INVESTIGATING THE GENOMIC COMPOSITION OF *OXYRRHIS MARINA,* AN EARLY DIVERGING DINOFLAGELLATE

by

Ronie Haro

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
April 2024

Dalhousie University is located in Mi'kma'ki, the
ancestral and unceded territory of the Mi'kmaq.
We are all Treaty people.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Dinoflagellates are microbial eukaryotes with unorthodox nuclear and mitochondrial genome configurations. They have giant nuclear genomes inflated in repeat elements and redundant gene copies, including retrogenes derived from the retrotransposition of processed mRNAs. Conversely, dinoflagellates exhibit significantly reduced, fragmented, and gene-impoverished mitochondrial genomes. However, due to limited genome sequencing, a comprehensive understanding of both nuclear and mitochondrial genomes is still elusive, with most knowledge derived from closely related sequenced symbiotic dinoflagellates. This thesis aims to address some of these gaps by sequencing and examining the genome of *Oxyrrhis marina*, a representative of an early diverging dinoflagellate, to investigate the organization of both nuclear and mitochondrial genomes. Simultaneously, the transcriptomes of 50 dinoflagellates species were analyzed to explore the prevalence of retrogenes and their functional implications. The *O. marina* nuclear genome survey yielded approximately 22% completeness and revealed around 40% repeat content, showing an expansion of long terminal repeat (LTR) retrotransposons and other repeats. Notable findings included genes organized in tandem arrays and a prevalence for unidirectional gene orientation. Endogenized viral elements were found in the *O. marina genome*, exhibiting characteristics indicative of a transposable element lifestyle. Furthermore, this research uncovered three mitochondrial chromosomes in *O. marina*, showing multiple gene copies and novel arrangements previously undescribed. Finally, the retrogene functional diversity was found to reflect the most common and active processes of the dinoflagellate cell, such as post-translational modification, cell signalling and transport. In summary, this thesis offers additional insights into the mechanisms associated with gene redundancy and provides a deeper understanding of the general mitochondrial and nuclear genome organization of *O. marina*, potentially representing the broader genome diversity of dinoflagellates.

LIST OF ABBREVIATIONS USED

BLAST	Basic local alignment search tool
BUSCO	Benchmarking universal single-copy orthologs
cob	Cytochrome B
cox1	Cytochrome C oxidase subunit I
cox3	Cytochrome C oxidase subunit III
DinoRL	Dinoflagellate spliced leader RNA relic
DinoSL	Dinoflagellate spliced leader RNA
DVNP	Dinoflagellate viral nucleoproteins
EST	Expressed sequencing tag
EVEs	Endogenous viral elements
Gbp	Gigabase pair
HGT	Horizontal gene transfer
HJR	Hollyday junction resolvase
HLP	Histone-like protein
HMW	High molecular weight DNA
Kbp	Kilobase pair
LINE	Long interspersed nuclear element
LSU	Large ribosomal RNA subunit
LTR	Long terminal repeat retrotransposon
Mbp	Megabse pair
MCP	Major capsid protein
mCP	Minor capsid protein
Mitogenome	Mitochondrial genome
MP	Mavericks/Polinton
mRNA	Messenger ribonucleic acid
mtDNA	Mitochondrial DNA molecule
NCBI	National Center for Biotechnology Information
NCLDV	Nucleocytoplasmic viruses
PLV	Polinton-like virus
pPolB	Protein-primed DNA polymerase B

PRO	Protease
rRNA	Ribosomal ribonucleic acid
RT	Reverse transcriptase
RVE-INT	Retrotransposon derived integrase
SD	Segmental duplication
SFH1	Helicase superfamily 1
SFH3	Helicase superfamily 3
SL	Spliced RNA
SLTS	Spliced leader trans-splicing
SSU	Small ribosomal RNA subunit
TE	Transposable elements
TIR	Terminal inverted repeats
WGD	Whole genome duplication

ACKNOWLEDGEMENT

I am deeply thankful to my supervisor, Claudio Slamovits, for providing me with the opportunity to pursue my doctoral degree. I'm grateful for his permanent guidance and mentorship, which encouraged critical thinking and an organismic perspective of my research endeavours. I extend my gratitude to Pia Elustodo and especially Susana Breglia, who introduced me to the fascinating world of microbial eukaryotes. Additionally, I would like to thank my supervisory committee members, Andrew Roger, John Archibald, Melanie Dobson, and Jamie Kramer, for their continuous support and direction.

Furthermore, I express my gratitude to Dayana Salas, Lucie Gallot, Shopie de Vries, Jon Jerlstrom Hultqvis, Joran Martijn, Dandan Zhao, Gina Filloramo, and Sergio Muñoz for their invaluable insights into the world of molecular biology and bioinformatics. Special thanks to my lab colleagues Alex Mora and Nick Brask, as well as fellow graduate students Shelby Williams, Kelsey Williamson, and Morgan Colp, for their support and camaraderie

I am thankful to my family and Karla for their boundless support and love and to my friends for their continual encouragement and shared laughter.

Finally, I extend my gratitude to the ICG community, Compute Canada, and the financial agencies Becas Chile, Moore Foundation, and NSERC for their contributions to my training and financial support.

CHAPTER 1 GENERAL INTRODUCTION

1.1 Background

Dinoflagellates are one of the most abundant unicellular eukaryotes in aquatic environments, encompassing roughly 6,000 described species (Gómez, 2005; Riding et al., 2023; Taylor et al., 2008). Approximately 2,000 species have been named from fossil records (i.e., dinocysts), extending back 200-400 million years (Penaud et al., 2018; Riding et al., 2023). Still, much more of the existing diversity remains to be discovered since it is estimated that dinoflagellates represent nearly half of the biodiversity in the world's surface ocean (Le Bescot et al., 2016). Dinoflagellates can be found in diverse environments, such as pelagic and benthic zones within both marine and freshwater habitats. They are metabolically very diverse and include species with different lifestyles, including autotrophic (photosynthetic), heterotrophic, mixotrophic, and less commonly parasitic or symbiotic forms (Jeong et al., 2010). Half of the known dinoflagellate species are photosynthetic and play an essential role as primary producers in marine ecosystems (Gordon & Leggat, 2010). They can be found in symbiosis with a wide variety of organisms. The most notable example is the endosymbiosis with coral reefs, which is essential for the survival of this ecosystem (Muller-Parker et al., 2015). Certain dinoflagellates form harmful algal blooms (red tides), producing secondaries metabolites with toxic effects for humans through shellfish food poisoning. These unicellular organisms can also display bioluminescence controlled by an endogenous circadian clock (Hastings, 1996).

Dinoflagellates have a series of peculiar morphological and cytological features. The size of dinoflagellate cells ranges between 10 to 2000 micrometres (μm). They have two flagella: the transverse one at the cingulum –a groove-like structure, and the longitudinal one at the sulcus (Spector, 1984). Initially, the dinoflagellate orders were established on morphological features such as the tabulation patterns of the thecal plate (i.e., cellulose armour) (Fensome, 1993). Later, molecular phylogenetics suggested that these orders are poly- and paraphyletic, and thecal plate tabulation patterns evolved multiple times within this group (H. Zhang, Bhattacharya, et al., 2007). Molecular data shows that dinoflagellates form a monophyletic group of protists within the Alveolates (Janouškovec et al., 2017), sharing a flattened vesicle system beneath the cell membrane known as the alveoli (Cavalier-Smith, 1993; Patterson et al., 1991). They are closely related to Apicomplexa, a parasitic phylum with some similar features, including reduced mitochondrial and plastid genomes (Keeling, 2010; Waller & McFadden, 2005). Remarkably, dinoflagellates have a complex evolutionary history of plastids involving secondary or tertiary endosymbiosis events. The most common is a peridinin-pigmented plastid likely derived from secondary endosymbiosis (i.e., photosynthetic eukaryote phagocytosed by non-photosynthetic protist) and related to the one found in Apicomplexa. However, in some dinoflagellates, the peridinin plastid has been replaced by other types of plastids derived from tertiary endosymbiosis (i.e., uptake of alga with secondary endosymbiont) (Dorrell & Howe, 2015; Waller & Kořený, 2017; Yoon et al., 2005). Notably, some dinoflagellates have photosensitive eye-like organelles called "ocelloids" composed of lipid vesicles (Gavelis et al., 2015).

The unconventional configuration of the dinoflagellate nucleus poses challenges for evolutionary interpretation (Figure 1.1). The nucleus shows diverse shapes such as round, triangular, square, and U-shaped. The number of chromosomes also varies significantly, from four in *Dymbiodinium borgerti* to more than two hundred for *Perdinium cinctum* and *Ceratium hirundinella* (Spector, 1984). Moreover, chromosomes resemble a liquid-crystalline structure and remain permanently condensed throughout the cell cycle, attached at least from one end to the nuclear envelope (Herzog et al., 1982; Livolant, 1978; Livolant & Bouligand, 1978; Moreno Díaz de la Espina et al., 2005; P. J. Rizzo, 1991; Wong, 2019). The packing architecture of the dinoflagellate chromosome relies on a complex system based on DNA superhelicity structure (Wong, 2019). The morphology and delimitation between euchromatin and heterochromatin in dinoflagellate chromosomes are poorly understood (Cuadrado et al., 2019a). Strikingly, mitosis (i.e., dinomitosis) is conducted without breaking the nuclear envelope (i.e., closed mitosis) by employing an extranuclear spindle apparatus (Boettcher & Barral, 2013; Drechsler & McAinsh, 2012). During dinomitosis, cytoplasmic tunnels or channels run through the nucleus along with the cytoplasmic spindle (Gavelis et al., 2019). The chromatids contact the spindle through a membrane-bound kinetochore attached to the inner membrane of the nuclear envelope. Finally, chromatids migrate to the end of cytoplasmic tunnels, and then the nucleus is divided. Notably, chromosomes remain permanently condensed and devoid of typical nucleosomal organization observed in eukaryotes. Due to the apparent lack of histones and nucleosomes, dinoflagellates were proposed to constitute a "mesokaryotic," intermediate state between prokaryotes and eukaryotes (Dodge, 1965). Later findings enabled by genome research and molecular phylogenetics demonstrated that dinoflagellates are actual eukaryotes with histones that were low in abundance and showed high sequence divergence (Hackett et al., 2005; Okamoto & Hastings, 2003; Roy & Morse, 2012). Far from being an isolated rarity, depletion of nucleosomes is but one of the many unusual aspects of dinoflagellate chromatin (i.e., dinochromain), generally reflected by a remarkably low ratio of protein to DNA, which is 1:10 in dinoflagellates, while in most eukaryotes is 1:1 (P. Rizzo & Nooden, 1973). Remarkably, a set of bacterial and viral-derived proteins, i.e., histone-like protein (HLP) and dinoflagellates viral proteins (DVNPs), were found in close association with the DNA and likely participate in the gene expression regulation (Chan & Wong, 2007; Janouškovec et al., 2017; Sala-Rovira et al., 1991).

Outside of the core dinoflagellates (i.e., characterized by the previously described nuclear features), there are several less studied lineages (Figure 1.1) (e.g., Oxyrrhinales and Syndiniales). *Oxyrrhis marina* is particularly interesting because its basal placement in the dinoflagellate tree is robustly supported (Janouškovec et al., 2017; Slamovits et al., 2007a). Consequently, *O. marina* has been extensively employed as a representative for dinoflagellates in evolutionary, ecological, behavioural, and biogeographical studies (Boakes et al., 2011; C. D. Lowe, Keeling, et al., 2011; Montagnes, Lowe, Roberts, et al., 2011; Roberts et al., 2011; Slamovits & Keeling, 2011; Watts et al., 2011; Yang et al., 2011). The heterotrophic free-living dinoflagellate *O. marina* stands apart from core dinoflagellates because of its unique cytological features. For instance, *O. marina* mitosis is considered an early diverging dinoflagellate, lacking the cytoplasmic tunnels seen in the core dinoflagellate (Gavelis et al., 2019). The spindle is intranuclear in contrast to the extranuclear spindle of the core dinoflagellates (Figure 1.1). Moreover, *O. marina's* chromosomes show reduced birefringence and are less condensed than in most dinoflagellates (Kato et al., 1997; Spector, 1984). These distinctive features position *O. marina* between the typical dinoflagellate and the canonical eukaryotic nucleus, making it a potentially key organism for studying the origin of these unique features and the diversification of dinoflagellates.

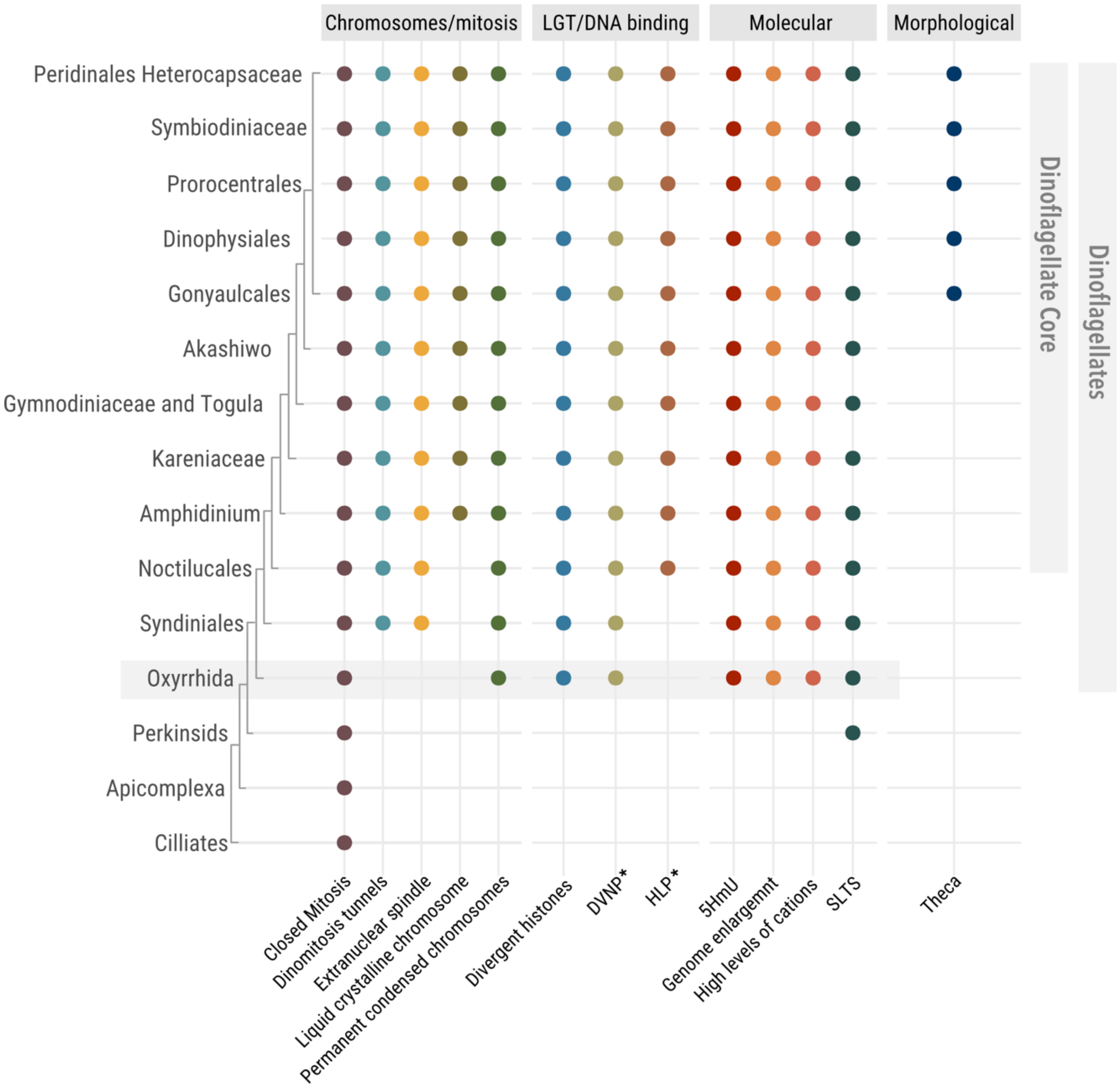


Figure 1.1. Dinoflagellate nuclear features mapped into the phylogeny.
Abbreviations: DVNP, dinoflagellate viral nucleoprotein; HLP, histone-like protein; 5HmU: 5-hydroxymethyl uracil; SLTS, spliced-leader trans-splicing. The asterisk (*) represented LGT events. The representation was based on earlier works (Gornik et al., 2019; Janouškovec et al., 2017).

1.2 Dinoflagellate DNA and genome

The nature of the chromatin of dinoflagellates differs substantially from canonical eukaryotes due to the prevalence of modified nucleotides and high levels of certain cations. A high-concentration Mg^{2+} and Ca^{2+} cations are present in association with the DNA of dinoflagellates compared to others eukaryotes. These elevated cation levels are thought to play a role in chromatin condensation, potentially compensating for the limited presence of histones (Koltover et al., 2000). Modified nucleotides such as 5-hydroxymethylcytosine (5-hmC) and 5-hydroxymethyluracil (5-hmU) are common in dinoflagellate DNA. 5-hmC levels are similar to other eukaryotes, and hypermethylation levels were detected in CG sites in *Symbiodinium*, although differential methylation was not seen under stress conditions (de Mendoza et al., 2018). Strikingly, a significant fraction of thymine (12-70%) is substituted by 5-hmU in some dinoflagellates (Herzog et al., 1982; Rae, 1973). These early observations based on buoyant density centrifugation also showed that 5-hmU is not randomly distributed and preferentially replaces thymine in dinucleotides TA and TC (Steele & Rae, 1980). It is likely that Ten-eleven translocation (TET) enzymes, primarily involved in DNA demethylation, take part in the 5-hmU synthesis (Gornik et al., 2019). Similarly, thymine is replaced by the unusual J-base (β-D-Glucopyranosyloxymethyluracil) in kinetoplastids, and 5-hmU is the precursor of the J-base (Borst & Sabatini, 2008). Recently, more insight associated with new sequencing methods shows that 5-hmU is often found at the edges of gene arrays and correlates with decreased chromatin accessibility in the dinoflagellate *Brevolium minutum* (Marinov et al., 2023). However, the functional significance of this modification is still unclear, and more validation with current approaches must be conducted.

Dinoflagellates are well known to have large nuclear DNA content (1.5-200 Gbp, human genome ~3 Gbp) (LaJeunesse et al., 2005). Early genomic surveys based on expressed sequence tags (EST) revealed that repeated sequences encompass more than 50% of the genome of *Alexandrium ostenfeldii* (Jaeckisch et al., 2011). Genes constitute a minority fraction of the dinoflagellate genome, estimated to be less than 1% (Hou & Lin, 2009). Nevertheless, genes are present in high copy numbers and organized in tandem. For instance, ~5,000 tandem copies of peridinin–chlorophyll α-protein were identified in

Gonyaulax polyedra (Le et al., 1997). Predictions based on genome size and gene content suggest that larger dinoflagellate genomes (~245 Gbp) may contain up to ~87,000-90,000 genes, which accounts for only 0.05-1.8% of the total genome (Hou & Lin, 2009). Consequently, the diversity of the proteome is predicted to be reduced due to the coding redundancy.

In recent years, genome assembly for small-size (1-4 Gbp) dinoflagellates (Suessiales and Symbiodiniaceae, Figure 1.1) was achieved through long-read sequencing technologies (Aranda et al., 2016a; González-Pech et al., 2021a; H. Liu et al., 2018; Shoguchi et al., 2013a; Stephens et al., 2020). A substantial portion of these genomes was identified as repeats and transposable elements (TEs), showing considerable variability depending on the species and organism's lifestyle (further details in the TEs section). These assemblies also predicted a large number of genes (~19,000-58,000), with a significant fraction (15-40%) being duplicated and arranged in tandem blocks (i.e., gene copies located next to each other) (H. Liu et al., 2018; Shoguchi et al., 2013a). Most notably, genes within tandem blocks are oriented unidirectionally, and it has been observed to influence both chromatin organization and gene transcription (Marinov et al., 2021; Nand et al., 2021). However, how tandem blocks are transcribed is still a matter of research. On the other hand, large-scale gene duplication has also been reported in the early diverging *O. marina* (Lee et al., 2014b); however, this has been inferred from transcripts alone, so that details about the genome organization of gene copies are unknown. Genome size varies drastically among dinoflagellates, and processes such as whole genome duplication or polyploidy and segmental duplication (i.e., unequal crossing over) have been proposed to explain the large fraction of genes duplicated in this group (Hou & Lin, 2009). *O. marina* has an intermediate genome size of about 30-50 Gbp with clear signatures for gene duplication (Lee et al., 2014a). Its early divergence in the dinoflagellate tree offers a strategic position to understand the mechanisms of genome enlargement in the core dinoflagellates compared to reduced genome size exhibited by other alveolates such as ciliates (~72-130 Mbp) and Apicomplexa (7-120 Mbp) (W. Chen et al., 2021; Swapna & Parkinson, 2017).

1.3 Gene expression and retrogenes

The control of gene expression in dinoflagellates predominantly relies on post-transcriptional and translational mechanisms rather than transcriptional regulation (Zaheri & Morse, 2022a). Unlike typical eukaryotes, cis-regulatory elements such as TATA-box promoters and transcription factors are notably scarce or divergent. Remarkably, genes arranged in tandem blocks tend to encode almost identical proteins and exhibit high expression levels (Bachvaroff & Place, 2008; L. Liu & Hastings, 2006). In contrast, the gene expression of individual genes remains relatively low. Gene amplification is proposed as the primary mechanism of gene expression regulation (Wisecaver & Hackett, 2011). This implies that genes in unidirectional tandem blocks are transcribed as polycistronic mRNA and further processed into monocistronic transcripts by trans-splicing, similar to trypanosomatids (Zaheri & Morse, 2022a). Despite of the exhaustive analysis of the cDNA, no evidence for polycistronic mRNA has been reported in dinoflagellate (Zhang, Hou, et al., 2007). However, further analysis using modern RNA sequencing methods (e.g., Nanopore RNA sequencing) has to be conducted in order to test this proposal.

Interestingly, all mRNAs in dinoflagellates undergo trans-splicing. A 22-nucleotide spliced-leader RNA motif (DinoSL-RNA) is consistently found at the 5' end of nuclear mRNAs across all dinoflagellate species (Lidie & Dolah, 2007; Slamovits & Keeling, 2008b; H. Zhang, Hou, et al., 2007). This observation also extends to the early diverging members *O. marina* and *Perkinsus*, suggesting the early emergence of this feature prior to the dinoflagellate diversification (Figure 1.1). On the other hand, DinoSL-RNA has been used to identify integrated cDNA, or retrogenes, in dinoflagellates genomes and transcriptomes . A large fraction of retrogenes (~30% of the total genes) have been identified in the genome of some symbiotic dinoflagellates. The emergence of retrogenes has been coupled with episodes of transposable element activity and the diversification of the symbiotic lineage (Jaeckisch et al., 2011; Slamovits & Keeling, 2008b; Song et al., 2017a). The abundance of retrogenes associated with photosynthesis, ion transport and cell adhesion functions has been proposed to facilitate the establishment of symbiosis (Song et al., 2017a). Furthermore, retrogenes potentially contribute to the observed overall gene redundancy in dinoflagellates. Retrotransposons have been proposed to be the mediator of retrogene formation, and retrogene survival depends on the acquisition or exploitation of distal promoters. However, these propositions have yet to be tested.

1.4 Transposable and viral elements

Dinoflagellates follow the general tendency of eukaryotic genomes: repeated DNA content correlates with the genome size (Wells & Feschotte, 2020). The first impressions of the repetitive nature of the dinoflagellate DNA emerged from reassociation kinetics (Allen et al., 1975) and S1 nuclease restriction enzyme studies (Hinnebusch et al., 1980), indicating that 50-60% of their genome comprises repeats. It was not until the advent of third-generation sequencing that complex elements such as retrotransposons emerged as prevalent components within the repeat fraction. Approximately 15-40% of the genome of symbiotic dinoflagellates is occupied by repeats (Y. Chen et al., 2022; González-Pech et al., 2021a), while it can reach up to ~64% in free-living dinoflagellates (Stephens et al., 2020). Likewise, the predominance of certain repeat elements likely changes depending on the lifestyle. So far, LINE retrotransposon are the most dominant transposable elements in symbiotic lineages (González-Pech et al., 2021a). In contrast, LTR elements and simple repeats are most prevalent in the free-living *Polarella glacialis* (Stephens et al., 2020). Nevertheless, the role of transposable elements in restructuring the dinoflagellate genome is unknown.

Very little is known about the acquisition of viral genes and the effects of viral infections on dinoflagellates. Viruses can modulate dinoflagellate density and disturb the symbiotic parentship. The ssRNA virus HcRNAV is responsible for the decline of blooms of *Heterocapsa circularisquama* (Tomaru et al., 2009). Notably, HcRNAV has been shown to acquire spliced-leaders (i.e., molecular mimicry) from dinoflagellates to evade the host immune response to the foreign nucleic acids. Additionally, one of the most significant instances of lateral gene transfer (LGT) involves the acquisition of Dinoflagellate/Viral Nucleoproteins (DVNPs), likely derived from the *Phycodnaviridae* family of giant viruses (Gornik et al., 2012). Giant viruses (NCLDV) infect symbiotic dinoflagellates, threatening the symbiosis with reef corals (Thurber & Correa, 2011), and participate in the termination of toxic dinoflagellate blooms (J. Wang et al., 2023). However, whether the virus persists in the cytoplasm or is integrated into the genome as a provirus is still being determined. On the other hand, small DNA viral elements (15-40 Kbp) known as Polinton/Maverick and virophages (parasites of NCLDVs) have been identified in dinoflagellate genomes (Bellas et al., 2023). Interestingly, Polinton/Maverick seems particularly abundant in the large genome of free-living dinoflagellates, suggesting a much larger diversity of these elements would be expected.

1.5 Dinoflagellate mitochondria

The mitochondrion is a double membrane bounded organelle found in the cells of virtually all eukaryotes. Mitochondria derived from an endosymbiotic alpha-proteobacterium acquired by early eukaryotic host cells, closely related to the Asgard archaea (Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017). Generally, mitochondria have been functionally associated with generating energy through aerobic respiration, resulting in ATP production. However, their functional roles are far broader and more complex, including participation in the apoptosis process, amino acid and nucleotide metabolism, calcium homeostasis, and lipid metabolism, among others (Roger et al., 2017a). Strikingly, mitochondria have different functions in highly specialized organelles, generally known as mitochondrion-related organelles (MROs) (Stairs et al., 2015). These were initially thought to be restricted to a few anaerobic parasitic protists (e.g., *Trichomonas vaginalis*) but are now known to be present in free-living anaerobic protists (Stairs et al., 2015). The mitochondrial genome (mitogenome) can encode dozens of proteins, including proteins involved in oxidative phosphorylation, as well as rRNA genes and mitochondrial-specific tRNAs. However, the gene content can vary drastically among organisms. Likewise, the mitochondrial genome varies in size and topology. It can be extremely large in gymnospermss (~4 Gbp in Siberian larch (Putintseva et al., 2020)) and highly reduced in some protists (6 Kbp in the apicomplexan *Plasmodium falciparum*). Gene transfer to the nuclear genome, also known as endosymbiotic gene transfer (EGT), may partially explain these drastic variations, although genome reduction and gene loss may be a preponderant factor.

The mitogenome of dinoflagellates has been of great interest due to its highly reduced and fragmented nature, like apicomplexans (Waller & Jackson, 2009a). It represents one of the most impoverished mitogenomes, encoding only three protein-coding genes, i.e., *cox1,* *cox3, and coxb*, in addition to several fragments of rRNA genes (Jackson et al., 2007; Nash et al., 2007; Norman & Gray, 2001a; Slamovits et al., 2007a). In contrast to the typical compact and circular mitogenome topology of most metazoa, the dinoflagellate mitogenome consists of a collection of DNA fragments in which genes can be found in multiple copies or fragments forming different arrangements. Genes lack canonical start and stop codons and are subject to RNA editing (reviewed in Waller & Jackson, 2009). This scenario poses challenges to generate a functional transcriptome. The features of the *O. marina* mitogenome are in concordance with the described tendency, but like apicomplexans, it lacks RNA editing (Slamovits et al., 2007a). Additional particularities can be found in the *O. marina* mitogenome, including a reduced gene complement with only two protein-coding genes and the presence of 5' oligo-U cap on mRNA (Slamovits et al., 2007a).The *O. marina* mitogenome peculiarities raise questions about its evolution in a broader context of protist mitogenome diversity.

This research is focused on characterizing the nuclear and mitochondrial genome content and organization of *O. marina,* as well as understanding the functional implications of retrogenes across dinoflagellates. Each chapter will include a detailed introduction and discussion, with broader aspects discussed in the general discussion chapter. Chapter 2 will describe the widespread abundance of retrogenes and their functional implications and diversity. Chapter 3 aims to reconstruct the fragmented mitochondrial genome of *O. marina* and understand the drivers that shape this genome. Chapter 4 will present the primary findings of the nuclear genome assembly, outlining the genome organization and prevalence of genetic elements. Finally, Chapter 5 will offer a detailed analysis of a notable viral element integrated into the *O. marina* genome.

CHAPTER 2 RETROGENES IN DINOFLAGELLATES

2.1 INTRODUCTION

Generating new genes significantly shapes molecular evolution, providing the raw material for the origination of evolutionary novelties. While well-documented DNA-based mechanisms like unequal crossing-over and segmental duplication are widely recognized as the main forces behind gene duplication (Kuzmin et al., 2022), less-explored RNA-based mechanisms (i.e., retroduplication) can also generate gene duplicates (Kaessmann et al., 2009a). This type of duplication eventually requires the off-target activity of reverse transcriptase derived from retrotransposons acting on the host mRNA (Casola & Betrán, 2017). Notably, LINE1 retrotransposon activity has been demonstrated to create retroposed gene copies in mammalian cell lines (Garcia-Perez et al., 2007; Klawitter et al., 2016). In retroduplication, the mRNA of a parent gene undergoes reverse transcription and is subsequently integrated into a new genomic locus (Kaessmann et al., 2009a) (Figure 2.1). The resultant duplicate consists of only exons devoid of cis-regulatory elements (e.g., promoters). Most retrocopies are non-functional (dead upon arrival) as they lack a promoter and lose their coding potential due to the accumulation of frameshift mutations causing premature stop codons (Mighell et al., 2000). However, retrocopies can sometimes escape this erosion, turning into *bona fide* genes, i.e., retrogenes (McCarrey & Thomas, 1987). Because of their potential to develop new functions, retrogenes were referred to as "seeds of evolution" (Brosius, 1991). Furthermore, it has been shown that retrogene functionality highly depends on the recruitment of regulatory sequences. They can be acquired from the new genomic neighbourhood or its retrotransposon mediator (reviewed in Kaessmann et al., 2009). As a result, retrogenes are prone to develop novel expression patterns that lead to new evolutionary trajectories and roles (Brosius & Gould, 1992; Long et al., 2003). In contrast, segmental duplication produces gene copies that primarily mirror the parental function. There is growing evidence supporting retrogene participation in a variety of processes, including subcellular relocalization of proteins (Rosso et al., 2008), neurotransmission (Burki & Kaessmann, 2004), tumor development (Staszak & Makałowska, 2021), and antiviral defence (Wilson et al., 2008). Additionally, since it is not only protein-coding transcripts that can become substrates for retrotranscription, some retrogenes have regulatory functions as non-coding RNA (Sasidharan & Gerstein, 2008;

Zheng & Gerstein, 2007). Retrogene formation represents a mechanism of gene emergence from *a priori* non-functional sequences and the recruitment of gene regulatory sequences from scratch. Nevertheless, understanding it is challenging, and surveying retrogenes in less-well-studied organisms may provide additional insight.

Retrogene research has predominantly focused on model organisms with an emphasis on mammals and *Drosophila melanogaster* because of the retention of young retrogenes (Bai et al., 2007; Emerson et al., 2004; Potrzebowski et al., 2008; Vinckenbosch et al., 2006). In contrast, there has been relatively less exploration of less-studied lineages such as green algae (Jąkalski et al., 2016) and dinoflagellates (Jaeckisch et al., 2011; Slamovits & Keeling, 2008b; Song et al., 2017a). Retrogenes are particularly abundant in the latter, but little is known about their persistence and significance for adaptation and genome evolution. The identification of retrogenes in dinoflagellates differs from other organisms because of a short spliced leader motif (DinoSL) upstream of the coding sequence (Figure 2.1). This DinoSL results from the trans-splicing of a short ~22 nucleotide SL RNA to the 5' end of the pre-mRNA of all protein-coding genes (Slamovits & Keeling, 2008b; H. Zhang, Hou, et al., 2007), a process termed Spliced leader trans-splicing (SLTS) that occurs in a handful of eukaryotic lineages and likely involved spliceosome machinery (Bitar et al., 2013). This DinoSL relic has been used as a "tag" to facilitate retrogenes identification from EST data (Jaeckisch et al., 2011; Slamovits & Keeling, 2008b) and genomic sequence (Song et al., 2017b). Retrogenes are abundant, accounting for 22-25% of the total genes in *Symbiodinium* genomes (Song et al., 2017b). Two massive retroposition episodes were inferred for *Breviolum minutum* and *Symbiodinium kawagutii,* leading to the enrichment of retroposed genes related to ion and transmembrane transport, photosynthesis and symbiosis establishment (Song et al., 2017b). These retrogenes have been proposed to be crucial for the adaptation to symbiotic life. Interestingly, the abundance of particular retrogenes may correlate with the expression level of their parental genes, meaning that highly expressed genes have higher chances of being the target of retroposition (Pavlicek et al., 2006). This may explain the accumulation of genes involved in stress response in *Symbiodinium* genomes stimulated by dramatic climate changes (Lin et al., 2015). The idea that highly expressed genes become retrogenes is a "self-reinforcing model of molecular evolution"

(Song et al., 2017b), although further corroboration is needed. Retroposition in dinoflagellate may be mediated by retrotransposons found particularly abundantly in several species (González-Pech et al., 2021a), but retroviruses cannot be ruled out. In terms of functionality, it is unclear how retrogenes persist and become functional. Transcriptional regulation in dinoflagellates is still poorly understood, but dinoflagellates appear to rely on fewer and simpler transcriptional regulatory elements compared to other eukaryotes (Roy et al., 2018; Zaheri & Morse, 2022a). An in-depth analysis of the presence and distribution of retrogenes in a wide range of dinoflagellates may help understand gene emergence and its diversity. The presence of DinoSL in retrogenes makes the search straightforward, even allowing the identification of retrogenes in intron-containing genes that are otherwise overlooked in other organisms.

Here, we analyze RNA sequencing data and transcriptome assemblies for 45 dinoflagellates species to conduct functional categorization, expression quantification and codon usage analysis on dinoflagellate retrogenes. Our findings show that most retrogenes originate from genes involved in essential housekeeping processes and other well-conserved core activities. A high level of retrogene expression and codon bias trends suggests that retrocopies easily become functional retrogenes. The prevalence of retrogenes across dinoflagellates can be associated with pervasive activity retrotransposon.

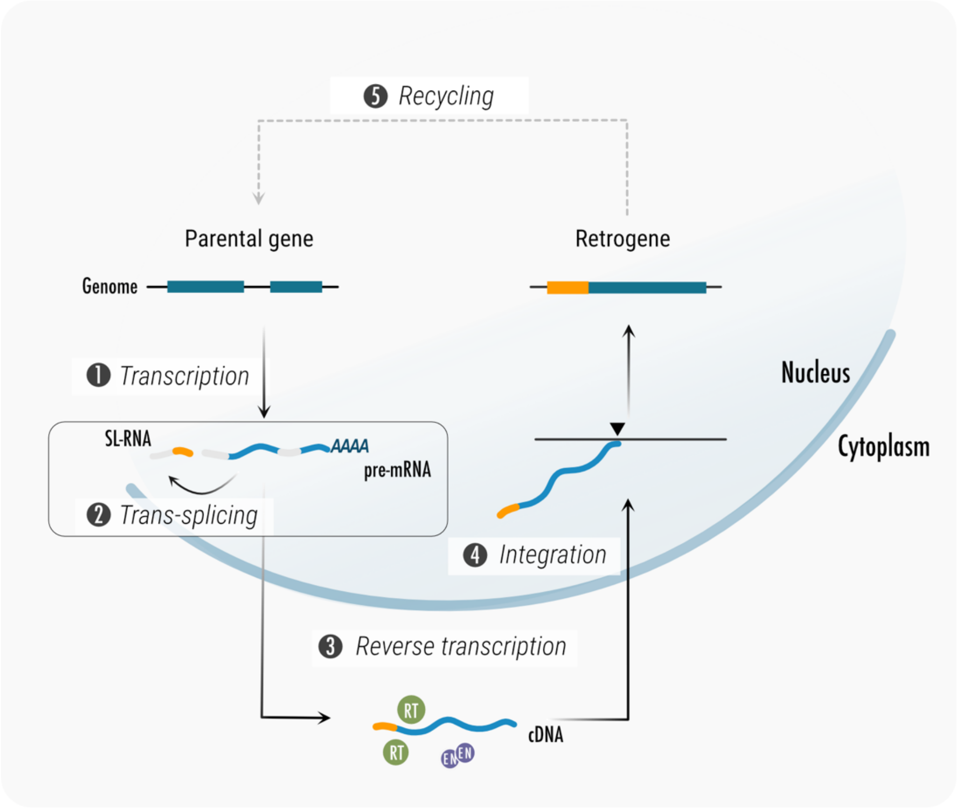


Figure 2.1. Conceptual framework for retrogene formation in dinoflagellates.
This model outlines the sequence of events initiating with the transcription of a parental gene (1). The distinctive feature in dinoflagellates is the occurrence of spliced-leader trans-splicing (2) during the process of retrogene formation. The processed mRNAs undergo reverse transcription (3) facilitated by reverse transcriptase (RT). Subsequently, these reverse-transcribed sequences are integrated into the genome (4). Eventually, retrogenes might be subject to retroduplication or recycling (5). This proposed model was based on Slamovits et al., 2008.

2.2 METHODS

2.2.1 Data collection and retrogene identification

Transcriptomes were generated by the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling et al., 2014), and further curated (Van Vlierberghe et al., 2021). Additional high-quality transcriptome assemblies of the Symbiodiniaceae family (Aranda et al., 2016b; Barshis et al., 2014; Bayer et al., 2012; Levin et al., 2016; Parkinson et al., 2016; Shoguchi et al., 2021) and *Pyrocystis lunula* (Menghini & Aubry, 2021) were included. All transcripts shorter than 200 bp were removed. Assembly completeness was assessed using Benchmarking Universal Single-Copy Orthologs BUSCO (v 3.0.0) (Simão et al., 2015a), using alveolate_odb10 (171 orthologs). Basically, BUSCO is a widely used method to assess the completeness of genome assembly based on the presence of a set of highly conserved single-copy orthologous that are expected to be present in the target lineage. We took advantage of the ubiquitous DinoSL (DCCGTAGCCATTTTGGCTCAAG, D: A, G, T) to find potential retrogenes. Transcripts having at least one DinoRL (after DinoSL relic, CCATTTTGGCTCAAG) (Slamovits & Keeling, 2008b) following DinoSL (CCGTAGCCATTTTGGCTCAAGCCATTTTGGCTCAAG) at their he 5' ends were targeted as potential retrogenes. Retrogene sequences were identified with seqkit (Shen et al., 2016), using the command seqkit grep (-s -m 5 -i). Retrogene redundancy was reduced by collapsing similar copies within species using CD-HIT (4.8.1, word size -8 and 90% identity) (W. Li & Godzik, 2006), keeping the largest isoform as representative.

2.2.2 Functional annotation

Hypothetical retrogenes were translated into amino acid sequences with TransDecoder v.5.5.0 (www.github.com/TransDecoder), and proteins were searched against the NCBI non-redundant database using diamond BLASTp (2.12.0, e-value 1E-5, -max-target-seq 10). PFAM domains were identified using HMMSCAN (HMMER v3.1b2) (Eddy, 2011) with an e-value cut-off of 1E-3. The amino acids sequences inferred from the retrogenes were also queried against our local PANTHER 14.0 database (Mi et al., 2017) with an e-value cut-off of 1E-3. Gene ontology terms (GO) were retrieved for the PFAM domains using the function bitr implemented in clusterProfiler (v 4.0) (Wu et al., 2021a) with *Saccharomyces cerevisiae* as a reference organism.

2.2.3 Enrichment analysis

Enrichment analysis was conducted for dinoflagellates with higher retrogenes counts and transcriptome completeness (BUSCO > 80%; *G. catenatum, A. molinatum*, and *B. nutricula*). The retrogenes PFAM domain annotation was tested for enrichment against their respective transcriptomes. Significance was determined using Fisher's exact test, and P-values were corrected for multiple comparisons using the Benjamini & Hochberg method. The GO enrichment was conducted in clusterProfiler v.4.0 (Wu et al., 2021a) using the function enrichGO. GO terms with a P-value less than 0.01 were considered enriched. Redundant GO terms were removed using the function simplify (cutoff=0.6) in clusterProfiler. We visualized the results using the function cnetplot of clusterProfiler and the R package ggplot2 (3.3.5) (Wickham, 2016).

2.2.4 Expression estimation

The raw reads (SRA) for *G. catenatum* (SRR1296705), *B. nutricula* (SRR1300537) and *A. molinatum* (SRR1296895, SRR1296896, SRR1296897, SRR1296898) were downloaded from NCBI under BioProject PRJNA231566 (Keeling et al., 2014). Reads were trimmed using the Trimmomatic v.0.39 software (Bolger et al., 2014) with a conservative setting (Johnson et al., 2019). Reads mapping and quantification were conducted by bowtie2 v.2.4.5 (Langmead & Salzberg, 2012) and RSEM software v.1.3.0 (B. Li & Dewey, 2011). The relative expression was normalized in Transcripts Per Millions (TPM), which normalizes the read count for the gene length divided by a million (scaling factor). Retrogenes and protein-coding CDS with TPM > 1 were retained. The Wilcoxon rank-sum test evaluated the comparison between retrogenes and protein-coding transcript expression levels.

2.2.5 Codon usage analysis

The codon usage indicators: GC3s, GC content, the effective number of codons (ENc), and RSCU values were estimated for hypothetical retrogenes and protein-coding CDS sequences using CodonW V1.4.4 software (http://codonw.sourceforge.net). Plots were generated in R using the function ggscatter of ggpubr V 0.4.0. These estimations were conducted for *G. catenatum*, *A. molinatum* and *B. nutricula*

2.3 RESULTS

2.3.1 Structure and distribution of retrogenes across dinoflagellate orders

We took advantage of the presence of a "relic" DinoSL (hereby DinoRL) to identify retrogenes in transcriptome assemblies from dinoflagellates, which makes the process straightforward and overcomes the general lack of genome sequencing (Song et al., 2018; Slamovits & Keeling, 2008b). We were particularly interested in studying retrogenes that include the SLTS (spliced-leader Trans-splicing) system as a part of their retroduplication cycle. Therefore, we searched for the DinoSL-DinoRL tandem (37 nucleotides) at the 5'-end of each transcript as the signature for retrogenes (see Methods). We retrieved 6,544 highly confident retrogenes across 37 of the 45 dinoflagellate species. DinoSL-DinoRL was the most prevalent arrangement in 95% of the retrogenes (Figure 2.2A). The remaining 5% have at least two DinoRLs tandemly arranged along with DinoSL. This implies that at least 5% of the retrogenes went through multiple recycling events or were subject to two retroduplication events. These findings are consistent with the fact that multiple rounds of retroduplication originated as part of the retrogene repertory of the *Symbiodinium* lineage (Song et al., 2017b). Additionally, the DinoSL (Figure 2.2B) consensus indicated that most of the retrogenes start with truncated DinoSL (GCTCAAG) followed by a DinoRL (CCATTTTGGCTCAAG). We observed this truncation in most of the dinoflagellate orders, with the exception of Noctilucales, where the canonical DinoSL prevailed (Supplementary Figure 1). Likely, RNA degradation and further processing and trimming of transcriptome sequences led to the truncation of DinoSL.

The transcriptome assembly dataset was highly complete; on average, 82% of BUSCO (Eukaryota_odb9) proteins for alveolates were identified, except for *Alexandrium andersonii*, *Alexandrium minutum*, *Cladocopium sp.*, and *Durusdinium trenchii* (< 50%).

