Université de Montréal

Transcriptional regulation in the dinoflagellates

Par

Bahareh Zaheri

Département des Sciences Biologiques, Faculté des Arts et Sciences

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de Ph.D. en Sciences Biologiques

May 2021

Université de Montréal

Département des Sciences Biologiques, Faculté des Arts et Sciences

Ce mémoire intitulé

Transcriptional regulation in the dinoflagellates

Présenté par

Bahareh Zaheri

A été évalué(e) par un jury composé des personnes suivantes

Daniel Kierzkowski
Président-rapporteur

David Morse
Directeur de recherche

Sophie Breton
Membre du jury

Senjie Lin
Examinateur externe

Jacques Belair
Représentant du doyen

Abstract

Dinoflagellates are a large family of unicellular eukaryotes found in marine and freshwater ecosystems and are important primary producers in marine ecosystem. They are famous for several distinctive behaviors including forming harmful algal blooms called “red tides”, emission of bioluminescence in the ocean, and contributing to the formation of coral reefs. They have an unusual genome structure with large amounts of DNA and permanently condensed chromosomes throughout all stages of the cell cycle. The chromatin lacks observable nucleosomes and has a liquid crystal structure. Some genes are encoded in multiple repeats located in tandem arrays producing virtually identical proteins without any known conserved elements detected in the upstream promoter regions or intergenic spacers. These unique features make it difficult to understand how gene expression is regulated. This thesis describes two experimental tests for the hypothesis that transcriptional regulation is difficult and is not the primary means of regulating gene expression in dinoflagellates.

Dinoflagellates show a paucity of transcription factors, and of these, cold shock domain proteins (CSPs) account for the majority of potential DNA binding proteins in the transcriptome. Here, the potential of dinoflagellate CSPs from free-living *Lingulodinium polyedra* and reef-forming *Symbiodinium kawagutii* (recently renamed to *Fugacium kawagutii*) to act as sequence specific transcription factors was tested. These studies using four different CSPs showed a preference for RNA over both single and double stranded DNA using electrophoretic mobility shift assays (EMSA). A second approach, testing for specific sequence binding by three cycles of selection and amplification binding (SAAB) did not enrich any consensus motif for any of the four proteins. We conclude dinoflagellate CSPs are more likely to function as RNA binding proteins than as transcription factors.

Expression of many genes in many dinoflagellate species has been reported to be regulated by light. This was tested for three genes whose expression was reported to be light-regulated in *Symbiodinium kawagutii*. The availability of a genome sequence for this species suggested that it might be possible to identify potential regulatory elements in the promoter of these genes. However, Northern blot analysis was unable to confirm differential expression of these three genes over a 24 hour light-dark cycle. Furthermore, RNA-Seq of samples taken at the end of the day and night also indicated these three genes were not light-induced. In total, only seven genes were found to be differentially expressed at dawn and dusk using RNA-Seq in triplicate with a false discovery rate (FDR) of 0.1. All were of lower abundance at the end of the light period on a 12:12 L:D cycle suggesting possible repression by light. Three of these seven, picked at random, were examined using qPCR analysis. Only two of the three had lower abundance at the end of the day by this technique, and the fold difference was less than what was observed with RNA-Seq. We conclude from this that there is little light regulation of gene expression in this dinoflagellate species.

Taken together, the studies described here support the hypothesis that dinoflagellates do not rely on regulation of genes at the transcriptional level to the same extent as other organisms.

Keywords: dinoflagellate, *Symbiodinium*, *Lingulodinium*, genome, gene expression, transcription, transcriptome, transcription factors, cold shock domain proteins, cold shock domains, light regulation

Résumé

Les dinoflagellés sont une famille d'eucaryotes unicellulaires trouvés dans les écosystèmes marins et d'eau douce et sont d'importants producteurs primaires. Ils sont réputés pour plusieurs comportements distinctifs, notamment la formation de proliférations d'algues nuisibles appelées « marées rouges », l'émission de bioluminescence dans l'océan et leur contribution à la formation de récifs coralliens. Leur structure génomique est inhabituelle avec de grandes quantités d'ADN et des chromosomes condensés en permanence à toutes les étapes du cycle cellulaire. L'ADN est sans nucléosome et se trouve dans une structure de cristaux liquides. Plusieurs gènes sont codés dans de multiples répétitions situées dans des réseaux en tandem produisant des protéines pratiquement identiques sans aucun élément conservé détecté dans les régions présumées promotrices en amont de la séquence codante. Ces caractéristiques uniques rendent difficile à comprendre comment les cellules régulent l'expression des gènes. Cette thèse examine l'hypothèse que la régulation de transcription est difficile et peu utilisée chez les dinoflagellés.

Les dinoflagellés présentent une rareté des facteurs de transcription, les protéines du domaine de choc froid (CSP) représentant la majorité des protéines de liaison à l'ADN potentielles dans le transcriptome de *Lingulodinium polyedra* et le génome de *Symbiodinium kawagutii*. Le potentiel des CSP de dinoflagellés à agir en tant que facteurs de transcription spécifiques à la séquence a été testé en utilisant des tests de déplacement de mobilité électrophorétique. Ces études ont révélé que quatre CSP différentes ont montré une préférence pour l'ARN par rapport à l'ADN simple et double brin. Une deuxième approche a examiné le ciblage de la séquence spécifique par des tests de sélection et de liaison d'amplification, et cela n'a révélé aucun motif consensus détectable dans la liaison à l'ADN. Nous concluons que les CSP dinoflagellés sont plus susceptibles de fonctionner comme des protéines de liaison à l'ARN que comme des facteurs de transcription.

Il a été rapporté que l'expression de nombreux gènes chez plusieurs espèces de dinoflagellés était régulée par l'exposition à la lumière. Cela a été testé pour trois gènes, dont l'expression régulée par la lumière chez l'espèce formant des récifs *Symbiodinium kawagutii*. La régulation de ces gènes a été rapportée dans la littérature suggérant la possibilité d'identifier les éléments régulateurs dans le promoteur. Cependant, l'analyse par transfert de Northern n'a pas pu valider le modèle d'expression de ces trois gènes chez *S. kawagutii*. De plus, le séquençage d'ARN à haut débit a confirmé que ces trois gènes n'étaient pas induits par la lumière. Au total, seuls sept gènes ont été exprimés de manière différentielle à l'aube et au crépuscule en utilisant RNA-Seq, et tous étaient de moindre abondance à la fin de la période de lumière sur un 12: 12 cycle L: D. Trois des sept ont également été examinés en utilisant une analyse qPCR, et seule deux des trois ont pu être confirmés comme étant altérés, mais avec une différence de facteur inférieure à celle observée avec RNA-Seq. Nous en concluons qu'il y a peu de régulation lumineuse de l'expression génique dans cette espèce dinoflagellé.

Dans l'ensemble, les études décrites ici appuient l'hypothèse que les dinoflagellés ont un moins grande dépendance sur la régulation transcriptionnelle que d'autres organismes.

Mots clés: dinoflagellé, *Symbiodinium*, *Lingulodinium*, génome, l'expression génique, transcription, transcriptome, facteurs de transcription, domaine de choc thermique froid, régulation par l'intensité lumineuse

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

LIST OF ABBREVIATIONS

3D:	3 dimensional
acpPC:	a-chlorophyll c2-peridinin-protein-complex
ADH:	alcohol dehydrogenase
AP:	alkaline phosphate
ATP:	adenosine 5’-triphosphate
bp:	base pair
BSA:	bovine albumin
cDNA:	complementary DNA
ChIP:	chromatin immunoprecipitation
CSD:	cold shock domain
Ct:	cycle threshold
DBP:	DNA binding domain
DEG:	differentially expressed genes
DIC:	differential interference contrast microscopy
DNA:	deoxyribonucleic acid
DTT:	dithiothreitol
EDTA:	ethylenediaminetetraacetic acid
EMSA:	electrophoretic mobility shift assay
FDR:	false discovery rate
GTP:	guanosine 5’-triphosphate
GST:	glutathione S-transferase
HAT:	histone acetyltransferases
HDAC:	histone deacetylases

Hi-C:	high throughput chromatin conformation capture
IPTG:	isopropyl β-D-1-thiogalactopyranoside
Kb:	kilobase
kDa:	kilodalton
KEGG:	Kyoto encyclopedia of genes and genomes
LB:	Luria-Bertoni
LBP:	Luciferin binding protein
LCF:	Luciferase
LD:	Light/dark
lncRNA:	long non-coding RNA
mg:	miligram
mL:	mililiter
mM:	millimolar
mRNA:	messenger RNA
MS:	mass spectrometry
NPC:	nuclear pore complex
nm:	nanomolar
OD:	optical density
OEE1:	oxygen evolving enzyme
ORF:	open reading frame
PBS:	phosphate-buffered saline
PCNA:	proliferating cell nuclear antigen
PCP:	peridinin Chlorophyll-a binding protein
PCR:	polymerase chain reaction
PEG:	polyethylene glycol

PFAM:	protein families
PPR:	pentatricopeptide repeat protein
pg:	picogram
PI:	propidium iodide
PMSF:	phenylmethanesulfonyl fluoride
PSII:	photosystem II
RBC:	Rubisco
RBCL:	Rubisco large subunit
RBPs:	RNA binding proteins
RCC1:	regulator of chromatin condensation
RNA:	ribonucleic acid
RNAP:	RNA polymerase
ROS:	reactive oxygen species
RRM:	RNA recognition motif
SAAB:	selection and amplification binding assay
SDS:	sodium dodecyl sulfate
SDS-PAGE:	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEM:	scanning electron microscopy
SL:	splice leader
snRNA:	small nuclear RNAs
snRNPs:	small nuclear ribonucleoprotein complexes
TAD:	topologically associating domains
TAF:	TBP-associated factor
TBP:	TATA-box binding-protein
TF:	transcription factor

TLF:	TATA-box like factor
TPM:	transcripts per million
Tris:	2-amino-2hydroxymethyl-1,3-prpanediol
tRNA:	transfer RNA
UTR:	untranslated region
UV:	ultra violet

DEDICATION

I would like to dedicate this thesis to my mother, with her great love, kindness and sacrifices have been always supporting me emotionally through my darkest days, and my father for always inspiring me with his strength and guidance.

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to my supervisor, Professor David Morse, whose insightful feedback and methodological advice helped me to sharpen my problem-solving outlooks and brought my work to a higher level. He has always motivated me with his enthusiasm, profound belief in my abilities, and constructive criticism to pursue my projects precisely. The completion of my dissertation would not have been possible without his continuous contribution. It was a great honor and precious experience for me to work with him.

I would like to extend my appreciation to Professor Daniel P. Matton, Professor Annie Angers and Dr. Jean Rivoal for their helpful contributions and practical suggestions. I am also grateful to all the faculty members, administrative staff and students of IRBV for their help.

I had a great pleasure of collaborating with an intelligent co-researcher, Dr. Steve Dagenais Bellefeuille, whose knowledge and expertise provided me an excellent opportunity to learn. I would also like to thank Dr. Mathieu Beauchemin for his help during the early days at the beginning of my PhD. Special thanks to dear Carl Bowazolo, for helpful advices and motivating discussions about scientific matters.

I am deeply indebted to my parents and my brother for their wise guidance and sympathy during the period of my PhD studies. Although I was far from home, their kindness and virtual presence kept me going and my success would not have been achieved without their love and support.

I gratefully acknowledge the financial help of the Department of Biological Science and the NSERC through grants, scholarships and funding in our laboratory. Without their generous support and funding, my projects could not have reached their goals.

CHAPTER 1- INTRODUCTION

Section 1.1 in the thesis has been submitted to the journal *Gene* with the title *“An overview of transcription in dinoflagellates”* by Bahareh Zaheri and David Morse and is currently under peer review.

Contributions

I wrote the first draft of the manuscript which was then reviewed and corrected by David Morse.

Funding

Funding was obtained from the Canadian National science and Engineering Research Council (grant number 171382 to D.M.). The funding body played no role in the design of the study and collection, analysis, and interpretation of data, or in writing the manuscript and decision to publish.

1.1.1. Abstract

Dinoflagellates are a vital diverse family of unicellular algae widespread in various aquatic environments. Typically large genomes and permanently condensed chromosomes without histones make these organisms unique among eukaryotes in terms of chromatin structure and gene expression. Genomic and transcriptomic sequencing projects have provided new insight into the genetic foundation of dinoflagellate behaviors. Genes in tandem arrays, trans-splicing of mRNAs and lower levels of transcriptional regulation compared to other eukaryotes all contribute to the differences seen. Here we present a general overview of transcription in dinoflagellates based on previously described work.

Key words: transcription, regulation, gene expression, dinoflagellate

1.1.2. Introduction

Dinoflagellates are an important group of unicellular protists living in marine and freshwater habitats. Their origin of presence on earth was confirmed in the Mesozoic and Cenozoic era with significant patterns in species diversity by fossil record analysis (MacRae et al., 1996). Although, the origin of modern dinoflagellates could be found in the Precambrian era by molecular phylogenetical and anatomical comparison analysis (MacRae et al., 1996, Fensome et al., 1994 & '95). Phylogenetically, dinoflagellates with two other groups of parasitic organisms, apicomplexan and ciliates, belong to the kingdom Alveolata, with the presence of flattened vesicles termed "cortical alveoli" in all the members (Gómez, 2012). The dinoflagellate family is sub-divided into two major groups, the syndinians and the core dinoflagellates (Adl et al., 2005, Bachvaroff et al., 2014). Roughly half the marine dinoflagellates are autotrophic (photosynthetic) and mostly found in the core dinoflagellate clade. The photosynthetic dinoflagellates can live freely or within different hosts as endosymbionts. Symbiotic dinoflagellates are now categorized as Symbiodiniaceae family (LaJeunesse et al., 2018). Many dinoflagellate species are reported as heterotrophic or mixotrophic (Field et al., 1998, Dagenais-Bellefeuille and Morse, 2013). However, Syndinian dinoflagellates are mostly parasitic (Taylor et al., 2008). On a global scale, dinoflagellates and diatoms contribute in roughly half of the marine carbon fixation, which roughly equals to 25% of the global totals (Field et al., 1998). Dinoflagellate are essential for the diversity and maintenance of coral reefs in the ocean, as they feed their host with organic materials produced through photosynthesis in exchange of a light-enriched shelter (Gordon and Leggat, 2010, Muscatine et al., 1981). Yet, some marine dinoflagellates produce potent toxins and in harmful algal bloom called "red tide" can damage both marine animal life and the economy (Glibert et al., 2005). Lastly, dinoflagellates are well-known for producing the “phosphorescence of the sea”, since they are source of nightly bioluminescence in the ocean (Schmitter et al., 1976). The probable purpose of light production could be to scare the predators away (White, 1979, Buskey et al., 1985) and/or to attract secondary predators to decrease the number of primary predators (Mensinger and Case, 1992, Fleisher KJaC, 1995).

A picture of the unusual nuclear feature of dinoflagellates is now emerging about their extraordinary genome structure with huge amounts of DNA condensed permanently in a form that resembles mitotic chromosomes without any appreciable levels of histones or visible nucleosomes (Lin, 2011). Given this unusual chromatin structure, it is interesting to study why these simple eukaryotes accumulate enormous amount of DNA and how they manage to control their gene expression at the transcriptional level.

1.1.3. Dinoflagellate biology

Dinoflagellates are an important group of unicellular eukaryotes found in various aquatic ecosystems. More than 2000 living dinoflagellate species have been reported, roughly half of which are photosynthetic (Field et al., 1998), either autotrophic or mixotrophic. The other half are exclusively heterotrophic (lacking plastids), feeding via osmotrophy and phagotrophy (Ignatiades and Gotsis-Skretas, 2010). Consequently, dinoflagellates represent a large fraction of both the phytoplankton and the zooplankton in both marine and freshwater ecosystems. Dinoflagellates are also common in benthic environments and polar waters, and they can host intracellular symbionts or be endosymbionts themselves. The majority of the photosynthetic zooxanthellae of invertebrate hosts are dinoflagellate symbionts, including many species of Symbiodiniaceae family (LaJeunesse et al., 2018), the symbionts vital for the survival of coral reefs (Gordon and Leggat, 2010). They supply a large part of their hosts' nutritional needs through photosynthesis and in return receive sanctuary, a light-rich environment, and inorganic nutrients, which enables the growth and proliferation of both partners. In the host cells, *Symbiodinium* cells are in a coccoid shape, surrounded by a membrane extended from the host cell plasmalemma during phagocytosis. This membrane prevents phagosome-lysosome fusion (Peng et al., 2010). Lastly, some dinoflagellates are predators on other protozoa and some are parasites of aquatic organisms (Taylor et al., 2008).

According to phylogenetic analysis, dinoflagellates, apicomplexans, and ciliates belong to the superphylum Alveolata (Gómez, 2012). The term alveolata refers to the presence of flattened vesicles called cortical alveoli, which create a discontinuous layer underneath the plasma membrane. Dinoflagellates differ from their relatives in various ways, of which the most important is a nucleus containing a huge quantity of DNA in permanently condensed chromosomes. Lacking nucleosomes, the chromatin assumes a liquid crystal structure with numerous genes organized in tandem gene arrays (Lin, 2011).

Dinoflagellates are biochemically different, containing diverse photosynthetic pigments and toxins. They are well known for forming harmful algal blooms called "red tides" (Glibert et al., 2005). About 75-80% of toxic phytoplankton species are dinoflagellates, and their toxins are among the most potent biotoxins known. Toxins may kill fish and shellfish either directly, such as the toxin produced by *Pfiesteria piscicida* (Peglar et al., 2004). Toxic effects can also be due to large numbers of cells that clog animal gills, deplete oxygen, etc. (Smayda, 1997).

Dinoflagellates can also be sources of bioluminescence in the ocean. Bioluminescent species produce a blue-green light at night, a phenomenon controlled by an endogenous circadian (daily) clock (Hastings, 1996). In addition to this nightly bioluminescence, in *Lingulodinium polyedra*, daily photosynthesis (Hastings et al., 1961), dawn cell division (Hastings and Sweeney, 1958) and diurnal vertical migration (Roenneberg et al., 1989), are also clock regulated. Bioluminescence is produced by small cytoplasmic organelles called scintillons, containing the luciferase enzyme, the substrate luciferin and a luciferin binding protein (LBP) as well (Johnson et al., 1985, Nicolas, 1991).

Another specific characteristic of dinoflagellates is their swimming behavior in response to various environmental signals including chemotaxis, phototaxis, and geotaxis, for which movement is organized by chemical stimuli, light, or gravity, respectively. Instead of moving randomly through the water column, dinoflagellates aggregate at specific depths, which differs according to the time of day. This vertical migration has proven to be an extremely complex process that, in addition to being clock controlled, can also vary depending on species, temperature or nutritional conditions. This behavior has been reported in *Alexandrium tamarense* (Fauchot, 2005), *Gymnodinium sanguineum* (Cullen and Horrigan, 1981), *Prorocentrum micans*, *Ceratium furca* and *Lingulodinium polyedra* (Kamykowski, 1981). Light affects the extent of vertical migration, however, it may not regulate the direction of the motion. It is thought dinoflagellates are able to place themselves in a position to take full advantage of both light and nutrients (Anderson and Stolzenbach, 1985).

Dinoflagellates are typically motile unicellular organisms with two flagella. One is a ribbon-like transverse flagellum, which encircles the cell like a belt in a groove called the cingulum, and this provides a revolving force for the cell. The other one is longitudinal flagellum directed posteriorly that lies in a second groove called the sulcus. The combined action of the two flagella gives the dinoflagellates their characteristic helical swimming motion.

The majority of photosynthetic species possess pigments such as chlorophylls a and c2, the carotenoid beta-carotene, and a group of xanthophylls unique to dinoflagellates, including peridinin, dinoxanthin, and diadinoxanthin. Peridinin in particular is responsible for the typical reddish-brown color of most photosynthetic dinoflagellates and is found bound to a soluble peridinin-chlorophyll *a*- protein (PCP) which is also unique to dinoflagellates. Other colors are due to the presence of other pigments such as fucoxanthin obtained by additional endosymbiotic events (Hackett et al., 2004a) (see below). Other organelles in dinoflagellate cell include mitochondria, rough and smooth endoplasmic reticulum, Golgi apparatus, lipid and starch particles, and nutrition vacuoles. Moreover, some species have a light-sensitive eye-like organelle, called ocelloid (Gavelis et al., 2015).

There are two visually different dinoflagellate cell types which are due to differences in the contents of the cortical alveolae. The so-called unarmored dinoflagellates have a single layer of flattened seemingly empty cortical vesicles underneath their outer plasmalemma, which distorts easily. Armored dinoflagellates on the other hand contain polysaccharides, principally cellulose, in their alveolae. The alveolae have specific shapes and are arranged in distinct species-specific patterns, and give the cells a more rigid, inflexible wall. Dinoflagellates show abundant diversity in external morphology described as horns, wings, spheres, collars, arms and hands with fingers (Hackett et al., 2004a).

Despite the fact that a huge proportion of plastid-containing dinoflagellates have the unique photopigment peridinin, dinoflagellates can also contain a variety of plastid types (Schnepf, 1999). The peridinin plastid is extremely different from that of other photosynthetic eukaryotes as it is surrounded by three membranes and lacks a typical plastid genome. Most plastids are surrounded by either two or four membranes and contain a ~150 kb circular genome encoding hundreds of genes essential for plastid function (Hackett et al., 2004a). In contrast, the plastid genome of peridinin-containing dinoflagellates has been broken into minicircles that encode a single, or at most only a few genes per mini circle. Currently, just 16 genes have been identified on minicircles (Barbrook and Howe, 2000, Zhang et al., 1999, Wang and Morse, 2006) meaning dinoflagellates have the smallest genome of any functional plastids. The remaining genes needed for photosynthesis have been transferred from the plastid to the nucleus. Studies on *Alexandrium tamarense* reported only 15 genes in the plastid while 48 photosynthetic genes found only in the plastids of all other eukaryotes have been moved to the nucleus (Hackett et al., 2004b). In dinoflagellates, nuclear-encoded plastid proteins are targeted to the plastid using a tripartite N-terminal targeting signal (Nassoury et al., 2003). However, four other types of plastids have been found in the dinoflagellates, and these have been acquired through endosymbiosis whose evolutionary lineages indicate a variety of different evolutionary origins such as from haptophytes, cryptophytes, diatoms, or prasinophytes. For example, *Karenia brevis, Karenia mikimotoi* and *Karlodinium micrum* have a fucoxanthin-containing plastid and do not contain peridinin. As these pigments are typically found in haptophyte algae, it is assumed that this plastid originated from a haptophyte alga through endosymbiosis (Tengs et al., 2000). The phylogeny of nuclear-encoded plastid-directed genes in *Lingulodinium* suggests a common ancestry with those of diatom plastids (Wang et al., 2008) although the origin of peridinin and the PCP to which it binds is unknown.

Dinoflagellates are generally haploid vegetative eukaryotes (Wisecaver and Hackett, 2011) and mitosis represents the most common form of reproduction. Asexual reproduction occurs by binary fission and while this is the major means of reproduction during optimal environmental conditions, sexual reproduction may occur under some conditions (Figueroa et al., 2015). A diploid zygote results from the fusion of two sexual haploid gametes, and vegetative reproduction can only start up again after the diploid zygote undergoes meiosis. Several dinoflagellates generate resting phases, called dinoflagellate cysts, as part of their life cycles (Lin et al., 2009), and a diploid zygote may form a dormant non-motile resting cyst (Uchida, 2001). The nuclear division of *Symbiodinium* occurs in darkness on culture media, according to light and electron microscopy and nuclear staining evidence (FITT and Trench, 1983). After a light exposure, two motile cells are produced through cytokinesis. At the onset of darkness, these motile flagellated cells lose their flagella and transform to a coccoid form without the ability to swim. Meiosis and sexual recombination has not yet been reported in *Symbiodinium* (Santos et al., 2004) although they have a gene complement consistent with the ability to undergo meiosis (Morse, 2019).

1.1.4. Chromatin structure

Eukaryotic cells all contain a nucleus in which the genetic material is surrounded by a double membrane nuclear envelope. The nucleoskeleton, including the nuclear lamina just underneath the envelope, creates a fibrous network within the nucleus to give mechanical support to the envelope and also plays a role in organizing the chromatin. Two matrix proteins, nuclear lamins and topoisomerase II have been found in dinoflagellates (Minguez et al., 1994). They may be involved in higher-order genome arrangement, chromatin regulation, transcription, DNA replication and DNA repair (Dechat et al., 2009).

The nucleus is one of the main sites in which there is regulation of gene expression. The acidic DNA molecules in a typical eukaryotic nucleus are typically surrounded by numerous basic proteins called histones, which fold and compact DNA into chromatin. A stretch of DNA containing 147 base pairs is wrapped around four pairs of the core histones H2A, H2B, H3 and H4 creating a nucleosome, the fundamental unit of chromatin (Luger et al., 1997). H1 acts as a linker histone allowing higher order assembly of nucleosomes (Hergeth and Schneider, 2015). Core histones contain two different domains, one a 20-35 N-terminal motif called the histone tail and the other a 80-90 C-terminal histone fold which interacts with other histones and DNA (Luger et al., 1997). The N-terminal histone tail can undergo numerous post-translational modifications including methylation, acetylation, phosphorylation, ubiquitination, SUMOylation, citrullination and ADP-ribosylation that may result in epigenetic regulation of gene expression (Mersfelder and Parthun, 2006). The chromatin in dinoflagellates is visibly different and has no detectable nucleosomes (Wisecaver and Hackett, 2011). Instead, the chromosomes have a characteristic banded pattern under the electron microscope and have been proposed to be in a liquid crystal state. There are two chromosomal regions in dinoflagellates, a main body containing transcriptionally inactive DNA and a peripheral dispersed region composed of DNA filaments involved in RNA transcription (Sigee, 1983). This conclusion, based on electron micrographs of chromatin after incorporation of radiolabelled adenine, may help explain how transcription can occur despite dinoflagellate chromosomes being condensed throughout all stages of the cell cycle. RNA binding proteins are among the major components of nuclear proteins (Beauchemin and Morse, 2018) suggesting that RNA may be involved in structural organization of the chromatin.

Dinoflagellates have a basic nuclear protein to DNA ratio of 1:10 (Rizzo et al., 1982), which is considerably lower than either the 1:1 ratio found in other eukaryotes (Kellenberger, 1988) or the 1:1.75 ratio seen in prokaryotes (Holck et al., 1987). The low levels of basic nuclear protein associated with dinoflagellate genome, and the correspondingly high DNA concentration, is the principal reason why the liquid crystal structure has been proposed for the chromatin structure (Kellenberger and Arnold-Schulz-Gahmen, 1992). Indeed, the concentration of DNA within the dinoflagellate nucleus corresponds to what would form cholesteric liquid crystals *in vitro* (Rill et al., 1991). This structure is so unusual that it was once proposed that dinoflagellates belonged to a kingdom intermediate between the prokaryotes and the eukaryotes, the Mesokaryota (Davies et al., 1988). Long thought to be completely absent, histones are now known to be present but at very low levels (Beauchemin and Morse, 2018, Gornik et al., 2012). Instead of histones, dinoflagellate nuclei contain two other groups of basic proteins. The first to be discovered were histone-like proteins (HLPs) (Vernet et al., 1990), initially identified in *Crypthecodinium cohnii* (Wong et al., 2003, Sala-Rovira et al., 1991). Various HLPs have been found in other dinoflagellates as well (Rizzo, 2003). For example, *L. polyedra* also contains an HLP with sequence specific DNA binding activity (Chudnovsky, 2002). Dinoflagellates HLPs are phylogenetically related to bacterial HLPs (Wong et al., 2003) and appear to be responsible for regulating the condensation of DNA loops and the access of genes to transcription factors (Chan and Wong, 2007). A second group of basic proteins includes dinoflagellate viral nucleoprotein (DVNP), which, like histones, can be modified post-translationally (Gornik et al., 2012). DVNP was proposed to have been gained by lateral gene transfer from an algal virus early in evolution of the dinoflagellates. These proteins appear in the early branching *Hematodinium* that have lost, apparently for the first time, the ability to form nucleosomes. They are absent in the closely related and more earlier branching *Perkinsus marina* which does form nucleosomes. When expressed in yeast, DVNP displaces histones and reduces the level of transcription (Irwin et al., 2018). Interestingly, DVNP expression in yeast results in growth inhibition, with the degree of inhibition mitigated by decreasing histone expression. This thus suggests a possible mechanism for histone replacement by DVNP.

The role played by histones in dinoflagellate chromatin structure is currently unknown, but it seems likely they are important since highly conserved histone transcripts have been found in *Pyrocystis lunula* (Okamoto and Hastings, 2003), *Symbiodinium* (Bayer et al., 2012) and *Lingulodinium* (Roy and Morse, 2012). In addition, several variants including H2A.X and H2A.Z (Hackett et al., 2005, Lin et al., 2010) as well as an extensive array of histone modifying enzyme transcripts have been reported (Roy and Morse, 2012). Since the sequences are conserved but the proteins themselves are virtually undetectable, this suggests core histones are likely to be present albeit at extremely low levels. In the model for chromatin structure described above, with transcriptionally active strands spreading out from a transcriptionally inactive core (Sigee, 1983), low levels of histones might be present in extrachromosomal regions of the DNA rather than in the bulk regions of the chromatin. However, dinoflagellate genome sequences do not show the presence of small gene rich regions interspersed among larger gene poor (potentially structural) regions (Lin et al., 2015).

1.1.5. Genome structure

Different species of eukaryotes vary widely in their genome content and size, and there is no direct correlation between genome size and the complexity of an organism. The genome size of some unicellular species can thus be higher than human cells, and dinoflagellates have many examples of this where DNA content ranges from 1–250 pg/cell, equivalent to approximately 1–250 Gbp (up to eighty times the 3 Gbp haploid human genome). The 3 pg (2.9 Gbp) genome of *Symbiodinium* (LaJeunese et al., 2005) is the nearest in size to that of a haploid human cell while the *L. polyedra* genome contains over 200 pg of DNA (Spector, 1984, Beauchemin et al., 2012). The large amount of DNA is one of the factors that may contribute to the liquid crystal structure proposed for the chromatin.

Some of the dinoflagellate genes have undergone enormous amplification and recombination in the genome resulting in multiple copies of each gene following one another in the DNA as tandem repeats. This has suggested a positive correlation between the genome size and the number of genes or the gene content in an organism. The predicted gene content for dinoflagellates based on genome size was estimated to lie between 37,000 and 87,000 (Hou and Lin, 2009), although the number of unique genes may be less than this estimate in species such as *Lingulodinium* where several thousand copies of some genes exist (Lee et al., 1993, Le et al., 1997). In 2015, an almost complete sequence was reported for *Symbiodinium kawagutii*, a species chosen because of its small 1.2 Gbp genome. This study identified 36,850 genes (Lin et al., 2015). Improvements to the original genome assembly suggested this number should be lower, at 26,609 genes (Liu et al., 2018), although after further sequencing using Hi-C (chromosome conformation capture) and a combination of Illumina and PacBio sequencing of the transcriptome, 45,192 genes were identified (Li et al., 2020). These three divergent values for the same species suggest that the development of dinoflagellate gene models may be at issue rather than sequencing and assembly (Chen et al., 2019b). A convenient and comprehensive online source called SAGER

(Symbiodiniaceae and Algal Genomic Resource Database) including collected genomic data of Symbiodiniaceae and marine algal species from other databases such as MMETSP and PhyloDB, is now available (Yu et al., 2020). For the small genome of the endoparasite dinoflagellate *Amoebophrya ceratii*, about 19,925 protein-coding genes was predicted (John et al., 2019). More than 50,000 protein coding genes have been revealed for diploid genome of the free-living dinoflagellate *Polarella glacialis* (Stephens et al., 2020). In general, among the ten dinoflagellate genome assemblies available [Table 1.1.], there are roughly 30,000-45,000 genes per haploid genome (Stephens et al., 2020).

	Total assembled bases (Mbp)	**G+C content (%) of assembly**	**Genes**	**Gene models supported by transcriptome (%)**	**Gene average length (CDS+introns)**	**Exons per gene**	**Average exon length (bp)**	**Genes with introns (%)**	**Average intron length (bp)**	**Average intergenic regions length (bp)**	**Reference**
Amoebophrya ceratii	87,69	55,92	19,925	24,4	2769	3,39	577,84	71,35	337,11	1 525	(John et al., 2019)
S. microadriaticum **Clade A**	808,24	50,51	49,109	0,763	12,898	21,8	109,5	98,2	504,7	3633	(Aranda et al., 2016)
F. kawagutii **Clade F**	935,07	43,97	36,850	72,82	3788	4,1	256	64,1	893	17888	(Lin et al., 2015)
	1048,48	46	26,609	64,4	6507	8,7	199	94,03	619	23041	(Liu et al., 2018)
	936,98	45,54	45,192	90,09	7242	12,6	126	92,5	479	12704	(Li et al., 2020)
B. minutum **Clade B**	615,52	43,6	41,925	77,2	11,959	19,6	99,8	95,3	499	2064	(Shoguchi et al., 2013)
Cladocopium spp. **Clade C92**	704,78	43,00	65,832	62,5	8192	11,27	130	80,3	622	2202	(Shoguchi et al., 2018)
Symbiodinium spp. **Clade A**	766,66	49,9	69,018	67,5	8834	13,38	105	83,4	561	2008	(Shoguchi et al., 2018)
C. goreaui **Clade C**	1030	43,76	35,913	67,02	6967	12,46	130,47	96,00	593,53	9 538	(Liu et al., 2018)
P. glacialis **CCMP2088**	2756,10	46,15	51,713	94,3	13,931	10,84	108,71	75,60	1295,99	20 922	(Stephens et al., 2020)
P. glacialis **CCMP1383**	2984,68	45,91	58,232	94,0	16,206	11,64	105,67	73,79	1408,00	21 625	(Stephens et al., 2020)
Hematodinium sp.	4769	47,31									(Gornik et al., 2015)

Table 1.1. predicted gene models. Summary statistics of predicted gene models in dinoflagellate genomes.

