) Efficient gene replacement by CRISPR/Cas-mediated homologous recombination in the model diatom Thalassiosira pseudonana. *New Phytologist* **n/a**.

Benoiston, A.S., Ibarbalz, F.M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S., and Bowler, C. (2017) The evolution of diatoms and their biogeochemical functions. *Philos Trans R Soc Lond B Biol Sci* **372**.

Berg, J.A., Belyeu, J.R., Morgan, J.T., Ouyang, Y., Bott, A.J., Quinlan, A.R. et al. (2020) XPRESSyourself: Enhancing, standardizing, and automating ribosome profiling computational analyses yields improved insight into data. *PLoS Comput Biol* **16**: e1007625.

Bertin, B., Renaud, Y., Aradhya, R., Jagla, K., and Junion, G. (2015) TRAP-rc, Translating Ribosome Affinity Purification from Rare Cell Populations of Drosophila Embryos. *J Vis Exp*.

Botzman, M., and Margalit, H. (2011) Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol* **12**: R109.

Bowler, C., Vardi, A., and Allen, A.E. (2009) Oceanographic and Biogeochemical Insights from Diatom Genomes. *Annual Review of Marine Science* **2**: 333-365.

Bowler, C., De Martino, A., and Falciatore, A. (2010) Diatom cell division in an environmental context. *Current Opinion in Plant Biology* **13**: 623-630.

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A. et al. (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456**: 239-244.

Brar, G.A., and Weissman, J.S. (2015) Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* **16**: 651-664.

Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T., and Weissman, J.S. (2012) High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335**: 552-557.

Brule, C.E., and Grayhack, E.J. (2017) Synonymous Codons: Choose Wisely for Expression. *Trends Genet* **33**: 283-297.

Brunet, C., and Lavaud, J. (2010) Can the xanthophyll cycle help extract the essence of the microalgal functional response to a variable light environment? *Journal of Plankton Research* **32**: 1609-1617.

Buchbender, A., Mutter, H., Sutandy, F.X.R., Körtel, N., Hänel, H., Busch, A. et al. (2020) Improved library preparation with the new iCLIP2 protocol. *Methods* **178**: 33-48.

Bussiere, C., Hashem, Y., Arora, S., Frank, J., and Johnson, A.W. (2012) Integrity of the P-site is probed during maturation of the 60S ribosomal subunit. *J Cell Biol* **197**: 747-759.

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M. et al. (2016) Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* **13**: 165-170.

Chartron, J.W., Hunt, K.C., and Frydman, J. (2016) Cotranslational signal-independent SRP preloading during membrane targeting. *Nature* **536**: 224-228.

Chassé, H., Boulben, S., Costache, V., Cormier, P., and Morales, J. (2017) Analysis of translation using polysome profiling. *Nucleic Acids Res* **45**: e15.

Chassé, H., Mulner-Lorillon, O., Boulben, S., Glippa, V., Morales, J., and Cormier, P. (2016) Cyclin B Translation Depends on mTOR Activity after Fertilization in Sea Urchin Embryos. *PLoS One* **11**: e0150318.

Chen, J., Melton, C., Suh, N., Oh, J.S., Horner, K., Xie, F. et al. (2011) Genome-wide analysis of translation reveals a critical role for deleted in azoospermia-like (Dazl) at the oocyte-to-zygote transition. *Genes Dev* **25**: 755-766.

Chepurnov, V.A., Mann, D.G., Sabbe, K., and Vyverman, W. (2004) Experimental studies on sexual reproduction in diatoms. *Int Rev Cytol* **237**: 91-154.

Chotewutmontri, P., and Barkan, A. (2018) Multilevel effects of light on ribosome dynamics in chloroplasts program genome-wide and psbA-specific changes in translation. *PLoS Genet* **14**: e1007555.

Chothani, S., Adami, E., Ouyang, J.F., Viswanathan, S., Hubner, N., Cook, S.A. et al. (2019) deltaTE: Detection of Translationally Regulated Genes by Integrative Analysis of Ribo-seq and RNA-seq Data. *Curr Protoc Mol Biol* **129**: e108.

Choudhary, S., Li, W., and A, D.S. (2020) Accurate detection of short and long active ORFs using Ribo-seq data. *Bioinformatics* **36**: 2053-2059.

Chung, B.Y., Hardcastle, T.J., Jones, J.D., Irigoyen, N., Firth, A.E., Baulcombe, D.C., and Brierley, I. (2015) The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. *RNA* **21**: 1731-1745.

Chung, B.Y.W., Deery, M.J., Groen, A.J., Howard, J., and Baulcombe, D.C. (2017) Endogenous miRNA in the green alga Chlamydomonas regulates gene expression through CDS-targeting. *Nature Plants* **3**: 787-794.

Clancy, S., and Brown, W. (2008) Translation: DNA to mRNA to Protein. *Nature Education* **1**.

Cribbs, A.P., and Perera, S.M.W. (2017) Science and Bioethics of CRISPR-Cas9 Gene Editing: An Analysis Towards Separating Facts and Fiction. *Yale J Biol Med* **90**: 625-634.

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

Dalal, V.K., and Tripathy, B.C. (2012) Modulation of chlorophyll biosynthesis by water stress in rice seedlings during chloroplast biogenesis. *Plant Cell Environ* **35**: 1685-1703.

Dard-Dascot, C., Naquin, D., d'Aubenton-Carafa, Y., Alix, K., Thermes, C., and van Dijk, E. (2018) Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *Bmc Genomics* **19**.

de Koning, W., Miladi, M., Hiltemann, S., Heikema, A., Hays, J.P., Flemming, S. et al. (2020) NanoGalaxy: Nanopore long-read sequencing data analysis in Galaxy. *Gigascience* **9**.

del Prete, M.J., Vernal, R., Dolznig, H., Müllner, E.W., and Garcia-Sanz, J.A. (2007) Isolation of polysome-bound mRNA from solid tissues amenable for RT-PCR and profiling experiments. *Rna* **13**: 414-421.

Deprost, D., Yao, L., Sormani, R., Moreau, M., Leterreux, G., Nicolaï, M. et al. (2007) The Arabidopsis TOR kinase links plant growth, yield, stress resistance and mRNA translation. *EMBO Rep* **8**: 864-870.

Diner, R.E., Bielinski, V.A., Dupont, C.L., Allen, A.E., and Weyman, P.D. (2016) Refinement of the Diatom Episome Maintenance Sequence and Improvement of Conjugation-Based DNA Delivery Methods. *Front Bioeng Biotechnol* **4**: 65.

Dolatabadi, J.E.N., and de la Guardia, M. (2011) Applications of diatoms and silica nanotechnology in biosensing, drug and gene delivery, and formation of complex metal nanostructures. *TrAC Trends in Analytical Chemistry* **30**: 1538-1548.

Domingues, N., Matos, A.R., Marques da Silva, J., and Cartaxana, P. (2012) Response of the diatom Phaeodactylum tricornutum to photooxidative stress resulting from high light exposure. *PLoS One* **7**: e38162.

Dong, H.P., Dong, Y.L., Cui, L., Balamurugan, S., Gao, J., Lu, S.H., and Jiang, T. (2016) High light stress triggers distinct proteomic responses in the marine diatom Thalassiosira pseudonana. *BMC Genomics* **17**: 994.

Dorrell, R.G., Liu, F., and Bowler, C. (2022) Reconstructing Dynamic Evolutionary Events in Diatom Nuclear and Organelle Genomes. In *The Molecular Life of Diatoms*. Falciatore, A., and Mock, T. (eds). Cham: Springer International Publishing, pp. 147-177.

dos Reis, M., Savva, R., and Wernisch, L. (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**: 5036-5044.

Doudna, J.A., and Charpentier, E. (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**: 1258096.

Dunn, J.G., and Weissman, J.S. (2016) Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC Genomics* **17**: 958.

Duret, L. (2000) tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends in Genetics* **16**: 287-289.

Eastman, G., Smircich, P., and Sotelo-Silveira, J.R. (2018a) Following Ribosome Footprints to Understand Translation at a Genome Wide Level. *Computational and Structural Biotechnology Journal* **16**: 167-176.

Eastman, G., Smircich, P., and Sotelo-Silveira, J.R. (2018b) Following Ribosome Footprints to Understand Translation at a Genome Wide Level. *Comput Struct Biotechnol J* **16**: 167-176.

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

Fenton, D.A., Kiniry, S.J., Yordanova, M.M., Baranov, P.V., and Morrissey, J.P. (2022) Development of a ribosome profiling protocol to study translation in Kluyveromyces marxianus. *Fems Yeast Research* **22**.

Ferenczi, A., Pyott, D.E., Xipnitou, A., and Molnar, A. (2017) Efficient targeted DNA editing and replacement in Chlamydomonas reinhardtii using Cpf1 ribonucleoproteins and single-stranded DNA. *Proc Natl Acad Sci U S A* **114**: 13567-13572.

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.

Floris, M., Bassi, R., Robaglia, C., Alboresi, A., and Lanet, E. (2013) Post-transcriptional control of light-harvesting genes expression under light stress. *Plant Molecular Biology* **82**: 147-154.

Francois, P., Arbes, H., Demais, S., Baudin-Baillieu, A., and Namy, O. (2021) RiboDoc: A Docker-based package for ribosome profiling analysis. *Computational and Structural Biotechnology Journal* **19**: 2851-2860.

Frigerio, S., Campoli, C., Zorzan, S., Fantoni, L.I., Crosatti, C., Drepper, F. et al. (2007) Photosynthetic Antenna Size in Higher Plants Is Controlled by the Plastoquinone Redox State at the Post-transcriptional Rather than Transcriptional Level*. *Journal of Biological Chemistry* **282**: 29457-29469.

Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., Schumann, F. et al. (2012) Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res* **22**: 2208-2218.

Fu, L., Wang, P., and Xiong, Y. (2020) Target of Rapamycin Signaling in Plant Stress Responses. *Plant Physiol* **182**: 1613-1623.

Fu, W., Shu, Y., Yi, Z., Su, Y., Pan, Y., Zhang, F., and Brynjolfsson, S. (2022) Diatom morphology and adaptation: Current progress and potentials for sustainable development. *Sustainable Horizons* **2**: 100015.

Gebauer, F., and Hentze, M.W. (2004) Molecular mechanisms of translational control. *Nature Reviews Molecular Cell Biology* **5**: 827-835.

Gerashchenko, M.V., and Gladyshev, V.N. (2014) Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res* **42**: e134.

Gerashchenko, M.V., and Gladyshev, V.N. (2017) Ribonuclease selection for ribosome profiling. *Nucleic Acids Res* **45**: e6.

Gerashchenko, M.V., Lobanov, A.V., and Gladyshev, V.N. (2012) Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc Natl Acad Sci U S A* **109**: 17394-17399.

Gingold, H., and Pilpel, Y. (2011) Determinants of translation efficiency and accuracy. *Mol Syst Biol* **7**: 481.

Goss, R., and Jakob, T. (2010) Regulation and function of xanthophyll cycle-dependent photoprotection in algae. *Photosynth Res* **106**: 103-122.

Gotsmann, V.L., Ting, M.K.Y., Haase, N., Rudorf, S., Zoschke, R., and Willmund, F. (2023) Utilizing high resolution ribosome profiling for the global investigation of gene expression in Chlamydomonas</i>. In: bioRxiv.

Grainger, S., Lonquich, B., Oon, C.H., Nguyen, N., Willert, K., and Traver, D. (2017) CRISPR Guide RNA Validation In Vitro. *Zebrafish* **14**: 383-386.

Greiner, A., Kelterborn, S., Evers, H., Kreimer, G., Sizova, I., and Hegemann, P. (2017) Targeting of Photoreceptor Genes in Chlamydomonas reinhardtii via Zinc-Finger Nucleases and CRISPR/Cas9. *Plant Cell* **29**: 2498-2518.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**: 835-840.

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.B. et al. (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol* **17**: 148.

Halse, G.R., and Syvertsen, E.E. (1996) Chapter 2 - Marine Diatoms. In *Identifying Marine Diatoms and Dinoflagellates*. Tomas, C.R. (ed). San Diego: Academic Press, pp. 5-385.

Hanson, G., and Coller, J. (2018) Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol* **19**: 20-30.

Heidcamp, W. (1995) Cell Biology Laboratory Manual. In. Biology Department, Gustavus Adolphus College, St. Peter, MN 56082.

Heiman, M., Kulicke, R., Fenster, R.J., Greengard, P., and Heintz, N. (2014) Cell type-specific mRNA purification by translating ribosome affinity purification (TRAP). *Nat Protoc* **9**: 1282-1291.

Hershberg, R., and Petrov, D.A. (2008) Selection on codon bias. *Annu Rev Genet* **42**: 287-299.

Hershey, J.W., Sonenberg, N., and Mathews, M.B. (2012) Principles of translational control: an overview. *Cold Spring Harb Perspect Biol* **4**.

Hiom, K. (2000) Homologous recombination. *Current Biology* **10**.

Hiraoka, Y., Kawamata, K., Haraguchi, T., and Chikashige, Y. (2009) Codon usage bias is correlated with gene expression levels in the fission yeast Schizosaccharomyces pombe. *Genes Cells* **14**: 499-509.

Hopes, A., and Mock, T. (2015) Evolution of Microalgae and Their Adaptations in Different Marine Ecosystems. In *eLS*, pp. 1-9.

Hopes, A., Nekrasov, V., Kamoun, S., and Mock, T. (2016) Editing of the urease gene by CRISPR-Cas in the diatom Thalassiosira pseudonana. *Plant Methods* **12**: 49.

Hopes, A., Nekrasov, V., Belshaw, N., Grouneva, I., Kamoun, S., and Mock, T. (2017) Genome Editing in Diatoms Using CRISPR-Cas to Induce Precise Bi-allelic Deletions. *Bio-Protocol* **7**.

Hornstein, N., Torres, D., Das Sharma, S., Tang, G.M., Canoll, P., and Sims, P.A. (2016) Ligation-free ribosome profiling of cell type-specific translation in the brain. *Genome Biology* **17**.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V. et al. (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**: 827-832.

Hsu, P.Y., Calviello, L., Wu, H.L., Li, F.W., Rothfels, C.J., Ohler, U., and Benfey, P.N. (2016) Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proc Natl Acad Sci U S A* **113**: E7126-E7135.

Huang, W., and Daboussi, F. (2017) Genetic and metabolic engineering in diatoms. *Philos Trans R Soc Lond B Biol Sci* **372**.

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y. et al. (2014) iCLIP: Protein-RNA interactions at nucleotide resolution. *Methods* **65**: 274-287.

Huysman, M.J.J., Martens, C., Vandepoele, K., Gillard, J., Rayko, E., Heijde, M. 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.

Höijer, I., Johansson, J., Gudmundsson, S., Chin, C.S., Bunikis, I., Häggqvist, S. et al. (2020) Amplification-free long-read sequencing reveals unforeseen CRISPR-Cas9 off-target activity. *Genome Biol* **21**: 290.

Ikemura, T. (1981) Correlation between the Abundance of Escherichia coli Transfer RNAs and the Occurrence of the Respective Codons in its Protein Genes- A Proposal for a Synonymous Codon Choice that is Optimal for the E. coli Translational System. *J Mol Biol* **151**: 389-409.

Ikemura, T. (1982) Correlation Between the Abundance of Yeast Transfer RNAs and the Occurrence of the Respective Codons in Protein Genes: Differences in Synonymous Codon Choice Patterns of Yeast and Escherichiu coli with Reference to the Abundance of Isoaccepting Transfer RNAs. *J Mol Biol* **158**: 573-597.

Imig, J., Kanitz, A., and Gerber, A.P. (2012) RNA regulons and the RNA-protein interaction network. **3**: 403-414.

Ingolia, N.T. (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet* **15**: 205-213.

Ingolia, N.T. (2016) Ribosome Footprint Profiling of Translation throughout the Genome. *Cell* **165**: 22-33.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**: 789-802.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218-223.

Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7**: 1534-1550.

Jeffryes, C., Campbell, J., Li, H., Jiao, J., and Rorrer, G. (2011) The potential of diatom nanobiotechnology for applications in solar cells, batteries, and electroluminescent devices. *Energy & Environmental Science* **4**.

Ji, Z., Song, R., Regev, A., and Struhl, K. (2015) Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* **4**: e08890.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012) A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**.

Juntawong, P., and Bailey-Serres, J. (2012) Dynamic Light Regulation of Translation Status in Arabidopsis thaliana. *Front Plant Sci* **3**: 66.

Juntawong, P., Girke, T., Bazin, J., and Bailey-Serres, J. (2014) Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. *Proc Natl Acad Sci U S A* **111**: E203-212.

Järvi, S., Suorsa, M., and Aro, E.M. (2015) Photosystem II repair in plant chloroplasts--Regulation, assisting proteins and shared components with photosystem II biogenesis. *Biochim Biophys Acta* **1847**: 900-909.

Kage, U., Powell, J.J., Gardiner, D.M., and Kazan, K. (2020) Ribosome profiling in plants: what is not lost in translation? *J Exp Bot* **71**: 5323-5332.

Kang, S., Jeon, S., Kim, S., Chang, Y.K., and Kim, Y.C. (2020) Development of a pVEC peptide-based ribonucleoprotein (RNP) delivery system for genome editing using CRISPR/Cas9 in Chlamydomonas reinhardtii. *Sci Rep* **10**: 22158.

Kapp, L.D., and Lorsch, J.R. (2004) The molecular mechanics of eukaryotic translation. *Annu Rev Biochem* **73**: 657-704.

Karas, B.J., Diner, R.E., Lefebvre, S.C., McQuaid, J., Phillips, A.P., Noddings, C.M. et al. (2015) Designer diatom episomes delivered by bacterial conjugation. *Nat Commun* **6**: 6925.

Karentz, D., Cleaver, J.E., and Mitchell, D.L. (1991) Cell survival characteristics and molecular responses of Antarctic phytoplankton to ultraviolet-B radiation. *Journal of Phycology* **27**: 326-341.

Karlsen, J., Asplund-Samuelsson, J., Thomas, Q., Jahn, M., and Hudson, E.P. (2018) Ribosome Profiling of Synechocystis Reveals Altered Ribosome Allocation at Carbon Starvation. *Msystems* **3**.

Kilian, O., Benemann, C.S., Niyogi, K.K., and Vick, B. (2011) High-efficiency homologous recombination in the oil-producing alga Nannochloropsis sp. *Proc Natl Acad Sci U S A* **108**: 21265-21269.

Kim, D., Kang, B.C., and Kim, J.S. (2021) Identifying genome-wide off-target sites of CRISPR RNA-guided nucleases and deaminases with Digenome-seq. *Nat Protoc* **16**: 1170-1192.

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H.R. et al. (2015) Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat Methods* **12**: 237-243, 231 p following 243.

Kim, H., and Kim, J.-S. (2014) A guide to genome engineering with programmable nucleases. **15**: 321-334.

King, H.A., and Gerber, A.P. (2016) Translatome profiling: methods for genome-scale analysis of mRNA translation. *Brief Funct Genomics* **15**: 22-31.

Kiniry, S.J., Judge, C.E., Michel, A.M., and Baranov, P.V. (2021) Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. *Nucleic Acids Research* **49**: W662-W670.

Koester, J.A., Berthiaume, C.T., Hiranuma, N., Parker, M.S., Iverson, V., Morales, R. et al. (2018) Sexual ancestors generated an obligate asexual and globally dispersed clone within the model diatom species Thalassiosira pseudonana. *Scientific Reports* **8**: 9.

Krasovec, M., and Filatov, D.A. (2019) Evolution of Codon Usage Bias in Diatoms. *Genes (Basel)* **10**.

Kronja, I., Yuan, B., Eichhorn, S.W., Dzeyk, K., Krijgsveld, J., Bartel, D.P., and Orr-Weaver, T.L. (2014) Widespread changes in the posttranscriptional landscape at the Drosophila oocyte-to-embryo transition. *Cell Rep* **7**: 1495-1508.

Lareau, L.F., Hite, D.H., Hogan, G.J., and Brown, P.O. (2014) Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife* **3**: e01257.

Lauria, F., Tebaldi, T., Bernabo, P., Groen, E.J.N., Gillingwater, T.H., and Viero, G. (2018) riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data. *Plos Computational Biology* **14**.

Le, H., Browning, K.S., and Gallie, D.R. (1998) The Phosphorylation State of the Wheat Translation Initiation Factors eIF4B, eIF4A, and eIF2 Is Differentially Regulated during Seed Development and Germination*. *Journal of Biological Chemistry* **273**: 20084-20089.

Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., and Qian, S.-B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences* **109**: E2424-E2432.

Lei, L., Shi, J., Chen, J., Zhang, M., Sun, S., Xie, S. et al. (2015) Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. *Plant J* **84**: 1206-1218.

Lesueur, L.L., Mir, L.M., and André, F.M. (2016) Overcoming the Specific Toxicity of Large Plasmids Electrotransfer in Primary Cells In Vitro. *Mol Ther Nucleic Acids* **5**: e291.

Li, G.W., Oh, E., and Weissman, J.S. (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**: 538-541.

Li, S., Zhang, Y., Xia, L., and Qi, Y. (2020) CRISPR-Cas12a enables efficient biallelic gene targeting in rice. In *Plant Biotechnol J*, pp. 1351-1353.

Liang, S.Q., Liu, P., Smith, J.L., Mintzer, E., Maitland, S., Dong, X. et al. (2022) Genome-wide detection of CRISPR editing in vivo using GUIDE-tag. *Nat Commun* **13**: 437.

Lisondro, I., Gómez Serrano, C., Sepúlveda, C., Batista Ceballos, A.I., and Acién Fernández, F.G. (2022) Influence of irradiance on the growth and biochemical composition of Nitzschia aff. pellucida. *Journal of Applied Phycology* **34**: 19-30.

Litchman, E., and Klausmeier, C.A. (2008) Trait-Based Community Ecology of Phytoplankton. *Annual Review of Ecology, Evolution, and Systematics* **39**: 615-639.

Liu, B., Molinaro, G., Shu, H., Stackpole, E.E., Huber, K.M., and Richter, J.D. (2019) Optimization of ribosome profiling using low-input brain tissue from fragile X syndrome model mice. *Nucleic Acids Res* **47**: e25.

Liu, M.J., Wu, S.H., Wu, J.F., Lin, W.D., Wu, Y.C., Tsai, T.Y., and Tsai, H.L. (2013) Translational landscape of photomorphogenic Arabidopsis. *Plant Cell* **25**: 3699-3710.

Liu, Q., Shvarts, T., Sliz, P., and Gregory, R.I. (2020) RiboToolkit: an integrated platform for analysis and annotation of ribosome profiling data to decode mRNA translation at codon resolution. *Nucleic Acids Research* **48**: W218-W229.

Liu, Y., Yang, Q., and Zhao, F. (2021) Synonymous but Not Silent: The Codon Usage Code for Gene Expression and Protein Folding. *Annual Review of Biochemistry* **90**: 375-401.

Long, S.P., Humphries, S., and Falkowski, P.G. (1994) Photoinhibition of Photosynthesis in Nature. *Annual Review of Plant Physiology and Plant Molecular Biology* **45**: 633-662.

Lopez-Gomollon, S., Beckers, M., Rathjen, T., Moxon, S., Maumus, F., Mohorianu, I. et al. (2014) Global discovery and characterization of small non-coding RNAs in marine microalgae. *BMC Genomics* **15**.

Lynn, D.J., Singer, G.A., and Hickey, D.A. (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res* **30**: 4272-4277.

Mahfouz, M.M., Kim, S., Delauney, A.J., and Verma, D.P. (2006) Arabidopsis TARGET OF RAPAMYCIN interacts with RAPTOR, which regulates the activity of S6 kinase in response to osmotic stress signals. *Plant Cell* **18**: 477-490.

Mann, D.G., and Vanormelingen, P. (2013) An inordinate fondness? The number, distributions, and origins of diatom species. *J Eukaryot Microbiol* **60**: 414-420.

Manuell, A.L., Quispe, J., and Mayfield, S.P. (2007) Structure of the chloroplast ribosome: novel domains for translation regulation. *PLoS Biol* **5**: e209.

McGlincy, N.J., and Ingolia, N.T. (2017) Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**: 112-129.

McInerney, J.O. (1998) GCUA- General Codon Usage Analysis. *Bioinformatics* **14**.

McKim, S.M., and Durnford, D.G. (2006) Translational regulation of light-harvesting complex expression during photoacclimation to high-light in Chlamydomonas reinhardtii. *Plant Physiology and Biochemistry* **44**: 857-865.

Mehravar, M., Shirazi, A., Mehrazar, M.M., and Nazari, M. (2019) In Vitro Pre-validation of Gene Editing by CRISPR/Cas9 Ribonucleoprotein. *Avicenna J Med Biotechnol* **11**: 259-263.

Meindl, A., Romberger, M., Lehmann, G., Eichner, N., Kleemann, L., Wu, J. et al. (2022) A rapid protocol for ribosome profiling of low input samples. *bioRxiv*: 2022.2009.2023.509038.

Merchante, C., Stepanova, A.N., and Alonso, J.M. (2017) Translation regulation in plants: an interesting past, an exciting present and a promising future. *The Plant Journal* **90**: 628-653.

Michel, A.M., Mullan, J.P., Velayudhan, V., O'Connor, P.B., Donohue, C.A., and Baranov, P.V. (2016) RiboGalaxy: A browser based platform for the alignment, analysis and visualization of ribosome profiling data. *RNA Biol* **13**: 316-319.

Mock, T., Hodgkinson, K., Wu, T., Moulton, V., Duncan, A., van Oosterhout, C., and Pichler, M. (2022) Structure and Evolution of Diatom Nuclear Genes and Genomes. In *The Molecular Life of Diatoms*. Falciatore, A., and Mock, T. (eds). Cham: Springer International Publishing, pp. 111-145.

Mock, T., Otillar, R.P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J. et al. (2017) Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. *Nature* **541**: 536-540.

Mohammad, F., and Buskirk, A.R. (2019) Protocol for Ribosome Profiling in Bacteria. *Bio Protoc* **9**.

Mohammad, F., Green, R., and Buskirk, A.R. (2019) A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *Elife* **8**.

Moore, E.R., Bullington, B.S., Weisberg, A.J., Jiang, Y., Chang, J., and Halsey, K.H. (2017) Morphological and transcriptomic evidence for ammonium induction of sexual reproduction in Thalassiosira pseudonana and other centric diatoms. *PLOS ONE* **12**: e0181098.

Moosburner, M.A., Gholami, P., McCarthy, J.K., Tan, M., Bielinski, V.A., and Allen, A.E. (2020) Multiplexed Knockouts in the Model Diatom Phaeodactylum by Episomal Delivery of a Selectable Cas9. *Front Microbiol* **11**: 5.

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.

Muench, D.G., Zhang, C., and Dahodwala, M. (2012) Control of cytoplasmic translation in plants. *WIREs RNA* **3**: 178-194.

Mussgnug, J.H., Wobbe, L., Elles, I., Claus, C., Hamilton, M., Fink, A. et al. (2005) NAB1 Is an RNA Binding Protein Involved in the Light-Regulated Differential Expression of the Light-Harvesting Antenna of Chlamydomonas reinhardtii. *The Plant Cell* **17**: 3409-3421.

Muto, M., Fukuda, Y., Nemoto, M., Yoshino, T., Matsunaga, T., and Tanaka, T. (2013) Establishment of a genetic transformation system for the marine pennate diatom Fistulifera sp. strain JPCC DA0580--a high triglyceride producer. *Mar Biotechnol (NY)* **15**: 48-55.

Müller, P., Li, X.P., and Niyogi, K.K. (2001) Non-photochemical quenching. A response to excess light energy. *Plant Physiol* **125**: 1558-1566.

Naduthodi, M.I.S., Mohanraju, P., Südfeld, C., D'Adamo, S., Barbosa, M.J., and van der Oost, J. (2019) CRISPR-Cas ribonucleoprotein mediated homology-directed repair for efficient targeted genome editing in microalgae Nannochloropsis oceanica IMET1. *Biotechnol Biofuels* **12**: 66.

Nagao, R., Kato, K., Kumazawa, M., Ifuku, K., Yokono, M., Suzuki, T. et al. (2022) Structural basis for different types of hetero-tetrameric light-harvesting complexes in a diatom PSII-FCPII supercomplex. *Nature Communications* **13**: 1764.

Ndah, E., Jonckheere, V., Giess, A., Valen, E., Menschaert, G., and Van Damme, P. (2017) REPARATION: ribosome profiling assisted (re-)annotation of bacterial genomes. *Nucleic Acids Res* **45**: e168.

Nelson, D.M., Treguer, P., Brzezinski, M.A., Lewynaert, A., and Queguiner, B. (1995) Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochemical Cycle* **9**: 359-372.

Niu, Y.F., Yang, Z.K., Zhang, M.H., Zhu, C.C., Yang, W.D., Liu, J.S., and Li, H.Y. (2012) Transformation of diatom Phaeodactylum tricornutum by electroporation and establishment of inducible selection marker. *Biotechniques* **52**.

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.

Nymark, M., Valle, K.C., Brembu, T., Hancke, K., Winge, P., Andresen, K. et al. (2009) An integrated analysis of molecular acclimation to high light in the marine diatom Phaeodactylum tricornutum. *PLoS One* **4**: e7743.

Palasin, K., Uechi, T., Yoshihama, M., Srisowanna, N., Choijookhuu, N., Hishikawa, Y. et al. (2019) Abnormal development of zebrafish after knockout and knockdown of ribosomal protein L10a. *Sci Rep* **9**: 18130.

Pechmann, S., and Frydman, J. (2013) Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat Struct Mol Biol* **20**: 237-243.

Peng, Z., Zaher, H., and Ben-Shahar, Y. (2018) Natural selection on gene-specific codon usage bias is common across eukaryotes. *bioRxiv*.

Perkins, P., Mazzoni-Putman, S., Stepanova, A., Alonso, J., and Heber, S. (2019) RiboStreamR: a web application for quality control, analysis, and visualization of Ribo-seq data. *BMC Genomics* **20**: 422.

Plotkin, J.B., and Kudla, G. (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**: 32-42.

Pop, C., Rouskin, S., Ingolia, N.T., Han, L., Phizicky, E.M., Weissman, J.S., and Koller, D. (2014) Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol Syst Biol* **10**: 770.

Popa, A., Lebrigand, K., Paquet, A., Nottet, N., Robbe-Sermesant, K., Waldmann, R., and Barbry, P. (2016) RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing. *F1000Res* **5**: 1309.

Pospíšil, P. (2016) Production of Reactive Oxygen Species by Photosystem II as a Response to Light and Temperature Stress. *Frontiers in Plant Science* **7**.

Poulsen, N., and Kröger, N. (2004) Silica morphogenesis by alternative processing of silaffins in the diatom Thalassiosira pseudonana. *J Biol Chem* **279**: 42993-42999.

Poulsen, N., Chesley, P.M., and Kröger, N. (2006) Molecular Genetic Manipulation of the Diatom Thalassiosira Pseudonana (Bacillariophyceae). *Journal of Phycology* **42**: 1059-1065.

Power, L. (2022) Beginners guide to ribosome profiling. *The Biochemist* **44**: 30-34.

Price, N.M., Harrison, G.I., Hering, J.G., Hudson, R.J., Nirel, P.M., Palenik, B., and Morel, F.M. (1989) Preparation and Chemistry of the Artificial Algal Culture Medium Aquil. *Aquil Biol Oceanogr* **6**: 443 – 461.

Puigbo, P., Guzman, E., Romeu, A., and Garcia-Vallve, S. (2007) OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res* **35**: W126-131.

Qi, F., and Zhang, F. (2019) Cell Cycle Regulation in the Plant Response to Stress. *Front Plant Sci* **10**: 1765.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Reid, D.W., Shenolikar, S., and Nicchitta, C.V. (2015) Simple and inexpensive ribosome profiling analysis of mRNA translation. *Methods* **91**: 69-74.

Rezayian, M., Niknam, V., and Ebrahimzadeh, H. (2019) Oxidative damage and antioxidative system in algae. *Toxicology Reports* **6**: 1309-1313.

Robertson, A.B., Klungland, A., Rognes, T., and Leiros, I. (2009) DNA repair in mammalian cells: Base excision repair: the long and short of it. *Cell Mol Life Sci* **66**: 981-993.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011) Integrative genomics viewer. In *Nat Biotechnol*, pp. 24-26.

Rodriguez, C.M., Chun, S.Y., Mills, R.E., and Todd, P.K. (2019) Translation of upstream open reading frames in a model of neuronal differentiation. *BMC Genomics* **20**: 391.

Roller, M., Lucić, V., Nagy, I., Perica, T., and Vlahovicek, K. (2013) Environmental shaping of codon usage and functional adaptation across microbial communities. *Nucleic Acids Res* **41**: 8842-8852.

Ruiz-Orera, J., and Alba, M.M. (2019) Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation. *Trends Genet* **35**: 186-198.

Saier, M.H., Jr. (2019) Understanding the Genetic Code. *J Bacteriol* **201**.

Salussolia, C.L., Winden, K.D., and Sahin, M. (2022) Translating Ribosome Affinity Purification (TRAP) of Cell Type-specific mRNA from Mouse Brain Lysates. *Bio Protoc* **12**: e4407.

Sancar, A. (2003) Structure and Function of DNA Photolyase and Cryptochrome Blue-Light Photoreceptors. *Chemical Reviews* **103**: 2203-2238.

Sander, J.D., and Joung, J.K. (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol* **32**: 347-355.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**: 671-675.

Schröder-Lorenz, A., and Rensing, L. (1987) Circadian changes in protein-synthesis rate and protein phosphorylation in cell-free extracts of Gonyaulax polyedra. *Planta* **170**: 7-13.

Schuller, A.P., and Green, R. (2018) Roadblocks and resolutions in eukaryotic translation. *Nat Rev Mol Cell Biol* **19**: 526-541.

Sendoel, A., Dunn, J.G., Rodriguez, E.H., Naik, S., Gomez, N.C., Hurwitz, B. et al. (2017) Translation from unconventional 5' start sites drives tumour initiation. *Nature* **541**: 494-499.

Serif, M., Dubois, G., Finoux, A.L., Teste, M.A., Jallet, D., and Daboussi, F. (2018) One-step generation of multiple gene knock-outs in the diatom Phaeodactylum tricornutum by DNA-free genome editing. *Nat Commun* **9**: 3924.

Sesma, A., Castresana, C., and Castellano, M.M. (2017) Regulation of Translation by TOR, eIF4E and eIF2α in Plants: Current Knowledge, Challenges and Future Perspectives. *Frontiers in Plant Science* **8**.

Shalgi, R., Hurt, J.A., Krykbaeva, I., Taipale, M., Lindquist, S., and Burge, C.B. (2013) Widespread regulation of translation by elongation pausing in heat shock. *Mol Cell* **49**: 439-452.

Sharma, A.K., Nymark, M., Sparstad, T., Bones, A.M., and Winge, P. (2018) Transgene-free genome editing in marine algae by bacterial conjugation - comparison with biolistic CRISPR/Cas9 transformation. *Sci Rep* **8**: 14401.

Sharma, N., Simon, D.P., Diaz-Garza, A.M., Fantino, E., Messaabi, A., Meddeb-Mouelhi, F. et al. (2021) Diatoms Biotechnology: Various Industrial Applications for a Greener Tomorrow. *Frontiers in Marine Science* **8**.

Shields, D.C., and Sharp, P.M. (1987) Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases. *Nucleic Acids Res* **15**.

Simbriger, K., Amorim, I.S., Chalkiadaki, K., Lach, G., Jafarnejad, S.M., Khoutorsky, A., and Gkogkas, C.G. (2020) Monitoring translation in synaptic fractions using a ribosome profiling strategy. *Journal of Neuroscience Methods* **329**.

siTOOLsBiotech Pan-Plant riboPOOL. URL https://www.sitoolsbiotech.com/pdf/PanPlantFlyer2020.pdf

Soma, F., Takahashi, F., Yamaguchi-Shinozaki, K., and Shinozaki, K. (2021) Cellular Phosphorylation Signaling and Gene Expression in Drought Stress Responses: ABA-Dependent and ABA-Independent Regulatory Systems. *Plants (Basel)* **10**.

Sonenberg, N., and Hinnebusch, A.G. (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136**: 731-745.

Starck, S.R., Tsai, J.C., Chen, K., Shodiya, M., Wang, L., Yahiro, K. et al. (2016) Translation from the 5' untranslated region shapes the integrated stress response. *Science* **351**: aad3867.

Steitz, J.A. (1969) POLYPEPTIDE CHAIN INITIATION - NUCLEOTIDE SEQUENCES OF 3 RIBOSOMAL BINDING SITES IN BACTERIOPHAGE R17 RNA. *Nature* **224**: 957-&.

Straka, L., and Rittmann, B.E. (2018) Light-dependent kinetic model for microalgae experiencing photoacclimation, photodamage, and photodamage repair. *Algal Research* **31**: 232-238.

Sun, Y., and Zerges, W. (2015) Translational regulation in chloroplasts for development and homeostasis. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1847**: 809-820.

Tirichine, L., Rastogi, A., and Bowler, C. (2017) Recent progress in diatom genomics and epigenomics. *Current Opinion in Plant Biology* **36**: 46-55.

Tomoiaga, D., Bubnell, J., Herndon, L., and Feinstein, P. (2022) High rates of plasmid cotransformation in E. coli overturn the clonality myth and reveal colony development. *Scientific Reports* **12**: 11515.

Tonn, T., Ozadam, H., Han, C., Segura, A., Tran, D., Catoe, D. et al. (2021) Single cell quantification of ribosome occupancy in early mouse development. *bioRxiv*: 2021.2012.2007.471408.

Trösch, R., Barahimipour, R., Gao, Y., Badillo-Corona, J.A., Gotsmann, V.L., Zimmer, D. et al. (2018) Commonalities and differences of chloroplast translation in a green alga and land plants. *Nat Plants* **4**: 564-575.

Tsai, S.Q., Zheng, Z., Nguyen, N.T., Liebers, M., Topkar, V.V., Thapar, V. et al. (2015) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* **33**: 187-197.

Tuller, T., Girshovich, Y., Sella, Y., Kreimer, A., Freilich, S., Kupiec, M. et al. (2011) Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Res* **39**: 4743-4755.

Vass, I., Cser, K., and Cheregi, O. (2007) Molecular mechanisms of light stress of photosynthesis. *Ann N Y Acad Sci* **1113**: 114-122.

Verbruggen, S., Ndah, E., Van Criekinge, W., Gessulat, S., Kuster, B., Wilhelm, M. et al. (2019) PROTEOFORMER 2.0: Further Developments in the Ribosome Profiling-assisted Proteogenomic Hunt for New Proteoforms. *Mol Cell Proteomics* **18**: S126-s140.

Volpi e Silva, N., Angelotti-Mendonça, J., Grisoste Barbosa, E., Gargioni Girotto, L., Correa Molinari, M., Daiane Mertz-Henning, L., and Marcia Lima Nepomuceno, A. (2021) Genome editing by CRISPR/Cas via homologous recombination. In *CRISPR technology in plant genome editing: Biotechnology applied to agriculture*. Correa Molinari, H., Bruno Rios Vieira, L., Volpi e Silva, N., Souza Prado, G., and Lopes Filho, J.H. (eds). Brasília: Embrapa, pp. 89-118.

Wang, H., McManus, J., and Kingsford, C. (2016) Isoform-level ribosome occupancy estimation guided by transcript abundance with Ribomap. *Bioinformatics* **32**: 1880-1882.

Wang, W., Zhao, S., Pi, X., Kuang, T., Sui, S.-F., and Shen, J.-R. (2020) Structural features of the diatom photosystem II–light-harvesting antenna complex. *The FEBS Journal* **287**: 2191-2200.

Warner, J.R., and McIntosh, K.B. (2009) How common are extraribosomal functions of ribosomal proteins? *Mol Cell* **34**: 3-11.

Warner, J.R., Knopf, P.M., and Rich, A. (1963) A multiple ribosomal structure in protein synthesis. *Proc Natl Acad Sci U S A* **49**: 122-129.

Weber, E., Engler, C., Gruetzner, R., Werner, S., and Marillonnet, S. (2011) A Modular Cloning System for Standardized Assembly of Multigene Constructs. *PLoS ONE* **6**.

Webster, C., Gaut, R.L., Browning, K.S., Ravel, J.M., and Roberts, J.K. (1991) Hypoxia enhances phosphorylation of eukaryotic initiation factor 4A in maize root tips. *Journal of Biological Chemistry* **266**: 23341-23346.

Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B., and Bartel, D.P. (2016) Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep* **14**: 1787-1799.

Wienert, B., Wyman, S.K., Richardson, C.D., Yeh, C.D., Akcakaya, P., Porritt, M.J. et al. (2019) Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. *Science* **364**: 286-289.

Wilhelm, B.T., and Landry, J.R. (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**: 249-257.

Wilson, D.N., and Doudna Cate, J.H. (2012) The structure and function of the eukaryotic ribosome. *Cold Spring Harb Perspect Biol* **4**.

Withers, J., and Dong, X. (2017) Post-translational regulation of plant immunity. *Current Opinion in Plant Biology* **38**: 124-132.

Wu, H.-Y.L., Song, G., Walley, J.W., and Hsu, P.Y. (2019) Translational landscape in tomato revealed by transcriptome assembly and ribosome profiling. *bioRxiv*.

Xiao, Z., Huang, R., Xing, X., Chen, Y., Deng, H., and Yang, X. (2018) De novo annotation and characterization of the translatome with ribosome profiling data. *Nucleic Acids Res* **46**: e61.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**: 134.

Yu, C.H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M.S., and Liu, Y. (2015a) Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol Cell* **59**: 744-754.

Yu, G., Fadrosh, D., Goedert, J.J., Ravel, J., and Goldstein, A.M. (2015b) Nested PCR Biases in Interpreting Microbial Community Structure in 16S rRNA Gene Sequence Datasets. *PLoS One* **10**: e0132253.

Zhang, C., and Hu, H. (2014) High-efficiency nuclear transformation of the diatom Phaeodactylum tricornutum by electroporation. *Mar Genomics* **16**: 63-66.