Next, we analyzed the distribution of retrogenes in the 46 dinoflagellate species belonging to seven taxonomic orders (Figure 2.2C). We identified retrogenes in most datasets. However, despite their assembly completeness, DinoSL-DinoRL retrogenes were not detected in nine species, most of them from the Suessiales order. This may be in part due to technical limitations of the sequencing process (e.g., RNA degradation and epigenetic modification), DinoRL degeneration and also unknown reasons likely associated with symbiotic lifestyle (low-frequency retrotransposition). On average, 182 retrogenes were identified per species (retrogenes isoforms were collapse within specie), with the highest count for *Pyrocystis lunula* (757) and the lowest for *Oxyrrhis marina* (1). This trend suggests that retroposition is widespread across the dinoflagellate diversity in concordance with the large scale of retroposition predicted for this lineage (Slamovits & Keeling, 2008b). However, retrogene abundance is generally similar among taxonomic orders, with the highest average in Gonyaulacales (271 per taxon) and the smallest in Suessiales (21 per taxon). Approximately half the species for order Suessiales tend to host fewer retrogenes, which might be associated with the reasons explained above. Additionally, no correlation between genome size and retrogene abundance was identified (Supplementary Figure 2).

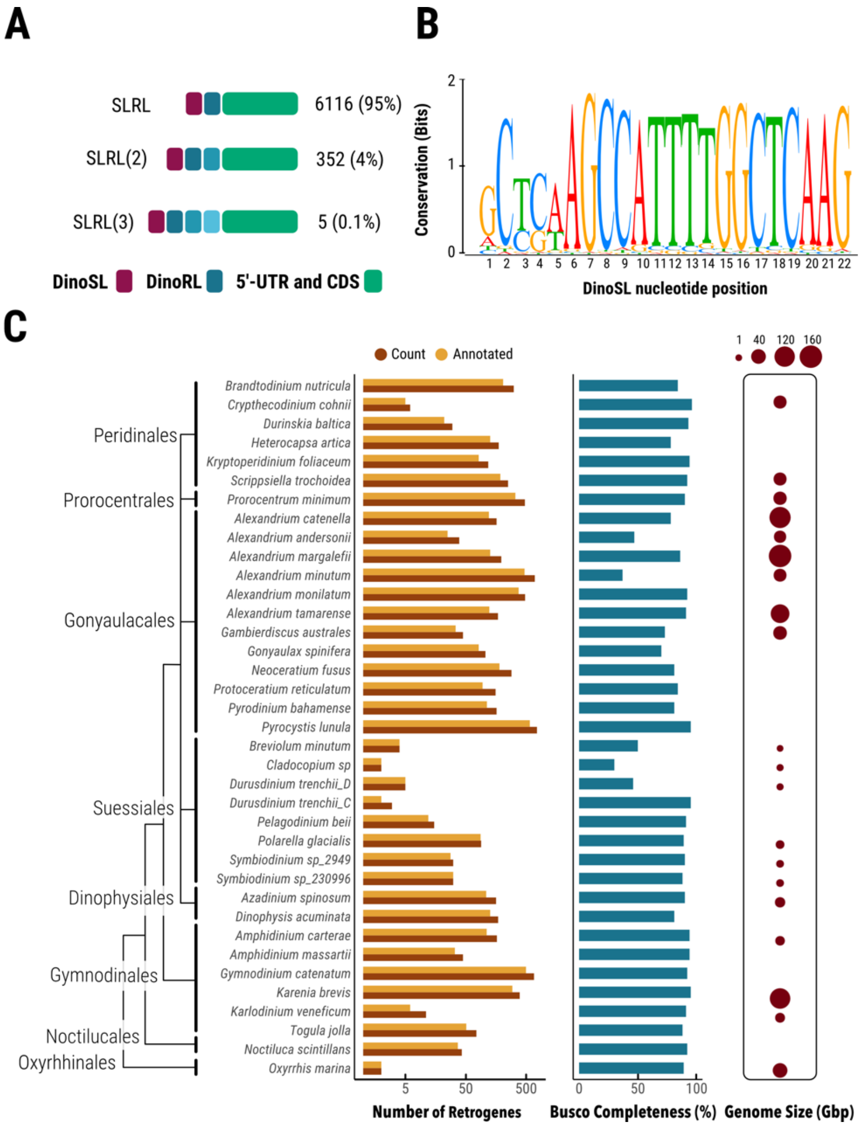


Figure 2.2. Retrogene survey in dinoflagellate transcriptome assemblies.

(A) number and type of DinoSL/DinoRL arrangements found in retrogenes. DinoSL-DinoRL was the most predominant tandem arrangement found (6116). (B) DinoSL logo consensus resulting from the alignment of DinoSL found in retrogenes of all dinoflagellates. Motif length and position are indicated. The height of the letters represents the sequence conservation and the nucleotide frequency in each position. (C) number of retrogenes (count), protein-coding retrogenes (annotated), and percentage of transcriptome assembly completeness (BUSCO) for each dinoflagellate species. An average of 182 retrogenes per species was identified, and the average assembly completeness was 82%. The dinoflagellate phylogeny sketch was based on Janouškovec. al. 2016.

2.3.2 Functional annotation and enrichment

We comprehensively annotated retrogenes against NCBI NR (non-redundant) and PANTHER and PFAM databases with HMM profiles. We were able to annotate 4,742 retrogenes (72%) to protein and highly conserved protein domains, suggesting that this considerable proportion potentially corresponds to functional retrogenes. Furthermore, the translated peptide size of annotated retrogenes was significantly larger (276 amino acids) compared with retrogenes without annotation (124 amino acids) (Supplementary Figure 3), suggesting premature stops codons and protein truncation may lead to pseudogenization of retrogenes. The top five most frequent PFAM domains identified were kinase, RRM1, Cold shock domain protein (CSD), Ubiquitin, and EF-hand domain (Supplementary Table 1). Similarly, the top 5 most abundant PANTHER proteins annotated were Chlorophyll a-b binding protein, RNA recognition motif, CDS, 60s ribosomal protein and Cytochrome b5 heme-binding domain (Supplementary Table 1).

We revealed that the retrogene dataset is representative of the functional core of dinoflagellates. We conducted a gene ontology annotation of the PFAM domains to understand the functional categorization of our retrogene dataset. About 30-40% of the Pfam domains encoded by the retrogenes do not affiliate with any gene ontology categories and terms. The GO category "biological process" was the most represented in the retrogene dataset to a lesser degree, "cellular component" and "molecular function", respectively (Figure 2.3A). In The "biological process" category, "metabolic processing of organic substances" (e.g., carbohydrates and amino acids), "nitrogen, biosynthesis", and "regulation" are highly represented (Figure 2.3A). We conducted an enrichment analysis of the dinoflagellates with the highest retrogene count: *Gymnodinium catenatum* (676), *Alexandrium monilatum* (481) and *Brandtondinium nutricula* (312). We tested the GO term over-representation (related to the transcriptomes) for each and found a core of 21 GO terms shared by the three dinoflagellates being the most enriched involved in post-translational modification (e.g., protein phosphorylation: GO:0006468, *p.adj.* < 1E-46, rich factor 0.48-0.5; phosphorylation: GO:0016310, *p.adj.* < 3E-26, rich factor 0.14-0.33: and protein serine/threonine kinase activity: GO:0004674, *p.adj.* < 5E-88, rich factor 0.87), cell signalling (e.g., intracellular signal transduction: GO:0035556, *p.adj.* < 2E-23, rich factor

0.34-0.37; and signalling: GO:0023052, *p.adj.* < 2E-20, rich factor 0.23-0.3) and xenobiotic transport (GO:0042910 *p.adj.* < 2E-06, rich factor 1, GO) (Figure 2.2B). We also found that housekeeping processes such as DNA, RNA metabolism, and post-transcriptional modification are commonly enriched (Figure 2.2B). On the other hand, we found unique enriched GO terms, mostly related to protein biosynthesis and glucose metabolism (Supplementary Figure 4). Additionally, several GO terms involved in DNA and RNA binding metabolism (e.g., regulation of DNA binding, DNA duplex unwinding, mRNA3'-UTR binding, and single-stranded RNA binding) are enriched because of retroduplication and splicing events like previous findings (Song et al., 2017b).

We wanted to know if the retroduplication process led to the amplification of particular protein domains and contributed to functional redundancy. Therefore, we compared the abundance of the most frequent protein domains encoded by retrogenes among the three selected dinoflagellates. We found that some domains are heavily represented in some species; for instance, RRM and EF-Hand motif are highly enriched in *B. nutricula,* which may result in reinforcing processes of signalling and post-transcriptional modification (Figure 2.2C). Similarly, in the case of *G. catenatum,* Pfam domains involved in similar functions, e.g., Pkinase and methyltransferases, were highly enriched. Additionally, some photosynthesis involved protein domains were enriched in *G. catenatum* (e.g., TPT transport) and *A. monilatum* (e.g., PsbK, PsbL, and PsbE). Most of these domains are annotated in the top 20 PFAM domains. Also, retroduplicated protein domains may contribute to adaptation; for instance, additional copies of glutathione transferases may help eliminate toxic compounds during symbiosis establishing in *B. catenatum*. The cold shock domain may benefit in enduring freezing conditions in *A. molinatum*. Additionally, Sulfotransferase domains may contribute to toxin production in *G. catenatum*.

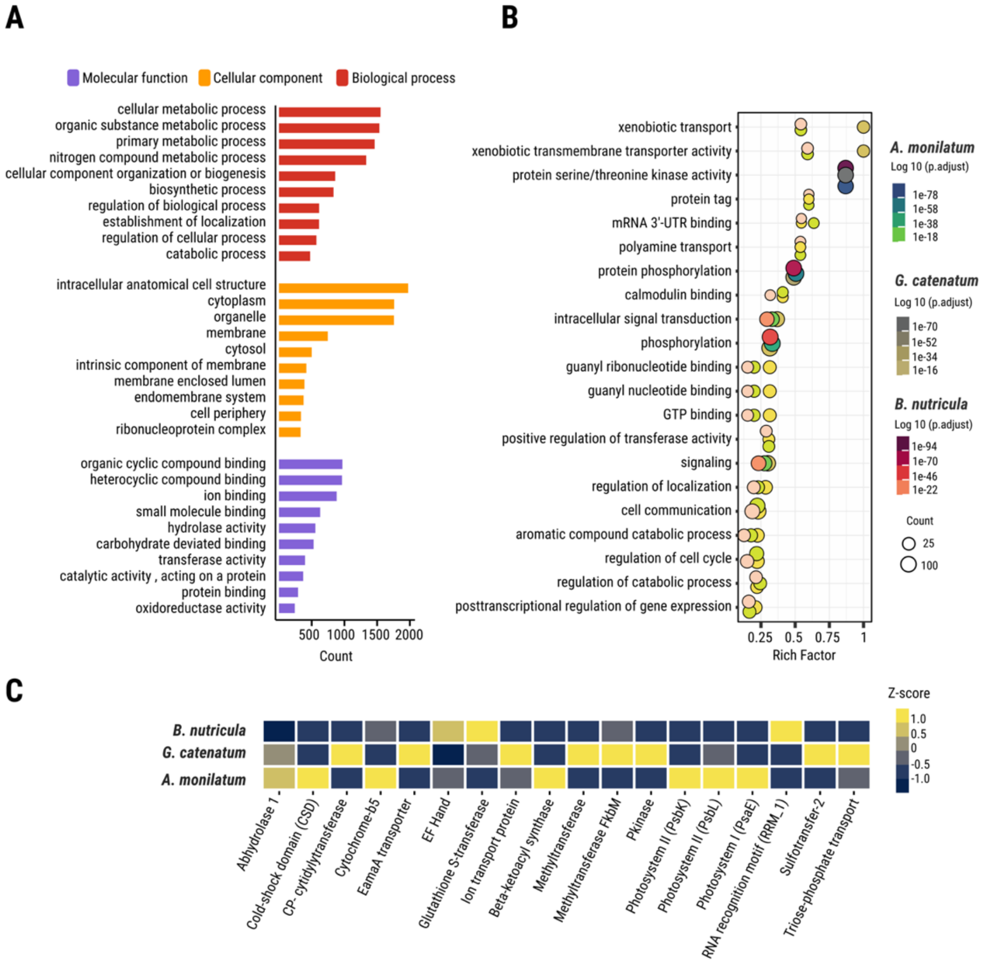


Figure 2.3. Functional annotation and enrichment analysis of retrogenes.

(A) gene ontology annotation with the top ten most abundant GO terms for each gene ontology category. Biological process was the most represented category in the retrogene dataset. (B) gene enrichment analysis with 21 commonly enriched GO terms for three selected dinoflagellates with the highest retrogene count: *A. molinatum*, *G. catenatum* and *B. nutricula.* Colour scales depict the $\log_{10}$ value of P-adjusted values, and high colour intensity indicates higher enrichment. The rich factor is the proportion of retrogenes to genes that are annotated in a particular GO term. The higher the rich factor, the higher the enrichment of the GO term. GO terms protein serine/threonine kinase activity, protein phosphorylation, intracellular signal transduction, signalling and phosphorylation resulted consistently enriched in the three species. (C) differentially enriched PFAM protein domains encoded by retrogenes for the three selected dinoflagellates. The frequency values were converted to z-scores to indicate the relative enrichment (yellow) and depletion (dark blue).

2.3.3 Retrogenes expression and codon usage trend

Retrogenes have been claimed to have a high survival rate in dinoflagellates, unlike other lineages, in where they are continuously lost or “dead on arrival” (Z. Zhang et al., 2003). Codon usage trends can be useful in investigating evolutionary processes affecting protein-coding genes and pseudogenes (Bi et al., 2023; X.-Y. Liu et al., 2020). We wanted to test if codon usage provides functional signatures to clarify the relationship between expression levels and the dynamics of gene duplication by retrotransposition. We compared the expression in $\log_2$ of transcript per million (TPM) between non-retrogenes and retrogenes for the three dinoflagellates with the largest retrogene counts (Figure 2.3A). We found that the median of retrogene expression is consistently higher in all three dinoflagellates, suggesting retrogenes remain functional and derived from the highly expressed core of genes. Furthermore, we compared 26,606 non-retrogenes coding sequences (CDSs) and 338 retrogenes for the three dinoflagellates for codon usage bias (Supplementary Figure 5). We found that the Nc (effective number of codon) was, on average, 44.5 for protein-coding genes (1,740 sequences Nc < 35) and 45.1 for retrogenes (34 sequences Nc < 35), suggesting absence of codon usage bias in the dataset. The GC content was generally high (62.7% non-retrogenes and 61.3% retrogenes) but not significantly different. High GC content at the third codon position was identified in both gene categories, but no significant difference in the GC3s was established (Supplementary Figure 5). We plotted Nc vs GC3s non-retrogenes and retrogenes CDSs along with the standard curve. We found that values clustered between 25-60 and 30-60 for non-retrogenes and retrogenes, respectively, suggesting some codon bias (Figure 2.3B). Most values lie around but not precisely on it, meaning that codon usage bias may depend on additional factors such as natural selection rather than pure mutational bias. The neutrality plot does not show the correlation between GC3 and GC12, suggesting 3rd codon position is less constrained compared with the first two (Figure 2.3C). Finally, the correspondence analysis of the relative synonymous codon usage (RSCU) indicated that the first axis explained 18% of the variation (Figure 2.3D).

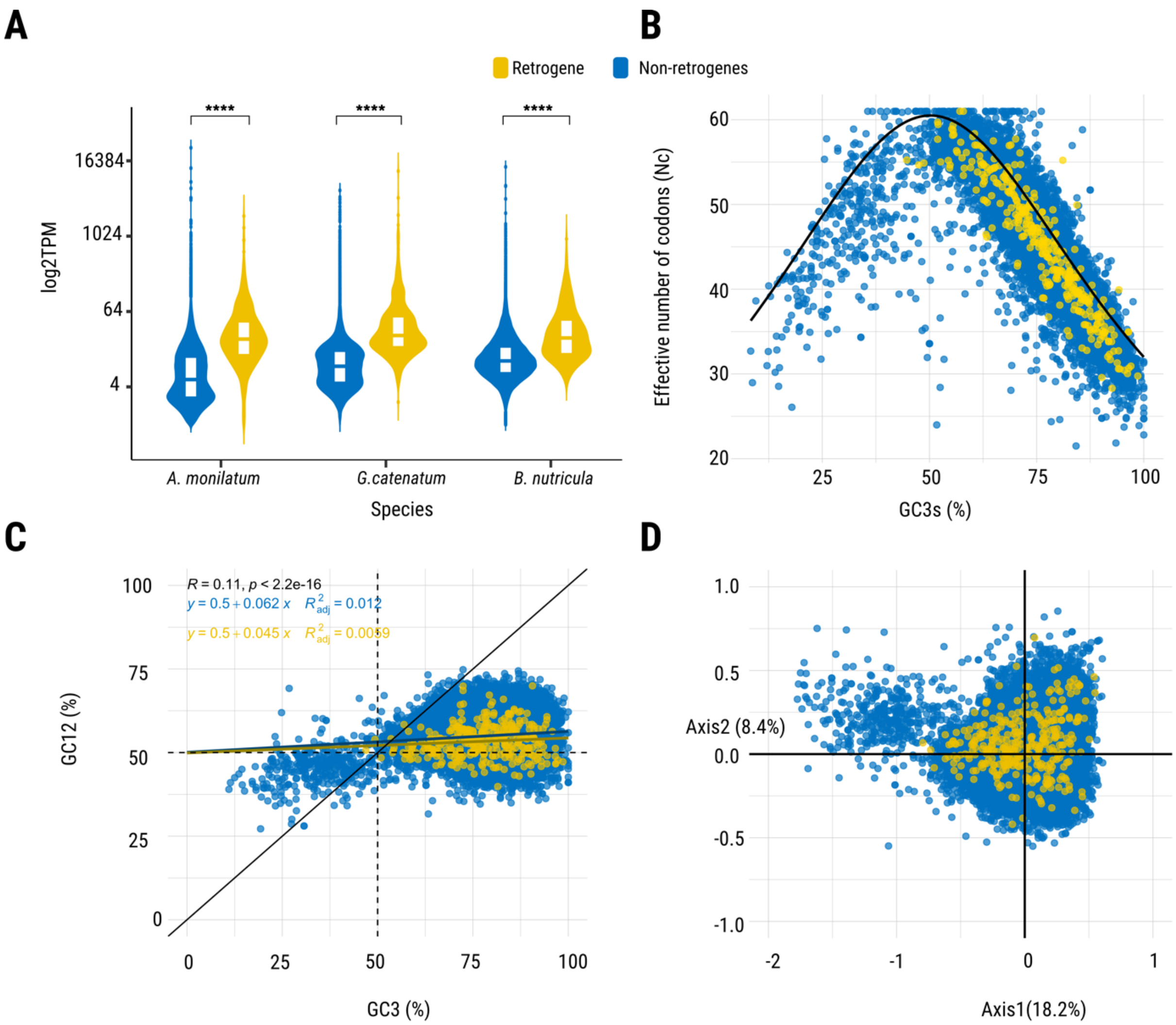


Figure 2.3. Retrogene expression and codon usage analysis.

(A) violin plot of expression level ($\log_2$ TPM) for retrogenes and protein-coding non-retrogenes. Wilcoxon rank-sum test was used to assess the significance. (B) the effective number of codons (Nc) relative to the GC percentage at synonymous third codon position (GC3s). The solid black curve represents the expected values for random codon usage. (C) neutrality analysis: The x-axis is the GC content at the third codon position. The Y-axis is the average GC content for the first and second codon positions. The diagonal represents neutrality (GC3=GC12), and genes/retrogenes lying on it indicate that codon usage is driven by neutral selection. Trend lines for retrogenes and genes are shown in yellow and blue, respectively**.** (D) correspondence analysis based on RSCU values for retrogenes and protein-coding genes.

2.4 DISCUSSION

Retroposition generates new gene copies from transcribed genes, contributing to genome evolution by directly or indirectly participating in gene birth-and-death processes, genome size dynamics, and adaptation. The dinoflagellates represent an ideal lineage to study retroposition due to the advantage of DinoRL as the hallmark of the retrogenes (Slamovits & Keeling, 2008b; Song et al., 2018). Consequently, we conducted a comprehensive survey of retrogenes in all available transcriptomes, focusing on understanding their functional categorization, distribution, and operational status.

In our survey, transcripts starting with full-length DinoSL followed by DinoRL were considered retrogenes; however, the actual number of retrogenes is likely underestimated because transcripts with partial or incomplete 5'-ends were discarded. Furthermore, technical limitations of RNA-seq, RNA degradation and epigenetic modification may negatively affect the recovery of retrogenes. But more significantly, degeneration by accumulation of random substitutions in the DinoRL eventually change the consensus beyond recognition, rendering detection impossible, thus contributing to underestimation of the number of retrogenes (Jaeckisch et al., 2011; Slamovits & Keeling, 2008b).

2.4.1 DinoSL conservation in retrogenes

SLTS is a distinctive phenomenon of dinoflagellates widespread in all nuclear mRNAs; recent studies indicated that this process could be preponderant in a certain fraction of the transcriptome, targeting specific functional categories of transcripts (Alacid et al., 2022; Stephens et al., 2020). We were interested in studying retrogenes targeted by the SLTS system where DinoSL-containing mRNAs are reverse transcribed and then integrated into the genome (Slamovits et al., 2011). We recovered a highly confident set of retrogenes confirming that DinoSL is conserved among dinoflagellate orders and distinctive from other Phylum as previously described. Interestingly, the DinoSL nucleotide consensus (Figure 2.2, Supplementary Figure 1) at the third, fourth, and fifth positions (TCA) differ from the expected CGT (canonical DinoSL), likely because of DinoSL 5'end truncation. Sample degradation may account for the partial degradation of DinoSL. No additional variants for DinoSL were identified in this analysis, but it cannot be ruled out that they may be present in low frequency. On the other hand, we found that retrogenes are widely distributed along dinoflagellate phylogeny, further cementing the notion that SLTS is a highly conserved and universal molecular feature of the entire clade. Dinoflagellate orders with large genome sizes, such as Gonyaulacales is richer in retrogenes content than smaller genome size orders (e.g., Suessiales), probably because the gene repertoire is broader in large genomes (Hou & Lin, 2009). Additionally, less selective constrictions for genome size may increase retrogene retention.

2.4.2 Post-translational modification pathways emerge as prominent features among retrogenes

We aimed to determine whether the set of retrogenes is enriched in any functional categories. We contrasted the retrogenes of three dinoflagellates belonging to three different orders (Gymnodiniales, Peridiniales and Gonyaulacales) with their corresponding transcriptome. We found that retrogenes are enriched in processes such as DNA metabolism (e.g., guanyl nucleotide binding, regulation of DNA binding, DNA duplex unwinding), RNA metabolism (e.g., guanyl nucleotide binding, mRNA3'-UTR binding, single-stranded RNA binding), protein biosynthesis (e.g., cytoplasmic translation, ribosomal small subunit assembly, regulation of cytoplasmic translation) and glucose metabolism (e.g., glucose metabolic process, glucan metabolic process, lactate metabolic process), (Supplementary Figure 4). Also, we were able to identify a shared core of highly enriched GO terms related to post-translational modification (e.g., protein phosphorylation, phosphorylation, and protein serine/threonine kinase activity), cell signalling (e.g., intracellular signal transduction, calmodulin binding and signalling) and transport (e.g., xenobiotic transport), (Figure 2.3A). We also detected that the enriched terms correlate positively with the abundance of specific retrogene-coding protein domains such as ion transport, Pkinase, and methyltransferases (Figure 2.3C). Therefore, retroduplicates contribute to the extensive gene redundancy that distinguishes dinoflagellate genomes (Hou et al., 2019). Previously, retrogenes were found to be enriched in housekeeping tasks in *Symbiodinium* species (Song et al., 2017b). Our findings demonstrate that this is broadly true for dinoflagellates and emphasize that post-transcriptional modification, cell signalling, and transport represent the predominant fraction of the dinoflagellate transcriptome that contributes to the retrogene repertory. The post-transcriptional modifications and signalling categories are especially significant as dinoflagellates rely heavily on translational levels to control protein abundance rather than transcriptional regulation (Roy et al., 2018; Zaheri & Morse, 2022a).

Stress response may promote the expression of pathways involved in cell signalling and xenobiotic transport. Certain stressors, such as heat, trigger retrotransposons' activation in dinoflagellates (J. E. Chen et al., 2018; Song et al., 2017b). Since highly expressed genes have a high likelihood of retroposition (Pavlicek et al., 2006), activation of retrotransposons may lead to the retroposition of genes related to pathways with high expression profiles. The retroposition mechanism invokes the off-target activity of the retrotransposon in the form of reverse transcriptase activity on cellular transcripts (Casola & Betrán, 2017; Kaessmann et al., 2009b). As a result, the primary mechanism of gene duplication in dinoflagellates may include a fine interplay between gene expression level and retrotransposon activity. The high number of gene copies is proposed to result in a self-reinforcing model of genome evolution in the dinoflagellate lineage (Song et al., 2018). The high expression profile of retrogenes that we observed (Figure 2.3A) might mirror the parental expression pattern, reinforcing or replacing it. In fact, most of the retrogenes in *Symbiodinium* are “orphans” with no parental gene found, and it has been shown that orphan retrogenes tend to recapitulate the expression pattern of the parental gene in mammals (Carelli et al., 2016). Interestingly, many functions were not consistently enriched in the three species analyzed (Figure 2.3C), which probably reflects the uniqueness of each lineage evolving independently under its own environmental constraints. Understanding how retrogenes acquired the already scarce regulatory sequences in dinoflagellates would help reveal the machinery behind the transcriptional profile of retrogenes.

Retrotransposed genes need to be expressed to be considered functional retrogenes (Carelli et al., 2016). First, we found that dinoflagellate retrogenes are consistently highly expressed (Figure 2.3A). Then, we hypothesized that functional retrogenes should have similar codon bias trends compared with protein-coding sequences (non-retrogene sequences). In general, we observed a similar tendency of codon bias between these two categories: similar GC3s composition and Nc values (Supplementary Figure 5). Therefore, we conclude that retrogenes represented a highly expressed section of the dinoflagellate transcriptome with similar codon bias compared with protein-coding sequences. On the other hand, promoters may be frequently present in the unicellular eukaryote genome, as is evidenced in yeast, where promoters are located at 50 bp on average from the transcription start site (TSS) (Qiu et al., 2020). A similar scenario might explain why retrotransposed genes become active retrogenes at such high rates. It could be that sequences that can drive transcription (even if at low levels) are ubiquitous. Therefore, a newly reverse transcribed sequence's probability of land near one is relatively high. Moreover, a TTTT box found in the DinoSL could itself be able to promote transcription (Song et al., 2017b), reinforcing the idea that retrocopies carry their own basal promoter. Additionally, transcripts undergoing multiple rounds of recycling are likely to carry more relics, enhancing the probability of being expressed upon integration.

.

CHAPTER 3 THE MITOCHONDRIAL GENOME OF *OXYRRHIS MARINA*

3.1 INTRODUCTION

The mitochondrial genomes of protists display remarkable diversity regarding genome size, gene richness and organization (Gray et al., 1998, 2004; Roger et al., 2017b). In some ways, this diversity spans extremes quite removed from the textbook examples of human or yeast mitochondrial genomes. For instance, the jakobid flagellate *Reclinomonas americana*, where the mitochondrial genome is densely packed with 97 proteins-coding genes within a 69 kbp DNA molecule (mtDNA) (Burger, Forget, et al., 2003; Burger et al., 2013). Conversely, the malaria parasite *Plasmodium falciparum* has a highly reduced 6 kbp mtDNA encoding only three proteins (*cox1*, *cox3*, and *cob*) along with fragments of SSU and LSU rRNAs (small and large ribosomal RNA subunit, respectively) (Schmedes et al., 2019; Tyagi et al., 2014). At the opposite side of the spectrum are the enormous multi-megabase mitochondrial genomes of some gymnosperms (Jackman et al., 2020). The significant disparities in gene content and genome size among protists likely result from gene loss and ongoing gene transfers from mitochondria to the nucleus, leading to mtDNA loss and functional changes (Burger, Forget, et al., 2003; Gray et al., 2001; Roger et al., 2017b). mtDNA loss and mitochondrion functional diversity are particularly evident in mitochondrion-related organelles (MROs) (Stairs et al., 2015), initially observed in anaerobic parasitic protists like *Giardia intestinalis* (mitosome) and *Trichomonas vaginalis* (hydrogenosome) (Makiuchi & Nozaki, 2014; Voleman & Doležal, 2019). However, they are further extended to free-living anaerobic protists inhabitants of low-oxygen environments (Stairs et al., 2015). The circular mitogenome topology is found to be dominant in protists, and it includes arrangements such as interlocked DNA rings in kinetoplastids of trypanosomatids (Morris et al., 2001). On the other hand, mitogenomes can be found as hundreds of linear DNA molecules ending in terminal repeats in *Amoebidium parasiticum* (putative telomeres) (Burger, Gray, et al., 2003). Despite protists' broad mitochondrial diversity spectrum, most mtDNA studies have focused on animals and plants (Bullerwell & Gray, 2004; Ladoukakis & Zouros, 2017).

Alveolates, which include apicomplexa, ciliates, and dinoflagellates, illustrate the diversity of mitochondrial genomes in protists. ciliates, such as *Paramecium aurelia* and *Tetrahymena thermophila*, typically possess a mitochondrial genome sized at around 40-77 Kbp, encoding approximately 50 proteins (Burger et al., 2000; Gray et al., 2004). On the other hand, the sister lineages apicomplexan and dinoflagellate (collectively called Myzozoa) are examples of highly reduced and fragmented mtDNA, having the smallest mitochondrial protein-coding gene repertory (Vaidya & Mather, 2009; Waller & Jackson, 2009b). The transition from gene-rich to gene-impoverished mitochondrial genomes in dinoflagellates and apicomplexa presumably involved transferring most genes to the nucleus.

The initial explorations of dinoflagellate mitochondrial genomes suggested a complex architecture with numerous recombined linear fragments and diverse gene arrangements (Norman & Gray, 1997). Early observations described the mitochondrial topology as a collection of heterogeneous small linear fragments, typically less than 10 Kbp in size. This observation was consistent across several dinoflagellate species (Chaput et al., 2002; Gray et al., 2004; Jackson et al., 2007; Nash et al., 2007; Norman & Gray, 1997, 2001; Slamovits et al. 2007a). Similarly to the apicomplexan case, dinoflagellate mtDNA exhibited reduced coding capacity (i.e., only *cox1*, *cox3*, and *coxb*) and fragmented rRNA genes. But uniquely to dinoflagellates, complex gene arrangements and structures can be found. For instance, in *Crypthecodinium cohnii*, multiple gene copies of *cox1* flanked by repeats were identified (Norman & Gray, 2001b). *Head-to-head* (*cob-cox1)* and *tail-to-tail* (*cox3-cob)* arrangements with variable spacers were frequently observed in *Amphidinium carterae*, suggesting the non-random occurrence of these arrangements (Nash et al., 2007). Unexpectedly, fused *cob-cox3* was reported in *O. marina* (Slamovits et al., 2007a). Moreover, extensive recombination was evidenced by the large fraction of non-coding DNA and gene fragments, along with inverted repeats (50-150 bp), capable of forming stem-loop structures (Nash et al., 2007). However, whether this structure facilitates recombination has not been investigated. Apparently, the expansion of non-coding DNA by recombination led to pseudogenization, gene fragmentation and insertion of full-length genes in different mtDNA contexts (Flegontov & Lukeš, 2012). On the other hand, identifying rRNA genes has been challenging due to their fragmented nature and lack of genomic data. Several fragments of LSU have been identified: LSUA, LSUD, LSUE, LSUF, LSUG, RNA2, and RNA10 (Jackson et al., 2007; Kamikawa et al., 2007). Only a few cases of SSU subunits (e.g., SSU RNA8) have been reported (Jackson et al., 2007). Initial comparisons detected substantial similarity between fragmented rRNA genes (i.e., LSU) of *C. cohnii* and small rRNA species (LSUE and LSUG) *of P. falciparum*, suggesting that fragmentation and rearrangement of mitochondrial rRNA genes began in the common ancestor of dinoflagellate and apicomplexa (Gray et al., 2004).

Later investigations uncovered additional oddities related to mitochondrial gene structure, including extensive RNA editing, trans-splicing, and complete loss of canonical start and stop codons (Imanian et al., 2012; Jackson et al., 2012a; Jackson & Waller, 2013; Nash et al., 2008). RNA editing has been observed for all three protein-coding genes and some rRNA genes. It often occurs in clusters, primarily affecting the first and second codon positions, involving approximately 2% of the sequence. The most common changes are A-G, U-C, and C-U substitutions (Flegontov & Lukeš, 2012). While editing has been reported for 25 dinoflagellate species, it has not been found in early branching members such as *Noctiluca scintillans* and *O. marina*. On the other hand, *cox3* exons trans-splicing is observed in diverse dinoflagellates (Jackson & Waller, 2013) but absent in early branching members such as *O. marina* (Slamovits et al., 2007a). Additionally, standard start (UAG) and stop codon (UAA) are absent in dinoflagellates transcripts of certain genes; instead, an in-frame stop codon UAA during the oligo-adenylation is found in *cox3* transcript of all dinoflagellates (Waller & Jackson, 2009b). Although the canonical start and stop codons were identified in the mitochondrial genome assembly of *B. minutum* (Shoguchi et al., 2015).

So far, only one mitochondrial genome assembly based on Next-Generation Sequencing (NGS) has been generated. Contrary to the fragmented topology observed for most dinoflagellates, a 326 Kbp mitogenome assembly was obtained for the coral-symbiont *Breviolum minutum*(Shoguchi et al., 2015). Most of its content is transcriptionally active non-coding mtDNA. Interestingly, rRNA genes, intergenic regions, and small RNA showed homology with *P. falciparum*, suggesting the conservation of functional non-coding DNA with alveolates. However, the current understanding of the mitogenome of dinoflagellates is hampered by limited sequencing data and poor taxon representation. Understanding mitogenome topology has primarily relied on limited techniques, such as pulsed-field electrophoresis and Southern blot analysis (Chaput et al., 2002; Jackson et al., 2007; Nash et al., 2007, 2008). Similarly, information about gene structures and arrangements was based on PCR, cloning, and expressed sequencing tags (EST), impeding intronic sequence identification, identification of gene arrangements, and the overall contiguity of the mitochondrial topology. Moreover, the scarcity of genomic sequencing data, mainly via NGS, has been a prevalent issue for most dinoflagellates as most genome sequencing efforts are primarily restricted to symbiotic species. Therefore, broader sampling will be required to understand the general topology of dinoflagellate mitogenomes as well as the conservation of non-coding sequences.

Oxyrrhis marina, an early dinoflagellate branch, shares mitochondrial genome characteristics with apicomplexa. Notably, it possesses an even more simplified gene repertoire comprising only two protein-coding genes: fused *cob-cox3* and *cox1* (Slamovits et al., 2007a). Several gene copies were identified as tandemly arranged, resembling the *P. falciparum* organization (Slamovits et al., 2007a). Additionally, *O. marina* differs from the rest of the dinoflagellates by lacking *cox3* trans-splicing and RNA editing. These findings suggest that *O. marina* represents an early stage of mitochondrial genome evolution compared to highly fragmented and recombinant genomes found in other dinoflagellates (Jackson et al., 2012b; Nash et al., 2007; Norman & Gray, 2001b). Further evidence supporting *O. marina* outsider status includes the presence of an oligo-U cap at the 5' end of mitochondrial transcripts (Slamovits et al., 2007a). The overall topology appears highly fragmented, with individual mtDNA fragments having copies of the same gene rather than different genes. However, technical limitations and sequencing sampling might impede obtaining a more detailed structure.

These early findings on revealing mtDNA features also raise new inquiries and challenges. For instance, are there remnants of single mtDNA molecules with the complete set of mitochondrial genes present in *O. marina*? To what extent is the organization of mtDNA conserved with apicomplexa, considering its position as early branching dinoflagellate? Can the fused *cob-cox3* be found as individual genes? A mitochondrial genome assembly was generated for *O. marina* to address these questions, combining long and short-read sequencing. This chapter describes highly resolved mitochondrial chromosomes for *O. marina,* providing an understanding of mtDNA recombination, gene rearrangement and conservation.

3.2 METHODS

3.2.1 Culturing and DNA Extraction

O. marina cells were isolated from Curaçao and maintained as a clonal culture at Slamovits Laboratory, Department of Biochemistry and Molecular Biology, Dalhousie University. *O. marina* is maintained in batches of 300-500 mL of F/2 media (1L artificial seawater, 1 mL $NaNO_3$, 1mL NaH_2PO_4, 1 mL trace metals solution, 0.5 mL vitamins solution) and regularly supplemented with a supplements solution composed of cholesterol (8 mg/mL^{-1}) and Coenzyme Q_{10} (100 μg/mL^{-1}) (C. D. Lowe, Mello, et al., 2011). Cultures were refreshed monthly by adding new F2 media and subjected to a 12-hour light/12-hour dark cycle at 22°C.

Several batches of DNA extraction were conducted using 300 mL of culture. First, DNA extraction was performed using the CTAB isolation protocol (Jagielski et al., 2017). Cells harvested during the exponential growth phase (approximately 15 days) were pelleted via centrifugation (3000 g for 15 mins at 4 °C). Since *O. marina* cultures are not axenic, the cell pellets were washed with sterile seawater at least twice to minimize bacterial presence by centrifugation at various speeds, and different size cells were separated. The cleaned pelleted cells were resuspended in prewarmed (60°C) 2% CTAB buffer (100 mM Tris pH 8.0, 20 mM EDTA, 1.4 M NaCl, 2% CTAB, and 1% PVP) and treated with Proteinase K (10 μg/mL) for 1.5 h at 60°C. Regular agitation (every 30 minutes) was employed to prevent cell clumping and enhance lysis. Later, samples were treated with RNase A (20 μg /mL) for 30 min at 37°C for RNA remotion. Proteins were removed by conducting two rounds of extraction with equal volumes of Phenol Chloroform Isoamyl Alcohol (Phe/Chl/IAA 25:24:1) and two additional rounds of Chloroform Isoamyl Alcohol (24:1) to remove phenol remnants (centrifugation: 10000 g for 15 min). DNA was precipitated by adding Isopropanol (12h, RT), and then samples were centrifuged (30 min, 14000 g at 4°C). The DNA pellet was washed twice with 70% ethanol and resuspended in TE buffer (pH 8). DNA samples were resuspended for at least 24h at 4°C before quality testing. Quantification was performed using a Qubit fluorometer and spectrophotometry, with an average of three measurements per sample. DNA fragments were visualized and evaluated using a 1% agarose gel. In cases of low-quality DNA samples (260/280 ≠ 1.8-2.0; 260/230 ≠ 2.0-2.2), additional rounds of Chloroform Isoamyl Alcohol were applied.

DNA was additionally cleaned up for long-read sequencing using gravity-flow columns (QIAGEN genomic-tip 20). Once DNA was clean, small-size DNA fragments (<10,000 bp) were removed using a short reads eliminator kit (Circulomics-Pacbio), following manufacturer guidelines. Essentially, this kit removes small DNA fragments through a selective precipitation procedure conducted by centrifugation. Finally, the high molecular weight DNA (HMW DNA) was resuspended in TE, and 24 h later, purity and integrity were evaluated. An additional round of Chloroform Isoamyl Alcohol was applied if it was needed. HMW DNAs were stored at -20°C.

3.2.2 Genome Sequencing, Decontamination, and Assembly

HMW DNA samples were prepared and sent for sequencing (Genome Quebec, Montreal) under the PacBio Sequell II platform (SMRTbell cell) for long-read sequencing. Briefly, PacBio SMARTbell technology is based on sequencing long circular DNA molecules formed by ligating harping adapters to each end of the DNA fragments. The DNA polymerase attaches to the single-stranded DNA molecule, adding complementary fluorescently labelled bases, and the light pulse emitted from this reaction is detected and recorded by a tiny well (zero-mode wavelength). Then, the collection of the light pulses is converted into actual nucleotides (base-calling). 9.1 million subreads were generated (93 Gbp) for PacBio sequencing.

For short-read sequencing, paired-end libraries (PE 150) were prepared at Genome Quebec and sequenced on Illumina novaseq 6000 platforms. In total, 780 million reads were generated (104 Gbp). The sequencing quality was analyzed with FastQC v.0.11.9, and adapters were removed using Trimmomatic v.0.39 (Bolger et al., 2014).

Long-read and short-read sequencing datasets were decontaminated using Centrifuge v.1.0.4 (Kim et al., 2016). Here, reads were taxonomically catalogued and classified based on a pre-build index system of microbial genomes. Bacterial contigs were removed using Recentrifuge (Martí, 2019).

A *de novo* genome assembly was generated for *O. marina*, based on the long-read dataset as a skeleton and using the highly accurate short-read dataset to correct this assembly because of the high error rate of long-reads (~7-10%). In detail, the *O. marina* genome assembly was generated from the decontaminated PacBio long-read dataset using Flye v.2.9.2 (Kolmogorov et al., 2019), followed by two rounds of polishing with Illumina short-read sequencing using Pilon v.1.24 (Walker et al., 2014). Finally, completeness was evaluated by searching for the presence of 303 highly conserved orthologues using BUSCO v.5.2.2 (Simão et al., 2015a), and by mapping RNA-seq data (*O. marina* LB1794; SRR1296907 and SRR1300472) using Minimap2 (H. Li, 2018).

3.2.3 Mitochondrial Contigs Binning, Assembly and Annotation

The initial assembly was performed using the Flye assembler with the decontaminated PacBio long reads. Mitochondrial contigs were extracted (binned) utilizing mitochondrial gene sequences of *O. marina* available on NCBI as bait. Contigs were binned using BLASTn (e-value ≤ 1E-5). This step helped to segregate mitochondrial contigs from nuclear and other non-mitochondrial sequences. To improve the quality of the contigs and the mitochondrial assembly, the binned contigs were further polished with Pilon v.1.24 (two rounds), resulting in more accurate contigs. Then, the polished contigs were used for a second fishing round using BLASTn. 125 contigs were obtained with this approach, ranging from ~7,000 bp to 130 kbp.

The Illumina short reads sequencing was also used to assemble mitochondrial contigs. The Illumina short reads were decontaminated as described for the long-read dataset. The clean reads were used to assemble the mitochondrial contigs using GetOrganelle (Jin et al., 2020). GetOrganelle is a "baiting and iterative mapping" approach that retrieves organelle associated reads and simultaneously conducts a *de novo* assembly. In our case, we used the following configuration: range of k-mer of 21, 55, 85, 105; and as seed, the option "embplant_mt" was selected. A total of 101 contigs ranging from 115 to 87,969 bp were obtained with this approach.

To reduce redundancy, contigs obtained with both approaches were reassembled and visually inspected using the assembler implemented in the Geneious R10.1 platform (option: high sensitivity). Additionally, merged contigs were interrogated for bacterial contamination using BLASTn and nr database (locally implemented). A local BLASTn search was conducted between merged contigs and *O. marina* mitochondrial genes (*cox1, cob, cox3-cob*) database implemented in Geneious to confirm the mitochondrial contigs. A total of eight hypothetical mitochondrial-related contigs were obtained.

Annotation of hypothetical mitochondrial contigs was conducted using the MITOS1 and MITOS2 automated annotators (Bernt et al., 2013), corroborated by BLASTn searches against mitochondrial genes. tRNA was further annotated using tRNAscan (T. M. Lowe & Eddy, 1997). The annotation files were transformed to gbk format using seqret tool from EMBOSS (Rice et al., 2000).

3.2.4 RNA-Seq Mapping

To study mitochondrial gene expression, RNA-seq sequencing data was retrieved from NCBI. The raw reads (SRA) for *O. marina* (SRR1296907 and SRR1300472) were downloaded under BioProject PRJNA231566 (Keeling et al., 2014). RNA reads were trimmed by Trimmomatic v.0.39 (Bolger et al., 2014) with a conservative setting (Johnson et al., 2019). Reads mapping was conducted with Hisat2 v. 2.2.1 (--max-intron-length 1000) (Kim et al., 2015). The mapped reads were calculated using SAMtools v.1.17 (H. Li et al., 2009).