The appearance of tandem gene arrays in the nuclear genomes of dinoflagellates is similar to what has been observed in trypanosomes. Tandem array genes in trypanosomes are expressed as polycistronic transcripts that are subsequently processed by *trans*-splicing. Interestingly, *trans*-splicing also occurs in dinoflagellates, and this has led to the proposal that dinoflagellate tandem array genes will also be expressed as polycistronic transcripts (Wisecaver and Hackett, 2011) (see below). The hypothesis of polycistronic transcripts is coherent with genome sequencing in several species, which show that changes in gene orientation are lower than for other organisms (Shoguchi et al., 2013, Stephens et al., 2020). They also agree with Hi-C experiments designed to characterize chromatin segments in physical proximity (Marinov et al., 2020) which has revealed topologically associated domains (TADs) where genes run in the same direction and whose boundaries correspond with changes in the direction of transcription. However, attempts to provide other types of experimental support for the polycistronic transcript hypothesis in dinoflagellates were unsuccessful (Beauchemin et al., 2012), including a thousand-fold less RNA from intergenic regions between mRNAs compared to regions removed during processing of primary rRNA transcripts into 28S and 18S rRNAs. In dinoflagellates, large tandem gene arrays appear to be composed of highly expressed genes, while lower expressed genes seemed to be encoded by only a single gene, often containing more introns (Bachvaroff and Place, 2008). However, this relationship does not hold in all cases (Beauchemin et al., 2012), suggesting that perhaps other post-transcriptional mechanisms may contribute to transcript abundance.

Dinoflagellate DNA contains the modified nucleotide 5-hydroxymethyluracil (5-HMeU), at levels corresponding to between 12–70% of the thymidine (Rae, 1976).

Also found are 5-methylcytosine and the rare N6–methyladenine (Rae and Steele, 1978). Lastly, between 0.5%–4% of cytosine is methylated (Blank R.J., 1988, Steele, 1980), and the methylation level appears dependent on the light condition (ten Lohuis and Miller, 1998). However, despite the many years following these observations, the functional significance of the various modifications is still unclear.

1.1.6. General control over transcription

Transcription is the first stage of gene expression in eukaryotes. This process generates a pre-mRNA transcript from the DNA, which is then spliced into a mature mRNA sequence, modified by addition of a 5' cap and a 3' poly A tail, and edited (if necessary) before being transported out of the nucleus. Transcription is followed by translation of mRNA into protein, and in some cases, by post-translational modification of the protein (Maston et al., 2006). Three different RNA polymerases, RNAP I, RNAP II and RNAP III, are responsible for the transcription of various sets of genes, generally ribosomal RNA, mRNA and tRNAs, respectively. To begin transcription, the polymerase must bind a promoter sequence. In eukaryotes, there are two main types of promoters regulating the expression of protein coding genes, those containing a TATA-box and those with CpG islands, which are regions rich in CG dinucleotides as their core domains (Carninci et al., 2006, Everett et al., 1983). TATA box promoters contain a TATA box sequence (consensus sequence TATAAA) located approximately 30 base pairs (bp) upstream from the transcription start site (TSS) in the DNA, the initiation box (INR; PyPyAN(T/A)PyPy) where transcription starts, a TFIIB recognition element or BRE (G/C)(G/C)(G/A)CGCC) and a downstream promoter element (DPE; (A/G)G(A/T)(C/T)(G/A/C)) site about 30 bases after the start site (Butler and Kadonaga, 2002). These sites allow binding of the

RNAP and a series of general transcription factors that assist the binding of the polymerase to these promoter sequences (Hernandez, 1993). The prevalent TATA box core promoter is the binding site for TATA-binding protein (TBP), a subunit of transcription factor II D (TFIID). In addition to TBP, TFIID contains a number of TBP-associated factors or TAFs (Dynlacht et al., 1991). The TFIID complex can mediate transcriptional activation, demonstrating that the TAFs have coactivator function (Zhou et al., 1993, Tanese et al., 1991).

The RNAP I promoters, which are responsible for transcription of the large rRNA genes, do not contain a TATA box even though RNAP I does require TBP (Engel et al., 2013). RNAP II promoters lacking a TATA box contain only an INR and a DPE, a combination also able to bind TFIID; the TBP-containing complex TFIID is thus utilized by promoters both with and without a TATA box. snRNAs are also transcribed by RNAP II, and these promoters consist of a proximal and a distal sequence element (PSE and DSE) (Hernandez, 2001), which are also found in what are termed class 3 RNAP III snRNA promoters (Lescure et al., 1991). The class 1 and class 2 RNAP III promoters consist of either an internal control region (ICR) containing an A box, an intermediate element and a C box, or of an A box in combination with a B box (Schramm and Hernandez, 2002). Transcription from both the TATA-less and TATA-containing RNAP III promoters requires a variety of TBP-containing complexes (Simmen et al., 1992). TBP either binds directly to the DNA in the case of TATA-containing promoters, or participates in protein-protein interactions in the case of TATA-less promoters.

Following binding and activation of the RNA polymerase at the promoter, the polymerase proceeds from 3' to 5' along the antisense strand and polymerizes ribonucleotide bases complementary to this template in the 5' to 3' direction to produce the sense RNA copy. Notably, all RNA contains uracil instead of thymine as the complementary base to adenine.

In eukaryotes, a number of different mechanisms contribute to transcription termination. These include specific and general termination factors as well as 3’ end processing enzymes that travel with the polymerase (Lykke-Andersen and Jensen, 2007).

As with other eukaryotes, dinoflagellates also contain three different forms of DNA-dependent RNA polymerases. This was first proposed for *C. cohnii*, since noticeable RNA polymerase activity still remained after inhibition of the RNA polymerase II (RNAP II) with the strong inhibitor α-amanitin (Rizzo, 1979). Sequence data confirms this, as all the necessary core and common elements for the three eukaryotic RNAPs have been found in the transcriptome of *L. polyedra*, excepting some elements that were also absent from the transcriptome of other members of the Alveolates (Roy and Morse, 2013). Interestingly, the number of RNAP components and general transcription factors in the *S. kawagutii* genome are low compared with those found in four other eukaryotes, including another member of the alveolates. Among RNAPs components, only half the number of expected core and specific subunits are present. Only 30% of the expected TFIIH components are present, and the genome lacks TFIIA, TFIIB, TFIIE and TFIIF. TBP is also absent, although *S. kawagutii* does contain a TBP-like factor called TLF which is similar to the TLF first found in *C. cohnii* (Guillebault et al., 2002) [Table 1.2.]. Clearly, the take-home message from this analysis is that the dinoflagellate genome encodes far fewer transcription factors than might be expected, even when compared to a phylogenetically close relative.

Transcription components	***H. sapiens***	***A. thaliana***	***P. falciparum***	***S. kawagutii***	***T. pseudonana***
RNA polymerase I, II and III					
Core	10	9	10	6	10
Specific	13	12	6	3	10
Common	5	5	4	0	4
Basal transcription factors					
TFIIA	2	2	0	0	0
TFIIB	1	1	1	0	0
TFIID	15	10	1	1	4
TBP	1	1	1	0	1
TFIIE	2	2	0	0	1
TFIIF	3	2	0	0	0
TFIIH (NER)	10	10	9	3	8

Table 1.2. RNA Polymerase components. Comparison of the number of components involved in transcription using KEGG pathway sequences in mammals (*H. sapiens*), plants (*A. thaliana*), alveolata (*P. falciparum* and *S. kawagutii*) and diatoms (*T. pseudonana*) with a cut off value of e^{-25}.

The *C. cohnii* TLF is homologous to TBP, but lacks the four phenylalanine residues that cooperate in binding to the TATA box (Guillebault et al., 2002). *In vitro* studies have shown that TLF does not bind the TATA box but binds instead a TTTT. This concords with the observation that genomic sequences upstream from coding sequences in dinoflagellates do not appear to have conserved TATA boxes. However, a caveat to studies of dinoflagellate promoters is that it is difficult to pinpoint their location in the genome. One reason that they are not well-defined is that (as will be discussed below) post-transcriptional *trans*-splicing adds a 22 nucleotide splice leader sequence to an acceptor site in the 5' untranslated region of all RNAPII transcripts. Since this removes all sequence between the transcriptional start site and the splice acceptor site, identifying the authentic transcriptional start site is difficult. One analysis, involving ~500 transcripts whose *trans* spliced leader acceptor site was unambiguously found in the genome, showed that a consensus sequence for the branch point was found in the genome about 20-30 bases upstream of the splice site, and that at various distances further upstream, a dinoflagellate TTTT box was found about 20-30 bases upstream from a consensus transcriptional start site (Lin et al., 2015). This would be consistent with the hypothesis that individual genes might be transcribed independently, as all sequence between the start site and the splice site would be removed during the trans splicing step. However, as mentioned above, a surprisingly large number of dinoflagellate genes are found in the same direction on the DNA, which supports the view that genes may be transcribed as polycistronic messages from a still undefined promoter element before being processing into individual transcripts by *trans*-splicing. This is clearly an important issue to resolve, as the presence of only a relatively small number of transcriptional start sites might help to explain the paucity of general transcription factors, and the identity of these sites may shed some light on why some components were not conserved.

1.1.7. Gene-specific control of transcription

Transcription factors (TFs) are the ensemble of proteins that act to activate or repress the transcription of downstream target genes by binding to the RNAP and/or to gene regulatory sequences. One group of TFs are the general transcription factors (GTFs) such as the TFII complexes described above, which are required by all mRNA genes and define the basal level of transcription (Juven-Gershon and Kadonaga, 2010).

In contrast to the GTFs, promoter-specific TFs can be different for each gene and are required for maximal level of transcription or for inducing the activated transcription (Hampsey, 1998). For example, the Hox transcription factor family is responsible for accurate construction of body parts in multicellular organisms (Lemons and McGinnis, 2006). Other transcription factors are involved in signaling cascades initiated by environmental stimuli, such as heat shock factors (HSF) which induce the expression of genes indispensable for growth at higher temperatures (Shamovsky and Nudler, 2008), cold shock proteins which permit cells to grow in temperatures below their optimum (Wistow, 1990) and hypoxia inducible factors (HIF) which regulate expression of genes allowing survival in low oxygen environments (Benizri et al., 2008).

The complement of promoter-specific dinoflagellate TFs appears to be small compared to other eukaryotes. The fraction of proteins annotated as TFs is roughly 4% in unicellular organisms such as yeast (Babu et al., 2004) but somewhere between 0.15% and 0.3% in dinoflagellates (Bayer et al., 2012, Beauchemin et al., 2012, Li et al., 2020) [Table 1.3.]. The situation is further complicated by the observation that about two thirds of the dinoflagellate proteins annotated as TFs are cold shock domain (CSD) proteins (CSPs) whose role in acting as functional TFs is not clear (Bayer et al., 2012, Beauchemin et al., 2012, Li et al., 2020). All dinoflagellate CSPs contain two RNA binding motifs, KGFGFI and VFVHF, within a highly conserved CSD of 70 amino acids (Beauchemin et al., 2016). Four divergent domain structures have been found in *Lingulodinium* CSPs, the most prevalent ones containing a CSD either alone or with a C-terminal G-rich domain. Less frequently observed are some structures containing a Zn-finger domain following the G-rich domain, as well as examples of sequences with multiple CSDs and one or more RNA recognition motifs (RRM). Many of the dinoflagellate CSPs have a structure similar to what is found in bacteria, as these typically contain only a CSD (Beauchemin et al., 2012). In *E. coli*, CSD proteins have a wide range of functions including binding DNA as transcription factors, binding to RNA, regulating transcription, splicing, and translation, and affecting mRNA stability as RNA chaperones (Mihailovich et al., 2010). Bacterial CSPs have a non-specific RNA binding function during cold stress, correlated to their chaperone activity, which helps transcription by acting as an antiterminator (Bae et al., 2000). However, the dinoflagellate proteins may be different from their bacterial counterparts as two *Lingulodinium* CSPs, both containing a single N-terminal CSD and a glycine-rich C-terminal region, were both unable to allow the growth at low temperature of an *E. coli* strain harboring a mutation in four different CSP genes (Beauchemin et al., 2016). Furthermore, cold temperatures did not induce the CSP transcripts in *L. polyedra*

(Roy et al., 2014c). In sum, dinoflagellate CSPs may not function as transcription factors at all, and if they do not, the number of sequence specific dinoflagellate TFs is remarkably small.

	H. sapiens	***A. thaliana***	***P. falciparum***	***S. kawagutii***	***T. pseudonana***
Fraction of genome as TFs	6.6%	5.5%	1.4%	0.4%	2.3%
Fraction of TFs as CSPs	0.82%	0.27%	2.6%	56%	1.6%

Table 1.3. CSPS and TFs fraction in the genome. Comparison of the fraction of the genome as specific TF and the fraction of TF as CSPs in mammals (*H. sapiens*), plants (*A. thaliana*), alveolata (*P. falciparum* and *S. kawagutii*) and diatoms (*T. pseudonana*).

The regulation of transcription factors is a crucial process, and there are some mechanisms that can lead to activating or deactivating the transcription factors, including ligand binding which can alter their subcellular localization (Whiteside and Goodbourn, 1993) and phosphorylation which may affect binding to DNA (Bohmann, 1990). Identification of post-translational modifications targeting TFs that affect transcription in dinoflagellates will have to await description of *bone fide* TFs. To date, there are no documented reports of a functional sequence specific dinoflagellate TF.

Not all TFs are DNA binding proteins. In many cases, an additional class of proteins called coactivators are responsible for augmenting gene expression by binding to a transcription factor, even though they are unable to bind DNA by themselves (Naar et al., 2001). As an example, interaction between TBP and TAFs is essential for the production of the TFIID complex. Mammalian or *Drosophila* TFIID can mediate basal and activated transcription *in vitro*, whereas TBP by itself can mediate only basal transcription, indicating that mammalian or *Drosophila* TAFs are necessary for activated transcription (Pugh and Tjian, 1990). While TAFs have been proposed to be coactivators that mediate activated transcription, recent studies have shown that TAFs can have multiple functions including core promoter-selective basal transcription (Martinez et al., 1998, Moqtaderi et al., 1996, Shen and Green, 1997), histone acetyltransferase activity (Mizzen et al., 1996), and TFIIF phosphorylation (Dikstein et al., 1996), demonstrating the significant role TAFs play in eukaryotic transcription. TBP plays a crucial role together with TAFs in communicating transcriptional regulatory factors and in the basic transcription machinery (Verrijzer and Tjian, 1996).

Many coactivators play a role in transcription not by binding the RNAP and altering its activity, but by changing chromatin structure. In most eukaryotes this can occur by post-translational modification of the histones in nucleosomes. Nucleosomes are the structures formed by histone protein octamers around which DNA is wound tightly, and the degree of nucleosome compaction affects accessibility of the DNA to sequence-specific transcription factors as well as to the RNA polymerase and the general transcription factors. Transcript elongation also involves melting and re-annealing of the double helix and extensive chromatin compaction inhibits these steps. Consequently, the more stable the chromatin structure, the more gene expression is repressed. Regulating chromatin structure is a vital basic step for regulation of gene expression, and acts in addition to the sequence-specific activators and repressors, coactivators and general transcription factors (Narlikar et al., 2002).

Typical eukaryotes regulate accessibility of the DNA template by two major group of complexes, one involving ATP-dependent complexes which hydrolyze ATP to move nucleosomes along the double helix and create accessible DNA on the surface of the histone octamer, and the other involving complexes that can alter nucleosome properties by adding or removing chemical modifying groups such as acetyl, phosphorus and methyl to histone N terminal tails (Dilworth et al., 2000). An important group of histone modifying proteins are the histone acetyltransferases (HAT) and histone deacetylases (HDAC). HAT are responsible for acetylation of histone tails, and in many cases increasing acetylation in the region of a promoter can enhance gene expression (Kuo et al., 2000). The opposite reaction, the removal of acetyl groups by HDAC, can be involved in gene repression (Khochbin et al., 2001). For example, histone acetyltransferase (HAT) activities, such as the GCN5-containing HATs and the nuclear hormone receptor HATs (Belotserkovskaya and Berger, 1999) act to decondense chromatin and activate gene expression. Similarly, the histone deacetylase complex Sin3-Rpd3 in yeast interacts with the transcription repressor Ume6p to deacetylate local histones and to repress gene expression (Fazzio et al., 2001). Thus, this type of coactivators are chromatin-remodeling complexes, which are responsible for altering the structure of chromatin and access of the transcriptional machinery to the DNA (Vignali et al., 2000). Another class of coactivators acts as a mediator complex linking a transcriptional activator to the general transcriptional machinery. This class is generally essential for transcription *in vivo* and can also stimulate high levels of activator-associated transcription *in vitro* (Boube et al., 2002).

Dinoflagellates also express a large complement of histone modifying proteins (Roy and Morse, 2012), further supporting the idea that histones, even at almost undetectable levels, are somehow involved in regulation of gene expression. Given that antibodies have not yet proven successful in detecting histones, it seems unlikely that histone modifications will be able to be linked to changes in gene expression by techniques such as chromatin immunoprecipitation (ChIP). However, reconciling the conservation of sequences encoding both histones and histone modifying enzymes with the paucity of histones remains an enigma. It is an intriguing possibility that histones may be assembled around the site of polycistronic message transcriptional initiation, as this would allow the dinoflagellates to use mechanisms for controlling transcription that are similar to those well-studied in other organisms and also account for the relative paucity of the histones themselves.

1.1.8. Differential regulation of transcription in dinoflagellates

The organization of the eukaryotic genome in the nucleus is important in DNA replication and transcription. Transcriptionally active genes are usually located in the nuclear interior while repressed genes in heterochromatic regions close to the nuclear envelope (Bickmore, 2013). Recently, Hi-C (High-throughput Chromatin Conformation Capture) has been performed on the coral symbiont *Breviolum minutum* (previously *Symbiodinium minutum*) (Marinov et al., 2020) in order to analyse the 3D structure of the genome. The dinoflagellate Hi-C maps lack the “loops” and “stripes” features found in other eukaryotes, however do contain robust topologically associating domains (TADs) between 200 to 2 Mbp in size. These TADs correspond to tandem gene arrays that are oriented in different directions from a central point to each of the TAD boundaries. These dinoTADs were degraded when transcriptional inhibitors such as triptolide and amanitin were added to the cell cultures, suggesting a transcription-induced supercoiling model formed by active polymerases which can alter the double helix structure of DNA in dinoflagellates. This important finding supports the hypothesis of a small number of transcriptional start sites, and underscores the role that transcription may play in regulating chromatin structure in dinoflagellates.

Environmental signals such as nutrition levels, temperature or light intensity can lead to different genetic responses by cells, which optimizes their survival using cell signaling pathways (Ruprecht et al., 2017) to influence transcript levels (Knijnenburg et al., 2008). For example, a microarray study on the nitrogen-depleted red tide dinoflagellate *Karenia brevis* (Morey et al., 2011) showed an increase in the level of nitrate and ammonium transporters and glutamine synthetases transcripts compared to addition of nitrogen or phosphorus supplements. Expression of pentatricopeptide repeat (PPR) protein transcripts with a role in chloroplast and mitochondria RNA processing increased up to 3-fold one hour after an addition of a nitrogen or phosphorus supplement (Morey et al., 2011). The expression of alkaline phosphatase (AP) gene was reported to be upregulated up to 6-fold following limitation of inorganic phosphorus in the culture medium (Morey et al., 2011, Lin, 2012). Regulation of alkaline phosphatase transcription in response to inorganic phosphorus stress was also described in *Amphidinium carterae*, with transcript levels increasing with a decrease in P and decreasing with an increase in P (Lin et al., 2011). In *P. lunula,* treating the cells with sodium nitrite resulted in 2-fold higher expression level of 204 genes and 4-fold upregulation of 37 genes based on microarray analysis (Okamoto and Hastings, 2003). One RNA-seq analysis in *S. kawagutii* (performed without replicates) revealed up to 50-fold differences in transcript levels for genes involved in molecular interaction, cell wall modulation, transport of nutrients such as iron and oxygen in response to heat stress (30 °C), perhaps suggesting a thermal tolerance mechanism (Lin et al., 2019). Higher transcript levels of photosystem and defence genes were induced in response to phosphate limitation or replacement of dissolved inorganic phosphate in the growth medium with glycerol-3-phosphate (Lin et al., 2019).

Temperature appears to be an important environmental cue for dinoflagellates. Rising temperatures can lead to the collapse of the *Symbiodinium*-coral symbiosis (Hoegh-Guldberg et al., 2007, Levin et al., 2016) perhaps because *Symbiodinium* cells produce more superoxide and hydrogen peroxide as a response to heat shock (Suggett et al., 2008, McGinty et al., 2012). These heat shocked dinoflagellate cells are expulsed from their coral hosts producing what is called coral bleaching (Downs et al., 2002, Krueger et al., 2015). Meiosis genes transcript levels increased up to 4-fold in both of two types of *Symbiodinium* cells (one sensitive to 32 °C and the other tolerant to 32 °C) following a nine days of adaptation at 32 °C (Levin et al., 2016). An increase of up to 4-fold in transcripts dealing with reactive oxygen species (ROS) as well as chaperone gene transcripts was also observed only for the tolerant *Symbiodinium* cells after 13 days at 32 °C (Levin et al., 2016). Another study on *Symbiodinium* (Gierz et al., 2016)*,* revealed increased level of expression of three *acpPC* (chlorophyll *a*-chlorophyll *c*$_2$-peridinin protein complex) genes during 16 days exposure to thermal stress (temperatures rising daily from 25 to 34 °C) with no changes observed for the two other *acpPC* genes. In this experiment, the efficiency of photosynthesis was reduced after eight days and after the whole 16 days period the density of *Symbiodinium* cells in the coral had decreased (Gierz et al., 2016). In *Prorocentrum minimum,* both Hsp70 and Hsp90 were upregulated by temperature increases and addition of copper (Guo et al., 2012, Guo and Ki, 2012). A recent qPCR approach testing the expression of cold shock domain protein transcripts in the harmful algal bloom forming dinoflagellate *Scrippsiella trochoidea* (*St*CSP) showed no transcriptional regulation in response to temperature stress. There was, however, a significant increase in *St*CSP transcript levels in resting cysts, suggesting a possible role for *St*CSP in encystment and cyst dormancy (Deng et al., 2019). Since there is likely to be little transcriptional activity in cysts, these CSPs may act to stabilize pre-existing transcripts in these cells.

Expression of many dinoflagellate genes have been reported to be influenced by light. Rubisco (*rbcL*) expression in cultured *Symbiodinium* under a 12:12 L:D cycle varied significantly (Mayfield et al., 2014) with *rbcL* expression increasing ~3 fold during the light phase. Levels of *acpPC* gene transcripts were higher in the dark phase comparing to light phase (Xiang et al., 2015). The acpPC protein sequence is highly homologous to the stress-related chlorophyll a/b binding proteins in *Chlamydomonas* which are also up-regulated when the cells are exposed to high light (Peers et al., 2009). *Symbiodinium* transcripts encoding the cryptochrome CRY2 decreased in high light and increased in the dark, while transcripts encoding the regulator of chromatin condensation (RCC1) protein decreased in the dark (Xiang et al., 2015) which is when many dinoflagellates begin S-phase. RCC1 binds to nucleosomes (Makde et al., 2010) and is involved in regulation of chromosome condensation in the S phase of the cell cycle (Ohtsubo et al., 1987), so it is curious that RCC1 is highly represented in *Symbiodinium* considering the permanently condensed chromatin and the lack of nucleosomes. Expression of the oxygen-evolving enhancer 1 (*OEE1*) gene of the photosystem II (PSII) complex, measured under 48 h of LL in cultured Symbiodinium (Sorek et al., 2016) showed that *OEE1* mRNA abundance increased about 3-fold during the light period and decreased during the subjective dark. However, an independent test of the expression levels of *rbcL*, *acpPC* and *OEE1* in *S. kawagutii* by both Northern blot analysis and RNA-Seq did not confirm these to be light regulated genes (Zaheri et al., 2019). This is consistent with the constant transcript level of Rubisco in *L. polyedra* over the daily cycle (Nassoury et al., 2003, Roy et al., 2014a). Clearly, the extent of light-induced transcription in dinoflagellates needs to be verified in more detail.

Meanwhile, a study in another dinoflagellate, the HAB forming *Prorocentrum donghaiense* (Shi et al., 2013), reported that *rbcII* transcript levels were higher at the cell cycle G2/M-phase under both light dark cycles and constant light. There was no rhythm when the cells were kept under constant darkness where the cell cycle remained blocked in G1 phase, and it was suggested there was a cell cycle related regulation of transcription of *rbcII* in this species (Shi et al., 2013). Expression of other genes may be linked to progression through the cell cycle as well. In the dinoflagellate *Pyrocystis lunula,* HLP transcript levels were upregulated throughout the S-phase (Wong J.T., 2005), while in *Alexandrium fundyense*, HLP transcript levels were higher during G1 phase (Wong J.T., 2005). In contrast, *Lingulodinium* histone and histone-like protein transcript levels did not increase while entering the cell S-phase thus differing from other eukaryotes where higher level of histone mRNA are found throughout S-phase (Roy and Morse, 2012). *Lingulodinium* does display higher levels of the clamp loading protein PCNA (proliferating cell nuclear antigen) at S-phase (Bowazolo et al., 2020) but this is not controlled transcriptionally as PCNA mRNA levels do not change (Roy et al., 2014b).

In some dinoflagellates, transcript levels of some genes could be different at various stages of cell growth. For example, in *A. tamarense,* a comparison between exponentially growing and stationary phase cultures showed that the former had higher levels for 489 sequences and lower levels for 4298 sequences based on microarray analysis. Higher expression of translation pathway genes and lower expression of intracellular signaling genes were observed for cells in their growth phase (Yang I., 2011). Different strains of dinoflagellates can also show differences in gene expression. For example, in a study of toxin related genes in *A. minutum,* microarray analysis has revealed higher expression of 145 sequences in toxic strains and 47 sequences in non-toxic strains (Salcedo et al., 2012).

Almost all living organisms, from prokaryotes to eukaryotes, show daily physiological rhythmic processes regulated by an endogenous circadian (about a day) clock (Kondo and Ishiura, 2000, McClung, 2006, Loros and Dunlap, 2001, Rivkees, 2007). The clock synchronizes to environmental signals such as light or temperature but clock controlled rhythms occur even in constant conditions (Roenneberg and Rehman, 1996, Roenneberg et al., 2007). For example, in *L. polyedra*, the daily photosynthesis rhythm continues to be rhythmic in constant light with a peak during the subjective day and bioluminescent light production continues rhythmically in constant darkness with a peak during the subjective night (Hastings, 2013). However, RNA sequencing revealed no rhythmic changes in transcript levels suggesting circadian changes in gene expression are regulated at translational and post-translational levels (Roy et al., 2014a). This was found to be the case for two proteins involved in the circadian bioluminescence, luciferase (LCF) (Johnson et al., 1984, Dunlap and Hastings, 1981) and luciferin binding protein (LBP) (Morse et al., 1989).

L. polyedra synthetizes both proteins at the beginning of the night and then degrades them at the end of the night. This has been proposed as a mechanism allowing cells to preserve nitrogen and recycle amino acids to produce other proteins (Hastings, 2013), perhaps due to a reduced access to nitrogen in the harmful algal blooms. However in the bioluminescent *Pyrocystis lunula,* luciferase is not degraded during the circadian rhythm of bioluminescence (Knaust R., 1998).

1.1.9. Splicing

Splicing of the primary mRNA after transcription is a critical processing step. As with other eukaryotes, precursor messenger RNA (pre-mRNA) can contain intervening sequences (introns) located among protein-coding sequences (exons), and the removal of these introns are required for efficient protein translation and function. Dynamic small nuclear ribonucleoprotein complexes (snRNPs) called spliceosomes composed of U1, U2, U4, U5, and U6 small nuclear RNAs (snRNA) bound to multiple proteins, are responsible for identifying the splice sites by base pairing, then removing introns and joining exons to release a mature mRNA (Will and Luhrmann, 2011, Wahl et al., 2009). Initial studies in dinoflagellate tandem repeat genes revealed few or no introns, such as in the *pcp*, *lbp* and *lcf* genes in *L. polyedra* (Le et al., 1997, Li and Hastings, 1998, Lee et al., 1993). The *rbc* gene in *Symbiodinium,* however, contains six introns (Rowan et al., 1996). Globally, the genome sequences from a number of different species indicate that between 73 and 96% of genes have introns, with between 11 to 19 exons per gene (Stephens et al., 2020). Introns mostly contain the conserved sequence of GU at their 5′ end and AG at their 3′ end which are recognised and spliced by spliceosomes, but in dinoflagellates there are some exceptions (Mount et al., 1992, Chow et al., 1977). For example, the Rubisco gene in

Symbiodinium contains introns with G(C/A)…..AG motifs at its ends (Rowan et al., 1996). The *sxtG* gene in *Alexandrium* contains an AG…..AG intron (Orr et al., 2013), while the *lcf C* gene intron has AT…..TC splice site in *P. lunula* (Okamoto et al., 2001). Lastly, the alcohol dehydrogenase (ADH) gene in *Crypthecodinium cohnii* contains an AGG at the 3′ and an AGG at the 5′ splice site (Mendez et al., 2015). Splice site analysis from the genome of *S. minutum* (Shoguchi et al., 2013) showed that overall, AG/GN was a conserved motif at both ends of the intron (the slash indicates the junction between introns and exons), although the AG was more highly conserved at the 3' end of the intron. *Symbiodinium* transcriptome analysis revealed that 85% of known splicing complex components were present (Bayer et al., 2012) while 70% of known components were identified in the *L. polyedra* transcriptome (Beauchemin et al., 2012). Also, in *C. cohnii*, four snRNPs were identified using antibodies for the Sm protein which is a part of all human snRNPs (Reddy et al., 1983). Here we have analyzed the *S. kawagutii* genome to retrieve the expected splicing components [Table 1.4.] and found the splicing machinery in dinoflagellates to be very well conserved when compared to other organisms.

Splicesome components	***H. sapiens***	***A. thaliana***	***P. falciparum***	***S. kawagutii***	***T. pseudonana***
General	9	8	8	9	7
U1	8	7	5	5	4
U2	12	10	7	8	10
U4/U6	7	7	6	7	7
U5	8	8	6	7	7
U5/U4/U6	5	5	4	2	5
Prp19 complex	9	8	7	5	7
Prp19 related	9	8	8	7	8
EJC/TREX	6	5	4	3	5
Common	3	3	1	1	1

Table 1.4. Splicesome components. Comparison of the number of splicesome components using KEGG pathway sequences in mammals (*H. sapiens*), plants (*A. thaliana*), alveolata (*P. falciparum* and *S. kawagutii*) and diatoms (*T. pseudonana*) with a cutoff value of e $^{-25}$.

Dinoflagellates also perform *trans*-splicing at the 5’ end of transcripts (Zhang et al., 2007), similar to what has been observed in trypanosomes (Sutton and Boothroyd, 1986). In trypanosomes, *trans*-splicing along with polyadenylation transforms polycistronic RNAs to single mRNAs. In fact, the trypanosome chromosomes contains a single bidirectional promoter, and the 5’ *trans*-splicing and 3’- polyadenylation are involved in regulating gene expression (Ouellette and Papadopoulou, 2009). Dinoflagellate *trans*- splicing adds a 22 nucleotide splice leader (SL) sequence (5′-DCCGUAGCCAUUUUGGCUCAAG-3′), originating from an SL-donor RNA (SL RNA), to an acceptor site in the 5’ end of all dinoflagellates mRNAs

(Sutton and Boothroyd, 1986, Zhang et al., 2007). The splice acceptor site appears to obey the same general rules as do cis-splicing sites, as in a study of 494 randomly selected transcripts in *S. kawagutii* whose splice acceptor site could be unambiguously mapped to the genome, there is a typical splice acceptor consensus sequence with a typical branch point consensus sequence located about 30 bases upstream (Lin et al., 2015).

SL RNAs in dinoflagellates are generally 50–60 nucleotides in length and contain an Sm binding sequence (AUUUUGG) within the exon, slightly different from other eukaryotes where the Sm binding motif is usually present in the intron (Zhang et al., 2007). However, a recent study (Song et al., 2019) has identified 18 SL genes in *S. kawagutii* between 103 to 292 bp in length, longer than the 50–60 bp previously reported. Furthermore, a new potential Sm-protein binding site, GUUUUC, has been found in the introns of these genes in *S. kawagutii*. It is possible that the only role of *trans*-slicing is to process polycistronic messages, as in the trypanosomes. However, it is also possible that *trans*-splicing plays a role in how well the transcript is translated, if only by addition of a cap to the 5' end. Further study in this area is clearly warranted.