Zhang, L., Chen, L., Lu, F., Liu, Z., Lan, S., and Han, G. (2020) Differentially expressed genes related to oxidoreductase activity and glutathione metabolism underlying the adaptation of Phragmites australis from the salt marsh in the Yellow River Delta, China. *PeerJ* **8**: e10024.

Zhang, S., Shen, J., Li, D., and Cheng, Y. (2021) Strategies in the delivery of Cas9 ribonucleoprotein for CRISPR/Cas9 genome editing. *Theranostics* **11**: 614-648.

Zhang, Y., Xiao, Z., Zou, Q., Fang, J., Wang, Q., Yang, X., and Gao, N. (2017) Ribosome Profiling Reveals Genome-wide Cellular Translational Regulation upon Heat Stress in Escherichia coli. *Genomics Proteomics Bioinformatics* **15**: 324-330.

Zhao, F., Yu, C.H., and Liu, Y. (2017) Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. *Nucleic Acids Res* **45**: 8484-8492.

Zhu, S.H., and Green, B.R. (2010) Photoprotection in the diatom Thalassiosira pseudonana: role of LI818-like proteins in response to high light stress. *Biochim Biophys Acta* **1797**: 1449-1457.

Zill, J.C., Kansy, M., Goss, R., Alia, A., Wilhelm, C., and Matysik, J. (2019) (15)N photo-CIDNP MAS NMR on both photosystems and magnetic field-dependent (13)C photo-CIDNP MAS NMR in photosystem II of the diatom Phaeodactylum tricornutum. *Photosynth Res* **140**: 151-171.

Zinshteyn, B., Wangen, J.R., Hua, B.Y., and Green, R. (2020) Nuclease-mediated depletion biases in ribosome footprint profiling libraries. *Rna* **26**: 1481-1488.

Appendix

7.1 Appendix A: Structure and evolution of diatom nuclear genes and genomes

Structure and Evolution of Diatom Nuclear Genes and Genomes

Thomas Mock, Kat Hodgkinson, Taoyang Wu, Vincent Moulton, Anthony Duncan, Cock van Oosterhout, and Monica Pichler

Abstract

Diatoms are one of the most successful eukaryotes. There are over 100,000 diatom species contributing nearly half of total algal abundance in the oceans. Diatoms have conquered almost all aquatic environments, with high abundance especially in coastal and polar oceans and inland waters. The first diatom genomes provided important insights into their genetic, metabolic, and morphological diversity, which is unmatched by any other algal class. However, the recent application of long-read sequencing in addition to population genomics and culture-independent approaches enables a step-change in our understanding of diatom genomes. This chapter synthesizes what we have learned about the structure and evolution of diatom nuclear genes and genomes since the genome of *Thalassiosira pseudonana* became available in 2004. We highlight some of the key findings and discuss mechanisms and drivers of diatom genome evolution and adaptation underpinning the success of the entire class. Considering that most of their genomic diversity is still unknown, large-scale genome projects and culture-independent methods such as metagenome-assembled and single-cell-amplified genomes hold great promise to reveal more of their inter- and intraspecific genomic diversity in an environmental context. Data from these studies will pave the way for novel insights into their genetic versatility, which will enable us to identify the key evolutionary innovations in diatoms, and their adaptive evolution to a wide variety of environments, including to some of the most extreme aquatic environments on Earth such as intertidal zones and polar oceans.

T. Mock (✉) · K. Hodgkinson · C. van Oosterhout · M. Pichler
School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, UK
e-mail: t.mock@uea.ac.uk

T. Wu · V. Moulton · A. Duncan
School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, UK

A. Falciatore, T. Mock (eds.), *The Molecular Life of Diatoms*,
https://doi.org/10.1007/978-3-030-92499-7_5

These insights are not only critical for advancing diatom-based biotechnology and synthetic biology, but will also improve our knowledge about how the various diatom lineages perform their important roles as key players for capturing CO_2 and as the foundation of diverse aquatic food webs, thus providing significant ecosystem services and maintaining the continued habitability on Earth.

Abbreviations

BAC	Bacterial Artificial Chromosome
DAE	Differential Allelic Expression
EGT	Endosymbiotic Gene Transfer
GO	Gene Ontology
HGT	Horizontal Gene Transfer
ISIPs	Iron Stress-Induced Proteins
JGI	Joint Genome Institute
lincRNAs	long intergenic non-coding RNAs
MAG	Metagenome-assembled genome
MGT	Metagenomics-based transcriptome
ncRNAs	non-coding RNAs
ONT	Oxford Nanopore Technology
ORF	Open Reading Frame
PacBio	Pacific Biosciences
PUFAs	Polyunsaturated Fatty Acids
SAG	Single-Amplified Genome
SMRT	Single-Molecule Real-Time
sRNAs	small non-coding RNAs

1 Introduction

Diatom genomics is a relatively young field, which has commenced by the publication of the genome of *Thalassiosira pseudonana* in 2004 (Armbrust et al. 2004). It was the first genome from the diverse group of photosynthetic stramenopiles, which represent one of the major lineages of eukaryotes (Burki et al. 2020). As it was also the first genome from a marine alga, *T. pseudonana* has been frequently used as a reference to study not only diatom-specific biology but also to address wider questions concerning microbial biodiversity and biotechnology, the role of endosymbiosis for the evolution of life on Earth, and for revealing intricacies of how phytoplankton orchestrate global biogeochemical cycles (Mock et al. 2008; Ashworth et al. 2013; Kustka et al. 2014; Delalat et al. 2015; Benoiston et al. 2017; Chen et al. 2018; Treguer et al. 2018). Today, approximately 17 years later, diatom research has significantly matured and not only leads the field of marine algal

research (Falciatore et al. 2020), but is on par with research using the 'green yeast' *Chlamydomonas reinhardtii* (Chlorophyta), which is a long-standing model for plant biology and biotechnology (Sasso et al. 2018). Although other algal groups such as haptophytes, dinoflagellates, and cryptophytes are nearly equally important considering their ecology and role in the evolution of eukaryotic life, only a few genomes from these groups are currently available, and the development of reverse genetics tools for experimental cell biology is still in its infancy. Reasons for the diatom success story can be found throughout this book; they are manifold, but most of them depend on the availability of diatom genomes, easy cultivation of diatom species under laboratory conditions, and the availability of diverse methods for their exploitation to address questions from ecology to biotechnology.

While multiple candidates were initially explored for genomic understanding of diatoms, research quickly focussed around two representative species, the centric diatom *Thalassiosira pseudonana* and the pennate diatom *Phaeodactylum tricornutum*. The latter was published only a few years later in 2008 (Bowler et al. 2008). Both genomes were used for comparative genomics and transcriptomics to provide first results on the evolution, ecology, and metabolic diversity of diatoms (Nisbet et al. 2004; Montsant et al. 2005, 2007; Dyhrman et al. 2012; Veluchamy et al. 2013; Levitan et al. 2015; Rastogi et al. 2018). Remarkable insights have been revealed, shedding light on some key features of diatom biology including the synthesis of their nanopatterned silica cell walls (Mock et al. 2008; Shrestha et al. 2012), the significance of the urea cycle (Allen et al. 2011), the acquisition of genes from bacteria and eukaryotic sources (Vancaester et al. 2020; Dorrell et al. 2021), and the role of transposable elements for driving metabolic versatility (Maumus et al. 2009), just to name a few. Thus, these genomes were used like a library to retrieve relevant information for addressing questions from different fields of biological research including molecular ecology and evolution, physiology, metabolism, and reverse genetics.

For instance, diverse gene expression studies not only revealed how genes were regulated under relevant growth conditions, but they were also used to identify physiological markers for ecological studies with natural diatom populations such as ISIPs (iron stress-induced proteins) (Marchetti et al. 2012; Caputi et al. 2019). Genes reporting on diverse nutrient limitations (e.g. nitrate reductase) were used to assess the physiological state of natural communities across environmental gradients (Bender et al. 2014; Alipanah et al. 2015; Amato et al. 2017; Lampe et al. 2018; Cohen et al. 2019), which provided deeper insights into how natural diatom communities respond to changes of environmental conditions. These early studies also contributed to extending the known functional diversity of diatom genes as the first diatom genomes were used as important references for identifying novel gene variants from natural diatom communities through either amplicon sequencing or metatranscriptomics.

Furthermore, the first diatom genomes were invaluable for mapping out the metabolism responsible for the success of diatoms in diverse ecosystems and under changing environmental conditions (Kroth et al. 2008; Rosenwasser et al. 2014; Kim et al. 2016; Levering et al. 2016). The main focus was on identifying

metabolic pathways involved in carbon acquisition (Kroth et al. 2008), the synthesis of lipids (Sayanova et al. 2017), signalling (Helliwell et al. 2021), vitamin auxotrophy (Helliwell et al. 2011), and cell-cycle progression (Huysman et al. 2010; Kim et al. 2017). Diatom genomes were used to identify genes involved in the response to nutrient limitations with strong emphasis on silicate, nitrate, and iron metabolism (Allen et al. 2008; Mock et al. 2008; Shrestha et al. 2012; Alipanah et al. 2015). Knowledge on the gene content in combination with genome-wide expression patterns revealed the first metabolic maps and metabolic pathway models (Fabris et al. 2012; Singh et al. 2015; Gruber and Kroth 2017; Levering et al. 2017). For details on the latter, please see "Constraint-based Modeling of Diatoms Metabolism and Quantitative Biology Approaches".

Once a minimal set of cyanobacterial genomes and diverse genomes from the algal tree of life became available, it was possible to reconstruct the evolutionary mosaicism of diatom genomes as the outcome of primary and secondary endosymbiosis (Moustafa et al. 2009; Prihoda et al. 2012; Benoiston et al. 2017; Dorrell et al. 2017). In particular, the first genomes from marine green and red algae (Matsuzaki et al. 2004; Worden et al. 2009; Bhattacharya et al. 2013; van Baren et al. 2016), which were published shortly after the genome of *T. pseudonana*, provided a stepchange in our understanding of how endosymbionts have shaped diatom genomes through gene acquisitions. One of the most remarkable discoveries, which is still strongly debated to date, are the genetic traces of a cryptic green-algal endosymbiont predating the acquisition of a red alga (Moustafa et al. 2009; Deschamps and Moreira 2012; Dorrell et al. 2017). The latter was known before diatom genomes became available because the plastid genome in diatom is derived from a red-algal endosymbiont. But only through the genomic lens and the availability of algal genomes from descendants of endosymbionts from the group of Archaeplastida (red and green algae), many genes were discovered in extant diatom genomes likely being of green- and red-algal origin (Moustafa et al. 2009; Dorrell et al. 2017). As there is no remnant organelle representing this green-algal endosymbiont, these genes were only discovered once the first green-algal genomes became available, i.e., *Ostreococcus tauri* and Micromonas species (Derelle et al. 2006; Worden et al. 2009).

Despite many taxonomic and phylogenetic studies have revealed the macro- and microevolution of diatoms, diatom research has only recently begun to reveal mechanisms responsible for genetic variations between and especially within a species (Koester et al. 2018; Rastogi et al. 2020), due to the availability of more diatom genomes (Lommer et al. 2012; Tanaka et al. 2015; Traller et al. 2016; Basu et al. 2017; Mock et al. 2017; Osuna-Cruz et al. 2020) and the use of the latest sequencing and assembly technologies. The latter mechanisms are likely to significantly drive the evolution of hyperdiversity (global number of species (species richness) differs by $\geq$ one order of magnitude compared to other classes of algae) in the class of diatoms. Thus, addressing the fundamental question as to how the forces of evolution have shaped the structure of diatom genomes as the consequence of natural selection is relevant for our understanding of the extraordinary diversity of diatoms, their evolvability, and their adaptability (Pinseel et al. 2020). Their

widespread distribution especially in coastal waters, subpolar and polar oceans, and in freshwater ecosystems accompanied by their hyperdiversity suggests a significant level of evolvability unseen in other microbial eukaryotes (Nakov et al. 2018). How this evolvability has evolved (e.g. mechanisms underpinning the evolution of evolvability) and how it contributes to the genomic and phenotypic plasticity in the class of diatoms are unknown, but these questions are at the heart of revealing fundamental mechanisms responsible for the success of diatoms.

This chapter not only provides basic information on the structure and evolution of diatom genomes, but it also critically reflects on how novel sequencing technologies (e.g. Oxford Nanopore, PacBio HiF, 10X Genomics) and bioinformatics tools (e.g. Hi-C-guided genome assemblies) have shaped and sometimes even revised our understanding of diatom genomes. Furthermore, it addresses the important question of evolutionary mechanisms, and it provides an outlook for diatom genomics. It is likely that the latter will be significantly shaped by culture-independent approaches and single-cell sequencing, both of which are still in their infancy, but these methods hold great promise in terms of revealing more of diatoms inter- and intraspecific genomic diversity in an environmental context (Delmont et al. 2020; Duncan et al. 2020).

2 The Basic Structure of Diatom Genes and Genomes

2.1 Genes

Coding genes in diatoms have a characteristic eukaryotic structure, although their length is relatively short due to the compactness of diatom genomes (e.g. Armbrust et al. 2004; Bowler et al. 2008; Basu et al. 2017; Mock et al. 2017). Approximately 50% of the average diatom genome space is occupied by protein-coding genes. There are a number of notable exceptions of species that possess a significant amount of repeats in the non-coding part of their genomes, such as in *Cyclotella cryptica* (Traller et al. 2016; Roberts et al. 2020) and *Fragilariopsis cylindrus* (Mock et al. 2017) (Table 1). For example, the repeat content of *C. cryptica* is 53%, and repeats have significantly increased the genome size (161.7 Mb) of this species. So far, no other diatom genome has been characterized by such a high repeat content, and in

Table 1. Genome properties of selected diatom genomes. (1) Armbrust et al. 2004; (2) Bowler et al. 2008; (3) Lommer et al. 2012; (4) Mock et al. 2017; (5) Tanaka et al. 2015; (6) Traller et al. 2016; (7) Osuna-Cruz et al. 2020; (8) Basu et al. 2017.

Species	*Thalassiosira pseudonana*[1]	*Phaeodactylum tricornutum*[2]	*Thalassiosira oceanica*[3]	*Fragilariopsis cylindrus*[4]	*Fistulifera solaris*[5]	*Cyclotella cryptica*[6]	*Seminavis robusta*[7]	*Pseudo-nitzschia multistriata*[8]
Genome Size	32 Mbp	27 Mbp	92 Mbp	61 Mbp	25 Mbp	171 Mbp	126 Mbp	59 Mbp
Repeats	≤ 2 %	≤ 6 %	N.D.	≤ 38 %	≤ 16 %	≤ 59 %	≤ 23 %	≤ 25 %
GC	48 %	51 %	53%	40 %	46 %	42 %	48 %	46 %
Gene count	11776	10402	10109	21066	11448	21250	36254	12008
Ploidy	Diploid	Diploid	Diploid	Partial Triploid	Allodiploid	Diploid	Diploid	Diploid
Haplotype Diversity (Polymorphisms)	Homozygous (≤ 1 %)	Homozygous (≤ 1 %)	N.D.	Heterozygous (≤ 6 %)	Heterozygous (≤ 38 %)	N.D.	Homozygous (≤ 1 %)	Homozygous (≤ 1 %)

general, repeats do not appear to have inflated the genome size of diatoms as much as that in dinoflagellates (Stephens et al. 2020). The relatively short coding genes in diatom genomes (mean gene length $\leq$ 2500 bps) have on average a single intron and two exons (Basu et al. 2017). Introns are usually being spliced out by the canonical eukaryotic splicing machinery. However, in some diatom species, the introns of genes are retained in mature mRNAs, and this alternative splicing mode is known as 'intron retention'. Over 95% of diatom genes possess canonical splice sites (Acceptor Sites: AG/CT and Donor sites: GT/AC), but only less than 20% of genes contain more than one intron. Intron retention and exon skipping have been observed and are assumed to contribute to the diversity of the protein space, facilitating phenotypic plasticity in a dynamic aquatic environment. In *Phaeodactylum tricornutum*, ca. 24% of genes undergo intron retention and ca. 20% exon skipping (Rastogi et al. 2018). A small percentage of genes (<15%) perform both, resulting in alternative splicing. Genes that undergo intron retention are more highly expressed than other genes, which has been observed previously in green algae. Interestingly, intron retention appears to be more common under stress conditions such as nutrient starvation, and especially for genes involved in mitigating stress. Possibly, intron retention helps to alleviate stress conditions, thereby contributing to a faster recovery once favourable conditions have resumed. This could support the typical boom and bust growth cycle of many diatom species. An alternative explanation is that intron retention is non-adaptive and simply a by-product of environmentally induced stress (Rastogi et al. 2018). Future research should examine the adaptive significance of intron retention, for example by comparing the population growth rates under stress conditions in lines with low and high intron retention.

The upstream non-coding parts of protein-coding diatom genes are variable in length due to alternative transcription and translation start sites. The latter usually is indicated by the conserved ATG start codon. Promoter analyses with *T. pseudonana* and *F. cylindrus* revealed a common motifs containing tandem repetition of the CAA triplet such as CAACAA and its derivatives (ACAACA, AACAAC) (Ashworth et al. 2013; Mock et al. 2017). Variable repeats in promoter regions may contribute to phenotypic plasticity due to either modifying the affinity of specific transcription factors or their diversity. If this is the case, they contribute to differential gene expression underpinning phenotypic plasticity (Ashworth et al. 2013; Mock et al. 2017). However, there are also genes with more canonical promoter-binding motifs such as the TATA box (Hopes et al. 2016). mRNAs of nuclear genes usually possess poly-A-tails at the 3′-prime end to increase stability of the transcript for RNA processing. Diatoms cannot only produce sense but also antisense transcripts (e.g. Dyhrman et al. 2012; de Carvalho and Bowler 2020). For instance, non-coding natural antisense transcripts (NATs) appear to be under tight regulation by nutrient and environmental stresses (Dyhrman et al. 2012; de Carvalho and Bowler 2020; George et al. 2020). In *P. tricornutum*, NATs cover 21.5% of annotated coding space. Thus, sense and antisense transcriptions appear to be common in diatom genomes, with canonical stop codons usually terminating the transcription process (e.g. Dyhrman et al. 2012; de Carvalho and Bowler 2020; George et al. 2020).

In addition to protein-coding genes and their non-coding antisense transcripts, there is evidence for a diverse set of intergenic non-coding regulatory genes in diatom genomes (e.g. Lopez-Gomollon et al. 2014; Rogato et al. 2014; de Carvalho et al. 2016). RNA sequencing revealed the presence of long and small non-coding RNAs (ncRNAs) originating from diverse loci. Long ncRNAs likely derive mostly from intergenic loci, producing long intergenic non-coding RNAs (lincRNAs) (e.g. de Carvalho et al. 2016; Basu et al. 2017). They appear to be highly responsive to P stress in *P. tricornutum* and homologs have been found in other diatoms and plants (de Carvalho et al. 2016), suggesting their importance for coping with P-stress. Their average length is significantly shorter than protein-coding mRNAs, and they either have no open reading frame (ORF), or only one single ORF. Most of them are intron-less (de Carvalho et al. 2016). Their specific function is still unclear but the limited information available suggests that they play regulatory roles in gene expression and/or translation as some of them were observed to act as precursors of short non-coding RNAs. However, an integrated proteome analysis revealed that at least some of the identified lincRNAs in *P. tricornutum* have coding potential as peptides were identified to be matching lincRNA ORFs (Yang et al. 2018). Equally uncertain is the identification and origin of small non-coding RNAs (sRNAs) in diatoms (Lopez-Gomollon et al. 2014; Rogato et al. 2014). Some of them appear to be originating from repeat regions, others from tRNAs and lincRNAs. Although early reports suggested the presence of canonical microRNAs in diatoms (Huang et al. 2011; Norden-Krichmar et al. 2011), experimental approaches were unable to validate any of the predicted canonical microRNAs at least for *T. pseudonana* (Lopez-Gomollon et al. 2014). tRNA-derived sRNAs appear to be the most abundant form of small non-coding RNAs in diatoms where they can contribute up to 20% of all sRNAs (Lopez-Gomollon et al. 2014). As many of them are differentially expressed under different growth conditions, they might be used for regulating translation to acclimatize to changing growth conditions. Regulatory mechanisms for non-coding RNAs in diatoms are still to be identified. For instance, it is largely unknown how the expression of non-coding genes is regulated, and there is only very preliminary information available how endoribonucleases such as Dicer might be involved in the process of their cleavage (e.g. De Riso et al. 2009; Rogato et al. 2014).

2.2 Nuclear Genomes

To generate the first diatom nuclear genomes at the beginning of this century was a significant methodological challenge. The only sequencing approach available for these genomes was shotgun Sanger sequencing complemented with BAC (Bacterial Artificial Chromosome) and FOSMID libraries to improve long-range contiguity and to provide haplotype information. *T. pseudonana* (Armbrust et al. 2004) and *P. tricornutum* (Bowler et al. 2008) were the first two diatom genomes sequenced by these approached at the Joint Genome Institute (JGI, DOE, USA). JGI provided

funding, technical and bioinformatics support to realize these pioneering sequencing projects. Both species were selected because of their small genome size (< 50 Mbp), rapid growth (>2 cell divisions per day) and extensively characterized physiology. Furthermore, *T. pseudonana* is a model for diatom cell-wall biology and biochemistry (e.g. Sumper and Brunner 2008; Hildebrand et al. 2018), and this centric species is a representative of the Thalassiosirales, an ecologically important and diverse diatom order (Malviya et al. 2016; Branco-Vieira et al. 2020). The pennate diatom *P. tricornutum* was established as a model for cell biology, biochemistry and reverse genetics (Bowler et al. 2008). However, it is an unusual diatom species because it has no absolute requirement for Si, but it grows well under different laboratory settings and without bacteria in co-culture. The latter two properties likely contributed to the rise of *P. tricornutum* as a model species. It is also the main species of the diatom biotechnology industry focussing on alternative fuels and high-value end products, such as essential polyunsaturated fatty acids (PUFAs) (e.g. Daboussi et al. 2014; Branco-Vieira et al. 2020; George et al. 2020). Considering the hyperdiversity of the class (Nakov et al. 2018), comprising ≥100,000 species, the availability of two diatom genomes was only the beginning of revealing the secrets of the molecular life of diatoms. Most of the subsequent genome projects were based on second and third generation sequencing technologies (e.g. Illumina, PacBio, Oxford Nanopore), which provided deeper insights into genome structure and diversity (Mock et al. 2017; Osuna-Cruz et al. 2020). Although resequencing of *T. pseudonana* and *P. tricornutum* has confirmed their diploid structure with a relatively low level of polymorphisms between the two haplotypes (Table 1) (Koester et al. 2018; Rastogi et al. 2020), other diatoms genomes are either comprised of more haplotypes, such as *Fistulifera solaris* (Tanaka et al. 2015), and/or significantly diverged haplotypes, such as *Fragilariopsis cylindrus* (Mock et al. 2017). Triploidy may play a role in the latter species, which only recently has been identified by combining Illumina with Oxford Nanopore sequencing and a haplotype-specific assembly strategy. Together with k-mer spectra, these data revealed that two out of three sub-genomes (haplotypes) were highly identical (up to 100% sequence identity), imposing challenges to differentiate them (Fig. 1). However, diverged alleles and extended genomic loci between the two most diverged haplotypes could still be identified (Mock et al. 2017). A small number of genomic loci even show divergence between all three haplotypes (Fig. 1). Unlike whole genome duplication, which is not uncommon in diatoms and which is considered to have significantly contributed to speciation (Parks et al. 2018), the allopolyploid genome of the coastal marine diatom *Fustulifera solaris* appears to be a consequence of introgressive hybridization, which has led to two sets of pseudo-parental sub-genomes (Tanaka et al. 2015). Furthermore, a near chromosome-scale assembly of the *F. solaris* genome provided first sequence based evidence of potential aneuploidy in diatom genomes (Maeda et al. 2021). Aneuploidy can arise from errors in chromosome segregation, leading to an abnormal number of chromosomes in a cell (Compton 2011).

Although allopolyploidy is prevalent in plants, it has not been reported before in algae. Transcriptome profiling with *F. solaris* revealed that both sub-genomes

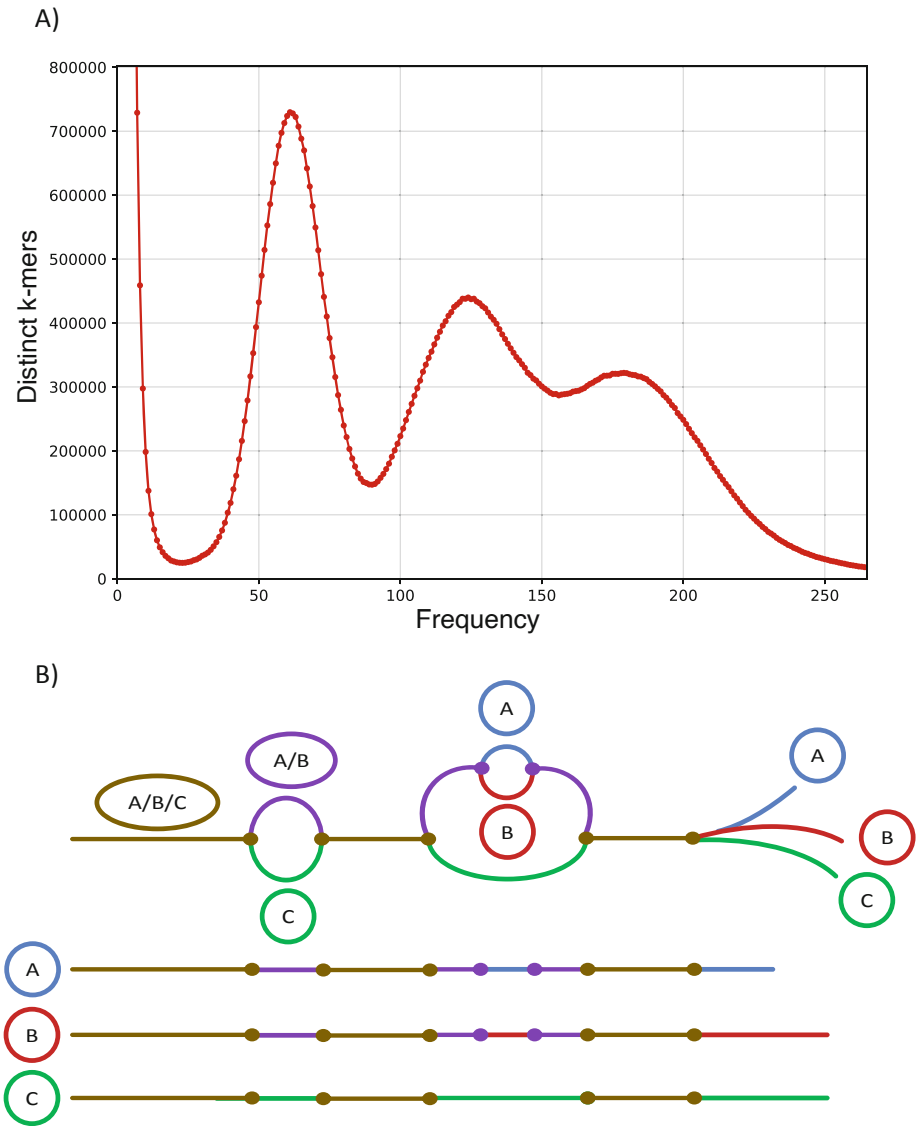


Fig. 1 (**a**) (i) The k-mer spectra for F. cylindrus CMP1102 shows three distinct distributions. Each distribution represents content that occurs once, twice and thrice, respectively, with the latter two occurring at harmonic frequencies to the first. The first distribution consists of unique content where the sub-genomes are diverged, while the third distribution consists of conserved content amongst all sub-genomes. (**b**) A de Bruijn graph of the assembly contains areas of varying coverage, as reflected in the k-mer spectra. Areas conserved between all sub-genomes (**a**, **b**, **c**) contain triple the coverage than the unique components. Where sub-genomes diverge, "bubbles" are formed. These may form between areas where 2 sub-genomes are very similar (**a**, **b**) opposing a unique sub-genome section (**c**). They may also occur in unique areas of the subgenomes (**a**, **b**, **c**), forming a "double-bubble". Unique content may diverge from double-coverage content or may diverge straight from triple-coverage content

contributed to global gene expression although the majority (61%) of homoeologous alleles were not equally expressed but showed an expression bias towards specific conditions (Nomaguchi et al. 2018). Differences in the promoter regions of homoeologous alleles might have contributed to this expression bias. A slightly more complex and not yet completely resolved genome structure was observed in the polar diatom *F. cylindrus* (Fig. 1) (Mock et al. 2017). The genome of *F. cylindrus* is characterized by significant allelic divergence between haplotypes for about 30% of the genome. Most of the polymorphisms have been identified in non-coding regions upstream of transcription start sites. Similar to *F. solaris*, diverged promoter regions appear to drive the differential expression of allelic pairs due to differences in binding affinities of transcription factors. Subsequent genome projects (e.g. *Skeletonema marinoi* (https://albiorix.bioenv.gu.se/Skeletonema_marinoi.html) confirmed that haplotype divergence is more common in diatoms than previously anticipated based on the genomes of *T. pseudonana* and *P. tricornutum*.

Taken together, our understanding of the structure of diatom genes and genomes has significantly altered over the last decade because of: a) sequencing more (non-model) diatom species, and b) increasingly advanced sequencing technologies and genome assembly tools, enabling the inclusion of long-read data resulting in (near) chromosome-level assemblies (Fig. 1). The latter has also been generated for *T. pseudonana* and *P. tricornutum*, which increased the contiguity of the assembly to the chromosome level from telomere to telomere compared to the original Sanger-based assemblies. However, the biggest revelations include the recent discovery of haplotype divergence in *F. cylindrus* and *F. solaris* and its influence on allelic expression bias (Mock et al. 2017; Hoguin et al. 2021). These insights were gained by sequencing species from different habitats using new sequencing and assembly approaches. The next steps could include the 'geography' of chromosomes to reveal if different parts of the chromatin between telomeres contribute differently to haplotype divergence and potentially the loss of heterozygosity underpinning adaptive processes to different habitats. Furthermore, how different levels of ploidy impact the structure of diatom genomes and the expression of alleles remains enigmatic. This is a question that needs to be addressed urgently in future genome projects, as there is mounting evidence that ploidy is an important driver of diatom evolution, adaption and speciation (e.g. Nakov et al. 2018; Parks et al. 2018). Currently, the most significant large-scale genome project addressing these questions is the '100 Diatom Genomes Project' funded by JGI (https://jgi.doe.gov/csp-2021-100-diatom-genomes/). A comparative analysis of 100 genomes from carefully selected diatom species representative of their structural, metabolic and evolutionary diversity covering the diatom-tree-of-life (Fig. 2) is likely to reveal novel insights into their genetic versatility, which will enable us to differentiate between what makes a diatom a diatom and what has evolved to underpin specific metabolic demands.

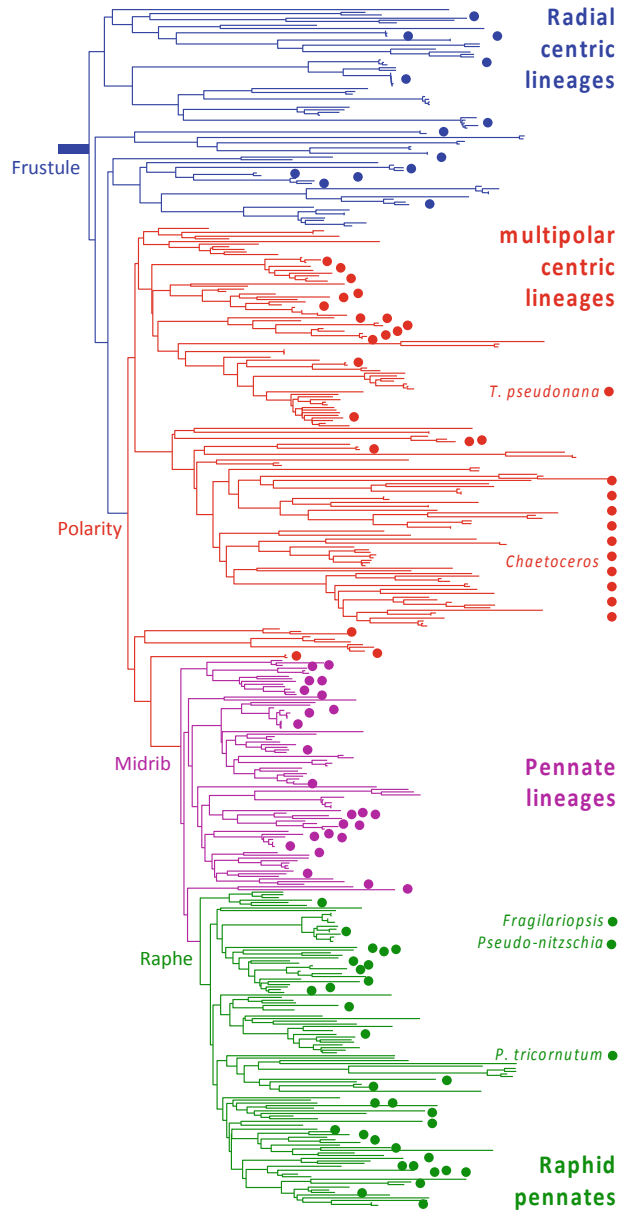


Fig. 2 18S rDNA diatom phylogeny. Dots indicate strains selected for whole-genome sequencing as part of the "100 Diatom Genomes Project"; dots to the right, strains already sequenced or in preparation. Features to the left: key acquisitions (Courtesy of Wiebe Kooistra)

3 The Evolutionary Mosaicism of Diatom Genomes Driven by Endosymbiotic and Horizontal Gene Transfer

3.1 Evolutionary Mosaicism

The era of phylogenomic studies with diatoms began once there were not only several diatom genomes and transcriptomes available but also genomes from the monophyletic supergroup of Archaeplastida, composed of primary plastid-bearing lineages, i.e., green and red algae, and the glaucophytes (e.g. Nisbet et al. 2004; Grossman 2005; Worden et al. 2009; Price et al. 2012). Genomes from these lineages provided the foundation for reconstructing major evolutionary events in eukaryotes, which possess complex plastids evolved by the engulfment of either red or green algae. These endosymbiotic gene transfer (EGT) events have left behind footprints in the genomes of extant eukaryotic lineages such as diatoms, as evidenced by gene loss and the transfer of genes from the endosymbiont to the host genome (Fig. 3) (e.g. Timmis et al. 2004; Li et al. 2006; Ponce-Toledo et al. 2019). Consequently, the genomes of diatoms can be considered a puzzle built by pieces from different sources acquired successively and over long periods of time (≥1 billion years) (e.g. Benoiston et al. 2017; Brodie et al. 2017). To identify the origin of each of these pieces still remains a major challenge as the ravages of time have had their

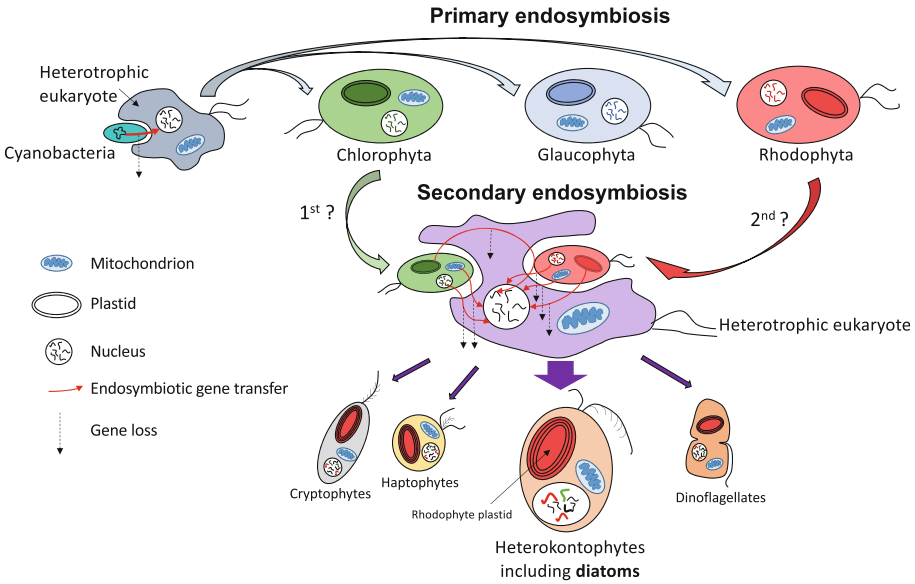


Fig. 3 The evolution of diatoms through primary and secondary endosymbiosis. Highly debated is the process by which a heterotrophic eukaryote acquired an endosymbiont from the group of Archaeplastida (Chlorophyta, glaucophyte, rhodophyta). Phylogenomics provided evidence of a cryptic endosymbiotic event with an ancient chlorophyte predating the acquisition of a red alga. The latter turned into the plastid of extant heterokontophytes including diatoms. Adapted from Hopes and Mock 2015

impact on the integrity of the puzzle pieces. This is especially the case for diatoms due to an ongoing controversy regarding a significant number of nuclear genes (> 1500) supposedly derived from a cryptic green algal endosymbiont (Moustafa et al. 2009). Although there is strong evidence that plastids in extant diatom species are derived from red algae, these 'green genes' have not only been identified in nuclear genomes of several diatoms, but also in their ancestors where they can contribute up to 25% of nucleus-encoded plastid targeted proteins (Dorrell et al. 2017). As to how these evolutionary processes have shaped the evolution of plastids in diatoms, please see "Reconstructing Dynamic Evolutionary Events in Diatom Nuclear and Organelle Genomes". These data suggest a massive genetic mosaicism in diatom genomes likely driven by successive endosymbiotic events with the green algal endosymbiont as being acquired prior to the acquisition of the red algal endosymbiont (Fig. 3) (Hopes and Mock 2015). High-frequency horizontal gene transfer (HGT) has been discussed and potential tree reconstruction artefacts to explain the high number of these cryptic 'green genes' in diatom genomes (e.g. Deschamps and Moreira 2012). Thus, the origin of the mosaicism in diatom genomes is still being debated, including the relative importance HGT and EGT. A global approach encompassing all genomes of potential 'donor' organisms responsible for the mosaicism of diatom genomes might help addressing this fundamental question.

However, as gene acquisition events took place over a period of more than a billion years, the endosymbiotic footprints likely have become eroded due to the evolutionary forces, such as mutation, genetic drift, gene flow, selection and recombination. Nonetheless, current data suggest that up to 20% of coding potential in diatom genomes has been derived from genomes of former endosymbionts, which is slightly more than in other major algal lineages (e.g. Moustafa et al. 2009). However, significant uncertainties still exist. For instance, undersampling and undersequencing of putative donors leave us with significant knowledge gaps in terms of the nature of the donors and their contribution to endosymbiosis and with respect to discriminating EGT from non-EGT derived genes.

While genes were transferred from the genomes of the endosymbionts to the genome of the host, they were either lost or undergoing significant modifications. Those genes that were retained are mostly enriched in photosynthesis-related processes, such as the synthesis of photosystem subunits, pigments and other processes essential for plastid maintenance (e.g. plastid division). However, recent data suggest that plastids from endosymbionts would not be viable without reprogramming and retargeting of host genes (Dorrell et al. 2017). For instance, plastid proteomes are composed of proteins that are not derived from endosymbionts, suggesting that a significant amount of genes of non-endosymbiont origin in the host nuclear genome have undergone structural modification to target their proteins to the plastids. In diatoms, this mainly is based on modifications at their 5′-prime ends such the evolution of specific plastid-targeting motifs, and sequences for signal and transit peptides. Thus, the acquisition of an endosymbiont had an impact on the structure of at least a subset of host genes and likely also their regulation as EGT contributed to the expansion of the redox sensing capabilities of the host. Those secondarily acquired genes that were not required for the maintenance of the plastid or other

organelles such as the mitochondria, contribute to the complexity of diatom metabolism by increasing the diversity and reticulate evolution of isoforms as described in "Reconstructing Dynamic Evolutionary Events in Diatom Nuclear and Organelle Genomes". Many of these isoforms are part of complex genetic networks underpinning core metabolism such as amino acid and lipid metabolism in diatoms (e.g. Benoiston et al. 2017; Brodie et al. 2017; Dorrell et al. 2017).