3.2.5 Additional Bioinformatic Analysis

GC content and sequencing depth are essential to identify the origin of genomic regions or organelles (e.g., mitochondrial and chloroplast). Estimation of the GC content of mitochondrial contigs was obtained with seqkit V2.3.1 (seqkit -g). The mean of GC content in a window of 10 bp was calculated with BEDTools v.2.31 (bedtools nuc) (Quinlan & Hall, 2010). On the other hand, the coverage and sequencing depth estimation was conducted with the mappers Minimap2 V2.24 (minimap2 -ax map-pb) (H. Li, 2018) and bwa-mem2 V2.2.1 (https://github.com/bwa-mem2/bwa-mem2.), for long-read and short-read data, respectively. Likewise, the mean of sequencing depth was estimated in a window of 10 bp using BEDTools (H. Li et al., 2009). A synteny analysis was conducted to compare highly similar regions among mitochondrial fragments. Syntenic blocks were obtained by reciprocal BLASTn (e-value ≤ 1E-5 and 70% identity). The final representation was conducted using Circlize v.0.4.15 (Gu et al., 2014) installed in R v.1.4.11. Similarly, synteny analysis was performed between mitochondrial chromosomes and the apicomplexan *P. falciparum* (GeneBank: M76611). Final representation was carried out in gggenes v.0.4.1 (https://github.com/wilkox/gggenes).

3.3 RESULTS

3.3.1 Mitochondrial genome assembly overview

The mitochondrial genome of *O. marina* demonstrates substantial dynamism, involving recombination, genes fragmentation and duplication. Assembling this genome was particularly challenging due to the significant presence of alphaproteobacteria sequences masking the mitochondrial genome. Nevertheless, the adopted approach generated three high-resolution mitochondrial chromosomes, including fragments with complete repertory for cytochrome oxidase genes (i.e., *cox1* and *cob-cox3*). Transcriptional evidence for protein-coding and rRNA genes was observed in all three chromosomes, suggesting they remain functional. Additionally, presumed similarity restricted to protein-coding genes was detected between *O. marina* mitochondrial chromosomes and the myzozoa member *Plasmodium falciparum*.

3.3.2 Challenges and strategy for mitochondrial genome assembly

To obtain a high-resolution mitochondrial genome for *O. marina*, we had to address not only the challenges arising from the reduced and fragmented nature of its mitochondrial genome (Slamovits et al., 2007a) but also the inherent complexity resulting from its potential symbiotic relationship with bacteria (Lee et al., 2014a). We combined short (Illumina) and long reads (PacBio) sequencing technologies to address these issues and tackle the genome fragmentation problem. The presence of symbiotic bacteria in the *O. marina* culture created difficulties in obtaining a pure fraction of eukaryotic DNA. To mitigate this, the specialized software Centrifuge (Kim et al., 2016) was used to remove as many bacterial sequences as possible. Subsequently, an initial short-read assembly was performed using the GetOrganelle software with a heteroplasmy-aware configuration to reduce the misassembly (Jin et al., 2020). The combination of these approaches allowed us to obtain high confidence mitochondrial chromosomes, and the strategy implemented here may be helpful for similar tasks in other dinoflagellates or organisms with fragmented mitochondrial genomes.

3.3.3 High-quality mitochondrial chromosomes

Three high-quality linear mitochondrial chromosome candidates were obtained, displaying distinctive GC content, similar coverage distribution, and gene content (Figure 3.1). These putative mitochondrial chromosomes varied in size, ranging from 16 to 42 kbp (chromosome_1 and chromosome_3, respectively), and exhibited a low and nearly identical GC content of approximately 37-38% (Table 1). The sequencing coverage for most cases was either 100% or close to it, with the lowest being 98% for chromosome_2 for the short-read coverage (Table 3.1). High sequencing depth was obtained for the mitochondrial chromosomes based on long-read data (averaged 645x). On the other hand, sequencing depth for short-reads averaged 31x (Table 3.1). Some general features, such as low GC, topology, and the presence of the expected protein-coding genes, were consistent with the previously assembled mitochondrial genome in *S. minutum*. Nevertheless, our findings revealed significant differences in size and number of gene copies (Table 3.2).

3.3.4 Mitochondrial chromosome synteny and topology

A synteny analysis was conducted to gain deeper insights into the structural similarity among the chromosomes (Figure 3.1). The results unveiled syntenic patterns and rearrangements encompassing coding and non-coding regions. For coding regions, most syntenic blocks involved regions linked to *cox1*. To a lesser degree, blocks include LSUE and LSUG. No synteny was observed involving full *cob-cox3*, except for a fraction of *cob* shared between chromosome_2 and 3. Additionally, syntenic blocks involving exclusively non-coding regions were frequently observed as well. Interestingly, blocks that included a *cox1* fragment (*cox1.3*) in chromosome_3 were observed along with non-coding counterparts (chromosome_2). These observations suggest that a substantial fraction of non-coding mtDNA might have originated as pseudogenes and gene fragments mutated beyond recognition. Still, a considerable fraction of non-coding mtDNA space showed no homology, suggesting large-scale recombination and additional recombinant fragments should be involved. On the other hand, inversions were the most frequently identified arrangement (see darker ribbons, Figure 3.1), suggesting that recombination may have an active role in remodelling mitochondrial chromosomes. Inverted repeats have been proposed for module recombination in dinoflagellate mitochondria (Nash et al., 2007), but they seem absent in *O. marina*.

Regarding topology, all mitochondrial chromosomes were found to be linear, and no circular topology was detected. According to these findings, *O. marina* mtDNA topology can be described as a collection of mtDNA fragments or chromosomes, including chromosomes with multiple gene copies and chromosomes with the full set of protein-coding genes (i.e., *cox1* and *cob-cox3*). Attempts to recover circular mitochondrial chromosomes resulted in spurious assemblies, especially merging inverted regions at the chromosome ends. The detection of frequent rearrangements and the significant non-coding mtDNA fraction without homology suggests true topology involves several chromosomes or mtDNA fragments.

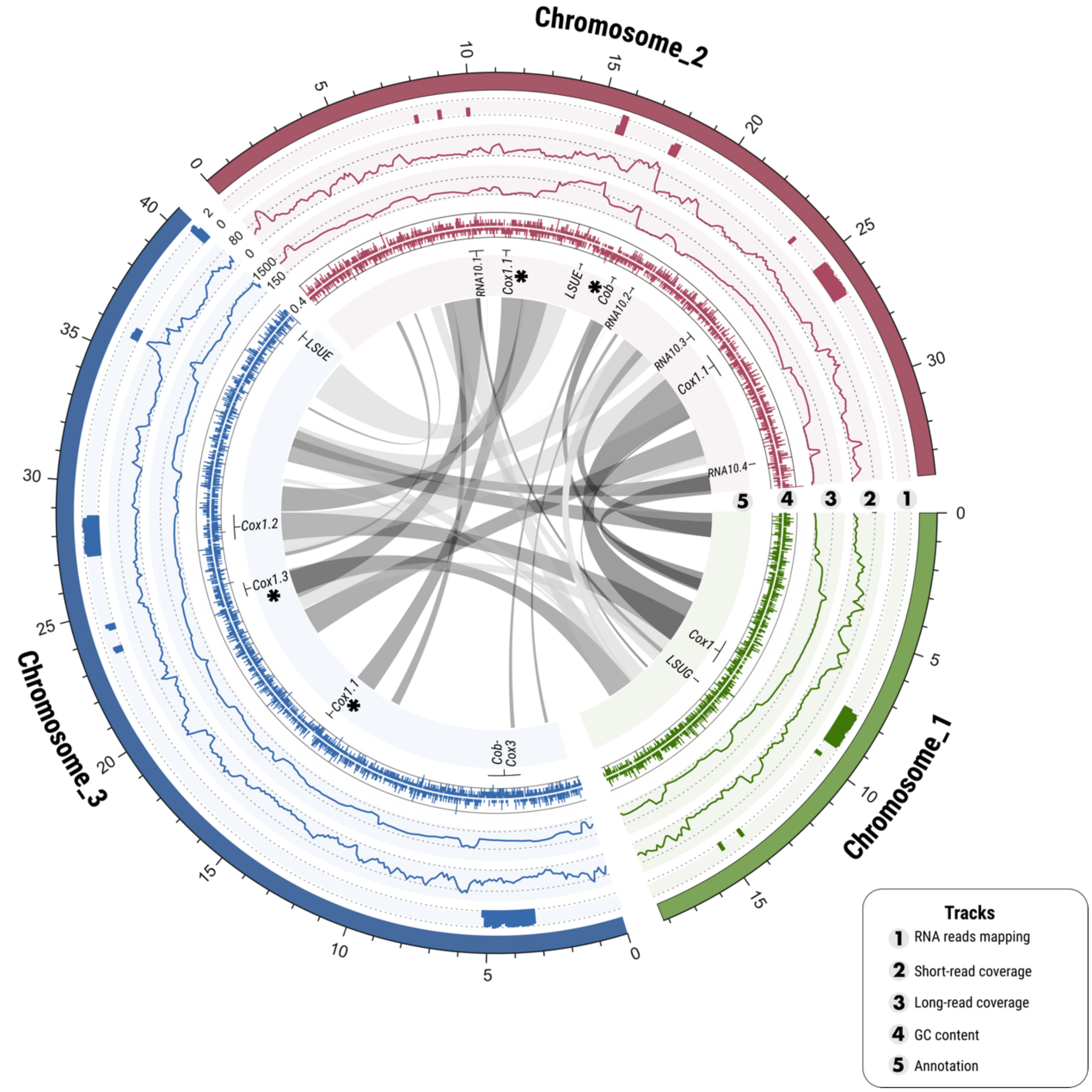


Figure 3.1. *O. marina* mtDNA shows extensive rearrangements and redundant gene content.

Circos plot displaying sequencing features for the three mitochondrial chromosomes. Each chromosome and its features are coloured differentially. From outermost to the innermost track: log10 of mapped RNA reads (track 1); average short-read sequencing depth (track 2) with average for chromosome_1 of 30x, chromosome_2 of 29x and chromosome_3 of 35x; average long-read sequencing depth (track 3) with average for chromosome_1 of 711x, chromosome_2 of 528x and chromosome_3 of 698x; GC-content (track 4) with average for chromosome_1 of 39.4%, chromosome_2 of 38.8%, and chromosome_3 of 37.7%. Annotation and location of protein-coding and rRNA genes (track 5). Stars represent fragmented genes or pseudogenes. Dark grey ribbons represent inverted homologous regions among chromosomes determined by reciprocal BLASTn (e-value ≤ 1E-5), and dark regions represent blocks with the same orientation. Supplementary Table 1 provides additional information, including annotation scores, p-values, and gene coordinates.

Table 3.1. Summary of sequencing features in assembled mitochondrial chromosomes of *O. marina.*

Fragment	GC (%)	Size (bp)	Illumina		PacBio	
			Coverage (%)	Mean Depth	Coverage (%)	Mean Depth
Chromsome_1	39.42	15926	99	30	100	711
Chromsome_2	38.85	33871	98	29	100	528
Chromsome_3	38.79	40613	100	35	100	698

Table 3.2. Comparison of mitochondrial genome assemblies among dinoflagellate and *Perkinsus*

	Dinoflagellates				Perkinsozoa
Feature	chro_1	chro_2	chro_3	*S. minutum*	*Perkinsus*[c]
Genome size (kbp)	16	33	41	326	40-95
Sequencing depth[a]	30;711	29;528	35;698	>100	2065-66225;38-1218
GC (%)	39.42	38.85	38.79	35.7	13-18
Topology	Linear	Linear	Linear	Linear	Linear(1)-Circular (3)
Protein coding genes[b]	1	2(3)	2(4)	3	3
rRNA gene[b]	1	2(5)	1	12	3(4)-6(8)
tRNA gene	–	–	–	5	–

[a] short read sequencing depth; Long read sequencing depth

[b] the total number of genes is shown in parenthesis

[c] range of sequencing features for four species of *Perkinsus*

3.3.5 Simple and fragmented gene repertory

O. marina showed reduced coding capacity encompassing multiple gene fragments. The protein-coding genes were annotated using the automatic mitochondrial genome annotator MITOS1-2. However, further curation was necessary because the lack of canonical start and stop codons in dinoflagellate genes may lead to false negatives. Despite the gene repertoire being limited to only two protein-coding genes, several additional copies and fragments were identified (Figure 3.1, Supplementary Table 2). Fragmented copies of *cox1* were commonly found as single copies or tandemly arranged (chromosome_3, Figure 3.1 and Figure 3.2). Most of these fragments were pseudogenes lacking expression patterns or showing diminished expression (e.g., *cox1.3* in chromosome_3). Likely, *cob* fragments identified in chromosome_2 resulted from a fragmentation of *cob-cox3* rather than a stand-alone gene copy. Among ribosomal RNA genes, several fragments of LSU rRNA, such as LSUE, LSUG, and RNA10, were identified. Specifically, LSUE was identified in chromosomes 2 and 3; in both cases, fragments appear to be transcriptionally active (Figure 3.1). LSUG was only identified in chromosome_1, and no evidence for expression was found. On the other hand, four fragments of RNA10 were identified on chromosome 2 (Figure 3.1 and Figure 3.2), and expression was detected in only one of these fragments (RNA10.1). These results are similar to those reported previously (Slamovits et al., 2007a) and confirm the fragmentation of rRNA genes in *O. marina*, likely derived from a myzozoan common ancestor shared with apicomplexa. No tRNA genes were identified, suggesting that *O. marina* mitochondria have lost the coding capacity for tRNA. It is possible that tRNA genes are now encoded in the nucleus and imported from the cytoplasm, which has been observed to be quite common among eukaryotes (LeBlanc et al., 1999; Mahapatra & Adhya, 1996). Several non-coding regions were identified, particularly between the coding regions. Consequently, repeat annotations were conducted using RepeatMasker to identify potential transposable elements or repeats that might account for these non-coding regions. As a result, approximately 1% of mtDNA were annotated as simple and low-complexity repeats, and all elements reported were found only once (Supplementary Table 3). This suggests that fragmented genes, rather than any mobile element, constitute the main component of non-coding DNA in *O. marina* mitochondria. Interestingly, the mitochondrial genome of *O. marina* has expanded its coding capacity with new gene copies that paradoxically end up fragmented and non-functional. Then, the genome size increased, but not the coding capacity.

3.3.6 Mitochondrial gene expression and homology with Apicomplexa

Motivated by previous findings reporting the expression of the entire mtDNA molecule (~300 kbp) (Shoguchi et al., 2015), the functional capacity of *O. marina's* mtDNA was determined through transcriptome reads mapping. The expression patterns were predominantly confined to coding regions, mostly to unfragmented copies of *cox1*, LSUE and *cob-cox3* loci (see Figure 3.1, track 1). Additionally, the continuous transcriptome mapping pattern observed for the *cox3-cob* locus confirmed the fused status of this gene. Marginal expression was detected for non-coding regions, such as the region between LSUE and *cox1.2* in chromosome 3 or flanking regions of cox1.3. These instances might be attributed to a mapping error or pseudogene residual transcriptional activity. These results agree with the expected decline of fragmented genes' transcriptional activity and EST data where only full-gene copies were obtained (Nash et al., 2007). This analysis suggests that mitochondrial chromosomes may still have functional genes, but it cannot rule out that most of the mitochondrial transcripts are generated by mitochondrial nuclear-encoded genes such as LSU and SSU identified from EST data (Lee et al., 2014a).

On the other hand, it is reported that *O. marina* exhibited striking similarities with members of Apicomplexa and includes reduced gene content and fragmented rRNA genes. To assess synteny, homology comparisons were conducted through reciprocal BLASTn (with 70% identity and e-value ≤ 1E-5) between *O. marina* chromosomes and the mitochondrial genome of *P. falciparum* (Figure 3.2). Limited similarity was detected, mostly restricted to *cox1* and LSUE. In detail, homology for *cox1* was shared among the *O. marina* chromosomes and *P. falciparum*; however, no similarity involving fragmented *cox1* copies was detected. Likewise, low similarity was detected for LSUE between the two species (Figure 3.2). Despite the large number of rRNA fragments (26) in *P. falciparum*, no additional homologue fragments were identified in *O. marina*. Additionally, no significant sequence similarity was detected involving non-coding DNA regions. These findings, including non-coding sequences, contrast with the numerous rRNA gene fragments shared between *B. minutum* and *P. falciparum* (Shoguchi et al., 2015), including non-coding sequences. Interestingly, these findings raise questions about how ancestral mtDNA organization of apicomplexa appears to be more conserved in derived dinoflagellates as opposed to the early branching such as *O. marina*.

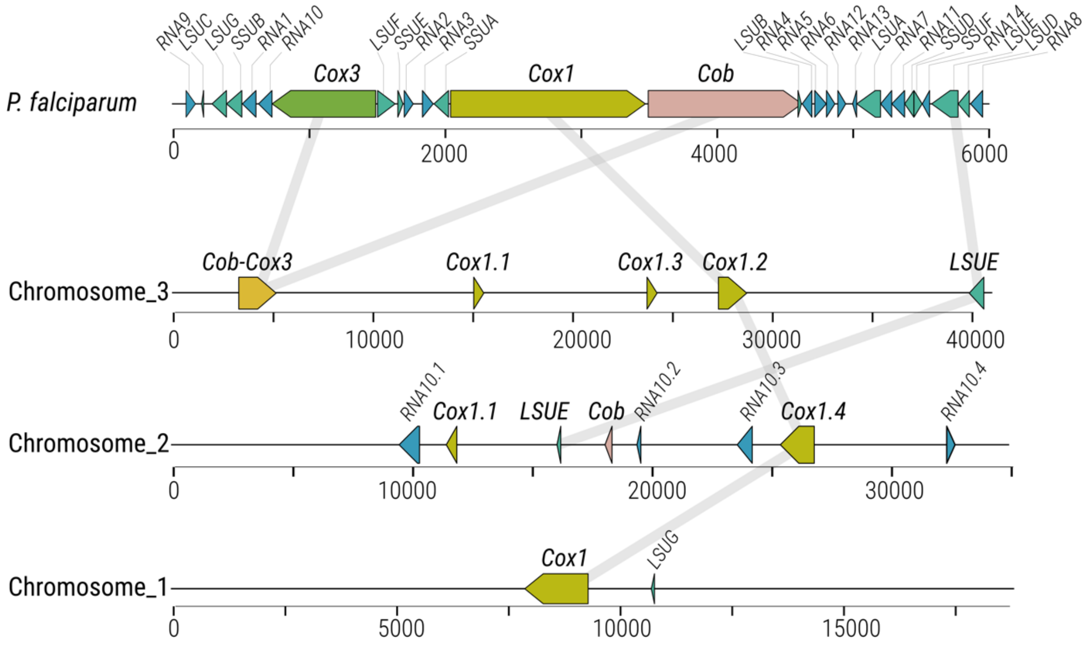


Figure 3.2. Mitochondrial chromosomes homology with *P. falciparum*.
Homology comparison among mitochondrial genes of *P. falciparum* and *O. marina* mitochondrial chromosomes through BLASTn (e-value ≤ 1E-5). Genes are coloured according to the gene legend. *P. falciparum* mitochondrial genome and its annotation were obtained from NCBI (GenBank accession number: M76611).

3.4 DISCUSSION

O. marina exhibited one of the most reduced and gene-impoverished mitochondrial genomes while retaining an organizational pattern similar to Apicomplexa. *O. marina* mtDNA coding capacity has been significantly minimized to only two protein-coding regions, i.e., *cox1* and *cob-cox3*, representing a unique characteristic for this lineage (Figure 3.3). We found that the arrangement of fragmented mitochondrial gene copies in tandem is consistent with previous findings (Slamovits et al., 2007b). Initially, *O. marina* mtDNA's topology was described as a collection of multiple mtDNA fragments, each encoding either single genes or multiple copies of the same gene (Slamovits et al., 2007b). However, current findings have expanded upon this initial description by revealing mtDNA fragments with both genes and multiple copies of different genes in the same fragment. Likely, this was not previously recognized due to the lack of sequencing. This discussion will be based on this previous finding and how the current research improves the mitochondrial genome's resolution, elucidating the general topology and providing new insight into these early findings.

3.4.1 Strategy implementation and difficulties

Long-read sequencing can effectively resolve complex mitochondrial genomes such as *O. marina*. Many fragments of the mitochondrial genome coexist in *O. marina*, and likely they have undergone recombination. PacBio long-read sequencing dealt with such complex structures by sequencing single DNA molecules. The average read length mapped to mitochondrial chromosomes was ~ 11 kbp, covering at least 30-60% of the chromosome length in a single DNA molecule. This contrasts with short-read sequencing with a hundred base pairs long, which are prone to misassembling and can lead erroneously to inferring heteroplasmy. However, implementing heteroplasmy-aware software such as GetOrganelle (Jin et al., 2020) and other alternatives such as NOVOplasty (Dierckxsens et al., 2017) helps reduce short-read incorrect mapping and assembly. On the other hand, implementing new approaches such as mitochondrial isolation/enrichment based on density-gradient centrifugation may help avoid bacterial contamination and reduce sequencing effort. In summary, PacBio long-read sequencing is a powerful method for sequencing fragmented mitochondrial genome configurations by generating long DNA molecules that accurately resolve complex structures.

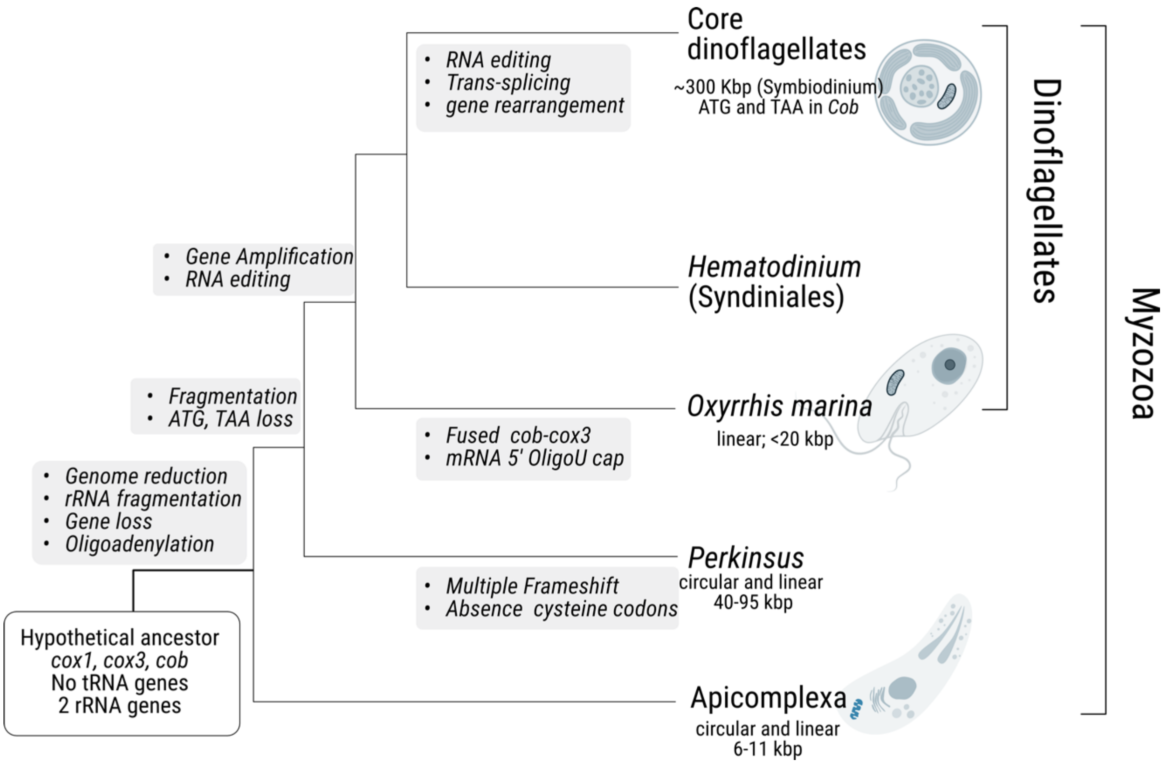


Figure 3.3. Schematic representation of mitogenome evolution in Myzozoa.
The tree roots represent the hypothetical mitogenome ancestor state (white text box), and subsequent modification in the hypothetical common ancestor as well as at individual lineages are shown in the gray text boxes. Additionally, information about mitogenome topology and size based on the genome assembly (Berná et al., 2021; Gornik et al., 2022; Shoguchi et al., 2015) is shown below the taxa or group name. Representations are based on previous reconstruction of the mitogenome evolution of alveolates (Gagat et al., 2017) and Apicomplexa (Berná et al., 2021).

3.4.2 General topology and genome architecture

O. marina displays fragmented mtDNA having a multicopy gene repertory organized in linear chromosomes. Mitochondrial chromosomes exhibit compact coding regions interspersed with extensive non-coding regions. Notably, one chromosome encompasses the complete gene set, i.e., *cob-cox3* and *cox1*, along with additional fragmented copies for *cox1*. This finding complements prior observations where mtDNA fragments carried single genes or copies of the same gene but never *cob-cox3* and *cox1* within the same mtDNA fragment (Slamovits et al., 2007a). The current findings are in concordance with the general organization of mtDNA (e.g., gene duplication, gene fragmentation, and recombination) found in the core of dinoflagellates such as *Alexandrium catenella*, *A. carterae*, and *C. cohnii* (Kamikawa et al., 2007; Nash et al., 2007; Norman & Gray, 2001b) (Figure 3), including the early member *Hematodinium* and more distantly *Perkinsus* (Gornik et al., 2022; Jackson et al., 2012b). Despite the initial investigation of mtDNA of dinoflagellates relying on poor sequencing sampling and limited methods (PCR and EST exploration), the current finding validates the early observations. Additionally, *O. marina* mtDNA topology suggests the presence of a complete mitochondrial gene set within a single mtDNA molecule is a conserved feature, and further fragmentation in the core dinoflagellates may have occurred.

There is a proposition that dinoflagellates may possess large mitochondria topologies (Waller & Jackson, 2009b). It assumes that recurrent recombination observed in dinoflagellates mitogenome could lead either to fragmentation or expansion. Such might be the case in *S. minutum*, reporting mitochondrial assembly of approximately ~ 326 kbp (Shoguchi et al., 2015). This topology contradicts the general fragmented nature of mtDNA described for the core of dinoflagellates. Because this assembly was based on short-read sequencing, it raises concerns about its contiguity, considering that a high level of recombination may mislead an assembly based solely on short reads. We carefully explored the idea of large mtDNA topology and assessed it on *O. marina*. However, all forced assembled variants inevitably led to long chimeric fragments with inconsistent sequencing depth. The size of *O. marina* mitochondrial chromosomes, approximately 30 kbp on average, corresponds with the estimated 30 kbp mtDNA fragments observed in dinoflagellates through Southern blot analysis (Jackson et al., 2007; Norman & Gray, 2001b). This suggests that similar processes underlie mitochondrial genome architecture in most dinoflagellates. The identification of various rearrangements, including inversions and relocations of genic regions, points to extensive recombination shaping the *O. marina* mitochondrial genome. This phenomenon is consistent with predictions for the core of dinoflagellates (Waller & Jackson, 2009b). The mitochondrial genome topology of *O. marina* likely consists of a collection of linear chromosomes that do not form part of a larger, integrated structure. Here, a fully resolved set of mitochondrial chromosomes is described, but the complete understanding of the overall *O. marina* mitochondrial genome topology remains to be solved.

3.4.3 Gene organization and non-coding DNA

Two distinct patterns of gene organization were observed: isolated single gene copies separated by non-coding DNA regions of approximately 5-10 kbp or genes closely spaced with longer non-coding regions in between. This arrangement mirrors prior findings (Slamovits et al., 2007a), but now with fragmented gene copies fully resolved. Moreover, genes were present as single copies or, in some cases, as discrete tandem arrays. For example, *cob-cox3* was found as a single copy on chromosome_3, while *cox1* copies were arranged of in tandem (chromosome_3). Despite several copies detected for *cox1*, only full-length copies were expressed, suggesting selective constraints might act differentially on the gene copies. Additional analysis of gene sequence structure (e.g., nucleotide substitution rates) may help elucidate the evolutionary processes underlaying these gene copies. The fusion of *cob-cox3* was confirmed through transcriptome read mapping, indicating significant expression as a single gene. However, considering their roles in different electron transport complexes, the generation of functionally distinct proteins remains unclear. Nevertheless, the presence of a *cob-cox3* polycistronic transcript cannot be ruled out (Slamovits et al., 2007a). A stand-alone *cob* fragment was detected (chromosome_2), likely representing a fragment of fused *cob-cox3*. Eventually, frequent recombination provides chances for this novel arrangement to occur, and the same process may lead to fragmented non-functional gene copies.

The non-coding portion of the *O. marina* mtDNA constitutes a substantial fraction of approximately 88% of the total mtDNA. This is comparable to the non-coding fraction observed in *A. carterae* (~85%) (Nash et al., 2007) and *B. minutum* (>90%) (Shoguchi et al., 2015). The repeat content annotation reveals the presence of simple repeats, accounting for about 1% of the mitochondrial chromosomes, like the 1.8% reported for *B. minutum* (Shoguchi et al., 2015). Inverted repeats, abundant in *A. carterae* and *C. cohnii* (Nash et al., 2007, 2008; Norman & Gray, 2001b) and posited to facilitate recombination by forming a stem-lop structure, are absent in *O. marina*, consistent with the early nature of its mitochondrial genome (Jackson et al., 2012b). Additionally, several syntenic blocks shared between coding regions and non-coding reciprocated regions were observed. This suggests that fragmented genes likely contribute to the overall non-coding DNA content of *O. marina* mitochondria. However, most of the non-coding DNA of dinoflagellate mitochondrial genomes remain poorly understood, as well as its sources and origin. Overall, chromosome structures allow us to get additional insight into non-coding DNA sources and processes, such as recombination, that ultimately seem to shape the mitochondrial genome of *O. marina*.

3.4.4 Fragmented ribosomal genes

A distinctive feature shared with Apicomplexa is the fragmented nature of SSU and LSU rRNA genes (Figure 3.3). This study confirmed the presence of fragments of LSU rRNA in *O. marina*, corresponding to previously identified fragments of LSUE, LSUG, and RNA10 (Slamovits et al., 2007a). These fragments are homologues to those observed in *Karlodinium micrum* and *P. falciparum* (Feagin et al., 1997; Jackson et al., 2007). The predicted arrangements were found to match homologous structure and interaction for LSU in dinoflagellates, *P. falciparum* and *Escherichia coli* (Slamovits et al., 2007a). Even though the resolution of the mitochondrial genome increased substantially, additional LSU fragments were not detected when *O. marina* chromosomes were compared to *F. falciparum* mitogenome. It is postulated that dinoflagellates are not forced to encode larger rRNA molecules. Instead, base-pair interactions among rRNA fragments seem to be sufficient to assemble a functional rRNA molecule like another organism (Adams & Palmer, 2003; Jackson et al., 2012b). We corroborated the ancestral state of rRNA gene fragmentation in all *O. marina* mitochondrial chromosomes, indicating that fragmentation observed in Apicomplexa has been preserved in all dinoflagellates (Nash et al., 2008; Waller & Jackson, 2009b) (Figure 3). On the other hand, we have no evidence for the presence of the SSU rRNA gene on *O. marina* mtDNA, and its presence has been elusive in dinoflagellates, only identified in *K. micrum* (Jackson et al., 2007). It is unclear whether SSU has been lost or translocated to the nucleus in the dinoflagellates (Waller & Jackson, 2009b). Likewise, no tRNA was identified and probably relied on nucleus-encoded tRNA to translate their proteins.

3.4.5 Future directions

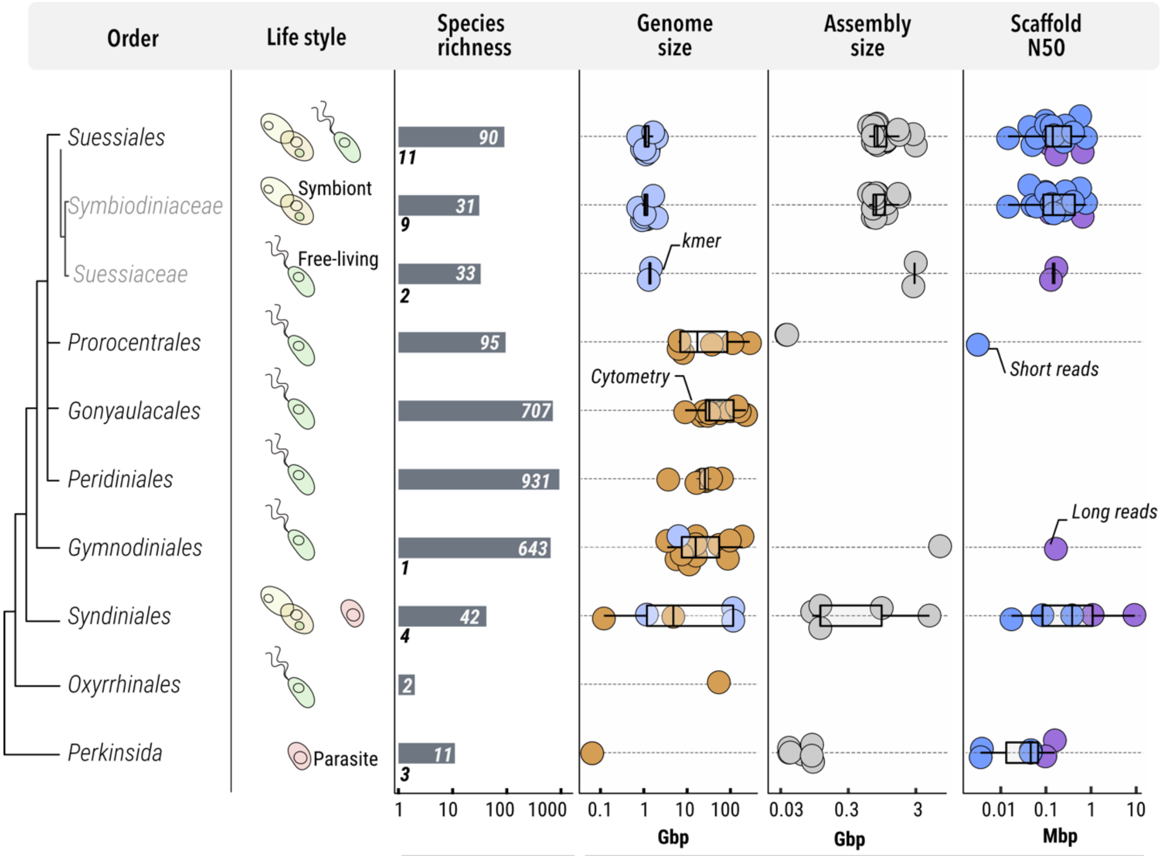
The results presented here aimed to provide a general view of the mitochondrial genome organization of *O. marina*. However, further analysis will be required to complement these findings. Firstly, additional investigation into the structure of mitochondrial genes and focusing specifically on identifying non-canonical start and stop codons will help delineate the boundaries of coding regions more precisely. Secondly, new efforts to identify additional mitochondrial fragments will be valuable to explore the evolutionary trends among different mitochondrial variants by analyzing the gene's nucleotide composition and codon bias. Expanding the scope of synteny analysis to include multiple species of dinoflagellates, as well as P. *falciparum*, will provide robust confirmation of the observed syntenic patterns. Lastly, further efforts to identify conserved repeats will provide additional insights into their potential role in recombination processes.

CHAPTER 4 NUCLEAR GENOME OF *OXYRRHIS MARINA*

4.1 INTRODUCTION

Dinoflagellates are unicellular eukaryotes with a fundamental role in aquatic environments and one of the largest groups of protists (i.e., ~50% of protists richness) in the world's oceans (Le Bescot et al., 2016). The nuclear structure of dinoflagellate cells diverges considerably from mainstream eukaryotes. They exhibit permanently condensed liquid-crystalline chromosomes (LCC) with birefringent properties (P. J. Rizzo, 2003). Although their chromosomes have, for decades, assumed to be devoid of nucleosomes, it was later found that they do possess a divergent set of histones, albeit in low abundance (Marinov & Lynch, 2015). In fact, emerging evidence indicates that a small fraction of the genome appears to show nucleosomal organization based on micrococcal nuclease digestion patterns (Gornik et al., 2012). Nevertheless, proteins such as HLP (histone-like protein) and DVNP (dinoflagellate viral nucleoprotein) are suggested to assume the function of packing the bulk of genomic DNA (Janouškovec et al., 2017). Additional molecular oddities include a reduced prevalence of transcriptional regulation (Zaheri & Morse, 2022), universal mRNA trans-splicing, the abundance of certain modified nucleotides such as 5-hmU (hydroxymethyluracil, which can substitute up to 68% of thymine), and remarkably large genomes with highly redundant gene sets. This unique nuclear configuration offers many opportunities to understand unconventional mechanisms of chromosome organization, DNA packing mechanisms, and transcriptional regulation, among other inquiries.

Over the past decade, new technologies such as long-read sequencing have enabled efficient sequencing and assembly of dinoflagellate genomes. Although these efforts have primarily focused on a limited number of symbiotic species with relatively small genome sizes (Figure 4.1) (i.e., 0.8 to 2.7 Gbp), (Aranda et al., 2016a; González-Pech et al., 2021; Gornik et al., 2015; John et al., 2019; Lin et al., 2015; Liu et al., 2018; Shoguchi et al., 2013, 2018; Stephens et al., 2020), these studies have provided valuable knowledge revealing the genomic basis for adaptation to symbiotic lifestyles and have also shed light on long-standing questions regarding chromatin organization and chromosome structure (Marinov & Lynch, 2015; Nand et al., 2021). Nevertheless, understanding of the genome organization in most free-living dinoflagellates is very limited because of a lack of sequencing data (Figure 4.1). Notably, some species may have genome sizes exceeding 200 Gbp, e.g., *Prorocentrum* and *Alexandrium* (Hou & Lin, 2009; Lin, 2011). However, these estimates based on fluorescence and flow cytometry (Figure 1) might be inflated due to the presence of a significant fraction of repetitive DNA and differences in chromatin that may affect the way in which fluorescent dyes interact with DNA (John et al., 2019; Stephens et al., 2020). Therefore, the actual genome sizes could be considerably smaller than initially calculated. This discrepancy encourages further endeavours of dinoflagellate sequencing projects.



Source: AlgaeBase *Source: NCBI-Sep-2022*

Figure 4.1. Genome sequencing bias in dinoflagellate.
Sequencing efforts have been focused primarily on symbiotic dinoflagellate (Suessiales). The bar chart showcases the distribution of species across different orders, with the number of genome assemblies indicated below each bar. Genome sizes, determined through cytometry and sequencing data, are represented in orange and light blue, respectively. Sequencing technologies are distinguished by blue for short reads and purple for long reads. Metadata was sourced from AlgaBase and NCBI (as of September 2022). Phylogenetic representation is based on Janouškovec et al., 2017.

Dinoflagellate genomes are characterized by their redundancy, with a substantial presence of repetitive elements of varying complexity and numerous gene copies. Large genome assemblies, i.e., *P. glacialis*, are substantially enriched in repetitive sequences, accounting for over 65% of the total genome content (Stephens et al., 2020). Gene redundancy is also a prevalent feature, where protein-coding genes are often found organized in extensive unidirectional tandem arrays (Nand et al., 2021; Shoguchi et al., 2013b; Stephens et al., 2020). Nevertheless, a more comprehensive sequencing effort targeting free-living dinoflagellate species is necessary to understand the prevalence of these genomic features across the diversity of the group.

Oxyrrhis marina, a free-living heterotrophic dinoflagellate with a broad distribution, is an ideal candidate for genome sequencing for several compelling reasons. Firstly, it can be easily cultured and maintained in laboratory conditions, ensuring a readily available source of DNA material for multiple rounds of sequencing. Secondly, it possesses a genome size of approximately ~30-50 Gb (Sano & Kato, 2009), which is more manageable than most free-living dinoflagellates' exceptionally large genomes (over 55 Gbp). Thirdly, it represents an early branch within the dinoflagellate lineage, making it a valuable model for investigating the origins of unique dinoflagellate features (e.g., genome enlargement, histone-like protein, spliced leader mRNA) (Figure 4.1). A previous study based on the analysis of expressed sequence tags (EST) from *O. marina* revealed critical features, including extensive gene redundancy and the presence of twenty variants of DVNPs (Lee et al., 2014b). However, the lack of genomic data has hindered a comprehensive understanding of *O. marina* genome organization.

This chapter introduces the initial genomic survey of *O. marina*, utilizing a combination of various sequencing technologies. It is one of the few instances where genomic data has been generated for a free-living dinoflagellate using long-read sequencing technology. The results presented here provide insights into the gene content, organization, and ploidy level, revealing a significant number of hypothetical genes with high redundancy. Genes are frequently organized in unidirectional blocks, and a notable dominance of LTR-retrotransposons was identified.

4.2 METHODS

4.2.1 DNA extraction, sequencing, and assembly

The procedures for DNA preparation, sequencing and assembly have been explained in more detail in Chapter 3. In summary, *O. marina* cultures were grown in F2 media and supplemented with cholesterol and coenzyme Q10. DNA extractions were conducted using the CTAB DNA isolation protocol, followed by purification and quantification. DNA underwent cleanup with gravity-flow columns, including selective removal of small fragments. HMW DNA was sequenced using PacBio SMARTbell technology, resulting in approximately 9 million reads. *De novo* assembly was conducted using Flye and polished with Illumina short reads via Pilon. Completeness was assessed using BUSCO and RNA-seq data mapping.

4.2.2 Decontamination of *O. marina* assembly

The *O. marina* culture is not axenic, and bacteria are commonly found in close contact with the cells (Lee et al., 2014b). Several washes with sterile seawater and PBS buffer were conducted to reduce the bacterial fraction during the DNA extraction process. Through the initial decontamination (chapter 3), reads taxonomically affiliated to bacteria were removed using the software Centrifuge. A tailored decontamination step was conducted to remove sequences and contigs from bacteria isolated from the *O. marina* culture. Eight bacterial genomes were sequenced by Alexander Mora as a part of his PhD research (at the Slamovits laboratory) using nanopore technologies ONT (Oxford Nanopore Technologies). Briefly, ~1 µg of HMW DNA was used to prepare ONT libraries using the ligation kit SQK-LSK114. Libraries were sequenced using the Flongle flow cell, and the basecalling was conducted using Guppy v.6.0.6 (Wick et al., 2019).

The bacterial genomes (i.e., *Oceanibaculum sp.*, *Halassospira sp.*, *Alcanivorax sp.*, *Alteromonas sp.*, *Maritalea sp.*, *Muricauda sp.*, *Oceanicaulis sp.*, and *Devosia sp.*) were mapped to assembly using Minimap2 v.2.24 (H. Li, 2018). The undesired bacterial contigs were removed employing SAMtools v.1.17 (H. Li et al., 2009). Given the common occurrence of bacterial and viral genes in dinoflagellate genomes, 156 contigs showing homology for bacterial genes were retained for further curation and LGT assessment. The assembly sequencing depth and coverage were estimated using Minimap2 and SAMtools.

4.2.3 Gene prediction and genome annotation

Genes models were predicted and annotated using the Funannotate package v1.8.15 (https://github.com/nextgenusfs/funannotate), which encompasses five ab-initio predictors: AUGUSTUS (Stanke & Morgenstern, 2005), SNAP (Korf, 2004), glimmerHMM (Majoros et al., 2004), CodingQuarry (Testa et al., 2015), and GeneMark-ES/ET (Ter-Hovhannisyan et al., 2008). Ab-initio predictors are statistical models that use gene templates and DNA sequence information such as start and stop codons, splice sites, polypyrimidine tracts, and codon usage patterns to predict genes (Z. Wang et al., 2004). These predictors were trained with the genomic data and combined with evidence-based predictions through EVidenceModeler (Haas et al., 2008), which aligned transcriptome data to the genome to compile accurate gene models. The Funannotate pipeline shell scripts were thoughtfully implemented and customized by Dr. Joran Martijn at Dr. A. Roger's computer cluster.