1.1.10. RNA transport

Nuclear pore complexes (NPCs) are complex structures in the nuclear envelope that facilitate the transport of macromolecules between the nucleus and cytoplasm. The size of NPCs is different in various eukaryotes. Yeast contain 50 MDa NPCs while mammalians NPCs are 125 MDa in size (Suntharalingam and Wente, 2003, Vasu and Forbes, 2001). Molecules smaller than 40 kDa can move freely through these complexes, although larger proteins and RNA molecules require specific transport receptors and signalling pathways (Fried and Kutay, 2003). In contrast to yeast and higher eukaryotes with highly conserved NPC components, many of the key NPC components have not been reported for apicomplexans (Frankel and Knoll, 2009). Higher eukaryotes also contain *trans*-acting nuclear and cytoplasmic factors to facilitate RNA passage (Kohler and Hurt, 2007). Here we show that many of nuclear transport and central channel components are apparently conserved in *S. kawagutii*, although a significant proportion of mammalian and plant cytoplasmic transport components are absent [Table 1.5.].

	H. sapiens	*A. thaliana*	*P. falciparum*	*S. kawagutii*	*T. pseudonana*
Nucleus	11	10	9	6	8
Central channel					
Nuclear basket	4	1	0	1	1
Symmetrical nups	11	9	1	3	6
Central channel	3	3	0	3	1
Spoke complex	5	5	0	2	2
Lumenal ring	3	1	0	0	0
Cytoplasmic tails	8	6	2	4	3
Cytoplasm	53	37	17	17	24

Table 1.5. Nuclear transport components. Comparison of the number of nuclear transport components using KEGG pathway sequences in mammals (*H. sapiens*), plants (*A. thaliana*), alveolata (*P. falciparum* and *S. kawagutii*) and diatoms (*T. pseudonana*), respectively, with a cutoff value of e $^{-25}$.

Messenger RNAs can contain nonsense codons as a result of mutation or frameshifts which might be expected to produce functionally defective proteins. However, an mRNA surveillance pathway, initiated by events occurring during transcript maturation, results in cytoplasmic degradation of nonsense transcripts (Zhang et al., 1998, Maquat, 1995, Peltz et al., 1994). In *L. polyedra*, similar to other alveolates, roughly a third of mRNA surveillance nuclear factors are conserved compared to mammalians and plants, although cytoplasmic factors are more highly conserved (Roy and Morse, 2013). Here the expected nucleus and cytoplasm components for mRNA surveillance is shown in *S. kawagutii* [Table 1.6.].

	H. sapiens	***A. thaliana***	***P. falciparum***	***S. kawagutii***	***T. pseudonana***
Nucleus					
Cap binding complex	2	2	1	0	1
Exon-junction complex	15	11	4	4	5
5' capping	2	2	0	0	2
Pre-mRNA processing	14	13	4	7	8
Cytoplasm					
Nonsense mediated decay	12	9	6	6	6
No-go decay	3	33	2	0	3

Table 1.6. mRNA surveillance components. Comparison of the number of mRNA surveillance components using KEGG pathway sequences in mammals (*H. sapiens*), plants (*A. thaliana*), alveolata (*P. falciparum* and *S. kawagutii)* and diatoms (*T. pseudonana*) with a cutoff value of e^{-25}.

1.1.11. Conclusion

Dinoflagellates possess a large nuclear genome organized in permanently condensed chromosomes, so how gene expression is regulated in these organisms is an intriguing problem in basic cell biology. Next-generation sequencing approaches have revealed aspects of genome-scale transcription related to species ecology, yet additional work is required to scrutinize transcriptional regulation. Clearly, understanding chromatin structure will be one important aspect to address. Analysis of the transcription factors that actually function in dinoflagellates will also be required to shed light on the mechanisms of gene expression in these unusual organisms.

1.2. Research objectives and hypothesis

Among different dinoflagellate species, *Symbiodinium* is emerging as an important model to study various aspects of dinoflagellate biology. Its small genome has allowed it to be sequenced, providing valuable information about genome structure and organisation. In addition, *Symbiodinium* has both intracellular symbiotic and free-living lifestyles which make it an interesting model for cell biology, evolution and the development of symbioses. Clearly, exploring the transcription system will significantly accelerate our ability to manipulate it, using in particular comprehensive *Symbiodinium* transcriptome analyses to shed light on its gene expression and transcription. *Symbiodinium* provides an opportunity to create a well-developed genetic system with wide-based knowledge about the genetic machinery in a dinoflagellate species.

Possessing a large nuclear genome is a prevalent characteristic in many dinoflagellates, but fortunately *Symbiodinium* species have greatly reduced genome sizes (LaJeunese et al., 2005). The availability of several genome sequence makes it possible to begin to scrutinize transcriptional regulation. It is currently unknown if dinoflagellates contain previously unknown transcription factor families which do not appear when *Symbiodinium* is subjected to gene ontology analysis. Transcription has always been assumed to be important, and based on the fact that the environment and life style of *Symbiodinium* completely change when they enter their invertebrate hosts, it has been proposed that a different assortment of transcription factors are expressed in the symbiotic phase (Bayer et al., 2012). Interestingly, *Lingulodinium polyedra* has long been a model system for study of the molecular mechanisms regulating the gene expression in dinoflagellates by a circadian clock (Hastings, 2007). To date, all examples studied in detail have been found to involve translational rather than transcriptional control.

The extraordinary features of dinoflagellate that caught my curiosity were the enormous amount of DNA and permanently condensed chromosome. I was extremely interested to understand the regulation of transcription as a fundamental mechanism in gene expression in these unusual organisms. I started my work by studying two dinoflagellate species, *Symbiodinium* and *Lingulodinium*. Initially, as both genome and transcriptome analyses show, DNA binding domain proteins are under-represented in dinoflagellates. Furthermore, the majority of annotated transcription factors are cold shock domain proteins. Thus, the challenge was to test *S. kawagutii* and *L. polyedra* CSPs for a possible role as functional transcription factors. My idea was that if CSPs were active as transcription factors, they should specifically bind defined sequence motifs in double stranded DNA. I performed EMSA and SAAB experiments to test these predictions. I found that four different CSPs were able to bind nucleic acids, both DNA and RNA, and that no sequence specific binding activity for double stranded DNA was observed.

A second series of experiments followed from the many recent reports of altered transcript abundance by changing environmental factors such as light. *S. kawagutii* was chosen as a model system since a genome sequence was available. My idea was that transcript abundance would be due to a transcriptional response to light in *S. kawagutii.* Once a transcriptional response was validated for specific gene candidates, this would then allow identification of *cis* and *trans* acting factors involved in light induction. However, I found that three genes that were previously reported to be light-regulated in *Symbiodinum* showed no difference in transcript abundance between light and dark conditions using either Northern blot or RNA-Seq analyses. RNA Seq revealed only seven genes with significant differences in transcript levels during light and dark conditions at a false discovery rate of 0.1. A qPCR analysis with three of these seven showed only two were confirmed to be differentially expressed, and these showed smaller differences than observed with RNA-Seq result.

1.2.1. Project 1

Common transcription factor domains found in other eukaryotes, including zinc fingers, helix-loop-helix, AP2, or homeobox domains either do not exist or are scarce in *Symbiodinium* and *Lingulodinium*, and this seems to be a distinguishing characteristic of the dinoflagellate clade. Only a small percentage of the transcriptome (between 0.15% and 0.3%) is annotated as transcription factors (TF) in several species, including *Lingulodinium* and *Symbiodinium* (Bayer et al., 2012, Beauchemin et al., 2012, Li et al., 2020). Furthermore, the majority of transcription factors in *Symbiodinium* and *Lingulodinium* are cold shock domain (CSD) containing proteins (CSPs) (Beauchemin et al., 2012, Bayer et al., 2012). However, while these might be responsible for the transcriptional regulation (Bayer et al., 2012), it is not yet known if dinoflagellate CSPs perform a role in transcription or not. In 2016, two *L. polyedra* CSPs showed binding to both DNA and RNA, but tests to determine sequence specific binding were not performed (Beauchemin et al., 2016). Indeed, only the TLFs, the transcription factor replacing the TATA-box binding protein in *Crypthecodinium cohnii*, have been shown to specifically bind a DNA sequence identified as TTTT (Guillebault et al., 2002). Accordingly, one project aimed to assess the contribution of CSPs in transcriptional control by examining the role and function of CSDs in *Symbiodinium kawagutii* and *Lingulodinium polyedra.* The specific hypothesis tested in this project is that dinoflagellate CSPs do not have the expected properties of specific transcription factors, in particular sequence specific DNA binding activity.

1.2.1.1. Project 1 experimental approach

A total of 4 dinoflagellate CSPs were cloned and expressed in bacteria as a GST fusion protein. All were tested for nucleic acid binding affinity by EMSA to ensure they were active. These CSPs were tested for sequence-specific DNA binding activities using three cycles of a selection and amplification binding assay (SAAB) starting with a set of random oligonucleotides (Chang et al., 1997, Magnani et al., 2004). Enrichment of a given motif by the three cycles of binding to the CSPs would support the hypothesis that they can act as transcription factors.

1.2.2. Project 2

Expression of many *Symbiodinium* genes is reported to be influenced by light and the promoters of these genes can be monitored for potential regulatory elements using the available genome sequence. For example, Rubisco (*rbcL*) transcript levels in cultured *Symbiodinium* under a 12D:12 L cycle increased ~3 fold during the light phase of a 12:12 L:D cycle (Mayfield et al., 2014). Similarly, *AcpPC* gene transcripts increased about ~2-fold when the cells were exposed to high light (Xiang et al., 2015). The AcpPC protein sequence is highly homologous to the stress-related chlorophyll a/b binding proteins in *Chlamydomonas* which are up-regulated when the cells are exposed to high light (Peers et al., 2009). Transcripts encoding the cryptochrome CRY2 decreased in high light and showed a > 2-fold increase in the transcript level in the dark (Xiang et al., 2015). The RCC1 domain, highly represented in *Symbiodinium* and capable of binding to both nucleosomes and double-stranded DNA (Makde et al., 2010), is involved in regulation of chromosome condensation in the S phase of the cell cycle (Ohtsubo et al., 1987). In dinoflagellates, *RCC1* transcript levels are modified under different light conditions, with some exhibiting elevated levels (~2-fold) in the dark (Xiang et al., 2015). Lastly, abundance of the oxygen-evolving enhancer 1 (*OEE1*) gene of the photosystem II (PSII) complex in cultured *Symbiodinium* increased about 3-fold during the light period and decreased during the subjective dark (Sorek et al., 2016). The hypothesis tested in this project was that light regulated genes contain promoter elements that act as *cis* regulatory sequences for specific transcription factors in dinoflagellates.

1.2.2.1. Project 2 experimental approach

RNA sequencing was be performed to compare the transcriptome profiles under two different light conditions in order to identify potential light regulated genes. Northern blot analysis was used to measure the RNA abundance of selected light regulated genes over the course of a light dark cycle. Quantitative PCR analysis was additionally performed to assess levels of selected genes under different conditions.

CHAPTER 2

This chapter has been published in *BMC Molecular and Cell Biology*, 2021, 22: 27 under the title *"Assessing nucleic acid binding activity of four dinoflagellate cold shock domain proteins from Symbiodinium kawagutii and Lingulodinium polyedra"* by Bahareh Zaheri and David Morse.

Contributions

The design of the different experiments presented here was developed during discussions with David Morse. I performed and analysed the results of all the experiments in this chapter. I wrote the first draft of the manuscript which was then reviewed and corrected by David Morse.

Funding

Funding was obtained from the Canadian National science and Engineering Research Council (grant number 171382-03 to D.M.). The funding body played no role in the design of the study and collection, analysis, and interpretation of data, or in writing the manuscript and decision to publish

2.1. Abstract

Background: Dinoflagellates have a generally large number of genes but only a small percentage of these are annotated as transcription factors. Cold shock domain (CSD) containing proteins (CSPs) account for roughly 60% of these. CSDs are not prevalent in other eukaryotic lineages, perhaps suggesting a lineage-specific expansion of this type of transcription factors in dinoflagellates, but there is little experimental data to support a role for dinoflagellate CSPs as transcription factors. Here we evaluate the hypothesis that dinoflagellate CSPs can act as transcription factors by binding double-stranded DNA in a sequence dependent manner.

Results: We find that both electrophoretic mobility shift assay (EMSA) competition experiments and selection and amplification binding (SAAB) assays indicate binding is not sequence specific for four different CSPs from two dinoflagellate species. Competition experiments indicate all four CSPs bind to RNA better than double-stranded DNA.

Conclusion: Dinoflagellate CSPs do not share the nucleic acid binding properties expected for them to function as *bone fide* transcription factors. We conclude the transcription factor complement of dinoflagellates is even smaller than previously thought suggesting that dinoflagellates have a reduced dependance on transcriptional control compared to other eukaryotes.

Key words: Transcription factors, cold shock domain proteins, dinoflagellates, RNA binding domain, DNA binding domain, transcription

2.2. Background

Dinoflagellates are an important group of unicellular eukaryotes perhaps best known for their large genomes and permanently condensed chromosomes. Surprisingly, little is known how gene expression is regulated in these organisms. Transcriptome analyses in several species, including *Lingulodinium* and *Symbiodinium,* have revealed a general paucity (typically 0.15%) of sequences annotated as transcription factors (TF). This is in sharp contrast to the roughly 6% of genes annotated as TF in plants (Riechmann et al., 2000) or animals (Zhang et al., 2012). In addition, a high proportion (~60%) of the annotated dinoflagellate TF in transcriptomes are cold shock domain (CSD) containing proteins (CSPs) (Beauchemin et al., 2012, Bayer et al., 2012) yet this class is typically less than 1% of the TF in other eucaryotes. CSDs are small (roughly 70 amino acid) nucleic acid binding domains containing two conserved RNA recognition motifs**,** KGFGFI and VFVHF, that are known to bind both DNA and RNA. All dinoflagellate CSPs contain the two RNA binding motifs characteristic of the CSD. Four divergent domain structures have been found in *Lingulodinium* and *Symbiodinium* proteins, the most prevalent ones containing a CSD either alone or with a C-terminal G-rich domain. Less frequently observed are some structures containing a Zn-finger domain following the G-rich domain, and also examples of sequences with multiple CSDs and one or more RNA recognition motifs (RRM). Thus, many of the dinoflagellate CSPs are similar to what are found in bacteria as these typically contain only a CSD (Beauchemin et al., 2016).

In *E. coli*, CSPs have a wide range of functions, including binding DNA as transcription factors, binding to RNA, regulating transcription, splicing, and translation, and affecting mRNA stability as RNA chaperones (Mihailovich et al., 2010, Budkina et al., 2020). Bacterial CSPs have a non-specific RNA binding function during cold stress, which is correlated to their chaperone activity, and this helps transcription by acting as an antiterminator (Bae et al., 2000, Budkina et al., 2020). However, the dinoflagellate proteins may be different from their bacterial counterparts as two *Lingulodinium* CSPs, both containing a single CSD followed by a glycine-rich C-terminal region, were both unable to complement the growth of an *E. coli* strain lacking four different CSP genes at low temperature (Beauchemin et al., 2016). Furthermore, cold temperatures did not induce the CSP transcripts in *L. polyedra* (Roy et al., 2014c). Previous work on *L. polyedra* CSPs showed binding to both single- and double-stranded DNA as well as to RNA, but it was unclear if binding would show any sequence specificity that would be likely if they were to function as transcription factors (Beauchemin et al., 2016). Here we performed two experimental approaches to assess the specific nucleic acid binding activity of *L. polyedra* CSP1 (*Lp*CSP1*)* and three *S. kawagutii* CSPs *(Sk*CSP1, *Sk*CSP2 and *Sk*CSP3*)*. Initially, these four CSPs were expressed, purified and used in electrophoretic mobility assays (EMSAs) to measure if they were active in binding nucleic acids. In a second approach, selection and amplification binding assays (SAAB) was used to determine if these proteins could bind a specific sequence on DNA. All these CSPs were able to bind to DNA and RNA, and no sequence specific binding activity toward DNA was observed.

2.3. Results

***2.3.1. SkCSP1*, *SkCSP2* and *SkCSP3* belong to a *Symbiodinium* unique clade**. The number of annotated DNA binding proteins in the genome of the *S. kawagutii* (Li et al., 2020) belonging either to CSD family or other TF [Fig 2.1] shows the relative importance of CSDs in dinoflagellates compared to plants and animals. All CSDs contain the two RNA recognition motifs (KGFGFI and VFVHF) shared with bacteria and plants (Beauchemin et al., 2012, Bayer et al., 2012). Phylogenetic analysis of CSDs from 12 predicted *Symbiodinium kawagutii* protein sequences was performed using RaxML, and all were found to cluster together within a single well defined clade together with some bacterial sequences [Fig 2.2] [Table 2.1. Supplementary]. This is slightly different from the situation in *Lingulodinium* where sequences are distributed among two different clades. The phylogenetic positions of the four CSPs examined here: *Lp*CSP1 (JO732587), *Sk*CSP1 (Skav223430), *Sk*CSP2 (Skav207008) and *Sk*CSP3 (Skav233957) are boxed.

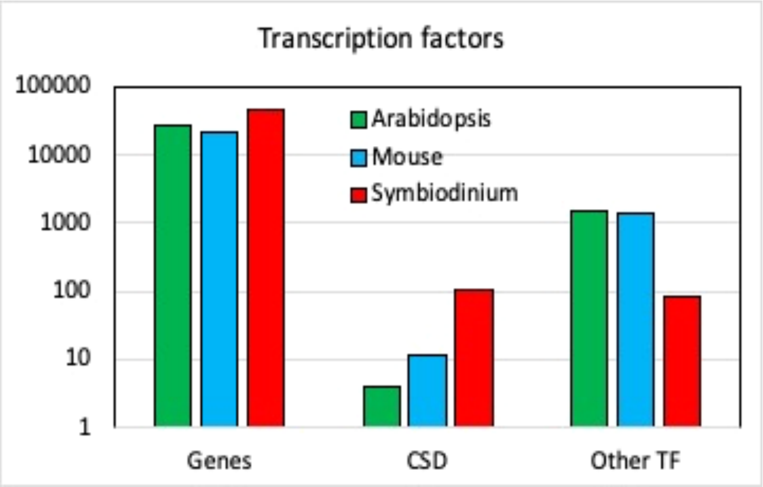


Figure 2.1. The abundance of DNA-binding domain families detected in *S. kawagutii* compared with plants and animals. The number of genes annotated as CSD and as other TF are shown for the most recent *S. kawagutii* genome. Note the log scale at left.

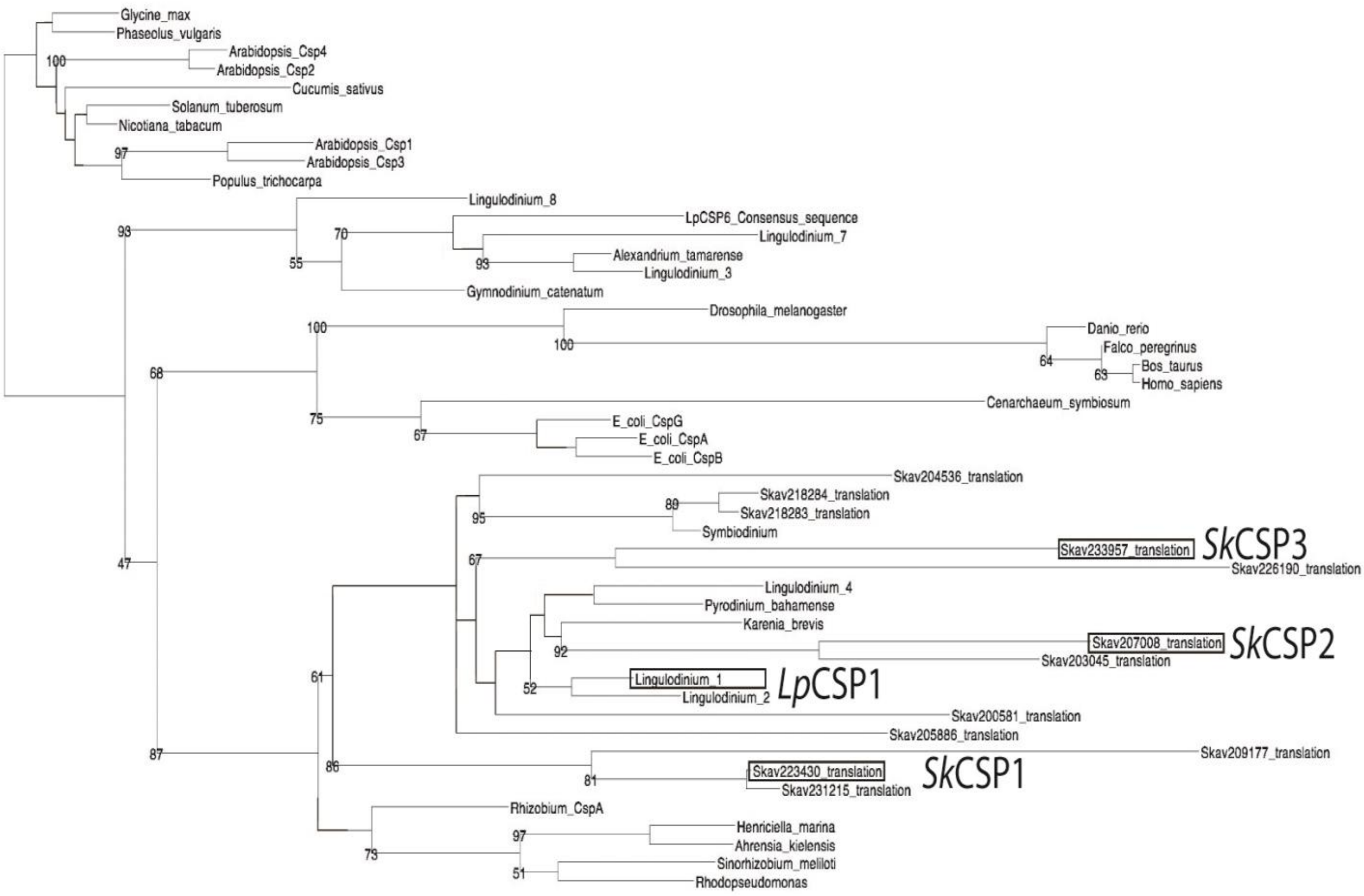
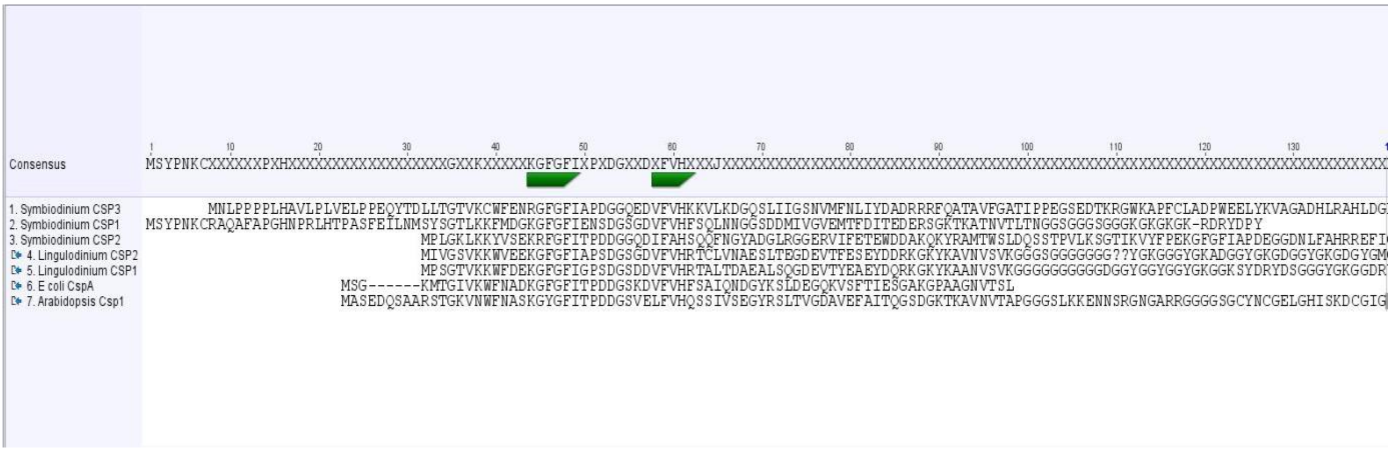


Figure 2.2. Phylogenetic reconstruction of a variety of dinoflagellate CSP. Sequences were aligned and the phylogeny reconstructed with RaxML. Numbers represent the bootstrap value in the tree (bootstrap values below 50 are not shown at the nodes). *Lp*CSP1*,* *Sk*CSP1, *Sk*CSP2 and *Sk*CSP3 sequences are boxed.

*Lp*CSP1 with a size of 113 amino acids has been previously cloned (Beauchemin et al., 2016). For this study, *Sk*CSP1, *Sk*CSP2 and *Sk*CSP3 were also cloned and have sizes of 128, 120 and 182 residues, respectively. All four CSPs were expressed as GST-tagged proteins and used for EMSA after removal of the GST tag [Fig 2.8. Supplementary]. Two of the *S. kawagutii* proteins contain an N-terminal extension [Fig 2.3].



```
Consensus                 MSYPNKCXXXXXXPXHXXXXXXXXXXXXXXXXXXGXXKXXXXXKGFGFIXPXDGXXDXFVHXXXJXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

1. Symbiodinium CSP3              MNLPPPPLHAVLPLVELPPEQYTDLLTGTVKCWFENRGFGFIAPDGGQEDVFVHKKVLKDGQSLIIGSNVMFNLIYDADRRRFQATAVFGATIPPEGSEDTKRGWKAPFCLADPWEELYKVAGADHLRAHLDG
2. Symbiodinium CSP1       MSYPNKCRAQAFAPGHNPRLHTPASFEILNMSYSGTLKKFMDGKGFGFIENSDGSGDVFVHFSQLNNGGSDDMIVGVEMTFDITEDERSGKTKATNVTLTNGGSGGGSGGGKGKGKGK-RDRYDPY
3. Symbiodinium CSP2                                      MPLGKLKKYVSEKRFGFITPDDGGQDIFAHSQQFNGYADGLRGGERVIFETEWDDAKQKYRAMTWSLDQSSTPVLKSGTIKVYFPEKGFGFIAPDEGGDNLFAHRREFI
4. Lingulodinium CSP2                                     MIVGSVKKWVEEKGFGFIAPSDGSGDVFVHRTCLVNAESLTEGDEVTFESEYDDRKGKYKAVNVSVKGGGSGGGGGGG??YGKGGGYGKADGGYGKGDGGYGKGDGYGM
5. Lingulodinium CSP1                                     MPSGTVKKWFDEKGFGFIGPSDGSDDVFVHRTALTDAEALSQGDEVTYEAEYDQRKGKYKAANVSVKGGGGGGGGGGDGGYGGYGGYGKGGKSYDRYDSGGGYGKGGDR
6. E coli CspA                                   MSG------KMTGIVKWFNADKGFGFITPDDGSKDVFVHFSAIQNDGYKSLDEGQKVSFTIESGAKGPAAGNVTSL
7. Arabidopsis Csp1                              MASEDQSAARSTGKVNWFNASKGYGFITPDDGSVELFVHQSSIVSEGYRSLTVGDAVEFAITQGSDGKTKAVNVTAPGGGSLKKENNSRGNGARRGGGGSGCYNCGELGHISKDCGIG
```

Figure 2.3. Alignment of CSD domains. CSD from the dinoflagellates *L. polyedra*, and *S. kawagutii*, the bacterium *E. coli* and the higher plant *Arabidopsis thaliana*. The two RNA recognition motifs are marked in green.

***2.3.2. Lingulodinium* and *Symbiodinium* CSPs bind to DNA and RNA**. EMSA experiments were conducted on *Lp*CSP1, *Sk*CSP1, *Sk*CSP2 and *Sk*CSP3 to analyze their binding to radiolabelled double-stranded (dsDNA), single-stranded (ssDNA) and RNA probes [Fig 2.4]. Fusion proteins still containing the glutathione *S*-transferase (GST) tags also bind nucleic acids but migrating slower on the gel, and all EMSA experiments used proteins after removal of the tag by thrombin.

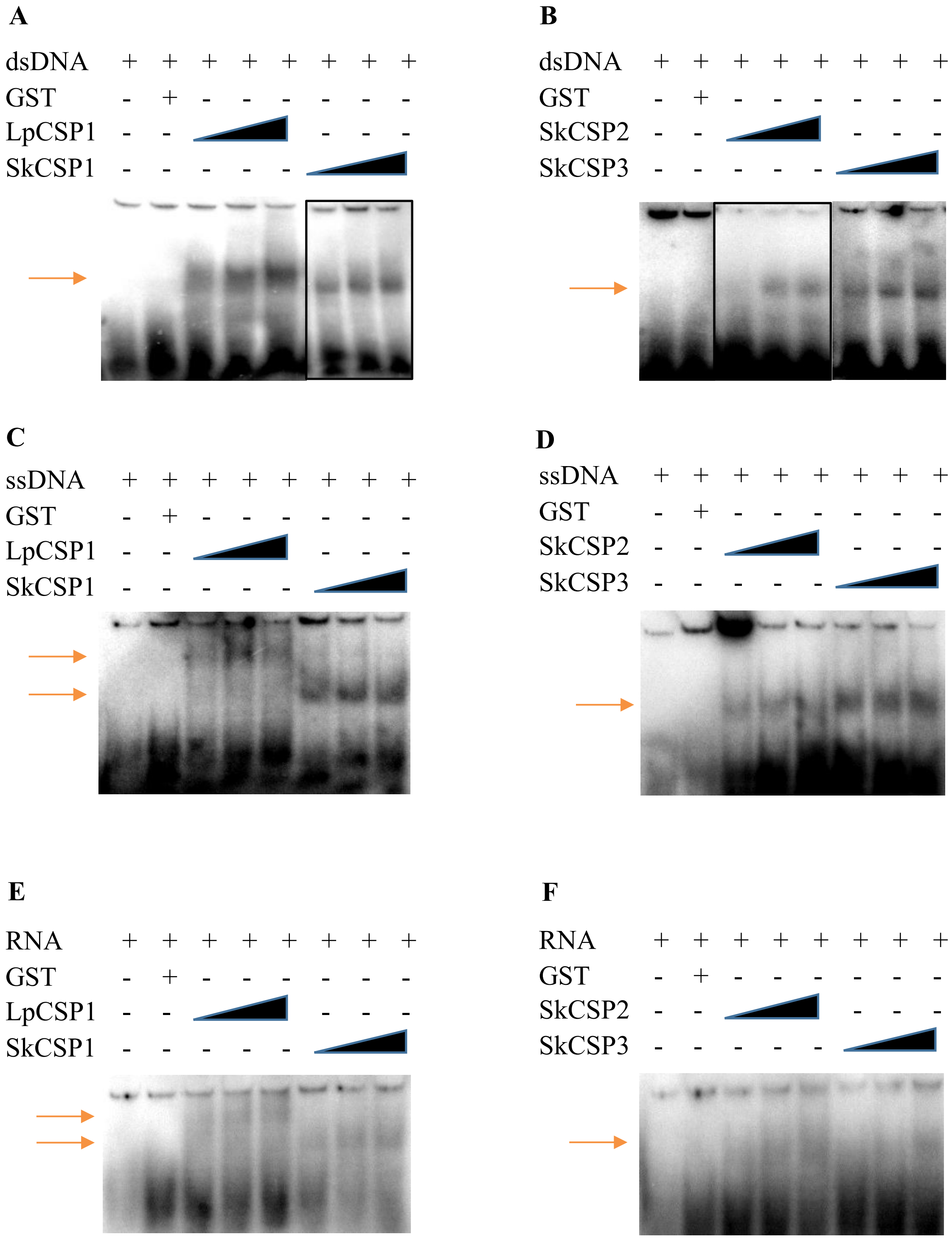


Figure 2.4. Nucleic acid binding activity of *L. polyedra* and *S. kawagutii* CSPs in EMSA. dsDNA (A, B), ssDNA (C, D) and RNA (E, F) probes were used. The black triangle shows the different concentrations of the CSPs (0.5, 1 and 3 μg in all the assays); position of the shifts are shown by arrows.

All proteins were able to bind dsDNA, ssDNA and RNA as seen by the presence of a radioactive band of lower mobility. The mobility of probe sequence was reduced to roughly the same extent with all proteins with the exception of *Lp*CSP1 binding to ssDNA or RNA. The amount of the reduced mobility band seemed to increase with increasing concentrations of the CSPs, although not precisely proportional to the amount of protein. We conclude that all four CSPs were able to bind to all three types of nucleic acids tested.

***2.3.3. Symbiodinium* CSPs prefer binding to single-stranded nucleic acids**. To assess the specificity of *Symbiodinium* CSPs interactions with different nucleic acid substrates, binding to dsDNA and ssDNA probes was evaluated using *Sk*CSP1 and unlabeled (cold) competitors [Fig 2.5]. When dsDNA was used as a probe, the intensity of the slowly migrating bands decreased dramatically when the amount of competing cold ssDNA was increased. In contrast, band intensity using ssDNA probes was mostly stable using increasing amounts of cold dsDNA. Furthermore, RNA appears to compete efficiently with both dsDNA and ssDNA. These results indicate that *Sk*CSP1 has a preference for single-stranded nucleic acids, with RNA preferred over DNA. This is consistent with a previous report for *Lingulodinium* CSP1 (Beauchemin et al., 2016). While the potential tendency to bind to ssDNA may support a role for these proteins in uncoiling the DNA structure during transcription, preferential binding to RNA suggests this may not be their primary role.