3.2 Horizontally Acquired Genes

Although the acquisition of genes via horizontal gene transfer (HGT) has been studied intensively in prokaryotes, the significance and especially the quantitative contribution of genes acquired via HGT in eukaryotes are still controversial (e.g. Van Etten and Bhattacharya 2020). However, the question is not so much if there is HGT in microbial eukaryotes but how much. Thus, the extent of which, and the differences in HGT between species, remains a subject of ongoing investigation. Identifying HGT in microbial eukaryotes such as diatoms is significantly more challenging compared to prokaryotes, and this is assumed to be the main reason for sometimes significant differences in estimates of HGT in the evolution of microbial eukaryotes (e.g. Bowler et al. 2008; Vancaester et al. 2020). Currently, challenges to estimate the contribution of HGT in eukaryotic microbial genomes are based on: a) genome size and complexity (e.g. repeats, heterozygosity, polyploidy), b) contaminants such as bacteria and viruses, and c) co-assembly of the contaminating DNA with the target DNA. The latest research on HGT in microbial eukaryotes suggests that horizontally acquired genes do not contribute more than 1.5% of the complete gene inventory, i.e., reflecting 'the 1% rule' (e.g. Van Etten and Bhattacharya 2020). This estimate is based on long-read sequencing and assembly-free approaches (Rossoni et al. 2019). Thus, genes are only considered to be of horizontal origin if they are physically linked with native genes on the same single read, which can be $\geq$50 kbps long in case of Oxford Nanopore, 10X Genomics or PacBio HiFi sequencing (e.g. Jain et al. 2018). This approach minimizes issues with incorporating DNA from contaminants in assemblies derived from short-read data. In particular, genomes with a considerable number of repeats suffer from fragmentation and co-assembly of contaminating reads, especially in conjunction with reads representing repeats (e.g. Schmid et al. 2018). If contaminating reads have similar GC content and k-mer frequency, they likely will be co-assembled. To avoid co-assembly, only reads from foreign organisms (HGT) as part of long reads should be accepted in addition to applying bioinformatics pipelines to remove contaminating sequences based on sequence similarity, read coverage and GC content (e.g. Fierst and Murdock 2017; Lu and Salzberg 2018). For instance, resequencing of the *Cyclotella cryptica* genome using a combination of long reads (MinION, Oxford Nanopore) and high-quality sort reads (Illumina HiSeq4000) in addition to applying 'BlobTools' removed up to 20% of genes in version 1 of the genome, which were considered to have been acquired via HGT (Roberts et al. 2020). If axenic cultures for the resequencing project would have been

used, it is likely that even more of the contaminating sequences would have been identified and therefore removed. However, some diatom species are known to live in a mutual relationship with bacteria (e.g. Amin et al. 2012; Monnich et al. 2020), which would require single-cell sequencing to avoid the inclusion of contaminants into the diatom genome assemblies. Although removing contaminating reads from diatom genome assemblies can be considered essential for estimating how HGT contributed to the evolutionary adaptation of diatoms, efforts to do so in the community of diatom researchers are still in their infancy. So far, the most commonly used approach to identify HGT events is to filter based on bootstrap support in phylogenetic trees (e.g. Bowler et al. 2008; Vancaester et al. 2020). However, depending on the cut-off used (e.g. 60%–80%), this may lead to misestimation and therefore false positives or false negatives (Van Etten and Bhattacharya 2020). As most diatom genomes so far have been assembled with short-read data because they were sequenced before long-read sequence technologies became available, it is likely that current estimates of how much HGT contributed to the evolution of diatoms will need to be revised (Roberts et al. 2020). Nevertheless, there is no doubt that HGT significantly contributed to the evolution of diatoms, but the suspicion is that the actual number might be much lower than currently estimated.

With our current tools, we are able to dissect and study some of the more recent HGT events. Indeed, if the time of their acquisition can be traced back before the split of centric and pennate diatoms (~90 million years ago), the signature of HGT is easier to unmask. Examples include genes essential for the urea cycle (e.g. carbamate kinase and ornithine cyclodeaminase) and nitrogen storage (e.g. allantoin synthase) (e.g. Allen et al. 2011; Vancaester et al. 2020). As most of them appear to be under purifying selection, they play important roles in central metabolism shared by the majority of diatom species (Vancaester et al. 2020). More recently acquired genes ($\leq$90 million years) from prokaryotes, viruses, or even other microbial eukaryotes appear to underpin specific adaptations required by diatom species that share a similar ecological niche (e.g. Raymond and Kim 2012; Nelson et al. 2021). However, we cannot exclude the possibility that our current tools only enable us to study those more recent HGT events, and that as science and technology advances, we will discover many more ancient HGT events. An example of recently acquired HGT genes is related to those of the family of ice-binding proteins (e.g. Janech et al. 2006; Sorhannus 2011). All of them appear to have been acquired from either cold-adapted bacteria or fungi as they convey freezing tolerance, a key trait required to thrive in polar ecosystems. Thus, this case of convergent evolution provides an example of how environmental conditions facilitated HGT in diatoms and therefore their ability to extend their global biogeographical distribution.

Recent large-scale microalgal genome sequencing encompassing all major groups (e.g. chlorophytes, haptophytes, bacillariophytes) provided some clues as to how virus genes contributed to the evolution of diatom genomes, potentially conferring niche-specific fitness benefits (Nelson et al. 2021). To identify genes acquired from viruses, this study identified virus-specific protein families (VFAMs) in microalgal genomes. After contamination screening, one of the main results was that VFAMs were enriched in marine microalgae and specifically diatoms and other

classes of the Ochrophyta (Photosynthetic stramenopiles). A negative correlation between the number of repeats and VFAMs in algal genomes suggests that less complex diatom genomes potentially benefit from viral genes, or that they can tolerate them better. Many of these VFAMs in marine species had elevated ratios of dN/dS indicating at least relaxed purifying selection. As a significant number of them were involved in membrane integrity, maybe they conferred halotolerance and therefore contributed to the evolutionary adaptation of diatoms to conditions of saltwater habitats (Nelson et al. 2021). An alternative 'neutralist' explanation is that during colonization of saltwater habitats, the early colonizers had not been exposed to the native marine viruses. These were able to invade the diatom genomes of the early colonizers, resulting in genetic hitchhiking at a genomic level.

Taken together, diatom genomes are a complex mix and match of genes not only from exo- and endosymbionts (EGT) due to their intertwined vertical evolution, but also from very distantly related species via horizontal gene transfer (HGT). Whereas the former appears to have provided basic toolsets for core metabolism, genes acquired via HGT seem to have conferred habitat-specific fitness benefits that enabled diatoms to conquer specific environments. Once (axenic) culturing techniques, long-read sequencing, bioinformatic assembly and QC methods have been optimized, we will be able to address questions about the nature of the HGT and EGT genes, whether the rate of HGT and EGT differed between species and between the ecological niche they are occupying.

4 Mechanisms and Drivers of Diatom Genome Evolution and Adaptation

Historically, the evolution of diatoms has been studied in the context of systematics and taxonomy reaching back to the pioneers in the late seventeenth century (e.g. Antoni van Leeuwenhoek) who likely discovered them by just using bead-like lenses (Lane 2015). However, it was the use of molecular tools and specifically the discovery of phylogenetic marker genes such as 18S that revolutionized our understanding of diatom evolution together with their fossilized remnants (Medlin et al. 1988). Although single-gene-based phylogenies provided first insights into the complexities of diatom evolution, the availability of the first algal genomes enabled us to extend this knowledge to all genes in their genomes and therefore to reconstruct the evolutionary origins of diatom metabolism underpinning their characteristic biology. However, compared to other fields of research, such as plant and animal sciences, population genetics and especially population genomics with diatoms is still in its infancy (e.g. Godhe and Rynearson 2017; Mock et al. 2017; Rengefors et al. 2017; Whittaker and Rynearson 2017; Koester et al. 2018; Parks et al. 2018; Postel et al. 2020; Rastogi et al. 2020). However, revealing how populations evolve is essential because some of the most fundamental questions (e.g. drivers of diversification and adaptation) can only be addressed if we understand how the evolutionary forces of mutation, recombination, selection, gene flow, and genetic drift shape genetic variation within and between species. For instance, we have insights into the

macroevolutionary diversity of diatoms but only a poor understanding of genetic variations within a diatom species. However, this knowledge is required to reveal how diatom genomes in a population change according to the evolutionary forces imposed by biotic and abiotic pressures.

One of the most fundamental properties of nuclear genomes in eukaryotic organisms is their size and level of ploidy. Both are linked as endoreduplication increases the DNA content of a cell, which can lead to an increase in cell volume (Connolly et al. 2008). Polyploidization has been shown to lead to reproductive isolation and eventually speciation (e.g. Koester et al. 2010; Parks et al. 2018). Thus, it potentially is an important mechanism driving speciation and therefore might contribute to the hyper-diversity in the group of diatoms, which is the youngest of all eukaryotic phytoplankton groups (Nakov et al. 2018). Estimates based on karyotyping suggest that chromosome counts range over several orders of magnitude, with flow cytometer measurements largely confirming these results with respect to genome size (Kociolek and Stoermer 1989). However, the mechanisms of chromosome fission and fusion, and whole or partial genome duplication, are not well understood (e.g. Parks et al. 2018). Although we assume the pennate diatom *F. solaris* has evolved allodiploidy based on hybridization events in distant parental lineages (Tanaka et al. 2015), our understanding of the frequency of these events and their evolutionary success is still limited (Amato and Orsini 2015). Usually, significant hybrid viability is seen with autopolyploidy, which is often caused by meiotic non-reduction (Mann 1994; von Dassow et al. 2008).

Recent phylogenomic approaches based on 37 diatom transcriptomes have shed light onto the importance of polyploidization in the group of diatoms. For instance, Parks et al. (2018) estimated the age distributions of duplicated genes. In combination with phylogenetically based reconciliation methods and gene counts, the authors showed that allopolyploidy as observed in *F. solaris* maybe as important as autopolyploidy. This study provided strong evidence for ancient allopolyploid events (>100 Myr) in the thalassiosiroid and pennate diatom clades. Although this work provides strong evidence that whole genome duplications have significantly contributed to the genome evolution in diatoms, their course sampling and macroevolutionary approach did not address intraspecific variation in genome size and ploidy as drivers of speciation. To gain insights into those processes requires a micro-evolutionary (i.e. population genomic) approach.

An excellent population genomic diatom model is *D. brightwellii* (e.g. Rynearson et al. 2006; Koester et al. 2010). This coastal species consists of distinct populations, some of which with a global distribution, and others only found locally in coastal embayments or estuaries. High F_{ST} (Proportion of the total genetic variance contained in a subpopulation (the S subscript) relative to the total genetic variance (the *T* subscript) values (can range from 0 to 1) suggest that they have been reproductively isolated for considerable time, and hence, *D. brightwellii* may actually represent a species-complex or meta-species. Interestingly, one local population is assumed to have diverged by a whole genome duplication, as is evidenced by its marked difference in genomes size (Koester et al. 2010). Furthermore, this population has a distinct phenotype, showing a slightly larger cell diameter. This population

co-exist with another *D. brightwellii* genotype at a single geographic location throughout a seasonal cycle. These observations suggest that this biodiversity can be maintained by ecological selection, and not only by geographic partitioning. Similar results have recently been found for the endemic Southern Ocean diatom *Fragilariopsis kerguelensis,* which can dominate phytoplankton communities in ice-free surface waters. Postel et al. (2020) identified three different genotypes across a transect in the Southern Ocean. Although two of them were separated geographically into a northern and southern genotype, the third genotype was omnipresent but reproductively isolated. Thus, diatom biodiversity can be maintained both by geographic isolation, as well as by reproductive isolation maintained across a large environmental envelope. Another recent study with *T. rotula* (Whittaker and Rynearson 2017) provided evidence that temporal genetic variation, as shown for *D. brightwellii* at a single geographical location, can be as similar as genetic variability observed over global distances (>10,000 km). These examples provide evidence that geographic structuring, local adaptations and environmental heterogeneity can all result in reproductive isolation, and that these are majors drivers for the genetic and genomic structure of diatom populations. Together, these processes can contribute to radiation and eventual speciation of diatoms, explaining the rich biodiversity of this taxon.

To identify the biological processes under selection, and therefore the mechanisms of diatom genome evolution and speciation, several research groups have begun to identify structural variations in diatom genomes from different populations (e.g. Koester et al. 2018; Osuna-Cruz et al. 2020; Rastogi et al. 2020). With whole diatom genomes available, we can start to identify how environmental conditions impact the evolution of genomes and therefore the divergence of populations as a consequence of adaptive evolution. Environmental and ecological variations are significant for generating and maintaining the diversity of diatoms (e.g. Whittaker and Rynearson 2017). Considering that temperature is both a strong selecting agent for microbes including diatoms (e.g. Thomas et al. 2012) and the environmental variable that changes most quickly, diatom populations might respond with significant changes in their structure and diversity with potential knock-on effects for aquatic food webs and biogeochemical cycles they drive.

The genome of the cold-adapted diatom *Fragilariopsis cylindrus* provided fundamental insights into how the environment drives changes in the structure of diatom genomes (Mock et al. 2017). The strong selection pressure imposed by the environmental conditions of the Southern Ocean likely contributed to a rare evolutionary mechanism of adaptation that has subsequently been confirmed in other diatom genome projects. The genome of *F. cylindrus* is characterized by approximately 29% of highly diverged alleles. The sequence divergence between alleles was up to 6%, a level of divergence that is typically observed only between the alleles of different species. Most remarkably, however, the most highly diverged alleles showed condition-specific differential expression (differential allelic expression, or DAE). Furthermore, the alleles appeared to have diverged by natural selection, rather than just by neutral evolution (i.e. genetic drift). A coalescence analysis showed that the majority of these alleles diverged shortly after the onset of the last glacial period,

which began about 110,000 years ago. The alleles of *F. cylindrus* appear to have adapted to accommodate different environmental conditions. In turn, DAE enables this diatom to express the adapted alleles under the right conditions, allowing it to thrive under the dramatically fluctuating conditions of the polar oceans.

Subsequent diatom genome projects including a re-sequencing project with *P. tricornutum* have confirmed that haplotype diversity and differential allelic expression appear to play a major role in the evolutionary adaptation of diatoms (Rastogi et al. 2020; Hoguin et al. 2021). Even the relatively homozygous genome of *P. tricornutum* (<1% polymorphisms) had 22% of genes with a moderate bias in allelic expression, and 1% of genes showed an almost complete monoallelic expression under different growth conditions (Hoguin et al. 2021). The most likely overall driver is the significant environmental variability of a 3-dimentional highly dynamic aquatic habitat in which they are suspended and which is being constantly modified by oceanic forces. Hence, a diversification of metabolism might provide a fitness advantage and therefore to outcompete competitors in an ever-changing environment. The latter has been known for a long time to be the preferred environment for diatoms because they primarily thrive in seasonally mixed surface ocean waters.

To test the hypothesis that environmental fluctuations drive the evolution of diverged alleles in diatoms and their differential expression, Tatman et al. (2021) developed a 2-D cellular automata model 'DAEsy-World' that builds on the classical Daisyworld model based on James Lovelock's Gaia theory of Earth as a self-regulating homeostatic system. (DAEsy-World was thus named to emphasize the differential allelic expression, DAE, in diatoms). This model quantified the effects of DAE on environmental homeostasis and examined whether DAE enables organisms to better adapt to strongly fluctuating environments. Using a gradient in the extent of temperature fluctuations, the model predicted that DAE increases the standing genetic variation, especially in sexually reproducing organisms. DAE allows for the build-up of genetic polymorphisms within genes driven by positive selection, and this can enhance phenotypic plasticity. This allows alleles to sub-functionalize and become adapted to different environmental or ecological niches. The model also showed that DAE is more likely to evolve under fluctuating environmental conditions, and extreme seasonal variation in the polar oceans may have led to the evolution of this remarkable adaptation (Tatman et al. 2021). Thus, the *in vivo* and *in silico* data suggest that DAE and the underlying allelic divergence is an evolutionary mechanism to adapt to a complex multidimensional adaptive landscape caused by significant spatiotemporal heterogeneity of the oceanic environment. This landscape can be better explored through the build-up of genetic diversity and phenotypic plasticity, both of which appears to prove a fitness advantage, which possibly contributes to the success of diatoms in these fluctuating environments.

Finally, the role of mitotic recombination and chromosomal rearrangements in diatom species is a subject that warrants further studies. The rapid population expansions of many diatom species are realized through periods with strict clonal reproduction (Chepurnov et al. 2004; Krueger-Hadfield et al. 2014). Small fitness differences between clonal lineage during this phase of exponential growth can have a big effect on the eventual population composition, and hence, the inclusive fitness

of each lineage. Given that environmental conditions fluctuate over time and space during such periods of clonal reproduction, the fitness of a clonal lineage is likely to deteriorate unless it can adapt to the changing environmental conditions. The epigenetic adaptation of DAE may accommodate some of this environmental change, but other genetic adaptations may also play a role. However, without sexual reproduction, each clonal lineage is constrained in its adaptive evolutionary response by the variation contained within its genome. A recent study suggests that mitotic recombination might be able to exploit the genetic variation that is present within each genome through mitotic recombination (i.e. gene conversion) (Bulankova et al. 2021). The mitotic recombination rate can make the genetic variation that is not available to selection (i.e. dominance and epistatic variation) available to natural selection, simply by replacing the dominant suppressor allele with a recessive variant. Other chromosomal rearrangements, possibly associated with transposable element activity (Schrader and Schmitz 2019), may also contribute to generating novel variation that is available to natural selection. This model draws a parallel with Nowell's hypothesis of clonal evolution in cancer cells (Nowell 1976) where chromosome instability results in cytogenetic heterogeneity, generating diversity that allows for the clonal expansion of increasingly aggressive tumour phenotypes (Chambers et al. 2002).

5 Genomes from Uncultured Diatom Species

5.1 Metagenome-Assembled Genomes (MAGs)

With $\geq$100,000 diatom species worldwide (Mann and Vanormelingen 2013), it will be challenging to assess their genomic diversity based on sequencing only isolated strains through the '100 diatom genomes project' (Fig. 2). Even so, this project will provide a step change in our understanding of diatom genomic diversity (Fig. 2) and an important reference dataset for novel culture-independent approaches based on natural diatom communities.

Metagenome-assembled genomes (MAGs) are a powerful approach to overcome the limitation from culture-dependent methods by recovering draft genomes directly from environmental samples. In this strategy, shotgun metagenomic sequencing reads are assembled into longer contigs and subsequently clustered into bins which represent contigs derived from the same taxa (Alneberg et al. 2018). Due to the smaller and simpler genomes of prokaryotic organisms and their high abundance in the ocean, most published MAGs have been assembled from bacteria and archaea (Hugerth et al. 2015; Tully et al. 2018). The reconstruction of diatom genomes is considerably more challenging due to their large and complex genomes characterized as mentioned above by, e.g., repeats, ploidy and genomic heterogeneity. Thus, diatom MAGs from shotgun metagenomic sequences are scarce and rarely complete. To the best of our knowledge, there are only two studies which contain diatom MAGs of medium quality or higher (Delmont et al. 2020; Duncan et al. 2020). Medium quality for a MAG is considered a completion of 50% or greater and

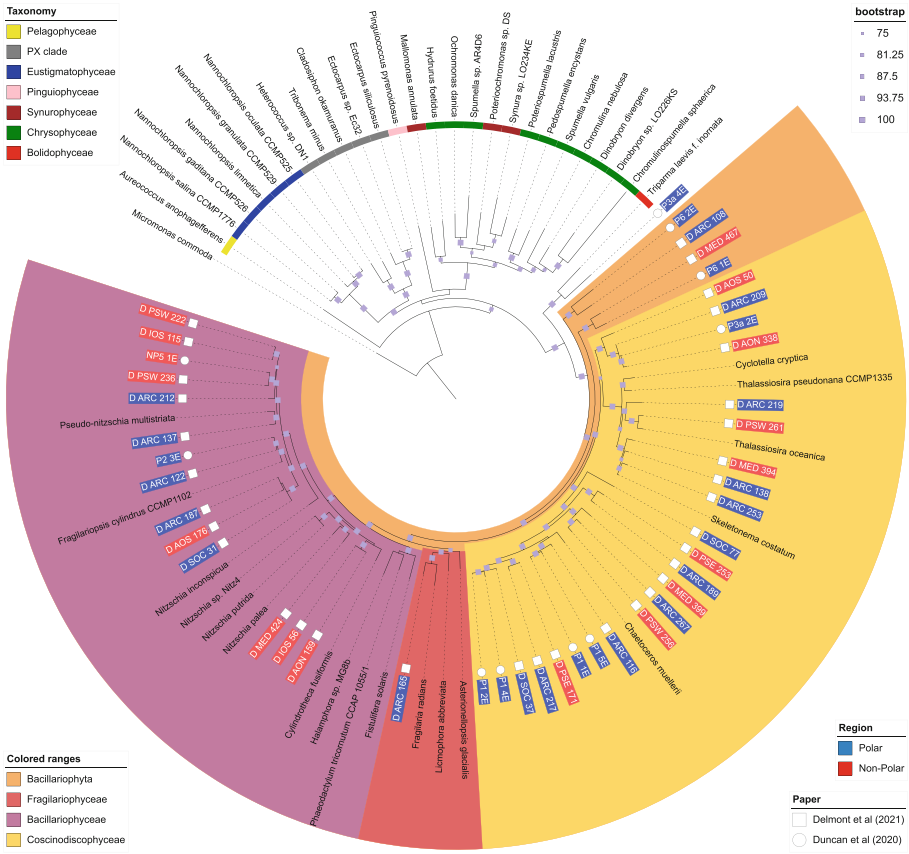


Fig. 4 Phylogenomic tree combining diatom MAGs from two sources (Delmont et al. 2020; Duncan et al. 2020), from the former non-redundant MAGs have been included, from the latter all MAGs. Tree was constructed using Phylosift (Darling et al. 2014) to identify and align marker genes from MAGs, and tree built using RaXML (Stamatakis 2014) with 100 bootstrap replicates. Leaf label colours indicate whether MAGs originated from Polar or Non-Polar data, blue and red respectively. The paper the MAG originates from is indicated by either a circle or star. The grayscale band below leaf labels shows the taxonomy assigned to the MAG in the originating paper. Violet squares show bootstrap values

contamination of 10% or less (Bowers et al. 2017). Binning of *Tara* Oceans data recovered 34 diatom MAGs of medium quality (Delmont et al. 2020); sequencing and binning of 12 Arctic and North Atlantic metagenomes resulted in the recovery of 9 medium quality diatom MAGs (Fig. 4) (Duncan et al. 2020). The highest completion among these MAGs is 87.1%, with an average completion of 68.78%. The longest is over 1 Gbp in length, though most are considerably smaller with a median size of 29.1 Mbp and 14,003 predicted genes. Combined these MAGs had 848,968 predicted genes: 730,506 and 118,480 respectively. Taxonomy could be assigned at a more specific level than phylum for some of these MAGs. Among the Tara Oceans

derived MAGs, genus was assigned to 28 of the 34 genomes, with the most common being *Chaetoceros*, *Fragilariopsis*, and *Pseudo-nitzschia*. Among the 9 MAGS from Duncan et al. (2020), 4 were assigned taxonomy at the rank of genus, from the genera *Fragilariopsis*, *Chaetoceros*, *Minutocellulus*, and *Leptocylindrus*.

An analogous gene-centred approach aims to identify metagenomics-based transcriptomes (MGT) by clustering genes which show similar patterns of abundance across metagenomic samples to identify transcriptomes of uncultured organisms (Vorobev et al. 2020). This approach identified 6 diatom MGTs with completion above 50%, although with quite high mean contamination of ca. 21%.

MAGs do not achieve the same level of quality as genomes sequenced from isolated cultures; however, they still advance our understanding of an organism's metabolic potential and ecology in a changing marine environment. A recent study has already shown this by combining MAG data with climate models to estimate biogeographical changes of eukaryotic plankton populations (Delmont et al. 2020). Gene Ontology (GO) enrichment analysis of the 9 MAGs from Duncan predicted a varying number of unique terms for each MAG, ranging from 2 (P1_4E) to 32 (NP5_1E) (Fig. 5a, b). All MAG-specific GO terms are displayed in detail in Fig. 5b. Thus, reconstructing MAGs can provide novel insight into diatoms' functional potential.

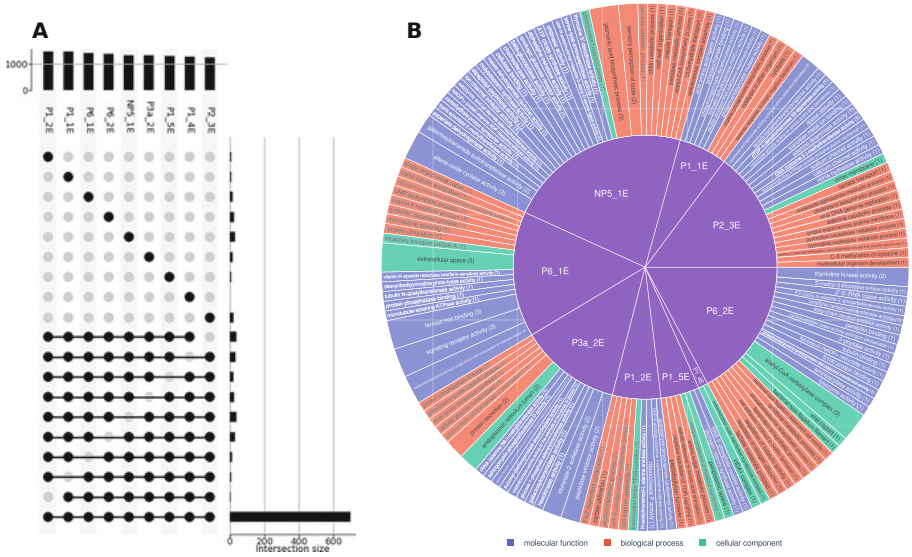


Fig. 5 An UpSet (Lex et al. 2014) showing number of Gene Ontology (GO) in a combination of MAGs from Duncan et al. (2020). Top bars indicate the number of GO terms observed in each MAG. Rows below relate to combinations of MAGs, and how many terms are unique to that combination. For example, the first row has a solid dot only for P1_2E, so the bar at the right shows how many GO terms occur only in that MAG. The final row correspondingly shows how many GO terms occur in all MAGs. The plot has been limited to show only a few combinations. B Gene Ontology (GO) terms which are unique to a MAG from Duncan et al. (2020). Inner ring shows MAG, and outer ring shows terms which are unique to that MAG

Sequencing of more diatom reference genomes will further improve the binning of diatom MAGs derived from environmental samples and open up the potential of this approach to exploring novel aspects of diatom metabolism and evolution, as well as the assessment of ecosystem changes.

5.2 Single-Amplified Genomes (SAGs)

Another state-of-the-art approach for recovering genomes from uncultivated microorganisms is the single-amplified genome strategy. In a nutshell, sorting of single cells is followed by whole-genome amplification, screening for SAGs and their subsequent sequencing (Kaster and Sobol 2020). As with MAGs, most SAGs are recovered from prokaryotes (Pachiadaki et al. 2019), yet only a few studies have exploited single-cell genomics to analyse the genome content eukaryotes, namely uncultured stramenopiles (Seeleuthner et al. 2018) and heterotrophic flagellated protists (Wideman et al. 2020). A recent study targeting small planktonic protists from Tara Ocean samples obtained more than 900 SAGs, of which 18 were assigned to diatoms using 18S rRNA gene identity. The low number of recovered diatom SAGs in this data set might be ascribed to inefficient cell lysis or a sorting bias which targeted cells <5 μm in size (Sieracki et al. 2019). Whole genome sequencing of a subset of these SAGs yielded three diatom genome assemblies with uneven completion rates ranging from ~7% to ~63% (Delmont et al. 2020). However, the SAG with the highest genome completion rate (~63%) was achieved by assembling 4 cells with identical 18S rRNA (Delmont et al. 2020).

Omics analyses alone are often not suited for recovering heterozygous genomes, which is a common strategy of microbial populations to respond to environmental stresses (Kaster and Sobol 2020). This is also seen in *Fragilariopsis cylindrus,* which shows high genomic heterogeneity with allele-specific expression across different environmental conditions (Mock et al. 2017). By capturing cell-to-cell heterogeneities, SAGs can overcome limitations of MAGs, which most likely represent consensus genomes derived from a mosaic of cells (Kaster and Sobol 2020). Comparative analysis from a collection of MAGs and SAGs concluded that both strategies provide accurate genomic information from uncultivated bacteria (Alneberg et al. 2018). While the level of genome completeness recovered by SAGs was consistently lower compared to MAGs from the same sample, the obtained functional gene categories were fairly similar (Alneberg et al. 2018). A study analysing choanoflagellate SAGs with moderate genomic completeness recovered phylogenetically-informative protein domains, indicating that SAG data can be used to place uncultured eukaryotic organisms within the tree of life while also revealing novel evolutionary insights (López-Escardó et al. 2017). The Darwin Tree of Life project launched in 2018 aims to sequence the genomes of 60,000 eukaryotic species in Britain and Ireland over the next 10 years (Darwin Tree of Life n.d.). This large-scale project will involve sequencing of 50 diatom species using single-cell genomics (personal communication). Phylogenetic analyses of these SAGs will greatly improve the evolutionary history of diatoms.

SAGs and MAGs generated from short-read sequencing are often constrained by repeat elements (Moss et al. 2020), resulting in highly fragmented genomes. Sequencing of large DNA fragments resolves repetitive elements and haplotypes by spanning across ambiguous regions and therefore substantially improves genomic assemblies (Laver et al. 2015; Moss et al. 2020). Long-read sequencing is currently dominated by Pacific Biosciences' (PacBio) single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies' (ONT) nanopore sequencing (Amarasinghe et al. 2020). One advantage of long-read sequencing is the detection of structural variants (e.g. insertions, deletions or duplications affecting $\geq$50 bp) which can greatly contribute to genome diversity (Amarasinghe et al. 2020). Thus, generating deep coverage diatom genomes with long reads would be beneficial. By including long-read sequences as well as metagenomic reads/contigs, weaknesses of single-cell genomics such as cell-sorting bias, chimera formation, and uneven read coverage can be mitigated and SAG assemblies can be greatly improved (Xu and Zhao 2018). Although assembling diatom SAGs is challenging, it can help identify novel metabolic and ecological functions, subspecies diversity, evolutionary histories, and host–virus interactions, and drive industrial bioprospecting (Labonte et al. 2015; Xu and Zhao 2018; Sieracki et al. 2019; Kaster and Sobol 2020).

6 Conclusions

The first genomes from diatoms enabled unprecedented insights into diatom biology and evolution. This discovery-based research also provided the blueprint for an interdisciplinary effort to study the oceans through the lens of genomics. A key driver is the fact that diatoms contribute approximately 40% of annual marine primary production and are thus keystone organisms for sustaining marine food webs and for driving global biogeochemical cycles of elements such as carbon. Diatom genomes have therefore served a wide community of researchers interested in organisms that can be considered models for Earth System Science as they have shaped our planet for millions of years and continue to do so. However, comparative evolutionary genomics with more diatom species from different environments and branches of the diatom-tree-of-life has made us realize that we are only at the beginning of revealing the molecular secretes of these remarkable organisms. Nevertheless, these pioneering studies have revealed the hyperdiversity of diatoms, laying bare our lack in understanding how this diversity has evolved and is maintained. The onset of large-scale diatom sequencing projects in combination with culture-independent sequencing approaches is appropriate to address these challenges. The aim of this novel work is to generate pan genomes for different diatom species, and possibly, a pan genome for the entire class. A pan genome is defined as a core genome containing genes present in all strains. It is distinct from the dispensable genome, which is composed of genes that may be absent from one or more strains. Thus, the pan genome captures the whole of the genetic content of a species, or even a class, if extended. Thus, it allows to establish a variant catalogue as a critical tool to study evolution and population genetics, and to identify species- and

taxon-specific adaptations that are essential for that group. This work requires sequencing technology that combines high throughput, accuracy and significant read length, all of which has recently become available. Thus, it seems inevitable that we soon will move beyond our reliance on a few diatom reference genomes from single strains, which are likely not representative for the most diverse group of algae. With the availability of pan genomes, we may even have a genomic lens that can be used to understand how these remarkable organisms support life on planet Earth.

References

Alipanah L, Rohloff J, Winge P, Bones AM, Brembu T (2015) Whole-cell response to nitrogen deprivation in the diatom Phaeodactylum tricornutum. J Exp Bot 66(20):6281–6296. https://doi.org/10.1093/jxb/erv340

Allen AE, Dupont CL, Obornik M, Horak A, Nunes-Nesi A, McCrow JP, Zheng H, Johnson DA, Hu HH, Fernie AR, Bowler C (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. Nature 473(7346):203. https://doi.org/10.1038/nature10074

Allen AE, LaRoche J, Maheswari U, Lommer M, Schauer N, Lopez PJ, Finazzi G, Fernie AR, Bowler C (2008) Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation. Proc Natl Acad Sci U S A 105(30):10438–10443. https://doi.org/10.1073/pnas.0711370105

Alneberg J, Karlsson CMG, Divne AM, Bergin C, Homa F, Lindh MV, Hugerth LW, Ettema TJG, Bertilsson S, Andersson AF, Pinhassi J (2018) Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. Microbiome 6(1):173. https://doi.org/10.1186/s40168-018-0550-0

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q (2020) Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21(1):30. https://doi.org/10.1186/s13059-020-1935-5

Amato A, Dell'Aquila G, Musacchia F, Annunziata R, Ugarte A, Maillet N, Carbone A, d'Alcala MR, Sanges R, Iudicone D, Ferrante MI (2017) Marine diatoms change their gene expression profile when exposed to microscale turbulence under nutrient replete conditions. Sci Rep 7:3826–3826. https://doi.org/10.1038/s41598-017-03741-6

Amato A, Orsini L (2015) Rare interspecific breeding in pseudo-nitzschia (Bacillariophyceae). Phytotaxa 217(2):145–154. https://doi.org/10.11646/phytotaxa.217.2.4

Amin SA, Parker MS, Armbrust EV (2012) Interactions between diatoms and bacteria. Microbiol Mol Biol Rev 76(3):667–684. https://doi.org/10.1128/MMBR.00007-12

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WW, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science 306(5693):79–86. https://doi.org/10.1126/science.1101156

Ashworth J, Coesel S, Lee A, Armbrust EV, Orellana MV, Baliga NS (2013) Genome-wide diel growth state transitions in the diatom Thalassiosira pseudonana. Proc Natl Acad Sci U S A 110(18):7518–7523. https://doi.org/10.1073/pnas.1300962110

Basu S, Patil S, Mapleson D, Russo MT, Vitale L, Fevola C, Maumus F, Casotti R, Mock T, Caccamo M, Montresor M, Sanges R, Ferrante MI (2017) Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. New Phytol 215(1):140–156. https://doi.org/10.1111/nph.14557

Bender SJ, Durkin CA, Berthiaume CT, Morales RL, Armbrust EV (2014) Transcriptional responses of three model diatoms to nitrate limitation of growth. Front Mar Sci 1. https://doi.org/10.3389/fmars.2014.00003

Benoiston AS, Ibarbalz FM, Bittner L, Guidi L, Jahn O, Dutkiewicz S, Bowler C (2017) The evolution of diatoms and their biogeochemical functions. Philos Trans R Soc Lond Ser B Biol Sci 372(1728). https://doi.org/10.1098/rstb.2016.0397

Bhattacharya D, Price DC, Chan CX, Qiu H, Rose N, Ball S, Weber APM, Arias MC, Henrissat B, Coutinho PM, Krishnan A, Zauner S, Morath S, Hilliou F, Egizi A, Perrineau MM, Yoon HS (2013) Genome of the red alga Porphyridium purpureum. Nat Commun 4:10. https://doi.org/10.1038/ncomms2931

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glockner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Genome Standards C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35(8):725–731. https://doi.org/10.1038/nbt.3893

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret JP, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, Fournet J, Haruta M, Huysman MJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kroger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin-Jezequel V, Lopez PJ, Lucas S, Mangogna M, McGinnis K, Medlin LK, Montsant A, Oudot-Le Secq MP, Napoli C, Obornik M, Parker MS, Petit JL, Porcel BM, Poulsen N, Robison M, Rychlewski L, Rynearson TA, Schmutz J, Shapiro H, Siaut M, Stanley M, Sussman MR, Taylor AR, Vardi A, von Dassow P, Vyverman W, Willis A, Wyrwicz LS, Rokhsar DS, Weissenbach J, Armbrust EV, Green BR, Van de Peer Y, Grigoriev IV (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 456(7219):239–244. https://doi.org/10.1038/nature07410

Branco-Vieira M, San Martin S, Agurto C, Freitas MAV, Martins AA, Mata TM, Caetano NS (2020) Biotechnological potential of Phaeodactylum tricornutum for biorefinery processes. Fuel 268:13. https://doi.org/10.1016/j.fuel.2020.117357

Brodie J, Ball SG, Bouget FY, Chan CX, De Clerck O, Cock JM, Gachon C, Grossman AR, Mock T, Raven JA, Saha M, Smith AG, Vardi A, Yoon HS, Bhattacharya D (2017) Biotic interactions as drivers of algal origin and evolution. New Phytol 216(3):670–681. https://doi.org/10.1111/nph.14760

Bulankova P, Sekulić M, Jallet D, Nef C, van Oosterhout C, Delmont TO, Vercauteren I, Osuna-Cruz CM, Vancaester E, Mock T, Sabbe K, Daboussi F, Bowler C, Vyverman W, Vandepoele K, De Veylder L (2021) Mitotic recombination between homologous chromosomes drives genomic diversity in diatoms. Curr Biol 31:1–12. https://doi.org/10.1016/j.cub.2021.05.013

Burki F, Roger AJ, Brown MW, Simpson AGB (2020) The New tree of eukaryotes. Trends Ecol Evol 35(1):43–55. https://doi.org/10.1016/j.tree.2019.08.008

Caputi L, Carradec Q, Eveillard D, Kirilovsky A, Pelletier E, Karlusich JJP, Vieira FRJ, Villar E, Chaffron S, Malviya S, Scalco E, Acinas SG, Alberti A, Aury JM, Benoiston AS, Bertrand A, Biard T, Bittner L, Boccara M, Brum JR, Brunet C, Busseni G, Carratala A, Claustre H, Coelho LP, Colin S, D'Aniello S, Da Silva C, Del Core M, Dore H, Gasparini S, Kokoszka F, Jamet JL, Lejeusne C, Lepoivre C, Lescot M, Lima-Mendez G, Lombard F, Lukes J, Maillet N, Madoui MA, Martinez E, Mazzocchi MG, Neou MB, Paz-Yepes J, Poulain J, Ramondenc S, Romagnan JB, Roux S, Manta DS, Sanges R, Speich S, Sprovieri M, Sunagawa S, Taillandier V, Tanaka A,

Tirichine L, Trottier C, Uitz J, Veluchamy A, Vesela J, Vincent F, Yau S, Kandels-Lewis S, Searson S, Dimier C, Picheral M, Bork P, Boss E, De Vargas C, Follows MJ, Grimsley N, Guidi L, Hingamp P, Karsenti E, Sordino P, Stemmann L, Sullivan MB, Tagliabue A, Zingone A, Garczarek L, d'Ortenzio F, Testor P, Not F, d'Alcala MR, Wincker P, Bowler C, Iudicone D, Gorsky G, Jaillon O, Karp-Boss L, Krzic U, Ogata H, Pesant S, Raes J, Reynaud EG, Sardet C, Sieracki M, Velayoudon D, Weissenbach J, Tara Oceans C (2019) Community-level responses to iron availability in Open Ocean plankton ecosystems. Glob Biogeochem Cycle 33(3):391–419. https://doi.org/10.1029/2018gb006022

Chambers AF, Groom AC, MacDonald IC (2002) Dissemination and growth of cancer cells in metastatic sites. Nat Rev Cancer 2(8):563–572. https://doi.org/10.1038/nrc865

Chen X-H, Li Y-Y, Zhang H, Liu J-L, Xie Z-X, Lin L, Wang D-Z (2018) Quantitative proteomics reveals common and specific responses of a marine diatom Thalassiosira pseudonana to different macronutrient deficiencies. Front Microbiol 9. https://doi.org/10.3389/fmicb.2018.02761

Chepurnov VA, Mann DG, Sabbe K, Vyverman W (2004) Experimental studies an sexual reproduction in diatoms. In: Jeon KW (ed) International review of cytology–a survey of cell biology, vol 237, p 91. https://doi.org/10.1016/S0074-7696(04)37003-8

Cohen NR, Gong W, Moran DM, McIlvin MR, Saito MA, Marchetti A (2019) Transcriptomic and proteomic responses of the oceanic diatom Pseudo-nitzschia granii to iron limitation (vol 20, pg 3109, 2018). Environ Microbiol 21(9):3527–3527. https://doi.org/10.1111/1462-2920.14771

Compton DA (2011) Mechanisms of aneuploidy. Curr Opin Cell Biol 23(1):109–113. https://doi.org/10.1016/j.ceb.2010.08.007

Connolly JA, Oliver MJ, Beaulieu JM, Knight CA, Tomanek L, Moline MA (2008) Correlated evolution of genome size and cell volume in diatoms (Bacillariophyceae). J Phycol 44(1):124–131. https://doi.org/10.1111/j.1529-8817.2007.00452.x

Daboussi F, Leduc S, Marechal A, Dubois G, Guyot V, Perez-Michaut C, Amato A, Falciatore A, Juillerat A, Beurdeley M, Voytas DF, Cavarec L, Duchateau P (2014) Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology. Nat Commun 5:3831. https://doi.org/10.1038/ncomms4831

Darling AE, Jospin G, Lowe E, Matsen FI, Bik HM, Eisen JA (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ 2:28. https://doi.org/10.7717/peerj.243

Darwin Tree of Life. (n.d.). https://www.darwintreeoflife.org/. February 18, 2021

de Carvalho MHC, Bowler C (2020) Global identification of a marine diatom long noncoding natural antisense transcripts (NATs) and their response to phosphate fluctuations. Sci Rep 10(1):11. https://doi.org/10.1038/s41598-020-71002-0

de Carvalho MHC, Sun HX, Bowler C, Chua NH (2016) Noncoding and coding transcriptome responses of a marine diatom to phosphate fluctuations. New Phytol 210(2):497–510. https://doi.org/10.1111/nph.13787

De Riso V, Raniello R, Maumus F, Rogato A, Bowler C, Falciatore A (2009) Gene silencing in the marine diatom Phaeodactylum tricornutum. Nucleic Acids Res 37(14):e96. https://doi.org/10.1093/nar/gkp448

Delalat B, Sheppard VC, Rasi Ghaemi S, Rao S, Prestidge CA, McPhee G, Rogers ML, Donoghue JF, Pillay V, Johns TG, Kroger N, Voelcker NH (2015) Targeted drug delivery using genetically engineered diatom biosilica. Nat Commun 6:8791. https://doi.org/10.1038/ncomms9791

Delmont TO, Gaia M, Hinsinger DD, Fremont P, Guerra AF, Eren AM, Vanni C, Kourlaiev A, d'Agata L, Clayssen Q, Villar E, Labadie K, Cruaud C, Poulain J, Da Silva C, Wessner M, Noel B, Aury J-M, de Vargas C, Bowler C, Karsenti E, Pelletier E, Wincker P, Jaillon O (2020) Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. bioRxiv:2020.2010.2015.341214. https://doi.org/10.1101/2020.10.15.341214

Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynie S, Cooke R, Saeys Y, Wuyts J, Jabbari K, Bowler C, Panaud O, Piegu B, Ball SG, Ral JP, Bouget FY, Piganeau G, De Baets B, Picard A, Delseny M, Demaille J, Van de Peer Y, Moreau H (2006) Genome analysis of the smallest free-living eukaryote Ostreococcus tauri

unveils many unique features. Proc Natl Acad Sci U S A 103(31):11647–11652. https://doi.org/10.1073/pnas.0604795103
Deschamps P, Moreira D (2012) Reevaluating the Green contribution to diatom genomes. Genome Biol Evol 4(7):795–800. https://doi.org/10.1093/gbe/evs053
Dorrell RG, Gile G, McCallum G, Meheust R, Bapteste EP, Klinger CM, Brillet-Gueguen L, Freeman KD, Richter DJ, Bowler C (2017) Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. elife 6:45. https://doi.org/10.7554/eLife.23717
Dorrell RG, Villain A, Perez-Lamarque B, de Kerdrel GA, McCallum G, Watson AK, Ait-Mohamed O, Alberti A, Corre E, Frischkorn KR, Karlusich JJP, Pelletier E, Morlon H, Bowler C, Blanc G (2021) Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes. Proc Natl Acad Sci U S A 118(4). https://doi.org/10.1073/pnas.2009974118
Duncan A, Barry K, Daum C, Eloe-Fadrosh E, Roux S, Tringe SG, Schmidt K, Valentin KU, Varghese N, Grigoriev IV, Leggett R, Moulton V, Mock T (2020) Metagenome-assembled genomes of phytoplankton communities across the Arctic Circle. bioRxiv:2020.2006.2016.154583. https://doi.org/10.1101/2020.06.16.154583
Dyhrman ST, Jenkins BD, Rynearson TA, Saito MA, Mercier ML, Alexander H, Whitney LP, Drzewianowski A, Bulygin VV, Bertrand EM, Wu ZJ, Benitez-Nelson C, Heithoff A (2012) The transcriptome and proteome of the diatom Thalassiosira pseudonana reveal a diverse phosphorus stress response. PLoS One 7(3):10. https://doi.org/10.1371/journal.pone.0033768
Fabris M, Matthijs M, Rombauts S, Vyverman W, Goossens A, Baart GJE (2012) The metabolic blueprint of Phaeodactylum tricornutum reveals a eukaryotic Entner-Doudoroff glycolytic pathway. Plant J 70(6):1004–1014. https://doi.org/10.1111/j.1365-313X.2012.04941.x
Falciatore A, Jaubert M, Bouly J-P, Bailleul B, Mock T (2020) Diatom molecular research comes of age: model species for studying phytoplankton biology and diversity. Plant Cell 32(3): 547–572. https://doi.org/10.1105/tpc.19.00158
Fierst JL, Murdock DA (2017) Decontaminating eukaryotic genome assemblies with machine learning. Bmc Bioinformatics 18. https://doi.org/10.1186/s12859-017-1941-0
George J, Kahlke T, Abbriano RM, Kuzhiumparambil U, Ralph PJ, Fabris M (2020) Metabolic engineering strategies in diatoms reveal unique phenotypes and genetic configurations with implications for algal genetics and synthetic biology. Front Bioeng Biotechnol 8:19. https://doi.org/10.3389/fbioe.2020.00513
Godhe A, Rynearson T (2017) The role of intraspecific variation in the ecological and evolutionary success of diatoms in changing environments. Philos Trans R Soc Lond B Biol Sci 372(1728): 10. https://doi.org/10.1098/rstb.2016.0399
Grossman AR (2005) Paths toward algal genomics. Plant Physiol 137(2):410–427. https://doi.org/10.1104/pp.104.053447
Gruber A, Kroth PG (2017) Intracellular metabolic pathway distribution in diatoms and tools for genome-enabled experimental diatom research. Philos Trans R Soc Lond B Biol Sci 372(1728). https://doi.org/10.1098/rstb.2016.0402
Helliwell KE, Kleiner FH, Hardstaff H, Chrachri A, Gaikwad T, Salmon D, Smirnoff N, Wheeler GL, Brownlee C (2021) Spatiotemporal patterns of intracellular Ca2+ signalling govern hypo-osmotic stress resilience in marine diatoms. New Phytol. https://doi.org/10.1111/nph.17162
Helliwell KE, Wheeler GL, Leptos KC, Goldstein RE, Smith AG (2011) Insights into the evolution of vitamin B-12 Auxotrophy from sequenced algal genomes. Mol Biol Evol 28(10):2921–2933. https://doi.org/10.1093/molbev/msr124
Hildebrand M, Lerch SJL, Shrestha RP (2018) Understanding diatom Cell Wall Silicification - moving forward. Front Mar Sci 5:19. https://doi.org/10.3389/fmars.2018.00125
Hoguin A, Rastogi A, Bowler C, Tirichine L (2021) Genome-wide analysis of allele-specific expression of genes in the model diatom Phaeodactylum tricornutum. Sci Rep 11(1):10. https://doi.org/10.1038/s41598-021-82529-1
Hopes A, Mock T (2015) Evolution of Microalgae and Their Adaptations in Different Marine Ecosystems. In: eLS. pp. 1–9. https://doi.org/10.1002/9780470015902.a0023744

Hopes A, Nekrasov V, Kamoun S, Mock T (2016) Editing of the urease gene by CRISPR-Cas in the diatom Thalassiosira pseudonana. Plant Methods 12:49. https://doi.org/10.1186/s13007-016-0148-0

Huang AY, He LW, Wang GC (2011) Identification and characterization of microRNAs from Phaeodactylum tricornutum by high-throughput sequencing and bioinformatics analysis. BMC Genomics 12:11. https://doi.org/10.1186/1471-2164-12-337

Hugerth LW, Larsson J, Alneberg J, Lindh MV, Legrand C, Pinhassi J, Andersson AF (2015) Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biol 16:279. https://doi.org/10.1186/s13059-015-0834-7

Huysman MJJ, Martens C, Vandepoele K, Gillard J, Rayko E, Heijde M, Bowler C, Inze D, Van de Peer Y, De Veylder L, Vyverman W (2010) Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. Genome Biol 11(2). https://doi.org/10.1186/gb-2010-11-2-r17

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36(4):338. https://doi.org/10.1038/nbt.4060

Janech MG, Krell A, Mock T, Kang JS, Raymond JA (2006) Ice-binding proteins from sea ice diatoms (Bacillariophyceae). J Phycol 42(2):410–416. https://doi.org/10.1111/j.1529-8817.2006.00208.x

Kaster AK, Sobol MS (2020) Microbial single-cell omics: the crux of the matter. Appl Microbiol Biotechnol 104(19):8209–8220. https://doi.org/10.1007/s00253-020-10844-0

Kim J, Brown CM, Kim MK, Burrows EH, Bach S, Lun DS, Falkowski PG (2017) Effect of cell cycle arrest on intermediate metabolism in the marine diatom Phaeodactylum tricornutum. Proc Natl Acad Sci U S A 114(38):E8007–E8016. https://doi.org/10.1073/pnas.1711642114

Kim J, Fabris M, Baart G, Kim MK, Goossens A, Vyverman W, Falkowski PG, Lun DS (2016) Flux balance analysis of primary metabolism in the diatom Phaeodactylum tricornutum. Plant J 85(1):161–176. https://doi.org/10.1111/tpj.13081

Kociolek JP, Stoermer EF (1989) Chromosome numbers in diatoms a review. Diatom Res 4(1):47–54. https://doi.org/10.1080/0269249X.1989.9705051

Koester JA, Berthiaume CT, Hiranuma N, Parker MS, Iverson V, Morales R, Ruzzo WL, Armbrust EV (2018) Sexual ancestors generated an obligate asexual and globally dispersed clone within the model diatom species Thalassiosira pseudonana. Sci Rep 8:9. https://doi.org/10.1038/s41598-018-28630-4

Koester JA, Swalwell JE, von Dassow P, Armbrust EV (2010) Genome size differentiates co-occurring populations of the planktonic diatom Ditylum brightwellii (Bacillariophyta). BMC Evol Biol 10. https://doi.org/10.1186/1471-2148-10-1

Kroth PG, Chiovitti A, Gruber A, Martin-Jezequel V, Mock T, Parker MS, Stanley MS, Kaplan A, Caron L, Weber T, Maheswari U, Armbrust EV, Bowler C (2008) A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole Genome analysis. PLoS One 3(1):14. https://doi.org/10.1371/journal.pone.0001426

Krueger-Hadfield SA, Balestreri C, Schroeder J, Highfield A, Helaouet P, Allum J, Moate R, Lohbeck KT, Miller PI, Riebesell U, Reusch TBH, Rickaby REM, Young J, Hallegraeff G, Brownlee C, Schroeder DC (2014) Genotyping an Emiliania huxleyi (prymnesiophyceae) bloom event in the North Sea reveals evidence of asexual reproduction. Biogeosciences 11(18):5215–5234. https://doi.org/10.5194/bg-11-5215-2014

Kustka AB, Milligan AJ, Zheng H, New AM, Gates C, Bidle KD, Reinfelder JR (2014) Low CO2 results in a rearrangement of carbon metabolism to support C-4 photosynthetic carbon assimilation in Thalassiosira pseudonana. New Phytol 204(3):507–520. https://doi.org/10.1111/nph.12926

Labonte JM, Swan BK, Poulos B, Luo H, Koren S, Hallam SJ, Sullivan MB, Woyke T, Wommack KE, Stepanauskas R (2015) Single-cell genomics-based analysis of virus-host interactions in

marine surface bacterioplankton. ISME J 9(11):2386–2399. https://doi.org/10.1038/ismej.2015.48

Lampe RH, Cohen NR, Ellis KA, Bruland KW, Maldonado MT, Peterson TD, Till CP, Brzezinski MA, Bargu S, Thamatrakoln K, Kuzminov FI, Twining BS, Marchetti A (2018) Divergent gene expression among phytoplankton taxa in response to upwelling. Environ Microbiol 20(8): 3069–3082. https://doi.org/10.1111/1462-2920.14361

Lane N (2015) The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'. Philos Trans R Soc Lond B Biol Sci 370(1666):10. https://doi.org/10.1098/rstb.2014.0344

Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ (2015) Assessing the performance of the Oxford Nanopore technologies MinION. Biomol Detect Quantif 3:1–8. https://doi.org/10.1016/j.bdq.2015.02.001

Levering J, Broddrick J, Dupont CL, Peers G, Beeri K, Mayers J, Gallina AA, Allen AE, Palsson BO, Zengler K (2016) Genome-scale model reveals metabolic basis of biomass partitioning in a model diatom. PLoS One 11(5):22. https://doi.org/10.1371/journal.pone.0155038

Levering J, Dupont CL, Allen AE, Palsson BO, Zengler K (2017) Integrated regulatory and metabolic networks of the marine diatom Phaeodactylum tricornutum predict the response to rising CO2 levels. Msystems 2(1). https://doi.org/10.1128/mSystems.00142-16

Levitan O, Dinamarca J, Zelzion E, Lun DS, Guerra LT, Kim MK, Kim J, Van Mooy BAS, Bhattacharya D, Falkowski PG (2015) Remodeling of intermediate metabolism in the diatom Phaeodactylum tricornutum under nitrogen stress. Proc Natl Acad Sci U S A 112(2):412–417. https://doi.org/10.1073/pnas.1419818112

Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H (2014) UpSet: visualization of intersecting sets. IEEE Trans Vis Comput Graph 20(12):1983–1992. https://doi.org/10.1109/tvcg.2014.2346248

Li SL, Nosenko T, Hackett JD, Bhattacharya D (2006) Phylogenomic analysis identifies red algal genes of endosymbiotic origin in the chromalveolates. Mol Biol Evol 23(3):663–674. https://doi.org/10.1093/molbev/msj075

Lommer M, Specht M, Roy AS, Kraemer L, Andreson R, Gutowska MA, Wolf J, Bergner SV, Schilhabel MB, Klostermeier UC, Beiko RG, Rosenstiel P, Hippler M, LaRoche J (2012) Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome Biol 13(7):20. https://doi.org/10.1186/gb-2012-13-7-r66

López-Escardó D, Grau-Bové X, Guillaumet-Adkins A, Gut M, Sieracki ME, Ruiz-Trillo I (2017) Evaluation of single-cell genomics to address evolutionary questions using three SAGs of the choanoflagellate Monosiga brevicollis. Sci Rep 7(1):11025. https://doi.org/10.1038/s41598-017-11466-9

Lopez-Gomollon S, Beckers M, Rathjen T, Moxon S, Maumus F, Mohorianu I, Moulton V, Dalmay T, Mock T (2014) Global discovery and characterization of small non-coding RNAs in marine microalgae. BMC Genomics 15:12. https://doi.org/10.1186/1471-2164-15-697

Lu J, Salzberg SL (2018) Removing contaminants from databases of draft genomes. PLoS Comput Biol 14(6):13. https://doi.org/10.1371/journal.pcbi.1006277

Maeda Y, Watanabe K, Kobayashi R, Yoshino T, Bowler C, Matsumoto M, Tanaka T (2021) Chromosome scale assembly of allopolyploid genome of the diatom Fistulifera solaris. bioRxiv. https://www.biorxiv.org/content/10.1101/2021.11.10.468027v1.full

Malviya S, Scalco E, Audic S, Vincenta F, Veluchamy A, Poulain J, Wincker P, Iudicone D, de Vargas C, Bittner L, Zingone A, Bowler C (2016) Insights into global diatom distribution and diversity in the world's ocean. Proc Natl Acad Sci U S A 113(11):E1516–E1525. https://doi.org/10.1073/pnas.1509523113

Mann DG (1994) Auxospore formation, reproductive plasticity and cell structure in navicula-ulvacea and the resurrection of the genus dickieia (bacillariophyta). Eur J Phycol 29(3): 141–157. https://doi.org/10.1080/09670269400650591

Mann DG, Vanormelingen P (2013) An inordinate fondness? The number, distributions, and origins of diatom species. J Eukaryot Microbiol 60(4):414–420. https://doi.org/10.1111/jeu.12047

Marchetti A, Schruth DM, Durkin CA, Parker MS, Kodner RB, Berthiaume CT, Morales R, Allen AE, Armbrust EV (2012) Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. Proc Natl Acad Sci U S A 109(6):E317–E325. https://doi.org/10.1073/pnas.1118408109

Matsuzaki M, Misumi O, Shin-I T, Maruyama S, Takahara M, Miyagishima SY, Mori T, Nishida K, Yagisawa F, Nishida K, Yoshida Y, Nishimura Y, Nakao S, Kobayashi T, Momoyama Y, Higashiyama T, Minoda A, Sano M, Nomoto H, Oishi K, Hayashi H, Ohta F, Nishizaka S, Haga S, Miura S, Morishita T, Kabeya Y, Terasawa K, Suzuki Y, Ishii Y, Asakawa S, Takano H, Ohta N, Kuroiwa H, Tanaka K, Shimizu N, Sugano S, Sato N, Nozaki H, Ogasawara N, Kohara Y, Kuroiwa T (2004) Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. Nature 428(6983):653–657. https://doi.org/10.1038/nature02398

Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, Vardi A, Grandbastien M-A, Bowler C (2009) Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. BMC Genomics 10. https://doi.org/10.1186/1471-2164-10-624

Medlin L, Elwood HJ, Stickel S, Sogin ML (1988) The characterization of enzymatically amplified eukaryotic 16s-like RRNA-coding regions. Gene 71(2):491–499. https://doi.org/10.1016/0378-1119(88)90066-2

Mock T, Otillar RP, Strauss J, McMullan M, Paajanen P, Schmutz J, Salamov A, Sanges R, Toseland A, Ward BJ, Allen AE, Dupont CL, Frickenhaus S, Maumus F, Veluchamy A, Wu T, Barry KW, Falciatore A, Ferrante MI, Fortunato AE, Glockner G, Gruber A, Hipkin R, Janech MG, Kroth PG, Leese F, Lindquist EA, Lyon BR, Martin J, Mayer C, Parker M, Quesneville H, Raymond JA, Uhlig C, Valas RE, Valentin KU, Worden AZ, Armbrust EV, Clark MD, Bowler C, Green BR, Moulton V, van Oosterhout C, Grigoriev IV (2017) Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature 541(7638):536–540. https://doi.org/10.1038/nature20803

Mock T, Samanta MP, Iverson V, Berthiaume C, Robison M, Holtermann K, Durkin C, Bondurant SS, Richmond K, Rodesch M, Kallas T, Huttlin EL, Cerrina F, Sussman MR, Armbrust EV (2008) Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. Proc Natl Acad Sci U S A 105(5):1579–1584. https://doi.org/10.1073/pnas.0707946105

Monnich J, Tebben J, Bergemann J, Case R, Wohlrab S, Harder T (2020) Niche-based assembly of bacterial consortia on the diatom Thalassiosira rotula is stable and reproducible. ISME J 14(6):1614–1625. https://doi.org/10.1038/s41396-020-0631-5

Montsant A, Allen AE, Coesel S, De Martino A, Falciatore A, Mangogna M, Siaut M, Heijde M, Jabbari K, Maheswari U, Rayko E, Vardi A, Apt KE, Berges JA, Chiovitti A, Davis AK, Thamatrakoln K, Hadi MZ, Lane TW, Lippmeier JC, Martinez D, Parker MS, Pazour GJ, Saito MA, Rokhsar DS, Armbrust EV, Bowler C (2007) Identification and comparative genomic analysis of signaling and regulatory components in the diatom Thalassiosira pseudonana. J Phycol 43(3):585–604. https://doi.org/10.1111/j.1529-8817.2007.00342.x

Montsant A, Jabbari K, Maheswari U, Bowler C (2005) Comparative genomics of the pennate diatom Phaeodactylum tricornutum. Plant Physiol 137(2):500–513. https://doi.org/10.1104/pp.104.052829

Moss EL, Maghini DG, Bhatt AS (2020) Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol 38(6):701–707. https://doi.org/10.1038/s41587-020-0422-6

Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D (2009) Genomic footprints of a cryptic plastid endosymbiosis in diatoms. Science 324(5935):1724–1726. https://doi.org/10.1126/science.1172983

Nakov T, Beaulieu JM, Alverson AJ (2018) Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (diatoms, Bacillariophyta). New Phytol 219(1):462–473. https://doi.org/10.1111/nph.15137

Nelson DR, Hazzouri KM, Lauersen KJ, Jaiswal A, Chaiboonchoe A, Mystikou A, Fu W, Daakour S, Dohai B, Alzahmi A, Nobles D, Hurd M, Sexton J, Preston MJ, Blanchette J, Lomas MW, Amiri KMA, Salehi-Ashtiani K (2021) Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution. Cell Host Microbe 29(2):250–266.e258. https://doi.org/10.1016/j.chom.2020.12.005

Nisbet RER, Kilian O, McFadden GI (2004) Diatom genomics: genetic acquisitions and mergers. Curr Biol 14(24):R1048–R1050. https://doi.org/10.1016/j.cub.2004.11.043

Nomaguchi T, Maeda Y, Yoshino T, Asahi T, Tirichine L, Bowler C, Tanaka T (2018) Homoeolog expression bias in allopolyploid oleaginous marine diatom Fistulifera solaris. BMC Genomics 19:17. https://doi.org/10.1186/s12864-018-4691-0

Norden-Krichmar TM, Allen AE, Gaasterland T, Hildebrand M (2011) Characterization of the small RNA transcriptome of the diatom, Thalassiosira pseudonana. PLoS One 6(8):e22870. https://doi.org/10.1371/journal.pone.0022870

Nowell PC (1976) Clonal evolution of tumor-cell populations. Science 194(4260):23–28. https://doi.org/10.1126/science.959840

Osuna-Cruz CM, Bilcke G, Vancaester E, De Decker S, Bones AM, Winge P, Poulsen N, Bulankova P, Verhelst B, Audoor S, Belisova D, Pargana A, Russo M, Stock F, Cirri E, Brembu T, Pohnert G, Piganeau G, Ferrante MI, Mock T, Sterck L, Sabbe K, De Veylder L, Vyverman W, Vandepoele K (2020) The Seminavis robusta genome provides insights into the evolutionary adaptations of benthic diatoms (vol 11, 3320, 2020). Nat Commun 11(1):1. https://doi.org/10.1038/s41467-020-19222-w

Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, Poulton NJ, Burkart MD, La Clair JJ, Chisholm SW, Stepanauskas R (2019) Charting the complexity of the marine microbiome through single-cell genomics. Cell 179(7):1623–1635 e1611. https://doi.org/10.1016/j.cell.2019.11.017

Parks MB, Nakov T, Ruck EC, Wickett NJ, Alverson AJ (2018) Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). Am J Bot 105(3):330–347. https://doi.org/10.1002/ajb2.1056

Pinseel E, Janssens SB, Verleyen E, Vanormelingen P, Kohler TJ, Biersma EM, Sabbe K, Van de Vijver B, Vyverman W (2020) Global radiation in a rare biosphere soil diatom. Nat Commun 11(1):12. https://doi.org/10.1038/s41467-020-16181-0

Ponce-Toledo RI, Lpez-Garca P, Moreira D (2019) Horizontal and endosymbiotic gene transfer in early plastid evolution. New Phytol 224(2):618–624. https://doi.org/10.1111/nph.15965

Postel U, Glemser B, Alekseyeva KS, Eggers SL, Groth M, Glockner G, John U, Mock T, Klemm K, Valentin K, Beszteri B (2020) Adaptive divergence across Southern Ocean gradients in the pelagic diatomFragilariopsis kerguelensis. Mol Ecol 29(24):4913–4924. https://doi.org/10.1111/mec.15554

Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber APM, Schwacke R, Gross J, Blouin NA, Lane C, Reyes-Prieto A, Durnford DG, Neilson JAD, Lang BF, Burger G, Steiner JM, Loffelhardt W, Meuser JE, Posewitz MC, Ball S, Arias MC, Henrissat B, Coutinho PM, Rensing SA, Symeonidi A, Doddapaneni H, Green BR, Rajah VD, Boore J, Bhattacharya D (2012) Cyanophora paradoxa Genome elucidates origin of photosynthesis in algae and plants. Science 335(6070):843–847. https://doi.org/10.1126/science.1213561

Prihoda J, Tanaka A, de Paula WBM, Allen JF, Tirichine L, Bowler C (2012) Chloroplast-mitochondria cross-talk in diatoms. J Exp Bot 63(4):1543–1557. https://doi.org/10.1093/jxb/err441

Rastogi A, Maheswari U, Dorrell RG, Vieira FRJ, Maumus F, Kustka A, McCarthy J, Allen AE, Kersey P, Bowler C, Tirichine L (2018) Integrative analysis of large scale transcriptome data draws a comprehensive landscape of Phaeodactylum tricornutum genome and evolutionary origin of diatoms. Sci Rep 8:14. https://doi.org/10.1038/s41598-018-23106-x

Rastogi A, Vieira FRJ, Deton-Cabanillas AF, Veluchamy A, Cantrel C, Wang GH, Vanormelingen P, Bowler C, Piganeau G, Hu HH, Tirichine L (2020) A genomics approach reveals the global genetic polymorphism, structure, and functional diversity of ten accessions of

the marine model diatom Phaeodactylum tricornutum. ISME J 14(2):347–363. https://doi.org/10.1038/s41396-019-0528-3

Raymond JA, Kim HJ (2012) Possible role of horizontal gene transfer in the colonization of sea ice by algae. PLoS One 7(5). https://doi.org/10.1371/journal.pone.0035968

Rengefors K, Kremp A, Reusch TBH, Wood AM (2017) Genetic diversity and evolution in eukaryotic phytoplankton: revelations from population genetic studies. J Plankton Res 39(2): 165–179. https://doi.org/10.1093/plankt/fbw098

Roberts WR, Downey KM, Ruck EC, Traller JC, Alverson AJ (2020) Improved reference Genome for Cyclotella cryptica CCMP332, a model for Cell Wall morphogenesis, salinity adaptation, and lipid production in diatoms (Bacillariophyta). G3-genes genomes. Genetics 10(9): 2965–2974. https://doi.org/10.1534/g3.120.401408

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. https://doi.org/10.1186/1471-2164-15-698

Rosenwasser S, van Creveld SG, Schatz D, Malitsky S, Tzfadia O, Aharoni A, Levin Y, Gabashvili A, Feldmesser E, Vardi A (2014) Mapping the diatom redox-sensitive proteome provides insight into response to nitrogen stress in the marine environment. Proc Natl Acad Sci U S A 111(7):2740–2745. https://doi.org/10.1073/pnas.1319773111

Rossoni AW, Price DC, Seger M, Lyska D, Lammers P, Bhattacharya D, Weber APM (2019) The genomes of polyextremophilic cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions. elife 8:57. https://doi.org/10.7554/eLife.45017

Rynearson TA, Newton JA, Armbrust EV (2006) Spring bloom development, genetic variation, and population succession in the planktonic diatom Ditylum brightwellii. Limnol Oceanogr 51(3): 1249–1261. https://doi.org/10.4319/lo.2006.51.3.1249

Sasso S, Stibor H, Mittag M, Grossman AR (2018) From molecular manipulation of domesticated Chlamydomonas reinhardtii to survival in nature. elife 7. https://doi.org/10.7554/eLife.39233

Sayanova O, Mimouni V, Ulmann L, Morant-Manceau A, Pasquet V, Schoefs B, Napier JA (2017) Modulation of lipid biosynthesis by stress in diatoms. Philos Trans R Soc Lond B Biol Sci 372(1728). https://doi.org/10.1098/rstb.2016.0407

Schmid M, Frei D, Patrignani A, Schlapbach R, Frey JE, Remus-Emsermann MNP, Ahrens CH (2018) Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. Nucleic Acids Res 46(17):8953–8965. https://doi.org/10.1093/nar/gky726

Schrader L, Schmitz J (2019) The impact of transposable elements in adaptive evolution. Mol Ecol 28(6):1537–1549. https://doi.org/10.1111/mec.14794

Seeleuthner Y, Mondy S, Lombard V, Carradec Q, Pelletier E, Wessner M, Leconte J, Mangot JF, Poulain J, Labadie K, Logares R, Sunagawa S, de Berardinis V, Salanoubat M, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Tara Oceans C, Pesant S, Poulton N, Stepanauskas R, Bork P, Bowler C, Hingamp P, Sullivan MB, Iudicone D, Massana R, Aury JM, Henrissat B, Karsenti E, Jaillon O, Sieracki M, de Vargas C, Wincker P (2018) Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. Nat Commun 9(1):310. https://doi.org/10.1038/s41467-017-02235-3

Shrestha RP, Tesson B, Norden-Krichmar T, Federowicz S, Hildebrand M, Allen AE (2012) Whole transcriptome analysis of the silicon response of the diatom Thalassiosira pseudonana. BMC Genomics 13:16. https://doi.org/10.1186/1471-2164-13-499

Sieracki ME, Poulton NJ, Jaillon O, Wincker P, de Vargas C, Rubinat-Ripoll L, Stepanauskas R, Logares R, Massana R (2019) Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. Sci Rep 9(1):6025. https://doi.org/10.1038/s41598-019-42487-1

Singh D, Carlson R, Fell D, Poolman M (2015) Modelling metabolism of the diatom Phaeodactylum tricornutum. Biochem Soc Trans 43:1182–1186. https://doi.org/10.1042/bst20150152

Sorhannus U (2011) Evolution of antifreeze protein genes in the diatom genus Fragilariopsis: evidence for horizontal gene transfer, gene duplication and episodic diversifying selection. Evol Bioinforma 7:279–289. https://doi.org/10.4137/ebo.S8321

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stephens TG, Gonzalez-Pech RA, Cheng YY, Mohamed A, Burt DW, Bhattacharya D, Ragan MA, Chan CX (2020) Genomes of the dinoflagellate Polarella glacialis encode tandemly repeated single-exon genes with adaptive functions. BMC Biol 18(1):21. https://doi.org/10.1186/s12915-020-00782-8

Sumper M, Brunner E (2008) Silica biomineralisation in diatoms: the model organism Thalassiosira pseudonana. Chembiochem 9(8):1187–1194. https://doi.org/10.1002/cbic.200700764

Tanaka T, Maeda Y, Veluchamy A, Tanaka M, Abida H, Marechal E, Bowler C, Muto M, Sunaga Y, Tanaka M, Yoshino T, Taniguchi T, Fukuda Y, Nemoto M, Matsumoto M, Wong PS, Aburatani S, Fujibuchi W (2015) Oil accumulation by the oleaginous diatom Fistulifera solaris as revealed by the Genome and transcriptome. Plant Cell 27(1):162–176. https://doi.org/10.1105/tpc.114.135194

Tatman B, Mock T, Wu T, Van Oosterhout C (2021) Significance of differential allelic expression (DAE) in phenotypic plasticity and evolutionary potential of microbial eukaryotes. Quant Biol 9(4):400–410. https://journal.hep.com.cn/qb/EN/10.15302/J-QB-021-0258

Thomas MK, Kremer CT, Klausmeier CA, Litchman E (2012) A global pattern of thermal adaptation in marine phytoplankton. Science 338(6110):1085–1088. https://doi.org/10.1126/science.1224836

Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5(2):123–U116. https://doi.org/10.1038/nrg1271

Traller JC, Cokus SJ, Lopez DA, Gaidarenko O, Smith SR, McCrow JP, Gallaher SD, Podell S, Thompson M, Cook O, Morselli M, Jaroszewicz A, Allen EE, Allen AE, Merchant SS, Pellegrini M, Hildebrand M (2016) Genome and methylome of the oleaginous diatom Cyclotella cryptica reveal genetic flexibility toward a high lipid phenotype. Biotechnol Biofuels 9:20. https://doi.org/10.1186/s13068-016-0670-3

Treguer P, Bowler C, Moriceau B, Dutkiewicz S, Gehlen M, Aumont O, Bittner L, Dugdale R, Finkel Z, Iudicone D, Jahn O, Guidi L, Lasbleiz M, Leblanc K, Levy M, Pondaven P (2018) Influence of diatom diversity on the ocean biological carbon pump. Nat Geosci 11(1):27–37. https://doi.org/10.1038/s41561-017-0028-x

Tully BJ, Graham ED, Heidelberg JF (2018) The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data 5:170203. https://doi.org/10.1038/sdata.2017.203

van Baren MJ, Bachy C, Reistetter EN, Purvine SO, Grimwood J, Sudek S, Yu H, Poirier C, Deerinck TJ, Kuo A, Grigoriev IV, Wong CH, Smith RD, Callister SJ, Wei CL, Schmutz J, Worden AZ (2016) Evidence-based green algal genomics reveals marine diversity and ancestral characteristics of land plants. BMC Genomics 17:22. https://doi.org/10.1186/s12864-016-2585-6

Van Etten J, Bhattacharya D (2020) Horizontal gene transfer in eukaryotes: Not if, but how much? Trends Genet 36(12):915–925. https://doi.org/10.1016/j.tig.2020.08.006

Vancaester E, Depuydt T, Osuna-Cruz CM, Vandepoele K (2020) Comprehensive and functional analysis of horizontal gene transfer events in diatoms. Mol Biol Evol 37(11):3243–3257. https://doi.org/10.1093/molbev/msaa182

Veluchamy A, Lin X, Maumus F, Rivarola M, Bhavsar J, Creasy T, O'Brien K, Sengamalay NA, Tallon LJ, Smith AD, Rayko E, Ahmed I, Le Crom S, Farrant GK, Sgro JY, Olson SA, Bondurant SS, Allen AE, Rabinowicz PD, Sussman MR, Bowler C, Tirichine L (2013) Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum. Nat Commun 4:2091. https://doi.org/10.1038/ncomms3091

von Dassow P, Petersen TW, Chepurnov VA, Armbrust EV (2008) Inter- and intraspecific relationships between nuclear DNA content and cell size in selected members of the centric diatom genus Thalassiosira (Bacillariophyceae). J Phycol 44(2):335–349. https://doi.org/10.1111/j.1529-8817.2008.00476.x

Vorobev A, Dupouy M, Carradec Q, Delmont TO, Annamale A, Wincker P, Pelletier E (2020) Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. Genome Res 30(4): 647–659. https://doi.org/10.1101/gr.253070.119

Whittaker KA, Rynearson TA (2017) Evidence for environmental and ecological selection in a microbe with no geographic limits to gene flow. Proc Natl Acad Sci U S A 114(10):2651–2656. https://doi.org/10.1073/pnas.1612346114

Wideman JG, Monier A, Rodriguez-Martinez R, Leonard G, Cook E, Poirier C, Maguire F, Milner DS, Irwin NAT, Moore K, Santoro AE, Keeling PJ, Worden AZ, Richards TA (2020) Unexpected mitochondrial genome diversity revealed by targeted single-cell genomics of heterotrophic flagellated protists. Nat Microbiol 5(1):154–165. https://doi.org/10.1038/s41564-019-0605-4

Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, Foulon E, Grimwood J, Gundlach H, Henrissat B, Napoli C, McDonald SM, Parker MS, Rombauts S, Salamov A, Von Dassow P, Badger JH, Coutinho PM, Demir E, Dubchak I, Gentemann C, Eikrem W, Gready JE, John U, Lanier W, Lindquist EA, Lucas S, Mayer KFX, Moreau H, Not F, Otillar R, Panaud O, Pangilinan J, Paulsen I, Piegu B, Poliakov A, Robbens S, Schmutz J, Toulza E, Wyss T, Zelensky A, Zhou K, Armbrust EV, Bhattacharya D, Goodenough UW, Van de Peer Y, Grigoriev IV (2009) Green evolution and dynamic adaptations revealed by genomes of the marine Picoeukaryotes micromonas. Science 324(5924):268–272. https://doi.org/10.1126/science.1167222

Xu Y, Zhao F (2018) Single-cell metagenomics: challenges and applications. Protein Cell 9(5): 501–510. https://doi.org/10.1007/s13238-018-0544-5

Yang MK, Lin XH, Liu X, Zhang J, Ge F (2018) Genome annotation of a model diatom Phaeodactylum tricornutum using an integrated Proteogenomic pipeline. Mol Plant 11(10): 1292–1307. https://doi.org/10.1016/j.molp.2018.08.005

Ribosome Profiling in the Model Diatom *Thalassiosira pseudonana*

Monica Pichler,[1,8] Andreas Meindl,[2,8] Markus Romberger,[2] Annemarie Eckes-Shephard,[1,3,9] Carl-Fredrik Nyberg-Brodda,[4] Claudia Buhigas,[5] Sergio Llaneza-Lago,[6] Gerhard Lehmann,[7] Amanda Hopes,[1] Gunter Meister,[7] Jan Medenbach,[2,10] and Thomas Mock[1,10]

[1]School of Environmental Sciences, University of East Anglia, Norwich, United Kingdom
[2]Regensburg Center for Biochemistry, University of Regensburg, Regensburg, Germany
[3]Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden
[4]Department of Computer Science, Université Gustave Eiffel, Paris, France
[5]School of Biological Sciences, University of East Anglia, Norwich, United Kingdom
[6]Norwich Medical School, University of East Anglia, Norwich, United Kingdom
[7]Biochemistry I, University of Regensburg, Regensburg, Germany
[8]These authors contributed equally to this work
[9]Current affiliation: Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden
[10]Corresponding authors: *jan.medenbach@vkl.uni-regensburg.de*; *t.mock@uea.ac.uk*

Published in the Molecular Biology section

Diatoms are an important group of eukaryotic microalgae, which play key roles in marine biochemical cycling and possess significant biotechnological potential. Despite the importance of diatoms, their regulatory mechanisms of protein synthesis at the translational level remain largely unexplored. Here, we describe the detailed development of a ribosome profiling protocol to study translation in the model diatom *Thalassiosira pseudonana,* which can easily be adopted for other diatom species. To isolate and sequence ribosome-protected mRNA, total RNA was digested, and the ribosome-protected fragments were obtained by a combination of sucrose-cushion ultracentrifugation and polyacrylamide gel electrophoresis for size selection. To minimize rRNA contamination, a subtractive hybridization step using biotinylated oligos was employed. Subsequently, fragments were converted into sequencing libraries, enabling the global quantification and analysis of changes in protein synthesis in diatoms. The development of this novel ribosome profiling protocol represents a major expansion of the molecular toolbox available for diatoms and therefore has the potential to advance our understanding of the translational regulation in this important group of phytoplankton. © 2023 The Authors. Current Protocols published by Wiley Periodicals LLC.

Basic Protocol: Ribosome profiling in *Thalassiosira pseudonana*
Alternate Protocol: Ribosome profiling protocol for diatoms using sucrose gradient fractionation

Keywords: diatoms • high-throughput sequencing • ribosome profiling • RNA • translation • *Thalassiosira pseudonana*

CURRENT PROTOCOLS
A Wiley Brand

Current Protocols e843, Volume 3
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpz1.843

How to cite this article:
Pichler, M., Meindl, A., Romberger, M., Eckes-Shephard, A., Nyberg-Brodda, C.-F., Buhigas, C., Llaneza-Lago, S., Lehmann, G., Hopes, A., Meister, G., Medenbach, J., & Mock, T. (2023). Ribosome profiling in the model diatom *Thalassiosira pseudonana*. *Current Protocols*, *3,* e843. doi: 10.1002/cpz1.843

INTRODUCTION

Diatoms are unicellular, eukaryotic microalgae, which comprise over 100,000 species throughout all aquatic environments. Their mosaic genomes have been shaped by secondary endosymbiosis and horizontal gene transfer, providing them with diverse regulatory mechanisms to adapt their protein synthesis in response to the changing environmental conditions (Mock et al., 2022). Due to their key role as primary producers in aquatic systems and their biotechnological potential, gene regulatory mechanisms in diatoms have been extensively studied using transcriptomic and proteomic data (e.g., Dong et al., 2016; Mock et al., 2008). The globally distributed species *Thalassiosira pseudonana* CCMP1335 was chosen for the first diatom genome sequencing project (Armbrust et al., 2004). Subsequently, it became a model organism due to the development of genetic manipulation tools such as the incorporation of recombinant DNA via transformation (Poulsen et al., 2006) and CRISPR/Cas-based genome editing (e.g., Belshaw et al., 2023; Hopes et al., 2016). However, despite these developments, regulation of protein synthesis at the translational level is largely unexplored in *T. pseudonana* and diatoms in general. The development of ribosome profiling or Ribo-Seq in 2009 (Ingolia et al., 2009) spurred the transcriptome-wide monitoring and quantification of translation in diatoms. This technique is based on the analysis of ∼30 nucleotide (nt) long mRNA fragments which are enclosed by translating ribosomes and protected from nuclease digestion (Steitz, 1969) (Fig. 1A). Deep sequencing of these generated footprints or ribosome-protected fragments (RPFs) thus provides a genome-wide and high-resolution snapshot of translation (Ingolia et al., 2009). Ribosome profiling was first developed in *Saccharomyces cerevisiae* (Ingolia et al., 2009) and has since been successfully applied to study translation in a wide range of organisms, such as bacteria (Li et al., 2012), fishes (Bazzini et al., 2012), mammals (Guo et al., 2010; Ingolia et al., 2011) and plants (Hsu et al., 2016; Juntawong et al., 2014). Ribosome profiling has revolutionized our understanding of gene expression by unveiling the full complexity and regulation of translation in both prokaryotes and eukaryotes. It has provided novel mechanistic insights into the translation mechanism, such as ribosomal pausing sites (Brar & Weissman, 2015; Ingolia et al., 2011; Karlsen et al., 2018; Shalgi et al., 2013) or previously unannotated (small) open reading frames (ORFs) (Aspden et al., 2014; Bazzini et al., 2014; Calviello et al., 2016; Hsu et al., 2016; Ji et al., 2015; Wu et al., 2019). However, to date, ribosome profiling has only been performed on one algal species, the green alga *Chlamydomonas reinhardtii* (Chung et al., 2015; Trosch et al., 2018). Here, we report the development of a ribosome profiling protocol for the model diatom *T. pseudonana* CCMP1335, representing the first application of this technique for any marine algae. Previously developed protocols were adapted (McGlincy & Ingolia, 2017; Meindl et al., 2023; Mohammad et al., 2019), which involved optimizations of harvesting strategy and cell lysis conditions, the amount of nuclease for mRNA digestion, and the implementation of a subtractive hybridization step for rRNA removal. Our optimized strategy ensures the generation of high-quality footprint data for this important class of organisms. Thus, the application of ribosome profiling in diatoms provides a powerful tool for studying their translational regulation. Moreover,

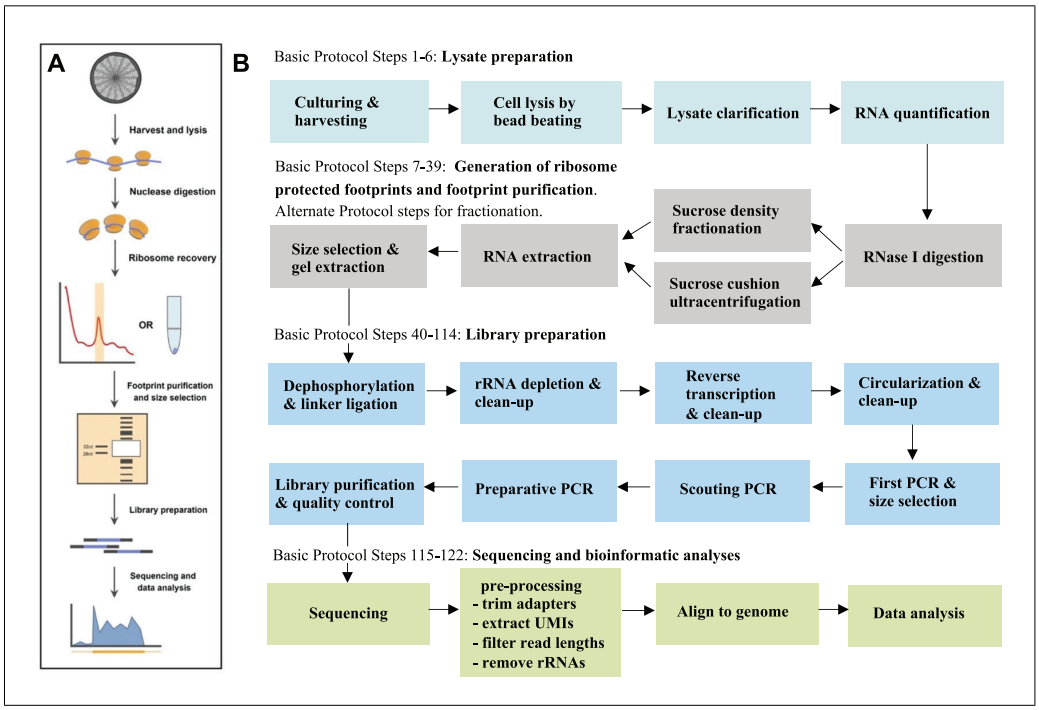


Figure 1 (**A**) Overview of the ribosome profiling protocol. The ribosome profiling protocol for *T. pseudonana* follows the same basic steps as for other cell types. Cells are harvested and lysed before nuclease digestion of mRNA unprotected by ribosomes is performed. mRNA fragments bound by ribosomes are recovered and purified prior to library preparation, sequencing, and computational data analysis. (**B**) Workflow diagram depicting all steps used in our ribosome profiling protocol. The four main parts include: (1) lysate preparation, (2) generation of ribosome protected footprints and footprint purification, (3) library preparation, and (4) sequencing and data processing. The corresponding steps in the Basic Protocol are listed. Additional steps outlined in the Alternate Protocol are necessary if sucrose density fractionation is performed.

this protocol can be adapted for use with other diatom species and may also facilitate the study of other marine algae.