In detail, the cleaned assembly underwent seven steps, starting with removing contigs smaller than 5000 bp using "funannotate clean," sorting and renaming the contigs with "funannotate sort." Then, soft-masking the repetitive elements using the RepeatModeler/RepeatMasker wrapper implemented in the "funannotate mask" script. For the training step, the RNA-seq reads retrieved from NCBI under SRR1296907 and SRR1300472 (BioProject PRJNA231566 (Keeling et al., 2014)) were trimmed with Trimmomatic v.0.39 (Bolger et al., 2014) and parsed with BBtools (BBMap - Bushnell B. - sourceforge.net/projects/bbmap/). Subsequently, "funannotate train" was executed, wherein transcriptome reads were mapped to the masked genome assembly, producing a splice-site aware alignment and a genome-guided transcriptome assembly via Trinity v2.13.2 (Grabherr et al., 2011). Gene prediction was carried out using "funannotate predict," beginning with the training of AUGUSTUS v.3.5.0 using evidence for gene models gathered in training steps. Additional ab-initio gene prediction was performed using glimmerHMM v. 3.0.4, SNAP v. 2013-07-28, CodingQuarry v.2.0, GeneMark-ES/ET v.4.7.1, and PASA. Finally, ab-initio predicted models were merged with EVidenceModeler v.2.1.0 to produce a final consensus set of gene models. UTR regions were predicted, and gene models were refined using "funannotate updated." Annotation involved searches against various databases, including UniProtDB v.2022_03, EggNog v. 1.0.3 (Huerta-Cepas et al., 2017), MEROPS v.12.0 (Rawlings et al., 2018), CAZYme, BUSCO2 (Simão et al., 2015b), PFAM-A 35.0 (Finn et al., 2014) and a local non-redundant (nr) Diamond protein database (02-03-2022) with an e-value threshold of ≤ 1E-5. Final gene prediction statistics were obtained using AGAT v1.2.0 (https://github.com/NBISweden/AGAT).

4.2.4 Repeat content annotation

The repeated content was initially annotated based on homology to repetitive elements using RepeatModeler v1.0.11 (https://www.repeatmasker.org/RepeatModeler/). Kimura distances for each repeat sequence were calculated using the "calcDivergenceFromAlign.pl" script, and the repeat distribution landscape was generated with the "createRepeatLandscape.pl" script. A refined repeat annotation, focusing on mobile element structures, was conducted using the EDTA software (Ou et al., 2019), which combines multiple tools, including LTR_FINDER (Xu & Wang, 2007), LTRharvest (Ellinghaus et al., 2008), HelitronScanner (Xiong et al., 2014), TIR-Learner (Su et al., 2019), LTR_retriever (Ou & Jiang, 2018). Additionally, 50 curated repeated elements identified with RepeatModeler were incorporated into the EDTA pipeline as a curated library for homology-based identification. A final non-redundant TE library was obtained for *O. marina*, which was visualized using the "geom_bar" function from the ggplot2 package (Wickham, 2016) within R v.4.0.3.

4.2.5 Assembly completeness and ploidy assessment

The assembly completeness was assessed by searching for highly conserved orthologs and through transcriptome mapping. Completeness evaluation involved Hmmsearch of BUSCO v.3.0.2 Eukaryota_odb9 database (303 genes), BUSCO v5.2.2 Alveolata_odb10 database (171 genes) (Simão et al., 2015b) and CEGMA v2.5 (248 genes) (Parra et al., 2007). Subsequently, RNA-seq data (SRR1296907 and SRR1300472) (Keeling et al.,

2014) were mapped to assembly using Hisat2 v. 2.2.1 (--max-intron-length 1000) (Kim et al., 2015). The percentage of mapped reads was determined using SAMtools (H. Li et al., 2009). Completeness was visualized using the geom_bar of ggplot2 (Wickham, 2016). The ploidy evaluation was conducted with NGSploidy v.3.1.3 (min_cov =7) (Augusto Corrêa dos Santos et al., 2017).

4.2.6 Analysis of gene organization and retrogene identification

The gene organization was investigated using information derived from the gene annotation. Gene arrangements and copies were detected through MCScanX (Y. Wang et al., 2012) by analyzing hits obtained from all-against-all BLASTP of the translated genes (e-value ≤ 1E-05). The scripts "duplicate_gene_classifier" and "gene_type" were employed to detect syntenic and homologous gene pairs (syntelogs) and classify them based on their organization, such as singleton, dispersed, proximal, tandem, and segmental (Y. Wang et al., 2012). Moreover, gene arrangement orientation was also evaluated by comparing it with *Plasmodium falciparum* 3D7 (GenBank *accession:* GCF_000002765.5) and *Symbiodinium microadriaticum* CCMP2467 (GenBank accession: LSRX00000000.1). Specifically, the frequency of changes in gene orientation within a sliding window of ten consecutive genes was examined (Shoguchi et al., 2013b; Stephens et al., 2020). Contigs having ten or more genes were evaluated; *O. marina*, 8,456 genes (518 contigs); *P. falciparum*, 6,505 genes (16 contigs); *S. microadriaticum*, 25,121 genes (601 contigs). Additionally, we identified retrogenes (genes resulting from cDNA integration) by searching for the 18-nucleotide spliced leader relic 5′-CCATTTTGGCTCAAG-3′ using the "seqkit" package v2.4.0 (Shen et al., 2016), allowing for three mismatches (grep -m3). The gene organization results were plotted using the function geom_bar, ggplot2 (Wickham, 2016). The retrogenes enrichment analysis was conducted using clusterProfiler v.3.17 (Wu et al., 2021b).

4.2.7 Telomeres detection using Fluorescence In Situ Hybridization (FISH)

To examine the distribution of telomeric repeats on *O. marina* chromosomes, Fluorescence In Situ Hybridization (FISH) was implemented. This process involved mitotic arrest induction by exposing the cells to colchicine, which inhibits the mitosis progression beyond metaphase by preventing microtubule polymerization. Cells then are treated with hypotonic solutions, and subsequently, the chromosomes are fixed and subjected to hybridization with a fluorescent probe.

First, 100 mL of culture was harvested and pelleted (3000 g for 15 min). Then, cells were placed in 10 mL F2 media with colchicine (1 μg/mL final concentration) and incubated at RT for 4 hours. Optimal conditions were obtained by testing colchicine concentrations ranging from 0.2 to 1 μg/mL and incubation time from 1 to 18 h.
Then, cells were centrifugated and resuspended in 10 mL of hypotonic solution (75 mM KCl, 10 mM $MgS0_4$, 0.2 mM spermine and 0.5 mM spermidine, pH 8) and incubated for 30 min at RT with vortexing every 10 min. After this, cells were centrifuged, supernatant discarded, and 1.5 mL polyamine isolation buffer was added (15 mM Tris pH 8.0, 1 mM EDTA, 0.5 mM EGTA, 80 mM KCl, 3 mM DTT, 0.25% Triton X-100, 0.2 mM spermine and 0.5 mM spermidine, pH 8). The incubation was conducted on ice for 15 min. Then, cells were vigorously vortexed for 10 min to disrupt the nuclear membrane and release the chromosomes. The density of the nuclei suspensions was verified by quick staining with DAPI (2 μg/mL), and the nuclei suspension was then stored at 4°C.

A FISH in suspension (FISH-IS) protocol was adapted from a previous study (Cuadrado et al., 2019b). 100 μL of the nuclei suspension was centrifuged (500 g for 10 min), the supernatant was removed, and the nuclei pellet was resuspended in 45% glacial acetic acid for 5 min. After centrifugation, the nuclei pellet was washed in 100 μL of 2×SSC for 5 min. Then, the nuclei pellet was resuspended in the hybridization solution: 20 μL of 2×SSC, 1% Triton X-100, 1% salmon sperm DNA (1mg/mL), 10 pmol of telomeric probe 5'-[Cy5]TTTAGGTTTAGGGTTTAGGG-3'. The hybridization was carried out in the dark at RT for 2 hours. Subsequently, 10 μL of the hybridization mixture was combined with 2 μL DAPI (2 μg/mL) and 3 μL of ProLong antifade solution. The mixture was then mounted on a slide, sealed, and allowed to settle for 6 h. Z-stack images (15-20 slices, 0.12 μm apart) were captured on Leica SP8 confocal with 100X objective (HC PL APO CS2, NA 1.40 OIL). Images were processed with ImageJ (Schneider et al., 2012).

4.3 RESULTS

4.3.1 *O. marina* genome sequencing

The sequencing process generated ~9.5 million long reads (PacBio, N50 ~15Kbp) and roughly 250 million short reads (Illumina paired-end). A substantial portion of the raw reads, approximately 30% from long and 60% from short reads were taxonomically associated with alphaproteobacterial families *Rhodospirillaceae* and *Rhodobacteraceae* (Supplementary Figure 6). Additional contigs were removed after targeting potential bacterial cohabitants from the genera *Oceanibaculum*, *Alcanivorax*, and *Flavobacteriales*.

4.3.2 *O. marina* genome assembly and ploidy

This is the first genome assembly generated at the survey level for *O. marina* using high throughput sequencing methods. Previous sequencing attempts based on 454 and Sanger sequencing only yield short gene-size contigs (C. D. Lowe, Mello, et al., 2011). The 212 Mbp draft assembly obtained here substantially improved the genome representation from the previous EST-based sequencing study (Lee et al., 2014b). The general assembly contiguity is low with N50 ~ 27 Kbp and maximum scaffold size ~279 Kbp (Table 4.1). Despite this, the general features of the assembly can be contrasted with other dinoflagellate assemblies (Table 4.1). The *O. marina* assembly showed a high GC content of ~59% only compared with ~56% of *Amoebophrya ceratti* and contrasted with the average 50% reported for symbiotic dinoflagellate assemblies (González-Pech et al., 2021a). Likely the *O. marina* assembly is more representative of the coding fraction of the genome (i.e., total GC% similar to CDSs GC%, table 4.1) where GC content is reported to be high (>55%) in most dinoflagellates (Williams et al., 2017). Even non-coding regions are also present, including TEs; most of this content remains unassembled due to its intricate complexity. Interestingly, the average sequencing depth was ~100x, with less than 50 contigs displaying <5x.

On the other hand, different estimations consistently recover similar proportions of the core of conserved orthologs, i.e., 19% (BUSCO Alveolate), 22% (BUSCO Eukaryotes) and 26% (CEGMA). Additionally, the transcriptome reads derived from a highly complete *O. marina* transcriptome assembly (~80 BUSCO completeness) were mapped to the assembly to estimate its completeness. Approximately 20% of the total reads mapped to the assembly (Figure 4.1B and C). However, it is important to note that BUSCO tends to underrepresent the genic content in the dinoflagellate genomes (González-Pech et al., 2021a; Stephens et al., 2020). A high proportion of orthologs are found duplicated (Supplementary Table 4), suggesting some prevalence of gene duplication. Only a few fragmented BUSCO orthologs were observed (Supplementary Table 4), implying low fragmentation and sequencing error rates were introduced during the sequencing. Despite the low BUSCO completeness, the draft assembly possesses enough quality to do further explorations about the gene content and genome organization.

Ploidy analysis revealed that *O. marina* possesses a haploid genome configuration. The ploidy level assessment showed a bimodal allele distribution, where the most common alleles (Figure 4.1A) had frequencies of approximately 95% (first allele) and 5% (second allele), likely attributable to sequencing errors in the case of the second allele (Augusto Corrêa dos Santos et al., 2017). Meanwhile, meiosis marker genes were identified (Supplementary Table 6) and reported previously by Lee et al. in 2014, suggesting the possibility of occasional sexual reproduction in *O. marina*. However, sexual reproduction could be rare, particularly in this hypothetical clonal culture, and thus, the presence of diploids is unlikely. Additionally, kmer-based ploidy and genome size estimation were conducted. Still, they were inconsistent primarily due to contamination, assembly fragmentation, and the large genome size expected for *O. marina* (30-50 Gbp), estimated based on fluorescence (Sano & Kato, 2009).

Table 4.1. ***O. marina*** **genome assembly statistics and gene annotation compared with available dinoflagellates assemblies.**

	O. marina **(this study)**	***Polarella glacialis*** **(Stephens et al., 2020)**	***Symbiodiniaceae***		***Amoebophrya ceratii*** **(John et al., 2019)**
			Symbiodinium microadriaticum **(Aranda et al., 2016a)**	***Cladocopium goreau*** **(Liu et al., 2018b)**	
%G+C	58.7	45.91	50.51	44.83	55.92
Total number of scaffolds	6,720	33,494	9,695	41,289	2,351
N50 length of scaffolds (bp)	27,304	170,304	573,512	98,034	83,970
Maximum scaffold length (bp)	278,933	2,170,995	3,144,590	8,337,000	536,776
Estimated genome size (Gbp)	——	3.02	1.1	1.19	0.12
Genes					
Number of genes	24,542	58,232	29,728	39,006	19,925
Gene models supported by transcriptome (%)	98	94	79.2	76.5	24.4
G+C content of CDS (%)	57.5	57.84	57.43	54.23	60.77
Exons					
Number of exons per gene	2.9	11.64	19.21	12.46	3.39
Average length (bp)	341	105.67	115.44	130.47	577.8
Total length (Mb)	24.54	71.6	65.92	63.42	39.08
Introns					
Proportion of genes with introns(%)	89	73.79	95.7	96	71.35
Average length (bp)	315	1408	387.92	593.53	337.11
Total length (Mb)	13.73	837.95	210	265.35	16.08
Intergenic regions					
Average length (bp)	1696	21,625	15,108	9538	1525

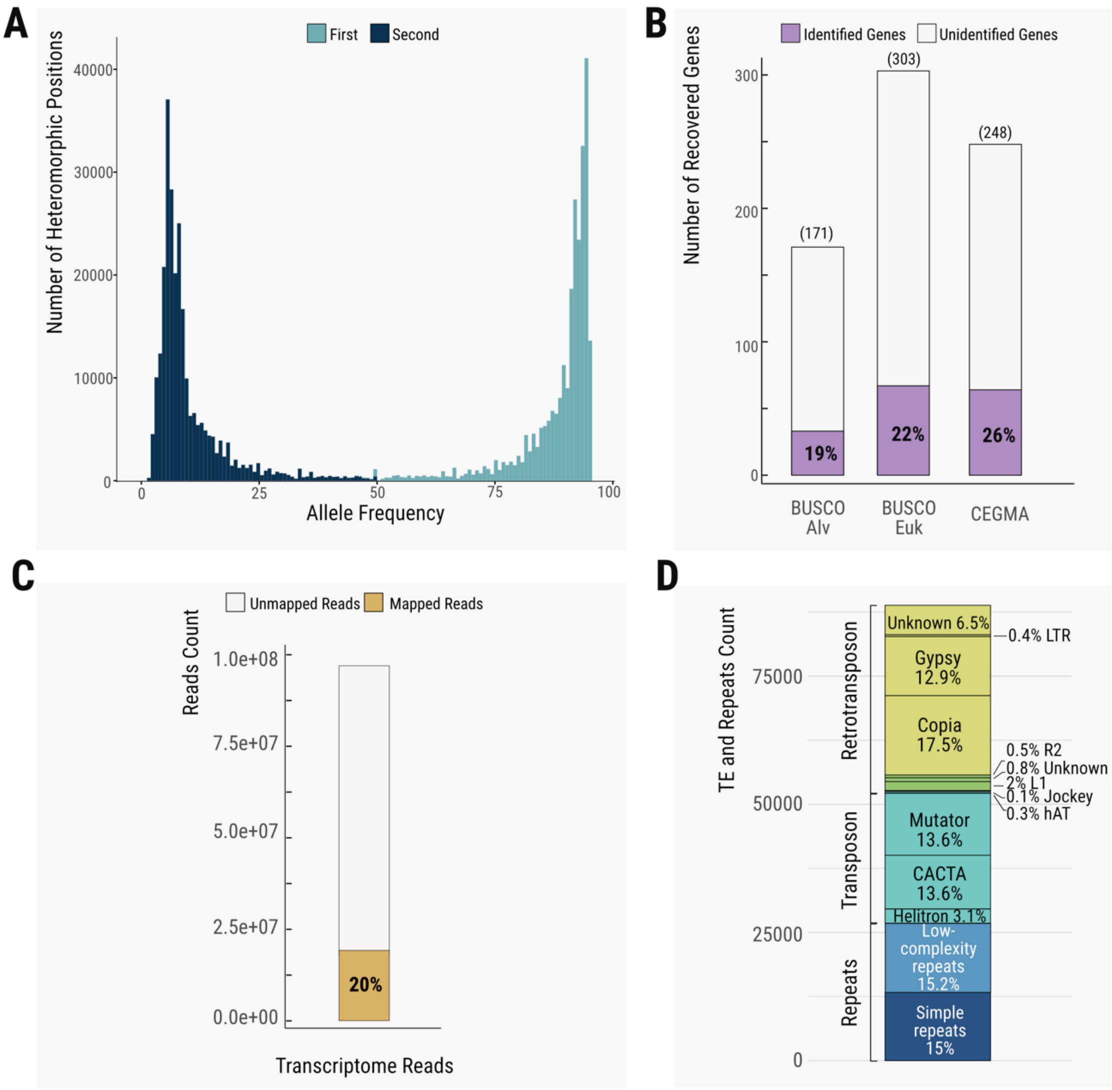


Figure 4.1. Completeness of *O. marina* draft assembly and repeat content.

(A) Histogram of two most frequent alleles suggesting haploid genome configurtion, estimated by ploidyNGS(Augusto Corrêa dos Santos et al., 2017). The dark and light blue bars represent the two most frequent alleles, as the legend displays. (B) completeness assessment based on the identification of conserved orthologues using three databases: Alveolate_odb10 (303 genes), Eukaryota_odb9 (171 genes), and CEGMA v2.5 (248 genes). Completeness is displayed as the percentage of genes identified. Detailed completeness assessment report can be found in supplementary table 6. (C) Completeness assessment based on transcriptomic reads mapping (SRR1296907 and SRR1300472). Completeness is indicated by the percentage of reads mapped to the draft assembly. (D) Repeat content of the draft assembly annotated by EDTA (Su et al., 2021). TE and repeats are categorized into retrotransposon, transposon, and repeats. Complementary repeat annotation based on Repeatmasker can be found in supplementary figure 7.

4.3.3 Repeat content and retrogenes

The draft assembly of *O. marina* is dominated by a significant proportion of repetitive elements, comprising 40% of the total assembly's length (Figure 4.1D). Within repeats, approximately 30% are attributed to low-complexity and simple repeats, while LTR retrotransposons comprise about 34% of these elements (Figure 4.1D). Such abundance of repeats can be dimensioned in the distribution of telomeric satellite repeats through the *O. marina* karyotype, including notably large intrachromosomal interstitial regions (Figure 4.2). Interstitial regions have also been observed in the karyotype of the dinoflagellate *Karenia brevis*, although to a small extent (Cuadrado et al., 2019b). This pattern implies the possibility of chromosome rearrangements within *O. marina* karyotype.

Comparatively, *O. marina* appears to fall within the middle range regarding repeat content among dinoflagellates. For instance, early genome surveys estimated ~5-6% of repeat content in *Prorocentrum minimum* and *Heterocapsa triquetra* (McEwan et al., 2008a; Ponmani et al., 2016), while higher proportions were estimated in the symbiotics *Symbiodinium microadriaticum* (28%) and *Fugacium kawagutii* (16%) (Aranda et al., 2016a). A higher proportion of repeats can be found in the free-living *P. glacialis* (~68%) and *Alexandrium ostenfeldi* (~58) (Jaeckisch et al., 2011; Stephens et al., 2020). Therefore, a more significant genomic fraction composed of repeats appears to be the signature of free-living dinoflagellates, and an increase of this portion would be expected for *O. marina.*

Most of the retrotransposons identified were LTR-retrotransposons. Retrotransposons account for ~34% of the repeat content, from which 30% are LTR-retrotransposons (Figure 4.1D). In contrast, the presence of LTR elements is significantly reduced in *P. glacialis* (10-11%) (Stephens et al., 2020) and in Symbiodiniaceae (< 5% of total assemblies) (González-Pech et al., 2021a). Non-LTR retrotransposons (like LINEs) are almost absent in *O. marina*, making up less than 3% of the total repeat content (Figure 4.1D); meanwhile, LINEs are found relatively more abundant in some members of Symbiodiniaceae (e.g., ~23% of the total assembly in *Symbiodinium pilosum*) (González-Pech et al., 2021a). The striking abundance of LTR elements is characteristic of *O. marina* and might contribute to its genome expansion, and the possibility of new LTR families may be expected.

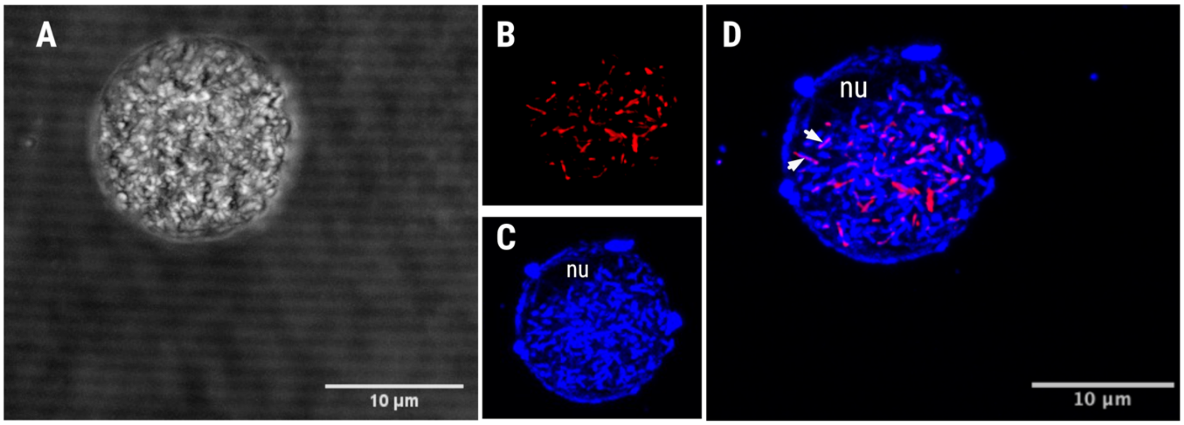


Figure 4.2. Images for FISH in suspension localizing telomeric repeats on *O. marina* nucleus.

O. marina nuclei resulting from confocal images stacks (n=20). (A) Brightfield image of an isolated nucleus showing a rounded shape. (B) Telomeric probe (red) hybridization of the labelled Cy5 oligonucleotide (TTTAGGTTTAGGGTTTAGGG) localizing telomeric repeats. (C) Chromosomes DAPI-stained DNA (blue). (D) DAPI-probe merge image localizing telomeric repeats on telomeric ends. Arrows indicate large interstitial signals covering a significant fraction of chromosomes. Inferred localization of nucleolus is labelled as "nu."

Retrogenes are identified by a DinoRL motif (spliced-leader relic detected on dinoflagellates transcriptomes, see Chapter 2) originating from SL-RNA trans-splicing and reverse transcription of mRNAs. Here, 200 retrogenes have been detected on the *O. marina* genome assembly, with most of them possessing a single DinoRL at their 5' end. Likely RNA degradation and epigenetic modification impeded the identification of these retrogene transcripts in the survey conducted on transcriptomes in Chapter 2 (only one retrogene identified in *O. marina*). Conversely, *S. kawagutti* and *B. minutum* genomes contain thousands of retrogenes directly linked to processes like symbiosis (Song et al., 2017b). *O. marina* retrogenes are enriched in DNA mobilization (e.g., RNA/DNA hybrid ribonuclease activity, DNA integration, transposition mediated) and post-translational modification and signalling (e.g., protein phosphorylation, transferase activity, intracellular signal transduction and signalling) (Supplementary Figure 8). These enriched GO categories are also prominent in the genome of *S. tridacnidorum* and *S. natans* (González-Pech et al., 2021a), indicating that mobile elements and dinoflagellate housekeeping genes likely contribute significantly to retrogene diversity. The scarcity of retrogenes is partially explained by the low assembly completeness. Still, their contribution to total genes is only ~0.8% compared with 22-23% observed in *S. kawagutti* and *B. minutum* (Song et al., 2017b). Differences in retrogene survival (i.e., transcription accessibility), genome size, trans-splicing rate, and the frequency of transposable activity may explain the variation in detectable retrogene abundance across dinoflagellates.

4.3.4 Gene prediction

A total of 24,542 gene models were predicted for the *O. marina* assembly, with 98% of these models supported by transcriptome data (Table 4.1). Interestingly, *O. marina* gene content is almost half of the total genes predicted for the diploid free-living *P. glacialis* (Table 4.1). It is possible that *O. marina* may have at least three times more genes based on the ~ 22% assembly completeness achieved (Figure 4.1B). However, assembly bias toward gene-rich sequences at the expense of complex repetitive sequences may explain the gene richness of the assembly. A high number of genes are expected for dinoflagellate genomes (60-90,000) based on regression models of these variables (Hou & Lin, 2009). Nevertheless, this large number of genes results in a notorious redundancy.

Regarding gene structure, 11% of the genes lack introns, and most of the genes have few exons (Table 1). This proportion of intronless genes may correspond to retrogenes whose DinoRLs have degenerated beyond recognition. Retrogenes are characterized by the lack of introns resulting from the retroposition of mature mRNA. Limited alternative splicing in *O. marina* might result in genes with a reduced number of exons compared with the rest of the dinoflagellates (Table 4.1). However, it cannot be ruled out that partial gene structures might be artifacts of the fragmented assembly.

4.3.5 Gene arrangement and annotation

Genes are in close proximity to each other, separated by intergenic regions averaging 1,692 bp (Figure 4.2A). A similar pattern of gene organization was observed in *P. glacialis*, where genes are often organized in clusters separated by intergenic regions of less than 5 Kbp. Remarkably, almost all of the genes are oriented unidirectionally (Stephens et al., 2020). This specific organization was assessed by examining the number of strand-orientation changes in the ten-genes window. The directionality in gene orientation of *O. marina* was tested and compared with dinoflagellate *S. microadriaticum* and the apicomplexan *Plasmodium falciparum*. *P. falciparum* was used to represent the random gene orientation of eukaryotes. Two strand-orientation changes every ten genes were more frequently observed for dinoflagellates contrasted with four and five detected for *P. falciparum* (Figure 4.2B). This trend confirms that the unidirectional gene cluster organization extends beyond the symbiotic dinoflagellates to the basal-branching *O. marina*. This gene organization has been suggested to represent a mechanism that optimizes gene proximity for simultaneous transcription (Stephens et al., 2020). However, this hypothesis has not been tested yet.

The analysis of duplicated genes in *O. marina* revealed that approximately 15,819 genes (64%) are categorized as dispersed duplicates, followed by 6,699 genes (27%) identified as singletons (Figure 4.2D). A similar proportion of duplicates and singletons were determined in *Cladocopium goreaui* (Y. Chen et al., 2020, 2022). Although the prevalence of dispersed duplicates in *O. marina* may be inflated due to the lack of continuity of the assembly, processes such as TE activity and chromosome rearrangements (e.g., inversion and translocation) may contribute to the dissemination of the gene copies. In fact, several episodes of large-scale TE activity have been inferred for *Symbiodinium* (Song et al., 2017b), and ongoing activity has been reported for *O. marina* (Lee et al., 2014b). On the other hand, gene annotation confirmed the abundance of repeat elements in the assembly, including protein domains encoded by retrotransposons such as reverse transcriptase (RT), RNase H, aspartic protease domain, and integrase among the most frequently annotated protein domains (Figure 4.2E). Interestingly, antifreeze and bacteriorhodopsin-like domains are also frequently found, and many of them are organized in tandem, particularly in the case of bacteriorhodopsin (Figures 4.2E and 4.2F). Bacteriorhodopsin is crucial in proton pumping and is posited to facilitate ATP synthesis by creating a proton gradient. It is present in photosynthetic and heterotrophic dinoflagellates (Slamovits et al., 2011) and is postulated to have a prominent role under poor nutritional conditions (Guo et al., 2014).

Several long-standing questions remain about chromatin organization, plastid ancestry, presence of meiosis and sexual reproduction in dinoflagellates. A preliminary search was conducted for genes that may shed light on these questions. Multiple copies of genes hypothetically associated with chromatin organization, including histones, DVNPs, and highly expanded (Regulator Chromosome Condensation 1) RCC1, were identified (Table Supplementary 6). It is worth noting that RCC1 is highly expanded in the *Symbiodinium* genomes and is likely involved in gene expression regulation (Shoguchi et al., 2013b). Additionally, eight plastid nuclear genes encoding plastid-localized proteins have been reported for *O. marina* (Slamovits & Keeling, 2008a); however, only two of them were identified in this study, i.e., ketol-acid reductoisomerase and glutamine synthetase. To reveal the potential occurrence of meiosis, 11-meiosis gene markers (Liu et al., 2018) were searched in the *O. marina* predicted proteins and only three of them were identified (Table supplementary 6). Multiple copies (13) of Mei2-like (master regulator of meiosis) were identified, supporting the eventual occurrence of meiosis, and additional meiosis-gene markers would be expected.

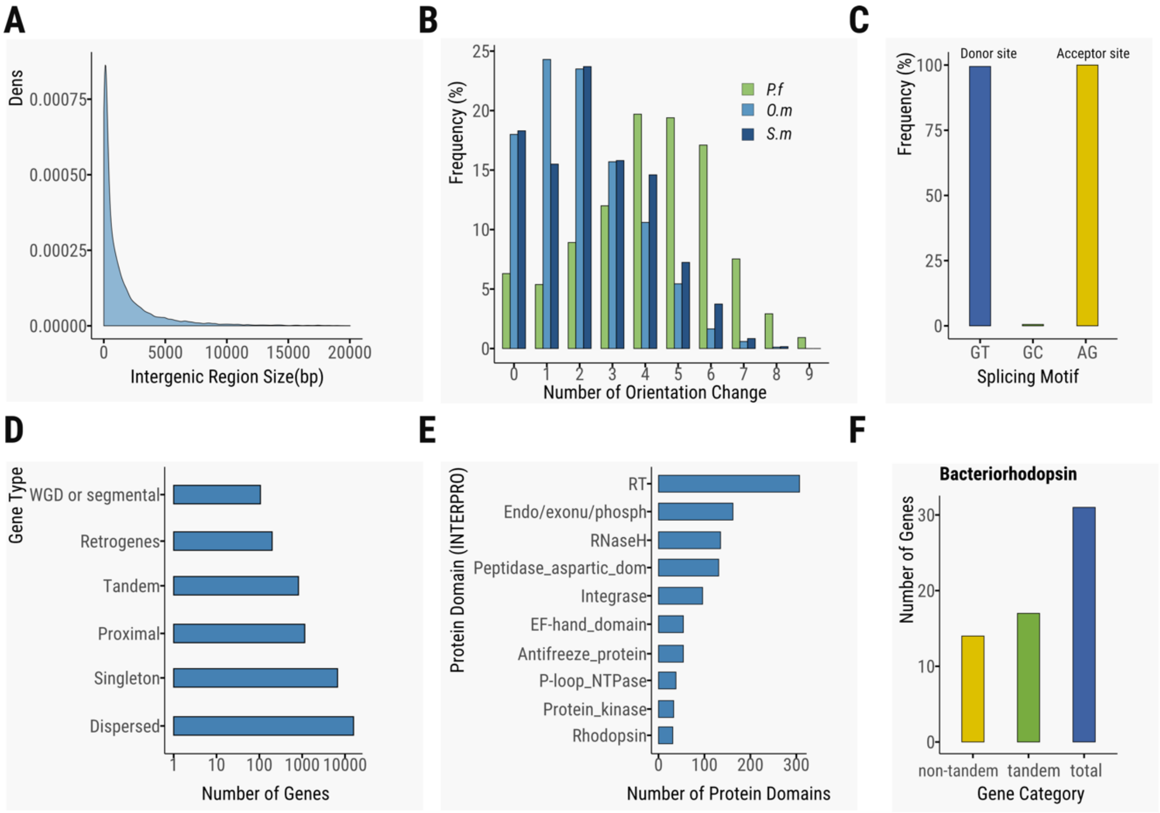


Figure 4.2. *O. marina* genome organization, gene architecture and arrangement. (A) Distribution of the size of intergenic regions (<20,000 bp), derived from the gene prediction for the *O. marina* assembly. (B) Comparison of gene orientation changes across alveolates. The frequency of gene orientation change was estimated in a ten-gene window using the predicted genes for *O. marina*, and the gff3 files for *Symbiodinium microadriaticum* CCMP2467 (GenBank accession: LSRX00000000.1), and *Plasmodium falciparum* 3D7 (GenBank accession: GCF_000002765.5). (C) Splicing sites motif for *O. marina* genes. (D) Classification of duplicated genes according to their localization and origin. Segmental (genes in syntenic blocks), tandem (continuous repeat), proximal (close chromosomal region but not adjacent), dispersed (other than segmental, tandem and proximal) (Y. Wang et al., 2012). (E) Top 10 most frequent protein domains annotated sorted by abundance: RT, IPR000477; Endo/exonu/phosph, IPR036691; RNaseH, IPR012337; Peptidase_aspartic_dom, IPR021109; Integrase, IPR001584; Antifreeze protein, IPR000104; EF-hand domain, IPR002048; P-loop_NTPase, IPR027417; Protein kinase, IPR000719; Rhodopsin, IPR001425. The top 30 most frequent protein domains are in the supplementary table 5. (F) Number of genes encoding Bacteriorhodopsin, categorized as tandemly organized and non-tandem.

4.4 DISCUSSION

4.4.1 Bacterial content and assembly

This is the first instance of a genome assembly generated for free-living dinoflagellates outside the Suessiales order, primarily consisting of symbiotic species with smaller genomes. Despite several efforts to remove the bacterial content from the cultures and assembly, a significant fraction of the genomic sequences recovered was attributed to bacteria. Within this, a substantial portion included uncharacterized bacterial groups closely associated with *O. marina*, as previously documented by Lee et al., 2014. These physical associations have also been observed between *Alexandrium spp*. and *Gyrodinium instriatum* and planktonic bacteria (Biegala et al., 2002). Nevertheless, the precise nature and extent of this potential symbiotic relationship remain enigmatic. After two cleaning steps, most of the prokaryotic DNA sequences were removed, and the presence of bacterial gene hits still needs to be evaluated as hypothetical LGT. Overall, our assembly represents a moderated well-assembled gene space fraction with a high-depth coverage, similar to the initial drafts obtained for polyploid repeat-reach genomes in plants (Ou et al., 2020). Future strategies to improve assembly contiguity and completeness include additional rounds of long-read sequencing (e.g., PacBio Hifi), read correction and assembly polishing.

4.4.2 Repeat content and LTR-retrotransposon bursts

Repetitive elements have a crucial role in shaping the genome of dinoflagellates. A large fraction of repeats could be considered an intrinsic feature of the free-living dinoflagellates (Stephens et al., 2020), and *O. marina* is not the exception. However, more genomic data from free-living dinoflagellates is needed to test this trend. Micro-satellite repeats have been significantly expanded in dinoflagellate chromosomes (e.g., AG-rich chromosome in *Karenia mikimotoi*), although telomeric interstitial sequences were rarely observed (Cuadrado et al., 2019a). Expansion of telomers into interstitial chromosomal space, also known as interstitial telomeric sequence (ITS), may result from chromosome rearrangements (inversion or fusion), uneven crossing-over, and insertion of telomeric repeats during DNA repair (Bolzán, 2012). Therefore, this pattern and the large fraction of dispersed gene copies identified suggest that genome rearrangement has occurred in *O.*

marina, and ongoing TE activity (discussed below) may be involved in scattering the gene content throughout the genome.

Similar to the finding for the free-living *P. glacialis* (Stephens et al., 2020), *O. marina* showed a notable proliferation of LTR-retrotransposon compared to non-LTR (e.g., LINE). This surge in LTR elements may result from exposure to environmental stressors in free-living conditions, as seen in diatoms in response to nitrate deprivation (Maumus et al., 2009). Likewise, heat stress has been documented to induce transposition activity in dinoflagellates (J. E. Chen et al., 2018). LTR retrotransposon is a type of Class I retrotransposon that uses a replication mechanism like retrovirus and an integration mechanism of copy-and-paste similar to transposons (Wells & Feschotte, 2020). Retrotransposons usually remain inactive or transcriptionally silenced, although instances of high and moderate expression have been detected for LINE and Ty1/copia retrotransposon in *Symbiodinium* and *O. marina*, respectively (de Mendoza et al., 2018; Lee et al., 2014b). In fact, reverse transcriptase (RT) encoded by retrotransposon was the most prevalent domain identified in the genome, supporting the ongoing transposition hypothesis in *O. marina* (Lee et al., 2014b). The transcriptional silencing of transposable elements is mediated by epigenetic mechanisms such as CG and histone methylation. Eventually, incomplete silencing of these elements due to non-canonical histones and scarcity of nucleosome organization may contribute to their prevalence in the genome. Alternatively, dinoflagellate genomes may exhibit a non-canonical methylation pattern in which retrotransposon may be essential (de Mendoza et al., 2018). Nevertheless, the observed difference in the abundance of retrotransposon subclasses remains unclear (i.e., LTR vs non-LTR) and might be related to class-specific and host-specific roles. Additional examination of the retrotransposon life cycle in dinoflagellates and the prevalence and distribution of the types of retroelement families are needed.

4.4.3 Gene Redundancy and Organization

O. marina displays a significant number of duplicated genes that are primarily found as dispersed gene copies. A similar proportion of duplicates and singletons were determined in *Cladocopium goreaui* (Y. Chen et al., 2020, 2022). The prevalence of dispersed duplicates suggests TE activity and karyotype arrangements may contribute to the scatter of the gene copies through the genome of *O. marina*. However, assembly contiguity must improve to corroborate the organization of the gene copies. Although multicopy genes are more frequently found in dinoflagellates compared to other eukaryotic groups (Hou & Lin, 2009), their prevalence is unclear. It is unclear whether dinoflagellate gene redundancy resulted from WGD events, segmental or individual gene duplication (including retroposition) or a combination of these mechanisms (Hou & Lin, 2009). WGD events have been shown to be the most efficient mechanism for generating redundancy and increasing plant genome size (i.e., autopolyploidy and allopolyploidy) (Panchy et al., 2016). Polyploidy has been proposed as a speciation mechanism in the dinoflagellate *Heterocapsa pygmaea* (Loeblich et al., 1981). The small detectable fraction of WGD/segmental duplication in *O. marina* (Figure 4.2D) and other dinoflagellate assemblies (González-Pech et al., 2021a) may represent remnants of ancient large-scale duplication events. Recent evidence of highly conserved syntenic blocks involving ~22% of the total genes suggests recent events of WGD in *Durusdinium trenchii* (Dougan et al., 2022). In this scenario, the pervasive activity of TE likely has contributed to erasing WGD signatures over time. Paradoxically, WGD likely induced TE activity due to silencing relaxation produced by cellular stress (Marburger et al., 2018). Several mechanisms determine genome evolution, and the prevalence of each may explain the observed differences in genome size across dinoflagellates.

In *O. marina*, genes are predominantly oriented unidirectionally throughout the genome (Figure 4.2B), which is consistent with observations in other dinoflagellates genome assemblies (Y. Chen et al., 2022; Shoguchi et al., 2013b; Stephens et al., 2020). This is the first evidence for this pattern outside of the Suessiales order, suggesting an early origin of this feature during dinoflagellate evolution. In this pattern, genes tend to be oriented similarly to neighbouring genes, which is highly infrequent in eukaryotes but is more commonly observed in prokaryotes. The only exceptions include kinetoplastids such as *Trypanosoma brucei*, which exhibit this type of gene arrangement (Daniels et al., 2010; Kolev et al., 2010). In trypanosomatids, gene blocks are transcribed as polycistronic mRNA, then excised by SL-RNA trans-splicing, resulting in monocistronic mRNA

(Clayton, 2019). This mechanism has been proposed in dinoflagellates to ensure efficient transcription (Stephens et al., 2020). Nevertheless, the presence of gene blocks transcribed as polycistronic transcripts in dinoflagellates remains uncertain (Wisecaver & Hackett, 2011). It is worth noting that unidirectional gene blocks are typically associated with genes sharing similar functions (Marinov et al., 2023; Nand et al., 2021; Stephens et al., 2020). Alternating orientation between gene blocks is linked to the chromatin folding structure of the dinoflagellate (Nand et al., 2021). Further insights into the functional roles of gene blocks in *O. marina* could provide a better understanding of this functional association.

4.4.4 Chromatin and chromosome organization and meiosis

Historically, dinoflagellate chromatin organization has been described as lacking the typical nucleosome organization composed by the octamer of histones (Gornik et al., 2019; Wisecaver & Hackett, 2011). Unusual absence of nucleosomal organization patterns inferred through the nuclease digestion (Gornik et al., 2012). Unusual reduced abundance at proteomics and transcriptomic levels (Riaz et al., 2018; Roy & Morse, 2012). However, more recent data have revealed the presence of histones and various viral and bacterial alternatives, such as DVNPs, HLP (histone-like protein), and RCC (regulator of chromosome condensation) proteins with a strong affinity for DNA (Gornik et al., 2019). Twenty DVNP variants have been previously identified in *O. marina* (Lee et al., 2014b) and likely represent the Np23, the most abundant basic protein detected biochemically in previous nuclear extracts (Kato et al., 1997). Prior to this work, the presence of histones has remained elusive in *O. marina*, with no clear homologues for histones or histone-related proteins (Janouškovec et al., 2017; Lee et al., 2014b). Homologues for core and linker histones have now been identified in transcriptomes, suggesting nucleosome organization is likely in *O. marina*. Additionally, the gene encoding for RCC1 (regulator of chromosome condensation 1) is present in many copies, which also has been found highly expanded in the genome of *Breviolum minutum* (Shoguchi et al., 2013b), and it has been proposed as an essential regulator of the gene expression (Hadjebi et al., 2008). RCC1 from eukaryotic and prokaryotic origin has been identified in the dinoflagellates (Shoguchi et al., 2013b) and phylogenetic analysis would help to understand the origin of these genes in *O. marina* and dinoflagellates at large.

While evidence of sexual reproduction in *O. marina* has been elusive, there have been limited observations of potential gametes (Montagnes, Lowe, Martin, et al., 2011) and additional confirmation of meiosis-related genes (Lee et al., 2014b). In addition to the previously identified SPO11 (meiotic recombination protein), MEIG1 (Meiosis-expressed gene 1 protein) and multiple copies of the Mei2-like genes were identified. Mei2-like genes have been identified in *Fugacium kawagutii* assembly (T. Li et al., 2020). They are also differentially expressed and potentially involved in the sexual reproduction and encystment of *Scripssiella trochoidea* (Deng et al., 2017). Therefore, the presence of these key meiosis gene regulators suggests the occurrence of sexual reproduction in *O. marina*. However, the low occurrence of meiosis or restricted to the small cell population may lead to the loss of meiosis-related genes in the genome. Additional understanding of life cycle and sexual reproduction remains to be explored.

4.4.5 Conclusion and future directions

Besides the large genome size estimated for free-living dinoflagellates, this study demonstrates that the PacBio long-read technology is a suitable approach for sequencing dinoflagellates. The initial draft genome assembly for *O. marina* indicates that transposable elements have an essential role in shaping its genome and gene content and redundancy. The strong tendency for groups of adjacent genes to be arranged unidirectionally can be further extended to basal free-living dinoflagellate lineage such as *O. marina*.

Additional rounds of genome sequencing using high-resolution long-read sequencing methods will help to improve the genome representation and assembly contiguity. Additional tailored gene annotations, including dinoflagellate proteomes, will improve the gene annotation.

CHAPTER 5 ENDOGENOUS VIRAL ELEMENTS IN THE GENOME OF *O. MARINA*

5.1 INTRODUCTION

Endogenous viral elements (EVEs) include all types of viruses integrated into the genomes of eukaryotes (Feschotte & Gilbert, 2012). Retroviruses constitute most EVEs, and their replicative activity has profound implications for structural variation. Non-retroviral EVEs (e.g., dsDNA viruses) appear more abundant than previously thought. Accruing evidence suggest that these elements play a substantial role in shaping the genome in many groups of eukaryotes, including diverse protists (Bellas et al., 2023) and cnidarians (Filée, 2014). The endogenization of large dsDNA viruses NCLDVs (i.e., Nucleocytoplasmic viruses) is deeply rooted in the tree of eukaryotes and results in a substantial reshuffling of the host genome. For instance, the endogenization of NCLDVs (genome size ~2 Mbp) in green algae resulted in assimilated regions rich in duplication and spliceosomal introns (Moniruzzaman et al., 2020). How endogenization occurs is not entirely understood; nonetheless, errors during DNA repair or non-target activity of reverse transcriptase encoded by retrotransposon are proposed to be involved (Geuking et al., 2009; Holmes, 2011). On the other hand, there is growing evidence supporting the pervasive presence of small-size, dsDNA endogenous viruses (i.e., 15-40 Kbp) in the genome of single-cell eukaryotes (Bellas et al., 2023; C. M. Bellas & Sommaruga, 2021; Yutin et al., 2013; Yutin, Shevchenko, et al., 2015). Small-size endogenous viruses include virus parasites of other viruses, endogenized virus-like transposable elements widespread in the genome of eukaryotes, and a variety of dual-life style viruses identified from metagenomic datasets. It has been proposed that this diversity might represent the tip of the iceberg of a rather extensive diversity of endogenous viral elements (Bellas et al., 2023).