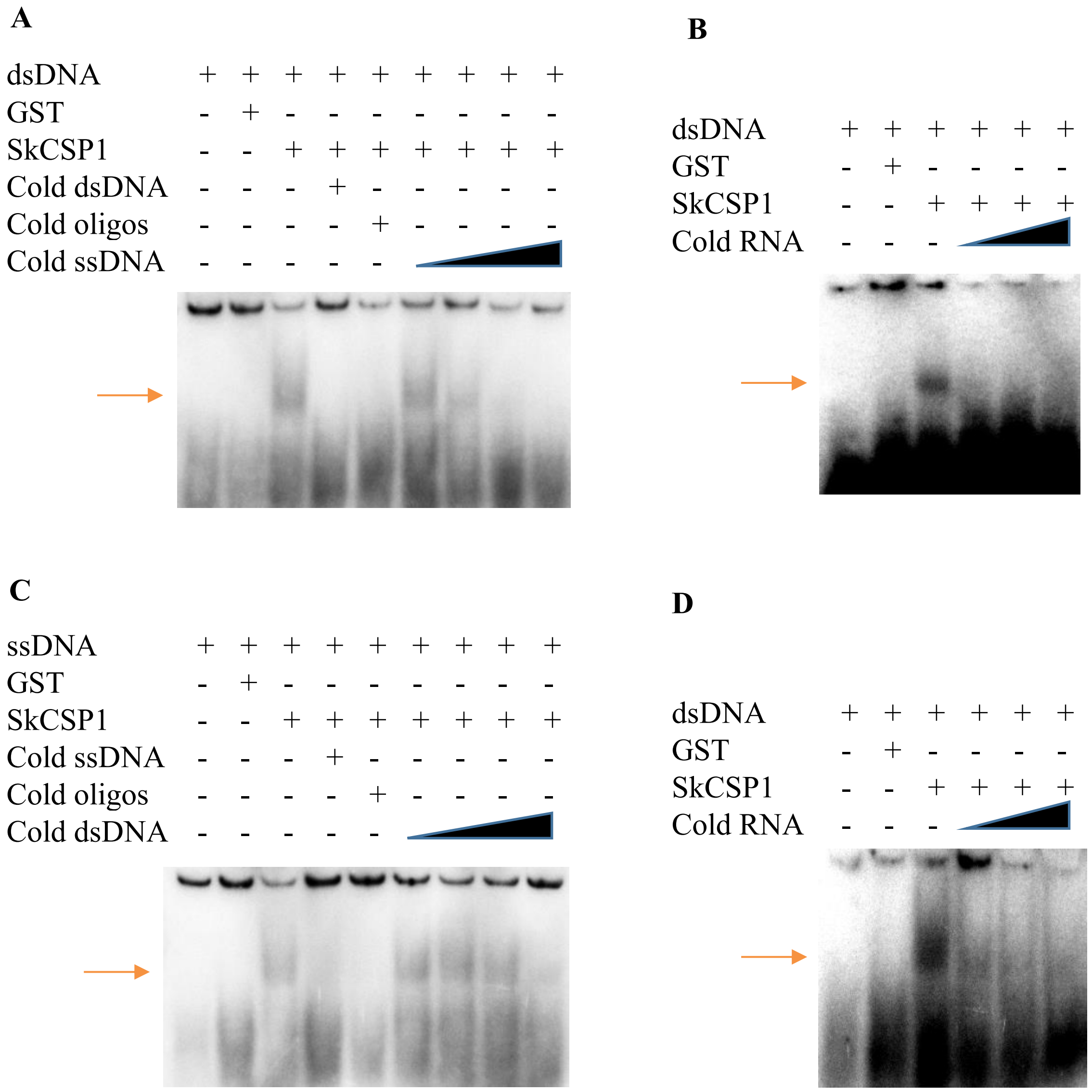


Figure 2.5. Competition assays of *Sk*CSP1 with ssDNA, dsDNA and RNA. *Symbiodinium* CSP1 binds to ssDNA better than dsDNA (A, C). Cold oligos have a different sequence than the ssDNA. Concentration of *Sk*CSP1 is 0.5 μg in all the assays; positions of the shifts are shown by arrows. RNA competes efficiently with both dsDNA and ssDNA (B, D). Cold DNA and Cold oligos concentrations are 40x the probe concentration. The black triangle shows the different concentrations of the unlabeled DNA (1x, 10x, 30x, and 80x) and the unlabeled RNA (1x, 30x and 80x).

***2.3.4. L. polyedra* and *S. kawagutii* CSPs bind non-specifically to DNA**. To assess the possibility of sequence specific binding of *Lingulodinium* and *Symbiodinium* CSPs to dsDNA, we performed a selection and amplification binding enrichment (SAAB) with DNA containing 9 random nucleotides (N9) flanked by PCR primers. These experiments used the fusion proteins directly to facilitate purification of bound DNA sequences, as the presence of the GST tag did not affect DNA binding on EMSA assays. After 3 rounds of SAAB, samples containing double-stranded N9 enriched by binding to *Lp*CSP1, *Sk*CSP1, *Sk*CSP2 and *Sk*CSP3 were sequenced [Fig 2.6]. Over 12,000 sequences were been obtained for each CSP, but sequence alignments after binding to all four shows no evidence for a consensus motif for any of the CSPs [Fig 2.7]. We conclude that there is no specific dsDNA which can be enriched by binding to *Lingulodinium* or *Symbiodinium* CSPs.

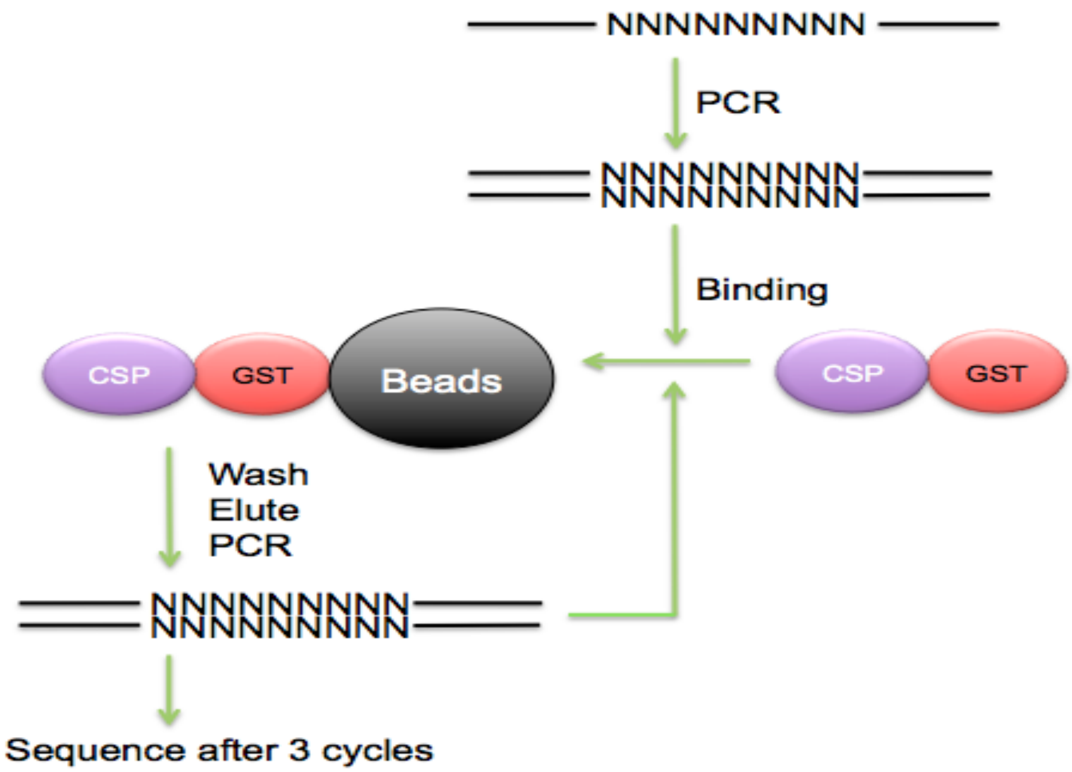


Figure 2.6. SAAB Schematic model. The protocol for analyzing the specificity of DNA sequence binding by selection and amplification binding assays (SAAB) involves repeated cycles of DNA binding to an immobilized protein followed by elution and amplification by PCR.

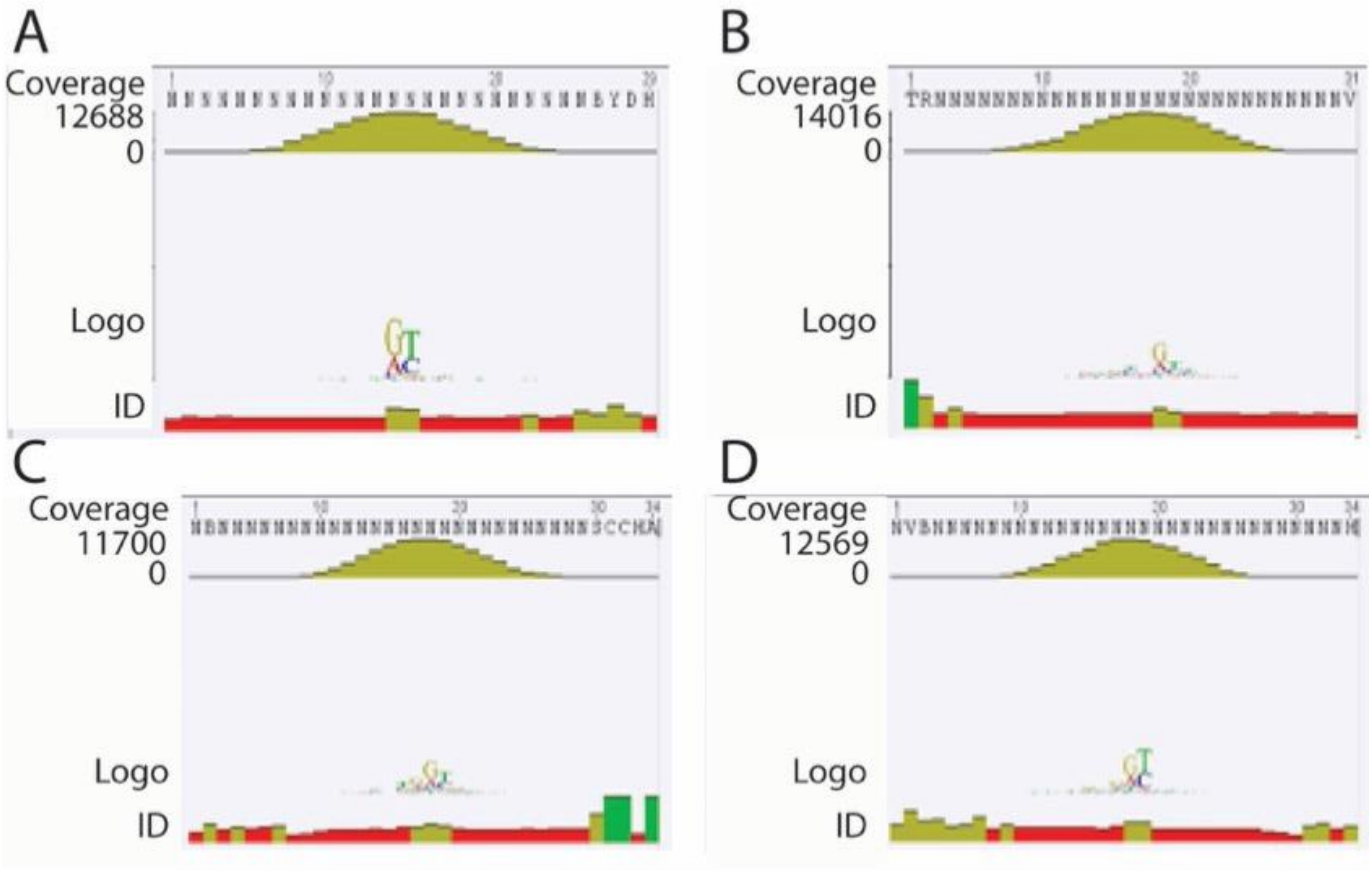


Figure 2.7. Consensus nucleotide binding activity of 4 different dinoflagellates CSPs. Over 12,000 different N9 sequences bound by *Lp*CSP1 (A), *Sk*CSP1 (B), *Sk*CSP2 (C), and *Sk*CSP3 (D) were aligned and used to prepare a sequence logo showing the frequency of each nucleotide at each position.

2.4. Discussion

Cold shock domain (CSD) proteins were recognized in *Escherichia coli* during cold shock stress (Jones et al., 1987, Budkina et al., 2020, Heinemann and Roske, 2021). The conservation of CSD in these proteins was discovered in bacteria, archaea, plants, and animals. In prokaryotes, CSPs containing only a CSD act mainly as RNA chaperones. *E. Coli* CSPs are cold inducible and act as RNA chaperons disrupting RNA secondary structures (Graumann and Marahiel, 1998, Budkina et al., 2020). They are also involved in the transcription regulation by binding specifically to *gyrA* promoter (CspA) (Jones et al., 1992, Heinemann and Roske, 2021). In eukaryotes, CSPs are composed of CSD and additional domains and aid in responding to cold stress, nutrient limitation and growth (Karlson and Imai, 2003, Nakaminami et al., 2006, Wistow, 1990, Graumann and Marahiel, 1998, Budkina et al., 2020). Plants CSPs are engaged in regulation of translation during cold stress and also complicated physiological processes such as seed and flower germination (Fusaro et al., 2007, Kim et al., 2013, Budkina et al., 2020). In *A. thaliana*, CSP3 interacts with other proteins involved in mRNA processing path (Kim et al., 2013). A vertebrate CSP called YB1 (Y-box binding protein) is responsible for the regulation of transcription by binding to the Y-box specific sequence, regulation of translation and RNA processing (Izumi et al., 2001, Lasham et al., 2003, Sommerville, 1999, Kleene, 2018, Mordovkina et al., 2020) and DNA repair (Budkina et al., 2020, Heinemann and Roske, 2021, Sangermano et al., 2020). YB1 prefers to bind to ssDNA rather than dsDNA, thus disentangling the double helix structure of DNA has been proposed for the activation of transcription (MacDonald et al., 1995, Budkina et al., 2020). YB1 also prefers RNA over ssDNA (Budkina et al., 2020) with the consensus CA(U/C)C sequence as the RNA-binding site (Wei et al., 2012, Yang et al., 2019). In dinoflagellates, CSPs contain the conserved CSD, mostly in the form of one CSD either alone or with a C-terminal G-rich domain (Beauchemin et al., 2016). Previously, a Y-box sequence (CTGATTGGCT) was used to study the binding specify of *L. polyedra* CSPs (Beauchemin et al., 2016), here we used different random C-rich sequences to test the possibility of sequence privileged targeting. For the SAAB assay, we synthesized a DNA sequence with 9 random nucleotides (N9) nestled between flanking PCR primers. The goal of this experiment was to see if several cycles of binding, elution and amplification would enrich for a particular sequence motif that could constitute a potential promoter element. However, no sequence motifs were enriched by binding to any of the four CSPs indicating that these proteins are unlikely to function as conventional sequence-specific transcription factors. It is not possible to rule out a role in DNA unwinding similar to what has been proposed for YB1, in which non-specific binding of CSPs to ssDNA was thought to help stabilize the structure, but it must be noted CSPs have no known helicase activity.

The importance of examining the nucleic acid binding properties of CSPs is due to the finding that the majority of the proteins annotated as transcription factors in the transcriptome of *Lingulodinium* (Beauchemin et al., 2012), *Symbiodinium* (Bayer et al., 2012) and the genome of *Symbiodinium* (Li et al., 2020, Yu et al., 2020) [Fig 2.1] are CSDs. Our hypothesis was that to act as transcription factors, dinoflagellates CSPs should bind to dsDNA in a sequence specific manner. We assessed nucleic acid binding activity of *Lp*CSP1, *Sk*CSP1, *Sk*CSP2 and *Sk*CSP3 using two different approaches. In one approach, electrophoretic mobility shifts assays (EMSA) were used to show that all four CSPs could bind both double- and single-stranded DNA as well as RNA [Figure 2.4]. When tested in competition EMSA experiments, RNA was found to compete with binding to DNA probes better than DNA competed with binding to DNA probes [Figure 2.5]. These characteristics are not what would be predicted for a transcription factor. In a second approach, selection and amplification binding (SAAB) experiments showed none of the four CSPs tested enriched a specific motif after three cycles of binding and PCR amplification, again inconsistent with a role as a sequence specific transcription factor.

Our results indicate that *Lp*CSP1, *Sk*CSP1, *Sk*CSP2 and *Sk*CSP3 binding to nucleic acids does not depend on sequence. We infer that the dinoflagellate CSPs in general are unlikely to act as sequence-specific transcription factors. Although only one *S. kawagutii* CSP (*Sk*CSP1) was extensively analyzed by competition EMSA, the similarity to the *Lingulodinium* CSP1 suggests the nucleic acid binding properties found may be a consistent lineage-specific feature. The balance of the evidence thus suggests that CSPs do bind nucleic acids, thus explaining why they were annotated as transcription factors. However, the details of the binding suggest they are unlikely to play this role *in vivo.* Additional characterization studies of dinoflagellate CSPs would be essential to recognize more about their function and possible interaction with other partners.

2.5. Conclusions

The four CSPs examined here do not bind DNA in a sequence specific manner. Furthermore, SkCSP1 prefers binding to single-stranded RNA. CSPs are unlikely to function as transcription factors in dinoflagellates.

2.6. Methods

2.6.1. **Cell cultures**. Cultures of *Symbiodinium kawagutii* (strain CCMP2468) and *Lingulodinium polyedra* (strain CCMP1936) were obtained from the National Center for Marine Algae (Boothbay Harbor, Maine). Cells were grown in f/2 sea water medium prepared from Instant Ocean under 12 h cool white fluorescent light and 12 h darkness as described (Wang et al., 2005) except that the temperature was 25 ± 1 °C for *S. kawagutii*.

2.6.2. **Phylogenetic reconstruction and primer design.** The CSP sequences for *Lingulodinium* and *Symbiodinium* were obtained from the dinoflagellate transcriptomes deposited at NCBI and from the *Symbiodinium kawagutii* genome at the Symbiodiniaceae and Algal Genomic Resource (SAGER) database (Yu et al., 2020). Phylogenetic analysis of CSDs from the predicted protein sequences [Table 2.1. Supplementary] was performed using a webserver for alignments (http://www.phylo.org/sub_sections/portal/) (Dereeper et al., 2008). The server performs sequence alignments using MUSCLE, and curation using GBlocks. Phylogenetic reconstructions were built with RaxML using the CIPRES portal (http://www.phylo.org/sub_sections/portal/). Trees were visualized by TreeDyn. Primers were designed using Geneious software (Kearse et al., 2012) or BLAST integrated into Galaxy (Cock et al., 2015) for amplification and subsequent cloning of the CSPs. Geneious software (Kearse et al., 2012) was also used for sequence alignments.

2.6.3. **Cloning, expression and purifying of CSPs**. Harvested *Symbiodinium* cultures were pelleted and then snap frozen in liquid nitrogen. Frozen pellets were crushed into a fine powder using a pre-chilled mortar and pestle, and the powder was added to Trizol (Invitrogen). Primer pairs based on sequences from the *Symbiodinium* transcriptome or genome were used to amplify CSPs from a first strand cDNA reaction product using the total RNA extracted from *Symbiodinium* cells as described (Beauchemin et al., 2016). For cDNA amplification, the reverse transcription reaction was performed with ProtoScript II first strand cDNA synthesis kit (New England BioLabs). The sequences were cloned into the pGEM-T vector (Promega) and sequenced. A second PCR was performed on the insert in the pGEM-T plasmid using primers containing restriction sites required for directional cloning into the bacterial expression vectors pGEX-4T2 (GE Healthcare) (Xia et al., 2001) [Table 2.2. Supplementary]. The reading frame of all clones were confirmed by sequencing and the size of the CSP fusion protein verified by SDS PAGE [Fig 2.8. Supplementary]. The pGEX4T2 vectors containing CSP sequences were used to transform the chemically competent cells of BL21. Liquid Luria Bertani (LB) medium was used to grow one colony of transformed *E. coli* overnight at 37°C with vigorous shaking in the presence of ampicillin to maintain selection for the plasmid. Protein expression were induced using Isopropyl β-D-1-thiogalactopyranoside (IPTG). Cells were collected by centrifugation, resuspended in PBS buffer and broken in a French pressure cell (Fisher Scientific). The cell lysates were then centrifuged and the supernatants were incubated with Glutathione Sepharose 4B beads (Promega) for 45 min at room temperature with end-over-end agitation. Beads were washed 4 times in PBS and resuspended in PBS supplemented with thrombin to cleave the GST tag. The size, and purity of the single CSPs were then analyzed by SDS-PAGE on acrylamide gel [Fig 2.8. Supplementary] and the Bradford assay (BioRad) was used to assess the protein concentration.

***2.6.4.* CSP electrophoretic mobility shift assays**. [γ-^{32}P]ATP (PerkinElmer) was used to 5′-end-label 32 nt ssDNA 5′-TCCGCCCTCCCTCCCCCCGCCCTCCCTCCCCA-3′ and 25 bp dsDNA 5′-GGCGCCCTCCCTCCGCCCTCCCTCA-3′ C-rich sequences using a T4 polynucleotide kinase kit (NEB). A QIAquick nucleotide removal kit (Qiagen) was used for removing the unincorporated nucleotides and purifying the probes. Either dsDNA or ssDNA 32P-labelled probes (1 ng) and increasing concentrations of CSPs (0.5–3 μg) were incubated in 20 μL of 2x binding buffer (20 mM Tris-Cl [pH 7.0], 20 mM MgCl2, 50 mM KCl, 10% glycerol and 1 mM DTT) for 30 min at room temperature. The CSP/DNA complexes were run through a 5% native polyacrylamide gels for 45 mins at 80 V in 1× Tris-borate-EDTA (TBE) buffer at room temperature. The gels were dried immediately and exposed overnight at -80 °C with a phosphorimager screen (Amersham). The images were analyzed with a Typhoon Trio+ (Amersham) using ImageQuant 5.2. Competition reactions were prepared by incubation of the CSPs and increasing amounts of cold unlabeled ssDNA, dsDNA or RNA probes (described below) for specific binding and a 40x excess of random 22 nt single-stranded oligonucleotide (TTATTGGGGCACACCGCATGCT) for non-specific competition in the binding buffer for 15 mins before adding the radiolabeled probes.

40 nt RNAs were synthesized by T7 RiboMAX RNA production kit (Promega) using dsDNA templates containing the N9 and T7 promoter sequences. Thereafter, RQI RNase-free DNase (Promega) was used for degradation of the dsDNA templates. The *in vitro* transcribed RNAs were quantitated using spectrophotometry (1.2 μg/μL), end-labeled using [γ-^{32}P]ATP (PerkinElmer) (see

above) and purified using filtration chromatography on a Bio-Gel P10 column (Bio-Rad). 1 ng labelled probe was incubated with increasing concentrations of CSPs in the binding reactions as described above.

***2.6.5.* Selection and amplification binding assays**. *Symbiodinium* and *Lingulodinium* CSPs were cloned and expressed as a fusion protein with a C-terminal GST tag as described above. The BL21 cell lysates were centrifuged and the supernatants containing GST tagged CSPs were incubated with Glutathione Sepharose 4B beads (Promega) for 45 min at room temperature with end-over-end agitation. Beads were washed 4 times in PBS. Immobilized *Lp*CSP1, *Sk*CSP1, *Sk*CSP2 and *Sk*CSP3 were tested for sequence-specific DNA binding activities against a set of degenerate oligonucleotides using a selection and amplification binding assay (SAAB) (Chang et al., 1997, Magnani et al., 2004). A set of single-stranded oligonucleotides with PCR primer sequences flanking nine random nucleotides (N9) were synthesized and used to produce double-stranded DNA by a single PCR cycle using the reverse primer. 15 μg of double-stranded DNA (N9) was allowed to bind to 10 μL of immobilized CSPs in a 100 μL total volume solution containing 75 mM NaCl, 1 mM DTT, 1 mM phenylmethylsulfonyl fluoride, 0.1% Triton X-100, 10 ng of poly(dI-dC) per μL, 10mM Tris-HCl (pH 7), 6% glycerol and 1% BSA. After 1 hour of agitation at 4 °C, the supernatant containing unbound oligonucleotides were removed. Following 3 times of washing with binding buffer, DNA was released from the protein by boiling in water (Chang et al., 1997). DNA was amplified in a PCR reaction to repeat the protein binding step. Three rounds of SAAB were performed before sending out the PCR products for sequencing [Fig 2.6].

Figure 2.8 (Supplementary). Purification of *Lp*CSP1, *Sk*CSP1, *Sk*CSP2 and *Sk*CSP3

A shows recombinant *Lp*CSP1-GST, *Sk*CSP1-GST, *Sk*CSP2-GST and *Sk*CSP3-GST analyzed on an 18% acrylamide SDS-PAGE gel after affinity purification. **B** shows *Lp*CSP1, *Sk*CSP1, *Sk*CSP2 and *Sk*CSP3 after removal of the GST tag by thrombin digestion and binding to glutathione-Sepharose 4B beads. The sizes of the proteins are shown in kilodaltons.

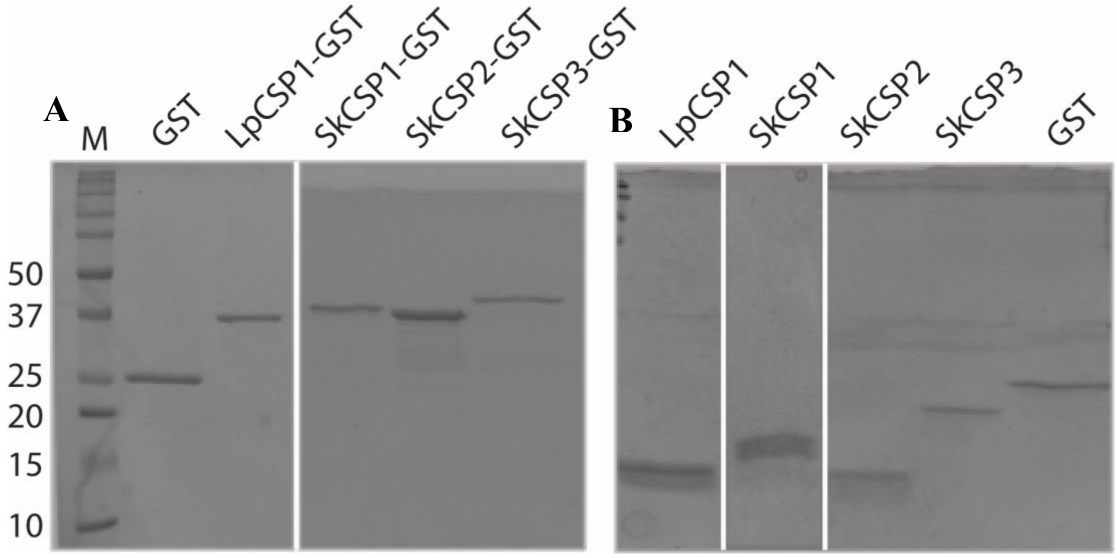


Table 2.1 (Supplementary). List of proteins selected for phylogenetic reconstruction

Alexandrium tamarense:	*Alexandrium tamarense* 1743629 transcribed RNA sequence GAIT01073544
Ahrensia kielensis:	Cold-shock protein [*Ahrensia kielensis*] gi|517517514|ref|WP_018687722.1|
***Arabidopsis* Csp1:**	Cold shock protein 1 [*Arabidopsis thaliana*] gi|332661203|gb|AEE86603.1|
***Arabidopsis* Csp2:**	Glycine rich protein 2 [*Arabidopsis thaliana*] NP_195580
***Arabidopsis* Csp3:**	Cold shock domain protein 3 [*Arabidopsis thaliana*] NP_565427
***Arabidopsis* Csp4:**	Full=Cold shock domain-containing protein 4; Short=AtCSP4 Q38896
Bos Taurus:	TPA: Lin-28 homolog B-like [*Bos taurus*] gi|296484122|tpg|DAA26237.1|
Cenarchaeum symbiosum:	Cold-shock protein [*Cenarchaeum symbiosum* A] ABK77130
Cucumis sativus:	PREDICTED: Cold shock domain-containing protein 4-like [*Cucumis sativus*] gi|449445142|ref|XP_004140332.1|
Danio rerio:	PREDICTED: protein Lin-28 homolog A-like [*Danio rerio*] gi|528503039|ref|XP_001340141.2|
Drosophila melanogaster:	Lin-28 [*Drosophila melanogaster*] NP_647983
***E. coli* CspA:**	Cold shock protein CspA [*Escherichia coli* CFT073] AAN82813
***E. coli* CspB:**	CspB [*Escherichia coli*] str. K-12 AAB61739
***E. coli* CspG:**	Cold shock protein CspG NP_309172
Falco peregrinus:	PREDICTED: protein Lin-28 homolog B isoform X1 [*Falco peregrinus*] gi|529448821|ref|XP_005244100.1|
Glycine max:	PREDICTED: glycine-rich protein 2-like [*Glycine max*] XP_003540832
Gymnodinium catenatum:	Gymnodinium catenatum GcatSW0_c188 transcribed RNA sequence GAIL01018775
***Henriciella marina*:**	Cold-shock protein [*Henriciella marina*] gi|516884417|ref|WP_018146825.1|
Homo sapiens:	Lin-28 homolog B (*C. elegans*), isoform CRA_a [*Homo sapiens*] gi|119568818|gb|EAW48433.1|
***Karenia brevis*:**	K05492D08 *Karenia brevis* Multi-strain Library *Karenia brevis* cDNA 5', mRNA sequence gi|194490792|gb|FK848095.1|FK848095
***Lingulodinium* 1:**	JO733348
***Lingulodinium* 2:**	JO734870
***Lingulodinium* 3:**	JO730956
***Lingulodinium* 4:**	JO729000
***Lingulodinium* 7:**	JO766444
***Lingulodinium* 8:**	JO761018
LpCSP6 Consensus Sequence	_
Nicotiana tabacum:	Full=Glycine-rich protein 2 P27484

Phaseolus vulgaris:	Hypothetical protein PHAVU_009G025100g [*Phaseolus vulgaris*] gi|561009269|gb|ESW08176.1|
Populus trichocarpa:	Hypothetical protein POPTR_0009s13460g [*Populus trichocarpa*] gi|566187811|ref|XP_002313723.2|
Pyrodinium bahamense:	TSA: *Pyrodinium bahamense* var. compressum F4W4PV301CKO5W transcribed RNA sequence gi|509887131|gb|GAIO01020278.1|
***Rhizobium* CspA:**	Cold shock protein CspA [*Rhizobium leguminosarum* bv. viciae 3841] gi|116254513|ref|YP_770349.1|
Rhodopseudomonas:	Cold shock DNA binding protein [*Rhodopseudomonas palustris* CGA009] gi|39936462|ref|NP_948738.1|
Sinorhizobium meliloti:	CspA [*Sinorhizobium meliloti*] AAC64672
Skav200581 translation:	Skav200581 [mRNA] locus=scaffold1051:18458:24228:-
Skav203045 translation:	Skav203045 [mRNA] locus=scaffold845:65342:72188:+
Skav204536 translation:	Skav204536 [mRNA] locus=scaffold1211:212853:213298:-
Skav205886 translation:	Skav205886 [mRNA] locus=scaffold123:16538:18973:+
Skav207008 translation:	Skav207008 [mRNA] locus=scaffold1554:61201:64400:-
Skav209177 translation:	Skav209177 [mRNA] locus=scaffold1137:469928:472431:-
Skav218283 translation:	Skav218283 [mRNA] locus=scaffold2035:589603:594139:-
Skav218284 translation:	Skav218284 [mRNA] locus=scaffold2035:596104:599782:-
Skav220717 translation:	Skav220717 [mRNA] locus=scaffold1850:102915:105042:-
Skav223430 translation:	Skav223430 [mRNA] locus=scaffold350:502771:503148:+
Skav224338 translation:	Skav224338 [mRNA] locus=scaffold1353:319050:321488:-
Skav226190 translation:	Skav226190 [mRNA] locus=scaffold2212:105531:115729:+
Skav228973 translation:	Skav228973 [mRNA] locus=scaffold671:194838:200215:+
Skav231215 translation:	Skav231215 [mRNA] locus=scaffold2958:225978:226268:-
Skav233957 translation:	Skav233957 [mRNA] locus=scaffold1382:273360:273902:+
Skav234280 translation:	Skav234280 [mRNA] locus=C9163801:2089:2301:+
Solanum tuberosum:	PREDICTED: glycine-rich protein 2-like [*Solanum tuberosum*] gi|565387789|ref|XP_006359670.1|
Symbiodinium:	TSA: *Symbiodinium* sp. clade D d_sym_30274 mRNA sequence gi|452175549|gb|GAFP01006036.1|

Table 2.2 (Supplementary). List of primers used for PCR amplification and cloning of *LpCSP1*, *SkCSP1*, *SkCSP2* and *SkCSP3* sequences in pGEX4T2 plasmid

Proteins	Accession number	Primers name	Primers sequence (5' to 3')
LpCSP1	JO733348	LpCSP1 F	GCAGCAATGCCTTCCGGCACTGTGAAGAAG
		LpCSP1 F BamHI	TGACACGGATCCATGCCTTCCGGCACTGTGAAGAAG
		LpCSP1 F NdeI	TGACCATATGCCTTCCGGCACTGTGAAGAAG
		LpCSP1 R	ACCCTCAGCTCAGAAACCTGAGGAGGGTCC
SkCSP1	Skav223430	SkCSP1 F	ATGTCATATCCGAACAAATGTCGGG
		SkCSP1 R XhoI	ATTCACTCGAGTCAGTAGGGGTCATAGCGATCAC
		SkCSP1 F SmaI	TTGATCCCGGGATGTCATATCCGAACAAATGTCGGG
		SkCSP1 R	TCAGTAGGGGTCATAGCGATCAC
SkSCP2	Skav207008	SkCSP2 F	ATGCCACTGGGGAAATTGAAAAA
		SkCSP2 R XhoI	ATTCACTCGAGTCAGTCCTTGCACCAATCGG
		SkCSP2 F SmaI	TTGATCCCGGGATGCCACTGGGGAAATTGAAAAA
		SkCSP2 R	TCAGTCCTTGCACCAATCGG
SkCSP3	Skav233957	SkCSP3 F	ATGAATCTTCCTCCGCCTCC
		SkCSP3 R XhoI	ATTCACTCGAGTCAAATCATTGAGTCCTCGAAGAA
		SkCSP3 F SmaI	TTGATCCCGGGATGAATCTTCCTCCGCCTCC
		SkCSP3 R	TCAAATCATTGAGTCCTCGAAGAA

CHAPTER 3

This chapter has been published in *Microorganisms*, 2019, 7: 261 under the title *"Assessing transcriptional responses to light by the dinoflagellate Symbiodinium"* by Bahareh Zaheri, Steve Dagenais-Bellefeuille, Bo Song and David Morse.