The Basic Protocol involves a series of steps for monosome preparation, isolation of ribosome-protected fragments, and subsequent sequencing library preparation (Fig. 1B). We have also developed an Alternate Protocol that yields comparable results but utilizes a sucrose gradient instead of a sucrose cushion, therefore requiring additional instrumentation. This Alternate Protocol enables quality control of the samples prior to sequencing library preparation, facilitating the adaption of ribosome profiling to other diatom species, which typically requires the optimization of reaction conditions, such as those used for limited nucleolytic digestion.

STRATEGIC PLANNING

To ensure optimal results, ribosome profiling should be carried out using exponentially growing *T. pseudonana* cells. For all experiments, we recommend a minimum of three biological replicates per condition to enhance data robustness. Additionally, to limit ribosomal run-off and RNA degradation, several steps of the protocol must be carried out in a cold room or on ice. As this protocol utilizes potentially hazardous chemicals such as cycloheximide and organic solvents, careful handling and access to a fume hood is crucial. Following monosome purification, it is imperative to maintain an RNase-free environment, using dedicated workspaces and RNase-free reagents. To minimize variation due to batch effects, it is advised to process samples from different conditions together by the

same individual and by using the same lot of reagents. We suggest a sequencing depth of 20 million reads per sample for standard gene expression analyses.

BASIC PROTOCOL

RIBOSOME PROFILING IN THALASSIOSIRA PSEUDONANA

The following protocol is designed to generate high-quality ribosome profiling data to quantify translation in *T. pseudonana* CCMP1335. Initially, cells are promptly harvested and flash-frozen in liquid nitrogen. After mechanical disruption of the cells, optimized RNase digestion conditions are employed to create ribosome-protected footprints. Monosomes are subsequently isolated using sucrose-cushion ultracentrifugation. Alternatively, ribosomes can be purified via sucrose density gradient fractionation as detailed in the Alternate Protocol. This protocol typically yields ∼10 to 50 ng of 26 to 31 nt RNA fragments (corresponding to ∼1 to 5 pmol of RNA), which serves as the optimal amount of starting material for sequencing library preparation. We have incorporated an rRNA removal step via subtractive hybridization with the aim to increase the fraction of informative reads that map to mRNAs.

Materials

Thalassiosira pseudonana clone CCMP1335 (Bigelow; see Internet Resources)
Aquil medium (Price et al., 1989; see Internet Resources)
Polysome resuspension buffer (PRB) (see recipe)
Sucrose cushion solution (see recipe)
Qubit RNA BR assay kit (ThermoFisher Scientific, cat. no. Q10210)
100 U/µl RNase I (ThermoFisher Scientific, cat. no. AM2294)
20 U/µl SUPERase-In RNase inhibitor (ThermoFisher Scientific, cat. no. AM2694 or AM2696)
TRizol reagent (ThermoFisher Scientific, cat. no. 15596026)
Direct-zol RNA MiniPrep kit (Zymo Research, cat. no. R2050)
3 M sodium acetate (NaOAc), pH 5.5 (ThermoFisher Scientific, cat. no. AM9740)
GlycoBlue (ThermoFisher Scientific, cat. no. AM9516)
Isopropanol (Carl Roth, cat. no. 9781)
1 M Tris·HCl, pH 8 (ThermoFisher Scientific, cat. no. 15568025)
RNaseZap (ThermoFisher Scientific, cat. no. AM9780)
MilliQ H_2O
15% urea-PAGE gel (see recipe)
1× TBE prepared from 10× TBE stock solution (Promega, cat. no. V4251)
2× denaturing sample loading buffer (see recipe)
NI-800 and NI-801 size marker oligos (see Table 1)
10,000× SYBR Gold (ThermoFischer Scientific, cat. no. S11494)
RNA gel extraction buffer (see recipe)
Ethanol (Carl Roth, cat. no. 9065)
Qubit microRNA assay kit (ThermoFisher Scientific, cat. no. Q32881)
T4 polynucleotide kinase (PNK) and buffer (New England Biolabs, cat. no. M0201)
RNaseOUT recombinant ribonuclease inhibitor (ThermoFisher Scientific, cat. no. 10777019)
Pre-adenylated rApp-L7 (see Table 2)

Table 1 Size Marker Oligonucleotides

Size marker oligos	Sequence[a]
Upper size marker, NI-800:	5′-AUGUACACUAGGGAUAACAGGGUAAUCAACGCGA/3Phos/
Lower size marker, NI-801:	5′-AUGUUAGGGAUAACAGGGUAAUGCGA/3Phos/

[a]Marker oligonucleotides used in this study are the same as described in McGlingy and Ingolia (2017) and are ordered as RNA oligonucleotides and sourced from Eurofins Genomics.

T4 RNA Ligase 2, truncated KQ and buffer (New England Biolabs, cat. no. M0373)
PEG400 (Sigma Aldrich, cat. no. 202398)
Dimethylsulfoxide (DMSO) (Sigma Aldrich, cat. no. D8418)
Depletion oligonucleotides (see Table 3)
20× SSC buffer (ThermoFisher Scientific, cat. no. 15557044)
Dynabeads MyOne streptavidin C1 (ThermoFisher Scientific, cat. no. 65001)
1× and 2× binding and washing buffer (see recipe)
Dynabeads MyOne silane (ThermoFisher Scientific, cat. no. 37002D)
RLT buffer (Qiagen, cat. no. 79216)
Phusion High-Fidelity DNA polymerase & dNTP mix (ThermoFisher Scientific, cat. no. F530N)
RT Oligo P7-circ (see Table 2)
SuperScript III reverse transcriptase and kit components (ThermoFisher Scientific, cat. no. 18080093 or 18080044)
1 M NaOH (Carl Roth, cat. no. 9356)
1 M HEPES, pH adjusted to 7.3 with NaOH (Carl Roth, cat. no. 9195)

Table 2 Oligonucleotides for Library Preparation

Name	Sequence[a,b,c]
rApp-L7 Adapter[d]	/5rApp/NNNAGATCGGAAGAGCACACGTCTGA/3ddC/
RT Oligo P7 - circ[e]	/5Phos/NNNNNNNAGATCGGAAGAGCGTCGTGT/iSp9/GATTCAGACGTGTGC
P5Solexa_*s[f]	ACACGACGCTCTTCCGATC*T
P7Solexa_*s[f]	GGAGTTCAGACGTGTGCTCTTCCGATC*T
P5Solexa	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
TrueSeq_P7_Index[g]	CAAGCAGAAGACGGCATACGAGAT-X-GTGACTGGAGTTCAGACGTGTGCTCTTCC

[a]Oligonucleotides used for library preparation are the same as described in Meindl et al. (2023).
[b]All oligonucleotides are HPLC-purified.
[c]Store and dilute primers in a buffered solution (e.g., 10 mM Tris·HCl, pH 8.0-8.5). To limit the number of freeze-thaw cycles, store as aliquots at −20°C.
[d]5' ribo-adenylated, 3' protected by di-deoxy nucleotide (ddC), contains a 3 nt 5' randomized sequence to minimize ligation bias and to serve as a UMI for the bioinformatic identification of PCR duplicates.
[e]This RT oligonucleotide serves as a circularization adapter containing a short sequence on its 3' end that serves as a primer for reverse transcription. Separated by a PEG spacer (denoted iSp9 in the sequence) is a ligation adapter for the P7 side of the amplicon. It contains a 5' phosphate group followed by seven degenerate nucleotides that minimize ligation bias and that serve as UMIs for the bioinformatic identification of PCR duplicates.
[f]To prevent degradation by the proof-reading activity of the polymerase, it is recommended to include a phosphorothioate bond at the 3' terminal end of the primer (highlighted as asterisk).
[g]X marks the position of a 6 nt barcode for experimental multiplexing. See Internet Resources, Illumina TrueSeq single indexes.

Table 3 Set of 8 Biotinylated Oligonucleotides for rRNA Depletion in *T. pseudonana*

Depletion oligo[a,b]	Sequence	Target
#1	/5biotin-TEG/ATTCACACGGAATTCCACGTGCTCCATGCTACTT	23S chloroplast
#2	/5biotin-TEG/ACCTCAGCCCGACCCAGGACATGAACAGAAGGCACT	28S
#3	/5biotin-TEG/TACTTCTCCAGACCACAATTCGATGCAC	5.8S
#4	/5biotin-TEG/GGCTACTCTCATGCTTACAGTACCACGTTCCTT	28S
#5	/5biotin-TEG/GGCTAGCTTAGCCTTCTCCGTCCCTCGA	23S chloroplast
#6	/5biotin-TEG/AAAAGCGCAATGTGCGTTCAAAAGTTTGATGATTCAC	5.8S
#7	/5biotin-TEG/GACACAATGATAGGAAGAGCCGACATCGA	28S
#8	/5biotin-TEG/AAAGTTTGAGAATAGGCCGAGGGCA	28S

[a]The primers are used for rRNA depletion via subtractive hybridization.
[b]The DNA oligonucleotides contain a 5' biotin tag and are HPLC purified to ensure highest quality.

CircLigase II ssDNA ligase (Lucigen, cat. no. CL4111K or CL4115K)
10 mM ATP solution (ThermoFisher Scientific, cat. no. AM8110G)
P5Solexa_*s oligonucleotide (see Table 2)
P7Solexa_*s oligonucleotide (see Table 2)
ProNex size-selective purification system (Promega, cat. no. NG2001)
P5Solexa oligonucleotide (see Table 2)
TrueSeq P7 Index 'X' oligonucleotide (see Table 2)
6× TBE loading buffer (see recipe)
7% PAA-TBE gel (see recipe)
Ultra low range DNA ladder (ThermoFischer Scientific, cat. no. 10597012)
Ethidium bromide (Carl Roth, cat. no. 2218)
5 mg/ml linear acrylamide (ThermoFisher Scientific, cat. no. AM9520)
High Sensitivity D1000 reagents (Agilent Technologies, cat. no. 5067-5585)

Vacuum pump for cell filtration (Welch Vacuum, model 2534C-02)
Nalgene reusable bottle top filters (ThermoFisher Scientific, cat. no. DS0320-5045)
47-mm diameter 1.2-µm isopore filters (Merck, cat. no. RTTP04700)
Laminar flow hood (Walker Safety Cabinets, model: Class II 1290 Recirc Gen 6 or equivalent)
Tweezers (ThermoFisher Scientific, cat. no. 15699310)
1.5-ml cryogenic vials (Carl Roth, cat. no. AET4)
Liquid nitrogen
425- to 600-µm glass beads (Merck, cat. no. G8772)
50-ml centrifuge tubes (ThermoFisher Scientific, cat. no. 339652)
Bead beater (BioSpec 3110BX Mini-BeadBeater-1, or equivalent)
Ice
Refrigerated microcentrifuge 5418R, 4°C (Eppendorf, cat. no. 5401000010)
Pipettes and tips (ThermoFisher Scientific, cat. no. 05-403-151)
NanoDrop One (ThermoFisher Scientific, cat. no. ND-ONE-W)
Qubit 4 fluorometer (ThermoFisher Scientific, cat. no. Q33238)
Qubit assay tubes (ThermoFisher Scientific, cat. no. Q32856)
11 × 34-mm polycarbonate ultracentrifuge tubes (Beckman, cat. no. 343778)
Optima TLX ultracentrifuge (Beckman, cat. no. 361545)
TLA 120.2 rotor (Beckman, cat. no. 357656)
1.5-ml microcentrifuge tubes (Fisher Scientific, cat. no. 11926955)
Dry ice
Heat block (VWR, cat. no. 12621-104)
Mini-Protean vertical electrophoresis cell (BioRad, cat. no. 1658004)
Blue light transilluminator, e.g., DarkReader (Clare Chemical Research, cat. no. DR46B)
Scalpel blades (Merck, cat. no. S2646)
18-G needle (VWR, cat. no. BD305195)
0.5-ml tubes (ThermoFisher Scientific, cat. no. AB0533)
Corning Costar Spin-X column centrifuge tube filters (Sigma Aldrich, cat. no. CLS8160)
DNA LoBind tubes (Eppendorf, cat. no. 0030108051)
Vortex
DynaMag-2 magnetic separation rack (Invitrogen, cat. no. 12321D)
Tube shaker
0.2-ml tubes (ThermoFisher Scientific, cat. no. AB0620)
T100 thermal cycler (BioRad, cat. no. 1861096)
Automated gel purification system, e.g., BluePippin (optional)
High Sensitivity D1000 ScreenTape (Agilent Technologies, cat. no. 5067-5584)

Agilent 2200 TapeStation nucleic acid system (Agilent Technologies, cat. no. G2965AA)
Illumina sequencing system
Cutadapt (see Internet Resources)
UMI-tools (see Internet Resources)
bowtie2 and genome sequences (see Internet Resources)
STAR aligner (see Internet Resources)
R software (see Internet Resources)
Rstudio software (see Internet Resources)
riboWaltz (see Internet Resources)

Cell growth

1. Grow *T. pseudonana* CCMP1335 to mid-exponential phase (5–7 $\times$ 10^5 cells/ml) in Aquil medium under desired experimental conditions.

   T. pseudonana cells are routinely grown in 24 hr light (100-140 μE) at 20°C.

Cell harvest

2. Collect 2 $\times$ 200 ml of culture by gentle vacuum filtration (using a vacuum pump and a reusable bottle top filter) onto a 47-mm filter under a laminar flow hood. Use sterile tweezers to roll up the filter membrane and insert it into a previously labeled 1.5-ml cryogenic vial. Immediately flash-freeze in liquid nitrogen (<1 min).

   Two replicates of 200 ml culture each are needed per sample to provide enough starting material. Replicates can be pooled again at a later stage of the protocol (step 12). Keep the time between filtration and flash-freezing under 60 s to preserve the translation state of the cell as accurately as possible. Choose a gentle filtration speed to avoid posing another stress factor or breaking of the cells.

Cell lysis

3. Add 400 µl of pre-chilled polysome resuspension buffer (PRB) to the tube with slightly thawed cells. Rinse off filter and remove with sterile tweezers. Add two scoops of glass beads (425- to 600-µm).

   Lysis is done by bead beating of the frozen cell pellets. Prepare enough PRB (at least 1150 µl per tube) for lysis and sucrose cushion (step 10) using a 50-ml centrifuge tube.

4. Treat sample in bead beater for 2 min (max speed) with a break on ice after 1 min to prevent overheating.

5. Clarify the lysate by centrifugation for 10 min at 14,000 $\times$ *g*, 4°C using a microcentrifuge. Transfer the supernatant using a pipette in a new 1.5-ml microcentrifuge tube and place on ice.

6. Measure UV absorption at 260 nm of the lysate using a NanoDrop to estimate the RNA concentration.

   NanoDrop measurements only provide a rough estimate of the RNA concentration due to metabolites and light harvesting complexes present in the crude diatom lysate which absorb UV light at 260 nm. If a more precise quantification of the RNA in the lysate is required, use a Qubit Fluorometer and assay tubes with the Qubit RNA BR Assay Kit (note that in this protocol the conditions for RNA digestion have been optimized based on UV absorption measurements).

RNA digestion

7. Take ~60 µg of total RNA and dilute to 300 µl with PRB if necessary. Add 7 U RNase I (100 U/µl) and incubate for 45 min at room temperature with gentle mixing.

8. Put on ice and add 2 µl SUPERase-In RNase inhibitor (20 U/µl) to stop nuclease digestion.

Ribosome recovery

9. Transfer the sample to a 11 × 34-mm polycarbonate ultracentrifuge tube.
10. Underlay sample with 750 µl 1 M sucrose cushion solution by positioning a pipette tip at the bottom of the tube and gently dispensing the sucrose cushion solution.

    The lysate should float above the sucrose solution, with a visible interface between the layers. Ensure the tubes are balanced prior to centrifugation and add PRB, if necessary.

11. Pellet ribosomes by centrifugation in an ultracentrifuge using a TLA120.2 rotor, 2 hr at 199,000 × *g* (75,000 rpm), 4°C.

    While pelleted ribosomes are translucent, the pellet may also contain aggregates of the photosystems, which can aid in their identification. As a precaution, prior to removing the tubes from the rotor, it is recommended to mark their outside where the pelleted ribosomes are expected to be located.

12. Remove the supernatant and resuspend the pellet in 350 µl of TRizol reagent. Transfer to a 1.5-ml microcentrifuge tube. Replicates of the same sample can be pooled at this point.
13. Purify resuspended RNA by using the Direct-zol RNA MiniPrep kit following the manufacturer's instruction for purification of total RNA. Elute RNA in 50 µl RNase-free H_2O provided with the kit.
14. Precipitate eluted RNA by adding 10 µl 3 M NaOAc pH 5.5, 1.5 µl GlycoBlue (to facilitate precipitation) and 38.5 µl RNase-free H_2O, followed by 150 µl isopropanol.
15. Carry out precipitation on dry ice >30 min or at −20°C overnight.
16. Pellet RNA by centrifugation for 30 min at 14,000 × *g*, 4°C using a microcentrifuge. Pipette all liquid from the tube, place sideways and air dry for 10 min.
17. Resuspend pellet in 5 µl 10 mM Tris·HCl, pH 8 prepared from the 1 M stock solution.

    Samples can be stored overnight at −20˚C or for several months at −80°C.

Ribosomal footprint fragment purification

It is critical to work in an RNase-free environment from this step onwards. Decontaminate the electrophoresis apparatus and other equipment with RNaseZap. MilliQ H_2O can be used to prepare the running buffer.

18. Prepare 15% urea-PAGE gel.

    Large gels of ∼20 cm × 20 cm × 1 mm are recommended for maximal resolution. If not available, smaller gels can be used.

19. Pre-heat a heating block to 80°C.
20. Pre-run gel in the electrophoresis cell at 15 V/cm for 15 min in 1× TBE.
21. Prepare RNA sample from step 17 by adding 5 µl of 2× denaturing sample loading buffer. To prepare marker oligos for two lanes, mix 1 µl lower marker oligo NI-800 (10 µM), 1 µl upper marker oligo NI-801 (10 µM) (see Table 1), 10 µl 2× denaturing sample loading buffer and 8 µl 10 mM Tris·HCl, pH 8.
22. Denature sample and marker oligos for 90 s at 80°C using the heat block. Keep on ice until loading.
23. Load sample onto the gel with marker oligo samples framing the sample lane.
24. Run gel at 15 V/cm until the lower dye (light blue) front has reached the lower third of the gel.

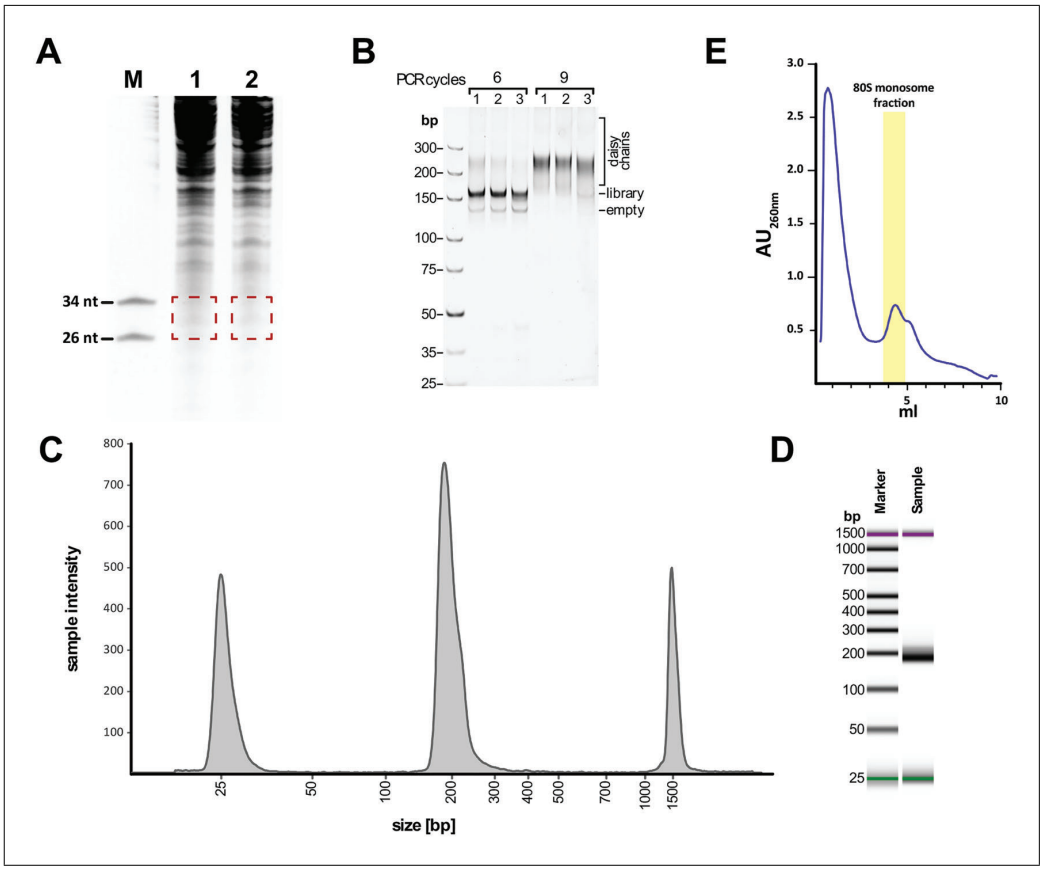


Figure 2 (**A**) Image of a typical ribosome footprint size selection gel from Basic Protocol step 26. Marker oligonucleotides (26 nt and 34 nt) are indicated on the left (M). Lane 1 and 2: digested RNA prepared after sucrose cushion ultracentrifugation. The red boxes highlight the regions excised from the gel containing ribosome-protected fragments (RPFs). (**B**) Results of a scouting PCR from Basic Protocol step 108. Three sequencing libraries were prepared from 26-34 nt RNA fragments and subjected to either 6 or 9 cycles of PCR amplification (as indicated at the top) followed by PAGE analysis, demonstrating that 6 cycles were the optimal number for library amplification. The sizes of empty amplicons, amplicons with insert (library), and heteroduplex DNA ('daisy chains') are indicated on the right. The size of the final amplicon (with a ~30 nt insert) is ~170 bp. (**C**) and (**D**) Tape station analysis of the final, purified sequencing library. (**E**) Density gradient analysis of a *T. pseudonana* extract treated with RNase I (7 U/µg of RNA) following the Alternate Protocol (step 7). Plotted is the UV absorbance ($\lambda = 260$ nm) of the sample; the fraction containing 80S ribosomal monosomes is highlighted in yellow, while the other fractions contain ribosomal subunits and polysomes.

25. Stain gel for 5 min with 1× SYBR gold in 1× TBE with gentle shaking.
26. Visualize gel on a blue light transilluminator and cut out the 26 to 34 nt region of the gel using a scalpel blade (see Fig. 2A).

Gel extraction

The gel extraction step has been adapted from Mohammad and Buskirk (2019).

27. Using an 18-G needle, poke holes into a 0.5-ml tube. Place the gel cut out into the tube.
28. Place the 0.5-ml tube into a 1.5 ml-tube.
29. Spin 5 min at 14,000 × *g*, room temperature, in a microcentrifuge until gel extrudes into bottom tube.
30. Add 400 µl RNA extraction buffer and 2 µl SUPERase-In.

31. Incubate overnight at 4˚C with gentle shaking.
32. Spin 5 s at 14,000 × *g*, room temperature, and transfer gel slurry into a Spin-X column.
33. Spin 3 min at 14,000 × *g*, room temperature, and transfer the eluate into a new low bind (LoBind) tube.
34. Precipitate RNA by adding 1.5 µl GlycoBlue, mix well and add two volumes 100% EtOH. Vortex briefly.
35. Incubate at −80°C for 1 hr.
36. Pellet by centrifugation for 30 min at 14,000 × *g*, 4°C.
37. Wash pellet with 300 µl cold 80% EtOH, centrifuge 5 min at 14,000 × *g*, 4°C, remove all liquid and briefly let air dry (<10 min).
38. Resuspend in 10 µl nuclease-free H_2O.
39. Measure concentration using the Qubit microRNA Assay Kit.

Samples can be stored overnight at −20˚C or for extended periods of time at −80°C.

Library preparation

For preparation of sequencing libraries, we followed the protocol by Meindl et al. (2023). With this protocol sequencing libraries can be prepared from as little as 0.1 pmol of RNA fragments; an optimal amount of starting material is 1 to 5 pmol of RNA (from step 38). To minimize sample loss, all reactions until PCR amplification of the final library (step 97) should be carried out in low bind reaction tubes.

We additionally implemented a subtractive hybridization step for rRNA removal using eight biotinylated depletion oligonucleotides representing abundant rRNA contaminants in *T. pseudonana*. The custom depletion step was optimized from Zinshteyn et al., and Green (2020).

RNA dephosphorylation to remove of terminal phosphates

40. For removal of the 2'-3'-cyclic phosphate that is generated by the RNase I cleavage, set up the following reaction:

Dephosphorylation mix		
x	µl	RNA in H_2O (from step 38)
2	µl	10 × T4 PNK buffer (supplied with the enzyme)
0.5	µl	40 U/µl RNaseOUT
0.5	µl	T4 PNK (with 3´phosphatase activity)
	add nuclease-free H_2O to 20 µl	

41. Mix all components well and incubate for 30 min at 37°C.
42. Inactivate the enzyme for 20 min at 65°C.

3´-adapter ligation

43. Prepare ligation mix (35 µl) using 3´dephosphorylated RNA adapters:

Ligation of adapter to 3' end of RNA		
20	µl	reaction from previous step
1	µl	20 µM pre-adenylated rApp-L7 adapters (see Table 2)
1.5	µl	10 × T4 RNA-ligase buffer (supplied with the enzyme)
1	µl	T4 RNA ligase 2, truncated KQ
6	µl	50% PEG400
4.5	µl	DMSO
1	µl	nuclease-free H_2O

44. Incubate overnight at 16°C.
45. Incubate 1 hr at 37°C.
46. Inactivate the enzyme for 15 min at 65°C.

Ribosomal RNA depletion

47. Dilute depletion oligos (see Table 3) in 4× SSC buffer (prepared from 20× stock solution) to concentration of 2.5 μM for each oligo and store at −20˚C.
48. Add 10 μl of depletion oligo mix solution to 35 μl of 3′-adapter ligation product from step 46.
49. Incubate the mixture at 80°C for 2 min to denature RNA.
50. Allow to cool down slowly to 25°C (temperature ramp 3°C/min).
51. Resuspend Dynabeads MyOne streptavidin C1 by vortexing the tube at medium speed.
52. Transfer 80 μl of bead suspension into a new low bind tube.
53. Put on a magnetic rack and wait for 1 min.
54. Aspirate and discard all supernatant.
55. Add 80 μl of 1× binding and washing buffer and agitate tube to resuspend beads. Place on magnetic rack and wait for 1 min.
56. Aspirate and discard all supernatant.
57. Repeat wash step twice.
58. Resuspend beads in 45 μl 2× binding and washing buffer.
59. Add the cooled sample (45 μl) from step 50 to the prepared beads (45 μl) from step 58 and mix well to homogeneously resuspend the beads.
60. Incubate at 25°C for 15 min (with shaking at 500 rpm).
61. Place on magnetic rack for >1 min to attract the beads to the magnet, then carefully transfer supernatant into a new tube.

At this point, RNA can be stored at −20°C overnight or −80°C for up to a month.

Clean-up (1)

62. Thoroughly mix the solution of magnetic MyONE Silane beads, using 20 μl beads per sample.
63. Wash the beads with 500 μl RLT buffer. Then, resuspend them in 650 μl RLT buffer per sample and add them to the sample. Mix the solution well.
64. Add 720 μl 100% ethanol and gently mix the solution by pipetting. Incubate the mixture for 5 min, mix again and incubate for another 5 min.
65. Magnetically attract the beads by placing the sample on the magnetic rack and discard the supernatant.
66. Resuspend the beads in 1 ml of freshly prepared cold 80% ethanol and transfer the mix into a new tube.
67. Wash again with 80% ethanol. Incubate 30 s before putting the sample on the magnet. Repeat this wash step.

68. Spin 5 s at 14,000 × *g*, room temperature, magnetically attract the beads and discard the supernatant. Air-dry the beads for 5 min at room temperature and resuspend them in 12 µl nuclease-free H_2O.

69. Incubate the mix for 5 min at room temperature. Magnetically attract the beads and add the eluate to the reverse transcription (RT) mix of the next step (step 70).

Reverse transcription

70. Prepare the following mix in 0.2-ml tubes:

RNA and primer mix		
12	µl	pooled RNA in H_2O (from step 69)
1	µl	10 mM dNTP mix
1	µl	0.5 pmol/µl RT Oligo P7-circ (see Table 2)

71. In a thermal cycler, heat the sample to 70°C for 5 min and incubate at 25°C until RT mix added, mix by pipetting. Do not put on ice to prevent unspecific annealing of the oligonucleotide used to prime reverse transcription.

72. Prepare the following RT mix and add it to the sample from the previous step (resulting in a total volume of 20 µl):

RT mix		
4	µl	5× first strand buffer (supplied with the enzyme)
1	µl	0.1 M DTT (supplied with the enzyme)
0.5	µl	200 U/µl Superscript III RT
0.5	µl	40 U/µl RNaseOUT

73. Incubate as follows:

25	°C	5	min
42	°C	20	min
50	°C	40	min
4	°C	∞	

74. To hydrolyze the RNA, add 1.65 µl 1 M NaOH and incubate at 98°C for 20 min. Then add 20 µl 1 M HEPES, pH 7.3 to neutralize the solution.

Clean-up (2)

75. Thoroughly mix the solution of magnetic MyONE Silane beads, using 10 µl of beads per sample.

76. Wash beads with 500 µl RLT buffer. Then, resuspend them in 650 µl RLT buffer per sample and add them to sample. Mix the solution well.

77. Add 720 µl 100% ethanol and gently mix the solution by pipetting. Incubate the mixture for 5 min, mix again and incubate another 5 min.

78. Magnetically attract the beads by placing the sample on the magnetic rack and discard the supernatant.

79. Resuspend the beads in 1 ml of freshly prepared cold 80% ethanol and transfer the mix into a new tube.

80. Wash again with 80% ethanol. Incubate 30 s before putting the sample on the magnet. Repeat this wash step.

81. Spin the mix 5 s at 14,000 × *g*, room temperature, in a microcentrifuge, magnetically attract the beads and discard the supernatant. Air-dry the beads for 5 min at room temperature and resuspend them in 14 µl nuclease-free H_2O.

82. Incubate the mix for 5 min at room temperature. Magnetically attract the beads and add the eluate to the reactions mix of the next step (step 83).

Circularization of the RT product

83. Prepare the following ligation mix (amounts given per sample):

Ligation mix		
14	µl	cDNA (from step 82)
2	µl	10 × CircLigase buffer (supplied with the enzyme)
1	µl	10 mM ATP
1	µl	50 mM $MnCl_2$ (supplied with the enzyme)
2	µl	CircLigase II

84. To ensure homogeneity, mix the ligation master-mix by vigorous stirring, pipetting, and flicking. Centrifuge the mix 5 s at 14,000 × *g*, room temperature, in a microcentrifuge to collect all drops.

85. Incubate 2 hr at 60°C in a thermocycler.

86. Incubate at 80°C for 10 min to inactivate the CircLigase.

Clean-up (3)

87. Thoroughly mix the solution of magnetic MyONE Silane beads, using 10 µl MyONE Silane beads per sample.

88. Wash beads with 500 µl RLT buffer. Then, resuspend them in 650 µl RLT buffer per sample and add them to sample. Mix the solution well.

89. Add 720 µl of 100% ethanol and gently mix the solution by pipetting. Incubate the mixture for 5 min, mix again and incubate another 5 min.

90. Magnetically attract the beads by placing the sample on the magnetic rack and discard the supernatant.

91. Resuspend the beads in 1 ml of freshly prepared cold 80% ethanol and transfer the mix into a new tube.

92. Wash again with 80% ethanol. Incubate 30 s before putting the sample on the magnet. Repeat this wash step.

93. Briefly spin the mix 5 s at 14,000 × *g*, room temperature, in a microcentrifuge, magnetically attract the beads and discard the supernatant. Air-dry the beads for 5 min at room temperature and resuspend them in 23 µl nuclease-free H_2O.

94. Incubate the mix for 5 min at room temperature. Magnetically attract the beads and transfer the solution to a fresh tube.

First PCR (cDNA pre-amplification)

95. Prepare 2× Phusion HF master mix on ice (25 µl required per reaction, volumes provided are for 100 µl):

2× Phusion HF master mix		
40	µl	5× Phusion HF buffer
4	µl	10 mM dNTPs
54	µl	nuclease-free H_2O
2	µl	Phusion polymerase

96. Prepare the following PCR mix:

PCR mix		
15	µl	circularized RT product (from step 94)
7.5	µl	nuclease-free H_2O
2.5	µl	primer mix of P5Solexa_*s and P7Solexa_*s, 10 µM each (see Table 2)
25	µl	2× Phusion HF master mix

97. Run the following PCR:

98	°C	30	s	
98	°C	10	s	
65	°C	30	s	perform amplification (steps 2 to 4) 6×
72	°C	30	s	
72	°C	3	min	
16	°C	∞		

Size selection to remove primer-dimers

To remove excess primer dimers, size-select your samples with ProNex size-selective purification system.

98. Equilibrate the ProNex size-selective chemistry (beads) to room temperature for 30 min and resuspend by vigorous vortexing.

99. For 50 µl of sample, add 147.5 µl ProNex size-selective chemistry. This is a 1:2.95 (v/v) ratio of sample to beads. Mix by pipetting 10× up and down.

100. Incubate the mixture at room temperature for 10 min.

101. Place the samples on a magnetic stand for 2 min. Discard the supernatant.

102. Leave the tube on the magnetic stand so that the beads remain attracted to the side of the tube, then add 300 µl ProNex wash buffer. If necessary, add more ProNex wash wuffer to cover all beads on the magnet. While the beads are magnetically attracted, incubate the ProNex wash buffer for 30 to 60 s before removal.

    When resuspending the samples in the wash buffer, do not remove ProNex beads from the magnet. This can cause up to 20% sample loss. For larger samples, increase the volume of ProNex wash buffer proportionally to the volume of sample and beads.

103. Repeat the last wash of the magnetically attracted beads with another 300 µl ProNex wash buffer for 40 to 60 s. Discard the supernatant and allow the samples to air-dry for ∼8 to 10 min (<60 min) until cracks are visible in the bead pellet.

104. Remove the beads from the magnetic stand and elute samples in 23 µl ProNex elution buffer. Resuspend all samples by pipetting and let them stand for 5 min at room temperature.

    Samples can be stored at −20˚C for several months.

Second PCR amplification-cycle optimization

Try two different cycle numbers for amplification of the sample, as described in Buchbender et al. (2020). A good starting point for cycle optimization is a range of 6-12 cycles. Ideally, you should observe enough product without overamplification. Overamplification is indicated by the appearance of large assemblies of improperly annealed, partially double-stranded, heteroduplex DNA migrating above the library (known as 'daisy chains', Fig. 2B; see also Huppertz et al., 2014).

105. Prepare the following master mix (10 µl per cycle number to be scouted):

PCR mix (per reaction)		
3.7	µl	nuclease-free H_2O
0.3	µl	primer mix of P5Solexa and TrueSeq P7 Index 'X', 10 µM each (see Table 2)
1	µl	cDNA (from step 104)
5	µl	2× Phusion HF master mix (see step 95)

106. Split into 10 µl reactions and run the following PCR with different cycle numbers:

98	°C	30	s
98	°C	10	s
65	°C	30	s perform amplification (steps 2 to 4) 6-12× *
72	°C	30	s
72	°C	3	min
16	°C	∞	

**Or more cycles as desired (see Buchbender et al., 2020).*

107. Combine 10 µl PCR product with 2 µl 6× TBE loading buffer. Run 6 µl of the amplified product on a small (∼7 × 10 cm) 7% PAA-TBE gel for 30 min at 200 V. Use 3 µl ultra low range DNA ladder (diluted 1:4 in loading dye), as a marker.

108. Stain gel for 10 min with ethidium bromide and visualize gel (see Fig. 2B).

Adhere to the necessary safety measures when working with ethidium bromide.

Preparative PCR

From your cycle optimization PCR results, estimate the minimum number of PCR cycles to use to amplify the library. Consider that the template for the reaction will be 2.5-times more concentrated (see PCR mix below), therefore one cycle less is needed than in the scouting PCR (steps 105 to 108: PCR cycle optimization).

For experimental multiplexing during sequencing, barcodes are introduced with the P7 primer. For additional information please refer to the Illumina TrueSeq Single Indexes (see Internet Resources). Ensure usage of compatible indices for all libraries to be sequenced on the same lane.

109. Prepare the following PCR mix:

PCR mix (per reaction)		
6.5	µl	nuclease-free H_2O
1	µl	primer mix of P5Solexa and TrueSeq P7 Index 'X', 10 µM each (see Table 2)
7.5	µl	cDNA (from step 104)
15	µl	2× Phusion HF master mix (see step 95)

110. Run the same PCR program as in step 106, but with the adjusted cycle number.

Purification of the sequencing library

Prior to sequencing, all primers and amplicons that do not contain an insert need to be removed. Empty amplicons have a length of 140 bp, while amplicons with ribosome-protected fragment inserts exhibit a length of ∼170 bp. Here we provide instructions for library purification using an automated agarose gel electrophoresis system. If this is not available in the lab, the library can be purified via a native 8% polyacrylamide gel as described previously (McGlincy & Ingolia, 2017).

111. Purify the PCR reaction using the automated agarose gel electrophoresis system (PippinPrep, or similar) using a 3% agarose cassette according to the manufacturer's instructions with the following settings: target fragment length 168 bp with the setting 'tight'.

112. Precipitate the purified library (30 µl) by the addition of 3 µl 3 M NaOAc pH 5.5, 2 µl linear acrylamide, and 90 µl ethanol. Incubate at least 30 min at −20°C, then centrifuge 30 min at 14,000 × *g*, 4°C. speed in a microcentrifuge. Aspirate supernatant, air dry pellet, and resuspend the pellet in 12 µl H_2O.

Quantitative and qualitative analysis of the purified sequencing library

For quality control and determination of the concentration of the libraries, a screen tape assay using the High Sensitivity D1000 ScreenTape and the High Sensitivity D1000 Reagents is employed on a TapeStation system.

113. Determine library concentration by measuring UV absorbance at 260 nm using the Nanodrop System.

114. Analyze an appropriate amount of the purified library using a high sensitivity screen tape assay on a TapeStation according to the manufacturer's instructions (Fig. 2C and D).

Sequencing and data analysis

115. Pool libraries as desired and subject to sequencing on an appropriate machine (e.g., Illumina NextSeq2000, P3 Reagents (50 Cycles), 80-85 cycles, single end read using Read1 primer, plus 6 nt Indexread).

    For a standard experiment, we recommend to sequence at least 20 million reads per library of which typically 15 ± 5% map uniquely to the genome.

116. After de-multiplexing, pre-process the sequencing reads by trimming of adapters using Cutadapt (version 2.8, adapter=AGATCGGAAGAGCACACGTCT, overlap=10, minimum-length=10, discard-untrimmed).

117. Extract unique molecular identifiers (UMIs) and append them to the read name using UMI-tools (version 1.0.1, extract-method=regex).

118. Filter reads by lengths and keep reads between 16 and 40 nt.

119. Align trimmed and filtered reads to an rRNA reference (e.g., from RNAcentral database) using bowtie2 (version 2.3.5.1).

120. Align all non-rRNA aligned reads to the Joint Genome Institutes (JGI) *T. pseudonana* reference genome (Thaps3_chromosomes_ assembly_ chromosomes_repeatmasked.fasta) with genome annotations (Thaps3.filtered_ proteins.FilteredModels2.gff3) using STAR (version, –quantMode Transcriptome-SAM GeneCounts, --outSAMattributes All, outFilterMismatchNmax 2).

121. Remove PCR duplicates using UMI-tools (extract-umi-method, read_id-method unique).

122. Analyze and interpret ribosome profiling data in R (version 3.5.1) and Rstudio using the riboWaltz package (version 1.2.0).

ALTERNATE PROTOCOL

RIBOSOME PROFILING PROTOCOL FOR DIATOMS USING SUCROSE GRADIENT FRACTIONATION

This protocol includes the same steps as the Basic Protocol, but with the addition of a sucrose density gradient instead of a sucrose cushion. This allows users to optimize nuclease digestion conditions to suit a wider range of diatom species as monitored by the collapse of polysomes into monosomes after nucleolytic digestion.