Small EVEs, including Mavericks/Polinton (MP), Polinton-like viruses (PLVs), and virophages, share a core of genes involved in capsid morphogenesis (Krupovic & Koonin, 2015). This core comprises a packing-ATPase and double and single jelly-roll major and minor capsid proteins (MCP and mCP). The presence of these genes is essential to predict viral status (i.e., active, dormant, or mobile element lifestyle). For instance, MP and Tlr1 elements in *Tetrahymena thermophila* were initially classified as large self-replicating transposons until the identification of capsid proteins led to their reclassification as bona fide viruses (Krupovic et al., 2014). A similar case was the MP element in *Trichomonas vaginalis*, which comprised one-third of the genome (Pritham et al., 2007) and was initially considered a large transposon until MCP was identified (Bellas et al., 2023). In addition to the morphogenesis genes, a set of four other genes are commonly found in MP: protein-primed DNA polymerase B (pPolB), integrase derived from retroviruses (RVE-INT), helicase (SF1H and SF3H) and maturation protease (PRO) (Kapitonov & Jurka, 2006; Pritham et al., 2007). However, MP genomic structure and gene content are not fixed and differ substantially across organisms (Yutin et al., 2013). Several genes occasionally found in MP (e.g., helicase, Bro-N) are shared with distantly related viruses (i.e., Megavirales) (Krupovic & Koonin, 2015). In this sense, MP has been proposed to represent an ancestral hub of most dsDNA viruses (Krupovic & Koonin, 2015). Two hypotheses—the "nuclear-scape" and "virophage-first"—aim to explain the origin of Bamfordvirae (MP, PLV, virophage, and NCLDVs). The “Nuclear scape hypothesis” postulates that an endogenous MP-like ancestor escaped from the nucleus of early eukaryotes, giving rise to adenoviruses and NCLDVs (Koonin & Krupovic, 2017; Krupovic & Koonin, 2015). The virophage-first hypothesis posits that NCLDVs and an ancestral virophage co-evolved and MP originated from an endogenized virophage (Campbell et al., 2017; Fischer & Suttle, 2011). However, there is still limited data supporting the tentative hypothesis.

Polinton-like viruses (PLVs) are distantly related to MP and usually exhibit reduced gene sets, lacking integrase and pPolB. Their abundance and diversity have recently been recognized through metagenomic studies of aquatic ecosystems, with most hosts remaining unidentified (C. M. Bellas & Sommaruga, 2021). Moreover, they share little sequence similarity, suggesting that they represent a fraction of a broader and diverse group, with potentially many PLVs to be discovered (C. M. Bellas & Sommaruga, 2021; Chase et al., 2022; Yutin, Shevchenko, et al., 2015). It has been proposed that PLVs may adopt a dual lifestyle as free viral particles as well as endogenized in the host genome, coupled to the cellular expression (C. M. Bellas & Sommaruga, 2021; Chase et al., 2022). On the other hand, virophages (family *Lavidaviridae*) rely on giant viruses (Mimiviridae) for their replication and are usually found coinfecting protists (Blanc et al., 2015; Hackl et al., 2021; Roitman et al., 2023). Virophages possess low GC content (27-39%) and a genome size of around 15-30 kbp (Fischer, 2021). It is posited that endogenized virophages in the unicellular eukaryote genome may act as a defence system against giant virus infection (Fischer & Hackl, 2016). Virophages differ from MP and PLVs in the type of helicase (SF3H), but the distinction between these elements can be challenging. For instance, recent evidence demonstrated that PLVs could adopt a virophage-like lifestyle in the algae *Phaeocystis globosa* (Roitman et al., 2023). Genomic screening on *P. globosa* revealed that PLV (PLVGezel-14T) was found integrated, suggesting a dual lifestyle (i.e., viral particle and provirus) like virophages (Fischer & Hackl, 2016).

A large-scale survey of small endogenous viral elements revealed a hidden diversity of these elements in the genome of protists (Bellas et al., 2023). This unprecedented finding was facilitated by assemblies based on long-read sequencing, leading to the identification of thousands of intact endogenous viruses. The genomes of dinoflagellates are highly enriched viral elements, suggesting that the giant genomes of dinoflagellates may act as a reservoir of these elements. This may be particularly interesting in those species capable of forming large blooms (i.e., red tides), where viruses are the significant drivers of bloom decline (i.e., viral shunt) (Kuhlisch et al., 2021). However, the diversity of dinoflagellates is poorly represented in terms of genome sequencing and is mostly limited to symbiotic species (González-Pech et al., 2021a). Despite the significant impact of the acquisition of viral genes in dinoflagellates such as DVNPs (dinoflagellate viral nuclear protein) and its implication for genome organization (Janouškovec et al., 2017), there is limited evidence for viral presence in dinoflagellate genomes (Benites et al., 2022; Correa et al., 2013). In this chapter, we report the genomes of endogenous viral elements in the dinoflagellates *Oxyrrhis marina* and the previously undescribed isolates TGD ("T" green dinoflagellate) and MGD ("M" green dinoflagellate) (Sarai et al., 2020), used to understand the prevalence of this type of elements in other free-living dinoflagellates. We analyze the viral signatures as well as the gene content of these elements. Characterizing these viral elements provides additional evidence to understand their impact on the genome of dinoflagellates.

5.2 METHODS

The processed draft assemblies of *O. marina,* TGD and MGD were interrogated for the presence of endogenous viral elements. Detailed information about the sequencing, assembly, and annotation procedures for *O. marina* can be found in earlier chapters. Genome assemblies for TGD and MGD dinoflagellates were generously facilitated by Dr. Yuji Inagaki from the University of Tsukuba, Tsukuba, Ibaraki, Japan. TGD and MGD are two strains of dinoflagellates previously undescribed and likely belong to the order Peridinales. Both were sequenced under Illumina PE, resulting in fragmented draft assemblies with low N50. It is worth noting that these assemblers were generated with the purpose of studying their plastid genomes.

5.2.1 Identification of Viral contigs and Annotation

We searched viral contigs on the draft assembly of *O. marina*. The genome assemblies of two new undescribed dinoflagellate isolates (MGD and TGD) (Sarai et al., 2020) were also interrogated. Viral contigs were primarily identified using Vibrant v.1.2.1 (Kieft et al., 2020), using the default configuration. In brief, Vibrant selected contigs with a minimum size of 10 Kbp and containing more than four open reading frames (ORFs), predicted through Prodigal v.2.6.3 (Hyatt et al., 2010). Putative viral contigs were annotated via HMMs profiles HHsearch v.3.1 (Eddy, 2011) using three databases: virus orthologous group (VOGDB (release 94)) (https://vogdb.org/), PFAM v.32 (2019) (Finn et al., 2014), KEGG (March-2023 release) (Kanehisa & Goto, 2000). Contig regions that gathered the above criteria were extracted as proviral elements and further interrogated for viral-markers genes such as MCP, PolB, ATPase and integrase (e-value ≤ E-5 and bit score 30). Further identification and curation were conducted through HHpred (Zimmermann et al., 2018) and BLASTp using predicted amino acid sequences as queries (e-value ≤ E-5). The boundaries of viral elements were determined based on GC content and the presence of TIR (terminal inverted repeats). TIRs were localized by self-BLAST analysis of the last 1000 bp at the end of regions with significant drops in GC content. TIRs were used to define the gene repertory of the viral elements and assess the general completeness of the viral elements. Neighbouring retrotransposons were annotated through HHsearch (e-value ≤ E-5) using gypsy database v.2.0 (Llorens et al., 2011). Viral genome representation was generated using the package gggenes (https://github.com/wilkox/gggenes) implemented in ggplot2 (Wickham, 2016).

5.2.2 GC Content and Codon Usage

We manually searched for detectable drops in GC content by inspecting GC-skew implemented in Geneious v.R11. Regions that deviated from *O. marina* assembly GC content (GC:58%) and associated with viral genes were subtracted and compared with the rest of the host contig. The viral and host regions' GC content was obtained using SAMtools v.1.18 (H. Li et al., 2009) and estimated using the sliding window approach (window size 300 bp). Statistical comparisons were calculated using the Wilcoxon test. On the other hand, codon usage bias analysis was estimated by comparing synonymous codon usage orderliness (SCUO) statistics between viral and host ORFs. The analysis was conducted using the coRdon (https://github.com/BioinfoHR/coRdon) implemented in R. Statistical comparisons were estimated using the Wilcoxon test. The results were graphed using ggplot and the function ggviolin implemented in R.

5.2.3 Phylogenetic analysis and placement

Sequences generated by Yutin et al. 2013, Yutin et al. 2015, Blanc et al. 2015 and Bellas and Sommauraga 2021 were used for the phylogenetic placement of the viral elements. This group of studies was used as a reference to keep consistency with the viral clusters robustly described. Additionally, MCP genes recently reported for dinoflagellates in Bellas et al. 2013 were included in the analysis. The alignment of the sequences was conducted with MAFFT v.7.471 (Katoh & Standley, 2013) and trimmed with trimAI v.1.4 (Capella-Gutiérrez et al., 2009). Maximum likelihood (ML) reconstruction was carried out using IQ-TREE (Nguyen et al., 2015) with 1000 ultrafast bootstrap (ufb) replicates. Best substitution models were estimated by ModelFinder (Kalyaanamoorthy et al., 2017). Phylogenetic trees were edited using Figtree V1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/)

5.2.4 RNA-seq mapping

Previously generated RNA-seq sequencing was used to estimate the expression of the viral elements. The raw reads (SRA) for *O. marina* (SRR1296907 and SRR1300472) were downloaded from NCBI under BioProject PRJNA231566 (Keeling et al., 2014). Reads were trimmed by Trimmomatic v.0.39 (Bolger et al., 2014) with a conservative setting (Johnson et al., 2019). Reads mapping and quantification were conducted by bowtie2 v.2.4.5 (Langmead & Salzberg, 2012) and rsem software v. 1.3.0 (B. Li & Dewey, 2011), respectively. Coverage per nucleotide was obtained using SAMtools (H. Li et al., 2009). Coverage plots were obtained with the ggplot2 function geom_point.

5.3 RESULTS

5.3.1 Identification of endogenous viral elements in dinoflagellates

The de *novo* genome assemblies of the dinoflagellates *O. marina*, TGD and MGD were interrogated for the presence of endogenous virus. The initial analysis of the *O. marina* assembly indicated that repeat content encompasses approximately 40% (detailed in Chapter 4), including elements such as MP. Further exploration was conducted targeting endogenous virus gene markers such as ATPase, pPolB, MCP, and helicase (see detail in the methods section). Initial MCP candidates were inspected for double jelly-roll secondary structure and used to interrogate the draft assemblies further. In total, 28 endogenized viral elements were identified in *O. marina* and five in both TGD and MGD assemblies (Figure 5.2).

5.3.2 Integrated viral elements in dinoflagellate genomes

We comprehensively analyzed the genomic regions associated with viral genes to determine whether these genes were part of viral elements or randomly scattered throughout the genome. In all three dinoflagellates, viral genes were found in conjunction, as expected for endogenous viral elements (C. M. Bellas & Sommaruga, 2021; Yutin et al., 2013; Yutin, Shevchenko, et al., 2015). In *O. marina*, the viral elements were in conspicuous low-GC content regions, and the contigs were long enough (2-92 kbp, average ~40 kbp) to retrieve full-length viral elements. The size of these endogenized regions ranged from 5.1 to 27 kbp in length, averaging approximately 16 kbp, with a distinct GC content compared to the host genome (Figure 5.1A). In support of this observation, we found that the GC content of the viral region was approximately ~25%, which significantly differed from ~58% of the host (Figure 5.1B). Moreover, the codon usage (SCUO) of viral ORFs significantly diverged compared to the host (Figure 5.1C). These analyses suggest the endogenization of viral elements into the *O. marina* genome. A different endogenization pattern was observed for MGD and TGD viral elements. In this case, the elements were fragmented (4-8 kbp) and with a GC content similar to the host genome (~44%), suggesting amelioration of the viral elements. Additionally, TIRs (terminal inverted repeats) were only found in *O. marina* (in four elements) with lengths ranging from 150 to 350 bp, suggesting the detection of the complete entity. The elements identified in MGD and TGD assemblies are likely incomplete due to the limitation of a short-read based assembly. Conversely, the long-read based assembly revealed contiguous viral elements in *O. marina*, displaying distinct signatures of integration into the genome.

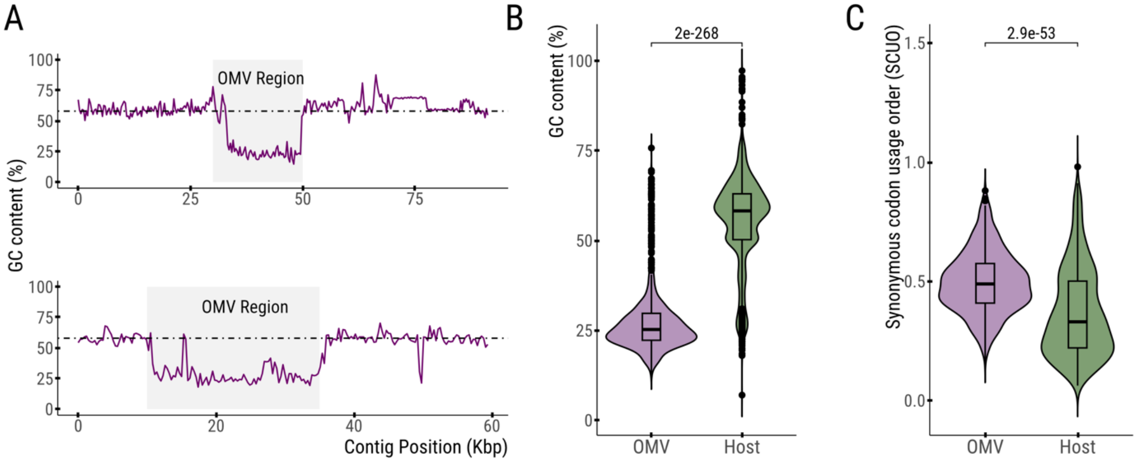


Figure 5.1. Comparison of GC content and codon usage bias between integrated viral regions and host.

(A) Two representative examples illustrate the difference in GC content between *O. marina* viral elements and the host. The black dotted line denotes the average GC content for *O. marina* (58%). (B) Comparison of the GC content (300 bp window) between *O. marina* viral elements (892) and the host *O. marina* (1464). Additional information about GC content can be found in supplementary figure 9.(C) Codon usage bias (SCUO) comparison of *O. marina* viral elements ORFs (684) and the host (872). P-values were estimated using the Wilcoxon test.

5.3.3 The gene complement of viral elements

The viral elements exhibited a gene content that resembled that of PLV, MP, and virophages while also containing genes commonly found in large dsDNA viruses (i.e., Megavirales). We conducted a detailed analysis of the gene content within these elements to assess their similarity with PLV, MP, and virophages. Among the three dinoflagellates, 12 genes were identified in the viral elements (Figure 5.2). Several genes were found fragmented into consecutive ORFs because of frameshift (i.e., the introduction of premature stop codons and sequence degeneration) (Figure 5.2). These genes were categorized based on their hypothetical functions, such as viral replication (SF1H and PolB), viral DNA packing (MCP and A32_ATPase), and viral integration (RVE-INT, GIY-YIG and HJR) (Krupovic & Koonin, 2016; Yutin, Shevchenko, et al., 2015). The accessory category included genes with unknown function in endogenous viruses (i.e., VRR-NUC, N6_Mtase, collagen-like protein, fused BRO-N–VRR, and Y-rec). Remarkably, the packing module was incomplete, lacking mCP (minor capsid protein) in all three dinoflagellates. Probably in the cases of TGD and MGD, the mCP identification is limited due to the incompleteness of the elements. However, for *O. marina*, the absence of mCP suggests that these elements might be incapable of forming virion particles. While viral elements in MGD and TGD were presumably incomplete, their gene content differed from *O. marina* elements due to the lack of the retroviral-integrase (RVT-INT) and the distinct gene arrangement within the viral elements. For instance, the DNA packing ATPase (A32_ATPase) was typically found in tandem with GIY-YIG endonuclease or MCP in TGD and MGD, respectively (Figure 5.2). Whereas it was found arranged head-to-head with the fused protein domains BRO-N–VRR-NUC in *O. marina* viral elements, suggesting that they are two separated viral entities.

Notably, the gene content of *O. marina* viral elements differs from the currently known eukaryotic small endogenous viruses due to the absence of retroviral-integrase (RVE-INT) (present in most MP) and the maturation protease (typical in most MP and virophages), while containing pPolB (usually absent in most PLV and virophages). Despite the absence of retroviral-integrase, the endonuclease GIY-YIG and tyrosine recombinase (Y-rec) were commonly found in *O. marina* viral elements and might be involved in viral DNA integration. Strikingly, *O. marina* viral elements consistently showed the presence of uncommon genes for this type of viral element, such as Holliday junction resolvase (HJR) and a BRO-N terminal domain fused with VRR-NUC endonuclease. HJR is typically found in Megavirales such as poxvirus and iridovirus (Garcia et al., 2000) and may be involved in DNA recombination. This enzyme has not been found in MP, PLV, or virophages. Similarly, the BRO-N terminal domain has been commonly found in Megavirales and phages (Krupovic & Koonin, 2015; Mönttinen et al., 2021). On the other hand, different types of DNA methylases are frequently found in PLV (Roitman et al., 2023; Yutin, Shevchenko, et al., 2015) and collagen-like repeats are observed in virophages (Yau et al., 2011). The chimeric gene content of viral elements in *O. marina* makes it difficult to determine their origins; however, it is important to notice that the gene content of endogenous viruses is highly dynamic.

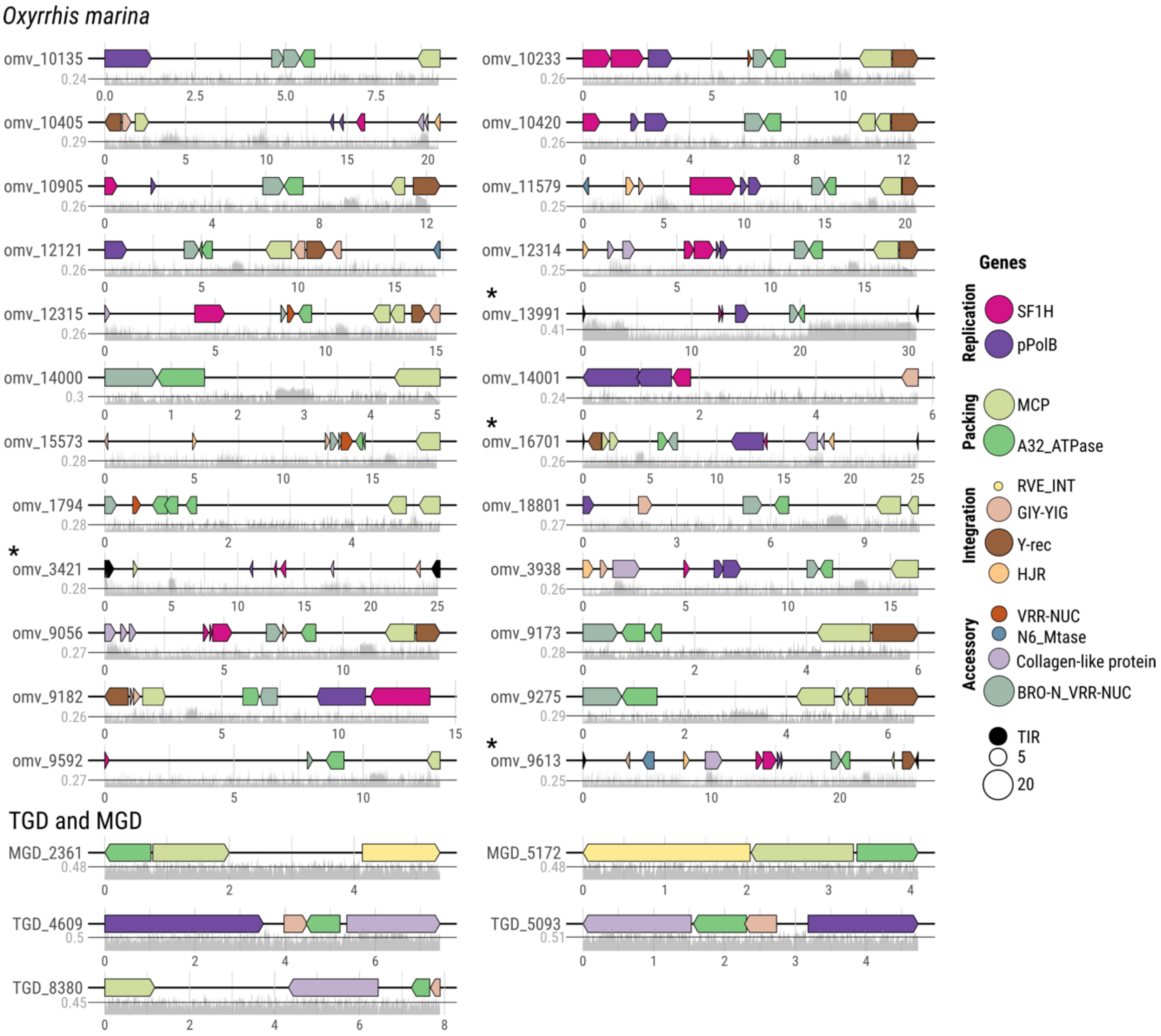


Figure 5.2. Partial and full-length viral elements in dinoflagellates *O. marina* and MGD and TGD.

Representation of the viral gene content for the three dinoflagellates: Each box depicts a particular gene coloured according to the legend. Legend abbreviation: SF1H (PF05970), superfamily 1 helicase; pPolB (PF03175), protein-primed DNA polymerase type B; HJR (PF04848), Holliday Junction Resolvase; MCP, major capsid protein; A32_ATPase (PF04665), DNA packing ATPase; RVE-INT (PF00665), retroviral integrase; GIY-YIG (PF01541), endonuclease; VRR-NUC (PF08774), VRR-NUC endonuclease domain; N6_MTase (PF02384), N6-methyltransferase; Collagen-like protein (PF01391); BRO-N_VRR-NUC fused Bro-N (PF02498) and VRR-NUC; Y-rec, tyrosine recombinase; TIR, terminal inverted repeat. The stars (*) depicted hypothetical full-length elements. In the legend, genes are grouped according to their putative function (packing, replication, integration, and accessory), and the circle size represents the gene frequency. The GC content is displayed below each element, and the black line represents the average. GC content for *O. marina* and TGD-MGD genomes were ~58% and ~44%, respectively.

5.3.4 Transcription of viral genes and their genomic neighbourhood

To investigate the integrity and transcriptional activity of the *O. marina* viral elements, their expression pattern was compared along with the host genome. Two sets of RNA-seq transcriptomes were mapped to viral elements and host contigs (Figure 5.3, Supplementary Figure 10). Most viral elements appeared to remain transcriptionally silent, with most of the transcriptional activity concentrated in gene-rich regions (Supplementary Figure 10). Additionally, sporadic instances of transcription were observed in non-coding regions. Nevertheless, a few cases of moderate expression for unknown proteins were detected within the viral regions, such as omv_3421 and omv_15573 (Figure 5.3A and B). A similar pattern has been detected in PLV *Tetraselmis striata;* genes are transcriptionally inactive except for a gene coding for an unknown protein (TVSG_00024) (Chase et al., 2022). It has been suggested that these genes might serve as sentinel genes, initiating the transcription of viral genes under unidentified triggers (Chase et al., 2022). Furthermore, Ty1/copia LTR-retrotransposons were frequently found in proximity to the integration regions (Figure 5.3C, Supplementary Figure 10). In fact, some retrotransposon genes showed high transcriptional activity, suggesting that they may be involved in the mobilization and propagation of these viral elements. The general absence of transcriptional activity and the lack of minor capsid protein (mCP) genes suggest that *O. marina* viral elements may adopt a transposable element-like lifestyle.

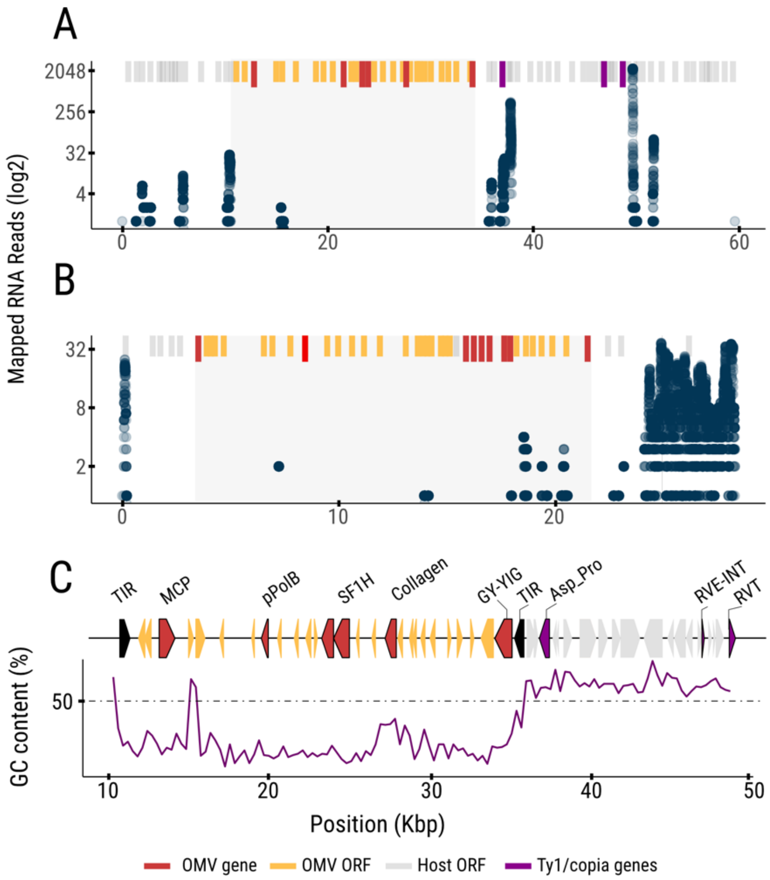


Figure 5.3. Neighbouring elements and expression pattern of viral elements and the host.

(A, B) two examples illustrating expression patterns on the endogenized regions (grey area) compared with flanking regions. The ribbons at the top represent ORFs and genes, as the legend describes. Mapped RNA reads are represented by blue dots (y-axis). The contigs distance is in kilo-basepair (Kbp). Additional information about the expression pattern of the OMV and host contigs can be found in supplementary figure 10. (C) endogenization of omv_3421 illustrates proximity with Ty1/copia LTR-retrotransposon. Genes are depicted according to the legend at the bottom. The dotted line represents 50% of GC content, and the purple skew indicates the fluctuation of GC along the contig

5.3.5 Phylogenetic placement of dinoflagellate viral elements

To understand the evolutionary relationship of the viral elements decribed here, a phylogenetic analysis was conducted using a set of core proteins previously described for endogenous viruses and NCLDVs (Bellas et al., 2023; C. M. Bellas & Sommaruga, 2021; Blanc et al., 2015; Yutin, Shevchenko, et al., 2015, 2015). We used protein sequences of MCP, ATPase, Helicase family 1, and pPolB to place the dinoflagellate viral elements within the phylogenetic context of small endogenous viruses. Our analysis successfully reproduced the major PLV, MP and virophage clades obtained in previous studies. The viral elements found in MGD and TGD were assigned within the MP group and clustered together with other dinoflagellate MP in all the phylogenetic reconstructions (Dino, Figures 5.4 and 5.5). On the other hand, the placement of *O. marina* viral elements turned out to be more challenging and no consistent affiliation to any particular viral groups was retrieved.

First, the viral status of elements was confirmed by the double jelly roll structure predicted for MCP (Figure 5.4B), typically described for most of dsDNA viruses (Krupovic & Koonin, 2015). According to MCP phylogeny, *O. marina* viral elements group with Tlr1 virus-like transposable element of *Tetrahymena thermophila*, whereas Mimivirus and phycodnaviruses cluster in a sister clade (Figure 5.4A). The phylogenetic tree of MCP suggests a phylogenetic scenario of proximity between *O. marina* viral element, Tlr1, and Megavirales. A similar conclusion can be inferred from the phylogenetic reconstruction based on ATPase (Figure 5.5A). According to the morphogenesis module (i.e., MCP and ATPase), *O. marina* viral element and Tlr1 may represent an ancient divergence from MP, as previously noticed for Tlr1 (Krupovic et al., 2014). Moreover, helicase family1 failed to recover the Tlr1 and *O. marina* viral elements group. Instead, the *O. marina* viral element appeared as an outgroup of a blended clade, including mobile elements and bacterial and eukaryotic viral elements (Figure 5B). As a sister group, the Tlr1 element is found in a group with Transpovirons, a novel type of mobile element exclusively found in Mimiviruses (Megavirales). Typically, both Tlr1 and PLV lack pPolB, and the phylogenetic reconstruction using pPolB confirmed that *O. marina* viral elements and Dino MP shared a common evolutionary origin (Figure 5C). In summary, the phylogenetic placement of *O. marina* viral elements remains enigmatic, with no apparent affiliation to any previously described groups. Like Tlr1, the *O. marina* element appears to represent a unique type of virus-like transposable element. An expected increase in sampling will help clarify those fascinating elements' evolutionary history.

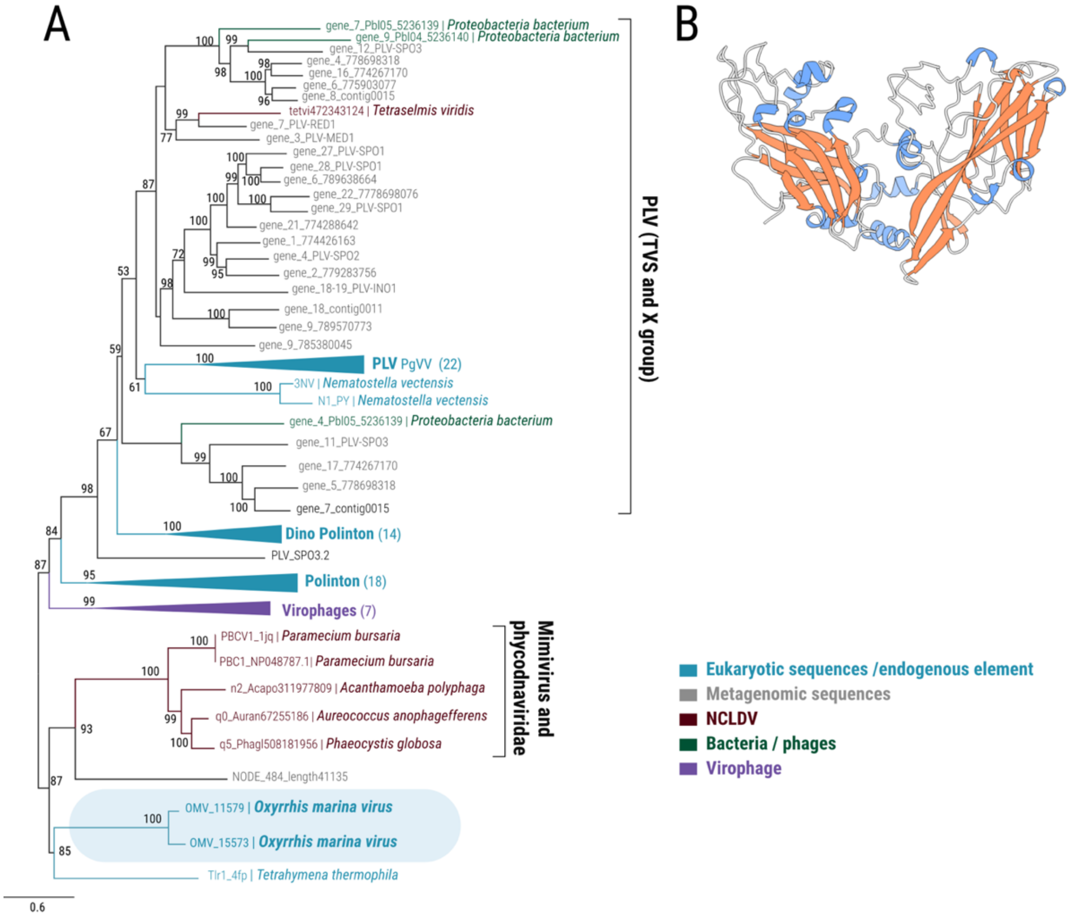

A
gene_7_Pbl05_5236139 | Proteobacteria bacterium
gene_9_Pbl04_5236140 | Proteobacteria bacterium
gene_12_PLV-SPO3
gene_4_778698318
gene_16_774267170
gene_6_775903077
gene_8_contig0015
tetvi472343124 | Tetraselmis viridis
gene_7_PLV-RED1
gene_3_PLV-MED1
gene_27_PLV-SPO1
gene_28_PLV-SPO1
gene_6_789638664
gene_22_7778698076
gene_29_PLV-SPO1
gene_21_774288642
gene_1_774426163
gene_4_PLV-SPO2
gene_2_779283756
gene_18-19_PLV-INO1
gene_18_contig0011
gene_9_789570773
gene_9_785380045
PLV PgVV (22)
3NV | Nematostella vectensis
N1_PY | Nematostella vectensis
gene_4_Pbl05_5236139 | Proteobacteria bacterium
gene_11_PLV-SPO3
gene_17_774267170
gene_5_778698318
gene_7_contig0015
PLV (TVS and X group)
Dino Polinton (14)
PLV_SPO3.2
Polinton (18)
Virophages (7)
PBCV1_1jq | Paramecium bursaria
PBC1_NP048787.1| Paramecium bursaria
n2_Acapo311977809 | Acanthamoeba polyphaga
q0_Auran67255186 | Aureococcus anophagefferens
q5_Phagl508181956 | Phaeocystis globosa
Mimivirus and phycodnaviridae
NODE_484_length41135
OMV_11579 | Oxyrrhis marina virus
OMV_15573 | Oxyrrhis marina virus
Tlr1_4fp | Tetrahymena thermophila
0.6
B
Eukaryotic sequences /endogenous element
Metagenomic sequences
NCLDV
Bacteria / phages
Virophage

Figure 5.4. Phylogenetic placement of viral elements on the phylogenetic context of MP, PLVs, and virophages.

(A) Phylogenetic placement of the viral elements based on ML of MCP (LG+G+F, 451 amino acids). The placement of *O. marina* viral elements is highlighted in light blue. MGD and TGD are nested in the Dino Polinton group. Branches with high support (>70%) representing a single group of elements were collapsed. Support values below 50% are not shown. The colour of the branch tip indicates the origin of the sequence/element according to the legend. Each taxon is labelled with the hostname and the original sequence's code. The sequences were obtained from a set of studies described in the methods section. (B) Structural model prediction for MCP of *O. marina* showing a double jelly-roll capsid structure. The colour indicates the secondary structure: orange beta-strand, and blue: alpha-helix.

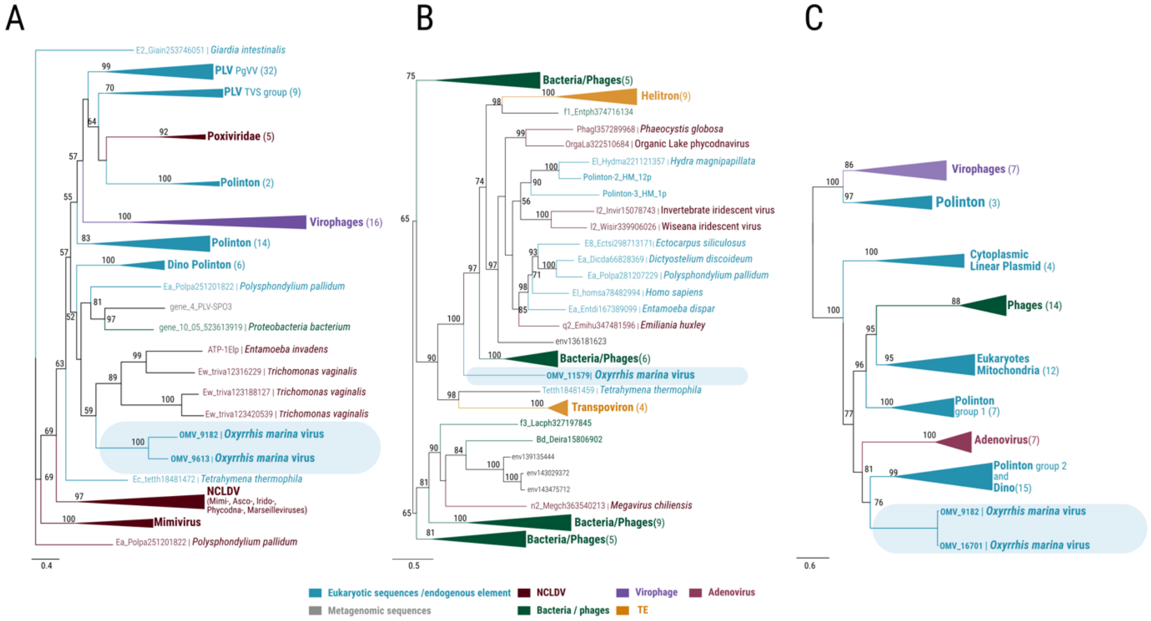

A
E2_Giain253746051 | Giardia intestinalis
99
PLV PgVV (32)
70
PLV TVS group (9)
64
92
Poxiviridae (5)
57
100
Polinton (2)
55
100
Virophages (16)
83
Polinton (14)
57
100
Dino Polinton (6)
Ea_Polpa251201822 | Polysphondylium pallidum
81
gene_4_PLV-SPO3
52
97
gene_10_05_523613919 | Proteobacteria bacterium
ATP-1Elp | Entamoeba invadens
99
Ew_triva12316229 | Trichomonas vaginalis
63
89
Ew_triva123188127 | Trichomonas vaginalis
100
59
Ew_triva123420539 | Trichomonas vaginalis
69
OMV_9182 | Oxyrrhis marina virus
100
OMV_9613 | Oxyrrhis marina virus
69
Ec_tetth18481472 | Tetrahymena thermophila
97
NCLDV
(Mimi-, Asco-, Irido-,
Phycodna-, Marseilleviruses)
100
Mimivirus
Ea_Polpa251201822 | Polysphondylium pallidum
0.4
B
75
Bacteria/Phages(5)
100
Helitron(9)
98
f1_Entph374716134
99
Phagl357289968 | Phaeocystis globosa
OrgaLa322510684 | Organic Lake phycodnavirus
100
El_Hydma221121357 | Hydra magnipapillata
74
Polinton-2_HM_12p
90
Polinton-3_HM_1p
56
65
100
I2_Invir15078743 | Invertebrate iridescent virus
I2_Wisir339906026 | Wiseana iridescent virus
E8_Ectsi298713171 | Ectocarpus siliculosus
93
100
Ea_Dicda66828369 | Dictyostelium discoideum
97
97
71
Ea_Polpa281207229 | Polysphondylium pallidum
98
El_homsa78482994 | Homo sapiens
85
Ea_Entdi167389099 | Entamoeba dispar
q2_Emihu347481596 | Emiliania huxley
env136181623
100
100
Bacteria/Phages(6)
90
OMV_11579| Oxyrrhis marina virus
Tetth18481459 | Tetrahymena thermophila
98
100
Transpoviron (4)
f3_Lacph327197845
Bd_Deira15806902
90
84
env139135444
100
100
env143029372
81
env143475712
n2_Megch363540213 | Megavirus chiliensis
65
100
Bacteria/Phages(9)
81
Bacteria/Phages(5)
0.5
C
86
Virophages (7)
100
97
Polinton (3)
100
Cytoplasmic
Linear Plasmid (4)
100
88
Phages (14)
95
95
Eukaryotes
Mitochondria (12)
96
100
Polinton
group 1 (7)
77
100
Adenovirus(7)
81
99
Polinton group 2
and
Dino(15)
76
OMV_9182 | Oxyrrhis marina virus
OMV_16701 | Oxyrrhis marina virus
0.6
Eukaryotic sequences /endogenous element
NCLDV
Virophage
Adenovirus
Metagenomic sequences
Bacteria / phages
TE

Figure 5.5. Phylogenetic placement of viral elements on the phylogenetic based on ATPase, PolB, and Helicase family 1.

(A) ATPase ML phylogeny (LG model, 238 amino acids) of viral elements. MGD and TGD branch in the Polinton group. (B) ML phylogeny of Helicase family 1 (LG model, 216 amino acids). (C) ML phylogeny based on pPolB aminoacidic sequence (model LG+G+F, 665 amino acids). *O. marina* viral element placements are highlighted in light blue. Nodes with high bootstrap (>70%) were collapsed. Nodes with bootstrap support values below 50% are not shown. Elements aggrupation was defined according to the set of studies indicated in the methods section. The colour code indicates the origin of the sequence/element according to the legend.

5.4 DISCUSSION

This survey identified endogenous DNA viruses as common inhabitants of the genomes of dinoflagellates. 33 viral elements ranging from 5 to 26 kbp long were identified in the preliminary genome assemblies of lesser-studied free-living dinoflagellates (i.e., *O. marina*, TGD, and MGD) with large genomes. These findings align with a recent study that revealed the presence of endogenous viruses hosted in dinoflagellate genomes (Bellas et al., 2023), indicating they are more abundant than previously thought. Likewise, marine viruses such as Mimiviridae and Phycodnaviridae are abundant during dinoflagellate blooms and persist as chronic infections by exploiting the cell machinery (J. Wang et al., 2023). However, the specific role of viral elements shaping the genomes of dinoflagellates remains to be seen due to the scarcity of sequencing data for most dinoflagellates, especially free-living with large genomes. Furthermore, limitations related to sequencing technology and the complex nature of endogenous viral elements pose challenges to their detection. First, most of the assemblies available for dinoflagellates are based on short-read sequencing that tends to be biased against low-GC DNA, leading to misrepresenting viral elements (Hackl et al., 2021). Secondly, poorly annotated dinoflagellates genome and transcriptomes, predominantly comprised of uncharacterized proteins ("dark proteins") (Stephens et al., 2018), impede the identification of endogenous viral proteins. Third, the high number of unknown and chimeric viral genes hinders the identification of viral elements based on functional annotation. Here, we demonstrate that genomic surveys based on long-read sequencing can resolve and assemble complex viral elements and accurately detect their insertions. The prevalence of endogenous viruses in the genome of these free-living dinoflagellates encourages further sequencing efforts in less-studied dinoflagellates to uncover potentially larger fractions of these elements.

The analysis of the gene repertory of *O. marina* viral elements revealed a significantly reduced morphogenesis module, suggesting a transposable element lifestyle. The morphogenesis module (i.e., ATPase, cysteine protease, MCP, and mCP) is highly conserved in endogenous viruses and the kingdom of Bamfordvirae (Krupovic & Koonin, 2015). Notably, the *O. marina* viral elements morphogenesis module only included MCP and ATPase. All viral elements lacked mCP, and additional efforts to identify it failed. The absence or degeneration of mCP has likely contributed to the transition to a mobile element lifestyle. However, the possibility that mCP is highly divergent and that an unidentified protein may assume its role cannot be ruled out. Furthermore, the *O. marina* viral element does not encode maturation protease, which is ubiquitous in virophages and MP and essential for the proteolytic processing of immature virions. This strongly suggests that the virion assembly of the *O. marina* viral element is no longer conducted, thus remaining as a provirus. This, however, does not preclude replication and movement of the element as a transposable element, A dual lifestyle has been preserved in MP and PLV, like the observed in Mu-like bacteriophages that acquired transposition capabilities while remaining as a bona fide virus (Koonin & Dolja, 2014). However, transitioning from bona fide viruses to mobile elements in viral endogenous elements remains unclear.