Contributions

I contributed to design of the experiments and writing the initial drafts of the manuscript. I performed the RNA extractions, PCR and quantitative PCR as well as statistical analysis of gene expression. I performed the northern blot experiments in collaboration with S. Dagenais-Bellefeuille. B. Song performed bioinformatic analyses, notably read counts from RNA-Seq experiments. D. Morse performed the microscopy and corrected the manuscript

3.1. Abstract

The control of transcription is poorly understood in dinoflagellates, a group of protists whose permanently condensed chromosomes are formed without histones. Furthermore, while transcriptomes contain a number of proteins annotated as transcription factors, the majority of these are cold shock domain proteins which are also known to bind RNA, meaning the number of true transcription factors is unknown. Here we have assessed the transcriptional response to light in the photosynthetic species *Symbiodinium kawagutii*. We find that three genes previously reported to respond to light using qPCR do not show differential expression using northern blots or RNA-Seq. Interestingly, global transcript profiling by RNA-Seq at LD 0 (dawn) and LD 12 (dusk) found only seven light-regulated genes (FDR = 0.1). qPCR using three randomly selected genes out of the seven was only able to validate differential expression of two. We conclude that there is likely to be less light regulation of gene expression in dinoflagellates than previously thought and suggest that transcriptional responses to other stimuli should also be more thoroughly evaluated in this class of organisms.

Keywords: dinoflagellate; transcriptional control; light regulation

3.2. Introduction

Dinoflagellates are protists with an unusual chromatin structure (Spector, 1984). The dinoflagellate chromosomes are permanently condensed, and can be observed with light microscopy using fluorescent DNA stains such as DAPI or propidinium iodide (Roy and Morse, 2013). When observed using the electron microscope, individual chromosomes display a characteristic whorled banding pattern reminiscent of the bacterial nucleoid (Soyer and Haapala, 1974), and nucleosomes have never been observed (Bodansky et al., 1979). The unusual chromatin structure has a number of molecular correlates. The histone proteins are at very low levels (Roy and Morse, 2012), and while one or two histones have been detected in several species (Gornik et al., 2012, Beauchemin and Morse, 2018), all four core histones have not yet been detected in any species. Instead of histones, dinoflagellates are thought to compact their DNA with a high level of divalent cations (Levi-Setti et al., 2008), histone-like proteins (HLP) (Wong et al., 2003) and a dinoflagellate/viral nucleoprotein (DVNP) (Gornik et al., 2012).

The unusual dinoflagellate nuclear structure raises problems with respect to the mechanisms of both DNA replication and transcription. Little is known about replication, but many studies have examined changes in gene expression in response to light. Some of these studies use qPCR to examine specific genes. For example, rhodopsin in *Prorocentrum* was followed over a 14:10 L:D cycle and was observed to vary three-fold between LD 0 and LD 14 (Shi et al., 2015). Similarly, transcripts encoding the oxygen evolving enzyme OEE1 in *Symbiodinium* were 2.5 fold more abundant at LD 12 than at LD 0 (Sorek et al., 2013), while transcript levels encoding the large rubisco subunit *rbcL* were three fold higher at LD 12 than LD 0 (Mayfield et al., 2014) suggesting higher levels of transcription during the light. Levels of the thylakoid chlorophyll a-chlorophyll c2-peridinin-protein-complex (acpPC) were reported to be higher in dark phase than in light phase (Boldt et al., 2008a) suggesting that lack of light promotes expression of the light harvesting gene transcript.

Other experimental approaches have used high throughput expression measures such as microarrays or RNA Seq. One of the earliest studies on differential transcription between day and night was carried out with *Pyrocystis* using microarrays programmed with about 3500 cDNAs (Okamoto and Hastings, 2003). About 80 differentially expressed genes (DEG) (~ 2%) were found to have a >2-fold difference between day and night in this species, with a maximum observed change of 2.5-fold. A similar microarray study comparing genes expressed during the day and night in *Karenia brevis* found 458 DEG among the 4629 genes examined (10%), with a significance threshold of $p < 0.0001$ and > 1.7 fold change (Lidie and van Dolah, 2007). RNA Seq studies in *Symbiodinium microadriaticum* found 67 DEG (0.1%) between day and night using DESeq with a false discovery rate (FDR) of 0.1 (Baumgarten et al., 2013) and a maximum fold change of 160. A much more substantial number of DEG were noted in a study using *Symbiodinium* strain SSB01 24 hours after a transfer from light to dark (Xiang et al., 2015). There were 1334 DEG (2.2%) when cells were grown phototrophically and 1739 DEG (2.9%) when cells were grown mixotrophically. These studies used duplicates (phototrophic growth) or triplicates (mixotrophic growth), but instead of an FDR = 0.1, the cutoff values for significance were $p < 0.05$ and a > 1.5-fold change. Lastly, 131 DEG (0.17%) were found when samples of *Lingulodinium polyedra* taken every six hours were compared using an FDR of 0.1 (Roy et al., 2014a), but Northern blots analyses of a random selection of these showed no changes suggesting all were likely to be false positives.

The initial goal of our experiments was to identify a light regulated gene in *S. kawagutii*, so that potential regulatory elements in the promoter could be determined from the genome sequence (Lin et al., 2015), dissected and the potential transcription factors involved identified. In one approach, we selected three genes whose transcripts had been previously been reported to be light regulated in *Symbiodinium*, and verified their expression levels using Northern blots. In a second approach, we analysed global transcript levels at dawn and dusk by RNA-Seq. However, neither of these approaches successfully identified a light regulated gene, consistent with what has been observed with the dinoflagellate *L. polyedra*. This suggests that previous reports of light responsive genes may have overestimated their number, and further suggests that other reports of transcriptional responses may also benefit from additional verification.

3.3. Materials and Methods

3.3.1. Cell cultures

Symbiodinium kawagutii (CCMP2468) was obtained from the National Center for Marine Algae and Microbiota (Boothbay Harbor, Maine) and cultured at 24˚C under a 12:12 light: dark cycle (40 µE m^{-2} s^{-1}) in standard f/2 medium lacking silicate (Guillard and Ryther, 1962). *S. kawagutii* has recently been renamed *Fugacium kawagutii* (LaJeunesse et al., 2018).

3.3.2. Microscopy

Cells were concentrated by centrifugation, then resuspended in a solution of 3% freshly made formaldehyde in seawater for 10 minutes then washed three times with fresh seawater. Cells were finally resuspended in phosphate buffered saline containing 0.05% Tween 20 and 1 µg/mL propidium iodide for 30 minutes. Images were taken using a Zeiss confocal microscope using a 63X objective in green (PI) and red (chlorophyll) channels. 3D reconstructions were made using Fiji (Schindelin et al., 2012).

3.3.3. RNA extraction and Northern blots

For the high light condition, *S. kawagutii* cells in fresh normal culture medium were transferred to 350 µmol of photons m^{-2} s^{-1} high light (HL) for 24 hours. *S. kawagutii* cells were harvested from LD0 (beginning of light), LD4, LD8, LD12 (beginning of darkness), LD16, LD20, LD24 and HL cultures. Total RNA was extracted with Trizol as described (Beauchemin et al., 2012), quantified and stored at -80. *S. kawagutii* RBCL, AcpPC, OEE1 and Actin sequences were acquired from the genome sequence (http:web.malab.cn/symka_new). Primers were used to amplify the sequence from a first strand cDNA reaction product using *S. kawagutii* total RNA. The identify of all PCR products was confirmed by sequencing.

Northern blotting analysis was performed as described (Roy et al., 2014a), 10 μg total RNA was electrophoresed on a denaturing agarose gel. The RNAs were transferred onto a nylon membrane (HybondTM-H+; Amersham Pharmacia Biotechnology, Piscataway, N.J., U.S.A.) and cross-linked by UV. PCR generated probes were labeled with [α-32P] ATP (BLU512H, Perkin Elmer) for hybridization. Membranes were hybridized at 65°C for 16 h and were then washed twice at 65°C for 15 min. The radiolabeled membranes were exposed to a phosphoscreen for 24 hours and revealed by Typhoon Imager.

3.3.4. RNA sequencing

Quality control, library construction and Illumina sequencing were performed at the McGill University and Genome Quebec Innovation Centre (Montreal, Quebec). Between 36 and 57 million paired end reads were recovered for each of the six samples. Raw sequence reads are available from NCBI using the accession number PRJNA517819.

The unigene list used for read mapping was downloaded from the *S. kawagutii* genome resources (http://web.malab.cn/symka_new/). This unigene list, containing 70,987 sequences, as well as the six paired-end Illumina sequence reads, were uploaded to the Galaxy web platform at usegalaxy.org. The reads were trimmed using TrimGalore and read counts for all sequences in the unigene list were determined using Salmon (Patro et al., 2017). Statistical significance was estimated using DESeq2 running in R (Anders and Huber, 2010).

3.3.5. Quantitative PCR

cDNAs were prepared from *S. kawagutii* RNAs extracted from the cells collected at four-hour intervals over an LD cycle plus high light cultures using ProtoScript First Strand cDNA Synthesis Kit (Invitrogen). Specific primers were designed for SymkaALLUN13501, SymkaALLUN19088, SymkaALLUN64909 and Actin. qPCR analysis was performed in a ViiA7 Real-Time PCR System (Applied Biosystem) using SYBR green qPCR Master Mix (Thermo Fisher). Gene specific primers (250 nM) and cDNA (150 ng) were used in a total volume of 10 µl. Triplicate samples from each of three biological replicates amplified using 10 min at 95 °C, followed by 35 cycles of 15 s at 95 °C, 1 min at 60°C, and 35 s at 68 °C, followed by a melt curve stage from 60 °C to 95 °C to verify the absence of non-specific amplification.

For gene expression analysis, Cycle threshold (Ct) values were obtained from the ViiA7 Real-Time PCR software. Expression levels of three target genes (ΔCt) were obtained relative to Actin as a reference. Student's *t*-test was used to verify the statistical significance of the data.

3.4. Results

S. kawagutii has a typical dinoflagellate chromosome structure. Cells at all times have visibly condensed chromosomes (Figure 3.1) that appear superficially similar to mitotic chromosomes in other cells. This compact structure suggests that transcription is likely to be challenging, since for more typical eukaryotic cells transcription rates decrease during mitosis when the chromatin is more condensed (Palozola et al., 2017).

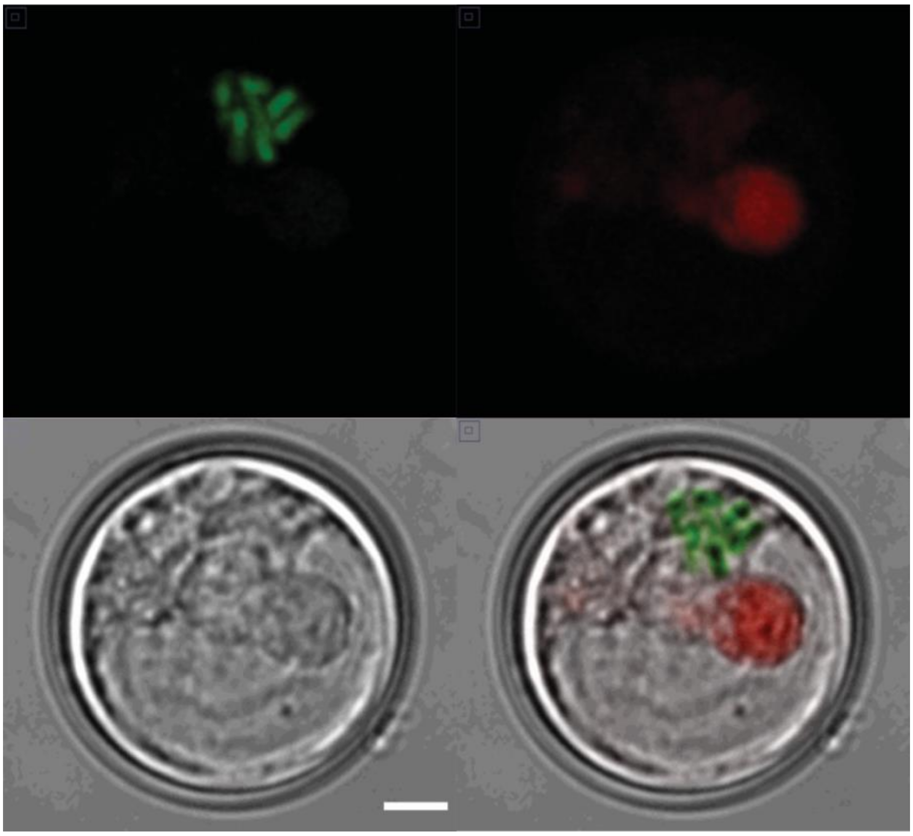
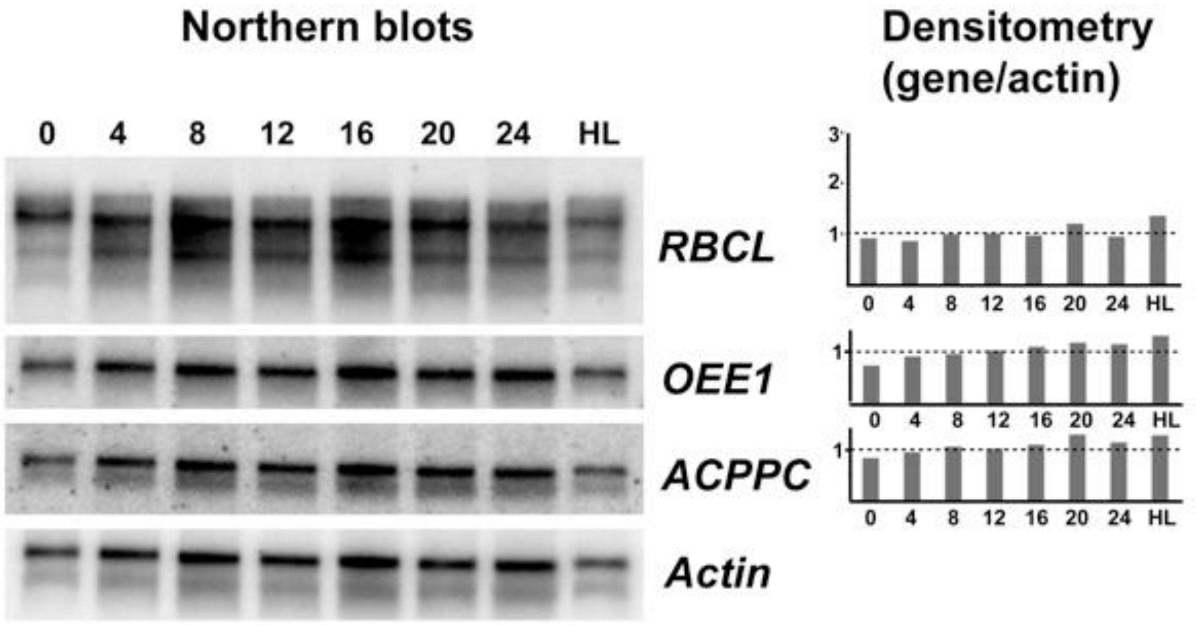


Figure 3.1. Condensed chromosomes in an interphase *S. kawagutii* cell. A confocal image of a single cell taken in (upper left) the green channel (PI staining of chromosomes), (upper right) the red channel (natural chlorophyll fluorescence) and (lower left) a DIC image (scale bar 1 µm). A merged image is shown in the lower right.

In a first attempt to identify light responsive genes in *S. kawagutii*, examples were selected from the literature. We selected oxygen evolving enzyme (OEE1) where transcript levels changed in abundance by 2.5 fold between LD 0 and LD 12 (Sorek et al., 2013), the large rubisco subunit *rbcL* where transcript levels were three-fold higher at LD 12 than LD 0 (Mayfield et al., 2014), and the thylakoid chlorophyll a-chlorophyll c2-peridinin-protein-complex (acpPC) where transcript levels were roughly three-fold higher in dark phase than in light phase (Boldt et al., 2008b). Actin was chosen as a reference because it is not regulated by light in *Lingulodinium* (Roy et al., 2014a) or as shown here by RNA-Seq in *S. kawagutii*. We amplified probes for these sequences from *S. kawagutii*, and used the probes to asses transcript levels at four-hour intervals over an LD cycle as well as a culture left under high light conditions. In no case were different transcript levels observed (Figure 3.2). We conclude there is no support for the hypothesis that transcription of these three genes responds to light.



⍰

Figure 3.2. Northern blot analysis of three potentially light regulated transcripts. A representative sample of Northern blots (n=4) using either an *rbcL*, an *oee1*, an *acppc* or an actin cDNA as a probe. RNA was prepared from samples taken every four hours from cells grown under a normal 12:12 LD cycle as well as from cells grown under high light (note that LD 0 and LD 24 should be identical). At right, densitometric scans for the top three probes are shown relative to the actin signal.

As a second attempt to identify light responsive genes, we prepared RNA samples in triplicate from *S. kawagutii* at LD 0 (dawn) and LD 12 (dusk). We reasoned that any light responsive genes would accumulate during the light period, and these would thus have higher levels at the end of the light phase. We compared read counts using the DESeq with a Benjamini-Hochberg correction (FDR = 0.1) to determine significant changes. A total of 7 changes (0.01%) were observed, all with higher levels at LD 0 than at LD 12 (Figure 3.3). Since all seven were higher at LD 0, this suggested that if these were truly light-regulated genes they would be induced by darkness or inhibited by light. These seven sequences were identified by BLAST searches (Table 3.1), and none correspond to the three sequences tested by Northern blots. When the stringency of statistical significance was increased by setting the FDR to 0.05, only one of these was observed to display a statistically significant change. When a Bonferroni correction was applied instead of the Benjamini-Hochberg correction, four genes showed significant changes with p < 0.05, and one with p < 0.01. We conclude the number of significant changes in transcript levels is very low.

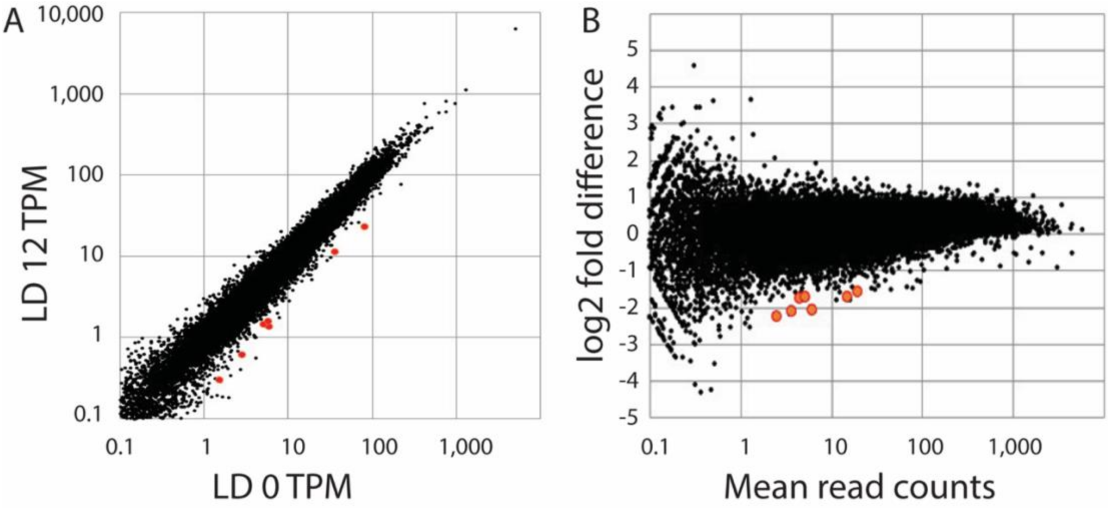


Figure 3.3. Comparison of transcript levels at LD 0 and LD 12. (A) A plot of read counts (as TPM, or transcripts per million) as the average of three samples at LD 12 are compared with the average of samples at LD 0. (B) An MA plot (fold-difference as a function of mean read count) is shown for triplicate samples at each of the two times as determined by DESeq2. The 7 sequences determined to be significantly different (p-adjust < 0.05; FDR = 0.1) are shown in red in both plots and are higher at LD 0 than at LD 12.

Gene ID	Best BLAST hit	E-value	Fold Change
SymkaALLUN26766	aminomethyl transferase family protein [Halobellus limi]	1.6	0.33
SymkaALLUN13501	putative alanine aminotransferase, mitochondrial	3e-13	0.23
SymkaALLUN70319	Hypothetical	9.7	0.24
SymkaALLUN19088	putative E3 ubiquitin-protein ligase HERC1	2e-21	0.2
SymkaALLUN64909	LysM domain-containing protein	2.9	0.3
SymkaALLUN23766	No Sig Hits	-	0.3
SymkaALLUN19996	No Sig hits	-	0.29

Table 3.1. Best BLAST hit for the seven potentially light regulated genes identified (FDR = 0.1)

To validate the differential expression of the seven genes detected by RNA Seq, we performed qPCR to assess the relative levels of three randomly selected genes (SymkaALLUN13501, SymkaALLUN19088, and SymkaALLUN64909). Assays were performed in triplicate for each of three biological replicates, and when compared to actin levels in each sample, two genes (ALLUN13501 and ALLUN19088) showed a significant difference between the two times (Figure 3.4). Since lower ΔCt values reflect higher transcript levels (i.e., transcript levels for these two genes are higher at LD 0, as found by RNA Seq), we conclude that at least some of the seven genes with different levels as measured by RNA Seq may reflect real differences in transcript levels. We note, however, that the fold difference appears smaller than that predicted by RNA Seq.

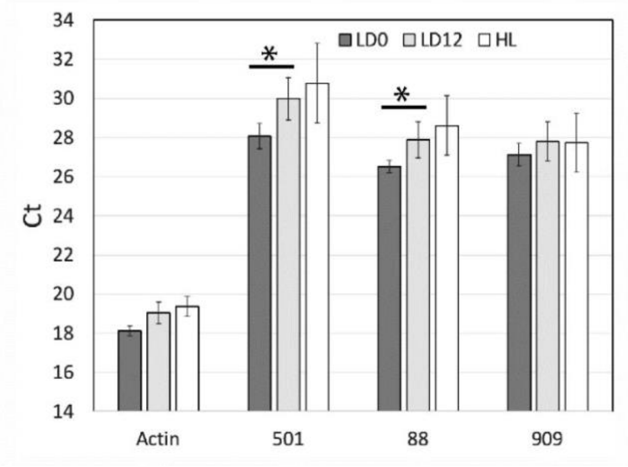


Figure 3.4. qPCR analysis of 3 selected light-regulated genes from RNA sequencing analysis. Ct values were obtained for three RNA-Seq predicted regulated genes (501, SymkaALLUN13501; 88, SymkaALLUN19088; 909, SymkaALLUN64909) as well as Actin as a control for the amount of cDNA. Triplicate samples from each of three biological replicates were averaged for LD 0 (samples were in the dark for 12 h), LD 12 (samples were in the light for 12 h), and for samples kept under constant high light for 24 h. Comparisons marked with * are significant at $p < 0.01$ using student's t-test, respectively.

Finally, to gain a global picture of the different fold changes detected, significant or not, we plotted the number of times different fold changes were observed as a function of the fold change (Figure 3.5). This analysis reveals a normal distribution of fold changes within the data set. To test the symmetry of the bell curve, positive fold changes were plotted as a function of negative fold changes (Figure 3.5 inset). The resulting curve is essentially a straight line with a slope of -1. The few exceptions to the linear relationship do not correspond to the genes classified as significant by DESeq. We conclude there is no overall bias for either positive or negative changes in transcript abundance between the two times examined.

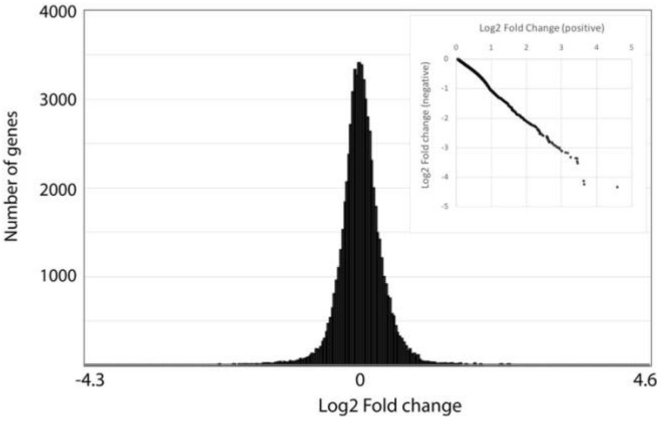


Figure 3.5. Fold changes are equally distributed. A histogram showing of the number of times a given log2 fold change is found in the data shows a normal distribution. The inset, showing a plot of the positive vs negative log2 fold changes, is essentially an unbiased straight line.

3.5. Discussion

In many of the studies reporting differential gene expression as a result of light, only a single method was used to measure transcript abundance. For example, qPCR, Northern blots, microarrays or RNA Seq have been used in individual studies but were not, with few exceptions, combined in the same study. One notable exception in *Lingulodinium* first used RNA Seq to identify DEG and then verified a random selection of these using Northern blots. Since Northern blots failed to confirm the RNA Seq-derived DEG, it was concluded all were likely to be false positives. This underscores the importance of validating high-throughput approaches, and suggests that it would be beneficial when several methods are combined to test for DEG.

The RNA-Seq experiments reported here used DESeq2 to identify DEG, with the threshold for significance determined by a false discovery rate (FDR) of 0.1. The FDR method, developed by Benjamini and Hochberg, uses a statistical method to restrain the number of false positives to a fixed percentage of the total positives, and thus provides increased confidence that what are termed significant changes are in large datasets are likely true positives (Benjamini and Hochberg, 1995). The FDR can be thought of as a method for using lower p-values to determine when datasets become larger. For example, using a dataset with 100 values, of which 5 are really significant, a p-value of 0.05 would mean there are 5 false positives detected among 95 non-significant values, thus corresponding to a false discovery rate of 50%. The false discovery rate climbs when either the number of really significant values decreases or the number of non-significant values increases, the latter being a direct consequence of using large datasets such as those produced by RNA Seq. In our study, when the FDR was fixed at 0.1, seven genes with significant difference were found. However, the number of significant differences decreases to 1 using a more stringent FDR of 0.05. It has been shown that the number of false positives recovered is considerably higher than the number expected (Rocke et al., 2015). This would agree with our observation that only two thirds of the DEG tested by qPCR were also found to show significant differences. Thus, in the light of the small number of significant changes found in our RNA Seq experiment, we suggest that there are likely no real significant changes in transcript levels brought about by the changes in light intensity in our experiment. This would then agree with the lack of significant changes in transcript abundance over the course of the daily LD cycle using the dinoflagellate *L. polyedra* (Roy et al., 2014a).

Our RNA Seq experiment indicating there are no light induced transcripts has methodological differences with other reports in the literature suggesting the opposite. For example, an RNA Seq study with *Symbiodinium microadriaticum* that showed 67 DEG when day and night were compared using DESeq with a false discovery rate (FDR) of 0.1 (Baumgarten et al., 2013) used single samples rather than triplicate samples (Table 3.2). When we perform DESeq with an FDR = 0.1 using only one of three samples for each of the two time points, DESeq recovers 55 DEG instead of the seven DEG found when triplicate samples are used. Thus, in the *S. microadriaticum* study, insufficient replication may have exaggerated the number of light responsive transcripts. Another RNA Seq study using *Symbiodinium* strain SSB01 looked at the number of DEG 24 hours after a transfer from light to dark (Xiang et al., 2015). Here, 1334 DEG were found using cells grown phototrophically and 1739 DEG when cells were grown mixotrophically. These studies used duplicates (phototrophic growth) or triplicates (mixotrophic growth), but, instead of an FDR = 0.1, the cut-off values for significance were $p < 0.05$ and a fold change > 1.5-fold. In our experiment, using triplicate samples with a similar cut-off value would result in 789 DEG instead of seven. Thus the *Symbiodinium* SSB01 study had an exaggerated number of DEG because the cut-off criteria were not as stringent as using an FDR of 0.1. Both replicated samples and appropriate statistical analysis of significance are required for correct interpretation of RNA Seq data.

It is important to emphasize that we do not propose dinoflagellates are incapable of transcriptional responses. However, in view of the experiments reported here, we believe it may be worthwhile re-examining the transcriptional response of dinoflagellates to stimuli other than light. A logical prediction from the permanently condensed chromatin that characterises dinoflagellate chromosomes is that transcriptional regulation is likely to be more difficult than in other cells. We thus suggest it may be important to verify transcriptional responses observed by a single method by using a complementary technique. Certainly, the finding of a true transcriptional response will be an important part in dissecting the molecular machinery that underpins this process in the dinoflagellates.

Species	Method	Comparisons	Replicates	FDR	p-value	DEG	Reference
S. kawagutii	Illumina/DESeq	LD0/ LD12	3	0.05		1	This study
	Illumina/DESeq	LD0/ LD12	3	0.1		7	This study
	Illumina/DESeq	LD0/ LD12	3		<0.05	789	This study
	Illumina/DESeq	LD0/ LD12	1	0.1		55	This study
S. microadriaticum	Illumina/DESeq	LD0/ LD12	1	0.1		67	Baumgarten 2013
	Illumina/DESeq	Normal/ 4h 4°C	1	0.1		119	Baumgarten 2013
	Illumina/DESeq	Normal/ 4h 36°C	1	0.1		2465	Baumgarten 2013
	Illumina/DESeq	Normal/ 12h 34°C	1	0.1		246	Baumgarten 2013
	Illumina/DESeq	Normal/ 4h 20g/L NaCl	1	0.1		138	Baumgarten 2013
	Illumina/DESeq	Normal/ 4h 60g/L NaCl	1	0.1		48	Baumgarten 2013
Symbiodinium SSB01	Illumina/DESeq	Light/ 24h dark	3		<0.05	1334	Xiang 2015
Symbiodinium	Illumina/DESeq	29.2°C/3d 31.9°C	2	0.05		0	Barshis 2014
	Illumina/DESeq	29.2°C/3d 31.9°C	2		<0.05	541	Barshis 2014
Symbiodinium sp	Illumina/Student's t test	Normal/ 4d 31°C	5	0.05		9471	Gierz 2017
	Illumina/Student's t test	Normal/ 19d 31°C	5			12701	Gierz 2017
	Illumina/Student's t test	Normal/ 28d 31°C	5			13269	Gierz 2017
Lingulodinium polyedra	Illumina/DESeq	Normal/ 1d 4°C	1	0.05		132	Roy 2014a
Lingulodinium polyedra	Illumina/DESeq	LD6/ LD18	1	0.05		5	Roy 2014b
Scrippsiella trochoidea	Illumina/DESeq	Normal/ N-limited	1	0.1		382	Cooper 2016
Scrippsiella trochoidea	Illumina/DESeq	Normal/ P-limited	1	0.1		17	Cooper 2016
Alexandrium tamarense	MPSS/Fisher's exact test	Normal/ N-limited	1		<1E-10	20	Moustafa 2010
Alexandrium tamarense	MPSS/Fisher's exact test	Normal/ P-limited	1		<1E-10	30	Moustafa 2010
Alexandrium tamarense	MPSS/Fisher's exact test	Normal/ Xenic	1		<1E-10	505	Moustafa 2010
Oxyrrhis marina	454/Fisher's exact test	30/50 practical saline units	1		@ < 0.05	29	Lowe 2011
Karenia brevis	Microarray	Normal/ N-limited	3		< 0.0001	456	Morey 2011
Karenia brevis	Microarray	Normal/ P-limited	3		< 0.0001	425	Morey 2011

Table 3.2. DEG identified in different dinoflagellates after different treatments

CHAPTER 4 – GENERAL DISCUSSION

4. 1. General Discussion

Dinoflagellates are a unique family of microscopic plankton in marine and freshwater environments. In the oceanic ecosystem, many are photosynthetic and are vital primary producers at the base of the food chain. Some species are bioluminescent, some can form harmful algal blooms releasing potent biotoxins and others are endosymbionts with anthozoans forming coral reefs. These features have encouraged scientists to study the regulation of gene expression that underlies these physiological processes. However, even after half a century of experimentation, it is still unclear how gene expression is regulated in these unusual eukaryotic organisms with their often gigantic genomes and permanently condensed chromosomes without nucleosomes.

