1-22-2021 2:00 PM

Sequencing and Assembling the Nuclear Genome of the Antarctic Psychrophilic Green Alga Chlamydomonas sp. UWO241: Unravelling the Evolution of Cold Adaptation

Xi Zhang, *The University of Western Ontario*

Supervisor: Smith, David R., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Biology

Abstract

DNA sequencing technologies have undergone tremendous advancements in recent years, but assembling, annotating, and analyzing a nuclear genome is still a huge undertaking, especially for small laboratory groups, partly because many eukaryotic genomes are repeat-rich and contain thousands of genes and introns. The Antarctic harbors a variety of algae that can withstand extreme cold but do not grow at warmer temperatures (psychrophiles), including the unicellular green alga *Chlamydomonas* sp. UWO241 (a.k.a. UWO241). Little is known, however, about how psychrophilic algae evolved from their respective mesophilic ancestors by adapting to particular cold environments. To present insights into this issue, I critically determined the draft nuclear genome (~212 Mb, 16,325 protein-coding genes) sequence of UWO241 and performed comparative genomic analyses. Firstly, an assembly pipeline was developed for processing high throughput sequencing (DNA-Seq) reads into genomic contigs. These contigs, alongside transcriptome sequencing (RNA-Seq) reads, were fed into an annotation pipeline, containing the commonly used bioinformatics gene-profiling software. Computational analyses were carried out on a powerful in-house computer. Finally, comparative genomic analyses were performed between UWO241 and its close green algal relatives in the Chlamydomonadales revealing: (1) UWO241 harbors hundreds of highly similar duplicate genes involved in diverse cellular processes, some of which I argue are aiding its survival in the Antarctic via gene dosage; (2) UWO241 encodes a large number (≥37) of ice-binding proteins (IBPs), putatively originating from horizontal gene transfer; and (3) UWO241 appears to have an expanded set of orthologous gene families for reverse transcriptase, IBPs and antenna proteins. These investigations deepen our understanding of evolution between psychrophilic and mesophilic algae and help unravel the existence of common mechanisms in the adaptation to cold environments.

Keywords

Next-generation sequencing, green algae, gene duplication, *Chlamydomonas* sp. UWO241, psychrophile, *Chlamydomonas* sp. ICE-L, ice-binding proteins, genome evolution, cold adaptation

Summary for Lay Audience

Most of the Earth exists at or below the freezing point of water. Such extreme environments can harbour a variety of organisms, including psychrophiles, which can withstand intense cold and cannot survive at more moderate temperatures. Lake Bonney is a permanently ice-covered lake in the McMurdo Dry Valleys of Antarctica. It is home to many cold-adapted microbes including the unicellular green alga *Chlamydomonas* sp. UWO241 (a.k.a. UWO241). Several important aspects of its biology, including physiology, molecular biology of photosynthesis and comparative genomics, have been studied in detail over the past 25 years. Here, the draft genome of UWO241 was determined, including a highly contiguous genome assembly and well-annotated coding regions. Furthermore, a comparative genomic framework between UWO241 and its close green algal relatives as well as other cold-adapted algae was built. Remarkably, UWO241 is unique in many ways. For example, the genome has large size of noncoding regions. On the other hand, many genes are duplicated and some gene families encoding important functions even contain more genes in the UWO241 genome than other green algal relatives, such as antenna protein genes, ribosome genes, and ice-binding protein genes. These features deepen our understanding of evolution between psychrophilic and mesophilic algae and help unravel the existence of common mechanisms in the adaptation to cold environments.

Co-Authorship Statement

Chapter 3 and 4 were adapted from a previous manuscript, which was submitted to the journal of iScience in 2020 with Xi Zhang (XZ), Marina Cvetkovska (MC), Rachael Morgan-Kiss (RMK), Norman P. A. Hüner (NPAH) and David Roy Smith (DRS). The study was conceptualized by MC, DRS and NPAH. The data were analyzed by MC and XZ. DRS and XZ drafted the manuscript. RMK and NPAH provided editorial comments. DRS and NPAH provided funding and assisted with the revisions of the manuscript. All authors commented to produce the manuscript for peer review. The Cell Press authorized to include the article in part in this thesis for non-commercial purpose.