Additional Materials *(also see Basic Protocol)*

10% sucrose solution (see recipe)
50% sucrose solution (see recipe)
Sucrose chase solution (see recipe)
Phenol/chloroform/isoamyl alcohol 25:24.1 (Carl Roth, cat. no. A156)

Gradient Master gradient forming device and tube holders (Biocomp Instruments, cat. nos. B108 and 105-914A)
Ultracentrifuge (e.g., Beckman Optima L-100 K Ultracentrifuge) with SW41 rotor (Beckman, cat. no. 331362)
14 × 89-mm tubes for SW41 rotor (Beckman, cat. no. 331372)
siFractor (siTOOLs, cat. no. eq-F001-F0SW41)
ÄKTA FPLC (or similar)

Cell growth and harvesting, lysate preparation, and nucleolytic digestion

1. Follow steps 1 to 8 of the Basic Protocol.

Preparation of sucrose gradient

We recommend using a gradient master device (Biocomp Instruments) for rapid and reproducible formation of sucrose gradients. Here, 12 ml gradients are used with a SW41 rotor (Beckman). Conditions for monosome purification using different rotors are described in Meindl et al. (2023).

2. Pre-cool the ultracentrifuge and cool down the rotor and the buckets. Thaw all necessary reagents.
3. Prepare 10% and 50% sucrose solutions.

   It is recommended to prepare the sucrose solutions well in advance to allow air bubbles to dissipate.

4. Prepare linear sucrose density gradients using the Biocomp gradient master according to the manufacturer's instructions. Use the 'long caps' provided by the manufacturer with the following program: SW41 rotor, sucrose, long caps, 10%-50%, 11 steps. If required, the density gradients can be stored at 4°C for several hours.

Sucrose gradient centrifugation

5. Remove the caps from the gradient tubes and gently overlay the solution with the sample. Pool replicates of the same sample at this point. Balance the tubes carefully with PRB buffer.
6. Carefully transfer tubes into the buckets, close lids and insert buckets into the rotor.
7. Insert rotor into the centrifuge and spin 3 hr at 151,000 × *g* (35,000 rpm) at 4°C.

Density gradient fractionation

This protocol uses a gradient fractionation system which employs a tube piercing unit to deliver a dense chase solution to displace the gradient. It is connected to an FPLC system for analysis and fractionation. Further details are described in Meindl et al. (2023).

8. Connect the siFractor (siTOOLs Biotech) to an FPLC machine according to the manufacturer's instructions and thoroughly rinse the system with H_2O. Fill a super loop (or similar) with sucrose chase solution and connect it to the FPLC system. Prime the tubes with the chase solution.
9. Follow the manufacturer's instructions for the fractionation of the samples from the gradient tubes while continuously monitoring conductivity and UV absorbance at 260 nm. Collect fractions of 1 ml each. Figure 2E depicts a typical UV profile obtained with nuclease treated *T. pseudonana* cell extract.

   Pause point. Gradient fractions can be stored at −80°C for extended periods of time.

RNA isolation from density gradient fractions

10. Select the gradient fractions that contain your complexes of interest (e.g., 80S monosomes) and dilute 1:1 (v/v) with nuclease-free H_2O to prevent phase inversion during organic extraction (due to the high sugar concentration of the sample).
11. For organic extraction, transfer the samples to a fume hood and add an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1) and mix thoroughly by vortexing.
12. Separate the phases by centrifugation in a bench-top centrifuge for 10-30 min, 14,000 × *g*, room temperature, then transfer the upper aqueous phase into new tube. Make sure not to transfer any of the organic solution.

    IMPORTANT: make sure not to transfer any of the interphase.

13. Precipitate nucleic acids by adding 1/10 volume 3 M NaOAc (pH 5.2), 2.5 µl linear acrylamide, and 0.8 volumes of isopropanol. Mix by inverting the tubes several times. Incubate samples at −20°C for at least 30 min, then centrifuge in a microcentrifuge 30 min at 14,000 × *g*, 4°C.

14. Aspirate supernatant and wash the pellet with cold 80% ethanol. Aspirate the ethanol, air-dry the pellet and resuspend the RNA in 5 µl 10 mM Tris, pH 8.

Samples can be stored overnight at −20˚C or for several months at −80°C.

15. Continue with footprint fragment purification as outlined in the Basic Protocol step 18.

REAGENTS AND SOLUTIONS

Binding and washing buffer, 1×

5 mM Tris·HCl, pH 8 (ThermoFisher Scientific, cat. no. 15568025)
0.5 mM EDTA (Carl Roth, cat. no. 8043)
1 M NaCl (Carl Roth, cat. no. 3957)
Double the concentrations for a 2× buffer
Store indefinitely at room temperature

Denaturing sample loading buffer, 2×

98% (v/v) formamide (Carl Roth, cat. no. P040)
10 mM EDTA (Carl Roth, cat. no. 8043)
300 µg/ml bromophenol blue (Carl Roth, cat. no. A512)
Store indefinitely at −20°C

PAA-TBE gel, 7%

Mix in a fume hood:
3.5 ml 30% acrylamide/bis-acrylamide solution (19:1) (BioRad, cat. no. 1610154)
1.5 ml 10× TBE (Promega, cat. no. V4251)
10 ml MilliQ H_2O
For polymerization add:
120 µl 10% APS (prepare fresh) (Carl Roth, cat. no. 9592)
12 µl TEMED (Carl Roth, cat. no. 2367)
Mix gently and pour gel
Insert well comb and allow 30 min for gel to polymerize before use

Polysome resuspension buffer (PRB)

50 mM Tris·HCl, pH 8 (ThermoFisher Scientific, cat. no. 15568025)
100 mM NH_4Cl (Carl Roth, cat. no. K298)
200 mM sucrose (Carl Roth, cat. no. 9097)
10.5 mM magnesium acetate (Carl Roth, cat. no. P026)
0.5 mM EDTA (Carl Roth, cat. no. 8043)
1 mM DTT (Carl Roth, cat. no. 6908)
0.01% cycloheximide (Carl Roth, cat. no. 8682)
0.1% Triton X-100 (Carl Roth, cat. no. 3051)
25 U/ml Turbo DNase (ThermoFisher Scientific, cat. no. AM2238 or AM2239)
Mix all reagents in a fume hood
Prepare fresh before every use and keep on ice

RNA gel extraction buffer

300 mM NaOAc, pH 5.5 (ThermoFisher Scientific, cat. no. AM9740)
1 mM EDTA (Carl Roth, cat. no. 8043)
0.25% (w/v) SDS (Carl Roth, cat. no. CN30)
Store indefinitely at room temperature

Sucrose chase solution

60 mM Tris·HCl, pH 7.4 (Carl Roth, cat. no. 4855)
450 mM NaCl (Carl Roth, cat. no. 3957)
15 mM $MgCl_2$ (ThermoFisher Scientific, cat. no. R0971)
0.0001% Ponceau S (Carl Roth, cat. no. 5938)
60% sucrose (w/v) (Carl Roth, cat. no. 9097)
Store up to 3 days at 4°C

50 ml of solution are sufficient for four gradients

Sucrose cushion solution

1 M sucrose (Carl Roth, cat. no. 9097) in PRB (see recipe)
20 U/ml SUPERase-In (ThermoFisher Scientific, cat. no. AM2694, AM2696)
Prepare fresh before every use and keep on ice

Sucrose solution, 10%

20 mM Tris·HCl, pH 7.4 (Carl Roth, cat. no. 4855)
150 mM NaCl (Carl Roth, cat. no. 3957)
5 mM $MgCl_2$ (ThermoFisher Scientific, cat. no. R0971)
10% sucrose (w/v) (Carl Roth, cat. no. 9097)
Store up to 3 days at 4°C
Prior to use, add:
1 mM DTT (Carl Roth, cat. no. 6908)
100 µg/µl cycloheximide (Carl Roth, cat. no. 8682)

25 ml of solution is sufficient for four gradients

Sucrose solution, 50%

20 mM Tris·HCl, pH 7.4 (Carl Roth, cat. no. 4855)
150 mM NaCl (Carl Roth, cat. no. 3957)
5 mM $MgCl_2$ (ThermoFisher Scientific, cat. no. R0971)
50% sucrose (w/v) (Carl Roth, cat. no. 9097)
Store up to 3 days at 4°C
Prior to use, add:
1 mM DTT (Carl Roth, cat. no. 6908)
100 µg/µl cycloheximide (Carl Roth, cat. no. 8682)

25 ml of solution is sufficient for four gradients

TBE loading buffer, 6×

In 6.7 ml nuclease-free H_2O, dissolve 25 mg bromophenol blue (Carl Roth, cat. no. A512) and 25 mg xylene cyanole (Carl Roth, cat. no. A513). Add 3.3 ml glycerol (Carl Roth, cat. no 3783) and mix. Store indefinitely at −20°C.

Urea-PAGE gel, 15%

Mix in a fume hood:
3.75 ml 40% acrylamide/bis-acrylamide solution (19:1) (BioRad, cat. no. 1610144)
4.8 × g urea (Carl Roth, cat. no. 3941)
1 ml 10× TBE (Promega, cat. no. V4251)
Bring up to 10 ml with MilliQ H_2O
Warm at 37°C until dissolved before adding the following:
120 µl 10% APS (prepare fresh) (Carl Roth, cat. no. 9592)
12 µl TEMED (Carl Roth, cat. no. 2367)
Mix gently and pour gel
Insert well comb and allow 30 min for gel to polymerize before use

COMMENTARY

Background Information

Ribosome profiling provides comprehensive snapshots of cellular translation based on the sequencing of mRNA fragments that are protected by translating ribosomes from nucleolytic digestion (Ingolia et al., 2009). These ribosome-protected fragments (RPFs) yield positional information of ribosomes on mRNA and their quantification allows the determination of protein synthesis rates. Moreover, by normalization to RNA abundance, ribosome loading scores can be derived that reflect the translation efficiency of mRNA species. Ribosome profiling monitors the final step of gene expression and therefore allows detection of gene expression changes that occur at different levels (including transcriptional regulation) (Ingolia, 2016). Changes in protein synthesis rates (as measured by changes in RPFs from a locus) can hence be driven, e.g., by transcriptional regulation, RNA turnover, differential RNA processing, or regulated translation. This allows to detect changes in gene expression programs, e.g., induced by a changing environment.

A wide range of ribosome profiling protocols are available for different species. Due to the different nature of cells and ribosomes, these protocols vary in their methods of harvesting, lysis conditions, and the selected nuclease and digestion conditions. In eukaryotic organisms, ribonuclease (RNase) I, A, T1 and micrococcal S7 are commonly used but their cutting efficiencies are species-dependent (Gerashchenko & Gladyshev, 2017). In this protocol, we used RNase I which performed well for *T. pseudonana* (Fig. 2E).

Different methods have been used in ribosome profiling studies to isolate digested monosomes, the most prevalent ones being ultracentrifugation via sucrose gradient and sucrose cushion (e.g., Ingolia et al., 2012; Meindl et al., 2023). As described in the Alternate Protocol, density gradient centrifugation was performed to identify optimal nuclease digestion conditions for *T. pseudonana* which had not been established before. Moreover, density gradient centrifugation allows the purification of ribosomal monosomes without carrying over a large portion of messenger ribonucleoproteins (mRNPs) to the next steps in the protocol (McGlincy & Ingolia, 2017). However, it requires additional instrumentation that may not be readily available in all labs.

Ribosome profiling sequencing libraries are typically dominated by rRNA fragments. Despite measures to limit these contaminations, e.g., precise excision of RNA in the right size after denaturing poly-acrylamide gel electrophoresis (PAGE) using appropriate markers, contaminations typically represent ∼90% of the reads. Hence, different measures have been implemented for their depletion. To specifically target the contaminants, they are experimentally identified typically by shallow sequencing of ribosome profiling libraries prepared under the same conditions from the respective organism/sample. Several commercially available rRNA depletion kits have been tested in a recent study and were found generally unsuitable for use with ribosome footprint libraries (Zinshteyn et al., 2020), especially kits based on targeted nuclease cleavage can lead to ribosome footprint degradation, reduction of mappable reads, interference with global gene expression measurements and blurred nucleotide resolution. Moreover, commercial kits that target *T. pseudonana* rRNA contaminations is not available. Thus, for our protocol we designed biotinylated antisense oligonucleotides that target the common contaminants found in sequencing libraries generated from ribosomal profiling of *T. pseudonana* and implemented a subtractive hybridization step in the protocol.

Various protocols for library preparation from ribosome profiling samples have been described that differ, e.g., in aspects of linker sequence and ligation, or the purification of the reaction intermediates. More recently, single-pot reactions have been introduced that exploit the template switching activity of reverse transcriptase (Ferguson et al., 2023; Ozadam et al., 2023). Here, we employ a recently developed protocol for sequencing library preparation that is particularly suited for low input samples (Meindl et al., 2023) and hence for the processing of samples derived from *T. pseudonana* that can have a low yield. In brief, for sequencing library preparation, an adapter sequence is ligated to the ribosome protected fragments, followed by reverse transcription, circularization of the cDNA and PCR amplification. Purification of the reaction intermediates occurs via solid phase reversible immobilization. Degenerate nucleotides on the ends of the adapters used in the ligation and circularization reactions reduce ligation bias and are employed for the identification of PCR duplicates during bioinformatic analyses.

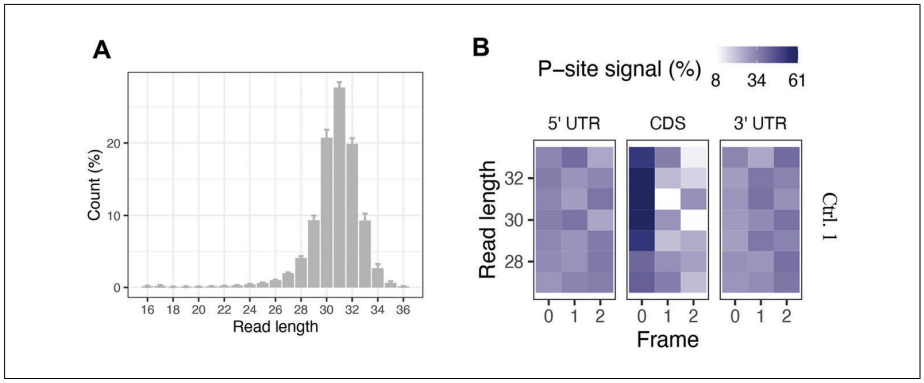


Figure 3 Quality control of ribosome profiling data. (**A**) Average read length distribution of all tested samples. Our results showed a peak of ribosome-protected fragments at 31 nt which is typical for ribosome profiling data. (**B**) Percentage of P-sites in the three translation reading frames along the 5' UTR, CDS and 3' UTR, stratified for read length (Sample Ctrl. 1), demonstrating that fragments of 30-32 nt have a strong frame preference along CDS.

Bioinformatic analysis of ribosome profiling data has been described in detail elsewhere (e.g., McGlincy & Ingolia, 2017). We mapped the sequencing reads against the *T. pseudonana* genome instead of the transcriptome due to its higher quality.

Taken together, the adjustments made to previously existing protocols result in high-quality sequencing data of actively translating ribosomes from *T. pseudonana*. Our protocol is robust and can now be used for future gene expression analysis in this organism, as well as a starting point for the adaption of the protocol to other related species.

Critical Parameters

For robust and reproducible results from ribosome profiling experiments several parameters are important, e.g., the way cells are harvested and lysed can have a profound effect on ribosome distribution by triggering of cellular stress responses (e.g., Mohammad et al., 2019). Also, the use of translation inhibitors has been debated since it can introduce artefacts in particular species (Gerashchenko & Gladyshev, 2014; Sharma et al., 2021). Furthermore, the nuclease(s) and conditions for the generation of ribosome protected fragments need to be carefully chosen. Nucleases that work well in one species, might result in a complete loss of the sample in others (Gerashchenko & Gladyshev, 2017). Hence, the conditions under which ribosome profiling is performed must be carefully chosen and adapted to the model system and research question pursued. Our protocol has been optimized for *T. pseudonana* and yields high-resolution footprints with decent frame information (see Fig. 3B). For other species, conditions suitable for ribosome profiling need to be experimentally established.

In all cases, ribosome profiling experiments need to be performed under conditions that prevent nucleolytic degradation of the samples, since shortening of the RPFs during library preparation compromises data quality. This includes working at low temperatures whenever possible, as well as the use of nuclease-free reagents and RNase inhibitors. Moreover, since the amount of input material for sequencing library preparation is typically rather limited, the use of low bind reaction tubes and low retention tips is highly recommended, once the RPFs have been excised from the gel. Experimental precision during critical steps, e.g., the size selection of the RPFs by denaturing page, limits contaminations and yields high quality data.

Proper amplification of the sequencing library is crucial for obtaining good results, since over-amplification increases the fraction of PCR duplicates in the final library. Yet, insufficient amplification of the sequencing library will produce only minute amounts of material that are often not sufficient for sequencing, the protocol provides a simple and robust step (scouting PCR) to identify optimal PCR conditions for the amplification of the library.

Troubleshooting

Since ribosome profiling is a complex technique with numerous factors contributing to success, troubleshooting of the experiments is complex. In the following, we provide a list of frequently encountered problems and how to deal with them.

Little yield after cell lysis

Incomplete recovery of sample from the filter: ensure to recover as much material as possible from the filter prior to cell lysis.

Incomplete lysis of the cells: check for complete lysis of the cells under the microscope, adjust lysis conditions, if required by, e.g., adding additional cycles in the bead beater.

Partial degradation of the sample during cell lysis: make sure to keep the sample cooled during cell lysis to prevent unwanted degradation.

Loss of sample or incomplete digestion after nuclease treatment of the extract

Amount of RNase used for digestion: we frequently observe batch to batch variation of the RNase I. Carefully titrate the enzyme to achieve optimal conditions.

RNase activity in the extract: supplement the extract with RNase inhibitors such as RNaseOUT (which does not affect RNase I) to inhibit endogenous nucleases.

Lysate too concentrated or too dilute: to ensure optimal nucleolytic digestion, ensure a total reaction volume of 300 µl when digesting 60 µg of total RNA.

Incorrect RNA quantification: quantify extracts as described in the protocol, adjust amount enzyme when required.

Loss of sample during sequencing library preparation or too many PCR cycles required to produce the sequencing library

Sample has been degraded by nucleases: use nuclease-free reagents, use nuclease inhibitors as indicated in the protocol, work on ice whenever possible, use filter tips to prevent nuclease carry-over from other reactions.

Sample loss due to unspecific adsorption onto surfaces: use low bind reaction tubes and low retention tips during library preparation.

Sample loss during clean-up: try not to lose beads when aspirating supernatant during the clean-up steps, do not remove beads from magnet during some of the washing steps (as indicated in the protocol); for ProNex size exclusion ensure the correct 1:2.95 v/v ratio of beads to sample as outlined in the protocol.

Low quality oligonucleotides: we observe batch to batch variation of the oligonucleotides used for library preparation; ensure that the oligonucleotides are of good quality (e.g., by running them on a denaturing PAGE), order new oligonucleotides if necessary.

Enzymatic reaction not efficient: test activity of the enzymes used for library preparation, order new batches if required.

Sequencing library contains too many empty amplicons

Input material for the library preparation was too low. Use more input material, and/or gel-purify the final library as outlined in the protocol.

Sequencing library dominated by a single experimental barcode

Contamination from a previous experiment: we highly recommend to physically separate pre-PCR and post-PCR work in two separate labs or use laminar flow hoods to prevent carryover between experiments.

Pooling of samples for sequencing: prior to pooling, quantify the amplicons as precisely as possible using a Qubit fluorimeter, or TapeStation analysis.

Understanding Results

Yield

The protocol provided provides instructions for the processing of a crude cell lysate containing ∼60 µg of total RNA. After RNA extraction, typically 10 to 15 µg of RNA can be recovered for gel purification, resulting in a yield of ∼10 to 50 ng of ∼30 nt RNA fragments (∼1 to 5 pmol of RNA), an ideal starting quantity for library preparation.

PCR cycles required for library amplification

The number of PCR cycles needed for the final amplification can be used as a crude indicator for the yield after sequencing library preparation. Low cycle numbers are observed when using large amounts of starting material, or they indicate a good overall performance during library preparation. Typically, libraries are obtained after 6 to 9 PCR cycles (Fig. 2B); higher PCR cycle numbers can result in increased fractions of PCR duplicates and empty amplicons.

Non-mRNA reads in the libraries (contamination)

Despite the depletion of contaminating rRNA sequences, it is expected to observe ∼60%-80% of total reads mapping to rRNA.

Data processing

We used the riboWaltz R package to demonstrate the quality of the ribosome profiling data and to visualize it. Typically, eukaryotic ribosome profiling data exhibits a read length of 28 to 32 nt; in our results we also found a peak of ribosome-protected fragments at 31 nt across all tested samples (Fig. 3A). By identifying the ribosomal P-site in each ribosome-protected fragment (RPF), riboWaltz allows for visual inspection of triplet periodicity, a distinct feature of ribosome profiling data due to the ribosome's movement along the mRNA 3-nt at a time. As expected for high-quality data, we observed a strong enrichment of sequencing reads in canonical coding sequences (CDSs). Our footprint fragments of 30 to 32 nt in length have a strong frame preference for one of the three subcodon positions on the CDS, but not on the 5' and 3' UTRs (Fig. 3B).

Time Considerations

The protocol (from harvest to library preparation) is expected to take ∼5 days. This could vary depending on the number of samples analyzed and the individuals experience with the techniques involved.

Acknowledgments

This work was supported by the UKRI-BBSRC Norwich Research Park Biosciences Doctoral Training Partnership (grant BB/M011216/1 to MP) and the German Research Foundation (grant SFB960 TP B11 to JM). We thank EMBL GeneCore for excellent support with sequencing.

Author Contributions

Monica Pichler: Data curation, methodology, resources, validation, writing—original draft; **Andreas Meindl:** Data curation, investigation, methodology; **Markus Romberger:** Data curation, investigation, methodology; **Annemarie Eckes-Shephard:** Data curation, investigation, methodology; **Carl-Fredrik Nyberg-Brodda:** Data curation, investigation, methodology; **Claudia Buhigas:** Data curation, investigation, methodology; **Sergio Llaneza-Lago:** Data curation, investigation, methodology; **Gerhard Lehmann:** Data curation, investigation, methodology; **Amanda Hopes:** Data curation, investigation, methodology, supervision; **Gunter Meister:** Funding acquisition; **Jan Medenbach:** Conceptualization, investigation, methodology, supervision, writing—review and editing; **Thomas Mock:** Conceptualization, funding acquisition, project administration, supervision, writing - review and editing.

Conflict of Interest

G.M. is a founder of, and J.M. a consultant to siTOOLs Biotech GmbH, Martinsried.

Data Availability Statement

The data that support the protocol are openly available on figshare at *https://doi.org/10.6084/m9.figshare.22717591* and *https://doi.org/10.6084/m9.figshare.23635512*.

Literature Cited

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S. G., Allen, A. E., Apt, K. E., Bechner, M., Brzezinski, M. A., Chaal, B. K., Chiovitti, A., Davis, A. K., Demarest, M. S., Detter, J. C., Glavina, T., Goodstein, D., Hadi, M. Z., … Rokhsar, D. S. (2004). The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science*, *306*, 79–86. https://doi.org/10.1126/science.1101156

Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., Amin, U., Mumtaz, M. A., Brocard, M., & Couso, J. P. (2014). Extensive translation of small open reading frames revealed by poly-ribo-seq. *Elife*, *3*, e03528. https://doi.org/10.7554/eLife.03528

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., Vejnar, C. E., Lee, M. T., Rajewsky, N., Walther, T. C., & Giraldez, A. J. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO Journal*, *33*(9), 981–993. https://doi.org/10.1002/embj.201488411

Bazzini, A. A., Lee, M. T., & Giraldez, A. J. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science*, *336*(6078), 233–237. https://doi.org/10.1126/science.1215704

Belshaw, N., Grouneva, I., Aram, L., Gal, A., Hopes, A., & Mock, T. (2023). Efficient gene replacement by CRISPR/Cas-mediated homologous recombination in the model diatom *Thalassiosira pseudonana*. *New Phytologist*, *238*, 438–452. https://doi.org/10.1111/nph.18587

Brar, G. A., & Weissman, J. S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature Reviews Molecular Cell Biology*, *16*(11), 651–664. https://doi.org/10.1038/nrm4069

Buchbender, A., Mutter, H., Sutandy, F. X. R., Kortel, N., Hanel, H., Busch, A., Ebersberger, S., & Konig, J. (2020). Improved library preparation with the new iCLIP2 protocol. *Methods (San Diego, Calif.)*, *178*, 33–48. https://doi.org/10.1016/j.ymeth.2019.10.003

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., & Ohler, U. (2016). Detecting actively translated open reading frames in

ribosome profiling data. *Nature Methods*, *13*(2), 165–170. https://doi.org/10.1038/nmeth.3688

Chung, B. Y., Hardcastle, T. J., Jones, J. D., Irigoyen, N., Firth, A. E., Baulcombe, D. C., & Brierley, I. (2015). The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. *RNA*, *21*(10), 1731–1745. https://doi.org/10.1261/rna.052548.115

Dong, H. P., Dong, Y. L., Cui, L., Balamurugan, S., Gao, J., Lu, S. H., & Jiang, T. (2016). High light stress triggers distinct proteomic responses in the marine diatom Thalassiosira pseudonana. *BMC Genomics*, *17*(1), 994. https://doi.org/10.1186/s12864-016-3335-5

Ferguson, L., Upton, H. E., Pimentel, S. C., Mok, A., Lareau, L. F., Collins, K., & Ingolia, N. T. (2023). Streamlined and sensitive mono- and diribosome profiling in yeast and human cells. *BioRxiv*, https://doi.org/10.1101/2023.02.01.526718

Gerashchenko, M. V., & Gladyshev, V. N. (2014). Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Research*, *42*(17), e134. https://doi.org/10.1093/nar/gku671

Gerashchenko, M. V., & Gladyshev, V. N. (2017). Ribonuclease selection for ribosome profiling. *Nucleic Acids Research*, *45*(2), e6. https://doi.org/10.1093/nar/gkw822

Guo, H., Ingolia, N. T., Weissman, J. S., & Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, *466*(7308), 835–840. https://doi.org/10.1038/nature09267

Hopes, A., Nekrasov, V., Kamoun, S., & Mock, T. (2016). Editing of the urease gene by CRISPR-Cas in the diatom *Thalassiosira pseudonana*. *Plant Methods*, *12*, 49. https://doi.org/10.1186/s13007-016-0148-0

Hsu, P. Y., Calviello, L., Wu, H. L., Li, F. W., Rothfels, C. J., Ohler, U., & Benfey, P. N. (2016). Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(45), E7126–E7135. https://doi.org/10.1073/pnas.1614788113

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L. E., Sibley, C. R., Sugimoto, Y., Tajnik, M., Konig, J., & Ule, J. (2014). iCLIP: Protein-RNA interactions at nucleotide resolution. *Methods*, *65*(3), 274–287. https://doi.org/10.1016/j.ymeth.2013.10.011

Ingolia, N. T. (2016). Ribosome footprint profiling of translation throughout the genome. *Cell*, *165*(1), 22–33. https://doi.org/10.1016/j.cell.2016.02.066

Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M., & Weissman, J. S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature Protocols*, *7*(8), 1534–1550. https://doi.org/10.1038/nprot.2012.086

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, *324*(5924), 218–223. https://doi.org/10.1126/science.1168978

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, *147*(4), 789–802. https://doi.org/10.1016/j.cell.2011.10.002

Ji, Z., Song, R., Regev, A., & Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife*, *4*, e08890. https://doi.org/10.7554/eLife.08890

Juntawong, P., Girke, T., Bazin, J., & Bailey-Serres, J. (2014). Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(1), E203–212. https://doi.org/10.1073/pnas.1317811111

Karlsen, J., Asplund-Samuelsson, J., Thomas, Q., Jahn, M., & Hudson, E. P. (2018). Ribosome profiling of synechocystis reveals altered ribosome allocation at carbon starvation. *mSystems*, *3*(5), https://doi.org/10.1128/mSystems.00126-18

Li, G. W., Oh, E., & Weissman, J. S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, *484*(7395), 538–541. https://doi.org/10.1038/nature10965

McGlincy, N. J., & Ingolia, N. T. (2017). Transcriptome-wide measurement of translation by ribosome profiling. *Methods*, *126*, 112–129. https://doi.org/10.1016/j.ymeth.2017.05.028

Meindl, A., Romberger, M., Lehmann, G., Eichner, N., Kleemann, L., Wu, J., Danner, J., Boesl, M., Mesitov, M., Meister, G., Koenig, J., Leidel, S. A., & Medenbach, J. (2023). A rapid protocol for ribosome profiling of low input samples. *Nucleic Acids Research*, gkad459. Advance online publication. https://doi.org/10.1093/nar/gkad459

Mock, T., Hodgkinson, K., Wu, T., Moulton, V., Duncan, A., van Oosterhout, C., & Pichler, M. (2022). Structure and evolution of diatom nuclear genes and genomes. In A. Falciatore & T. Mock (Eds.), *The molecular life of diatoms* (pp. 111–145). Springer International Publishing.

Mock, T., Samanta, M. P., Iverson, V., Berthiaume, C., Robison, M., Holtermann, K., Durkin, C., Bondurant, S. S., Richmond, K., Rodesch, M., Kallas, T., Huttlin, E. L., Cerrina, F., Sussman, M. R., & Armbrust, E. V. (2008). Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(5), 1579–1584. https://doi.org/10.1073/pnas.0707946105

Mohammad, F., & Buskirk, A. R. (2019). Protocol for ribosome profiling in bacteria.

Bio-Protocol, *9*(24), e3468. https://doi.org/10.21769/BioProtoc.3468

Mohammad, F., Green, R., & Buskirk, A. R. (2019). A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *Elife*, *8*, e42591. https://doi.org/10.7554/eLife.42591

Ozadam, H., Tonn, T., Han, C. M., Segura, A., Hoskins, I., Rao, S., Ghatpande, V., Tran, D., Catoe, D., Salit, M., & Cenik, C. (2023). Single-cell quantification of ribosome occupancy in early mouse development. *Nature*, *618*(7967), 1057–1064. https://doi.org/10.1038/s41586-023-06228-9

Poulsen, N., Chesley, P. M., & Kröger, N. (2006). Molecular genetic manipulation of the diatom *Thalassiosira pseudonana* (Bacillariophyceae). *Journal of Phycology*, *42*(5), 1059–1065. https://doi.org/10.1111/j.1529-8817.2006.00269.x

Price, N. M., Harrison, G. I., Hering, J. G., Hudson, R. J., Nirel, P. M., Palenik, B., & Morel, F. M. (1989). Preparation and chemistry of the artificial algal culture medium aquil. *Biological Oceanography*, *6*, 443–461. https://doi.org/10.1080/01965581.1988.10749544

Shalgi, R., Hurt, J. A., Krykbaeva, I., Taipale, M., Lindquist, S., & Burge, C. B. (2013). Widespread regulation of translation by elongation pausing in heat shock. *Molecular Cell*, *49*(3), 439–452. https://doi.org/10.1016/j.molcel.2012.11.028

Sharma, P., Wu, J., Nilges, B. S., & Leidel, S. A. (2021). Humans and other commonly used model organisms are resistant to cycloheximide-mediated biases in ribosome profiling experiments. *Nature Communications*, *12*(1), 5094. https://doi.org/10.1038/s41467-021-25411-y

Steitz, J. A. (1969). Polypeptide chain initiation: Nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. *Nature*, *224*(5223), 957–964. https://doi.org/10.1038/224957a0

Trosch, R., Barahimipour, R., Gao, Y., Badillo-Corona, J. A., Gotsmann, V. L., Zimmer, D., Muhlhaus, T., Zoschke, R., & Willmund, F. (2018). Commonalities and differences of chloroplast translation in a green alga and land plants. *Nature Plants*, *4*(8), 564–575. https://doi.org/10.1038/s41477-018-0211-0

Wu, H. L., Song, G., Walley, J. W., & Hsu, P. Y. (2019). The Tomato Translational Landscape Revealed by Transcriptome Assembly and Ribosome Profiling. *Plant Physiology*, *181*(1), 367–380. https://doi.org/10.1104/pp.19.00541

Zinshteyn, B., Wangen, J. R., Hua, B. Y., & Green, R. (2020). Nuclease-mediated depletion biases in ribosome footprint profiling libraries. *RNA*, *26*(10), 1481–1488. https://doi.org/10.1261/rna.075523.120

Internet Resources

https://ncma.bigelow.org/ccmp1335

A source of the Thalassiosira pseudonana clone CCMP1335.

https://ncma.bigelow.org/PDF%20Files/NCMA%20algal%20medium%20Aquil.pdf

A source for the full recipe of Aquil medium.

https://cutadapt.readthedocs.io/en/stable/guide.html

Manual for Cutadapt.

https://umi-tools.readthedocs.io/en/latest/QUICK_START.html

Manual for UMI-tools.

https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

Manual for bowtie2.

https://mycocosm.jgi.doe.gov/Thaps3/Thaps3.home.html

Data source of the T. pseudonana reference genome and genome annotations.

https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

Manual for STAR aligner.

https://cran.r-project.org/bin/windows/base/

R software download source.

https://www.rstudio.com/products/rstudio/download/

Rstudio software download source.

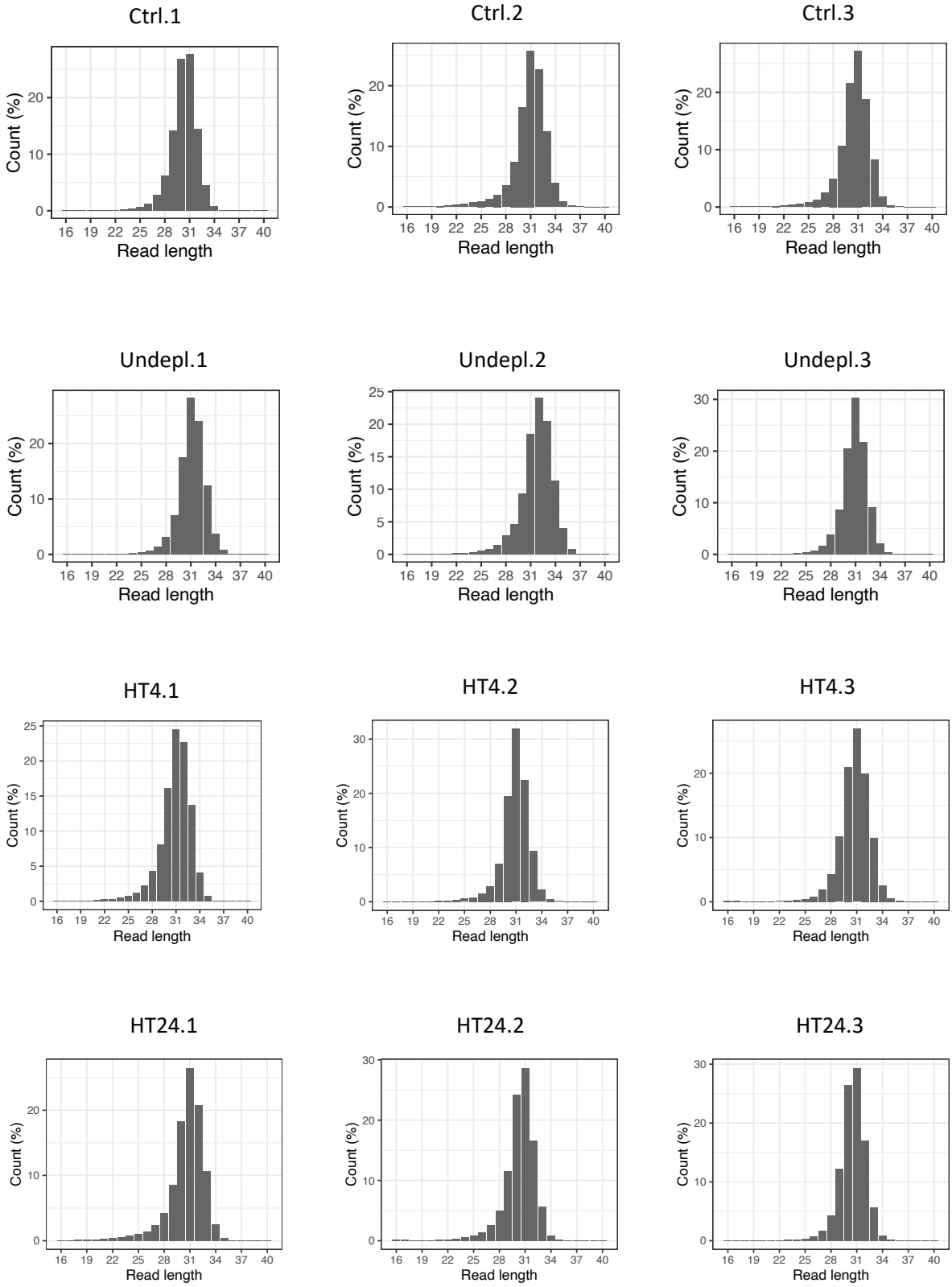
https://github.com/LabTranslationalArchitectomics/riboWaltz

Manual for riboWaltz.

www.illumina.com

Illumina TrueSeq single indexes online resource.