The *O. marina* viral elements possess a redundant integration module encoding proteins capable of resolving complex recombinant structures. Unlike all MP, PLV, and most virophages (C. M. Bellas & Sommaruga, 2021; Yutin et al., 2013; Yutin, Shevchenko, et al., 2015), *O. marina* viral elements lack the typical reverse transcriptase-derived integrase (RVE-INT). Instead, they encode a set of alternative proteins potentially involved in the viral integration, including GIY-YIG endonuclease, Y-rec (tyrosine recombinase), and the VRR-NUC endonuclease domain, which can be stand-alone or fused with the DNA-binding BRO-N domain. Although the precise mechanism of integrating these elements remains unclear, a Y-rec (OLV-11) homolog has been frequently associated with integration hotspots in the virophage (Santini et al., 2013). It has been proposed to mediate the integration of PLV (Yutin, Shevchenko, et al., 2015). Furthermore, the *O. marina* viral element encodes HJR, potentially involved in resolving specific complex recombinant structures formed during viral integration (i.e., the holiday junctions). HJR has not been found until now in small endogenous eukaryotic viruses. However, it has been reported in Megavirales (e.g., poxvirus), resolving concatemers of the viral genome into unit-length molecules (Garcia et al., 2000). Apparently, *O. marina* elements can encode several proteins and domains that may optimize viral DNA integration into the host genome.

Typically, the replication module comprehends pPolB in all MP and some virophages (C. M. Bellas & Sommaruga, 2021; Yutin et al., 2013; Yutin, Shevchenko, et al., 2015). However, pPolB is absent or inactivated in PLVs due to the fusion with helicase superfamily 1 (SFH1) (Krupovic et al., 2016). Additionally, in some PLVs and virophages, SF1H and SF3H are fused with bacterial DNA polymerase (TVpol), potentially involved in viral replication (Iyer et al., 2008). In the case of *O. marina* viral elements, SF1H and pPolB were found in proximity but never fused. The high prevalence of both enzymes indicates that they are essential for viral genome replication, but their specific function is unknown.

Accessory or cargo genes of *O. marina* viral elements are commonly observed in small and large endogenous viruses and are predicted to facilitate protein interactions, attachment, and recombination (Mougari et al., 2020). DNA methylase, endonucleases, and tyrosine recombinase are commonly observed in PLVs (C. M. Bellas & Sommaruga, 2021; Yutin et al., 2013; Yutin, Kapitonov, et al., 2015). Moreover, fused BRO-N with the endonuclease VRR-NUC was found in almost all *O. marina* viral elements, suggesting that it may be essential for viral genome recombination. The BRO-N domain has a high affinity for DNA and is usually found expanded in members of Megavirales (i.e., Ascoviridae, Iridoviridae and Poxiviridae) (Iranzo et al., 2016; Krupovic & Koonin, 2015). The exact function of this protein is unknown, but the BRO-N domain has been proposed to be involved in the transcriptional regulation of viral or host genes (Zemskov et al., 2000).

The complex and entangled web of interactions among endogenous viral elements poses significant challenges for the *O. marina* viral element phylogenetic placement. First, most gene content and modules are shared with different mobile elements, linear plasmids and bacteriophages. However, phylogenetic reconstructions based on MCP closely link MP, PLV virophages and NCLDV, suggesting a shared common ancestor with phages (Krupovic & Koonin, 2015). Secondly, the distribution of pPolB across viral elements is scattered, exhibiting functional constraints in most PLVs and absence in NCLDVs. However, pPolB phylogenetic analysis distinguishes between eukaryotic and prokaryotic viruses, positioning MP as a basal and paraphyletic group (Koonin & Krupovic, 2017;

Krupovic et al., 2016; Krupovic & Koonin, 2015). Thirdly, chimeric elements and highly divergent lineages of PLVs are susceptible to phylogenetic artifacts such as long-branch attraction (Yutin, Shevchenko, et al., 2015). Fourth, there needs to be more evidence to support the hypothesis of the origin and evolution of the endogenous virus (i.e., virophage first and nuclear scape). This scenario and the lack of viral representation make the *O. marina* viral element precise identification difficult.

While TGD and MGD belong to the specific group of dinoflagellate MP (Bellas et al., 2023), consistently recovered in all phylogenetic reconstructions, the placement of *O. marina* viral elements remains less defined. Phylogenetic analyses based on MCP and ATPase indicate the proximity of *O. marina* viral elements to Tlr1 PLV and Megavirales. This suggests they share a more recent common ancestor from which the morphogenetic module was inherited. This proximity might be supported by the presence of HJR and BRO-N domain in Megavirales and *O. marina* viral elements. This scenario aligns with the nuclear-scape hypothesis, proposing that Megavirales evolved and inherited most of the conserved core of genes from MP (Koonin & Krupovic, 2017; Krupovic & Koonin, 2015). Alternatively, ancestral *O. marina* PLV could have parasitized large DNA viruses and acquired HJR and BRO-N. The SF1H phylogeny positions *O. marina* viral elements as a deeply diverging lineage rather than nested or affiliated to any group. Furthermore, the pPolB phylogeny supports a scenario in which the pPolB of *O. marina* viral elements and dinoflagellate MP originate from a common ancestral MP or phage. Likely, *O. marina* viral elements represent a different PLV lineage with a transposable element lifestyle, which has partially lost its morphogenesis toolbox, resulting in the inability to form virion particles and complete an infective life cycle.

5.4.1 Conclusion

Endogenous viral elements are prevalent in the giant dinoflagellate genomes, which might serve as a "shelter" for viral genomes. The prevalence of dinoflagellates in aquatic ecosystems and the high plasticity of their genome, including a significant fraction of active TE, may facilitate viral integration and propagation. Consequently, the genomes of dinoflagellates constitute a unique genetic environment that could favour recombination between diverse viral elements and TE, resulting in new chimeric combinations. From this point of view, dinoflagellates could constitute oceanic-scale crucibles of viruses that influence the evolution of countless life forms. Additionally, novel sequencing technology has been essential for identifying endogenous viruses. The *O. marina* viral element potentially represents a novel PLV lineage that has adopted a transposable element-like lifestyle. Despite lacking essential components for assembling virions, this element displays a set of genes potentially involved in viral DNA integration. The diverse gene content and intricate phylogenetic placement of this viral element highlight its mosaic nature and complex evolutionary history. Additional studies and sequencing efforts are required to explore the free-living dinoflagellate genome and potentially reveal more extensive and diverse fractions of endogenized viral elements.

CHAPTER 6 GENERAL DISCUSSION

6.1 The functional core dinoflagellates and the mechanism behind rampant gene duplication

Retrogenes have been broadly studied in model organisms and were found to contribute significantly to genetic diversity, phenotypic evolution, and disease research (Casola & Betrán, 2017; Staszak & Makałowska, 2021, 2021). However, few studies have been conducted on non-model organisms, including dinoflagellates. The presence of DinoSL and DinoRL allows the detection of retrogenes and often reveals additional retroduplication or recycling that otherwise would be overlooked (Slamovits & Keeling, 2008b).

The lack of genome sequences limits retrogenes research in dinoflagellates, but since the hallmark of past retroposition events (DinoRL) appears in the mRNA, this limitation can be mitigated using a large transcriptome dataset (Keeling et al., 2014). Our comprehensive analysis confirms the widespread presence of retrogenes across the dinoflagellate diversity (Slamovits & Keeling, 2008b). We revealed that the functional diversity of retrogenes is primarily associated with fundamental cellular processes such as post-translational modification, cell signalling, and transport (Figure 6.1). Unlike in model organisms, dinoflagellate retrogenes have tended to reinforce existing expression profile rather than introducing lineage-specific features or novel functions (Makałowska & Kubiak, 2023). In this case, retrogenes are mostly fated to mirror the parental gene function. Dinoflagellate retrogenes challenge the conventional notion that the vast majority of retrotransposed genes are "dead upon arrival" (Kaessmann et al., 2009b), which suggests pseudogenization as the inevitable outcome due to the absence of regulatory sequences. Contrary to this, most retrogenes are highly expressed and conserve protein similarity. A plausible account for the high survival rates of retrogenes in this lineage is that self-regulatory sequences included in the DinoSL-DinoRL motif (i.e., TTTT) might be involved in the transcription activation (Song et al., 2017b); however, further understanding of how retrogenes are transcribed must be conducted.

Retrogenes are formed by a duplication mechanism that uses the mRNA molecule as an intermediary (Kaessmann et al., 2009b). Reverse transcriptase is required for this process, with retrotransposons likely being the primary contributors. LTR and non-LTR retrotransposon are highly abundant in dinoflagellate genomes (González-Pech et al., 2021b; Stephens et al., 2020) and have been involved in recurrent retroposition events coupled with dramatic climate change episodes (Song et al., 2017b). This is further supported by the overrepresentation of retrogenes associated with reverse transcription function in genomic surveys (Song et al., 2017b). LTR and non-LTR retrotransposon abundance might vary according to each species, lifestyle, and exposure to environmental stressors. Despite these insights, the precise mechanism of retrogene duplication remains unknown. Additional surveys for retroposition signatures, such as polyA tails and target site duplication (TSD), may help to unveil the type of integration and nature of the retrotransposon behind this mechanism.

6.2 OMV: a chimeric PLV with a transposable element lifestyle

A substantial portion of the eukaryotic genome comprises endogenous viral elements (EVEs) originating from RNA viruses and retroelements (Holmes, 2011). These elements have profound implications for genetic disorders and antiviral defence (Chuong et al., 2016; Frank et al., 2022). Emerging evidence suggests that numerous EVEs originate from dsDNA viruses, potentially influencing the structure of the eukaryotic genome (Xia et al., 2024). Recently, it has been demonstrated that small dsDNA viruses are prevalent in the genomes of many unicellular eukaryotes (Bellas et al., 2023), including dinoflagellates. However, their impact on genome organization remains to be seen.

Information regarding small viral endogenous elements in dinoflagellates is scarce due to the absence of genomic sequencing, and the inherent complexity of these viral elements presents challenges for their detection. Through long-read sequencing, we successfully identified numerous EVEs within the genome of *O. marina*. The *O. marina* viral element described here likely represents a unique polinton-like virus (PLV) lineage with proximity to members of Megavirales. Potentially, *O. marina* PLV may provide additional support to the “Nuclear scape hypothesis”, which predicts that most of the conserved core of genes of large dsDNA virus is derived from small dsDNA viruses such as Polinton (Krupovic & Koonin, 2015). Alternatively, frequent events of HGT between different viral entities, including Megavirales, may account for the chimeric gene repertory observed in *O. marina* PLV. Notably, *O. marina* PLV appears to have quickly transitioned to a transposable element's lifestyle (i.e., transcriptional repressed and partial loss of its morphogenesis module) (Figure 6.1), stressing the thin line between these two modes (i.e., as free viral particle versus “captive” transposable element). On the other hand, since stable viral infections were observed in dinoflagellates involved in bloom formation (i.e., *Prorocentrum shikokuense*) and likely regulate the bloom termination (J. Wang et al., 2023), it predicts that a broader diversity of EVEs remains unrevealed in the genome of these protists.

It has been well established that the contribution of bacteriophage integration in the prokaryotic genome is a major driver of innovation and acquisition of critical physiological traits (Ochman et al., 2000). On the other hand, it is traditionally thought that this evolutionary trend is less common in eukaryotic genomes, and these primarily evolved by modification of the existing genetic information (Keeling & Palmer, 2008). However, examples such as *O. marina* PLV challenge this perspective, suggesting that the influence of the dsDNA virus in the genome of microbial eukaryotes could be more significant than initially thought. The impact of the acquisition of genes from dsDNA viruses, such as DVNPs, and its consequences for adaptation is starting to be recognized in dinoflagellates. Therefore, this study underscores the implications of endogenizing small viral elements in dinoflagellate genomes and highlights the need for further research in this emerging area.

6.3 Mitochondrial genome dynamics

The size and content of mitochondrial genomes (mitogenomes) vary significantly among organisms. Unicellular eukaryotes, such as dinoflagellates and apicomplexans, display remarkably reduced and fragmented mitogenomes (Berná et al., 2021; Flegontov & Lukeš, 2012). The mitogenome of the dinoflagellate *O. marina* includes only two protein-coding genes and a series of unique features that prompted questions about its evolution, highlighting its distinctive characteristics in the broader context of mitochondrial genomic diversity. Nevertheless, challenges such as genome fragmentation, a high proportion of non-coding regions, and bacterial contamination complicate the analysis. Although long-read sequencing effectively generated high-quality mtDNA fragments, additional confirmation might be necessary.

O. marina exhibits fragmented mtDNA with a multicopy gene repertory organized in linear chromosomes or mtDNA fragments, containing compact coding regions interspersed with extensive non-coding regions (Slamovits et al., 2007a). Likely, fragmented gene copies have contributed significantly to the overall non-coding DNA content. However, the possibility that transposable elements might also contribute to non-coding fractions cannot be ruled out. On the other hand, the mitogenome topology of *O. marina* involved a more extensive collection of mtDNA fragments due to high recombination levels. The topology observed in *O. marina* aligns with the common fragmented topology characteristic of dinoflagellates, often described as a collection of mitochondrial DNA fragments (reviewed in Flegonotov & Lukeš, 2012; Waller & Jackson, 2009). This contrasts starkly with the large linear, single molecule mitogenome assembly reported for *S. minutum*, which retains conserved homology for non-coding DNA with Apicomplexa (*P. falciparum*) (Shoguchi et al., 2015). It would be expected that this homology should also be conserved in a more early diverging dinoflagellate such as *O. marina*; however, such homology was not detected. This suggests that different recombination rates of mitogenomes among dinoflagellate lineages could significantly influence sequence conservation and the overall mitogenome topology, leading to either fragmentation or enlargement (Waller & Jackson, 2009b). Ultimately, relaxed selective constraints have played a vital role in originating the striking molecular features in the dinoflagellates mitogenomes (Flegontov & Lukeš, 2012).

6.4 *O. marina* genomic survey, gene redundancy and potential mechanism

One of the most prominent characteristics of dinoflagellate genomes is their significant gene redundancy, providing genetic robustness and reinforcement to particular biological functions. Dinoflagellate multicopy genes can be organized in two manners: either arranged in tandem or as isolated gene copies (Mendez et al., 2015; Wisecaver & Hackett, 2011).

The sequencing of dinoflagellates has been limited to the Suessiales order (primarily symbiotic), revealing that tandem gene organization is predominant (~50%) compared to single genes (~12-16%) (Stephens et al., 2020). This organizational pattern has profound implications for chromatin domain formation and transcription (Nand et al., 2020). Although the tandem gene organization was frequently observed in the *O. marina* assembly, it might be underestimated due to the lack of depth and contiguity of the assembly. Nevertheless, this provides sufficient evidence to confirm the prevalence of this gene organization beyond the symbiotic lineage, representing a broader dinoflagellate genome organization that originated early in the history of the lineage. Several propositions attempt to explain gene tandem formation. It has been proposed that RNA-mediated duplication could generate cDNA gene copies inserted into the genome close to the original copy (Slamovits & Keeling, 2008b). Eventually, the integration of reversed-transcribed pre-mature transcripts may account for the high frequency of intronless genes identified in the genomes. Alternatively, concerted evolution and frequent unequal crossing-over may explain the rise of gene tandems in the dinoflagellates (Mendez et al., 2015). However, a comprehensive understanding of the evolution of multigene families in dinoflagellates requires further investigation.

Dinoflagellate genome sizes vary widely (~1-250 Gbp), averaging ten times larger than the human genome. This contrasts with most microbial eukaryotes, whose genomes are typically in the order of Mbp (Hou, 2008). A positive correlation between dinoflagellate cell size and genome size suggests that larger cells can harbour larger genomes, potentially protecting the coding DNA fraction from physical damage (Petrov, 2001). Despite this correlation, neither extremely polyploidy nor expansion of repetitive elements seems to satisfactorily explain the genome size variation among dinoflagellates. In cases such as the *Heterocapsa triquetra* genome (28-23 Gbps), the repeat fraction of the genome represents a minority (McEwan et al., 2008b). Additionally, symbiotic dinoflagellate genomes have shown little evidence for WGD or SD (Lin et al., 2015; Shoguchi et al., 2013b). However, recent findings in the symbiotic dinoflagellate *Durusdinium trenchii* suggest that extensive syntenic blocks may correlate with WGD events (Dougan et al., 2022). Additionally, the expansion of specific TE families has been identified as a leading factor in genome enlargement for *P. glacialis* and *O. marina*. Further sequencing is required to determine if this holds true for most free-living dinoflagellates. Overall, the dinoflagellate genome appears highly plastic, with multiple mechanisms shaping its structure and size, and no single mechanism seems universally dominant.

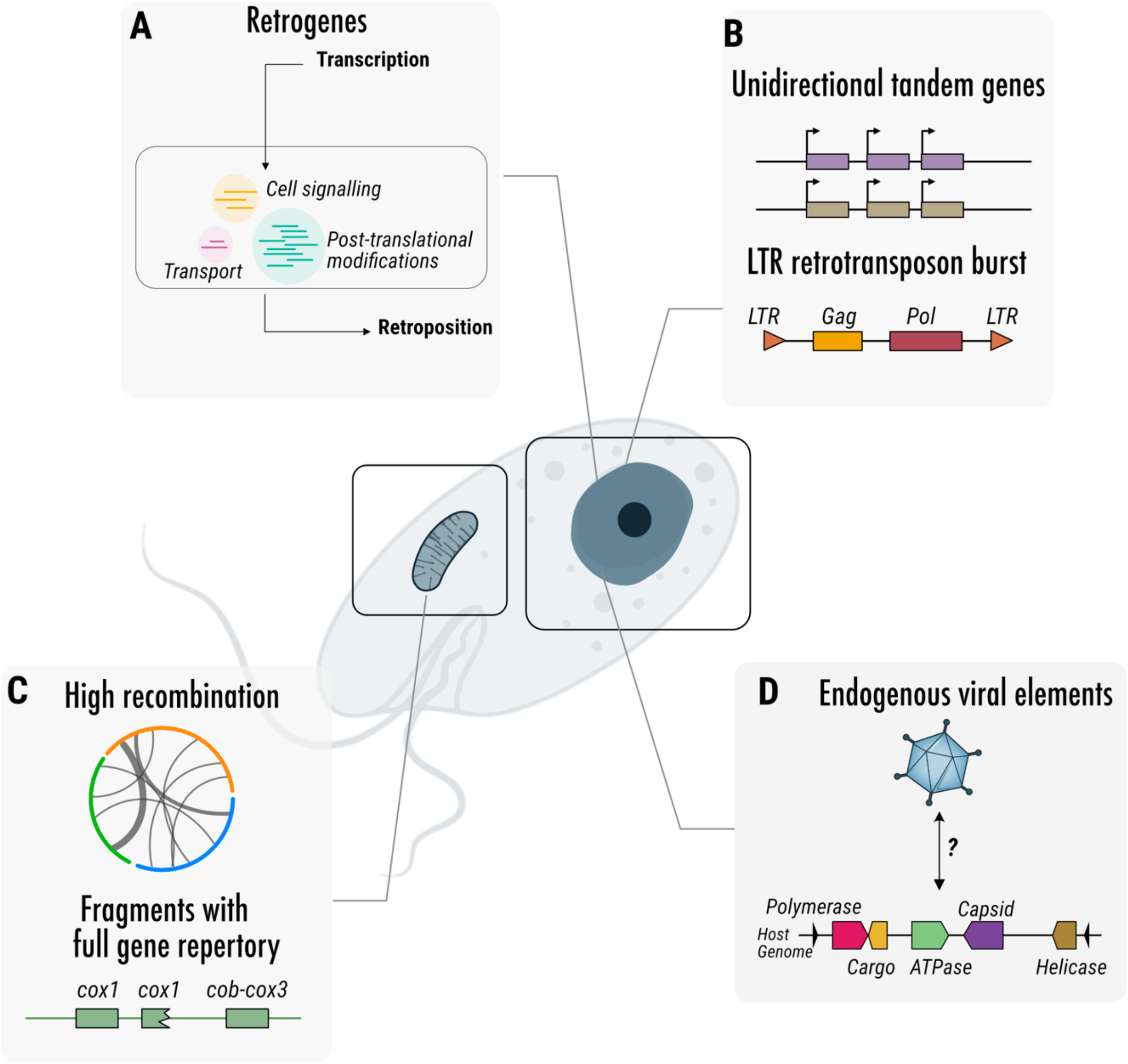


Figure 6.1. Schematic representation of the main findings of this research.
The main findings associated with the retrogenes survey in dinoflagellates (A), *O. marina* nuclear genome survey (B), mitochondrial genome (C), and endogenous viral elements (D).

6.5 Conclusion

Dinoflagellates exhibit a unique genome configuration among eukaryotes, generating considerable interest in unravelling this distinctive genomic makeup. The sequencing of the smaller genomes of symbiotic dinoflagellates has been a significant milestone, revealing striking features such as tandemly organized genes with unidirectional transcription. Despite the challenges posed by their extreme genome size, recent advancements in sequencing and bioinformatics have made genomic exploration of dinoflagellates, such as *O. marina*, feasible. The *O. marina* genomic survey provides valuable information about its genome organization and composition that can be further extrapolated to the entire group (Figure 6.1). Initially observed in the symbiotic lineage, gene tandem organization and redundancy of mobile elements now emerge as genomic signatures for the entire lineage originated at the early stage of dinoflagellate radiation. Endogenous DNA viruses, typically overlooked in dinoflagellate genomes, emerge as preponderant elements with significant implications for genome organization and evolutionary novelties.

While retrogenes have been extensively studied in model organisms, their exploration of dinoflagellates has been limited. This thesis tested the assumption that transcript abundance correlates with retrogene diversity, demonstrating that the functional core of dinoflagellates determines retrogene functional diversity. Retrotransposons are pointed to as mediators of gene duplication, and environmental stressors are suggested as triggers for gene retroduplication. Eventually, genes whose expression is stimulated by environmental stressors may end up as retrogenes, establishing a positive feedback loop that reinforces functional responses. These observations constitute further evidence for the critical role of gene duplication in the regulation of gene expression, in addition to its contribution to driving the genome size. Additionally, comprehensive genomic surveys are crucial for a deeper understanding of this mechanism and its contribution to dinoflagellate gene redundancy and genome size.

The dinoflagellate mitogenome is characterized by its reduced and fragmented nature but also because of its unconventional set of oddities. Despite this, the dinoflagellate mitogenome has received limited attention and remains unexplored using modern sequencing methods. *O. marina* genomic survey has been powerful enough to detect its highly fragmented mitogenome. This data consolidates and completes previous knowledge about the nature of this genome but also expands its organization, including the unveiled set of gene arrangements in the mtDNA molecules. Recombination emerges as the primary driver of dinoflagellate mitogenomes. Further endeavours to explore additional mitogenomes are essential to comprehend the prevalence of recombination and the diversity of mitogenome topology.

APPENDIX A SUPPLEMENTARY TABLES

Table 1. The top ten most abundant PFAM and Panther protein domains were identified in the retrogene dataset.

	PFAM	
PFAM ID	Description	Count
PF00069.27	Protein kinase domain	53
PF00076.24	RNA recognition motif (RRM_1)	46
PF00313.24	Cold-shock DNA-binding domain (CSD)	46
PF00240.25	Ubiquitin	28
PF00036.34	EF hand domain	27
PF00520.33	Ion transport protein	22
PF00025.23	ADP-ribosylation factor (Arf)	22
PF01423.24	LSM domain	18
PF00504.23	Chlorophyll A-B binding protein	15
PF12796.9	Ank_2	13
	PANTHER	
PANTHER FAMILY ID	Description	Count
PTHR21649:SF99	Chlorophyll a-b binding protein, chloroplastic	21
PTHR23003	Rna recognition motif rrm domain containing protein	16
PTHR46565	Cold shock domain protein 2	16
PTHR11143:SF7	60s ribosomal protein l26-related	11
PTHR43400:SF8	Cytochrome b5 heme-binding domain-containing protein	11
PTHR10098:SF106	Tetratricopeptide repeat protein 28	10
PTHR10666:SF238	gh17761p	10
PTHR10442	40s ribosomal protein s21	10
PTHR47170	Malonyl-coa acp transacylase, acp-binding	9
PTHR10037:SF62	Sodium channel protein 60e	9

Table 2. Protein-coding genes and rRNA annotation for *O. marina* mitochondrial genomes

Start	End	Annotation	MITOS [a]	Strand	Blast [b]	size (bp)
			Score/pvalue		evalue	
chromosome_1						
7856	9274	Cox1	347391474.6	-		1418
10688	10759	LSUG	–	-	1.10E-10	71
chromosome_2						
9414	10268	RNA10_1	–	-	1.27E-166	854
11385	11826	Cox1_1	–	-	4.60E-10	441
16011	16166	LSUE	3.10E-08	-	–	155
18016	18307	Cob	12456185	-	–	291
19347	19512	RNA10_2	0.4229	-	2.31E-80	165
23536	24163	RNA10_3	0.1804	-	1.63E-162	627
25335	26754	Cox1_2	346788847.8	+	–	1419
32287	32333	RNA10_4	–	+	1.18E-18	46
chromosome_3						
3260	5132	Cob-Cox3	133680838.6	+	–	1713
15027	15531	Cox1	–	+	8.40E-12	504
23870	24670	Cox	0.1109	+	–	800
27281	28700	Cox1_2	346591072.6	+	–	1419
39839	40572	LSUE	4.80E-07	-	–	733

[a] mitos1 quality values for protein coding genes and pvalues for RNA

[b] blastn and blastp e-values rRNA and protein coding genes search, respectively

Table 3. Annotation of repeat content for mitochondrial chromosomes based on RepeatMasker approach.

Query sequence	Repeat	Class/family
chro_19126	(TCTTCTC)n	Simple_repeat
chro_19126	(G)n	Simple_repeat
chro_19126	C	MamTip2
chro_19126	(G)n	Simple_repeat
chro_19126	(T)n	Simple_repeat
chro_19126	(CTTAT)n	Simple_repeat
chro_19126	(TCTT)n	Simple_repeat
chro_15870	(AAGAT)n	Simple_repeat
chro_15870	(CAATAG)n	Simple_repeat
chro_15870	(TCATT)n	Simple_repeat
chro_15870	(TCTTCTC)n	Simple_repeat
chro_15870	A-rich	Low_complexity
chro_15886	(TCTTTTCA)n	Simple_repeat
chro_15886	(CTTTCT)n	Simple_repeat
chro_15886	(CGATA)n	Simple_repeat
chro_15886	(TTGCTA)n	Simple_repeat
chro_15886	(ATTTT)n	Simple_repeat
chro_15886	(TCTTCTC)n	Simple_repeat
chro_15886	(TCATT)n	Simple_repeat
chro_15886	(TCGTA)n	Simple_repeat
chro_15800	(G)n	Low_complexity
chro_15800	(G)n	Low_complexity
chro_15800	(T)n	Low_complexity
chro_15800	(TG)n	Low_complexity
chro_15800	(TCTTTAT)n	Simple_repeat
chro_15800	(TTCTATT)n	Simple_repeat
Total repeats:		27 (0.90 %)

Table 4. Detailed genome assembly completeness assessment report using BUSCO Eukaryote_db9, Alveolate_db10, and CEGMA.

Categories	Orthologs
Eukaryote_db9	
Complete BUSCOs (C)	67
Complete and single-copy BUSCOs (S)	29
Complete and duplicated BUSCOs (D)	38
Fragmented BUSCOs (F)	8
Missing BUSCOs (M)	228
Total BUSCO groups searched	303
Alveolate_db10	
Complete BUSCOs (C)	33
Complete and single-copy BUSCOs (S)	14
Complete and duplicated BUSCOs (D)	19
Fragmented BUSCOs (F)	6
Missing BUSCOs (M)	132
Total BUSCO groups searched	171
CEGMA	
Complete	64
%Ortho*	79
Total	248

*Percentage of detected CEGS that have more than 1 ortholog

Table 5. Top 30 InterPro domains

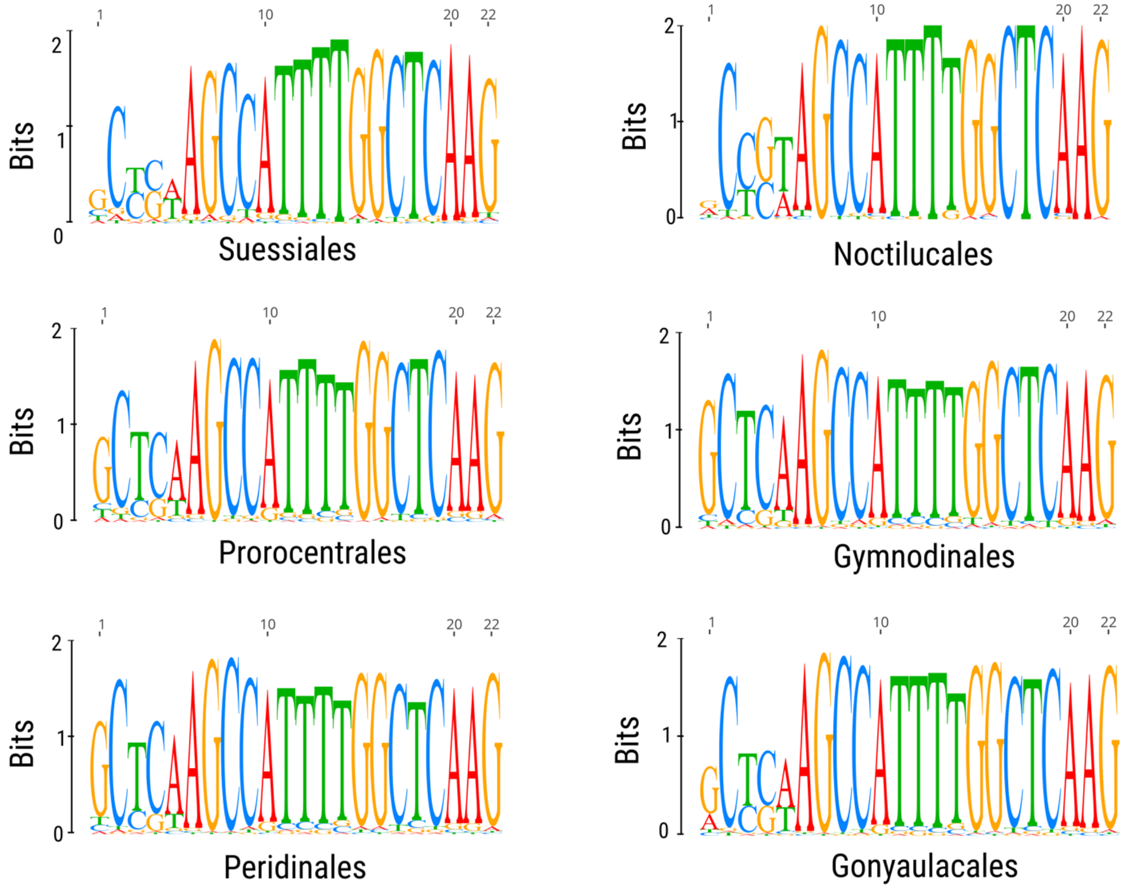
Description	InterPro ID	Protein domain frequency
Reverse transcriptase domain	IPR000477	307
Endonuclease/exonuclease/phosphatase superfamily	IPR036691	162
Ribonuclease H-like superfamily	IPR012337	135
Aspartic peptidase domain superfamily	IPR021109	131
Integrase, catalytic core	IPR001584	96
Antifreeze protein, type I	IPR000104	54
EF-hand domain	IPR002048	54
Reverse transcriptase, RNA-dependent DNA polymerase	IPR013103	51
P-loop containing nucleoside triphosphate hydrolase	IPR027417	38
Protein kinase domain	IPR000719	33
Archaeal/bacterial/fungal rhodopsins	IPR001425	31
Ribonuclease H superfamily	IPR036397	30
HNH nuclease	IPR003615	30
Ankyrin repeat	IPR002110	25
Tetratricopeptide-like helical domain superfamily	IPR011990	25
Zinc finger, CCHC-type superfamily	IPR036875	20
RNA recognition motif domain	IPR000504	20
Regulator of chromosome condensation, RCC1	IPR000408	20
Zinc finger, CCHC-type superfamily	IPR036875	20
Leucine-rich repeat domain superfamily	IPR032675	19
ABC transporter type 1, transmembrane domain MetI-like	IPR000515	18
SRCR domain	IPR001190	17
ABC transporter-like, ATP-binding domain	IPR003439	17
AMP-dependent synthetase/ligase	IPR000873	16
Peptidase A2A, retrovirus, catalytic	IPR001995	14
Mei2-like, C-terminal RNA recognition motif	IPR007201	14
Peptidase A2A, retrovirus, catalytic	IPR001995	14
Heat shock protein Hsp90 family	IPR001404	14
EF-Hand 1, calcium-binding site	IPR018247	13
Signal transduction response regulator	IPR001789	13

Table 6. Genes of interest predicted for *O. marina*.

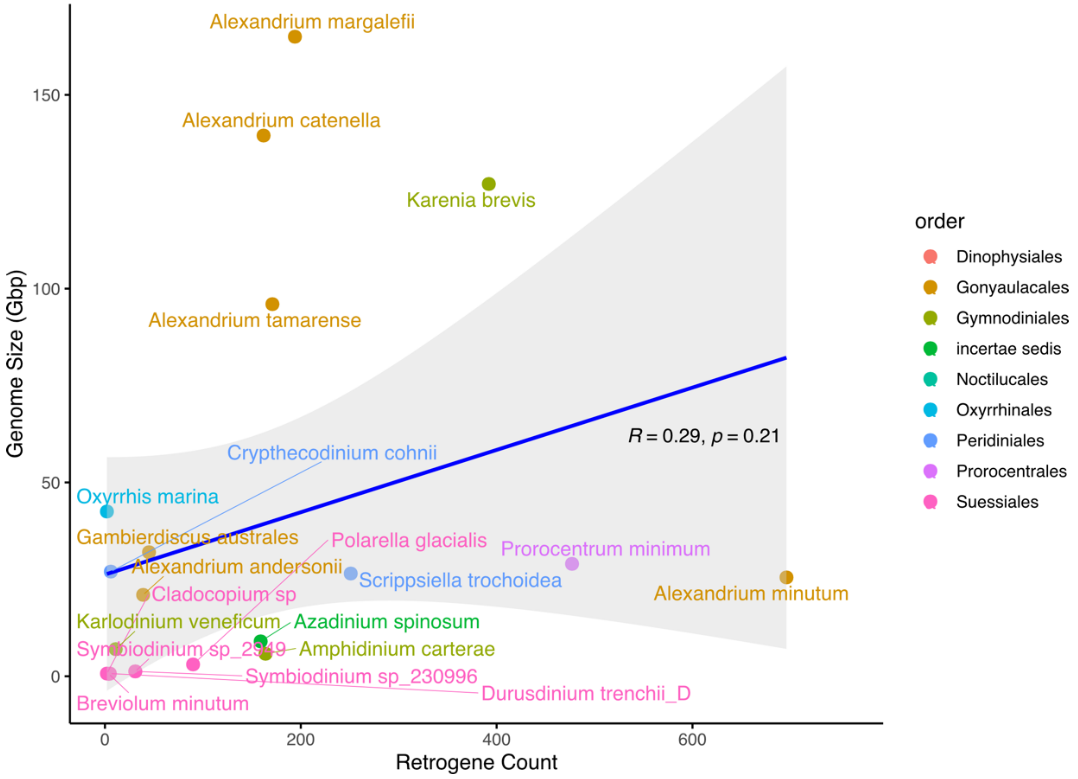
Contig	**Description**	**InterPro**	**Copy number**
	Chromatin and chromosome organization		
contig_7746	Linker histone, H1/H5	IPR005819	4
contig_14110	Histone H4	IPR001951	1
contig_16271	DVNP family	IPR043928	6
contig_15401	Regulator of chromosome condensation, RCC1	IPR000408	20
contig_13753	Regulator of chromosome condensation, RCC2	IPR009091	3
	Plastid-origin genes		
contig_16587	Ketol-acid reductoisomerase	IPR000506	1
contig_18843	Glutamine synthetase	IPR008146	1
	Meiosis-related genes		
scaffold_16835	Mei2-like	IPR007201	13
contig_9753	Meiosis-expressed gene 1 protein, MEIG1	IPR020186	1
contig_11471	SPO11	IPR002815	1

E-value cutoff of PFAM domain search:1e-5

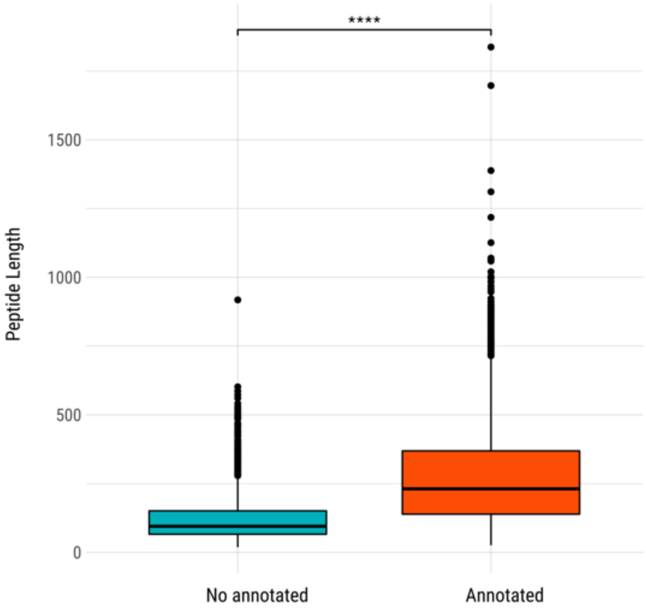
APPENDIX B SUPPLEMENTARY FIGURES



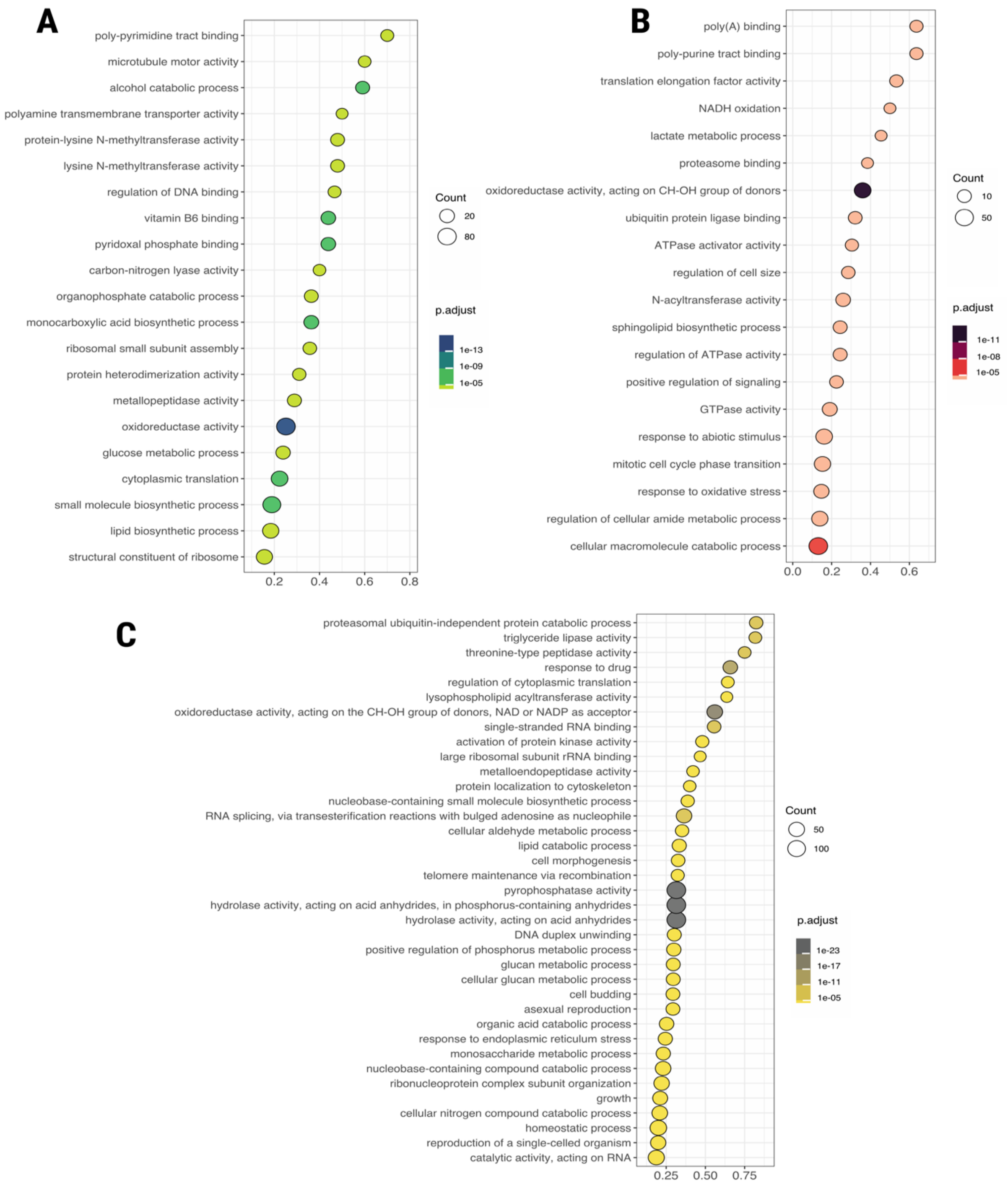
Supplementary figure 1. Retrogenes SL consensus for each dinoflagellate family. Noctilucales logo consensus agrees with the canonical motif for dinoflagellates (DCCGTAGCCATTTTGGCTCAAG, D = A, G, T). In contrast, the rest of the orders show the same DinoSL variant (GCTCAAGCCATTTTGGCTCAAG)



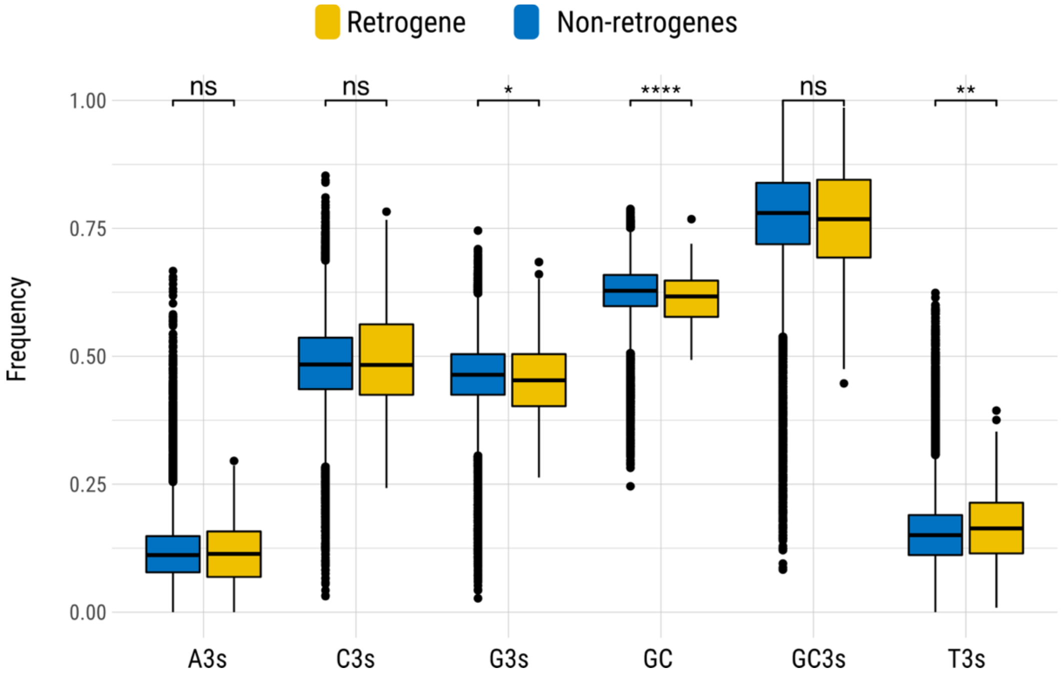
Supplementary figure 2. Regression between genome size and retrogene number. Dinoflagellates orders are depicted in different colours. The shaded area represents the confidence interval.