Techniques of forward and reverse genetics are not yet generally available in dinoflagellates. There have been reports of successful transformation in some species, including *Symbiodinium*, but so far we and others have been unable to replicate these reports (Chen et al., 2019a). For example, one report showed green fluorescent *Symbiodinium* following transformation with a GFP transgene (Ortiz-Matamoros, 2015). However, in our hands, green fluorescent “transformed” *S. kawagutii* cells could be seen even without addition of the GFP transgene, clearly indicating transformation did not produce the green fluorescent phenotype. Thus, instead of transformation or classical genetic studies, researchers have turned to microarray studies and high throughput sequencing technologies, including RNA-Seq, as means of assessing gene expression. This latter is the logic employed here. I have also used recombinant DNA technology to express dinoflagellate genes in other organisms to study the properties of individual proteins *in vitro*.

Since transcription of different genes is regulated by promotor-specific transcription factors (Hampsey, 1998) and dinoflagellate cold shock domain (CSD) proteins form two third of the annotated transcription factors in the transcriptomes and the *S. kawagutii* genome (Bayer et al., 2012, Beauchemin et al., 2012, Li et al., 2020), my first interest was to test the hypothesis that dinoflagellate CSPs were involved in transcriptional regulation. I examined one *L. polyedra* CSP and 3 different *S. kawagutii* CSPs for binding to nucleic acids in a sequence-specific manner to assess a possible role as transcription factors for these proteins. The first observation was that all of these CSPs were active in binding nucleic acids in EMSA assays. They were able to bind RNA better than single stranded and double stranded DNA using competition EMSA experiments, which does not support a role as transcription factors. Additionally, I performed selection and amplification binding (SAAB) experiments to determine if these proteins could target a specific sequence on DNA. After performing three cycles of binding and PCR amplification, none of the four CSPs enriched a specific motif, again inconsistent with a sequence specific transcription factor role. Taken all together, dinoflagellate CSPs do bind nucleic acids, thus while they were annotated as transcription factors, they are unlikely to play a sequence specific role *in vivo*.

What role might CSPs play in dinoflagellates? Their strong binding to RNA could be comparable to the association of many transcription factors with diverse types of RNA that regulates the gene expression in various ways including binding to mRNA products of transcription as RNA chaperones to regulate mRNA translation rate and direct interaction with long non-coding RNAs (lncRNAs) which act as a scaffold for transcription factor assembly. Thus, leading to control over gene expression at the posttranscriptional level. Given the high representation of RNA binding proteins in the chromatin-associated protein catalogue in dinoflagellates, the RNA binding property of CSPs could also assist in regulating the chromatin structure. Alternatively, their binding to ssDNA could be suggestive of aiding transcription by facilitating unwinding the DNA double helix structure.

Regulation of gene expression by extracellular signals such as light has been frequently studied in dinoflagellates. I selected three candidate light regulated genes in *Symbiodinium* for further study. My ultimate goal was to analyse the *cis* and *trans* acting factors mediating light-regulated transcriptional responses. I first used RNA-Seq to validate the transcriptional response of these three genes to light in our *S. kawagutii* cultures. Using an FDR of 0.1, seven genes with significant difference in transcript abundance at the end of the day compared to the end of the night were found, although the three original candidates selected from the literature were not among these seven. Interestingly, one of three genes, randomly selected from the seven, did not show significant differences when examined by qPCR. I conclude that, in *S. kawagutii,* transcriptional regulation by light is rare. This is supported by the observation that if the FDR is decreased to a more stringent value of 0.05, only one gene would be considered to vary in a significant manner, and this one would have been confirmed by qPCR. This analysis is consistent with the idea there are few light regulated genes in *Symbiodinium*, and also agrees with the absence of significant differences in transcript level during a day-night cycle in *L. polyedra*.

To sum up, regulation of gene expression at a transcriptional level does not appear to be as prevalent in dinoflagellates as it is in other organisms. This would thus help to explain the under-representation of DNA binding domains and components of the general transcription apparatus. Moreover, the low level of protein associated with dinoflagellate genomes and an apparent lack of nucleosomes suggests epigenetic regulation of transcription may also be infrequent. It does seem likely that there is a role played by histones given the conservation of histone genes and histone-modifying enzymes, however what exactly this role might be is unclear. *S. kawagutii* probably lacks a real response to light at the transcriptional level. This is consistent with the discovery of few promoter elements upstream of dinoflagellate genes, which decreases a need for specific transcription factors and their cognate *cis* regulatory sequences to initiate transcription.

4. 2. Future perspectives

Updated genome assemblies and transcriptomes of *S. kawagutii* (Li et al., 2020) deposited at SAGER (Symbiodiniaceae and Algal Genomic Resource Database) as well as other dinoflagellate transcriptomes deposited at NCBI are wide-ranging databases available for further studies. The transcripts encoding proteins can be selected, amplified and expressed in bacteria to further investigate the role and function of the many unique proteins in dinoflagellates. As an example, it is possible that dinoflagellate sequence banks may contain previously unidentified transcription factors, similar to the novel group of ApiAP2 transcription factors recently discovered in apicomplexans, yet these will have to be characterized biochemically.

It will also be important to validate the transcriptional responses of dinoflagellates to other extracellular stimuli such as nutrition deficiency and temperature. Our results suggest using several different techniques instead of relying on one. In contrast with previous positive reports, my work reported here showed neither Northern blots nor RNA-seq recognized a light regulated gene in *S. kawagutii*, previously identified by qPCR. The necessity of validating the any given approach by additional experiments is thus evident. Given that regulation of gene expression at transcriptional level seem to be uncommon in dinoflagellate for low levels of transcription regulators and their promoter elements, examining posttranscriptional mechanisms mediating transcripts level should be more focused for future studies.

5. Bibliography

ADL, S. M., SIMPSON, A. G., FARMER, M. A., ANDERSEN, R. A., ANDERSON, O. R., BARTA, J. R., BOWSER, S. S., BRUGEROLLE, G., FENSOME, R. A., FREDERICQ, S., JAMES, T. Y., KARPOV, S., KUGRENS, P., KRUG, J., LANE, C. E., LEWIS, L. A., LODGE, J., LYNN, D. H., MANN, D. G., MCCOURT, R. M., MENDOZA, L., MOESTRUP, O., MOZLEY-STANDRIDGE, S. E., NERAD, T. A., SHEARER, C. A., SMIRNOV, A. V., SPIEGEL, F. W. & TAYLOR, M. F. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol,* 52**,** 399-451.

ANDERS, S. & HUBER, W. 2010. Differential expression analysis for sequence count data. *Genome Biol,* 11**,** R106.

ANDERSON, D. M. & STOLZENBACH, K. D. 1985. Selective retention of two dinoflagellates in a well-mixed estuarine embayment: the importance of diel vertical migration and surface avoidance. *Marine Ecology Progress Series* 25**,** 39–50.

ARANDA, M., LI, Y., LIEW, Y. J., BAUMGARTEN, S., SIMAKOV, O., WILSON, M. C., PIEL, J., ASHOOR, H., BOUGOUFFA, S., BAJIC, V. B., RYU, T., RAVASI, T., BAYER, T., MICKLEM, G., KIM, H., BHAK, J., LAJEUNESSE, T. C. & VOOLSTRA, C. R. 2016. Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. *Sci Rep,* 6**,** 39734.

BABU, M. M., LUSCOMBE, N. M., ARAVIND, L., GERSTEIN, M. & TEICHMANN, S. A. 2004. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol,* 14**,** 283-91.

BACHVAROFF, T. R., GORNIK, S. G., CONCEPCION, G. T., WALLER, R. F., MENDEZ, G. S., LIPPMEIER, J. C. & DELWICHE, C. F. 2014. Dinoflagellate phylogeny revisited: using ribosomal proteins to resolve deep branching dinoflagellate clades. *Mol Phylogenet Evol,* 70**,** 314-22.

BACHVAROFF, T. R. & PLACE, A. R. 2008. From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate Amphidinium carterae. *PLoS One,* 3**,** e2929.

BAE, W., XIA, B., INOUYE, M. & SEVERINOV, K. 2000. Escherichia coli CspA-family RNA chaperones are transcription antiterminators. *Proc Natl Acad Sci U S A,* 97**,** 7784-9.

BARBROOK, A. C. & HOWE, C. J. 2000. Minicircular plastid DNA in the dinoflagellate Amphidinium operculatum. *Mol Gen Genet,* 263**,** 152-8.

BAUMGARTEN, S., BAYER, T., ARANDA, M., LIEW, Y. J., CARR, A., MICKLEM, G. & VOOLSTRA, C. R. 2013. Integrating microRNA and mRNA expression profiling in Symbiodinium microadriaticum, a dinoflagellate symbiont of reef-building corals. *BMC Genomics,* 14**,** 704.

BAYER, T., ARANDA, M., SUNAGAWA, S., YUM, L. K., DESALVO, M. K., LINDQUIST, E., COFFROTH, M. A., VOOLSTRA, C. R. & MEDINA, M. 2012. Symbiodinium transcriptomes: genome insights into the dinoflagellate symbionts of reef-building corals. *PLoS One,* 7**,** e35269.

BEAUCHEMIN, M. & MORSE, D. 2018. A proteomic portrait of dinoflagellate chromatin reveals abundant RNA-binding proteins. *Chromosoma,* 127**,** 29-43.

BEAUCHEMIN, M., ROY, S., DAOUST, P., DAGENAIS-BELLEFEUILLE, S., BERTOMEU, T., LETOURNEAU, L., LANG, B. F. & MORSE, D. 2012. Dinoflagellate tandem array gene transcripts are highly conserved and not polycistronic. *Proc Natl Acad Sci U S A,* 109**,** 15793-8.

BEAUCHEMIN, M., ROY, S., PELLETIER, S., AVERBACK, A. & MORSE, D. 2016. Characterization of two dinoflagellate cold shock domain proteins. *mSphere* e00034-15.

BELOTSERKOVSKAYA, R. & BERGER, S. L. 1999. Interplay between chromatin modifying and remodeling complexes in transcriptional regulation. *Crit Rev Eukaryot Gene Expr,* 9**,** 221-30.

BENIZRI, E., GINOUVES, A. & BERRA, E. 2008. The magic of the hypoxia-signaling cascade. *Cell Mol Life Sci,* 65**,** 1133-49.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the fasle discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.,* 57**,** 289–300.

BICKMORE, W. A. 2013. The spatial organization of the human genome. *Annu Rev Genomics Hum Genet,* 14**,** 67-84.

BITAR, M., BORONI, M., MACEDO, A. M., MACHADO, C. R. & FRANCO, G. R. 2013. The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. *Front Genet,* 4**,** 199.

BLANK R.J., H. V. A. R., KERSTEN W 1988. Base composition of DNA from symbiotic dinoflagellates: A tool for phylogenetic classification. *Arch. Microbiol.,* 149**,** 515-520.

BODANSKY, S., MINTZ, L. B. & HOLMES, D. S. 1979. The mesokaryote Gyrodinium cohnii lacks nucleosomes. *Biochem Biophys Res Commun,* 88**,** 1329-36.

BOHMANN, D. 1990. Transcription factor phosphorylation: a link between signal transduction and the regulation of gene expression. *Cancer Cells,* 2**,** 337-44.

BOLDT, L., YELLOWLEES, D. & LEGGAT, W. 2008a. Measuring symbiodinium sp. Gene expression patterns with quantitative real-time pcr. *Proceedings of the 11th International Coral Reef Symposium (ICRS).* USA: Lauderdale, FL.

BOLDT, L., YELLOWLEES, D. & LEGGAT, W. 2008b. Measuring symbiodinium sp. Gene expression patterns with quantitative real-time pcr. *Proceedings of the 11th International Coral Reef Symposium (ICRS); Ft. Lauderdale, FL, USA***,** 118–122.

BOUBE, M., JOULIA, L., CRIBBS, D. L. & BOURBON, H. M. 2002. Evidence for a mediator of RNA polymerase II transcriptional regulation conserved from yeast to man. *Cell,* 110**,** 143-51.

BOWAZOLO, C., TSE, S. P. K., BEAUCHEMIN, M., LO, S. C., RIVOAL, J. & MORSE, D. 2020. Label-free MS/MS analyses of the dinoflagellate Lingulodinium identifies rhythmic proteins facilitating adaptation to a diurnal LD cycle. *Sci Total Environ,* 704**,** 135430.

BUCHHEIM, M. A. & CHAPMAN, R. L. 1991. Phylogeny of the colonial green flagellates: a study of 18S and 26S rRNA sequence data. *Biosystems,* 25**,** 85-100.

BUDKINA, K. S., ZLOBIN, N. E., KONONOVA, S. V., OVCHINNIKOV, L. P. & BABAKOV, A. V. 2020. Cold Shock Domain Proteins: Structure and Interaction with Nucleic Acids. *Biochemistry (Mosc),* 85**,** S1-S19.

BUSKEY, E. J., REYNOLDS, G. T., SWIFT, E. & WALTON, A. J. 1985. Interaction between copepods and bioluminescent dinoflagellates: direct observation using image intensification *Biol Bull,* 169**,** 530.

BUTLER, J. E. & KADONAGA, J. T. 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev,* 16**,** 2583-92.

CARNINCI, P., SANDELIN, A., LENHARD, B., KATAYAMA, S., SHIMOKAWA, K., PONJAVIC, J., SEMPLE, C. A., TAYLOR, M. S., ENGSTROM, P. G., FRITH, M. C., FORREST, A. R., ALKEMA, W. B., TAN, S. L., PLESSY, C., KODZIUS, R., RAVASI, T., KASUKAWA, T., FUKUDA, S., KANAMORI-KATAYAMA, M., KITAZUME, Y., KAWAJI, H., KAI, C., NAKAMURA, M., KONNO, H., NAKANO, K., MOTTAGUI-TABAR, S., ARNER, P., CHESI, A., GUSTINCICH, S., PERSICHETTI, F., SUZUKI, H., GRIMMOND, S. M., WELLS, C. A., ORLANDO, V., WAHLESTEDT, C., LIU, E. T., HARBERS, M., KAWAI, J., BAJIC, V. B., HUME, D. A. & HAYASHIZAKI, Y. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet,* 38**,** 626-35.

CHAN, Y. H. & WONG, J. T. 2007. Concentration-dependent organization of DNA by the dinoflagellate histone-like protein HCc3. *Nucleic Acids Res,* 35**,** 2573-83.

CHANG, C., JACOBS, Y., NAKAMURA, T., JENKINS, N. A., COPELAND, N. G. & CLEARY, M. L. 1997. Meis proteins are major in vivo DNA binding partners for wild-type but not chimeric pbx proteins. *Molecular and Cellular Biology,* 17**,** 56795687.

CHEN, J. E., BARBROOK, A. C., CUI, G., HOWE, C. J. & ARANDA, M. 2019a. The genetic intractability of Symbiodinium microadriaticum to standard algal transformation methods. *PLoS One,* 14**,** e0211936.

CHEN, Y., GONZÁLEZ-PECH, R., STEPHENS, T., BHATTACHARYA, D. & CHAN, C. 2019b. Evidence That Inconsistent Gene Prediction Can Mislead Analysis of Dinoflagellate Genomes. *J. Phycol.,* 56**,** 6-10.

CHOW, L. T., GELINAS, R. E., BROKER, T. R. & ROBERTS, R. J. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell,* 12**,** 1-8.

CHUDNOVSKY, Y. L., J.F.; RIZZO, P.J.; HASTINGS, J.W.; FAGAN, T. 2002. Cloning, expression, and characterization of a histone-like protein from the marine dinoflagellate Lingulodinium polyedrum. *J. Phycol. ,* 38**,** 543–550.

COCK, P. J., CHILTON, J. M., GRUNING, B., JOHNSON, J. E. & SORANZO, N. 2015. NCBI BLAST+ integrated into Galaxy. *Gigascience,* 4**,** 39.

CULLEN, J. J. & HORRIGAN, S. G. 1981. Effects of nitrate on the diurnal vertical migration, carbon to nitrogen ratio, and the photosynthetic capacity of the dinoflagellate Gymnodinium splendens. *Mar. Biol. ,* 62**,** 81-89.

DAGENAIS-BELLEFEUILLE, S. & MORSE, D. 2013. Putting the N in dinoflagellates. *Front Microbiol,* 4**,** 369.

DAVIES, W. S., JAKOBSEN, K. & NORDBY, O. 1988. Characterization of DNA from the dinoflagellate Woloszynskia bostoniensis. *Journal of Protozoolology,* 35**,** 418–422.

DE MENDOZA, A., BONNET, A., VARGAS-LANDIN, D. B., JI, N., HONG, F., YANG, F., LI, L., HORI, K., PFLUEGER, J. & BUCKBERRY, S. 2018. Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons. *Nat Commun,* 9**,** 1341.

DECHAT, T., ADAM, S. A. & GOLDMAN, R. D. 2009. Nuclear lamins and chromatin: when structure meets function. *Adv Enzyme Regul,* 49**,** 157-66.

DENG, Y., HU, Z., CHAI, Z. & TANG, Y. Z. 2019. Cloning and Partial Characterization of a Cold Shock Domain-Containing Protein Gene from the Dinoflagellate Scrippsiella trochoidea. *J Eukaryot Microbiol,* 66**,** 393-403.

DEREEPER, A., GUIGNON, V., BLANC, G., AUDIC, S., BUFFET, S., CHEVENET, F., DUFAYARD, J. F., GUINDON, S., LEFORT, V., LESCOT, M., CLAVERIE, J. M. & GASCUEL, O. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res,* 36**,** W465-9.

DIKSTEIN, R., RUPPERT, S. & TJIAN, R. 1996. TAFII250 is a bipartite protein kinase that phosphorylates the base transcription factor RAP74. *Cell,* 84**,** 781-90.

DILWORTH, F. J., FROMENTAL-RAMAIN, C., YAMAMOTO, K. & CHAMBON, P. 2000. ATP-driven chromatin remodeling activity and histone acetyltransferases act sequentially during transactivation by RAR/RXR In vitro. *Mol Cell,* 6**,** 1049-58.

DOWNS, C. A., FAUTH, J. E., HALAS, J. C., DUSTAN, P., BEMISS, J. & WOODLEY, C. M. 2002. Oxidative stress and seasonal coral bleaching. *Free Radic Biol Med,* 33**,** 533-43.

DUNLAP, J. C. & HASTINGS, J. W. 1981. The biological clock in *Gonyaulax* controls luciferase activity by regulating turnover. *J Biol Chem,* 256**,** 10509-18.

DYNLACHT, B. D., HOEY, T. & TJIAN, R. 1991. Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. *Cell,* 66**,** 563-76.

ENGEL, C., SAINSBURY, S., CHEUNG, A. C., KOSTREWA, D. & CRAMER, P. 2013. RNA polymerase I structure and transcription regulation. *Nature,* 502**,** 650-5.

EVERETT, R. D., BATY, D. & CHAMBON, P. 1983. The repeated GC-rich motifs upstream from the TATA box are important elements of the SV40 early promoter. *Nucleic Acids Res,* 11**,** 2447-64.

FAUCHOT, J. L., M.; ROY, S. 2005. Daytime and nighttime vertical migrations of Alexandrium tamarense in the St. Lawrence estuary (Canada). *Mar. Ecol. Prog. Ser.,* 296**,** 241-250.

FAZZIO, T. G., KOOPERBERG, C., GOLDMARK, J. P., NEAL, C., BASOM, R., DELROW, J. & TSUKIYAMA, T. 2001. Widespread collaboration of Isw2 and Sin3-Rpd3 chromatin remodeling complexes in transcriptional repression. *Mol Cell Biol,* 21**,** 6450-60.

FENSOME, R. A., MACRAE, R. A. & WILLIAMS, G. L. 1994 & '95. Dinoflagellate Evolution and Diversity Through Time. *Science Review,* 1996**,** 45-50.

FIELD, C. B., BEHRENFELD, M. J., RANDERSON, J. T. & FALKOWSKI, P. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science,* 281**,** 237-40.

FIGUEROA, R. I., DAPENA, C., BRAVO, I. & CUADRADO, A. 2015. The Hidden Sexuality of Alexandrium Minutum: An Example of Overlooked Sex in Dinoflagellates. *PLoS One,* 10**,** e0142667.

FITT, W. K. & TRENCH, R. K. 1983. The Relation of Diel Patterns of Cell Division to Diel Patterns of Motility in the Symbiotic Dinoflagellate Symbiodinium microadriaticum Freudenthal in Culture *The New Phytologist,* 94**,** 421-432

FLEISHER KJAC, J. F. 1995. Cephalopod prediction facilitated by dinoflagellate luminescence *Biol Bull,* 189**,** 263-271.

FRANKEL, M. B. & KNOLL, L. J. 2009. The ins and outs of nuclear trafficking: unusual aspects in apicomplexan parasites. *DNA Cell Biol,* 28**,** 277-84.

FRIED, H. & KUTAY, U. 2003. Nucleocytoplasmic transport: taking an inventory. *Cell Mol Life Sci,* 60**,** 1659-88.

FUSARO, A. F., BOCCA, S. N., RAMOS, R. L., BARROCO, R. M., MAGIOLI, C., JORGE, V. C., COUTINHO, T. C., RANGEL-LIMA, C. M., DE RYCKE, R., INZE, D., ENGLER, G. & SACHETTO-MARTINS, G. 2007. AtGRP2, a cold-induced nucleo-cytoplasmic RNA-binding protein, has a role in flower and seed development. *Planta,* 225**,** 1339-51.

GAVELIS, G. S., HAYAKAWA, S., WHITE, R. A., 3RD, GOJOBORI, T., SUTTLE, C. A., KEELING, P. J. & LEANDER, B. S. 2015. Eye-like ocelloids are built from different endosymbiotically acquired components. *Nature,* 523**,** 204-7.

GIERZ, S. L., GORDON, B. R. & LEGGAT, W. 2016. Integral Light-Harvesting Complex Expression In Symbiodinium Within The Coral Acropora aspera Under Thermal Stress. *Sci Rep,* 6**,** 25081.

GLIBERT, P. M., ANDERSON, D. M., GENTIEN, P. & SELLNER, K. 2005. The global, complex phenomena of harmful algal blooms. *Oceanography (Wash DC)* 18**,** 132–141.

GÓMEZ, F. 2012. A CHECKLIST AND CLASSIFICATION OF LIVING DINOFLAGELLATES (DINOFLAGELLATA, ALVEOLATA). *CICIMAR Oceánides,* 27**,** 65-140.

GORDON, B. R. & LEGGAT, W. 2010. Symbiodinium-invertebrate symbioses and the role of metabolomics. *Mar Drugs,* 8**,** 2546-68.

GORNIK, S. G., FEBRIMARSA, CASSIN, A. M., MACRAE, J. I., RAMAPRASAD, A., RCHIAD, Z., MCCONVILLE, M. J., BACIC, A., MCFADDEN, G. I., PAIN, A. & WALLER, R. F. 2015. Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. *Proc Natl Acad Sci U S A,* 112**,** 5767-72.

GORNIK, S. G., FORD, K. L., MULHERN, T. D., BACIC, A., MCFADDEN, G. I. & WALLER, R. F. 2012. Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. *Curr Biol,* 22**,** 2303-12.

GRAUMANN, P. L. & MARAHIEL, M. A. 1998. A superfamily of proteins that contain the cold-shock domain. *Trends Biochem Sci,* 23**,** 286-90.

GUILLARD, R. R. & RYTHER, J. H. 1962. Studies of marine planktonic diatoms. I. Cyclotella nana Hustedt, and Detonula confervacea (cleve) Gran. *Can J Microbiol,* 8**,** 229-39.

GUILLEBAULT, D., SASORITH, S., DERELLE, E., WURTZ, J. M., LOZANO, J. C., BINGHAM, S., TORA, L. & MOREAU, H. 2002. A new class of transcription initiation factors, intermediate between TATA box-binding proteins (TBPs) and TBP-like factors (TLFs), is present in the marine unicellular organism, the dinoflagellate Crypthecodinium cohnii. *J Biol Chem,* 277**,** 40881-6.

GUO, R., EBENEZER, V. & KI, J. S. 2012. Transcriptional responses of heat shock protein 70 (Hsp70) to thermal, bisphenol A, and copper stresses in the dinoflagellate Prorocentrum minimum. *Chemosphere,* 89**,** 512-20.

GUO, R. & KI, J. S. 2012. Differential transcription of heat shock protein 90 (HSP90) in the dinoflagellate Prorocentrum minimum by copper and endocrine-disrupting chemicals. *Ecotoxicology,* 21**,** 1448-57.

HACKETT, J. D., ANDERSON, D. M., ERDNER, D. L. & BHATTACHARYA, D. 2004a. Dinoflagellates: a remarkable evolutionary experiment. *Am J Bot,* 91**,** 1523-34.

HACKETT, J. D., SCHEETZ, T. E., YOON, H. S., SOARES, M. B., BONALDO, M. F., CASAVANT, T. L. & BHATTACHARYA, D. 2005. Insights into a dinoflagellate genome through expressed sequence tag analysis. *BMC Genomics,* 6**,** 80.

HACKETT, J. D., YOON, H. S., SOARES, M. B., BONALDO, M. F., CASAVANT, T. L., SCHEETZ, T. E., NOSENKO, T. & BHATTACHARYA, D. 2004b. Migration of the plastid genome to the nucleus in a peridinin dinoflagellate. *Curr Biol,* 14**,** 213-8.

HAMPSEY, M. 1998. Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol Mol Biol Rev,* 62**,** 465-503.

HASTINGS, J. W. 1996. Chemistries and colors of bioluminescent reactions: a review. *Gene,* 173**,** 5-11.

HASTINGS, J. W. 2007. The Gonyaulax clock at 50: translational control of circadian expression. *Cold Spring Harb Symp Quant Biol,* 72**,** 141-4.

HASTINGS, J. W. 2013. Circadian Rhythms in Dinoflagellates: What Is the Purpose of Synthesis and Destruction of Proteins? *Microorganisms,* 1**,** 26-32.

HASTINGS, J. W., ASTRACHAN, L. & SWEENEY, B. M. 1961. A persistent daily rhythm in photosynthesis. *J Gen Physiol,* 45**,** 69-76.

HASTINGS, J. W. & SWEENEY, B. M. 1958. A persistant diural rhythm of luminescence in Gonyaulax polyedra. *The Biological bulletin,* 115**,** 444—458.

HASTINGS, K. E. 2005. SL trans-splicing: easy come or easy go? *Trends Genet,* 21**,** 240-7.

HEINEMANN, U. & ROSKE, Y. 2021. Cold-Shock Domains-Abundance, Structure, Properties, and Nucleic-Acid Binding. *Cancers (Basel),* 13.

HERGETH, S. P. & SCHNEIDER, R. 2015. The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle. *EMBO Rep,* 16**,** 1439-53.

HERNANDEZ, N. 1993. TBP, a universal eukaryotic transcription factor? *Genes Dev,* 7**,** 1291-308.

HERNANDEZ, N. 2001. snRNA genes: a model system to study fundamental mechanisms of transcription. *J Biol Chem*.

HOEGH-GULDBERG, O., MUMBY, P. J., HOOTEN, A. J., STENECK, R. S., GREENFIELD, P., GOMEZ, E., HARVELL, C. D., SALE, P. F., EDWARDS, A. J., CALDEIRA, K., KNOWLTON, N., EAKIN, C. M., IGLESIAS-PRIETO, R., MUTHIGA, N., BRADBURY, R. H., DUBI, A. & HATZIOLOS, M. E. 2007. Coral reefs under rapid climate change and ocean acidification. *Science,* 318**,** 1737-42.

HOLCK, A., LOSSIUS, I., AASLAND, R., HAARR, L. & KLEPPE, K. 1987. DNA- and RNA-binding proteins of chromatin from Escherichia coli. *Biochim Biophys Acta,* 908**,** 188-99.

HOU, Y. & LIN, S. 2009. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. *PLoS One,* 4**,** e6978.

IGNATIADES, L. & GOTSIS-SKRETAS, O. 2010. A review on toxic and harmful algae in Greek coastal waters (E. Mediterranean Sea). *Toxins (Basel),* 2**,** 1019-37.

IRWIN, N. A. T., MARTIN, B. J. E., YOUNG, B. P., BROWNE, M. J. G., FLAUS, A., LOEWEN, C. J. R., KEELING, P. J. & HOWE, L. J. 2018. Viral proteins as a potential driver of histone depletion in dinoflagellates. *Nat Commun,* 9**,** 1535.

IZUMI, H., IMAMURA, T., NAGATANI, G., ISE, T., MURAKAMI, T., URAMOTO, H., TORIGOE, T., ISHIGUCHI, H., YOSHIDA, Y., NOMOTO, M., OKAMOTO, T., UCHIUMI, T., KUWANO, M., FUNA, K. & KOHNO, K. 2001. Y box-binding protein-1 binds preferentially to single-stranded nucleic acids and exhibits 3'-->5' exonuclease activity. *Nucleic Acids Res,* 29**,** 1200-7.

JAECKISCH, N., YANG, I., WOHLRAB, S., GLOCKNER, G., KROYMANN, J., VOGEL, H., CEMBELLA, A. & JOHN, U. 2011. Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate Alexandrium ostenfeldii. *PLoS One,* 6**,** e28012.

JOHN, U., LU, Y., WOHLRAB, S., GROTH, M., JANOUSKOVEC, J., KOHLI, G. S., MARK, F. C., BICKMEYER, U., FARHAT, S., FELDER, M., FRICKENHAUS, S., GUILLOU, L., KEELING, P. J., MOUSTAFA, A., PORCEL, B. M., VALENTIN, K. & GLOCKNER, G. 2019. An aerobic eukaryotic parasite with functional mitochondria that likely lacks a mitochondrial genome. *Sci Adv,* 5**,** eaav1110.

JOHNSON, C. H., INOUE, S., FLINT, A. & HASTINGS, J. W. 1985. Compartmentalization of algal bioluminescence: autofluorescence of bioluminescent particles in the dinoflagellate Gonyaulax as studied with image-intensified video microscopy and flow cytometry. *J Cell Biol,* 100**,** 1435-46.

JOHNSON, C. H., ROEBER, J. F. & HASTINGS, J. W. 1984. Circadian changes in enzyme concentration account for rhythm of enzyme activity in gonyaulax. *Science,* 223**,** 1428-30.

JONES, P. G., KRAH, R., TAFURI, S. R. & WOLFFE, A. P. 1992. DNA gyrase, CS7.4, and the cold shock response in Escherichia coli. *J Bacteriol,* 174**,** 5798-802.

JONES, P. G., VANBOGELEN, R. A. & NEIDHARDT, F. C. 1987. Induction of proteins in response to low temperature in Escherichia coli. *J Bacteriol,* 169**,** 2092-5.

JUVEN-GERSHON, T. & KADONAGA, J. T. 2010. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol,* 339**,** 225-9.

KAMYKOWSKI, D. 1981. Light, temperature and nitrogen as interacting factors affecting diel vertical migrations of dinoflagellates in culture. *J. Plankton. Res. ,* 3**,** 331-344.

KARLSON, D. & IMAI, R. 2003. Conservation of the cold shock domain protein family in plants. *Plant Physiol,* 131**,** 12-5.

KEARSE, M., MOIR, R., WILSON, A., STONES-HAVAS, S., CHEUNG, M., STURROCK, S., BUXTON, S., COOPER, A., MARKOWITZ, S., DURAN, C., THIERER, T., ASHTON, B., MEINTJES, P. & DRUMMOND, A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics,* 28**,** 1647-9.

KELLENBERGER, E. 1988. About the organisation of condensed and decondensed non-eukaryotic DNA and the concept of vegetative DNA (a critical review). *Elsevier,* 29**,** 51-62.

KELLENBERGER, E. & ARNOLD-SCHULZ-GAHMEN, B. 1992. Chromatins of low-protein content: special features of their compaction and condensation. *FEMS Microbiol Lett,* 100**,** 361-70.

KHOCHBIN, S., VERDEL, A., LEMERCIER, C. & SEIGNEURIN-BERNY, D. 2001. Functional significance of histone deacetylase diversity. *Curr Opin Genet Dev,* 11**,** 162-6.

KIM, M. H., SONODA, Y., SASAKI, K., KAMINAKA, H. & IMAI, R. 2013. Interactome analysis reveals versatile functions of Arabidopsis COLD SHOCK DOMAIN PROTEIN 3 in RNA processing within the nucleus and cytoplasm. *Cell Stress Chaperones,* 18**,** 517-25.

KLEENE, K. C. 2018. Y-box proteins combine versatile cold shock domains and arginine-rich motifs (ARMs) for pleiotropic functions in RNA biology. *Biochem J,* 475**,** 2769-2784.

KNAUST R., U. T., LI L., TAYLOR W., HASTINGS J.W. 1998. The circadian rhythm of bioluminescence in Pyrocystis is not due to differences in the amount of luciferase: A comparative study of three bioluminescent marine dinoflagellates. *J. Phycol.,* 34**,** 167-172.