7.3 Appendix C: Ribosome profiling and RNA-seq quality control



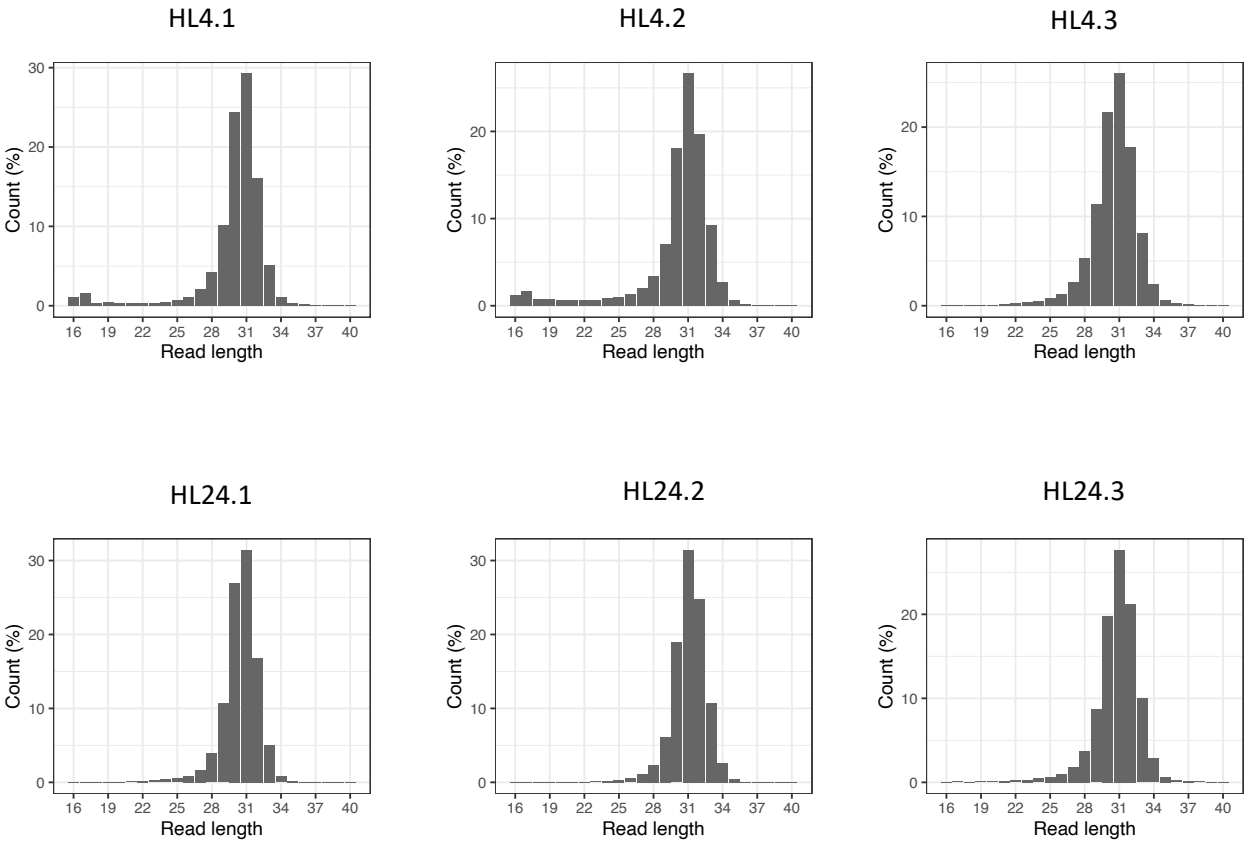
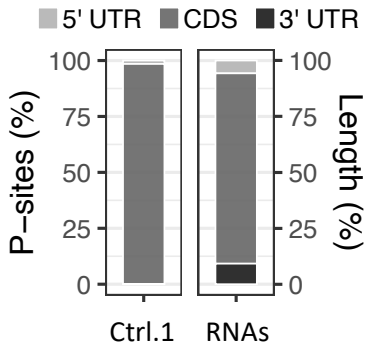
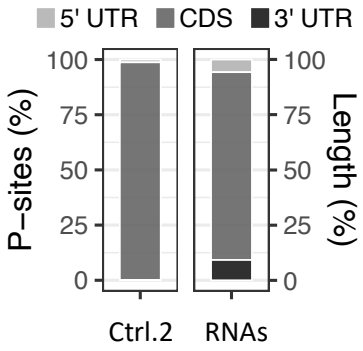
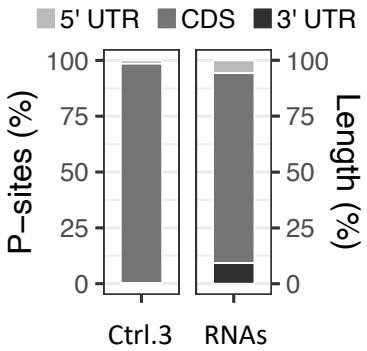
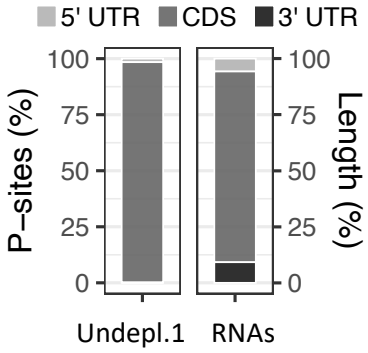


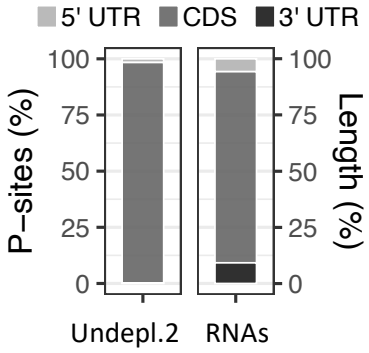
Figure 7.1: ***Read length distribution****. The number of reads according to length (between 25 and 35 nucleotides) for all 18 ribosome profiling samples.*

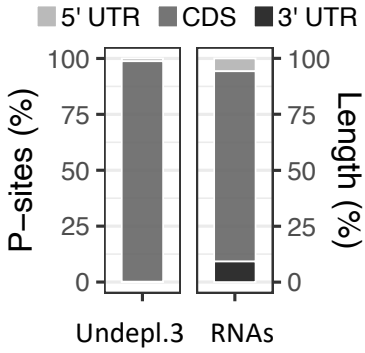

5' UTR
CDS
3' UTR
P−sites (%)
Length (%)
100
75
50
25
0
Ctrl.1
RNAs

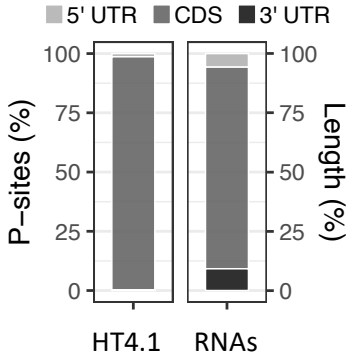

5' UTR
CDS
3' UTR
P−sites (%)
Length (%)
100
75
50
25
0
Ctrl.2
RNAs

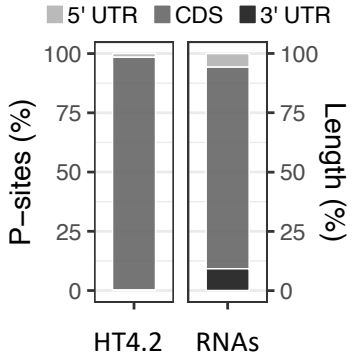

5' UTR
CDS
3' UTR
P−sites (%)
Length (%)
100
75
50
25
0
Ctrl.3
RNAs

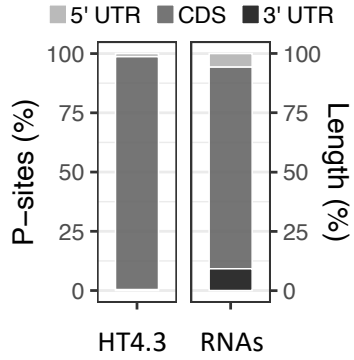

5' UTR
CDS
3' UTR
P−sites (%)
Length (%)
100
75
50
25
0
Undepl.1
RNAs

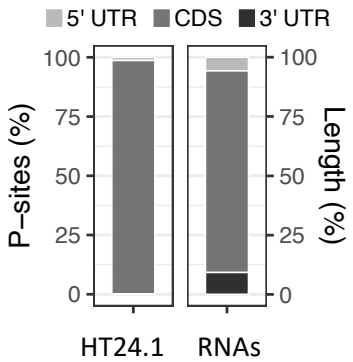

5' UTR
CDS
3' UTR
P−sites (%)
Length (%)
100
75
50
25
0
Undepl.2
RNAs

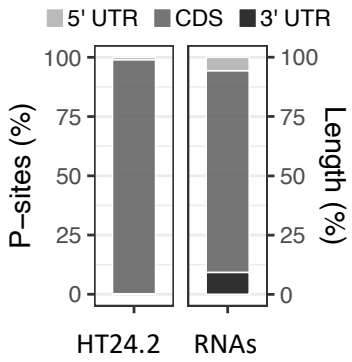

5' UTR
CDS
3' UTR
P−sites (%)
Length (%)
100
75
50
25
0
Undepl.3
RNAs

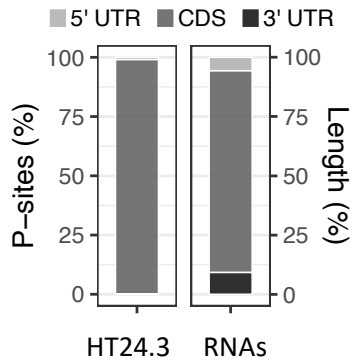

5' UTR
CDS
3' UTR
P−sites (%)
Length (%)
100
75
50
25
0
HT4.1
RNAs


5' UTR
CDS
3' UTR
P−sites (%)
Length (%)
100
75
50
25
0
HT4.2
RNAs


5' UTR
CDS
3' UTR
P−sites (%)
Length (%)
100
75
50
25
0
HT4.3
RNAs


5' UTR
CDS
3' UTR
P−sites (%)
Length (%)
100
75
50
25
0
HT24.1
RNAs


5' UTR
CDS
3' UTR
P−sites (%)
Length (%)
100
75
50
25
0
HT24.2
RNAs


5' UTR
CDS
3' UTR
P−sites (%)
Length (%)
100
75
50
25
0
HT24.3
RNAs

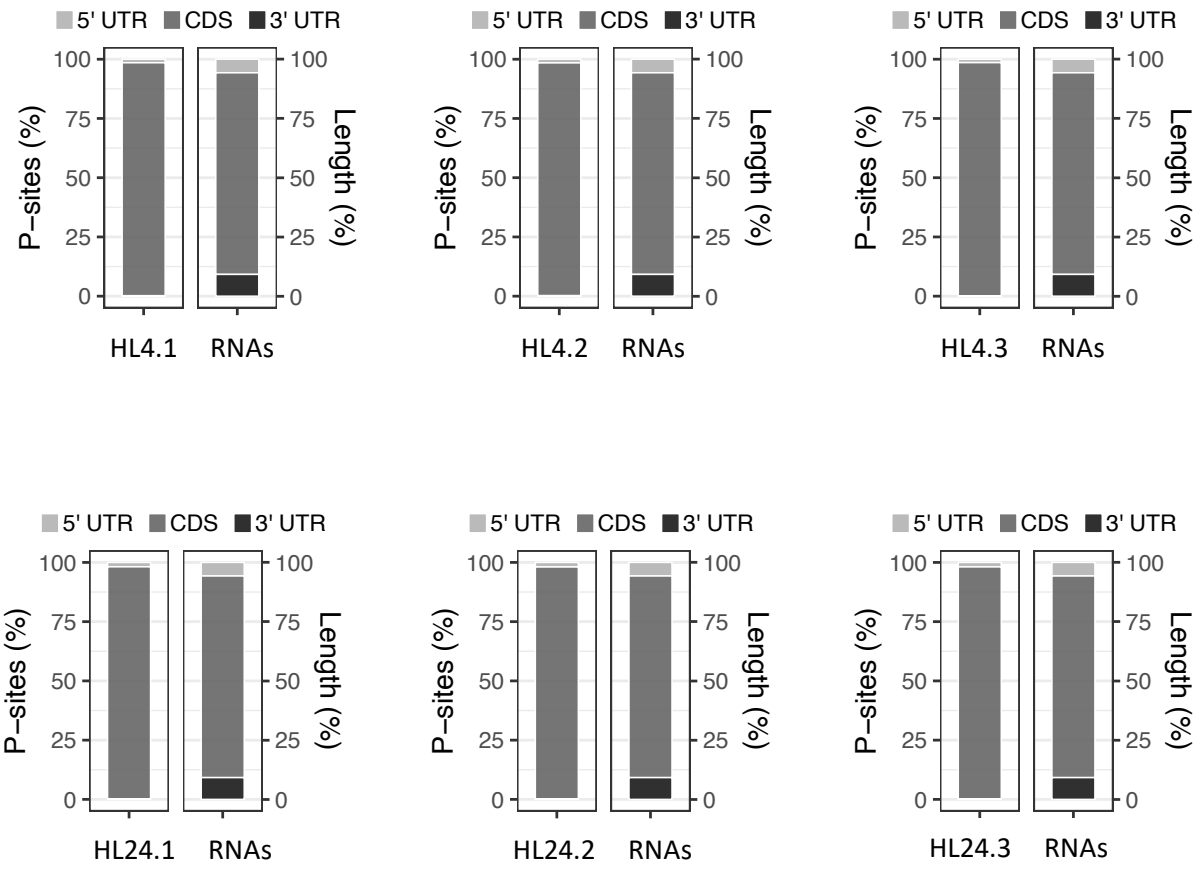
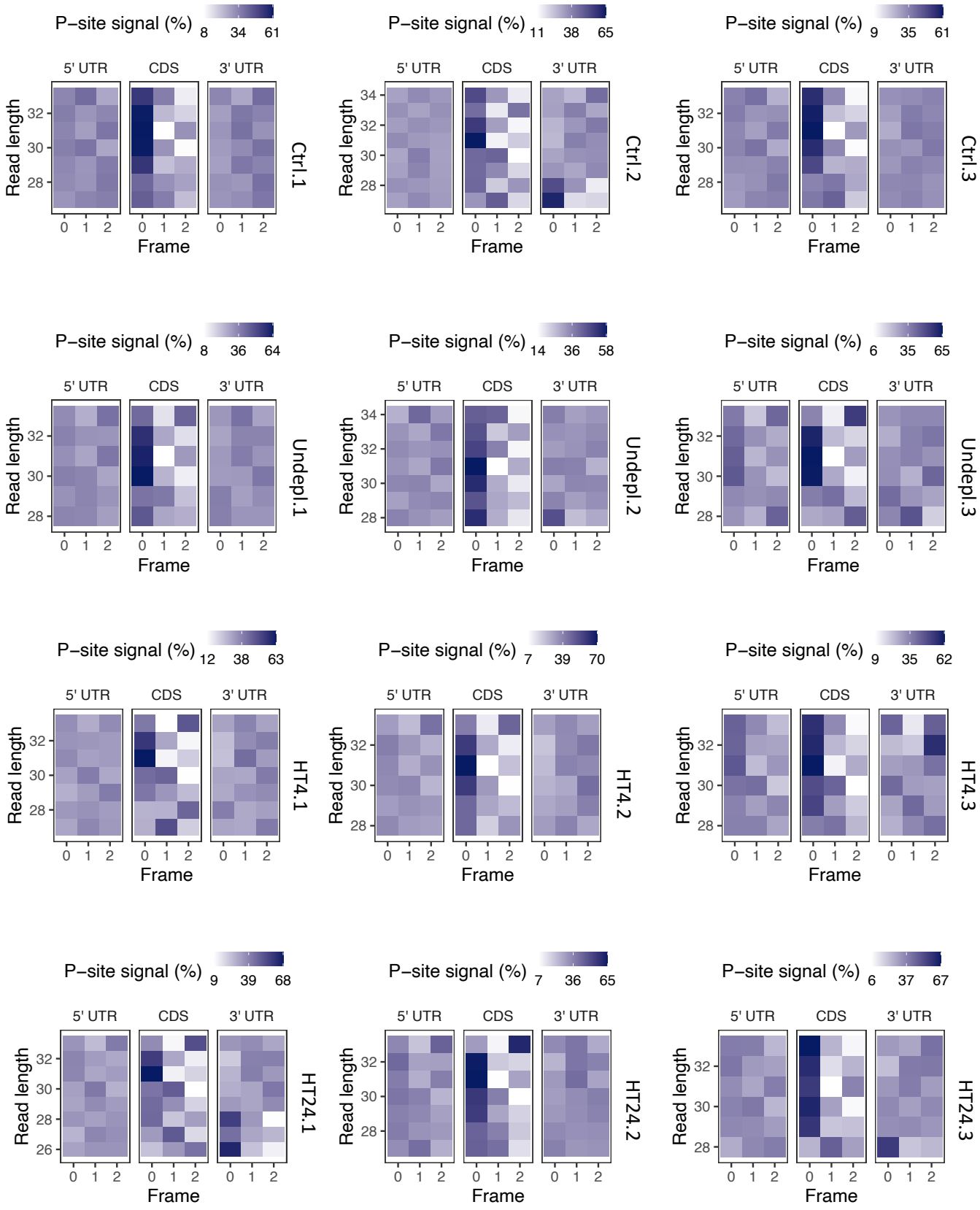


Figure 7.2: ***P-site analysis of all 18 ribosome profiling samples.*** *Left, percentage of P-sites in the 5'UTR, CDS and 3'UTR. Right, expected read distribution from random fragmentation or RNA. A clear enrichment of ribosome profiling data in CDS is shown.*


P−site signal (%) 8 34 61
5' UTR
CDS
3' UTR
Read length
32
30
28
0 1 2
Frame
Ctrl.1
P−site signal (%) 11 38 65
34
Ctrl.2
P−site signal (%) 9 35 61
Ctrl.3
P−site signal (%) 8 36 64
Undepl.1
P−site signal (%) 14 36 58
Undepl.2
P−site signal (%) 6 35 65
Undepl.3
P−site signal (%) 12 38 63
HT4.1
P−site signal (%) 7 39 70
HT4.2
P−site signal (%) 9 35 62
HT4.3
P−site signal (%) 9 39 68
26
HT24.1
P−site signal (%) 7 36 65
HT24.2
P−site signal (%) 6 37 67
HT24.3

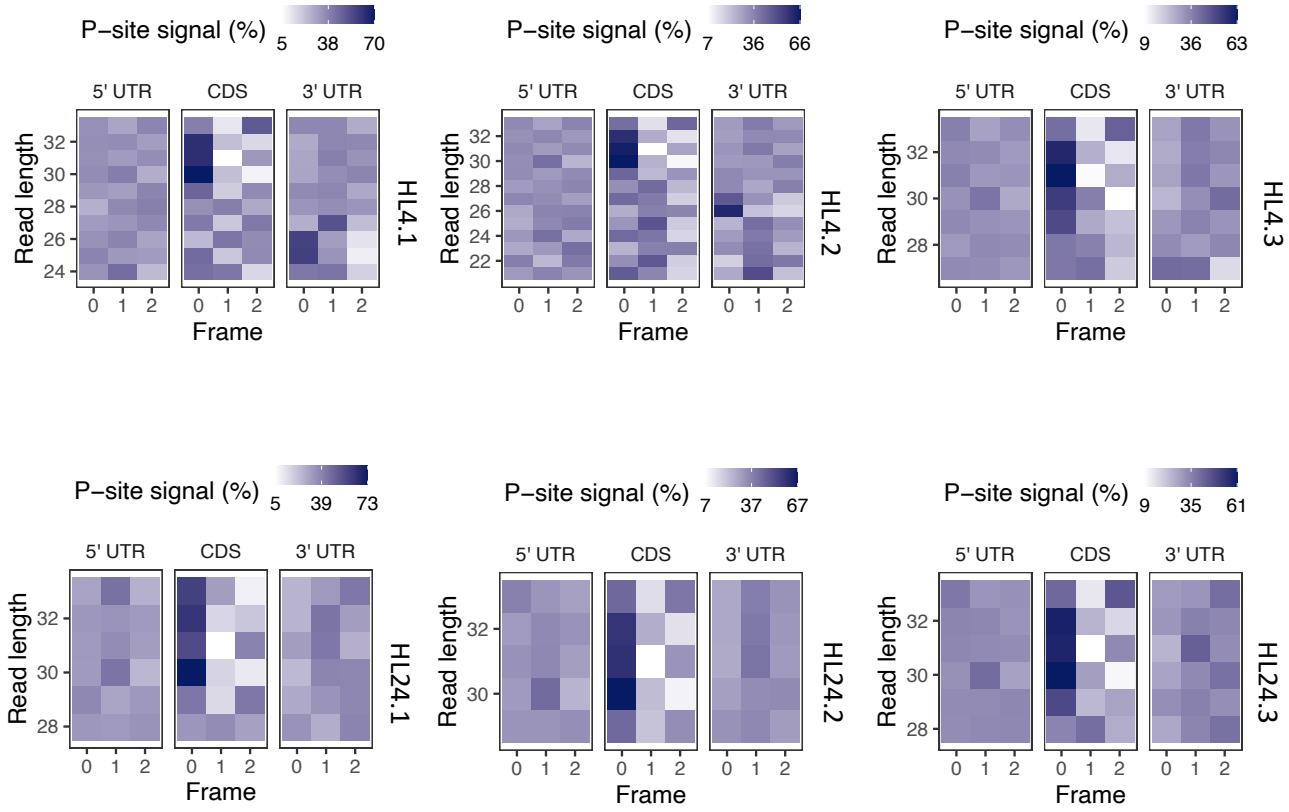
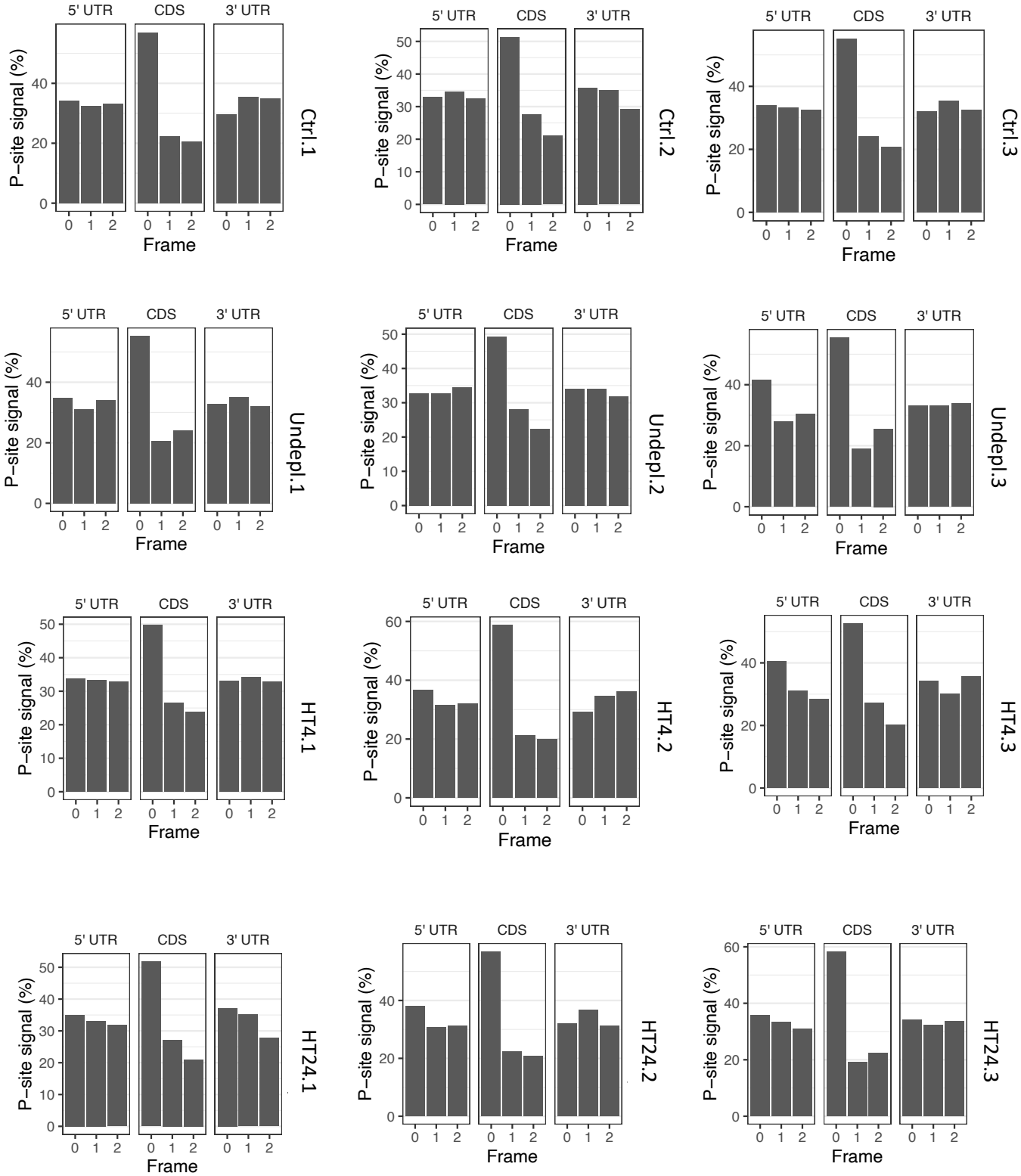


Figure 7.3: ***Reading frame preference of all 18 ribosome profiling samples****. Percentage of P-sites in the three frames along the 5' UTR, CDS and 3' UTR, stratified for read length. Reads between 30 and 32 nt in length show a clear frame preference in the CDS.*


5' UTR
CDS
3' UTR
P−site signal (%)
Frame
Ctrl.1
Ctrl.2
Ctrl.3
Undepl.1
Undepl.2
Undepl.3
HT4.1
HT4.2
HT4.3
HT24.1
HT24.2
HT24.3

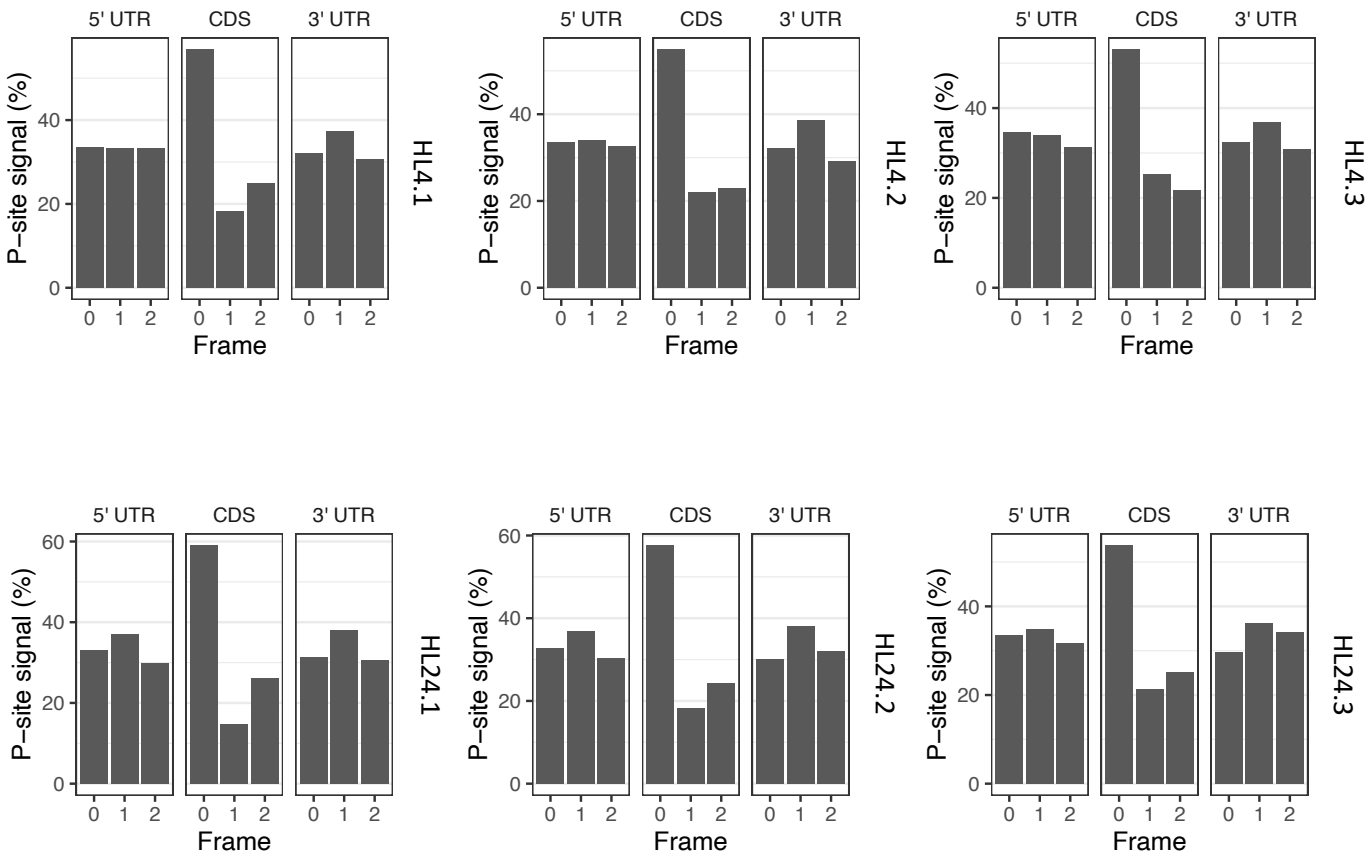


Figure 7.4: ***Triplet periodicity analysis of the 18 ribosome profiling samples.*** *Percentage of P-sites falling into one of the three reading frames for 5'UTR, CDS and 3'UTR. The majority of reads map to CDS and show a clear frame preference.*

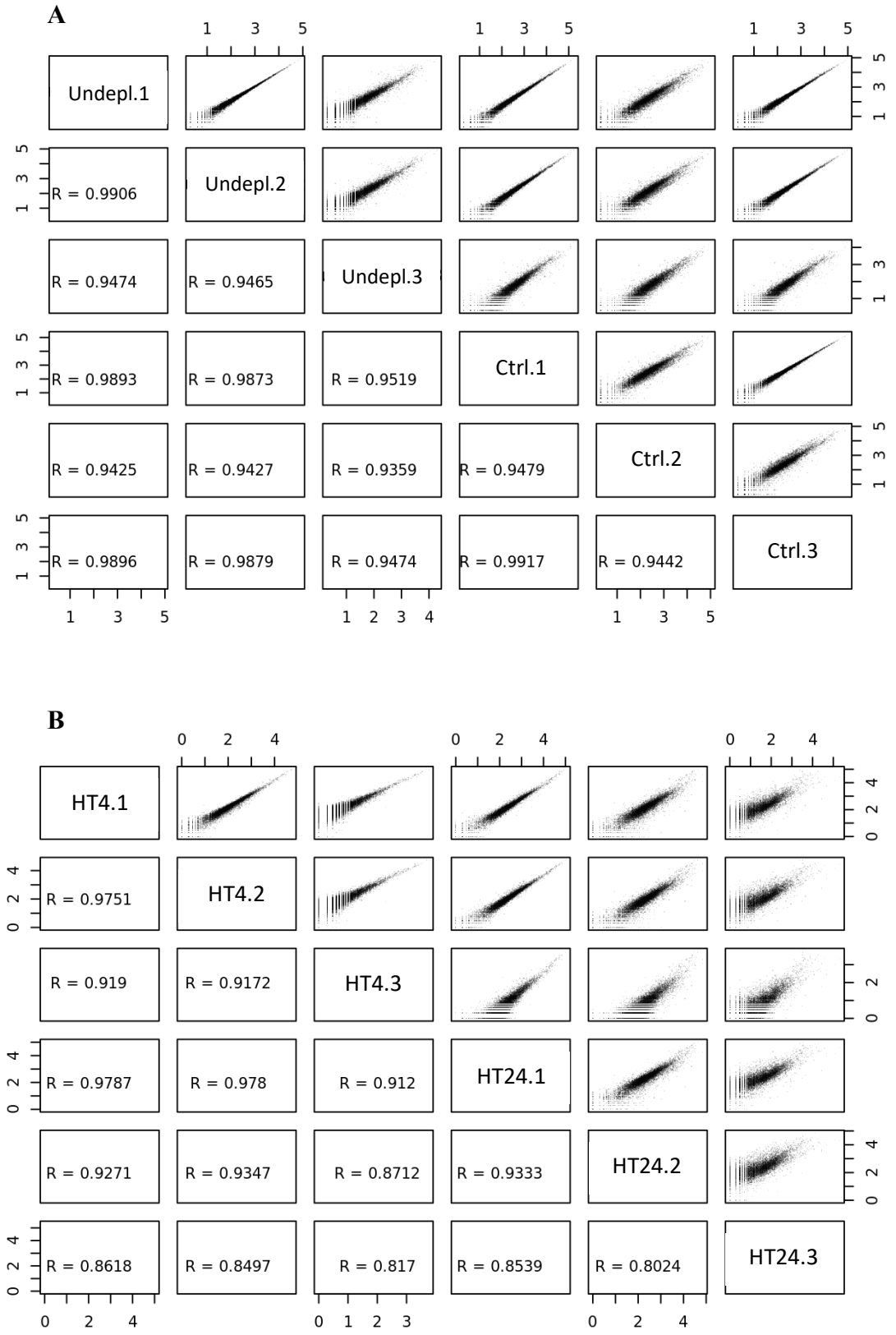


Figure 7.5: ***Statistical validation****. **A** Scatter plots of undepleted and rRNA depleted control samples. **B** Cells treated with high temperature (32°C) for 4 and 24 hours (HL= High light treatment for 4 h and 24 h) show high levels of reproducibility between biological replicates indicated by a Spearman's correlation coefficient >0.8. scatter plots show the log_{10} read counts per gene in a sample relative to another sample.*

Table 7.1: Results of Spearman's rank correlation analysis of RNA-Seq data. Biological replicates showed a high correlation coefficient, representing high correlation between the samples. HL= High light treatment for 4 h and 24 h, HT= High temperature treatment for 4 h and 24 h. 3 biological replicates per condition.

Sample	Ctrl.1	Ctrl.2	Ctrl.3	HL4.1	HL4.2	HL4.3	HL24.1	HL24.2	HL24.3	HT4.1	HT4.2	HT4.3	HT24.1	HT24.2	HT24.3
Ctrl.1	1	0.9692	0.9586	0.9033	0.9147	0.9152	0.8564	0.8659	0.8646	0.8726	0.8942	0.8913	0.8926	0.9032	0.8945
Ctrl.2	0.9692	1	0.9773	0.9342	0.9406	0.9280	0.8411	0.8473	0.8395	0.8477	0.8701	0.8759	0.8935	0.9038	0.9007
Ctrl.3	0.9586	0.9773	1	0.9413	0.9441	0.9238	0.8242	0.8286	0.8284	0.8586	0.8849	0.8894	0.8859	0.8972	0.8846
HL4.1	0.9033	0.9342	0.9413	1	0.9955	0.9669	0.8205	0.8355	0.8303	0.8182	0.8561	0.8398	0.8643	0.8765	0.8455
HL4.2	0.9147	0.9406	0.9441	0.9955	1	0.9716	0.8323	0.8476	0.8410	0.8240	0.8613	0.8447	0.8725	0.8851	0.8555
HL4.3	0.9152	0.9280	0.9238	0.9669	0.9716	1	0.9034	0.9203	0.9152	0.8548	0.8783	0.8712	0.9028	0.9093	0.8871
HL24.1	0.8564	0.8411	0.8242	0.8205	0.8323	0.9034	1	0.9862	0.9832	0.8338	0.8322	0.8428	0.8824	0.8847	0.8716
HL24.2	0.8659	0.8473	0.8286	0.8355	0.8476	0.9203	0.9862	1	0.9932	0.8584	0.8589	0.8665	0.8897	0.8903	0.8739
HL24.3	0.8646	0.8395	0.8284	0.8303	0.8410	0.9152	0.9832	0.9932	1	0.8706	0.8685	0.8742	0.8941	0.8933	0.8733
HT4.1	0.8726	0.8477	0.8586	0.8182	0.8240	0.8548	0.8338	0.8584	0.8706	1	0.9866	0.9780	0.9150	0.9068	0.8779
HT4.2	0.8942	0.8701	0.8849	0.8561	0.8613	0.8783	0.8322	0.8589	0.8685	0.9866	1	0.9806	0.9123	0.9133	0.8726
HT4.3	0.8913	0.8759	0.8894	0.8398	0.8447	0.8712	0.8428	0.8665	0.8742	0.9780	0.9806	1	0.9149	0.9164	0.8941
HT24.1	0.8926	0.8935	0.8859	0.8643	0.8725	0.9028	0.8824	0.8897	0.8941	0.9150	0.9123	0.9149	1	0.9925	0.9778
HT24.2	0.9032	0.9038	0.8972	0.8765	0.8851	0.9093	0.8847	0.8903	0.8933	0.9068	0.9133	0.9164	0.9925	1	0.9788
HT24.3	0.8945	0.9007	0.8846	0.8455	0.8555	0.8871	0.8716	0.8739	0.8733	0.8779	0.8726	0.8941	0.9778	0.9788	1

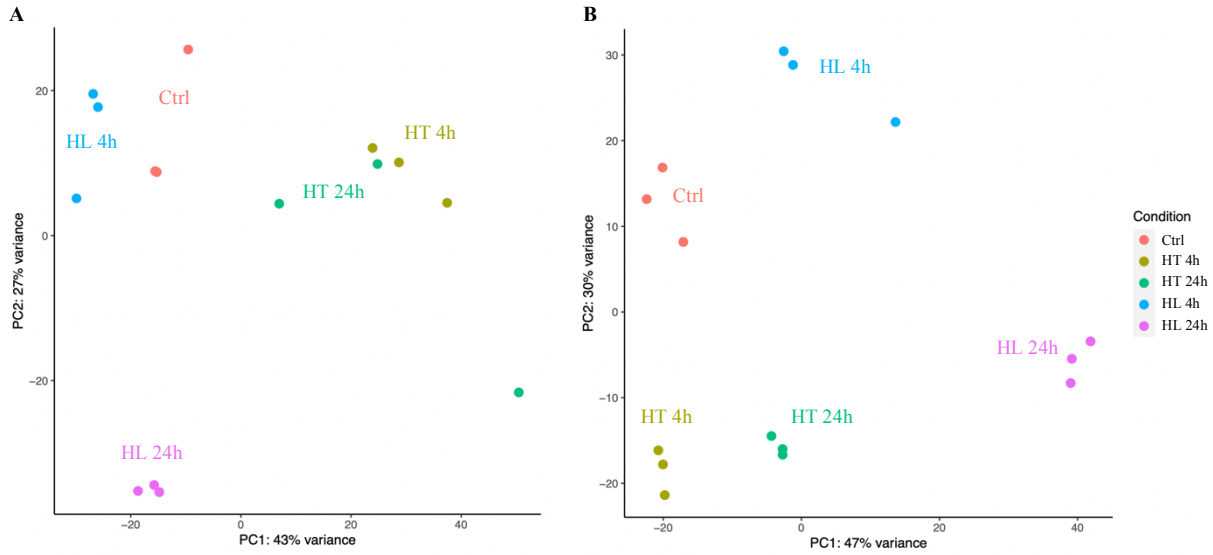


*Figure 7.6: Principal component analysis (PCA) of (**A**) Ribo-Seq and (**B**) RNA-Seq data to assess inter- and intragroup variability. Samples are displayed along the first component (PC1) and the second component (PC2), with PCA1 accounting for 43% of the variance in the Ribo-Seq data and for 47% of the variance in the RNA-Seq data. Biological replicates (indicated by data points in the same colour) group well together. Ctrl=Control, HL= High light stress for 4 h and 24 h, HT= High temperature stress for 4 h and 24 h.*

7.4 Appendix D: Functional analysis

Table 7.2: ***GO term enrichment analysis.*** *Biological processes of DTEGs under 4 h of high light stress.*

GO_term	n	GO_term	n
metabolism	27	ubiquitin cycle	3
regulation of transcription, DNA-dependent	17	aromatic compound metabolism	2
proteolysis and peptidolysis	16	ATP synthesis coupled proton transport	2
protein amino acid phosphorylation	15	biosynthesis	2
electron transport	11	biotin biosynthesis	2
transport	8	ciliary or flagellar motility	2
ion transport	5	folic acid and derivative biosynthesis	2
DNA repair	4	glucose metabolism	2
nucleosome assembly	4	intracellular protein transport	2
protein modification	4	intracellular signaling cascade	2
regulation of cell cycle	4	microtubule-based movement	2
cation transport	3	mRNA metabolism	2
cell adhesion	3	oxygen and reactive oxygen species metabolism	2
chromosome organization and biogenesis (sensu Eukaryota)	3	phosphoenolpyruvate-dependent sugar phosphotransferase system	2
gluconeogenesis	3	photosynthesis light harvesting	2
one-carbon compound metabolism	3	protein biosynthesis	2
phosphate transport	3	protein metabolism	2
potassium ion transport	3	protein transport	2
protein folding	3	RNA processing	2
protein ubiquitination	3	small GTPase mediated signal transduction	2
tRNA aminoacylation for protein translation	3	tricarboxylic acid cycle	2

Table 7.3: ***GO term enrichment analysis.*** *Molecular functions of DTEGs under 4 h of high light stress.*

GO_term	n	GO_term	n
ATP binding	39	MAP kinase 1 activity	3
catalytic activity	22	MAP kinase 2 activity	3
DNA binding	21	MAP kinase activity	3
oxidoreductase activity	16	MAP kinase kinase activity	3

transcription factor activity	15	**MAP kinase kinase kinase activity**	3
protein kinase activity	14	**MAP kinase kinase kinase kinase activity**	3
calcium ion binding	12	**MAP/ERK kinase kinase activity**	3
protein serine/threonine kinase activity	8	**MP kinase activity**	3
zinc ion binding	8	**mu DNA polymerase activity**	3
protein-tyrosine kinase activity	7	**multifunctional calcium- and calmodulin-regulated protein kinase activity**	3
ubiquitin-protein ligase activity	7	**NF-kappaB-inducing kinase activity**	3
iron ion binding	6	**non-membrane spanning protein tyrosine kinase activity**	3
nucleic acid binding	6	**nu DNA polymerase activity**	3
GTP binding	5	**phenol kinase activity**	3
ion channel activity	5	**phosphatidylinositol phosphate kinase activity**	3
metalloendopeptidase activity	5	**phosphofructokinase activity**	3
metallopeptidase activity	5	**phosphoinositide 3-kinase activity**	3
nucleoside-triphosphatase activity	5	**phosphopantetheine binding**	3
nucleotide binding	5	**phosphorylase kinase activity**	3
protein binding	5	**phosphorylase kinase regulator activity**	3
RNA binding	5	**phosphotransferase activity, alcohol group as acceptor**	3
ATPase activity	4	**protein kinase C activity**	3
exo-alpha-sialidase activity	4	**protein threonine/tyrosine kinase activity**	3
glucan 1,4-alpha-glucosidase activity	4	**receptor signaling protein serine/threonine kinase activity**	3
ligase activity	4	**receptor signaling protein tyrosine kinase activity**	3
3-phosphoinositide-dependent protein kinase activity	3	**ribosomal protein S6 kinase activity**	3
alpha DNA polymerase activity	3	**SAP kinase activity**	3
AMP-activated protein kinase activity	3	**sigma DNA polymerase activity**	3
atypical protein kinase C activity	3	**subtilase activity**	3
beta DNA polymerase activity	3	**theta DNA polymerase activity**	3
binding	3	**transmembrane receptor protein kinase activity**	3
calcium-dependent protein kinase C activity	3	**transmembrane receptor protein serine/threonine kinase activity**	3
calmodulin regulated protein kinase activity	3	**transmembrane receptor protein tyrosine kinase activity**	3
cAMP-dependent protein kinase regulator activity	3	**transporter activity**	3
casein kinase activity	3	**tRNA ligase activity**	3
casein kinase I activity	3	**zeta DNA polymerase activity**	3

cGMP-dependent protein kinase activity	3	acyltransferase activity	2
cobinamide kinase activity	3	adenosylhomocysteinase activity	2
cyclic nucleotide-dependent protein kinase activity	3	ATP-dependent helicase activity	2
cyclin-dependent protein kinase activating kinase activity	3	cAMP-dependent protein kinase activity	2
cyclin-dependent protein kinase activating kinase regulator activity	3	cation channel activity	2
cyclin-dependent protein kinase activity	3	chitin binding	2
cyclin-dependent protein kinase regulator activity	3	chymotrypsin activity	2
delta DNA polymerase activity	3	cysteine-type endopeptidase activity	2
deoxyribodipyrimidine photo-lyase activity	3	formate C-acetyltransferase activity	2
diacylglycerol-activated phospholipid-dependent protein kinase C activity	3	GTPase activity	2
DNA photolyase activity	3	heat shock protein binding	2
DNA-dependent protein kinase activity	3	hydrogen-transporting ATP synthase activity, rotational mechanism	2
DNA-directed DNA polymerase activity	3	hydrogen-transporting ATPase activity, rotational mechanism	2
electron transporter activity	3	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	2
epsilon DNA polymerase activity	3	inositol 2-dehydrogenase activity	2
eta DNA polymerase activity	3	magnesium ion binding	2
eukaryotic elongation factor-2 kinase activator activity	3	methyltransferase activity	2
eukaryotic elongation factor-2 kinase activity	3	microtubule motor activity	2
eukaryotic elongation factor-2 kinase regulator activity	3	molecular function unknown	2
eukaryotic translation initiation factor 2alpha kinase activity	3	monooxygenase activity	2
G-protein coupled receptor kinase activity	3	myosin ATPase activity	2
galactosyltransferase-associated kinase activity	3	peptidase activity	2
gamma DNA-directed DNA polymerase activity	3	phosphoenolpyruvate carboxylase activity	2
glycogen synthase kinase 3 activity	3	potassium channel activity	2
hydrolase activity	3	protein kinase CK2 activity	2
IkappaB kinase activity	3	protein kinase CK2 regulator activity	2
iota DNA polymerase activity	3	pyrophosphatase activity	2
Janus kinase activity	3	pyruvate carboxylase activity	2
JUN kinase activity	3	sodium-dependent phosphate transporter activity	2
JUN kinase kinase activity	3	sugar porter activity	2

JUN kinase kinase kinase activity	3	**transferase activity, transferring hexosyl groups**	2
kappa DNA polymerase activity	3	**trypsin activity**	2
lambda DNA polymerase activity	3	**unfolded protein binding**	2
		voltage-gated potassium channel activity	2

Table 7.4: ***GO term enrichment analysis.*** *Cellular components of DTEGs under 4 h of high light stress.*

GO_term	n	GO_term	n
membrane	25	**intracellular**	3
nucleus	22	**ubiquitin ligase complex**	3
integral to membrane	9	**microtubule associated complex**	2
cytoplasm	6	**mitochondrial inner membrane**	2
extracellular region	4	**protein kinase CK2 complex**	2
nucleosome	4	**proton-transporting two-sector ATPase complex**	2
cAMP-dependent protein kinase complex	3		