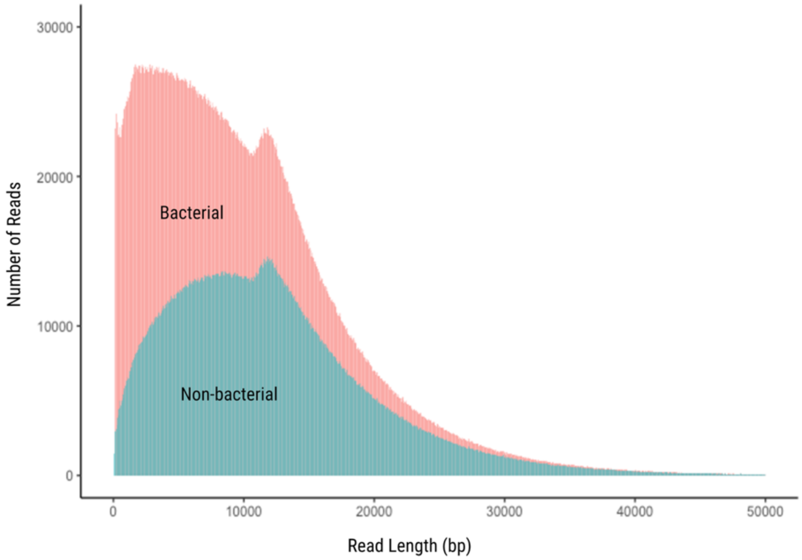
Supplementary figure 3. Size of peptides translated from retrogene. Annotated retrogenes show significantly larger (276 amino acids) sizes compared with no annotated retrogenes (124 amino acids).



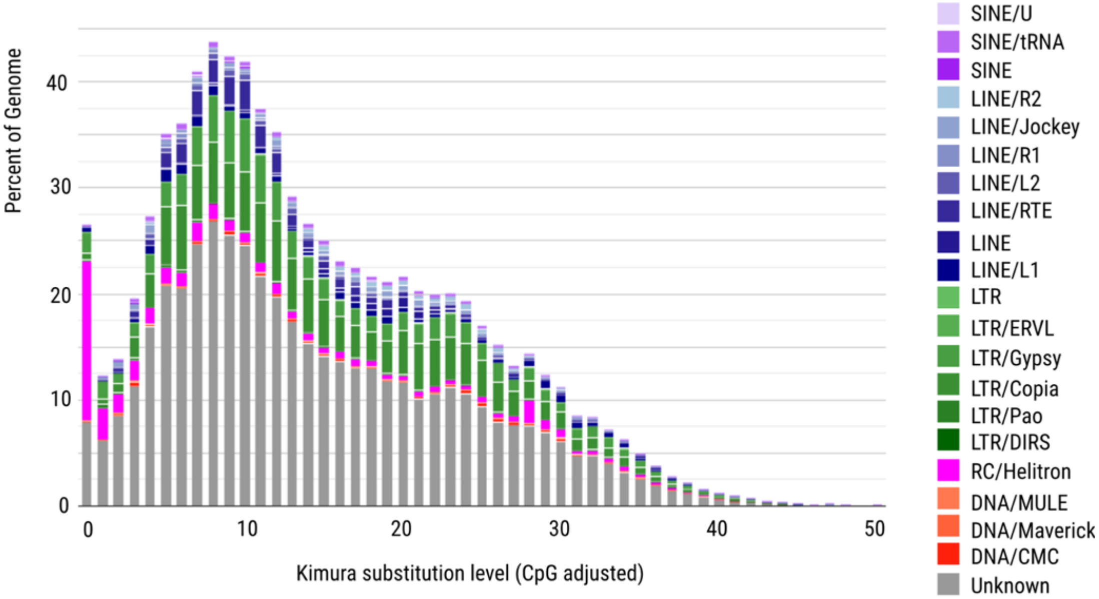
Supplementary figure 4. Unique enriched GO term for each of the selected dinoflagellates. A *A. minutum*, B *B. nutricula*. C *G. catenatum*



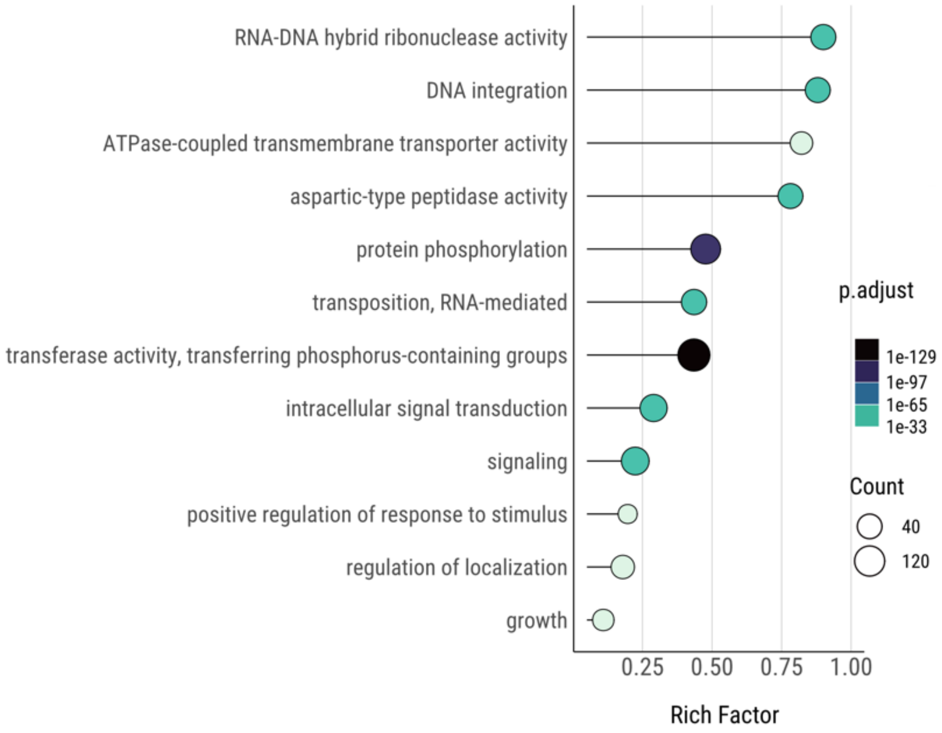
Supplementary figure 5. Comparison of base composition between retrogenes and protein-coding genes. Frequency of adenine, cytosine, guanine, and thymine in the third synonymous codon position (A3s, C3s, G3s, T3s). Also, the frequency of GC at the synonymous third codon position. The difference between the two categories was assessed by Student's t-test (p. adjust < 0.05).



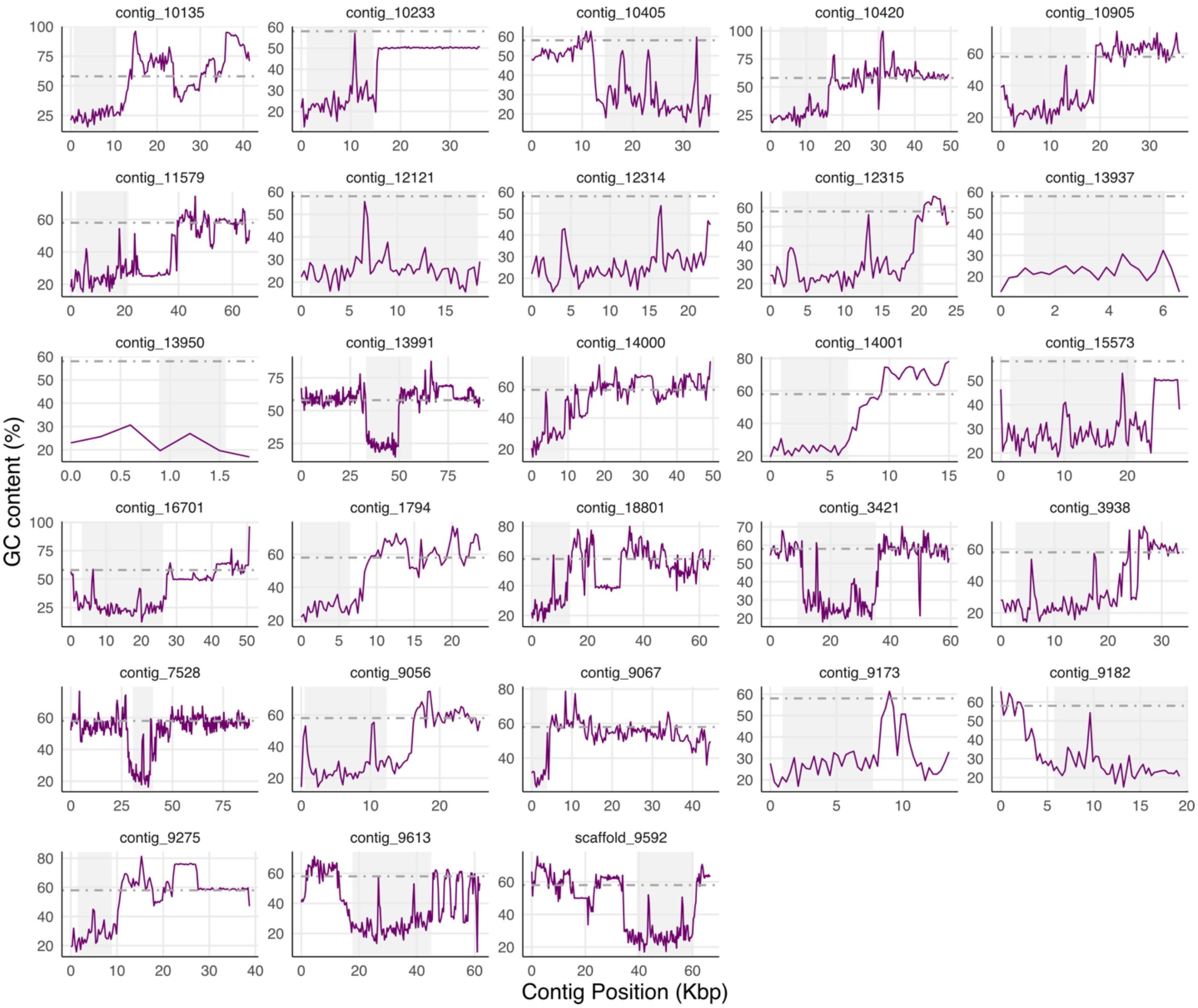
Supplementary figure 6. Proportion of bacterial and non-bacterial reads. Reads length <50 kbp, N50 15 kbp.



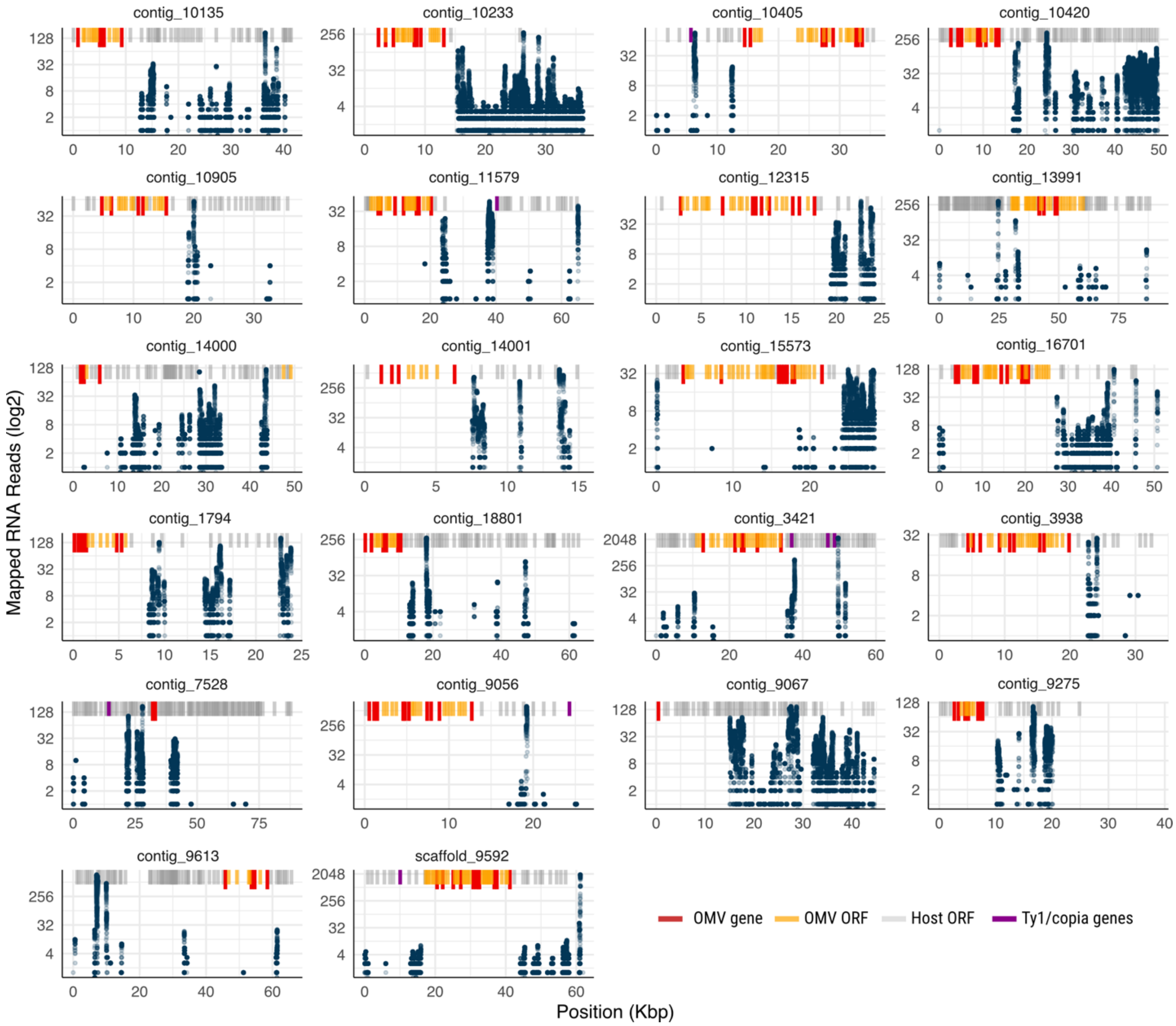
Supplementary figure 7. Repeat landscape derived from Repeatmasker. Distribution and proportion of repeat elements identified according to the kimura model of sequence divergence (sequence divergence from reference). Repeat elements are depicted according to the figure legend.



Supplementary figure 8. Enrichment analysis of retrogenes present in *O. marina*. GO-enriched categories for retrogenes found by DinoRL upstream of the coding region mostly involve DNA mobilization.



Supplementary figure 9. GC content distribution of *O. marina* viral elements and host. Purple skew represents GC content fluctuation along the contig, and the viral regions are highlighted in light gray.



Supplementary figure 10. RNA-seq reads mapping patterns on the host contigs. Blue dots are the number of RNA-seq reads (y-axis) mapped to a particular position of the *O. marina* contigs (x-axis). Red and orange genes depict integrated viral regions. The purple ribbon shows the location of Ty1/copia LTR-retrotransposon.

BIBLIOGRAPHY

Adams, K. L., & Palmer, J. D. (2003). Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. Molecular Phylogenetics and Evolution, 29(3), 380–395. https://doi.org/10.1016/S1055-7903(03)00194-5

Alacid, E., Irwin, N. A. T., Smilansky, V., Milner, D. S., Kilias, E. S., Leonard, G., & Richards, T. A. (2022). A diversified and segregated mRNA spliced-leader system in the parasitic Perkinsozoa. Open Biology, 12(8), 220126. https://doi.org/10.1098/rsob.220126

Allen, J. R., Roberts, T. M., Loeblich, A. R., & Klotz, L. C. (1975). Characterization of the DNA from the dinoflagellate crypthecodinium cohnii and implications for nuclear organization. Cell, 6(2), 161–169. https://doi.org/10.1016/0092-8674(75)90006-9

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. (2016a). Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. Scientific Reports, 6(1). https://doi.org/10.1038/srep39734

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. (2016b). Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. Scientific Reports, 6(1), Article 1. https://doi.org/10.1038/srep39734

Augusto Corrêa dos Santos, R., Goldman, G. H., & Riaño-Pachón, D. M. (2017). ploidyNGS: Visually exploring ploidy with Next Generation Sequencing data. Bioinformatics, 33(16), 2575–2576. https://doi.org/10.1093/bioinformatics/btx204

Bachvaroff, T. R., & Place, A. R. (2008). From Stop to Start: Tandem Gene Arrangement, Copy Number and Trans-Splicing Sites in the Dinoflagellate Amphidinium carterae. PLoS ONE, 3(8), e2929. https://doi.org/10.1371/journal.pone.0002929

Bai, Y., Casola, C., Feschotte, C., & Betrán, E. (2007). Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biology, 8(1), R11. https://doi.org/10.1186/gb-2007-8-1-r11

Barshis, D. J., Ladner, J. T., Oliver, T. A., & Palumbi, S. R. (2014). Lineage-Specific Transcriptional Profiles of Symbiodinium spp. Unaltered by Heat Stress in a Coral Host. Molecular Biology and Evolution, 31(6), 1343–1352. https://doi.org/10.1093/molbev/msu107

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(4), e35269. https://doi.org/10.1371/journal.pone.0035269

Bellas, C., Hackl, T., Plakolb, M.-S., Koslová, A., Fischer, M. G., & Sommaruga, R. (2023). Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses. Proceedings of the National Academy of Sciences, 120(16), e2300465120. https://doi.org/10.1073/pnas.2300465120

Bellas, C. M., & Sommaruga, R. (2021). Polinton-like viruses are abundant in aquatic ecosystems. Microbiome, 9(1), 13. https://doi.org/10.1186/s40168-020-00956-0

Benites, L. F., Stephens, T. G., & Bhattacharya, D. (2022). Multiple waves of viral invasions in Symbiodiniaceae algal genomes. Virus Evolution, 8(2), veac101. https://doi.org/10.1093/ve/veac101

Berná, L., Rego, N., & Francia, M. E. (2021). The Elusive Mitochondrial Genomes of Apicomplexa: Where Are We Now? Frontiers in Microbiology, 12. https://www.frontiersin.org/articles/10.3389/fmicb.2021.751775

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., & Stadler, P. F. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. Molecular Phylogenetics and Evolution, 69(2), 313–319. https://doi.org/10.1016/j.ympev.2012.08.023

Biegala, I. C., Kennaway, G., Alverca, E., Lennon, J.-F., Vaulot, D., & Simon, N. (2002). Identification of Bacteria Associated with Dinoflagellates (dinophyceae) Alexandrium Spp. Using Tyramide Signal Amplification–Fluorescent in Situ Hybridization and Confocal Microscopy1. Journal of Phycology, 38(2), 404–411. https://doi.org/10.1046/j.1529-8817.2002.01045.x

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. Frontiers in Genetics, 4. https://doi.org/10.3389/fgene.2013.00199

Blanc, G., Gallot-Lavallée, L., & Maumus, F. (2015). Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. Proceedings of the National Academy of Sciences, 112(38), E5318–E5326. https://doi.org/10.1073/pnas.1506469112

Boakes, D. E., Codling, E. A., Thorn, G. J., & Steinke, M. (2011). Analysis and modelling of swimming behaviour in Oxyrrhis marina. Journal of Plankton Research, 33(4), 641–649. https://doi.org/10.1093/plankt/fbq136

Boettcher, B., & Barral, Y. (2013). The cell biology of open and closed mitosis. Nucleus, 4(3), 160–165. https://doi.org/10.4161/nucl.24676

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

Bolzán, A. D. (2012). Chromosomal aberrations involving telomeres and interstitial telomeric sequences. Mutagenesis, 27(1), 1–15. https://doi.org/10.1093/mutage/ger052

Borst, P., & Sabatini, R. (2008). Base J: Discovery, Biosynthesis, and Possible Functions. Annual Review of Microbiology, 62(1), 235–251. https://doi.org/10.1146/annurev.micro.62.081307.162750

Brosius, J. (1991). Retroposons—Seeds of Evolution. Science. https://doi.org/10.1126/science.1990437

Brosius, J., & Gould, S. J. (1992). On "genomenclature": A comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proceedings of the National Academy of Sciences of the United States of America, 89(22), 10706–10710.

Bullerwell, C. E., & Gray, M. W. (2004). Evolution of the mitochondrial genome: Protist connections to animals, fungi and plants. Current Opinion in Microbiology, 7(5), 528–534. https://doi.org/10.1016/j.mib.2004.08.008

Burger, G., Forget, L., Zhu, Y., Gray, M. W., & Lang, B. F. (2003). Unique mitochondrial genome architecture in unicellular relatives of animals. Proceedings of the National Academy of Sciences, 100(3), 892–897. https://doi.org/10.1073/pnas.0336115100

Burger, G., Gray, M. W., Forget, L., & Lang, B. F. (2013). Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. Genome Biology and Evolution, 5(2), 418–438. https://doi.org/10.1093/gbe/evt008

Burger, G., Gray, M. W., & Lang, B. F. (2003). Mitochondrial genomes: Anything goes. Trends in Genetics, 19(12), 709–716. https://doi.org/10.1016/j.tig.2003.10.012

Burger, G., Zhu, Y., Littlejohn, T. G., Greenwood, S. J., Schnare, M. N., Lang, B. F., & Gray, M. W. (2000). Complete sequence of the mitochondrial genome of Tetrahymena pyriformis and comparison with Paramecium aurelia mitochondrial DNA11Edited by M. Yaniv. Journal of Molecular Biology, 297(2), 365–380. https://doi.org/10.1006/jmbi.2000.3529

Burki, F., & Kaessmann, H. (2004). Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nature Genetics, 36(10), Article 10. https://doi.org/10.1038/ng1431

Campbell, S., Aswad, A., & Katzourakis, A. (2017). Disentangling the origins of virophages and polintons. Current Opinion in Virology, 25, 59–65. https://doi.org/10.1016/j.coviro.2017.07.011

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Carelli, F. N., Hayakawa, T., Go, Y., Imai, H., Warnefors, M., & Kaessmann, H. (2016). The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Research, 26(3), 301–314. https://doi.org/10.1101/gr.198473.115

Casola, C., & Betrán, E. (2017). The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses? Genome Biology and Evolution, 9(6), 1351–1373. https://doi.org/10.1093/gbe/evx081

Cavalier-Smith, T. (1993). Kingdom protozoa and its 18 phyla. Microbiological Reviews, 57(4), 953–994.

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

Chaput, H., Wang, Y., & Morse, D. (2002). Polyadenylated Transcripts Containing Random Gene Fragments are Expressed in Dinoflagellate Mitochondria. Protist, 153(2), 111–122. https://doi.org/10.1078/1434-4610-00090

Chase, E. E., Desnues, C., & Blanc, G. (2022). Integrated viral elements suggest the dual lifestyle of Tetraselmis spp. Polinton-like viruses. Virus Evolution, 8(2), veac068. https://doi.org/10.1093/ve/veac068

Chen, J. E., Cui, G., Wang, X., Liew, Y. J., & Aranda, M. (2018). Recent expansion of heat-activated retrotransposons in the coral symbiont Symbiodinium microadriaticum. The ISME Journal, 12(2), 639–643. https://doi.org/10.1038/ismej.2017.179

Chen, W., Zuo, C., Wang, C., Zhang, T., Lyu, L., Qiao, Y., Zhao, F., & Miao, M. (2021). The hidden genomic diversity of ciliated protists revealed by single-cell genome sequencing. BMC Biology, 19, 264. https://doi.org/10.1186/s12915-021-01202-1

Chen, Y., González-Pech, R. A., Stephens, T. G., Bhattacharya, D., & Chan, C. X. (2020). Evidence That Inconsistent Gene Prediction Can Mislead Analysis of Dinoflagellate Genomes. Journal of Phycology, 56(1), 6–10. https://doi.org/10.1111/jpy.12947

Chen, Y., Shah, S., Dougan, K. E., van Oppen, M. J. H., Bhattacharya, D., & Chan, C. X. (2022). Improved Cladocopium goreaui Genome Assembly Reveals Features of a Facultative Coral Symbiont and the Complex Evolutionary History of Dinoflagellate Genes. Microorganisms, 10(8), 1662. https://doi.org/10.3390/microorganisms10081662

Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science, 351(6277), 1083–1087. https://doi.org/10.1126/science.aad5497

Clayton, C. (2019). Regulation of gene expression in trypanosomatids: Living with polycistronic transcription. Open Biology, 9(6), 190072. https://doi.org/10.1098/rsob.190072

Correa, A. M. S., Welsh, R. M., & Vega Thurber, R. L. (2013). Unique nucleocytoplasmic dsDNA and +ssRNA viruses are associated with the dinoflagellate endosymbionts of corals. The ISME Journal, 7(1), 13–27. https://doi.org/10.1038/ismej.2012.75

Cuadrado, Á., De Bustos, A., & Figueroa, R. I. (2019a). Chromosomal markers in the genus Karenia: Towards an understanding of the evolution of the chromosomes, life cycle patterns and phylogenetic relationships in dinoflagellates. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-018-35785-7

Cuadrado, Á., De Bustos, A., & Figueroa, R. I. (2019b). Chromosomal markers in the genus Karenia: Towards an understanding of the evolution of the chromosomes, life cycle patterns and phylogenetic relationships in dinoflagellates. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-018-35785-7

Daniels, J.-P., Gull, K., & Wickstead, B. (2010). Cell Biology of the Trypanosome Genome. Microbiology and Molecular Biology Reviews : MMBR, 74(4), 552–569. https://doi.org/10.1128/MMBR.00024-10

de Mendoza, A., Bonnet, A., Vargas-Landin, D. B., Ji, N., Li, H., Yang, F., Li, L., Hori, K., Pflueger, J., Buckberry, S., Ohta, H., Rosic, N., Lesage, P., Lin, S., & Lister, R. (2018). Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons. Nature Communications, 9(1). https://doi.org/10.1038/s41467-018-03724-9

Deng, Y., Hu, Z., Shang, L., Peng, Q., & Tang, Y. Z. (2017). Transcriptomic Analyses of Scrippsiella trochoidea Reveals Processes Regulating Encystment and Dormancy in the Life Cycle of a Dinoflagellate, with a Particular Attention to the Role of Abscisic Acid. Frontiers in Microbiology, 8, 2450. https://doi.org/10.3389/fmicb.2017.02450

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Research, 45(4), e18. https://doi.org/10.1093/nar/gkw955

Dorrell, R. G., & Howe, C. J. (2015). Integration of plastids with their hosts: Lessons learned from dinoflagellates. Proceedings of the National Academy of Sciences, 112(33), 10247–10254. https://doi.org/10.1073/pnas.1421380112

Dougan, K. E., Bellantuono, A. J., Kahlke, T., Abbriano, R. M., Chen, Y., Shah, S., Granados-Cifuentes, C., Oppen, M. J. H. van, Bhattacharya, D., Suggett, D. J., Chan, C. X., & Rodriguez-Lanetty, M. (2022). Whole-genome duplication in an algal symbiont serendipitously confers thermal tolerance to corals (p. 2022.04.10.487810). bioRxiv. https://doi.org/10.1101/2022.04.10.487810

Drechsler, H., & McAinsh, A. D. (2012). Exotic mitotic mechanisms. Open Biology, 2(12), 120140. https://doi.org/10.1098/rsob.120140

Eddy, S. R. (2011). Accelerated Profile HMM Searches. PLoS Computational Biology, 7(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics, 9(1), 18. https://doi.org/10.1186/1471-2105-9-18

Emerson, J. J., Kaessmann, H., Betrán, E., & Long, M. (2004). Extensive Gene Traffic on the Mammalian X Chromosome. Science, 303(5657), 537–540. https://doi.org/10.1126/science.1090042

Feagin, J. E., Mericle, B. L., Werner, E., & Morris, M. (1997). Identification of additional rRNA fragments encoded by the Plasmodium falciparum 6 kb element. Nucleic Acids Research, 25(2), 438–446.

Fensome, R. A. (1993). A Classification of Living and Fossil Dinoflagellates. American Museum of Natural History.

Feschotte, C., & Gilbert, C. (2012). Endogenous viruses: Insights into viral evolution and impact on host biology. Nature Reviews Genetics, 13(4), Article 4. https://doi.org/10.1038/nrg3199

Filée, J. (2014). Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: The visible part of the iceberg? Virology, 466–467, 53–59. https://doi.org/10.1016/j.virol.2014.06.004

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2014). Pfam: The protein families database. Nucleic Acids Research, 42(Database issue), D222–D230. https://doi.org/10.1093/nar/gkt1223

Fischer, M. G. (2021). The Virophage Family Lavidaviridae. Current Issues in Molecular Biology, 40, 1–24. https://doi.org/10.21775/cimb.040.001

Fischer, M. G., & Hackl, T. (2016). Host genome integration and giant virus-induced reactivation of the virophage mavirus. Nature, 540(7632), Article 7632. https://doi.org/10.1038/nature20593

Fischer, M. G., & Suttle, C. A. (2011). A virophage at the origin of large DNA transposons. Science (New York, N.Y.), 332(6026), 231–234. https://doi.org/10.1126/science.1199412

Flegontov, P., & Lukeš, J. (2012). Chapter Six—Mitochondrial Genomes of Photosynthetic Euglenids and Alveolates. In L. Maréchal-Drouard (Ed.), Advances in Botanical Research (Vol. 63, pp. 127–153). Academic Press. https://doi.org/10.1016/B978-0-12-394279-1.00006-5

Frank, J. A., Singh, M., Cullen, H. B., Kirou, R. A., Benkaddour-Boumzaouad, M., Cortes, J. L., Garcia Pérez, J., Coyne, C. B., & Feschotte, C. (2022). Evolution and antiviral activity of a human protein of retroviral origin. Science, 378(6618), 422–428. https://doi.org/10.1126/science.abq7871

Gagat, P., Mackiewicz, D., & Mackiewicz, P. (2017). Peculiarities within peculiarities—Dinoflagellates and their mitochondrial genomes. Mitochondrial DNA. Part B, Resources, 2(1), 191–195. https://doi.org/10.1080/23802359.2017.1307699

Garcia, A. D., Aravind, L., Koonin, E. V., & Moss, B. (2000). Bacterial-type DNA Holliday junction resolvases in eukaryotic viruses. Proceedings of the National Academy of Sciences, 97(16), 8926–8931. https://doi.org/10.1073/pnas.150238697

Garcia-Perez, J. L., Marchetto, M. C. N., Muotri, A. R., Coufal, N. G., Gage, F. H., O'Shea, K. S., & Moran, J. V. (2007). LINE-1 retrotransposition in human embryonic stem cells. Human Molecular Genetics, 16(13), 1569–1577. https://doi.org/10.1093/hmg/ddm105

Gavelis, G. S., Hayakawa, S., White III, R. A., Gojobori, T., Suttle, C. A., Keeling, P. J., & Leander, B. S. (2015). Eye-like ocelloids are built from different endosymbiotically acquired components. Nature, 523(7559), Article 7559. https://doi.org/10.1038/nature14593

Gavelis, G. S., Herranz, M., Wakeman, K. C., Ripken, C., Mitarai, S., Gile, G. H., Keeling, P. J., & Leander, B. S. (2019). Dinoflagellate nucleus contains an extensive endomembrane network, the nuclear net. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-018-37065-w

Geuking, M. B., Weber, J., Dewannieux, M., Gorelik, E., Heidmann, T., Hengartner, H., Zinkernagel, R. M., & Hangartner, L. (2009). Recombination of Retrotransposon and Exogenous RNA Virus Results in Nonretroviral cDNA Integration. Science, 323(5912), 393–396. https://doi.org/10.1126/science.1167375

Gómez, F. (2005). A list of free-living dinoflagellate species in the world's oceans. Acta Botanica Croatica. https://www.semanticscholar.org/paper/A-list-of-free-living-dinoflagellate-species-in-the-G%C3%B3mez/b6170d2829ca2203ef0765376074c1a65e54b86a

González-Pech, R. A., Stephens, T. G., Chen, Y., Mohamed, A. R., Cheng, Y., Shah, S., Dougan, K. E., Fortuin, M. D. A., Lagorce, R., Burt, D. W., Bhattacharya, D., Ragan, M. A., & Chan, C. X. (2021a). Comparison of 15 dinoflagellate genomes reveals extensive sequence and structural divergence in family Symbiodiniaceae and genus Symbiodinium. BMC Biology, 19(1), 73. https://doi.org/10.1186/s12915-021-00994-6

González-Pech, R. A., Stephens, T. G., Chen, Y., Mohamed, A. R., Cheng, Y., Shah, S., Dougan, K. E., Fortuin, M. D. A., Lagorce, R., Burt, D. W., Bhattacharya, D., Ragan, M. A., & Chan, C. X. (2021b). Comparison of 15 dinoflagellate genomes reveals extensive sequence and structural divergence in family Symbiodiniaceae and genus Symbiodinium. BMC Biology, 19(1), 73. https://doi.org/10.1186/s12915-021-00994-6

Gordon, B. R., & Leggat, W. (2010). Symbiodinium—Invertebrate Symbioses and the Role of Metabolomics. Marine Drugs, 8(10), Article 10. https://doi.org/10.3390/md8102546

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. Proceedings of the National Academy of Sciences, 112(18), 5767–5772. https://doi.org/10.1073/pnas.1423400112

Gornik, S. G., Flores, V., Reinhardt, F., Erber, L., Salas-Leiva, D. E., Douvropoulou, O., Lassadi, I., Einarsson, E., Mörl, M., Git, A., Stadler, P. F., Pain, A., & Waller, R. F. (2022). Mitochondrial Genomes in Perkinsus Decode Conserved Frameshifts in All Genes. Molecular Biology and Evolution, 39(10), msac191. https://doi.org/10.1093/molbev/msac191

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. Current Biology, 22(24), 2303–2312. https://doi.org/10.1016/j.cub.2012.10.036

Gornik, S. G., Hu, I., Lassadi, I., & Waller, R. F. (2019). The Biochemistry and Evolution of the Dinoflagellate Nucleus. Microorganisms, 7(8), Article 8. https://doi.org/10.3390/microorganisms7080245

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology, 29(7), 644–652. https://doi.org/10.1038/nbt.1883

Gray, M. W., Burger, G., & Lang, B. F. (2001). The origin and early evolution of mitochondria. Genome Biology, 2(6), reviews1018.1. https://doi.org/10.1186/gb-2001-2-6-reviews1018

Gray, M. W., Lang, B. F., & Burger, G. (2004). Mitochondria of protists. Annual Review of Genetics, 38, 477–524. https://doi.org/10.1146/annurev.genet.37.110801.142526

Gray, M. W., Lang, B. F., Cedergren, R., Golding, G. B., Lemieux, C., Sankoff, D., Turmel, M., Brossard, N., Delage, E., Littlejohn, T. G., Plante, I., Rioux, P., Saint-Louis, D., Zhu, Y., & Burger, G. (1998). Genome structure and gene content in protist mitochondrial DNAs. Nucleic Acids Research, 26(4), 865–878.

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). Circlize Implements and enhances circular visualization in R. Bioinformatics (Oxford, England), 30(19), 2811–2812. https://doi.org/10.1093/bioinformatics/btu393

Guo, Z., Zhang, H., & Lin, S. (2014). Light-Promoted Rhodopsin Expression and Starvation Survival in the Marine Dinoflagellate Oxyrrhis marina. PLOS ONE, 9(12), e114941. https://doi.org/10.1371/journal.pone.0114941

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology, 9(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

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. https://doi.org/10.1186/1471-2164-6-80

Hackl, T., Duponchel, S., Barenhoff, K., Weinmann, A., & Fischer, M. G. (2021). Virophages and retrotransposons colonize the genomes of a heterotrophic flagellate. eLife, 10, e72674. https://doi.org/10.7554/eLife.72674

Hadjebi, O., Casas-Terradellas, E., Garcia-Gonzalo, F. R., & Rosa, J. L. (2008). The RCC1 superfamily: From genes, to function, to disease. Biochimica Et Biophysica Acta, 1783(8), 1467–1479. https://doi.org/10.1016/j.bbamcr.2008.03.015

Hastings, J. W. (1996). Chemistries and colors of bioluminescent reactions: A review. Gene, 173(1), 5–11. https://doi.org/10.1016/0378-1119(95)00676-1

Herzog, M., Soyer, M. O., & Daney de Marcillac, G. (1982). A high level of thymine replacement by 5-hydroxymethyluracil in nuclear DNA of the primitive dinoflagellate Prorocentrum micans E. European Journal of Cell Biology, 27(2), 151–155.

Hinnebusch, A. G., Klotz, L. C., Immergut, E., & Loeblich, A. R. (1980). Deoxyribonucleic acid sequence organization in the genome of the dinoflagellate Crypthecodinium cohnii. Biochemistry, 19(9), 1744–1755. https://doi.org/10.1021/bi00550a004

Holmes, E. C. (2011). The Evolution of Endogenous Viral Elements. Cell Host & Microbe, 10(4), 368–377. https://doi.org/10.1016/j.chom.2011.09.002

Hou, Y. (2008). Isolation and characterization of proliferating cell nuclear antigen and the small subunit ribosomal RNA genes in dinoflagellates: Insights into dinoflagellate genome evolution. Doctoral Dissertations, 1–148.

Hou, Y., Ji, N., Zhang, H., Shi, X., Han, H., & Lin, S. (2019). Genome size-dependent pcna gene copy number in dinoflagellates and molecular evidence of retroposition as a major evolutionary mechanism. Journal of Phycology, 55(1), 37–46. https://doi.org/10.1111/jpy.12815

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(9), e6978. https://doi.org/10.1371/journal.pone.0006978

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Molecular Biology and Evolution, 34(8), 2115–2122. https://doi.org/10.1093/molbev/msx148

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 11(1), 119. https://doi.org/10.1186/1471-2105-11-119

Imanian, B., Pombert, J.-F., Dorrell, R. G., Burki, F., & Keeling, P. J. (2012). Tertiary endosymbiosis in two dinotoms has generated little change in the mitochondrial genomes of their dinoflagellate hosts and diatom endosymbionts. PloS One, 7(8), e43763. https://doi.org/10.1371/journal.pone.0043763

Iranzo, J., Krupovic, M., & Koonin, E. V. (2016). The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. mBio, 7(4), 10.1128/mbio.00978-16. https://doi.org/10.1128/mbio.00978-16

Iyer, L. M., Abhiman, S., & Aravind, L. (2008). A new family of polymerases related to superfamily A DNA polymerases and T7-like DNA-dependent RNA polymerases. Biology Direct, 3, 39. https://doi.org/10.1186/1745-6150-3-39

Jackman, S. D., Coombe, L., Warren, R. L., Kirk, H., Trinh, E., MacLeod, T., Pleasance, S., Pandoh, P., Zhao, Y., Coope, R. J., Bousquet, J., Bohlmann, J., Jones, S. J. M., & Birol, I. (2020). Complete Mitochondrial Genome of a Gymnosperm, Sitka Spruce (Picea sitchensis), Indicates a Complex Physical Structure. Genome Biology and Evolution, 12(7), 1174–1179. https://doi.org/10.1093/gbe/evaa108

Jackson, C. J., Gornik, S. G., & Waller, R. F. (2012a). The mitochondrial genome and transcriptome of the basal dinoflagellate Hematodinium sp.: Character evolution within the highly derived mitochondrial genomes of dinoflagellates. Genome Biology and Evolution, 4(1), 59–72. https://doi.org/10.1093/gbe/evr122

Jackson, C. J., Gornik, S. G., & Waller, R. F. (2012b). The Mitochondrial Genome and Transcriptome of the Basal Dinoflagellate Hematodinium sp.: Character Evolution within the Highly Derived Mitochondrial Genomes of Dinoflagellates. Genome Biology and Evolution, 4(1), 59–72. https://doi.org/10.1093/gbe/evr122

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

Jackson, C. J., & Waller, R. F. (2013). A widespread and unusual RNA trans-splicing type in dinoflagellate mitochondria. PloS One, 8(2), e56777. https://doi.org/10.1371/journal.pone.0056777

Jaeckisch, N., Yang, I., Wohlrab, S., Glöckner, 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(12), e28012. https://doi.org/10.1371/journal.pone.0028012

Jagielski, T., Gawor, J., Bakuła, Z., Zuchniewicz, K., Żak, I., & Gromadka, R. (2017). An optimized method for high quality DNA extraction from microalga Prototheca wickerhamii for genome sequencing. Plant Methods, 13, 77. https://doi.org/10.1186/s13007-017-0228-9

Jąkalski, M., Takeshita, K., Deblieck, M., Koyanagi, K. O., Makałowska, I., Watanabe, H., & Makałowski, W. (2016). Comparative genomic analysis of retrogene repertoire in two green algae Volvox carteri and Chlamydomonas reinhardtii. Biology Direct, 11(1), 35. https://doi.org/10.1186/s13062-016-0138-1

Janouškovec, J., Gavelis, G. S., Burki, F., Dinh, D., Bachvaroff, T. R., Gornik, S. G., Bright, K. J., Imanian, B., Strom, S. L., Delwiche, C. F., Waller, R. F., Fensome, R. A., Leander, B. S., Rohwer, F. L., & Saldarriaga, J. F. (2017). Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. Proceedings of the National Academy of Sciences, 114(2), E171–E180. https://doi.org/10.1073/pnas.1614842114

Jeong, H. J., Yoo, Y. D., Kim, J. S., Seong, K. A., Kang, N. S., & Kim, T. H. (2010). Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs. Ocean Science Journal, 45(2), 65–91. https://doi.org/10.1007/s12601-010-0007-2

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C. W., Yi, T.-S., & Li, D.-Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biology, 21(1), 241. https://doi.org/10.1186/s13059-020-02154-5

John, U., Lu, Y., Wohlrab, S., Groth, M., Janouškovec, 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., & Glöckner, G. (2019). An aerobic eukaryotic parasite with functional mitochondria that likely lacks a mitochondrial genome. Science Advances, 5(4), eaav1110. https://doi.org/10.1126/sciadv.aav1110

Johnson, L. K., Alexander, H., & Brown, C. T. (2019). Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. GigaScience, 8(4). https://doi.org/10.1093/gigascience/giy158

Kaessmann, H., Vinckenbosch, N., & Long, M. (2009a). RNA-based gene duplication: Mechanistic and evolutionary insights. Nature Reviews Genetics, 10(1), 19–31. https://doi.org/10.1038/nrg2487

Kaessmann, H., Vinckenbosch, N., & Long, M. (2009b). RNA-based gene duplication: Mechanistic and evolutionary insights. Nature Reviews Genetics, 10(1), 19–31. https://doi.org/10.1038/nrg2487

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. Nature Methods, 14(6), Article 6. https://doi.org/10.1038/nmeth.4285

Kamikawa, R., Inagaki, Y., & Sako, Y. (2007). Fragmentation of Mitochondrial Large Subunit rRNA in the Dinoflagellate Alexandrium catenella and the Evolution of rRNA structure in Alveolate Mitochondria. Protist, 158(2), 239–245. https://doi.org/10.1016/j.protis.2006.12.002

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 28(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Kapitonov, V. V., & Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. Proceedings of the National Academy of Sciences, 103(12), 4540–4545. https://doi.org/10.1073/pnas.0600833103

Kato, K. H., Moriyama, A., Huitorel, P., Cosson, J., Cachon, M., & Sato, H. (1997). Isolation of the major basic nuclear protein and its localization on chromosomes of the dinoflagellate, Oxyrrhis marina. Biology of the Cell, 89(1), 43–52.