KNIJNENBURG, T. A., WESSELS, L. F. & REINDERS, M. J. 2008. Combinatorial influence of environmental parameters on transcription factor activity. *Bioinformatics,* 24**,** i172-81.

KOHLER, A. & HURT, E. 2007. Exporting RNA from the nucleus to the cytoplasm. *Nat Rev Mol Cell Biol,* 8**,** 761-73.

KONDO, T. & ISHIURA, M. 2000. The circadian clock of cyanobacteria. *Bioessays,* 22**,** 10-5.

KRUEGER, T., FISHER, P. L., BECKER, S., PONTASCH, S., DOVE, S., HOEGH-GULDBERG, O., LEGGAT, W. & DAVY, S. K. 2015. Transcriptomic characterization of the enzymatic antioxidants FeSOD, MnSOD, APX and KatG in the dinoflagellate genus Symbiodinium. *BMC Evol Biol,* 15**,** 48.

KUO, M. H., VOM BAUR, E., STRUHL, K. & ALLIS, C. D. 2000. Gcn4 activator targets Gcn5 histone acetyltransferase to specific promoters independently of transcription. *Mol Cell,* 6**,** 1309-20.

LAJEUNESE, T., LAMBERT, G., ANDERSON, A., COFFROTH, M. A. & GALBRAITH, D. W. 2005. Symbiodinium (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates. *J Phycol,* 41**,** 880-886.

LAJEUNESSE, T. C., PARKINSON, J. E., GABRIELSON, P. W., JEONG, H. J., REIMER, J. D., VOOLSTRA, C. R. & SANTOS, S. R. 2018. Systematic Revision of Symbiodiniaceae Highlights the Antiquity and Diversity of Coral Endosymbionts. *Curr Biol,* 28**,** 2570-2580 e6.

LASHAM, A., MOLONEY, S., HALE, T., HOMER, C., ZHANG, Y. F., MURISON, J. G., BRAITHWAITE, A. W. & WATSON, J. 2003. The Y-box-binding protein, YB1, is a potential negative regulator of the p53 tumor suppressor. *J Biol Chem,* 278**,** 35516-23.

LE, Q. H., MARKOVIC, P., HASTINGS, J. W., JOVINE, R. V. & MORSE, D. 1997. Structure and organization of the peridinin-chlorophyll a-binding protein gene in Gonyaulax polyedra. *Mol Gen Genet,* 255**,** 595-604.

LEE, D. H., MITTAG, M., SCZEKAN, S., MORSE, D. & HASTINGS, J. W. 1993. Molecular cloning and genomic organization of a gene for luciferin-binding protein from the dinoflagellate Gonyaulax polyedra. *J Biol Chem,* 268**,** 8842-50.

LEE, R., LAI, H., MALIK, S. B., SALDARRIAGA, J. F., KEELING, P. J. & SLAMOVITS, C. H. 2014. Analysis of EST data of the marine protist Oxyrrhis marina, an emerging model for alveolate biology and evolution. *BMC Genomics,* 15**,** 122.

LEMONS, D. & MCGINNIS, W. 2006. Genomic evolution of Hox gene clusters. *Science,* 313**,** 1918-22.

LESCURE, A., CARBON, P. & KROL, A. 1991. The different positioning of the proximal sequence element in the Xenopus RNA polymerase II and III snRNA promoters is a key determinant which confers RNA polymerase III specificity. *Nucleic Acids Res,* 19**,** 435-41.

LEVI-SETTI, R., GAVRILOV, K. L. & RIZZO, P. J. 2008. Divalent cation distribution in dinoflagellate chromosomes imaged by high-resolution ion probe mass spectrometry. *Eur J Cell Biol,* 87**,** 963-76.

LEVIN, J. Z., YASSOUR, M., ADICONIS, X., NUSBAUM, C., THOMPSON, D. A., FRIEDMAN, N., GNIRKE, A. & REGEV, A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods,* 7**,** 709-15.

LEVIN, R. A., BELTRAN, V. H., HILL, R., KJELLEBERG, S., MCDOUGALD, D., STEINBERG, P. D. & VAN OPPEN, M. J. 2016. Sex, Scavengers, and Chaperones: Transcriptome Secrets of Divergent Symbiodinium Thermal Tolerances. *Mol Biol Evol,* 33**,** 3032.

LI, L. & HASTINGS, J. W. 1998. The structure and organization of the luciferase gene in the photosynthetic dinoflagellate Gonyaulax polyedra. *Plant Mol Biol,* 36**,** 275-84.

LI, T., YU, L., SONG, B., SONG, Y., LI, L., LIN, X. & LIN, S. 2020. Genome Improvement and Core Gene Set Refinement of Fugacium kawagutii. *Microorganisms,* 8.

LIDIE, K. B. & VAN DOLAH, F. M. 2007. Spliced leader RNA-mediated trans-splicing in a dinoflagellate, Karenia brevis. *J Eukaryot Microbiol,* 54**,** 427-35.

LIN, S. 2011. Genomic understanding of dinoflagellates. *Res Microbiol,* 162**,** 551-69.

LIN, S., CHENG, S., SONG, B., ZHONG, X., LIN, X., LI, W., LI, L., ZHANG, Y., ZHANG, H., JI, Z., CAI, M., ZHUANG, Y., SHI, X., LIN, L., WANG, L., WANG, Z., LIU, X., YU, S., ZENG, P., HAO, H., ZOU, Q., CHEN, C., LI, Y., WANG, Y., XU, C., MENG, S., XU, X., WANG, J., YANG, H., CAMPBELL, D. A., STURM, N. R., DAGENAIS-BELLEFEUILLE, S. & MORSE, D. 2015. The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis. *Science,* 350**,** 691-4.

LIN, S., YU, L. & ZHANG, H. 2019. Transcriptomic Responses to Thermal Stress and Varied Phosphorus Conditions in Fugacium kawagutii. *Microorganisms,* 7.

LIN, S., ZHANG, H., HOU, Y., ZHUANG, Y. & MIRANDA, L. 2009. High-level diversity of dinoflagellates in the natural environment, revealed by assessment of mitochondrial cox1 and cob genes for dinoflagellate DNA barcoding. *Appl Environ Microbiol,* 75**,** 1279-90.

LIN, S., ZHANG, H., ZHUANG, Y., TRAN, B. & GILL, J. 2010. Spliced leader-based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. *Proc Natl Acad Sci U S A,* 107**,** 20033-8.

LIN, X., ZHANG, H., HUANG, B. & LIN, S. 2011. Alkaline Phosphatase Gene Sequence And Transcriptional Regulation By Phosphate Limitation In Amphidinium Carterae (Dinophyceae)(1). *J Phycol,* 47**,** 1110-20.

LIN, X. Z., H.; HUANG, B.; LIN, S. 2012. Alkaline phosphatase gene sequence characteristics and transcriptional regulation by phosphate limitation in Karenia brevis (Dinophyceae) *Harmful Algae,* 17**,** 14-24.

LIU, H., STEPHENS, T. G., GONZALEZ-PECH, R. A., BELTRAN, V. H., LAPEYRE, B., BONGAERTS, P., COOKE, I., ARANDA, M., BOURNE, D. G., FORET, S., MILLER, D. J., VAN OPPEN, M. J. H., VOOLSTRA, C. R., RAGAN, M. A. & CHAN, C. X. 2018. Symbiodinium genomes reveal adaptive evolution of functions related to coral-dinoflagellate symbiosis. *Commun Biol,* 1**,** 95.

LOROS, J. J. & DUNLAP, J. C. 2001. Genetic and molecular analysis of circadian rhythms in Neurospora. *Annu Rev Physiol,* 63**,** 757-94.

LUGER, K., MADER, A. W., RICHMOND, R. K., SARGENT, D. F. & RICHMOND, T. J. 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature,* 389**,** 251-60.

LYKKE-ANDERSEN, S. & JENSEN, T. H. 2007. Overlapping pathways dictate termination of RNA polymerase II transcription. *Biochimie,* 89**,** 1177-82.

MACDONALD, G. H., ITOH-LINDSTROM, Y. & TING, J. P. 1995. The transcriptional regulatory protein, YB-1, promotes single-stranded regions in the DRA promoter. *J Biol Chem,* 270**,** 3527-33.

MACRAE, R. A., FENSOME, R. A. & WILLIAMS, G. L. 1996. Fossil dinoflagellate diversity, originations, and extinctions and their significance. *Can. J. Bot.,* 74**,** 1687- 1694.

MAGNANI, E., SJOLANDER, K. & HAKE, S. 2004. From endonucleases to transcription factors: evolution of the AP2 DNA binding domain in plants. *The Plant Cell,* 16**,** 2265-2277.

MAKDE, R. D., ENGLAND, J. R., YENNAWAR, H. P. & TAN, S. 2010. Structure of RCC1 chromatin factor bound to the nucleosome core particle. *Nature,* 467**,** 562–566.

MAQUAT, L. E. 1995. When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells. *RNA,* 1**,** 453-65.

MARINOV, G. K., TREVINO, A. E., XIANG, T., KUNDAJE, A., GROSSMAN, A. R. & GREENLEAF, W. J. 2020. Transcription-dependent domain-scale 3D genome organization in dinoflagellates. *bioRxiv*.

MARTINEZ, E., GE, H., TAO, Y., YUAN, C. X., PALHAN, V. & ROEDER, R. G. 1998. Novel cofactors and TFIIA mediate functional core promoter selectivity by the human TAFII150-containing TFIID complex. *Mol Cell Biol,* 18**,** 6571-83.

MASTON, G. A., EVANS, S. K. & GREEN, M. R. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet,* 7**,** 29-59.

MATSUMOTO, J., DEWAR, K., WASSERSCHEID, J., WILEY, G. B., MACMIL, S. L., ROE, B. A., ZELLER, R. W., SATOU, Y. & HASTINGS, K. E. 2010. High-throughput sequence analysis of Ciona intestinalis SL trans-spliced mRNAs: alternative expression modes and gene function correlates. *Genome Res,* 20**,** 636-45.

MAYFIELD, A. B., HSIAO, Y. Y., CHEN, H. K. & CHEN, C. S. 2014. Rubisco expression in the dinoflagellate Symbiodinium sp. is influenced by both photoperiod and endosymbiotic lifestyle. *Mar Biotechnol,* 16**,** 371–384.

MCCLUNG, C. R. 2006. Plant circadian rhythms. *Plant Cell,* 18**,** 792-803.

MCGINTY, E. S., PIECZONKA, J. & MYDLARZ, L. D. 2012. Variations in reactive oxygen release and antioxidant activity in multiple Symbiodinium types in response to elevated temperature. *Microb Ecol,* 64**,** 1000-7.

MENDEZ, G. S., DELWICHE, C. F., APT, K. E. & LIPPMEIER, J. C. 2015. Dinoflagellate Gene Structure and Intron Splice Sites in a Genomic Tandem Array. *J Eukaryot Microbiol,* 62**,** 679-87.

MENSINGER, A. F. & CASE, J. F. 1992. Dinoflagellate luminescence increases susceptibility of zooplankton to teleost predictaion. *Mar Biol,* 112**,** 207-210.

MERSFELDER, E. L. & PARTHUN, M. R. 2006. The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. *Nucleic Acids Res,* 34**,** 2653-62.

MIHAILOVICH, M., MILITTI, C., GABALDON, T. & GEBAUER, F. 2010. Eukaryotic cold shock domain proteins: highly versatile regulators of gene expression. *Bioessays,* 32**,** 109-18.

MINGUEZ, A., FRANCA, S. & MORENO DIAZ DE LA ESPINA, S. 1994. Dinoflagellates have a eukaryotic nuclear matrix with lamin-like proteins and topoisomerase II. *J Cell Sci,* 107 (Pt 10)**,** 2861-73.

MIZZEN, C. A., YANG, X. J., KOKUBO, T., BROWNELL, J. E., BANNISTER, A. J., OWEN-HUGHES, T., WORKMAN, J., WANG, L., BERGER, S. L., KOUZARIDES, T., NAKATANI, Y. & ALLIS, C. D. 1996. The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. *Cell,* 87**,** 1261-70.

MOQTADERI, Z., BAI, Y., POON, D., WEIL, P. A. & STRUHL, K. 1996. TBP-associated factors are not generally required for transcriptional activation in yeast. *Nature,* 383**,** 188-91.

MORDOVKINA, D., LYABIN, D. N., SMOLIN, E. A., SOGORINA, E. M., OVCHINNIKOV, L. P. & ELISEEVA, I. 2020. Y-Box Binding Proteins in mRNP Assembly, Translation, and Stability Control. *Biomolecules,* 10.

MOREY, J. S., MONROE, E. A., KINNEY, A. L., BEAL, M., JOHNSON, J. G., HITCHCOCK, G. L. & VAN DOLAH, F. M. 2011. Transcriptomic response of the red tide dinoflagellate, *Karenia brevis,* to nitrogen and phosphorus depletion and addition. *BMC Genomics,* 12**,** 346.

MORSE, D. 2019. A Transcriptome-based Perspective of Meiosis in Dinoflagellates. *Protist,* 170**,** 397-403.

MORSE, D., PAPPENHEIMER, A. M., JR. & HASTINGS, J. W. 1989. Role of a luciferin-binding protein in the circadian bioluminescent reaction of Gonyaulax polyedra. *J Biol Chem,* 264**,** 11822-6.

MOUNT, S. M., BURKS, C., HERTZ, G., STORMO, G. D., WHITE, O. & FIELDS, C. 1992. Splicing signals in Drosophila: intron size, information content, and consensus sequences. *Nucleic Acids Res,* 20**,** 4255-62.

MUSCATINE, L., MCCLOSKEY, L. R. & MARIAN, R. E. 1981. Estimating the daily contribution of carbon from zooxanthellae to coral animal respiration. *Limnol. Oceanogr.,* 26**,** 601-611.

NAAR, A. M., LEMON, B. D. & TJIAN, R. 2001. Transcriptional coactivator complexes. *Annu Rev Biochem,* 70**,** 475-501.

NAKAMINAMI, K., KARLSON, D. T. & IMAI, R. 2006. Functional conservation of cold shock domains in bacteria and higher plants. *Proc Natl Acad Sci U S A,* 103**,** 10122-7.

NARLIKAR, G. J., FAN, H. H. & E. KINGSTON, R. 2002. Cooperation between Complexes that Regulate Chromatin Structure and Transcription. *Cell,* 108**,** 475–487.

NASSOURY, N., CAPPADOCIA, M. & MORSE, D. 2003. Plastid ultrastructure defines the protein import pathway in dinoflagellates. *J Cell Sci,* 116**,** 2867-74.

NICOLAS, M. T. M., D.; BASSOT, J.M.; ET AL. 1991. Colocalization of luciferin binding protein and luciferase to the scintillons ofGonyaulax polyedra revealed by double immunolabeling after fast-freeze fixation. *Protoplasms,* 160**,** 159-166.

NILSSON, D., GUNASEKERA, K., MANI, J., OSTERAS, M., FARINELLI, L., BAERLOCHER, L., RODITI, I. & OCHSENREITER, T. 2010. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei. *PLoS Pathog,* 6**,** e1001037.

OHTSUBO, M., KAI, R., FURUNO, N., SEKIGUCHI, T., SEKIGUCHI, M., HAYASHIDA, H., KUMA, K., MIYATA, T., FUKUSHIGE, S. & MUROTSU, T. 1987. Isolation and characterization of the active cDNA of the human cell cycle gene (RCC1) involved in the regulation of onset of chromosome condensation. *Genes Dev,* 1**,** 585–593.

OKAMOTO, O. K. & HASTINGS, J. W. 2003. Genome-wide analysis of redox-regulated genes in a dinoflagellate. *Gene,* 321**,** 73-81.

OKAMOTO, O. K., LIU, L., ROBERTSON, D. L. & HASTINGS, J. W. 2001. Members of a dinoflagellate luciferase gene family differ in synonymous substitution rates. *Biochemistry,* 40**,** 15862-8.

ORR, R. J., STUKEN, A., MURRAY, S. A. & JAKOBSEN, K. S. 2013. Evolutionary acquisition and loss of saxitoxin biosynthesis in dinoflagellates: the second "core" gene, sxtG. *Appl Environ Microbiol,* 79**,** 2128-36.

ORTIZ-MATAMOROS, M. F. A. V., M.; ISLAS-FLORES, T. 2015. Transient transformation of cultured photosynthetic dinoflagellates (Symbiodinium spp.) with plant-targeted vectors. *Ciencias Marinas,* 41**,** 21-32.

OUELLETTE, M. & PAPADOPOULOU, B. 2009. Coordinated gene expression by post-transcriptional regulons in African trypanosomes. *J Biol,* 8**,** 100.

PALOZOLA, K. C., DONAHUE, G., LIU, H., GRANT, G. R., BECKER, J. S., COTE, A., YU, H., RAJ, A. & ZARET, K. S. 2017. Mitotic transcription and waves of gene reactivation during mitotic exit. *Science,* 358**,** 119-122.

PATRO, R., DUGGAL, G., LOVE, M. I., IRIZARRY, R. A. & KINGSFORD, C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods,* 14**,** 417-419.

PEERS, G., TRUONG, T. B., OSTENDORF, E., ELRAD, D., GROSSMAN, A. R., HIPPLER, M. & NIYOGI, K. K. 2009. An ancient light-harvesting protein is critical for the regulation of algal photosynthesis. *Nature,* 462**,** 518–521.

PEGLAR, M. T., NERAD, T. A., ANDERSON, O. R. & GILLEVET, P. M. 2004. Identification of amoebae implicated in the life cycle of Pfiesteria and Pfiesteria-like dinoflagellates. *J Eukaryot Microbiol,* 51**,** 542-52.

PELTZ, S. W., HE, F., WELCH, E. & JACOBSON, A. 1994. Nonsense-mediated mRNA decay in yeast. *Prog Nucleic Acid Res Mol Biol,* 47**,** 271-98.

PENG, S. E., WANG, Y. B., WANG, L. H., CHEN, W. N., LU, C. Y., FANG, L. S. & CHEN, C. S. 2010. Proteomic analysis of symbiosome membranes in Cnidaria-dinoflagellate endosymbiosis. *Proteomics,* 10**,** 1002-16.

POUCHKINA-STANTCHEVA, N. N. & TUNNACLIFFE, A. 2005. Spliced leader RNA-mediated trans-splicing in phylum Rotifera. *Mol Biol Evol,* 22**,** 1482-9.

PUGH, B. F. & TJIAN, R. 1990. Mechanism of transcriptional activation by Sp1: evidence for coactivators. *Cell,* 61**,** 1187-97.

RAE, P. M. 1976. Hydroxymethyluracil in eukaryote DNA: a natural feature of the pyrrophyta (dinoflagellates). *Science,* 194**,** 1062-4.

RAE, P. M. & STEELE, R. E. 1978. Modified bases in the DNAs of unicellular eukaryotes: an examination of distributions and possible roles, with emphasis on hydroxymethyluracil in dinoflagellates. *Biosystems,* 10**,** 37-53.

REDDY, R., SPECTOR, D., HENNING, D., LIU, M. H. & BUSCH, H. 1983. Isolation and partial characterization of dinoflagellate U1-U6 small RNAs homologous to rat U small nuclear RNAs. *J Biol Chem,* 258**,** 13965-9.

RIECHMANN, J. L., HEARD, J., MARTIN, G., REUBER, L., JIANG, C., KEDDIE, J., ADAM, L., PINEDA, O., RATCLIFFE, O. J., SAMAHA, R. R., CREELMAN, R., PILGRIM, M., BROUN, P., ZHANG, J. Z., GHANDEHARI, D., SHERMAN, B. K. & YU, G. 2000. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science,* 290**,** 2105-10.

RILL, L., STRZELECKA, T., DAVIDSON, M. & VAN WINKLE, D. 1991. Ordered phases in concentrated DNA solutions. *Physica A,* 176**,** 87-116.

RIVKEES, S. A. 2007. The Development of Circadian Rhythms: From Animals To Humans. *Sleep Med Clin,* 2**,** 331-341.

RIZZO, P. J. 1979. RNA synthesis in isolated nuclei of the dinoflagellate Crypthecodinium cohnii. *J Protozool,* 26**,** 290-4.

RIZZO, P. J. 2003. Those amazing dinoflagellate chromosomes. *Cell Res,* 13**,** 215-7.

RIZZO, P. J., JONES, M. & RAY, S. M. 1982. Isolation and properties of isolated nuclei from the Florida red tide dinoflagellate *Gymnodinium breve* (Davis). *J Protozool,* 29**,** 217-22.

ROCKE, D. M., RUAN, L., ZHANG, Y., GOSSETT, J. J., DURBIN-JOHNSON, B. & S., A. 2015. Excess False Positive Rates in Methods for Differential Gene Expression Analysis using RNA-Seq Data. *Biorxiv.*

ROENNEBERG, T., COLFAX, G. N. & HASTINGS, J. W. 1989. A circadian rhythm of population behavior in Gonyaulax polyedra. *J Biol Rhythms,* 4**,** 201-16.

ROENNEBERG, T., KUMAR, C. J. & MERROW, M. 2007. The human circadian clock entrains to sun time. *Curr Biol,* 17**,** R44-5.

ROENNEBERG, T. & REHMAN, J. 1996. Nitrate, a nonphotic signal for the circadian system. *FASEB J,* 10**,** 1443-7.

ROWAN, R., WHITNEY, S. M., FOWLER, A. & YELLOWLEES, D. 1996. Rubisco in marine symbiotic dinoflagellates: form II enzymes in eukaryotic oxygenic phototrophs encoded by a nuclear multigene family. *Plant Cell,* 8**,** 539-53.

ROY, S., BEAUCHEMIN, M., DAGENAIS-BELLEFEUILLE, S., LETOURNEAU, L., CAPPADOCIA, M. & MORSE, D. 2014a. The Lingulodinium circadian system lacks rhythmic changes in transcript abundance. *BMC Biol,* 12**,** 107.

ROY, S., BEAUCHEMIN, M., DAGENAIS-BELLEFEUILLE, S., LETOURNEAU, L., CAPPADOCIA, M. & MORSE, D. 2014b. The *Lingulodinium* circadian system lacks rhythmic changes in transcript abundance. *BMC Biol,* 12**,** 107.

ROY, S., LETOURNEAU, L. & MORSE, D. 2014c. Cold-induced cysts of the photosynthetic dinoflagellate *Lingulodinium polyedrum* have an arrested circadian bioluminescence rhythm and lower levels of protein phosphorylation. *Plant Physiol,* 164**,** 966-77.

ROY, S. & MORSE, D. 2012. A full suite of histone and histone modifying genes are transcribed in the dinoflagellate Lingulodinium. *PLoS One,* 7**,** e34340.

ROY, S. & MORSE, D. 2013. Transcription and maturation of mRNA in dinoflagellates. *Microorganisms,* 1**,** 71-99.

RUPRECHT, V., MONZO, P., RAVASIO, A., YUE, Z., MAKHIJA, E., STRALE, P. O., GAUTHIER, N., SHIVASHANKAR, G. V., STUDER, V., ALBIGES-RIZO, C. & VIASNOFF, V. 2017. How cells respond to environmental cues - insights from bio-functionalized substrates. *J Cell Sci,* 130**,** 51-61.

SALA-ROVIRA, M., GERAUD, M. L., CAPUT, D., JACQUES, F., SOYER-GOBILLARD, M. O., VERNET, G. & HERZOG, M. 1991. Molecular cloning and immunolocalization of two variants of the major basic nuclear protein (HCc) from the histone-less eukaryote Crypthecodinium cohnii (Pyrrhophyta). *Chromosoma,* 100**,** 510-8.

SALCEDO, T., UPADHYAY, R. J., NAGASAKI, K. & BHATTACHARYA, D. 2012. Dozens of toxin-related genes are expressed in a nontoxic strain of the dinoflagellate Heterocapsa circularisquama. *Mol Biol Evol,* 29**,** 1503-6.

SANGERMANO, F., DELICATO, A. & CALABRO, V. 2020. Y box binding protein 1 (YB-1) oncoprotein at the hub of DNA proliferation, damage and cancer progression. *Biochimie,* 179**,** 205-216.

SANTOS, S. R., SHEARER, T. L., HANNES, A. R. & COFFROTH, M. A. 2004. Fine-scale diversity and specificity in the most prevalent lineage of symbiotic dinoflagellates (Symbiodinium, Dinophyceae) of the Caribbean. *Mol Ecol,* 13**,** 459-69.

SCHINDELIN, J., ARGANDA-CARRERAS, I., FRISE, E., KAYNIG, V., LONGAIR, M., PIETZSCH, T., PREIBISCH, S., RUEDEN, C., SAALFELD, S., SCHMID, B., TINEVEZ, J. Y., WHITE, D. J., HARTENSTEIN, V., ELICEIRI, K., TOMANCAK, P. & CARDONA, A. 2012. Fiji: an open-source platform for biological-image analysis. *Nat Methods,* 9**,** 676-82.

SCHMITTER, R. E., NJUS, D., SULZMAN, F. M., GOOCH, V. D. & HASTINGS, J. W. 1976. Dinoflagellate bioluminescence: a comparative study of invitro components. *J Cell Physiol,* 87**,** 123-34.

SCHNEPF, E. E., M. 1999. Dinophyte chloroplasts and phylogeny—a review. *Grana* 38**,** 81–97.

SCHRAMM, L. & HERNANDEZ, N. 2002. Recruitment of RNA polymerase III to its target promoters. *Genes Dev,* 16**,** 2593-620.

SHAMOVSKY, I. & NUDLER, E. 2008. New insights into the mechanism of heat shock response activation. *Cell Mol Life Sci,* 65**,** 855-61.

SHEN, W. C. & GREEN, M. R. 1997. Yeast TAF(II)145 functions as a core promoter selectivity factor, not a general coactivator. *Cell,* 90**,** 615-24.

SHI, X., LI, L., GUO, C., LIN, X., LI, M. & LIN, S. 2015. Rhodopsin gene expression regulated by the light dark cycle, light spectrum and light intensity in the dinoflagellate Prorocentrum. *Front Microbiol,* 6**,** 555.

SHI, X., ZHANG, H. & LIN, S. 2013. Tandem repeats, high copy number and remarkable diel expression rhythm of form II RuBisCO in Prorocentrum donghaiense (Dinophyceae). *PLoS One,* 8**,** e71232.

SHOGUCHI, E., BEEDESSEE, G., TADA, I., HISATA, K., KAWASHIMA, T., TAKEUCHI, T., ARAKAKI, N., FUJIE, M., KOYANAGI, R., ROY, M. C., KAWACHI, M., HIDAKA, M., SATOH, N. & SHINZATO, C. 2018. Two divergent Symbiodinium genomes reveal conservation of a gene cluster for sunscreen biosynthesis and recently lost genes. *BMC Genomics,* 19**,** 458.

SHOGUCHI, E., SHINZATO, C., KAWASHIMA, T., GYOJA, F., MUNGPAKDEE, S., KOYANAGI, R., TAKEUCHI, T., HISATA, K., TANAKA, M., FUJIWARA, M., HAMADA, M., SEIDI, A., FUJIE, M., USAMI, T., GOTO, H., YAMASAKI, S., ARAKAKI, N., SUZUKI, Y., SUGANO, S., TOYODA, A., KUROKI, Y., FUJIYAMA, A., MEDINA, M., COFFROTH, M. A., BHATTACHARYA, D. & SATOH, N. 2013. Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Curr Biol,* 23**,** 1399-408.

SIGEE, D. C. 1983. Structural DNA and genetically active DNA in dinoflagellate chromosomes. *Biosystems,* 16**,** 203-10.

SIMMEN, K. A., WALDSCHMIDT, R., BERNUES, J., PARRY, H. D., SEIFART, K. H. & MATTAJ, I. W. 1992. Proximal sequence element factor binding and species specificity in vertebrate U6 snRNA promoters. *J Mol Biol,* 223**,** 873-84.

SLAMOVITS, C. H. & KEELING, P. J. 2008. Widespread recycling of processed cDNAs in dinoflagellates. *Curr Biol,* 18**,** R550-2.

SMAYDA, T. J. 1997. Harmful algal blooms: Their ecophysiology and general relevance to phytoplankton blooms in the sea. *Limnol. Oceanogr.,* 42**,** 1137-1153.

SOMMERVILLE, J. 1999. Activities of cold-shock domain proteins in translation control. *Bioessays,* 21**,** 319-25.

SONG, B., CHEN, S. & CHEN, W. 2018. Dinoflagellates, a Unique Lineage for Retrogene Research. *Front Microbiol,* 9**,** 1556.

SONG, B., MORSE, D., SONG, Y., FU, Y., LIN, X., WANG, W., CHENG, S., CHEN, W., LIU, X. & LIN, S. 2017. Comparative Genomics Reveals Two Major Bouts of Gene Retroposition Coinciding with Crucial Periods of Symbiodinium Evolution. *Genome Biol Evol,* 9**,** 2037-2047.

SONG, Y., ZAHERI, B., LIU, M., SAHU, S. K., LIU, H., CHEN, W., SONG, B. & MORSE, D. 2019. Fugacium Spliced Leader Genes Identified from Stranded RNA-Seq Datasets. *Microorganisms,* 7.

SOREK, M., YACOBI, Y. Z., ROOPIN, M., BERMAN-FRANK, I. & LEVY, O. 2013. Photosynthetic circadian rhythmicity patterns of Symbiodinium, [corrected] the coral endosymbiotic algae. *Proc Biol Sci,* 280**,** 20122942.

SOREK, M., YACOBI, Y. Z., ROOPIN, M., BERMAN-FRANK, I. & LEVY, O. 2016. Photosynthetic circadian rhythmicity patterns of Symbiodium, the coral endosymbiotic algae. *Proc R Soc B*.

SOYER, M. O. & HAAPALA, O. K. 1974. Electron microscopy of RNA in dinoflagellate chromosomes. *Histochemistry,* 42**,** 239-46.

SPECTOR, D. 1984. *Dinoflagellates,* New York, Academic.

STEELE, R. E., RAE, P. M. 1980. Ordered distribution of modified bases in the DNA of a dinoflagellate. *Nucleic Acids Res,* 8**,** 4709-4725.

STEPHENS, T. G., GONZALEZ-PECH, R. A., CHENG, Y., MOHAMED, A. R., BURT, D. W., BHATTACHARYA, D., RAGAN, M. A. & CHAN, C. X. 2020. Genomes of the dinoflagellate Polarella glacialis encode tandemly repeated single-exon genes with adaptive functions. *BMC Biol,* 18**,** 56.

STOVER, N. A. & STEELE, R. E. 2001. Trans-spliced leader addition to mRNAs in a cnidarian. *Proc Natl Acad Sci U S A,* 98**,** 5693-8.

SUGGETT, D. J., WARNER, M. E., SMITH, D. J., DAVEY, P., HENNIGE, S. & BAKER, N. R. 2008. Photosynthesis and Production of Hydrogen Peroxide by Symbiodinium (Pyrrhophyta) Phylotypes with Different Thermal Tolerances(1). *J Phycol,* 44**,** 948-56.

SUNTHARALINGAM, M. & WENTE, S. R. 2003. Peering through the pore: nuclear pore complex structure, assembly, and function. *Dev Cell,* 4**,** 775-89.

SUTTON, R. E. & BOOTHROYD, J. C. 1986. Evidence for trans splicing in trypanosomes. *Cell,* 47**,** 527-35.

TANESE, N., PUGH, B. F. & TJIAN, R. 1991. Coactivators for a proline-rich activator purified from the multisubunit human TFIID complex. *Genes Dev,* 5**,** 2212-24.

TAYLOR, F. J., HOPPENRATH, M. & F. SALDARRIAGA, J. 2008. Dinoflagellate diversity and distribution. *Biodiversity and Conservation,* 17**,** 407-418.

TEN LOHUIS, M. R. & MILLER, D. J. 1998. Light-regulated transcription of genes encoding peridinin chlorophyll a proteins and the major intrinsic light-harvesting complex proteins in the dinoflagellate amphidinium carterae hulburt (Dinophycae). Changes In cytosine methylation accompany photoadaptation. *Plant Physiol,* 117**,** 189-96.

TENGS, T., DAHLBERG, O. J., SHALCHIAN-TABRIZI, K., KLAVENESS, D., RUDI, K., DELWICHE, C. F. & JAKOBSEN, K. S. 2000. Phylogenetic analyses indicate that the 19'Hexanoyloxy-fucoxanthin-containing dinoflagellates have tertiary plastids of haptophyte origin. *Mol Biol Evol,* 17**,** 718-29.

UCHIDA, T. 2001. The role of cell contact in the life cycle of some dinoflagellate species. *J. Plankton Res. ,* 23**,** 889-891.

VANDENBERGHE, A. E., MEEDEL, T. H. & HASTINGS, K. E. 2001. mRNA 5'-leader trans-splicing in the chordates. *Genes Dev,* 15**,** 294-303.

VASU, S. K. & FORBES, D. J. 2001. Nuclear pores and nuclear assembly. *Curr Opin Cell Biol,* 13**,** 363-75.

VERNET, G., SALA-ROVIRA, M., MAEDER, M., JACQUES, F. & HERZOG, M. 1990. Basic nuclear proteins of the histone-less eukaryote Crypthecodinium cohnii (Pyrrhophyta): two-dimensional electrophoresis and DNA-binding properties. *Biochim Biophys Acta,* 1048**,** 281-9.

VERRIJZER, C. P. & TJIAN, R. 1996. TAFs mediate transcriptional activation and promoter selectivity. *Trends Biochem Sci,* 21**,** 338-42.