Table 7.5: ***GO term enrichment analysis.*** *Biological processes of DTEGs under 24 h of high light stress.*

GO_term	n	GO_term	n
metabolism	75	**DNA methylation**	3
proteolysis and peptidolysis	46	**DNA topological change**	3
electron transport	36	**fatty acid desaturation**	3
regulation of transcription, DNA-dependent	36	**folic acid and derivative biosynthesis**	3
protein amino acid phosphorylation	34	**gluconeogenesis**	3
transport	23	**ion transport**	3
protein biosynthesis	20	**metal ion transport**	3
protein ubiquitination	13	**mRNA processing**	3
chromosome organization and biogenesis (sensu Eukaryota)	11	**one-carbon compound metabolism**	3
nucleosome assembly	11	**phosphoenolpyruvate-dependent sugar phosphotransferase system**	3
RNA processing	11	**response to oxidative stress**	3
tRNA aminoacylation for protein translation	11	**two-component signal transduction system (phosphorelay)**	3
cation transport	10	**ubiquitin cycle**	3

carbohydrate metabolism	9	**aromatic compound metabolism**	2
protein folding	9	**biotin biosynthesis**	2
protein metabolism	9	**carbohydrate transport**	2
regulation of cell cycle	9	**carotenoid biosynthesis**	2
amino acid transport	7	**cell adhesion**	2
chitin metabolism	7	**cell cycle**	2
intracellular signaling cascade	7	**cellular protein metabolism**	2
DNA repair	6	**chromosome organization and biogenesis**	2
nitrogen compound metabolism	6	**DNA metabolism**	2
photosynthesis light harvesting	6	**DNA replication**	2
protein modification	6	**G-protein coupled receptor protein signaling pathway**	2
protein transport	6	**glycine catabolism**	2
ATP synthesis coupled proton transport	5	**isoprenoid biosynthesis**	2
biosynthesis	5	**microtubule-based movement**	2
fatty acid biosynthesis	5	**nucleotide-sugar transport**	2
glycolysis	5	**phosphate transport**	2
intracellular protein transport	5	**potassium ion transport**	2
regulation of oxidoreductase activity	5	**proline biosynthesis**	2
arginine biosynthesis	4	**protein amino acid dephosphorylation**	2
cell wall catabolism	4	**protein amino acid glycosylation**	2
chitin catabolism	4	**protein targeting**	2
lipid metabolism	4	**proton transport**	2
oxygen and reactive oxygen species metabolism	4	**pyrimidine base biosynthesis**	2
response to pest, pathogen or parasite	4	**signal transduction**	2
small GTPase mediated signal transduction	4	**sodium ion transport**	2
translational elongation	4	**SRP-dependent cotranslational protein-membrane targeting**	2
tricarboxylic acid cycle	4	**tetracycline transport**	2
amino acid biosynthesis	3	**transcription initiation**	2
aromatic amino acid family biosynthesis	3	**translational initiation**	2
cyclic nucleotide biosynthesis	3	**tryptophanyl-tRNA aminoacylation**	2
		ubiquitin-dependent protein catabolism	2

Table 7.6: ***GO term enrichment analysis.*** *Molecular functions of DTEGs under 24 h of high light stress.*

GO_term	n	GO_term	n
ATP binding	101	DNA topoisomerase (ATP-hydrolyzing) activity	3
oxidoreductase activity	67	epsilon DNA polymerase activity	3
DNA binding	63	eta DNA polymerase activity	3
catalytic activity	56	FAD binding	3
nucleic acid binding	34	gamma DNA-directed DNA polymerase activity	3
zinc ion binding	34	inositol or phosphatidylinositol phosphatase activity	3
protein kinase activity	33	inositol-1(or 4)-monophosphatase activity	3
transcription factor activity	25	ion channel activity	3
protein serine/threonine kinase activity	21	iota DNA polymerase activity	3
protein-tyrosine kinase activity	19	kappa DNA polymerase activity	3
ATPase activity	18	lambda DNA polymerase activity	3
RNA binding	18	mu DNA polymerase activity	3
ubiquitin-protein ligase activity	18	nu DNA polymerase activity	3
glucan 1,4-alpha-glucosidase activity	17	oxidoreductase activity, acting on CH-OH group of donors	3
hydrolase activity	16	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	3
protein binding	16	oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water	3
exo-alpha-sialidase activity	15	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	3
GTP binding	15	pancreatic elastase activity	3
nucleoside-triphosphatase activity	15	peptidase activity	3
nucleotide binding	15	phosphorus-oxygen lyase activity	3
ATP-dependent helicase activity	14	plasminogen activator activity	3
calcium ion binding	14	proprotein convertase activity	3
metallopeptidase activity	13	proteasome endopeptidase activity	3
S-adenosylmethionine-dependent methyltransferase activity	12	pyruvate carboxylase activity	3
structural constituent of ribosome	12	RNA methyltransferase activity	3
transporter activity	11	serine-type endopeptidase activity	3
tRNA ligase activity	11	serine-type signal peptidase activity	3
chitinase activity	10	sigma DNA polymerase activity	3

cyclophilin	10	small monomeric GTPase activity	3
cyclophilin-type peptidyl-prolyl cis-trans isomerase activity	10	structural molecule activity	3
FK506-sensitive peptidyl-prolyl cis-trans isomerase	10	theta DNA polymerase activity	3
GTPase activity	10	threonine endopeptidase activity	3
peptidyl-prolyl cis-trans isomerase activity	10	transferase activity, transferring nitrogenous groups	3
ATP-dependent DNA helicase activity	9	translation initiation factor activity	3
ATPase activity, coupled to transmembrane movement of substances	9	triacylglycerol lipase activity	3
chitin binding	9	zeta DNA polymerase activity	3
iron ion binding	9	1-phenanthrol glycosyltransferase activity	2
ATP-dependent RNA helicase activity	8	1-phenanthrol methyltransferase activity	2
ATPase activity, coupled	8	1-phenylethanol dehydrogenase activity	2
ATPase activity, coupled to transmembrane movement of ions	8	1,2-dihydroxy-phenanthrene glycosyltransferase activity	2
ATPase activity, uncoupled	8	2-hydroxyisobutyrate 3-monooxygenase activity	2
binding	8	2-hydroxytetrahydrofuran dehydrogenase activity	2
cysteine-type endopeptidase activity	8	2-polyprenyl-6-methoxy-1,4-benzoquinone methyltransferase activity	2
DNA translocase activity	8	3-keto sterol reductase activity	2
DNA-dependent ATPase activity	8	3-ketoglucose-reductase activity	2
helicase activity	8	3-oxoacyl-[acyl-carrier protein] reductase activity	2
ligase activity	8	4-chlorophenoxyacetate monooxygenase activity	2
phosphopantetheine binding	8	4-nitrocatechol 4-monooxygenase activity	2
protein-transporting ATPase activity	8	4-nitrophenol 2-monooxygenase activity	2
RNA-dependent ATPase activity	8	5-exo-hydroxycamphor dehydrogenase activity	2
single-stranded DNA-dependent ATP-dependent DNA helicase activity	8	9-phenanthrol glycosyltransferase activity	2
transferase activity, transferring hexosyl groups	8	9-phenanthrol UDP-glucuronosyltransferase activity	2
trypsin activity	8	acetylgalactosaminyltransferase activity	2
acid-amino acid ligase activity	7	actin binding	2
amino acid-polyamine transporter activity	7	adenosylhomocysteinase activity	2
electron transporter activity	7	adenylate cyclase activity	2
endochitinase activity	7	alkanesulfonate monooxygenase activity	2
histone acetyltransferase activity	7	alpha-1,2-mannosyltransferase activity	2

hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	7	alpha-1,3-mannosyltransferase activity	2
metalloendopeptidase activity	7	alpha-1,6-mannosyltransferase activity	2
methyltransferase activity	7	alpha-pinene dehydrogenase activity	2
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors	7	alpha-pinene monooxygenase activity	2
phosphotransferase activity, alcohol group as acceptor	7	alpha(1,3)-fucosyltransferase activity	2
ribosomal S6-glutamic acid ligase activity	7	alpha(1,6)-fucosyltransferase activity	2
UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate-D-alanyl-D-alanine ligase activity	7	amino-acid N-acetyltransferase activity	2
3-phosphoinositide-dependent protein kinase activity	6	aminomethyltransferase activity	2
AMP-activated protein kinase activity	6	ammonia monooxygenase activity	2
atypical protein kinase C activity	6	arginine N-methyltransferase activity	2
calcium-dependent protein kinase C activity	6	beta-1,4-mannosyltransferase activity	2
calmodulin regulated protein kinase activity	6	bile acid:sodium symporter activity	2
casein kinase activity	6	C-3 sterol dehydrogenase (C-4 sterol decarboxylase) activity	2
casein kinase I activity	6	C-methyltransferase activity	2
cGMP-dependent protein kinase activity	6	C-terminal protein carboxyl methyltransferase activity	2
chymotrypsin activity	6	calcium- and calmodulin-responsive adenylate cyclase activity	2
cobinamide kinase activity	6	cAMP-dependent protein kinase regulator activity	2
cyclic nucleotide-dependent protein kinase activity	6	cathepsin F activity	2
cyclin-dependent protein kinase activating kinase activity	6	cation transporter activity	2
cyclin-dependent protein kinase activating kinase regulator activity	6	cellulose synthase activity	2
cyclin-dependent protein kinase activity	6	chloral hydrate dehydrogenase activity	2
cyclin-dependent protein kinase regulator activity	6	deoxyribodipyrimidine photo-lyase activity	2
diacylglycerol-activated phospholipid-dependent protein kinase C activity	6	di-n-butyltin dioxygenase activity	2
DNA-dependent protein kinase activity	6	dimethylsilanediol hydroxylase activity	2
eukaryotic elongation factor-2 kinase activator activity	6	DNA photolyase activity	2
eukaryotic elongation factor-2 kinase activity	6	DNA-(apurinic or apyrimidinic site) lyase activity	2
eukaryotic elongation factor-2 kinase regulator activity	6	DNA-methyltransferase activity	2
eukaryotic translation initiation factor 2alpha kinase activity	6	dolichyl pyrophosphate Glc1Man9GlcNAc2 alpha-1,3-glucosyltransferase activity	2

G-protein coupled receptor kinase activity	6	dolichyl pyrophosphate Man9GlcNAc2 alpha-1,3-glucosyltransferase activity	2
galactosyltransferase-associated kinase activity	6	dolichyl-diphosphooligosaccharide-protein glycotransferase activity	2
glycogen synthase kinase 3 activity	6	dolichyl-phosphate-glucose-glycolipid alpha-glucosyltransferase activity	2
IkappaB kinase activity	6	epoxide dehydrogenase activity	2
Janus kinase activity	6	fluorene oxygenase activity	2
JUN kinase activity	6	FMN binding	2
JUN kinase kinase activity	6	fucosyltransferase activity	2
JUN kinase kinase kinase activity	6	galactosyltransferase activity	2
MAP kinase 1 activity	6	gamma-glutamyltransferase activity	2
MAP kinase 2 activity	6	gluconate dehydrogenase activity	2
MAP kinase activity	6	glucose dehydrogenase activity	2
MAP kinase kinase activity	6	glucosyltransferase activity	2
MAP kinase kinase kinase activity	6	glycolipid mannosyltransferase activity	2
MAP kinase kinase kinase kinase activity	6	hydrogen-transporting two-sector ATPase activity	2
MAP/ERK kinase kinase activity	6	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	2
MP kinase activity	6	hydrolase activity, hydrolyzing O-glycosyl compounds	2
multifunctional calcium- and calmodulin-regulated protein kinase activity	6	hydroxymethylmethylsilanediol oxidase activity	2
myosin ATPase activity	6	hydroxymethylsilanetriol oxidase activity	2
N-acetyltransferase activity	6	inorganic diphosphatase activity	2
NF-kappaB-inducing kinase activity	6	isocitrate dehydrogenase activity	2
non-membrane spanning protein tyrosine kinase activity	6	ketoreductase activity	2
phenol kinase activity	6	kynurenine 3-monooxygenase activity	2
phosphatidylinositol phosphate kinase activity	6	limonene 8-monooxygenase activity	2
phosphofructokinase activity	6	linoleoyl-CoA desaturase activity	2
phosphoinositide 3-kinase activity	6	lipopolysaccharide-1,6-galactosyltransferase activity	2
phosphorylase kinase activity	6	lysine N-methyltransferase activity	2
phosphorylase kinase regulator activity	6	mannosyltransferase activity	2
protein kinase C activity	6	metal ion transporter activity	2
protein threonine/tyrosine kinase activity	6	methyl tertiary butyl ether 3-monooxygenase activity	2
protein-synthesizing GTPase activity	6	methylarsonite methyltransferase activity	2
pyrophosphatase activity	6	methylsilanetriol hydroxylase activity	2

receptor signaling protein serine/threonine kinase activity	6	mevaldate reductase activity	2
receptor signaling protein tyrosine kinase activity	6	mono-butyltin dioxygenase activity	2
ribosomal protein S6 kinase activity	6	mRNA methyltransferase activity	2
SAP kinase activity	6	myrtenol dehydrogenase activity	2
subtilase activity	6	nuclease activity	2
transmembrane receptor protein kinase activity	6	nucleotide-sugar transporter activity	2
transmembrane receptor protein serine/threonine kinase activity	6	nutrient reservoir activity	2
transmembrane receptor protein tyrosine kinase activity	6	O-methyltransferase activity	2
bis(5'-nucleosyl)-tetraphosphatase activity	5	peroxidase activity	2
calcium-dependent protein serine/threonine phosphatase activity	5	phenanthrol glycosyltransferase activity	2
calcium-dependent protein serine/threonine phosphatase regulator activity	5	phosphoenolpyruvate carboxylase activity	2
carboxypeptidase A activity	5	phosphoric ester hydrolase activity	2
CTD phosphatase activity	5	pinocarveol dehydrogenase activity	2
dATP pyrophosphohydrolase activity	5	procollagen-proline 4-dioxygenase activity	2
diacylglycerol pyrophosphate phosphatase activity	5	protein kinase CK2 regulator activity	2
dihydroneopterin monophosphate phosphatase activity	5	protein methyltransferase activity	2
dihydroneopterin triphosphate pyrophosphohydrolase activity	5	protein serine/threonine phosphatase activity	2
DNA helicase IV activity	5	protein-arginine N-methyltransferase activity	2
hydrogen-transporting ATP synthase activity, rotational mechanism	5	protein-arginine N5-methyltransferase activity	2
hydrogen-transporting ATPase activity, rotational mechanism	5	protein-leucine O-methyltransferase activity	2
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	5	protein-lysine N-methyltransferase activity	2
magnesium-dependent protein serine/threonine phosphatase activity	5	rhodopsin-like receptor activity	2
molecular function unknown	5	rRNA (adenine-N6,N6-)-dimethyltransferase activity	2
myosin phosphatase activity	5	rRNA (adenine) methyltransferase activity	2
myosin phosphatase regulator activity	5	rRNA (cytosine-C5-967)-methyltransferase activity	2
N-methyltransferase activity	5	rRNA (cytosine) methyltransferase activity	2
phosphoprotein phosphatase activity	5	rRNA (guanine) methyltransferase activity	2
protein phosphatase type 2A activity	5	rRNA (uridine-2'-O-)-methyltransferase activity	2
protein phosphatase type 2B activity	5	rRNA (uridine) methyltransferase activity	2

protein phosphatase type 2C activity	5	rRNA methyltransferase activity	2
sugar porter activity	5	S-methyltransferase activity	2
thiamin-pyrophosphatase activity	5	selenocysteine methyltransferase activity	2
UDP-2,3-diacylglucosamine hydrolase activity	5	sigma factor activity	2
acyltransferase activity	4	site-specific DNA-methyltransferase (cytosine-specific) activity	2
ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism	4	sodium-dependent phosphate transporter activity	2
biotin binding	4	stearoyl-CoA 9-desaturase activity	2
cation channel activity	4	steroid dehydrogenase activity	2
chitin synthase activity	4	tert-butyl alcohol 2-monooxygenase activity	2
cysteine-type peptidase activity	4	tetracycline:hydrogen antiporter activity	2
disulfide oxidoreductase activity	4	thioredoxin-disulfide reductase activity	2
DNA ligase (ATP) activity	4	tocopherol O-methyltransferase activity	2
DNA-directed DNA polymerase activity	4	transaminase activity	2
DNA-directed RNA polymerase activity	4	transferase activity, transferring phosphorus-containing groups	2
DNA-directed RNA polymerase I activity	4	translation elongation factor activity	2
DNA-directed RNA polymerase II activity	4	tri-n-butyltin dioxygenase activity	2
DNA-directed RNA polymerase III activity	4	tRNA (adenine)-methyltransferase activity	2
heat shock protein binding	4	tRNA (cytosine)-methyltransferase activity	2
metal ion binding	4	tRNA (guanine) methyltransferase activity	2
monooxygenase activity	4	tRNA (guanosine) methyltransferase activity	2
oligosaccharyl transferase activity	4	tRNA (uracil) methyltransferase activity	2
pseudouridine synthase activity	4	tRNA (uridine) methyltransferase activity	2
pseudouridylate synthase activity	4	tRNA methyltransferase activity	2
unfolded protein binding	4	tRNA-pseudouridine synthase activity	2
1-alkyl-2-acetylglycerophosphocholine esterase activity	3	tryptophan-tRNA ligase activity	2
acetylglucosaminyltransferase activity	3	ubiquitin thiolesterase activity	2
alpha DNA polymerase activity	3	UDP-glucose:glycoprotein glucosyltransferase activity	2
beta DNA polymerase activity	3	UDP-N-acetylglucosamine-peptide N-acetylglucosaminyltransferase activity	2
delta DNA polymerase activity	3	versicolorin reductase activity	2
		voltage-gated potassium channel activity	2

Table 7.7: ***GO term enrichment analysis.*** *Cellular components of DTEGs under 24 h of high light stress.*

GO_term	n	GO_term	n
membrane	80	**cAMP-dependent protein kinase complex**	2
nucleus	53	**chromosome**	2
integral to membrane	22	**endoplasmic reticulum**	2
intracellular	21	**Golgi membrane**	2
cytoplasm	13	**oxygen evolving complex**	2
ubiquitin ligase complex	13	**procollagen-proline, 2-oxoglutarate-4-dioxygenase complex**	2
ribosome	12	**protein kinase CK2 complex**	2
nucleosome	11	**protein serine/threonine phosphatase complex**	2
extracellular region	10	**ribonucleoprotein complex**	2
proton-transporting two-sector ATPase complex	6	**signal recognition particle (sensu Eukaryota)**	2
magnesium-dependent protein serine/threonine phosphatase complex	5	**small nucleolar ribonucleoprotein complex**	2
mitochondrial inner membrane	5	**voltage-gated potassium channel complex**	2
myosin phosphatase complex	5		

Table 7.8: ***Genes with exclusive differential translational regulation upon 24 h of high light stress with adjusted p-value < 0.05.*** *FC, fold change, TE, translation efficiency, Protein ID from JGI.*

Category	Protein ID	Log_2FC TE	Adjusted p-value
Upregulated TE	10098	5,801501566	0,00064481
	10100	4,165249586	0,00732116
	11142	4,096196446	3,33E-05
	6097	3,77141	0,00553876
	38190	3,37965	2,23E-07
	261067	3,241323156	0,00390826
	11267	2,84722	2,87E-08
	11943	2,819242584	3,16E-06
	4349	2,768825194	1,94E-15
	262946	2,331162372	1,70E-17
	23937	2,077473397	2,13E-06
	263428	2,03383	1,13E-05
	3963	1,872519587	0,00081988
	260933	1,746612114	2,00E-07
	37258	1,70657	0,00175039
	21224	1,618377624	0,00048974

	5866	1,577441647	0,00705163
	25088	1,557742589	0,00107633
	8370	1,55774	0,00272036
	268486	1,55379	0,0254467
	20622	1,530731217	0,01933879
	11366	1,481719814	0,0002897
	35005	1,479844791	0,0064269
	31298	1,477679482	0,02088387
	3770	1,452362759	0,0019456
	9678	1,440230996	0,03367102
	6294	1,439312137	0,02296753
	25457	1,42127	0,00418085
	25872	1,41691	1,41E-05
	38513	1,411321802	7,95E-05
	23811	1,393355614	3,22E-07
	8420	1,33664	0,02219164
	4411	1,328772988	0,00012236
	8348	1,307734956	0,02554158
	42828	1,26242	9,34E-05
	2019	1,1897804	0,00459451
	35373	1,189591491	0,00823533
	9792	1,186564869	0,04717632
	23684	1,16387	1,33E-06
	1587	1,153597003	2,13E-05
	1649	1,15241	1,15E-08
	24046	1,117018338	0,0268555
	25026	1,090447582	0,00850199
	7313	1,084977753	0,00071529
	42992	1,07318	0,0135527
	10850	1,04151	0,00144786
	23872	1,035542579	0,00010697
	10092	1,030284258	0,01626646
	23820	1,0234	0,0004306
	4350	1,009686608	0,00089571
	7330	1,007978519	0,01981331
	23338	1,003151119	0,00327125
	11271	0,99554276	0,03168113
	2824	0,993778005	0,00179015
	3707	0,99136898	0,00330176
	2278	0,98423031	0,00404958
	33242	0,98201313	0,00236317
	2578	0,977385	0,01973209

24097	0,970562115	0,00626571
20640	0,967685901	0,00464926
4253	0,954248945	0,00095639
36527	0,95094575	0,00260817
7512	0,94723307	0,00159626
41106	0,935578285	0,03434803
4423	0,931314993	0,03389215
8459	0,915826618	0,01661729
262163	0,910318775	0,02280377
23066	0,898917376	0,00813057
2717	0,896299048	0,00076486
4578	0,887613143	0,00395939
24309	0,879432208	0,02489585
25262	0,878146306	0,00050903
38559	0,87798153	0,0040082
4630	0,848680546	0,04003297
28667	0,838634109	0,0003548
268217	0,838200914	0,00848008
23592	0,837430982	0,00015222
7727	0,833229292	6,48E-05
7926	0,819665745	0,00137419
35126	0,818371005	0,03599923
3230	0,811912369	0,03414582
268187	0,806905513	0,02203917
9514	0,79393017	0,01042907
11612	0,793745053	0,02963857
268817	0,790292218	0,04838614
11797	0,789036414	0,02497398
36099	0,782065208	0,04099013
2648	0,779573181	0,00183633
22135	0,778699806	0,02253388
18220	0,768483331	0,0227289
7678	0,768309904	0,00725847
17735	0,768037129	0,00368739
460	0,761142926	0,00814065
23895	0,743987901	0,03785057
9301	0,742154128	0,00257823
6974	0,735721827	0,00340526
25047	0,715353426	0,00971255
4007	0,702965067	0,02502552
22007	0,697009244	0,01523612
4858	0,695022402	0,45359208

	25127	0,69188665	0,01804726
	269474	0,688588975	0,0423707
	21592	0,684408553	0,01160371
	21756	0,678757573	7,14E-05
	30659	0,662910963	0,00813057
	950	0,658284025	0,03225042
	23379	0,655747521	0,01704453
	11478	0,65429823	0,03015731
	38460	0,638694437	0,00169117
	21282	0,631677217	0,02597261
	263562	0,622634328	0,02110613
	23230	0,613726705	0,0268555
	21668	0,606372638	0,00263969
	25506	0,603304589	0,04210933
	263321	0,601609279	0,03901496
	21177	0,577181623	0,0435352
	22862	0,553774754	0,03755591
	25218	0,551085908	0,04159886
	25164	0,531977653	0,01977704
	24494	0,491825666	0,02384376
	5153	0,478635433	0,01095765
	4224	0,464058	0,02298726
	23571	0,428285879	0,03025485
	9849	0,42730255	0,00937468
	20786	0,344778328	0,04501261
Downregulated TE	11278	-5,751124223	0,04946612
	7542	-4,296084745	0,01733354
	23160	-3,817438827	0,02066848
	8280	-3,62779	0,00092053
	22984	-3,272036318	0,00505877
	30980	-2,700191752	8,05E-06
	15424	-2,571779769	3,73E-05
	16378	-2,389107068	0,04408883
	1308	-2,205928865	1,25E-05
	19913	-2,101486482	1,31E-05
	26041	-1,927137177	0,00562745
	2698	-1,672011273	0,01535158
	20827	-1,658204426	0,00089571
	9099	-1,653307173	0,00782142
	6798	-1,645424503	0,000144
	264901	-1,595241852	0,00658345
	10589	-1,53608	6,02E-05

7895	-1,50437	0,02259866
22425	-1,48224	0,01523612
268009	-1,479045866	0,0240548
262083	-1,424815648	0,000395
1704	-1,411282479	4,67E-07
21279	-1,401137984	0,00128221
22749	-1,38618318	0,00379972
24247	-1,382321455	7,87E-14
4616	-1,33294323	0,04239708
23216	-1,32993	7,29E-06
11213	-1,293543637	0,00317287
35611	-1,289320585	0,03531129
22952	-1,28306623	2,46E-06
23556	-1,275387855	4,21E-05
34348	-1,270362538	7,73E-07
6947	-1,253744481	0,04239708
36576	-1,246549086	0,00692096
38578	-1,235449443	0,03495709
10772	-1,234625317	6,65E-07
13224	-1,215076187	0,00420026
35911	-1,20687	0,00705163
33500	-1,160742164	0,00052002
270120	-1,148554036	7,62E-08
8327	-1,133587135	0,03525879
11432	-1,12888906	0,00160506
269821	-1,12761	0,00562745
39550	-1,121119416	0,00326891
24233	-1,1152	0,00054551
264004	-1,097410251	0,04641229
42508	-1,087985243	7,12E-06
6953	-1,081553844	0,02608487
5026	-1,035694406	5,23E-05
20889	-1,030595443	1,32E-07
24525	-1,02793	1,32E-07
23129	-1,02718394	0,00483005
42971	-0,991547899	0,00031718
261710	-0,99047514	0,00097857
26548	-0,98226999	0,0471602
41038	-0,973710224	0,00137436
22659	-0,967142644	0,04841654
14719	-0,963975691	0,00111117
29217	-0,962289012	0,02180277

23622	-0,96041003	0,00693675
269232	-0,959000781	0,0007377
270116	-0,958978375	0,00038815
24416	-0,958065151	0,01195198
25866	-0,956152043	0,00057493
23735	-0,917328696	0,00171587
263622	-0,913710344	0,04469953
19826	-0,896579774	0,04621981
269937	-0,8928503	0,020037
37277	-0,892296552	0,00913412
25478	-0,88055793	0,00802218
23606	-0,869295935	0,00029541
10818	-0,865963192	0,02474776
9148	-0,857367634	0,00022896
36202	-0,84898548	0,02026918
25708	-0,846404024	5,25E-06
269080	-0,83879862	0,03452094
26221	-0,822042076	0,0397818
24295	-0,805312243	0,02781543
21455	-0,804575741	0,00813057
6780	-0,800205004	0,04641229
35274	-0,799565508	0,01898411
21067	-0,798098218	0,02346965
268526	-0,793801356	0,02666804
2779	-0,792701725	6,49772E-05
25869	-0,789775268	0,030698763
4138	-0,788129691	0,040062535
264387	-0,784472506	0,005627453
36037	-0,775527919	0,030067809
6533	-0,762853556	0,004232759
20653	-0,76258532	0,001373888
3416	-0,762567606	0,034573088
8680	-0,759061334	0,005315298
268398	-0,750518656	0,000734341
35726	-0,735447451	0,004649261
7475	-0,731379624	0,000131221
39864	-0,729187672	0,000928667
262091	-0,718008162	0,002526028
41117	-0,712088597	0,025869029
4160	-0,710370975	0,027594442
9073	-0,695534738	0,012659825
12030	-0,686781085	0,004311297

	34661	-0,67337317	0,039928634
	23031	-0,665976948	0,011754483
	9772	-0,653290802	0,004800676
	23315	-0,652134514	0,013648407
	10560	-0,641967522	0,011577263
	41300	-0,61535117	0,005012213
	6770	-0,613612102	0,02052603
	23753	-0,597854994	0,012716805
	22120	-0,594225608	0,011263022
	33432	-0,574865013	0,002219697
	4524	-0,572334054	0,00431296
	24483	-0,5566883	0,016266461
	268552	-0,544209379	0,020496327
	41333	-0,540366514	0,02088387
	30193	-0,538872637	0,025811448
	24042	-0,524880505	0,006728529
	25539	-0,494581628	0,024054796
	16075	-0,488852541	0,024747764
	22907	-0,443318424	0,008196566
	13393	-0,378955541	0,035016973

Table 7.9: ***Genes with exclusive differential translational regulation upon 4 h of high light stress with adjusted p-value < 0.05.*** *FC, fold change, TE, translation efficiency. Protein ID from JGI.*

Category	Protein ID	Log_2FC TE	Adjusted p-value
Upregulated TE	2034	8,08895066	0,00024703
	31875	7,03245647	0,03632755
	9674	6,82547881	0,00172338
	269828	6,50275139	0,00543954
	269871	6,10152809	0,00078
	8611	5,8473848	0,00554843
	3330	5,69307314	0,00241228
	7145	5,54479963	0,00182644
	24923	5,47959545	0,00064933
	1783	5,47143308	0,01178801
	3876	5,42777662	0,00104866
	24722	5,37961024	0,02377782
	20674	5,37118156	0,00104445
	1822	5,34860281	0,00296837

	2605	5,11701022	0,03995814
	10738	5,10209963	0,00655218
	9127	4,94248979	0,00433582
	10867	4,65011059	0,01677528
	10164	4,53476557	0,00825306
	7632	4,49676697	0,03485504
	35342	4,46265632	0,00044524
	6097	4,20890693	0,00615182
	25067	3,94301985	0,00073764
	5043	3,88546576	0,02615689
	9443	3,4263427	0,00187827
	11743	3,22971654	0,01325323
	22947	3,12991455	0,01137157
	4481	3,10476381	0,03006946
	2116	3,09327625	0,01011816
	25389	2,54120734	0,00104866
	24953	2,53929356	0,03006946
	24950	2,4056464	0,01320469
	22785	2,17962046	0,02101391
	7202	2,11168561	2,63E-05
	9963	2,06049331	0,01037108
	24714	2,00139621	0,03611898
	10806	1,92641357	0,00435276
	11943	1,79671333	0,0109093
	4349	1,74733153	0,00010389
	38513	1,63207998	7,70E-08
	27273	1,61898059	3,97E-05
	263428	1,44595697	0,01011816
	36045	1,43006892	0,03995814
	9163	1,41820824	0,00064933
	268714	1,32899602	5,19E-05
	260933	1,23328718	0,01018137
	23811	0,95847588	0,01946452
	20640	0,93837467	0,02055146
	1955	0,92820924	0,00482035
	263212	0,89969435	5,58E-05
	6564	0,89400025	0,00256502
	35005	0,87812215	0,04713293
	2270	0,87618451	0,0278967
	21250	0,776033	0,01325323
	1649	0,71504228	0,02257299
	25281	0,64540511	0,01385246

Downregulated TE	264804	-5,3529216	0,01788063
	12179	-3,59249	0,02654463
	9258	-3,038938	0,00727338
	9261	-2,9507554	0,0068117
	37791	-2,800654	0,00825306
	30990	-2,6268908	0,00189282
	10169	-2,587413	0,03809003
	25107	-2,3348348	0,04130958
	23503	-2,2187745	0,01767812
	7416	-2,181706	0,0301226
	31382	-1,7153954	0,00104866
	20952	-1,6323159	0,00313843
	6971	-1,5947476	0,0082688
	25436	-1,3881466	0,02022403
	34348	-1,3718832	0,00014732
	9401	-1,3615164	0,04909866
	25679	-1,3612262	0,04263591
	21620	-1,353638	0,04786133
	2039	-1,3434978	0,04066462
	38046	-1,2774876	0,00658664
	21333	-1,2664046	0,03660312
	37294	-1,2088699	1,08E-05
	262151	-1,1169696	0,02122421
	13459	-1,1136891	0,04479565
	26548	-1,0981571	0,00520306
	261925	-1,0817901	0,0164995
	263012	-1,0492631	0,01826695
	28413	-1,0192499	0,02304519
	21339	-1,0120945	0,01523303
	1049	-1,0118661	0,01230632
	39315	-0,8214757	0,03632755
	11746	-0,8099782	0,03179133
	37534	-0,7814126	0,04055297
	22952	-0,7057751	0,04254883
	11296	-0,643193	0,03006946
	36582	-0,6163864	0,02140671
	24171	-0,5637173	0,02531352

7.5 Appendix E: Sequences of codon modified genes

Lhcx6 WT vs **Lhcx6 MF**

```
WT    ATGAAATTCACTCTCCTCTCCTCGGCCATCGTAGCCACCTCTGCCTTCGTCGCTCCTTCA  60
      ||||| || || ||  | ||||||||||| ||||| ||||| || || || ||||| 
MF    ATGAAGTTTACCCTTTTGTCCTCGGCCATTGTAGCTACCTCAGCATTTGTTGCTCCGAGC  60

WT    CCCAGCACTATCGCCTCCACCGCTCTCTTCTCCACCGAAGAATCCACCGAGCAAGACATC  120
      ||    |||||||| || ||||| |||||||||||| ||||| || |||||||||||| |||
MF    CCTTCTACTATCGCATCTACCGCGCTCTTCTCCACGGAAGAGTCAACCGAGCAAGATATC  120

WT    ATCACACCCGTCTCACCATCAGTAGCAGCAATCAACGGATGGACACCCAATGAAACACAA  180
      ||||| || |||   || || || || ||||| |||||||||||| || || ||||||||
MF    ATCACTCCTGTCAGTCCCTCTGTGGCTGCAATTAACGGATGGACTCCAAACGAAACACAG  180

WT    AACTGTTTCGGACTTCCCGGAAGTGTTGCTCCCACTGGATACTTCGATCCTCTTGGATTT  240
      || ||||| ||| | || ||    |||||||| || |||||||||||| ||||||||||||
MF    AATTGTTTTGGATTGCCGGGCTCCGTTGCTCCTACGGGATACTTCGACCCTCTTGGATTT  240

WT    GCCCAAGATGGAATCACACTTAATGAGATCAAACGCAATCGTGAGGCAGAAGTCATGCAT  300
      ||||| ||||| || ||||| || || |||||  | ||||| |||||||| || |||||
MF    GCCCAGGATGGTATAACACTCAACGAAATCAAGAGGAATCGAGAGGCAGAGGTGATGCAC  300

WT    GGACGTGTTGCAATGTTGGCCACACTTGGATACTTTGCTGGAGAGGCTCTTCCCAGTCCA  360
      ||||| || || |||||||| ||  | |||||||| || || || || || ||||||||
MF    GGACGCGTGGCCATGTTGGCTACGTTAGGATACTTCGCCGGCGAAGCCCTGCCCAGTCCC  360

WT    TTTGGAATTACCGGACCTGCTAATGATCAACTTCAGCAAGTTCCTCTTCCTGCCTTCCTT  420
      || ||||| || |||||||| || |||||||| || ||||| || || || || ||  |
MF    TTCGGAATCACTGGACCTGCGAACGATCAACTCCAACAAGTCCCACTCCCCGCTTTTTTG  420

WT    CTCCTCACTGCCGGTATTGCCAGTGCGGAATTGAAACGTGCTAATATTGGATGGGTCGAG  480
      ||  | |||||||| |||||||| || ||  | || |||||||||||| |||||||| |||
MF    CTTTTGACTGCCGGCATTGCCAGCGCTGAGCTTAAGCGTGCTAATATCGGATGGGTAGAG  480
```

```
WT    CCTGACTTTGGAAACTGGACCAAGACTTTGTGGAAGCTTCGCGACAACTACTACCCTGGT  540
      || |||||  || |||||||||||  ||||||||||||||  ||  ||||||||||||  ||  ||
MF    CCAGACTTCGGTAACTGGACCAAAACTTTGTGGAAGCTGCGTGACAACTACTATCCCGGG  540

WT    GACGTTGGTTTTGATCCTTTGGGATTGAAGCCTACAGATGCCAAAGCATTTGCTGATATG  600
      || |||||  || |||||   | ||   | |||||  ||  |||||  ||||||||  |||||  |||
MF    GATGTTGGATTCGATCCACTCGGGCTCAAGCCAACGGATGCAAAAGCATTCGCTGACATG  600

WT    CAGACTCGTGAATTGCAGAATGGGCGGTTGGCTATGATTGGTGCTATTGGTATGATTTCT  660
      |||||   | ||   | |||||  ||  ||   | ||  |||||  ||  ||  ||||||||||||   |
MF    CAGACAAGAGAGCTACAGAACGGTCGTCTTGCCATGATCGGAGCGATTGGTATGATTAGT  660

WT    CAGGAGCTGGTGAATCACAGGACTATCATGGGAACTATCGATTTCTACAACAAGGTGTAC  720
      ||  |||  ||||  |||||  |||||  ||  |||||  ||  ||  |||||  |||||  |||||||||
MF    CAAGAGTTGGTTAATCATAGGACAATTATGGGCACCATTGATTTTTACAATAAGGTGTAC  720

WT    TCAGGTGTCAATCCGTATGAGGGATGTGGAGATGGTGTCATTTGCTAA  768
      ||  ||||||||  ||  |||||  ||  ||  ||  ||  ||  |||||  ||  |||
MF    TCGGGTGTCAACCCTTATGAAGGTTGCGGGGACGGAGTCATCTGTTAA  768
```

Figure 7.7: Alignment of Lhcx6 wild-type gene (WT, accession number XM_002295147.1 black) and Lhcx6 codon modified gene (MF, green). Gaps represent codon optimized regions.

RPL10a WT vs **RPL10a MF**

```
WT   ATGTCAAACAAGTTGAACTCCGCCCTCCTCGACAAGGCCGTCGAGGACATCTTGGCCTTC  118
     ||||| || || || || || ||  |  | ||||| || || |||||||| || || ||
MF   ATGTCGAATAAATTAAATTCGGCGTTATTAGACAAAGCGGTAGAGGACATATTAGCGTTT  60

WT   TCCGCCGGTGAAACCATCACCAAGGGTGGTGAGGAACTCAAGGGAAAGAAGCGTAACTTC  178
     || || || ||||| || || || || || || ||| | || || || || || || ||
MF   TCGGCGGGGGAAACGATAACGAAAGGGGGGGAAGAATTAAAAGGGAAAAAACGGAATTTT  120

WT   ACTGAAACCATTGAGATCCAAATCACCCTTAAGAACTACGATCCTCAGCGTGACAAGCGA  238
     || ||||| || || || || || ||  | || || || || || ||||| ||||| ||
MF   ACGGAAACGATAGAAATACAGATAACGTTAAAAAATTATGACCCGCAGCGGGACAAACGG  180

WT   TTCTCCGGAACCTTCCGTCTTCCTGCCATCCCCCGTCCCAACATCAAGTGCTGCATGCTC  298
     || || || || || ||  | || || || || || || || || || |||||||||| |
MF   TTTTCGGGGACGTTTCGGTTACCGGCGATACCGCGGCCGAATATAAAATGCTGCATGTTA  240

WT   GGAAATGCCGCCCATTGTGAGCAGGCCGATCGTATCGGCGTAGCTCACATGAGTACCGAG  358
     || ||||| || ||||| || ||||| || || || || ||||| || |||   || ||
MF   GGGAATGCGGCGCATTGCGAACAGGCGGACCGGATAGGGGTAGCGCATATGTCGACGGAA  300

WT   GATCTTAAGAAGCTCAACAAAAACAAAAAGTTGGTGAAGAAGCTTGCCAAGAAGTATGAC  418
     ||| | || ||  | || ||||| ||||| || || || ||  | || || || ||||||
MF   GATTTAAAAAAATTAAATAAAAATAAAAAATTAGTAAAAAAATTAGCGAAAAAATATGAC  360

WT   TTCTTCCTTGCCTCTGACAACATGATCAAGCAGATCCCCCGTCTTTTGGGACCCGGTCTT  478
     || ||  | || || ||||| ||||| || ||||| || ||  | || || || ||  |
MF   TTTTTTTTAGCGTCGGACAATATGATAAAACAGATACCGCGGTTATTAGGGCCGGGGTTA  420

WT   ACTAAGGCTGGTAAGTTCCCCACCCTTCTTGCTGGCGGAGAGGACATGCAGGAGAAGATT  538
     || || || || || || || ||  |  | || || || ||||||||||||| || ||
MF   ACGAAAGCGGGGAAATTTCCGACGTTATTAGCGGGGGGGGAGGACATGCAGGAAAAAATA  480

WT   GACGAGGTCAAGTCTACCATCAAGTTCCAGATGAAAAAGGTCATGTGCCTTAACGTTGCT  598
     ||||| || || || || || || || |||||||||| || |||||| | || || ||
MF   GACGAAGTAAAATCGACGATAAAATTTCAGATGAAAAAAGTAATGTGCTTAAATGTAGCG  540
```

```
WT   ATTGGAAATGTTGATATGGACAAACAGCAAATCATTGTCAACACTCAGTTGTCTGCTAAC  658
     || || ||||| || ||||||||||||||| || || || || || ||||| || || ||
MF   ATAGGGAATGTAGACATGGACAAACAGCAGATAATAGTAAATACGCAGTTATCGGCGAAT  600

WT   TTTTTGGCGTCGCTTCTTAAGAAGCAGTGGCAGAACATTGGACAGATGTTCATCAAGTCT  718
     ||||| |||||| |  | || || |||||||||||| || || |||||||| || || ||
MF   TTTTTAGCGTCGTTATTAAAAAAACAGTGGCAGAATATAGGGCAGATGTTTATAAAATCG  660

WT   ACCATGGGACCTTCTATCCAGATTTACTTCTAA  751
     || ||||| || || || ||||| || || |||
MF   ACGATGGGGCCGTCGATACAGATATATTTTTAA  693
```

Figure 7.8: Alignment of RPL10a wild-type gene (WT, accession number XM_002291341.1, black) and RPL10a codon modified gene (MF, red). Gaps represent codon sub-optimized regions..