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010

Keeling, P. J. (2010). The endosymbiotic origin, diversification and fate of plastids. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 365(1541), 729–748. https://doi.org/10.1098/rstb.2009.0103

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., … Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. PLOS Biology, 12(6), e1001889. https://doi.org/10.1371/journal.pbio.1001889

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics, 9(8), 605–618. https://doi.org/10.1038/nrg2386

Kieft, K., Zhou, Z., & Anantharaman, K. (2020). VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome, 8(1), 90. https://doi.org/10.1186/s40168-020-00867-0

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. Nature Methods, 12(4), 357–360. https://doi.org/10.1038/nmeth.3317

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. Genome Research, 26(12), 1721–1729. https://doi.org/10.1101/gr.210641.116

Klawitter, S., Fuchs, N. V., Upton, K. R., Muñoz-Lopez, M., Shukla, R., Wang, J., Garcia-Cañadas, M., Lopez-Ruiz, C., Gerhardt, D. J., Sebe, A., Grabundzija, I., Merkert, S., Gerdes, P., Pulgarin, J. A., Bock, A., Held, U., Witthuhn, A., Haase, A., Sarkadi, B., … Schumann, G. G. (2016). Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. Nature Communications, 7(1), Article 1. https://doi.org/10.1038/ncomms10286

Kolev, N. G., Franklin, J. B., Carmi, S., Shi, H., Michaeli, S., & Tschudi, C. (2010). The Transcriptome of the Human Pathogen Trypanosoma brucei at Single-Nucleotide Resolution. PLoS Pathogens, 6(9), e1001090. https://doi.org/10.1371/journal.ppat.1001090

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology, 37(5), Article 5. https://doi.org/10.1038/s41587-019-0072-8

Koltover, I., Wagner, K., & Safinya, C. R. (2000). DNA condensation in two dimensions. Proceedings of the National Academy of Sciences, 97(26), 14046–14051. https://doi.org/10.1073/pnas.97.26.14046

Koonin, E. V., & Dolja, V. V. (2014). Virus World as an Evolutionary Network of Viruses and Capsidless Selfish Elements. Microbiology and Molecular Biology Reviews, 78(2), 278–303. https://doi.org/10.1128/mmbr.00049-13

Koonin, E. V., & Krupovic, M. (2017). Polintons, virophages and transpovirons: A tangled web linking viruses, transposons and immunity. Current Opinion in Virology, 25, 7–15. https://doi.org/10.1016/j.coviro.2017.06.008

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics, 5(1), 59. https://doi.org/10.1186/1471-2105-5-59

Krupovic, M., Bamford, D. H., & Koonin, E. V. (2014). Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. Biology Direct, 9, 6. https://doi.org/10.1186/1745-6150-9-6

Krupovic, M., & Koonin, E. V. (2015). Polintons: A hotbed of eukaryotic virus, transposon and plasmid evolution. Nature Reviews Microbiology, 13(2), Article 2. https://doi.org/10.1038/nrmicro3389

Krupovic, M., & Koonin, E. V. (2016). Self-synthesizing transposons: Unexpected key players in the evolution of viruses and defense systems. Current Opinion in Microbiology, 31, 25–33. https://doi.org/10.1016/j.mib.2016.01.006

Krupovic, M., Yutin, N., & Koonin, E. V. (2016). Fusion of a superfamily 1 helicase and an inactivated DNA polymerase is a signature of common evolutionary history of Polintons, polinton-like viruses, Tlr1 transposons and transpovirons. Virus Evolution, 2(1), vew019. https://doi.org/10.1093/ve/vew019

Kuhlisch, C., Schleyer, G., Shahaf, N., Vincent, F., Schatz, D., & Vardi, A. (2021). Viral infection of algal blooms leaves a unique metabolic footprint on the dissolved organic matter in the ocean. Science Advances, 7(25), eabf4680. https://doi.org/10.1126/sciadv.abf4680

Kuzmin, E., Taylor, J. S., & Boone, C. (2022). Retention of duplicated genes in evolution. Trends in Genetics, 38(1), 59–72. https://doi.org/10.1016/j.tig.2021.06.016

Ladoukakis, E. D., & Zouros, E. (2017). Evolution and inheritance of animal mitochondrial DNA: Rules and exceptions. Journal of Biological Research-Thessaloniki, 24(1), 2. https://doi.org/10.1186/s40709-017-0060-4

LaJeunesse, T. C., Lambert, G., Andersen, R. A., Coffroth, M. A., & Galbraith, D. W. (2005). SYMBIODINIUM (PYRRHOPHYTA) GENOME SIZES (DNA CONTENT) ARE SMALLEST AMONG DINOFLAGELLATES1. Journal of Phycology, 41(4), 880–886. https://doi.org/10.1111/j.0022-3646.2005.04231.x

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), Article 4. https://doi.org/10.1038/nmeth.1923

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

Le, Q. H., Markovic, P., Hastings, J. W., Jovine, R. V. M., & Morse, D. (1997). Structure and organization of the peridinin-chlorophyll a-binding protein gene in Gonyaulax polyedra. Molecular and General Genetics MGG, 255(6), 595–604. https://doi.org/10.1007/s004380050533

LeBlanc, A. J., Yermovsky-Kammerer, A. E., & Hajduk, S. L. (1999). A Nuclear Encoded and Mitochondrial Imported Dicistronic tRNA Precursor in Trypanosoma brucei *. Journal of Biological Chemistry, 274(30), 21071–21077. https://doi.org/10.1074/jbc.274.30.21071

Lee, R., Lai, H., Malik, S. B., Saldarriaga, J. F., Keeling, P. J., & Slamovits, C. H. (2014a). Analysis of EST data of the marine protist Oxyrrhis marina, an emerging model for alveolate biology and evolution. BMC Genomics, 15(1), 122. https://doi.org/10.1186/1471-2164-15-122

Lee, R., Lai, H., Malik, S., Saldarriaga, J. F., Keeling, P. J., & Slamovits, C. H. (2014b). Analysis of EST data of the marine protist Oxyrrhis marina, an emerging model for alveolate biology and evolution. BMC Genomics, 15(1), 122. https://doi.org/10.1186/1471-2164-15-122

Levin, R. A., Beltran, V. H., Hill, R., Kjelleberg, S., McDougald, D., Steinberg, P. D., & van Oppen, M. J. H. (2016). Sex, Scavengers, and Chaperones: Transcriptome Secrets of Divergent Symbiodinium Thermal Tolerances. Molecular Biology and Evolution, 33(9), 2201–2215. https://doi.org/10.1093/molbev/msw119

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12(1), 323. https://doi.org/10.1186/1471-2105-12-323

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

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(1), Article 1. https://doi.org/10.3390/microorganisms8010102

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lidie, K. B., & Dolah, F. M. V. (2007). Spliced Leader RNA-Mediated trans-Splicing in a Dinoflagellate, Karenia brevis. Journal of Eukaryotic Microbiology, 54(5), 427–435. https://doi.org/10.1111/j.1550-7408.2007.00282.x

Lin, S. (2011). Genomic understanding of dinoflagellates. Research in Microbiology, 162(6), 551–569. https://doi.org/10.1016/j.resmic.2011.04.006

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., … Morse, D. (2015). The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis. Science, 350(6261), 691–694. https://doi.org/10.1126/science.aad0408

Liu, H., Stephens, T. G., González-Pech, R. A., Beltran, V. H., Lapeyre, B., Bongaerts, P., Cooke, I., Aranda, M., Bourne, D. G., Forêt, 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. Communications Biology, 1(1). https://doi.org/10.1038/s42003-018-0098-3

Liu, L., & Hastings, J. W. (2006). Novel and Rapidly Diverging Intergenic Sequences Between Tandem Repeats of the Luciferase Genes in Seven Dinoflagellate Species1. Journal of Phycology, 42(1), 96–103. https://doi.org/10.1111/j.1529-8817.2006.00165.x

Livolant, F. (1978). Positive and negative birefringence in chromosomes. Chromosoma, 68(1), 45–58. https://doi.org/10.1007/bf00330371

Livolant, F., & Bouligand, Y. (1978). New observations on the twisted arrangement of Dinoflagellate chromosomes. Chromosoma, 68(1), 21–44. https://doi.org/10.1007/BF00330370

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G. P., Maumus, F., Munoz-Pomer, A., Sempere, J. M., Latorre, A., & Moya, A. (2011). The Gypsy Database (GyDB) of mobile genetic elements: Release 2.0. Nucleic Acids Research, 39(suppl_1), D70–D74. https://doi.org/10.1093/nar/gkq1061

Loeblich, A. R., III, Schmidt, R. J., & Sherley, J. L. (1981). Scanning electron microscopy of Heterocapsa pygmaea sp. Nov., and evidence for polyploidy as a speciation mechanism in dinoflagellates. Journal of Plankton Research, 3(1), 67–79. https://doi.org/10.1093/plankt/3.1.67

Long, M., Betrán, E., Thornton, K., & Wang, W. (2003). The origin of new genes: Glimpses from the young and old. Nature Reviews Genetics, 4(11), Article 11. https://doi.org/10.1038/nrg1204

Long, M., VanKuren, N. W., Chen, S., & Vibranovski, M. D. (2013). New Gene Evolution: Little Did We Know. Annual Review of Genetics, 47(1), 307–333. https://doi.org/10.1146/annurev-genet-111212-133301

Lowe, C. D., Keeling, P. J., Martin, L. E., Slamovits, C. H., Watts, P. C., & Montagnes, D. J. S. (2011). Who is Oxyrrhis marina? Morphological and phylogenetic studies on an unusual dinoflagellate. Journal of Plankton Research, 33(4), 555–567. https://doi.org/10.1093/plankt/fbq110

Lowe, C. D., Mello, L. V., Samatar, N., Martin, L. E., Montagnes, D. J., & Watts, P. C. (2011). The transcriptome of the novel dinoflagellate Oxyrrhis marina (Alveolata: Dinophyceae): response to salinity examined by 454 sequencing. BMC Genomics, 12(1), 519. https://doi.org/10.1186/1471-2164-12-519

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research, 25(5), 955–964. https://doi.org/10.1093/nar/25.5.955

Mahapatra, S., & Adhya, S. (1996). Import of RNA into Leishmania Mitochondria Occurs through Direct Interaction with Membrane-bound Receptors *. Journal of Biological Chemistry, 271(34), 20432–20437. https://doi.org/10.1074/jbc.271.34.20432

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. Bioinformatics (Oxford, England), 20(16), 2878–2879. https://doi.org/10.1093/bioinformatics/bth315

Makałowska, I., & Kubiak, M. R. (2023). Novel functions of a retroposed gene. Trends in Genetics, 39(6), 439–441. https://doi.org/10.1016/j.tig.2023.03.006

Makiuchi, T., & Nozaki, T. (2014). Highly divergent mitochondrion-related organelles in anaerobic parasitic protozoa. Biochimie, 100, 3–17. https://doi.org/10.1016/j.biochi.2013.11.018

Marburger, S., Alexandrou, M. A., Taggart, J. B., Creer, S., Carvalho, G., Oliveira, C., & Taylor, M. I. (2018). Whole genome duplication and transposable element proliferation drive genome expansion in Corydoradinae catfishes. Proceedings of the Royal Society B: Biological Sciences, 285(1872), 20172732. https://doi.org/10.1098/rspb.2017.2732

Marinov, G. K., Chen, X., Swaffer, M. P., Xiang, T., Grossman, A. R., & Greenleaf, W. J. (2023). Genome-wide distribution of 5-hydroxymethyluracil and chromatin accessibility in the Breviolum minutum genome (p. 2023.09.18.558303). bioRxiv. https://doi.org/10.1101/2023.09.18.558303

Marinov, G. K., & Lynch, M. (2015). Diversity and Divergence of Dinoflagellate Histone Proteins. G3: Genes|Genomes|Genetics, 6(2), 397–422. https://doi.org/10.1534/g3.115.023275

Marinov, G. K., Trevino, A. E., Xiang, T., Kundaje, A., Grossman, A. R., & Greenleaf, W. J. (2021). Transcription-dependent domain-scale three-dimensional genome organization in the dinoflagellate Breviolum minutum. Nature Genetics, 53(5), Article 5. https://doi.org/10.1038/s41588-021-00848-5

Martí, J. M. (2019). Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. PLoS Computational Biology, 15(4), e1006967. https://doi.org/10.1371/journal.pcbi.1006967

Maumus, F., Allen, A. E., Mhiri, C., Hu, H., Jabbari, K., Vardi, A., Grandbastien, M.-A., & Bowler, C. (2009). Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. BMC Genomics, 10(1), 624. https://doi.org/10.1186/1471-2164-10-624

McCarrey, J. R., & Thomas, K. (1987). Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature, 326(6112), Article 6112. https://doi.org/10.1038/326501a0

McEWAN, M., Humayun, R., Slamovits, C. H., & Keeling, P. J. (2008a). Nuclear Genome Sequence Survey of the Dinoflagellate Heterocapsa triquetra. Journal of Eukaryotic Microbiology, 55(6), 530–535. https://doi.org/10.1111/j.1550-7408.2008.00357.x

McEWAN, M., Humayun, R., Slamovits, C. H., & Keeling, P. J. (2008b). Nuclear Genome Sequence Survey of the Dinoflagellate Heterocapsa triquetra. Journal of Eukaryotic Microbiology, 55(6), 530–535. https://doi.org/10.1111/j.1550-7408.2008.00357.x

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. The Journal of Eukaryotic Microbiology, 62(5), 679–687. https://doi.org/10.1111/jeu.12230

Menghini, D., & Aubry, S. (2021). De novo transcriptome assembly data of the marine bioluminescent dinoflagellate Pyrocystis lunula. Data in Brief, 37, 107254. https://doi.org/10.1016/j.dib.2021.107254

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., & Thomas, P. D. (2017). PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Research, 45(D1), D183–D189. https://doi.org/10.1093/nar/gkw1138

Mighell, A. j., Smith, N. r., Robinson, P. a., & Markham, A. f. (2000). Vertebrate pseudogenes. FEBS Letters, 468(2–3), 109–114. https://doi.org/10.1016/S0014-5793(00)01199-6

Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A., & Aylward, F. O. (2020). Widespread endogenization of giant viruses shapes genomes of green algae. Nature, 588(7836), Article 7836. https://doi.org/10.1038/s41586-020-2924-2

Montagnes, D. J. S., Lowe, C. D., Martin, L., Watts, P. C., Downes-Tettmar, N., Yang, Z., Roberts, E. C., & Davidson, K. (2011). Oxyrrhis marina growth, sex and reproduction. Journal of Plankton Research, 33(4), 615–627. https://doi.org/10.1093/plankt/fbq111

Montagnes, D. J. S., Lowe, C. D., Roberts, E. C., Breckels, M. N., Boakes, D. E., Davidson, K., Keeling, P. J., Slamovits, C. H., Steinke, M., Yang, Z., & Watts, P. C. (2011). An introduction to the special issue: Oxyrrhis marina, a model organism? Journal of Plankton Research, 33(4), 549–554. https://doi.org/10.1093/plankt/fbq121

Mönttinen, H. A. M., Bicep, C., Williams, T. A., & Hirt, R. P. (2021). The genomes of nucleocytoplasmic large DNA viruses: Viral evolution writ large. Microbial Genomics, 7(9), 000649. https://doi.org/10.1099/mgen.0.000649

Moreno Díaz de la Espina, S., Alverca, E., Cuadrado, A., & Franca, S. (2005). Organization of the genome and gene expression in a nuclear environment lacking histones and nucleosomes: The amazing dinoflagellates. European Journal of Cell Biology, 84(2), 137–149. https://doi.org/10.1016/j.ejcb.2005.01.002

Morris, J. C., Drew, M. E., Klingbeil, M. M., Motyka, S. A., Saxowsky, T. T., Wang, Z., & Englund, P. T. (2001). Replication of kinetoplast DNA: An update for the new millennium. International Journal for Parasitology, 31(5), 453–458. https://doi.org/10.1016/S0020-7519(01)00156-4

Mougari, S., Chelkha, N., Sahmi-Bounsiar, D., Di Pinto, F., Colson, P., Abrahao, J., & La Scola, B. (2020). A virophage cross-species infection through mutant selection represses giant virus propagation, promoting host cell survival. Communications Biology, 3, 248. https://doi.org/10.1038/s42003-020-0970-9

Muller-Parker, G., D'Elia, C. F., & Cook, C. B. (2015). Interactions Between Corals and Their Symbiotic Algae. In C. Birkeland (Ed.), Coral Reefs in the Anthropocene (pp. 99–116). Springer Netherlands. https://doi.org/10.1007/978-94-017-7249-5_5

Nand, A., Zhan, Y., Salazar, O. R., Aranda, M., Voolstra, C. R., & Dekker, J. (2020). Chromosome-scale assembly of the coral endosymbiont Symbiodinium microadriaticum genome provides insight into the unique biology of dinoflagellate chromosomes. bioRxiv, 2020.07.01.182477. https://doi.org/10.1101/2020.07.01.182477

Nand, A., Zhan, Y., Salazar, O. R., Aranda, M., Voolstra, C. R., & Dekker, J. (2021). Genetic and spatial organization of the unusual chromosomes of the dinoflagellate Symbiodinium microadriaticum. Nature Genetics, 53(5), Article 5. https://doi.org/10.1038/s41588-021-00841-y

Nash, E. A., Barbrook, A. C., Edwards-Stuart, R. K., Bernhardt, K., Howe, C. J., & Nisbet, R. E. R. (2007). Organization of the mitochondrial genome in the dinoflagellate Amphidinium carterae. Molecular Biology and Evolution, 24(7), 1528–1536. https://doi.org/10.1093/molbev/msm074

Nash, E. A., Nisbet, R. E. R., Barbrook, A. C., & Howe, C. J. (2008). Dinoflagellates: A mitochondrial genome all at sea. Trends in Genetics: TIG, 24(7), 328–335. https://doi.org/10.1016/j.tig.2008.04.001

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution, 32(1), 268–274. https://doi.org/10.1093/molbev/msu300

Norman, J. E., & Gray, M. W. (1997). The cytochrome oxidase subunit 1 gene (cox1) from the dinoflagellate, Crypthecodinium cohnii. FEBS Letters, 413(2), 333–338. https://doi.org/10.1016/S0014-5793(97)00938-1

Norman, J. E., & Gray, M. W. (2001a). A Complex Organization of the Gene Encoding Cytochrome Oxidase Subunit 1 in the Mitochondrial Genome of the Dinoflagellate, Crypthecodinium cohnii: Homologous Recombination Generates Two Different cox1 Open Reading Frames. Journal of Molecular Evolution, 53(4), 351–363. https://doi.org/10.1007/s002390010225

Norman, J. E., & Gray, M. W. (2001b). A Complex Organization of the Gene Encoding Cytochrome Oxidase Subunit 1 in the Mitochondrial Genome of the Dinoflagellate, Crypthecodinium cohnii: Homologous Recombination Generates Two Different cox1 Open Reading Frames. Journal of Molecular Evolution, 53(4), 351–363. https://doi.org/10.1007/s002390010225

Ochman, H., Lawrence, J. G., & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature, 405(6784), Article 6784. https://doi.org/10.1038/35012500

Ohno, S. (1970). Evolution by gene duplication.

Okamoto, O. K., & Hastings, J. W. (2003). Genome-wide analysis of redox-regulated genes in a dinoflagellate. Gene, 321, 73–81.

Ou, S., & Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiology, 176(2), 1410–1422. https://doi.org/10.1104/pp.17.01310

Ou, S., Liu, J., Chougule, K. M., Fungtammasan, A., Seetharam, A. S., Stein, J. C., Llaca, V., Manchanda, N., Gilbert, A. M., Wei, S., Chin, C.-S., Hufnagel, D. E., Pedersen, S., Snodgrass, S. J., Fengler, K., Woodhouse, M., Walenz, B. P., Koren, S., Phillippy, A. M., … Ware, D. (2020). Effect of sequence depth and length in long-read assembly of the maize inbred NC358. Nature Communications, 11(1), Article 1. https://doi.org/10.1038/s41467-020-16037-7

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N., & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biology, 20(1), 275. https://doi.org/10.1186/s13059-019-1905-y

Panchy, N., Lehti-Shiu, M., & Shiu, S.-H. (2016). Evolution of Gene Duplication in Plants. Plant Physiology, 171(4), 2294–2316. https://doi.org/10.1104/pp.16.00523

Parkinson, J. E., Baumgarten, S., Michell, C. T., Baums, I. B., LaJeunesse, T. C., & Voolstra, C. R. (2016). Gene Expression Variation Resolves Species and Individual Strains among Coral-Associated Dinoflagellates within the Genus Symbiodinium. Genome Biology and Evolution, 8(3), 665–680. https://doi.org/10.1093/gbe/evw019

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics, 23(9), 1061–1067. https://doi.org/10.1093/bioinformatics/btm071

Patterson, D. J., Larsen, J., & Systematics Association (Eds.). (1991). The Biology of free-living heterotrophic flagellates. Published for the Systematics Association by Clarendon Press ; Oxford University Press.

Pavlicek, A., Gentles, A. J., Pačes, J., Pačes, V., & Jurka, J. (2006). Retroposition of processed pseudogenes: The impact of RNA stability and translational control. Trends in Genetics : TIG, 22(2), 69–73. https://doi.org/10.1016/j.tig.2005.11.005

Penaud, A., Hardy, W., Lambert, C., Marret, F., Masure, E., Servais, T., Siano, R., Wary, M., & Mertens, K. N. (2018). Dinoflagellate fossils: Geological and biological applications. Revue de Micropaléontologie, 61(3), 235–254. https://doi.org/10.1016/j.revmic.2018.09.003

Petrov, D. A. (2001). Evolution of genome size: New approaches to an old problem. Trends in Genetics, 17(1), 23–28. https://doi.org/10.1016/S0168-9525(00)02157-0

Ponmani, T., Guo, R., & Ki, J.-S. (2016). Analysis of the genomic DNA of the harmful dinoflagellate Prorocentrum minimum: A brief survey focused on the noncoding RNA gene sequences. Journal of Applied Phycology, 28(1), 335–344. https://doi.org/10.1007/s10811-015-0570-0

Potrzebowski, L., Vinckenbosch, N., Marques, A. C., Chalmel, F., Jégou, B., & Kaessmann, H. (2008). Chromosomal Gene Movements Reflect the Recent Origin and Biology of Therian Sex Chromosomes. PLOS Biology, 6(4), e80. https://doi.org/10.1371/journal.pbio.0060080

Pritham, E. J., Putliwala, T., & Feschotte, C. (2007). Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene, 390(1–2), 3–17. https://doi.org/10.1016/j.gene.2006.08.008

Putintseva, Y. A., Bondar, E. I., Simonov, E. P., Sharov, V. V., Oreshkova, N. V., Kuzmin, D. A., Konstantinov, Y. M., Shmakov, V. N., Belkov, V. I., Sadovsky, M. G., Keech, O., & Krutovsky, K. V. (2020). Siberian larch (Larix sibirica Ledeb.) mitochondrial genome assembled using both short and long nucleotide sequence reads is currently the largest known mitogenome. BMC Genomics, 21(1), 654. https://doi.org/10.1186/s12864-020-07061-4

Qiu, C., Jin, H., Vvedenskaya, I., Llenas, J. A., Zhao, T., Malik, I., Visbisky, A. M., Schwartz, S. L., Cui, P., Čabart, P., Han, K. H., Lai, W. K. M., Metz, R. P., Johnson, C. D., Sze, S.-H., Pugh, B. F., Nickels, B. E., & Kaplan, C. D. (2020). Universal promoter scanning by Pol II during transcription initiation in Saccharomyces cerevisiae. Genome Biology, 21(1), 132. https://doi.org/10.1186/s13059-020-02040-0

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England), 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rae, P. M. M. (1973). 5-Hydroxymethyluracil in the DNA of a Dinoflagellate. Proceedings of the National Academy of Sciences, 70(4), 1141–1145. https://doi.org/10.1073/pnas.70.4.1141

Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., & Finn, R. D. (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Research, 46(D1), D624–D632. https://doi.org/10.1093/nar/gkx1134

Riaz, S., Sui, Z., Niaz, Z., Khan, S., Liu, Y., & Liu, H. (2018). Distinctive Nuclear Features of Dinoflagellates with A Particular Focus on Histone and Histone-Replacement Proteins. Microorganisms, 6(4). https://doi.org/10.3390/microorganisms6040128

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics, 16(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Riding, J. B., Fensome, R. A., Soyer-Gobillard, M.-O., & Medlin, L. K. (2023). A Review of the Dinoflagellates and Their Evolution from Fossils to Modern. Journal of Marine Science and Engineering, 11(1), Article 1. https://doi.org/10.3390/jmse11010001

Rizzo, P. J. (1991). The Enigma of the Dinoflagellate Chromosome. The Journal of Protozoology, 38(3), 246–252. https://doi.org/10.1111/j.1550-7408.1991.tb04437.x

Rizzo, P. J. (2003). Those amazing dinoflagellate chromosomes. Cell Research, 13(4), 215–217. https://doi.org/10.1038/sj.cr.7290166

Rizzo, P., & Nooden, L. D. (1973). Isolation and chemical composition of dinoflagellate nuclei. The Journal of Protozoology, 20(5), 666–672.

Roberts, E. C., Wootton, E. C., Davidson, K., Jeong, H. J., Lowe, C. D., & Montagnes, D. J. S. (2011). Feeding in the dinoflagellate Oxyrrhis marina: Linking behaviour with mechanisms. Journal of Plankton Research, 33(4), 603–614. https://doi.org/10.1093/plankt/fbq118

Roger, A. J., Muñoz-Gómez, S. A., & Kamikawa, R. (2017a). The Origin and Diversification of Mitochondria. Current Biology, 27(21), R1177–R1192. https://doi.org/10.1016/j.cub.2017.09.015

Roger, A. J., Muñoz-Gómez, S. A., & Kamikawa, R. (2017b). The Origin and Diversification of Mitochondria. Current Biology, 27(21), R1177–R1192. https://doi.org/10.1016/j.cub.2017.09.015

Roitman, S., Rozenberg, A., Lavy, T., Brussaard, C. P. D., Kleifeld, O., & Béjà, O. (2023). Isolation and infection cycle of a polinton-like virus virophage in an abundant marine alga. Nature Microbiology, 8(2), 332–346. https://doi.org/10.1038/s41564-022-01305-7

Rosso, L., Marques, A. C., Weier, M., Lambert, N., Lambot, M.-A., Vanderhaeghen, P., & Kaessmann, H. (2008). Birth and Rapid Subcellular Adaptation of a Hominoid-Specific CDC14 Protein. PLOS Biology, 6(6), e140. https://doi.org/10.1371/journal.pbio.0060140

Roy, S., Jagus, R., & Morse, D. (2018). Translation and Translational Control in Dinoflagellates. Microorganisms, 6(2), 30. https://doi.org/10.3390/microorganisms6020030

Roy, S., & Morse, D. (2012). A full suite of histone and histone modifying genes are transcribed in the dinoflagellate Lingulodinium. PLoS One, 7(4).

Sala-Rovira, M., Geraud, M. L., Caput, D., Jacques, F., Soyer-Gobillard, M. O., Vernet, G., & Herzog, M. (1991). Molecular cloning and immunolocalization of two variants of the major basic nuclear protein (HCc) from the histone-less eukaryote Crypthecodinium cohnii (Pyrrhophyta). Chromosoma, 100(8), 510–518.

Sano, J., & Kato, K. H. (2009). Localization and Copy Number of the Protein-Coding Genes Actin, α-Tubulin, and HSP90 in the Nucleus of a Primitive Dinoflagellate, Oxyrrhis marina. Zoological Science, 26(11), 745–753. https://doi.org/10.2108/zsj.26.745

Santini, S., Jeudy, S., Bartoli, J., Poirot, O., Lescot, M., Abergel, C., Barbe, V., Wommack, K. E., Noordeloos, A. A. M., Brussaard, C. P. D., & Claverie, J.-M. (2013). Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. Proceedings of the National Academy of Sciences, 110(26), 10800–10805. https://doi.org/10.1073/pnas.1303251110

Sarai, C., Tanifuji, G., Nakayama, T., Kamikawa, R., Takahashi, K., Yazaki, E., Matsuo, E., Miyashita, H., Ishida, K., Iwataki, M., & Inagaki, Y. (2020). Dinoflagellates with relic endosymbiont nuclei as models for elucidating organellogenesis. Proceedings of the National Academy of Sciences, 117(10), 5364–5375. https://doi.org/10.1073/pnas.1911884117

Sasidharan, R., & Gerstein, M. (2008). Genomics: Protein fossils live on as RNA. Nature, 453(7196), 729–731. https://doi.org/10.1038/453729a

Schmedes, S. E., Patel, D., Kelley, J., Udhayakumar, V., & Talundzic, E. (2019). Using the Plasmodium mitochondrial genome for classifying mixed-species infections and inferring the geographical origin of P. falciparum parasites imported to the U.S. PLoS ONE, 14(4), e0215754. https://doi.org/10.1371/journal.pone.0215754

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nature Methods, 9(7), Article 7. https://doi.org/10.1038/nmeth.2089

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE, 11(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Shoguchi, E., Beedessee, G., Hisata, K., Tada, I., Narisoko, H., Satoh, N., Kawachi, M., & Shinzato, C. (2021). A New Dinoflagellate Genome Illuminates a Conserved Gene Cluster Involved in Sunscreen Biosynthesis. Genome Biology and Evolution, 13(evaa235). https://doi.org/10.1093/gbe/evaa235

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(1), 458. https://doi.org/10.1186/s12864-018-4857-9

Shoguchi, E., Shinzato, C., Hisata, K., Satoh, N., & Mungpakdee, S. (2015). The Large Mitochondrial Genome of Symbiodinium minutum Reveals Conserved Noncoding Sequences between Dinoflagellates and Apicomplexans. Genome Biology and Evolution, 7(8), 2237–2244. https://doi.org/10.1093/gbe/evv137

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., … Satoh, N. (2013a). Draft Assembly of the Symbiodinium minutum Nuclear Genome Reveals Dinoflagellate Gene Structure. Current Biology, 23(15), 1399–1408. https://doi.org/10.1016/j.cub.2013.05.062

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., … Satoh, N. (2013b). Draft Assembly of the Symbiodinium minutum Nuclear Genome Reveals Dinoflagellate Gene Structure. Current Biology, 23(15), 1399–1408. https://doi.org/10.1016/j.cub.2013.05.062

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015a). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015b). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Slamovits, C. H., & Keeling, P. J. (2008a). Plastid-derived genes in the nonphotosynthetic alveolate Oxyrrhis marina. Molecular Biology and Evolution, 25(7), 1297–1306. https://doi.org/10.1093/molbev/msn075

Slamovits, C. H., & Keeling, P. J. (2008b). Widespread recycling of processed cDNAs in dinoflagellates. Current Biology, 18(13), R550–R552. https://doi.org/10.1016/j.cub.2008.04.054

Slamovits, C. H., & Keeling, P. J. (2011). Contributions of Oxyrrhis marina to molecular biology, genomics and organelle evolution of dinoflagellates. Journal of Plankton Research, 33(4), 591–602. https://doi.org/10.1093/plankt/fbq153

Slamovits, C. H., Okamoto, N., Burri, L., James, E. R., & Keeling, P. J. (2011). A bacterial proteorhodopsin proton pump in marine eukaryotes. Nature Communications, 2(1). https://doi.org/10.1038/ncomms1188

Slamovits, C. H., Saldarriaga, J. F., Larocque, A., & Keeling, P. J. (2007a). The Highly Reduced and Fragmented Mitochondrial Genome of the Early-branching Dinoflagellate Oxyrrhis marina Shares Characteristics with both Apicomplexan and Dinoflagellate Mitochondrial Genomes. Journal of Molecular Biology, 372(2), 356–368. https://doi.org/10.1016/j.jmb.2007.06.085

Slamovits, C. H., Saldarriaga, J. F., Larocque, A., & Keeling, P. J. (2007b). The Highly Reduced and Fragmented Mitochondrial Genome of the Early-branching Dinoflagellate Oxyrrhis marina Shares Characteristics with both Apicomplexan and Dinoflagellate Mitochondrial Genomes. Journal of Molecular Biology, 372(2), 356–368. https://doi.org/10.1016/j.jmb.2007.06.085

Song, B., Chen, S., & Chen, W. (2018). Dinoflagellates, a Unique Lineage for Retrogene Research. Frontiers in Microbiology, 9. https://doi.org/10.3389/fmicb.2018.01556

Song, B., Morse, D., Song, Y., Fu, Y., Lin, X., Wang, W., Cheng, S., Chen, W., Liu, X., & Lin, S. (2017a). Comparative Genomics Reveals Two Major Bouts of Gene Retroposition Coinciding with Crucial Periods of Symbiodinium Evolution. Genome Biology and Evolution, 9(8), 2037–2047. https://doi.org/10.1093/gbe/evx144

Song, B., Morse, D., Song, Y., Fu, Y., Lin, X., Wang, W., Cheng, S., Chen, W., Liu, X., & Lin, S. (2017b). Comparative Genomics Reveals Two Major Bouts of Gene Retroposition Coinciding with Crucial Periods of Symbiodinium Evolution. Genome Biology and Evolution, 9(8), 2037–2047. https://doi.org/10.1093/gbe/evx144

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature, 521(7551), 173–179. https://doi.org/10.1038/nature14447

Spector, D. L. (1984). Dinoflagellates. Academic Press.

Stairs, C. W., Leger, M. M., & Roger, A. J. (2015). Diversity and origins of anaerobic metabolism in mitochondria and related organelles. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1678), 20140326. https://doi.org/10.1098/rstb.2014.0326

Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research, 33(Web Server issue), W465-467. https://doi.org/10.1093/nar/gki458

Staszak, K., & Makałowska, I. (2021). Cancer, Retrogenes, and Evolution. Life, 11(1), Article 1. https://doi.org/10.3390/life11010072

Steele, R. E., & Rae, P. M. M. (1980). Ordered distribution of modified bases in the DNA of a dinoflagellate. Nucleic Acids Research, 8(20), 4709–4726. https://doi.org/10.1093/nar/8.20.4709

Stephens, T. G., González-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 Biology, 18(1), 56. https://doi.org/10.1186/s12915-020-00782-8

Stephens, T. G., Ragan, M. A., Bhattacharya, D., & Chan, C. X. (2018). Core genes in diverse dinoflagellate lineages include a wealth of conserved dark genes with unknown functions. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-35620-z

Su, W., Gu, X., & Peterson, T. (2019). TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. Molecular Plant, 12(3), 447–460. https://doi.org/10.1016/j.molp.2019.02.008

Su, W., Ou, S., Hufford, M. B., & Peterson, T. (2021). A Tutorial of EDTA: Extensive De Novo TE Annotator. Methods in Molecular Biology (Clifton, N.J.), 2250, 55–67. https://doi.org/10.1007/978-1-0716-1134-0_4

Swapna, L. S., & Parkinson, J. (2017). Genomics of Apicomplexan Parasites. Critical Reviews in Biochemistry and Molecular Biology, 52(3), 254–273. https://doi.org/10.1080/10409238.2017.1290043

Taylor, F. J. R., Hoppenrath, M., & Saldarriaga, J. F. (2008). Dinoflagellate diversity and distribution. Biodiversity and Conservation, 17(2), 407–418. https://doi.org/10.1007/s10531-007-9258-3

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., & Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Research, 18(12), 1979–1990. https://doi.org/10.1101/gr.081612.108

Testa, A. C., Hane, J. K., Ellwood, S. R., & Oliver, R. P. (2015). CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics, 16(1), 170. https://doi.org/10.1186/s12864-015-1344-4

Thurber, R. L. V., & Correa, A. M. S. (2011). Viruses of reef-building scleractinian corals. Journal of Experimental Marine Biology and Ecology, 408(1), 102–113. https://doi.org/10.1016/j.jembe.2011.07.030

Tomaru, Y., Mizumoto, H., & Nagasaki, K. (2009). Virus resistance in the toxic bloom-forming dinoflagellate Heterocapsa circularisquama to single-stranded RNA virus infection. Environmental Microbiology, 11(11), 2915–2923. https://doi.org/10.1111/j.1462-2920.2009.02047.x

Tyagi, S., Pande, V., & Das, A. (2014). Whole Mitochondrial Genome Sequence of an Indian Plasmodium falciparum Field Isolate. The Korean Journal of Parasitology, 52(1), 99–103. https://doi.org/10.3347/kjp.2014.52.1.99

Vaidya, A. B., & Mather, M. W. (2009). Mitochondrial evolution and functions in malaria parasites. Annual Review of Microbiology, 63, 249–267. https://doi.org/10.1146/annurev.micro.091208.073424

Van Vlierberghe, M., Di Franco, A., Philippe, H., & Baurain, D. (2021). Decontamination, pooling and dereplication of the 678 samples of the Marine Microbial Eukaryote Transcriptome Sequencing Project. BMC Research Notes, 14(1), 306. https://doi.org/10.1186/s13104-021-05717-2

Vinckenbosch, N., Dupanloup, I., & Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. Proceedings of the National Academy of Sciences, 103(9), 3220–3225. https://doi.org/10.1073/pnas.0511307103

Voleman, L., & Doležal, P. (2019). Mitochondrial dynamics in parasitic protists. PLOS Pathogens, 15(11), e1008008. https://doi.org/10.1371/journal.ppat.1008008

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE, 9(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Waller, R. F., & Jackson, C. J. (2009a). Dinoflagellate mitochondrial genomes: Stretching the rules of molecular biology. BioEssays, 31(2), 237–245. https://doi.org/10.1002/bies.200800164

Waller, R. F., & Jackson, C. J. (2009b). Dinoflagellate mitochondrial genomes: Stretching the rules of molecular biology. BioEssays, 31(2), 237–245. https://doi.org/10.1002/bies.200800164

Waller, R. F., & Kořený, L. (2017). Chapter Four - Plastid Complexity in Dinoflagellates: A Picture of Gains, Losses, Replacements and Revisions. In Y. Hirakawa (Ed.), Advances in Botanical Research (Vol. 84, pp. 105–143). Academic Press. https://doi.org/10.1016/bs.abr.2017.06.004

Waller, R. F., & McFadden, G. I. (2005). The apicoplast: A review of the derived plastid of apicomplexan parasites. Current Issues in Molecular Biology, 7(1), 57–79.

Wang, J., Li, L., & Lin, S. (2023). Active viral infection during blooms of a dinoflagellate indicates dinoflagellate-viral co-adaptation. Applied and Environmental Microbiology, 0(0), e01156-23. https://doi.org/10.1128/aem.01156-23

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & Paterson, A. H. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Research, 40(7), e49. https://doi.org/10.1093/nar/gkr1293

Wang, Z., Chen, Y., & Li, Y. (2004). A Brief Review of Computational Gene Prediction Methods. Genomics, Proteomics & Bioinformatics, 2(4), 216–221. https://doi.org/10.1016/S1672-0229(04)02028-5

Watts, P. C., Martin, L. E., Kimmance, S. A., Montagnes, D. J. S., & Lowe, C. D. (2011). The distribution of Oxyrrhis marina: A global disperser or poorly characterized endemic? Journal of Plankton Research, 33(4), 579–589. https://doi.org/10.1093/plankt/fbq148

Wells, J. N., & Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. Annual Review of Genetics, 54, 539–561. https://doi.org/10.1146/annurev-genet-040620-022145

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biology, 20(1), 129. https://doi.org/10.1186/s13059-019-1727-y

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer.

Williams, E., Place, A., & Bachvaroff, T. (2017). Transcriptome Analysis of Core Dinoflagellates Reveals a Universal Bias towards "GC" Rich Codons. Marine Drugs, 15(5), Article 5. https://doi.org/10.3390/md15050125

Wilson, S. J., Webb, B. L. J., Ylinen, L. M. J., Verschoor, E., Heeney, J. L., & Towers, G. J. (2008). Independent evolution of an antiviral TRIMCyp in rhesus macaques. Proceedings of the National Academy of Sciences, 105(9), 3557–3562. https://doi.org/10.1073/pnas.0709003105

Wisecaver, J. H., & Hackett, J. D. (2011). Dinoflagellate Genome Evolution. Annual Review of Microbiology, 65(1), 369–387. https://doi.org/10.1146/annurev-micro-090110-102841

Wong, J. T. Y. (2019). Architectural Organization of Dinoflagellate Liquid Crystalline Chromosomes. Microorganisms, 7(2), 27. https://doi.org/10.3390/microorganisms7020027

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021a). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation, 2(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021b). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation, 2(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Xia, B., Zhang, W., Zhao, G., Zhang, X., Bai, J., Brosh, R., Wudzinska, A., Huang, E., Ashe, H., Ellis, G., Pour, M., Zhao, Y., Coelho, C., Zhu, Y., Miller, A., Dasen, J. S., Maurano, M. T., Kim, S. Y., Boeke, J. D., & Yanai, I. (2024). On the genetic basis of tail-loss evolution in humans and apes. Nature, 626(8001), 1042–1048. https://doi.org/10.1038/s41586-024-07095-8

Xiong, W., He, L., Lai, J., Dooner, H. K., & Du, C. (2014). HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proceedings of the National Academy of Sciences, 111(28), 10263–10268. https://doi.org/10.1073/pnas.1410068111

Xu, Z., & Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research, 35(Web Server issue), W265–W268. https://doi.org/10.1093/nar/gkm286

Yang, Z., Jeong, H. J., & Montagnes, D. J. S. (2011). The role of Oxyrrhis marina as a model prey: Current work and future directions. Journal of Plankton Research, 33(4), 665–675. https://doi.org/10.1093/plankt/fbq112

Yau, S., Lauro, F. M., DeMaere, M. Z., Brown, M. V., Thomas, T., Raftery, M. J., Andrews-Pfannkoch, C., Lewis, M., Hoffman, J. M., Gibson, J. A., & Cavicchioli, R. (2011). Virophage control of antarctic algal host–virus dynamics. Proceedings of the National Academy of Sciences, 108(15), 6163–6168. https://doi.org/10.1073/pnas.1018221108

Yoon, H. S., Hackett, J. D., Van Dolah, F. M., Nosenko, T., Lidie, K. L., & Bhattacharya, D. (2005). Tertiary Endosymbiosis Driven Genome Evolution in Dinoflagellate Algae. Molecular Biology and Evolution, 22(5), 1299–1308. https://doi.org/10.1093/molbev/msi118

Yutin, N., Kapitonov, V. V., & Koonin, E. V. (2015). A new family of hybrid virophages from an animal gut metagenome. Biology Direct, 10(1), 19. https://doi.org/10.1186/s13062-015-0054-9

Yutin, N., Raoult, D., & Koonin, E. V. (2013). Virophages, polintons, and transpovirons: A complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. Virology Journal, 10(1), 158. https://doi.org/10.1186/1743-422X-10-158

Yutin, N., Shevchenko, S., Kapitonov, V., Krupovic, M., & Koonin, E. V. (2015). A novel group of diverse Polinton-like viruses discovered by metagenome analysis. BMC Biology, 13(1), 95. https://doi.org/10.1186/s12915-015-0207-4

Zaheri, B., & Morse, D. (2022a). An overview of transcription in dinoflagellates. Gene, 829, 146505. https://doi.org/10.1016/j.gene.2022.146505

Zaheri, B., & Morse, D. (2022b). An overview of transcription in dinoflagellates. Gene, 829, 146505. https://doi.org/10.1016/j.gene.2022.146505

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K. W., Anantharaman, K., Starnawski, P., Kjeldsen, K. U., Stott, M. B., Nunoura, T., Banfield, J. F., Schramm, A., Baker, B. J., Spang, A., & Ettema, T. J. G. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature, 541(7637), Article 7637. https://doi.org/10.1038/nature21031

Zemskov, E. A., Kang, W., & Maeda, S. (2000). Evidence for Nucleic Acid Binding Ability and Nucleosome Association of Bombyx mori Nucleopolyhedrovirus BRO Proteins. Journal of Virology, 74(15), 6784–6789. https://doi.org/10.1128/jvi.74.15.6784-6789.2000

Zhang, H., Bhattacharya, D., & Lin, S. (2007). A Three-Gene Dinoflagellate Phylogeny Suggests Monophyly of Prorocentrales and a Basal Position for Amphidinium and Heterocapsa. Journal of Molecular Evolution, 65(4), 463–474. https://doi.org/10.1007/s00239-007-9038-4

Zhang, H., Hou, Y., Miranda, L., Campbell, D. A., Sturm, N. R., Gaasterland, T., & Lin, S. (2007). Spliced leader RNA trans-splicing in dinoflagellates. Proceedings of the National Academy of Sciences, 104(11), 4618–4623. https://doi.org/10.1073/pnas.0700258104

Zhang, Z., Harrison, P. M., Liu, Y., & Gerstein, M. (2003). Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. Genome Research, 13(12), 2541–2558. https://doi.org/10.1101/gr.1429003

Zheng, D., & Gerstein, M. B. (2007). The ambiguous boundary between genes and pseudogenes: The dead rise up, or do they? Trends in Genetics: TIG, 23(5), 219–224. https://doi.org/10.1016/j.tig.2007.03.003

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A. N., & Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of Molecular Biology, 430(15), 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007