VIGNALI, M., HASSAN, A. H., NEELY, K. E. & WORKMAN, J. L. 2000. ATP-dependent chromatin-remodeling complexes. *Mol Cell Biol,* 20**,** 1899-910.

WAHL, M. C., WILL, C. L. & LUHRMANN, R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell,* 136**,** 701-18.

WANG, Y., JENSEN, L., HOJRUP, P. & MORSE, D. 2005. Synthesis and degradation of dinoflagellate plastid-encoded psbA proteins are light-regulated, not circadian-regulated. *Proc Natl Acad Sci U S A,* 102**,** 2844-9.

WANG, Y., JOLY, S. & MORSE, D. 2008. Phylogeny of dinoflagellate plastid genes recently transferred to the nucleus supports a common ancestry with red algal plastid genes. *J Mol Evol,* 66**,** 175-84.

WANG, Y. & MORSE, D. 2006. Rampant polyuridylylation of plastid gene transcripts in the dinoflagellate Lingulodinium. *Nucleic Acids Res,* 34**,** 613-9.

WEI, W. J., MU, S. R., HEINER, M., FU, X., CAO, L. J., GONG, X. F., BINDEREIF, A. & HUI, J. 2012. YB-1 binds to CAUC motifs and stimulates exon inclusion by enhancing the recruitment of U2AF to weak polypyrimidine tracts. *Nucleic Acids Res,* 40**,** 8622-36.

WHITE, H. H. 1979. Effects of dinoflagellate bioluminescence on the ingestion rates of herbivorous zooplankton. *Journal of Experimental Marine Biology and Ecology,* 36**,** 217-224.

WHITESIDE, S. T. & GOODBOURN, S. 1993. Signal transduction and nuclear targeting: regulation of transcription factor activity by subcellular localisation. *J Cell Sci,* 104 (Pt 4)**,** 949-55.

WILL, C. L. & LUHRMANN, R. 2011. Spliceosome structure and function. *Cold Spring Harb Perspect Biol,* 3.

WISECAVER, J. H. & HACKETT, J. D. 2011. Dinoflagellate genome evolution. *Annu Rev Microbiol,* 65**,** 369-87.

WISTOW, G. 1990. Cold shock and DNA binding. *Nature,* 344**,** 823-4.

WONG J.T., K. A. C. 2005. Proliferation of dinoflagellates: Blooming or bleaching. *Bioessays,* 27**,** 730-740.

WONG, J. T., NEW, D. C., WONG, J. C. & HUNG, V. K. 2003. Histone-like proteins of the dinoflagellate Crypthecodinium cohnii have homologies to bacterial DNA-binding proteins. *Eukaryot Cell,* 2**,** 646-50.

XIA, B., KE, H. & INOUYE, M. 2001. Acquirement of cold sensitivity by quadruple deletion of the cspA family and its suppression by PNPase S1 domain in Escherichia coli. *Mol Microbiol,* 40**,** 179-88.

XIANG, T., NELSON, W., RODRIGUEZ, J., TOLLETER, D. & GROSSMAN, A. R. 2015. Symbiodinium transcriptome and global responses of cells to immediate changes in light intensity when grown under autotrophic or mixotrophic conditions. *The Plant Journal,* 82**,** 67–80.

YANG I., B. S., TILLMANN U., CEMBELLA A., JOHN U 2011. Growth- and nutrient-dependent gene expression in the toxigenic marine dinoflagellate Alexandrium minutum. *Harmful Algae,* 12**,** 55-69.

YANG, X. J., ZHU, H., MU, S. R., WEI, W. J., YUAN, X., WANG, M., LIU, Y., HUI, J. & HUANG, Y. 2019. Crystal structure of a Y-box binding protein 1 (YB-1)-RNA complex reveals key features and residues interacting with RNA. *J Biol Chem,* 294**,** 10998-11010.

YU, L., LI, T., LI, L., LIN, X., LI, H., LIU, C., GUO, C. & LIN, S. 2020. SAGER: a database of Symbiodiniaceae and Algal Genomic Resource. *Database (Oxford),* 2020.

ZAHERI, B., DAGENAIS-BELLEFEUILLE, S., SONG, B. & MORSE, D. 2019. Assessing Transcriptional Responses to Light by the Dinoflagellate Symbiodinium. *Microorganisms,* 7.

ZHANG, H., CAMPBELL, D. A., STURM, N. R. & LIN, S. 2009. Dinoflagellate spliced leader RNA genes display a variety of sequences and genomic arrangements. *Mol Biol Evol,* 26**,** 1757-71.

ZHANG, H., HOU, Y., MIRANDA, L., CAMPBELL, D. A., STURM, N. R., GAASTERLAND, T. & LIN, S. 2007. Spliced leader RNA trans-splicing in dinoflagellates. *Proc Natl Acad Sci U S A,* 104**,** 4618-23.

ZHANG, H., ZHUANG, Y., GILL, J. & LIN, S. 2013. Proof that dinoflagellate spliced leader (DinoSL) is a useful hook for fishing dinoflagellate transcripts from mixed microbial samples: Symbiodinium kawagutii as a case study. *Protist,* 164**,** 510-27.

ZHANG, H. M., CHEN, H., LIU, W., LIU, H., GONG, J., WANG, H. & GUO, A. Y. 2012. AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res,* 40**,** D144-9.

ZHANG, J., SUN, X., QIAN, Y., LADUCA, J. P. & MAQUAT, L. E. 1998. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Mol Cell Biol,* 18**,** 5272-83.

ZHANG, Z., GREEN, B. R. & CAVALIER-SMITH, T. 1999. Single gene circles in dinoflagellate chloroplast genomes. *Nature,* 400**,** 155-9.

ZHENG, Y., ZHAO, L., GAO, J. & FEI, Z. 2011. iAssembler: a package for de novo assembly of Roche-454/Sanger transcriptome sequences. *BMC Bioinformatics,* 12**,** 453.

ZHOU, Q., BOYER, T. G. & BERK, A. J. 1993. Factors (TAFs) required for activated transcription interact with TATA box-binding protein conserved core domain. *Genes Dev,* 7**,** 180-7.

Appendix

Fugacium Spliced Leader Genes Identified from Stranded RNA-Seq Datasets

Yue Song [2], Bahareh Zaheri [4], Min Liu [3], Sunil Kumar Sahu [3], Huan Liu [3], Wenbin Chen [3], Bo Song [1] and David Morse [4]

Published in Microorganisms, 2019, *7*, 171

doi: 10.3390/microorganisms7060171

I prepared all RNA samples for this article.

Abstract

Trans-splicing mechanisms have been documented in many lineages that are widely distributed phylogenetically, including dinoflagellates. The spliced leader (SL) sequence itself is conserved in dinoflagellates, although its gene sequences and arrangements have diversified within or across different species. In this study, we present 18 *Fugacium kawagutii* SL genes identified from stranded RNA-seq reads. These genes typically have a single SL but can contain several partial SLs with lengths ranging from 103 to 292 bp. Unexpectedly, we find the SL gene transcripts contain sequences upstream of the canonical SL, suggesting that generation of mature transcripts will require additional modifications following *trans*-splicing. We have also identified 13 SL-like genes whose expression levels and length are comparable to Dino-SL genes. Lastly, introns in these genes were identified and a new site for Sm-protein binding was proposed. Overall, this study provides a strategy for fast identification of SL genes and identifies new sequences of *F. kawagutii* SL genes to supplement our understanding of *trans*-splicing.

Keywords: dinoflagellates; *Symbiodinium*; *Fugacium*; trans-splicing; spliced leader

Introduction

Dinoflagellates are a large group of ecologically important unicellular algae. Many members in this lineage play critical roles in marine ecosystems as primary producers, contributors to red tides, and symbionts of reef corals and other invertebrates. Dinoflagellates are also known for their distinct genomic features which include large genome sizes, permanently condensed chromosomes, lack of nucleosomes (Lin, 2011). Furthermore, the maturation of mRNAs in dinoflagellates has been proposed to require *trans*-splicing of a spliced leader (SL) sequence (Zhang et al., 2007) (Figure 1A).

The mechanism of *trans*-splicing had been reported in many other lineages including nematodes, flatworms, cnidarians, rotifers, chordates, and euglenozoans. The sequences of the spliced leader are conserved within each lineage but varies among different groups. Different roles of *trans*-splicing have been proposed, which include translation regulation (enhancing or blocking), mRNA stabilization, 5′UTR sanitization, protein retargeting, as well as creating or destroying upstream open reading frames (Hastings, 2005, Matsumoto et al., 2010, Nilsson et al., 2010). Several lineages have more than one consensus SL sequence; for example, euglenozoans can have as many as 14 SL sequences (Bitar et al., 2013). On the other hand, dinoflagellates have only one, (DCCGUAGCCAUUUUGGCUCAAG (D = U, A or G) (Zhang et al., 2007). Attempts to clone SL genes have been made using *Prorocentrum minimum*, *Karenia brevis*, *Polarella glacialis*, *Heterocapsa arctica*, *Karlodinium veneficum*, and *Pfiesteria piscicida* (Zhang et al., 2007, Zhang et al., 2009). Despite the conservation of SL within this phylum, the arrangements of SL

RNA genes are rather diverse. In these organisms, SL genes were clustered alone or mixed with 5S rRNA. The introns in these genes also showed substantial differences.

The presence of SL on the 5′ end of all mature mRNAs has greatly advanced research on dinoflagellates, particularly for Symbiodiniaceae, a group of symbiotic dinoflagellates, by facilitating the identification of dinoflagellate transcripts from mixed samples (Zhang et al., 2013) and identification of retrogenes in the transcriptome (Slamovits and Keeling, 2008, Jaeckisch et al., 2011, Lee et al., 2014) and the genome (Song et al., 2017). However, the SL RNA genes themselves had still not been cloned in Symbiodiniaceae even though several genome assemblies had been released (Shoguchi et al., 2013, Lin et al., 2015, Aranda et al., 2016, Liu et al., 2018). Previous attempts of SL RNA gene cloning relied on polymerase chain reaction with a SL-derived forward primer, which limited the identification of sequences upstream of SLs. High-throughput sequencing of cDNA libraries provided an alternative approach for the identification of SL transcripts, but without strand information there would be many mistakes and uncertainties. This problem can be solved by strand-specific sequencing, in which the 5′ terminus of transcripts will be certainly found in the forward or reverse reads depending on the strategy of library construction (Levin et al., 2010).

In this study, we identified 18 SL and 13 SL-like genes from stranded RNA-seq reads, and found extra sequences upstream of SLs. The SL genes identified in this study are generally longer than those previously reported. Introns from these genes are also identified and a new potential site of Sm-protein binding is proposed.

Materials and Methods

Genomic and Transcriptomic Data

The datasets of stranded RNA-seq reads of *F. kawagutii* (previously called *Symbiodinium kawagutii*) (LaJeunesse et al., 2018) were downloaded from NCBI under the accessions of SRP182908 and SRP119222 (de Mendoza et al., 2018). Previously unpublished transcriptome sequences were prepared from TRIzol-purified RNA samples taken at two times during a 12:12 light/dark cycle (lights on and lights off). Quality control, TruSeq stranded mRNA sample preparation (including poly(A) RNA purification), and Illumina sequencing using a HiSeq 4000 was performed at the McGill University and Genome Quebec Innovation Centre (Montreal, Canada). The draft genome sequences of *F. kawagutii* (Lin et al., 2015) were also downloaded from NCBI.

Clustering and Assembly of SL-Containing Reads

SL-containing reads were identified by searching for SL sequences (CCGTAGCCATTTTGGCTCAAG) in the reverse reads (R2). Clustering and assembly were performed by iAssembler (version v1.3.3) software (Zheng et al., 2011) with a minimum overlap of length >30 bp and identity >95%.

Identification of Sm-Protein Binding Sites and Structural Analysis of SL Genes

Five U-rich motifs in introns of Dino-SL genes were selected and their appearances in the introns of *F. kawagutii* genes were counted and normalized to their expected appearance as a random occurrence (1/n^4, where n is the length of motif).

The secondary structure of SL genes was simulated using MFOLD online service (http://unafold.rna.albany.edu/?q=mfold) with the folding temperature set at 20 °C (Zhang et al., 2009).

Results

Searches for SL-Containing Reads

We identified SL-containing reads by searching for SL sequences in the reverse reads (R2) in the stranded RNA-seq datasets. In total, we obtained 195 SL-containing reads from 8 libraries constructed for this study and from a previously published dataset (de Mendoza et al., 2018). Among these, 154, 31, and 10 contained one, two, and three units of SL, respectively. We also found several SL-containing genes that also contained SL relicts (AGCCATTTTGGCTCAAG) (Supplementary Figure S1). The sequences were clustered into 18 groups according to their similarities, and their consensus sequences were obtained by aligning and assembling the sequences in each group (Supplementary Table S1). These sequences are thus likely to be derived from SL genes. There are 16 sequences with a single copy of the SL, one with 2 SL units and one with 3 SL units (Figure 2; Supplementary Table S1). Remarkably, we note the presence of sequence upstream from the canonical SL which, after *trans*-splicing, constitutes the 5′ end of all dinoflagellate mRNA, suggesting extra steps are required to remove these sequences before mRNA maturation (Figure 1B).

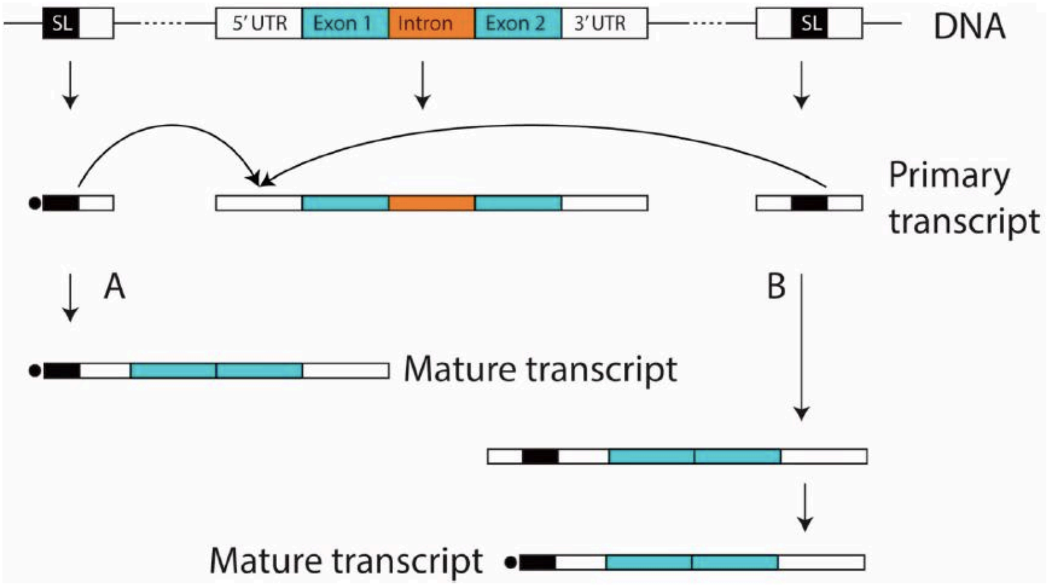


Figure 1. Spliced leader (SL) *trans*-splicing mechanism in dinoflagellates. (A) the mechanism previously thought; (B) a mechanism proposed based on the findings reported in this study, in which extra steps are needed to remove the sequences upstream from the SL before mRNA maturation

Tandem-SL Genes

Among the reads containing both SL and multiple SL relicts (AGCCATTTTGGCTCAAG), 31 have two SLs and 10 have three SLs. Interestingly, reads containing two SLs clustered into one group while reads with three SLs clustered in another (Supplementary Table S1). We noticed that unlike the sequence diversity observed downstream of SL in reads with only a single SL (Supplementary Figure S1), the 3′ downstream sequences in these reads were conserved (Figure 2). These similar/identical reads were not the products of PCR duplicates in the libraries because duplicates had already been removed before the analysis. Moreover, these reads were found in different libraries, both in this study and in a previously reported work (de Mendoza et al., 2018). One possible explanation is that they were from highly expressed genes, which would have a greater chance of being modified by addition of multiple SLs (Song et al., 2017, Song et al., 2018). However, we excluded this possibility because they were also found in the genome assembly of *F. kawagutii*. Therefore, these different consensus sequences represent different loci of tandem SL.

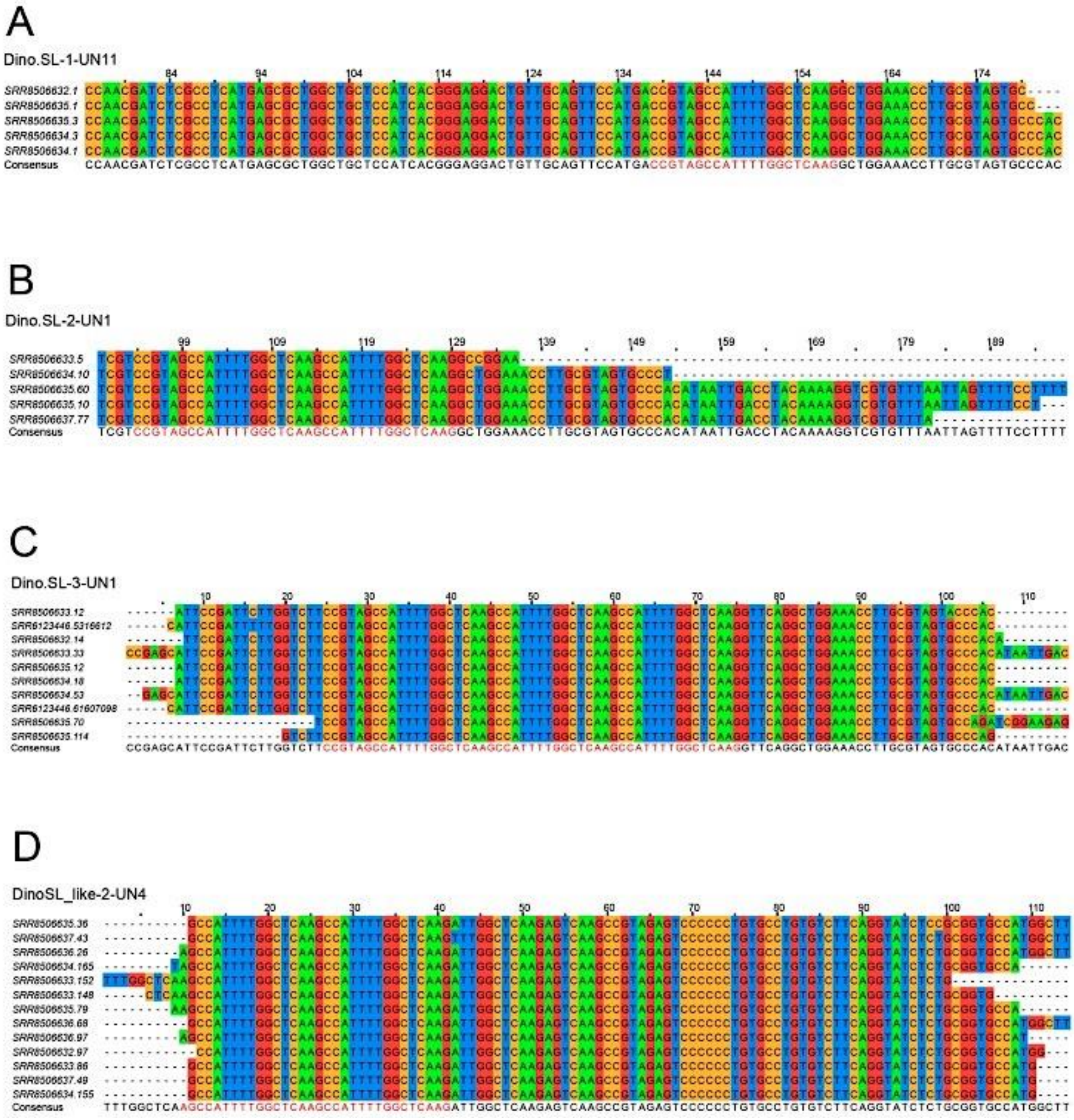


Figure 2. Alignment and assembly of SL-containing reads. The consensus sequence is shown at the bottom of each alignment. Examples include (**A**) single SL genes, and tandem SL genes with (**B**) two and (**C**) three units, and (**D**) an example of SL-like genes. The sequences of SL and its relicts were colored in red in the consensus sequences. More examples can be found in the Supplementary Figure S1.

Introns in SL Genes

We identified 18 types of introns from the different SL loci. In the loci of single SL genes, 16 different sequences (Supplementary Table S1) were found after the 3′ end of SL (AAG). In the loci of tandem SL, two types of different sequences downstream of SL were found (Supplementary Table S1 and Figure S2). All these downstream sequences started with GU or GC, the motifs characteristic of the 5′ end of dinoflagellate introns (Shoguchi et al., 2013).

Sm-protein binding sites, which always have oligo(U) motifs, are conserved in different organisms (Pouchkina-Stantcheva and Tunnacliffe, 2005, Stover and Steele, 2001, Vandenberghe et al., 2001) , and oligo(U) presence constitutes evidence of an intron. We have analyzed the sequences of these introns to identify the potential Sm-binding sites. Previous works proposed that the Sm-protein binding site was the AUUUUGG located in the SL exon, instead of being found in the intron (Zhang et al., 2007). However, this seems unlikely given that Sm-protein binding sites are usually located in introns—indeed, there is no precedent as yet for Sm-protein binding to exons. We found reads with several U-rich motifs (CUUUUG, GUUUA, GUUUUC, GUUUA, GUUUUA, and UUUAA) in the introns. None of these motifs was identical to the known Sm-protein binding sites. These results suggest that the Sm-protein binding sites in dinoflagellates may be different from any of the sites known in other organisms. We further counted the appearances of these motifs in the introns of *F. kawagutii* genes and found that CUUUG, CUUUUG, and GUUUUC are more frequent in introns (Figure 3A). We then examined their location in the predicted RNA structure of the SL gene transcripts. Interestingly, despite differences in sequences and length, the different structures share some conserved features. In particular, the SL sequence is found in the stalk of a "Y" shape structure formed from three stem–loops (Figure 3B,C). A bulge, a feature of Sm-protein binding sites, appeared at a GUUUUC motif in the fork of the Y in the SL transcript. Therefore, we propose GUUUUC as a potential Sm-protein binding site in *F. kawagutii*.

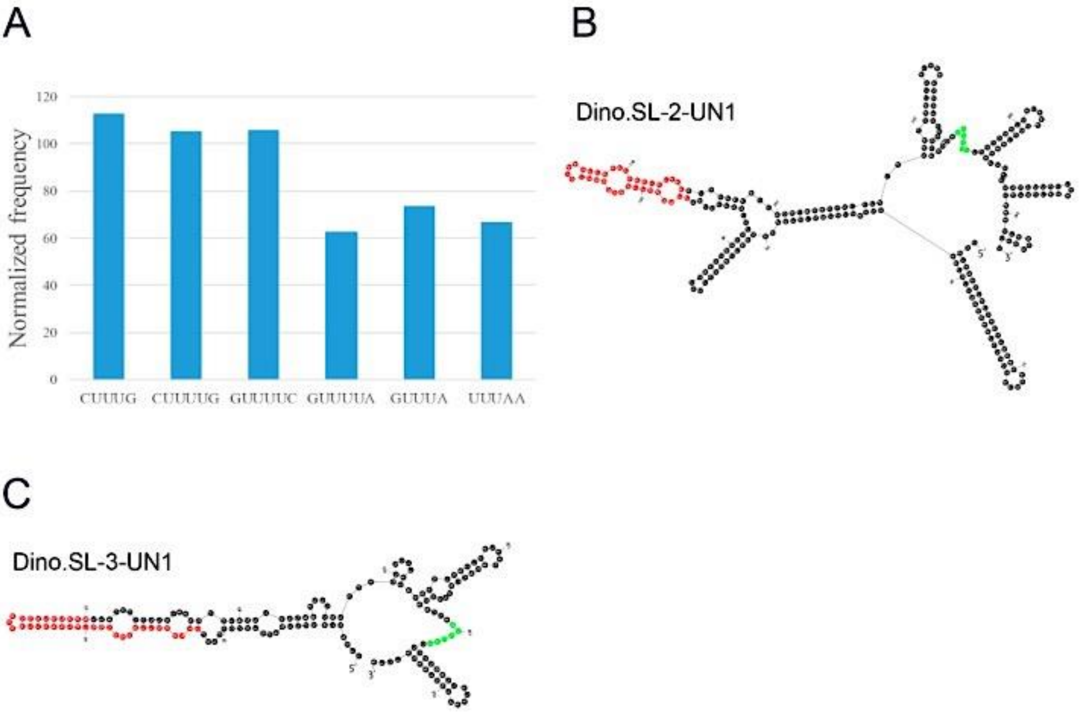


Figure 3. Potential Sm-protein binding sites. (**A**) The appearances of U-rich motifs in introns of *F. kawagutii* genes; the secondary structure of SL genes, (**B**) Dino.SL-2-UN1, (**C**) Dino.SL-3-UN1. The proposed Sm-protein sites (GUUUUC) were colored in green and the sequences of SL were colored in red in (**B**,**C**). The sequences of these SL genes can be found in the Supplementary Table S1.

SL-like Genes

Besides these SL loci with single or tandem SLs, we also found several reads bearing multiple repeats of SL relicts (AGCCATTTTGGCTCAAG). For example, there was a transcript having three tandem SL relicts but from which no bone fide SL could be recovered. Aligning and assembling of these sequences resulted in 13 consensus sequences (Supplementary Table S1). We inspected these sequences closely and found a minor difference at the start of the first unit of SL relict, in which the TCCG of a canonical SL was replaced by TCG (Figure 2). This is unlikely to be a sequence error because these sequences were repeatedly found in 8 different independently constructed libraries. We also confirmed that these reads corresponded to sequences found in the genome. Although the sequences at these loci are different from loci containing SL, the relict sequences were also transcribed at a higher level (17 reads/million) than the authentic SL gene sequences (4.6 reads/million). It is possible that these SL-like loci are located near or at the same cluster of SL loci and shared the same mechanism of transcription and regulation. However, they may not be functional because, unlike the identified SL loci, intron donor sites ($G^{U}/_{C}$) and putative Sm-protein binding sites (GUUUUC) were absent from the sequences downstream of tandem SLs (Figure 2; Supplementary Table S1). This suggests that SL or its relicts may not be able to be spliced from these loci.

Discussion

Despite being found in numerous species, the evolution of *trans*-splicing machinery and SL genes are enigmatic. In dinoflagellates, several SL genes have been cloned in species ranging from *Polarella* and *Heterocapsa* to *Prorocentrum* and *Pfiesteria* (Zhang et al., 2007, Zhang et al., 2009). As shown by these reported SL genes, the sequences and genomic organizations of SL genes are diverse (Zhang et al., 2009). Therefore, more information about these SL genes—particularly those in species other than those aforementioned—is needed to further our understanding of the character and evolution of these genes. Indeed, the *F. kawagutii* SL genes identified in this study displayed several features different from those previously reported.

The Lengths of SL Genes in *F. kawagutii* Are Longer

The lengths of SL genes identified in this study range from 103 to 292 bp, with an average of 164 bp (Supplementary Table S1), which is remarkably longer than those reported in *P. minimum*, *K. brevis*, *P. glacialis*, *H. arctica*, *K. veneficum*, and *P. piscicida*, which are predominantly 50–60 bp in length (Zhang et al., 2009). In previous studies (Zhang et al., 2007, Zhang et al., 2009) , SL genes had been cloned using a 3′ RACE strategy with SL-derived forward primers, assuming SL was the 5′ terminus of each unit. As a consequence, the sequences upstream of SL were undetectable. Therefore, the lengths of SL genes may have been underestimated in previous studies. We also remeasured the length of *F. kawagutii* SL genes excluding these upstream sequences. However, the lengths of *F. kawagutii* SL gene averaged to 89 bp, which is still longer than those reported. As these *F. kawagutii* SL genes were identified using single-ended or pair-ended stranded reads from libraries with insertion sizes of ~200 bp (de Mendoza et al., 2018), the lengths of these SL genes may have also been underestimated. Transcript assembling may partially mitigate this problem but may also possibly introduce errors by, for example, linking reads derived from different loci.

The Fate of the Upstream Sequences

It has been generally assumed that the canonical SL sequence was located at the 5′ end of the transcript derived from the SL gene. Since these transcripts were also thought to be capped at their 5′ end, the *trans*-splicing mechanism was thus responsible for providing the cap structure on all other transcripts. However, the presence of sequence upstream of the SL as shown here suggests the generation of mature transcripts may be more complicated than previously thought (Figure 1B). The sequences immediately upstream of the SL do not show any conserved features in the different versions of the SL gene transcript, yet these must clearly be removed as no transcripts have yet been detected in any dinoflagellate transcriptome to date with sequence upstream from the SL. Lastly, a mechanism must have evolved for capping the transcripts after the excess sequences have been removed. These upstream sequences were not included in the previously cloned SL genes because they were cloned using 3′ RACE strategy with SL-derived forward primers; SL was assumed to be the 5′ terminus of each unit during design of this cloning strategy (Zhang et al., 2007, Zhang et al., 2009).

Novel Sm-Protein Binding Sites

The presence of Sm-protein binding sites, which are usually found in introns of genes in various organisms, is in fact a criterion used for the determination of intronic sequences of genes. Sm-protein binding sites were also found in introns of various SL genes in different organisms (Bitar et al., 2013). Dinoflagellates constitute an exception in that canonical Sm-protein binding sites (AUUUUGG) were found in the exons of SL genes (Zhang et al., 2007, Zhang et al., 2009). Given that the Sm-protein binding sites would normally be spliced off, blocking further Sm protein binding, its presence in exons is puzzling. In this work, we found a U-rich motif, GUUUUC, which was found in the introns of the identified SL genes in *F. kawagutii*. We have therefore proposed this sequence—which is different from the one proposed in other studies (Zhang et al., 2007, Zhang et al., 2009)—as a potential Sm-protein binding site. One possible reason for the differences may lie in the fact that the species studied here (*F. kawagutii*) was different from those used in previous studies (*P. minimum*, *K. brevis*, *P. glacialis*, *H. arctica*, *K. veneficum*, and *P. piscicida*). Another explanation could be the different strategies used for SL gene cloning in different works. In previous studies, SL genes were cloned using RACE techniques from non-poly(A) RNA, which after removal of poly(A) containing transcripts, were ligated with oligo(A) before reverse transcription (Zhang et al., 2007, Zhang et al., 2009). During this procedure, the oligo(A) tail might have been ligated with trimmed or fragmented transcripts leading to cloning of partial length SL genes. The 3′ regions bearing the true Sm-protein binding sites might thus have been missed if only the 5′ part of the SL genes were cloned. Lastly, different fractions of transcripts were selected for analysis in this compared to previous studies. SL genes cloned using 3′ RACE were derived from non-poly(A) transcripts while those identified from stranded RNA datasets in this study were derived from polyadenylated transcripts. However, we cannot know with certainty that the transcripts sequenced in this study were polyadenylated, as some residual fraction of unmodified RNA may still be present after the poly(A) purification step.

SL-like Genes in *F. kawagutii*

We also identified 13 SL-like genes in *F. kawagutii*. All of them have multiple full or partial SLs. This is due to the fact that candidates with only one SL or its relict were removed to exclude the possibility of false discoveries caused by random appearance of the SL sequence. This would thus lead to a failure to identify any SL-like genes with only one SL unit. These genes recovered are very similar in sequences to the real SL genes, but differ in that the donor sites for intron splicing are lost. Furthermore, they also lack Sm-protein binding sites. We speculate that these genes cannot be accurately spliced and represent pseudo-SL genes. According to a rough estimation of their expression levels based on their read counts in the different libraries, the expression levels of these SL-like genes are comparable to that of real SL genes.

Tandem-SL Genes in *F. kawagutii*

Multiple tandem SLs or their relicts had been found in transcripts of many genes of various organisms ranging from *Perkinsus marinus* and *Oxyrrhis marina* to *Alexandrium tamarense* (Slamovits and Keeling, 2008, Jaeckisch et al., 2011). The appearance of relicts was interpreted as an accumulation resulted from multiple rounds of *trans*-splicing during the evolution of dinoflagellates. It is interesting that all SL relicts found in these transcripts start with “CCA”—they are thus missing the upstream CCGTAG of SL. One possibility to account for this is that CCGTAG may have been spliced off, given that “AG” could provide a site for intron splicing. However, the SL relicts identified in genomes are longer and contain CCGTAG. The sequence of the tandem partial SLs in the SL genes suggests an alternative interpretation of their appearance in transcripts: there was only one *trans*-splicing event derived from a tandem-SL gene, instead of multiple rounds of *trans*-splicing from single copy SL transcripts.

Overall, we identified 18 SL genes including 16 single and 2 multiple-SL genes, as well as 13 multiple SL-like genes which have degenerated due to the loss of intron splicing donors. This study provides important supplements to our knowledge of SL genes and illustrates how SL genes can be identified from stranded RNA-seq datasets. The number of SL genes identified in this work is rather limited and presumably only represents a small fraction of the SL genes in *F. kawagutii* genome, at least in part because the datasets used were obtained from polyadenylated transcripts which are likely to be underrepresented in SL genes. Nevertheless, the success identification of SL genes suggests stranded RNA sequencing is a feasible and efficient approach for SL gene identification.