The introduction of Chapter 4 was adapted in part from a manuscript of web tool (HSDFinder) to be submitted in 2021 by Xi Zhang (XZ), Yining Hu (YH) and David Roy Smith (DRS). The study was conceptualized by XZ and DRS. The data were analyzed by XZ. YNH implemented the HSDFinder website. XZ and DRS drafted the manuscript and all authors commented to produce the manuscript for peer review.

As the first author (XZ) of the two manuscripts, I significantly contributed to most aspects of the work including study design, data collection and analysis as well as manuscript preparation and submission. The Chapter 3 and 4 were conceived of, performed by and written by me with the following exception: Dr. Marina Cvetkovska isolated the UWO241 sample and collaborated with the nuclear genome sequencing and transcriptome datasets.

Acknowledgments

To my supervisor, Dr. David Roy Smith, I owe an immense debt of gratitude for his support, mentorship and scientific insights during my studies. I could still remember the words he encouraged me “What often differentiates a good scientist from a great scientist is being able to find the story from the data and tell that story in a compelling way.” Thank you, David, you have my heartful thanks. I thank my advisory committee members, Norman Peter Andrew Hüner and Kathleen Allen Hill, for their encouragement and suggestions. In particular, thanks to Norm, who instructed me with the words that “The great scientist is to know the boundary between the things you know, and the things you don’t know”. I also want to address my thanks to the attendance of my Ph.D. proposal assessment assessors: Ryan Austin, Vera Tai and Kathleen Allen Hill, and Ph.D. comprehensive exam assessors: Marc-André Lachance, Denis Maxwell and Jim Karagiannis. Thanks for the suggestions from Zaichao Zhang and Jingpu Song in the discussion of thesis chapters. I thank Marina Cvetkovska for her valuable advice and I enjoyed working with her for three years. And specifically, Yining Hu is my life partner and provides me with great support for computer programming. Thanks for her unconditional love and endless encouragement.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ABC transporters	ATP-Binding-Cassette transporters
BLAST	Basic Local Alignment Search Tool
BUSCO	Benchmarking Universal Single-Copy Orthologs
CBW	Canadian Bioinformatics Workshop
CDS	Coding Sequence
DBG	De Bruijn Graph
DPOR	Light-independent Protochlorophyllide Reductase
Fd	Ferredoxin protein
GO	Gene Ontology
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
HSDs	Highly Similar Duplicates
HSP	Heat Shock Protein
IBPs	Ice-Binding Proteins
ICE-L	*Chlamydomonas* sp. ICE-L
KEGG	Kyoto Encyclopedia of Genes and Genomes
LHCII	Light Harvesting Antenna of PSII
LINEs	Long Interspersed Nuclear Elements
LPOR	Light-dependent Protochlorophyllide Oxidoreductase
NA	Not Applicable
NCBI	National Center for Biotechnology
NGS	Next Generation Sequencing
PE	Paired-End
QC	Quality Control
SINEs	Short Interspersed Nuclear Elements
TE	Transposable Elements
TGS	Third Generation Sequencing
tRNA	transfer RNA
UWO241	*Chlamydomonas* sp. UWO241
WGS	Whole Genome Sequencing

List of Appendices

Chapter 1

1 General Introduction

Next-generation sequencing (NGS) and third-generation sequencing (TGS) technologies have made it easy to obtain huge amounts of raw high-throughput DNA and RNA sequencing data (DNA-Seq and RNA-Seq) from green algae, but downstream nuclear genomic analyses, such as genome assembly, gene annotation, and comparative genomic analysis, remain time-consuming, and complicated, especially for smaller laboratory groups with limited computing infrastructure. A variety of algae from Antarctica can tolerate cold but do not grow at warmer temper (psychrophiles), including the unicellular green alga *Chlamydomonas* sp. UWO241 (a.k.a. UWO241). But how psychrophilic algae evolved from their respective mesophilic ancestors by adapting to particular cold environments is little known. UWO241 as a psychrophile is emerging as model to explore this issue after years' research on several important aspects, including its physiology, molecular biology of photosynthesis, and comparative genomics. In Chapter 1, I present the background information on sequencing technologies and green algal genomics, with a particular focus on cold-adapted algae as well as members of the Chlamydomonadales. Chapter 2 is a step-by-step user guide offering researchers a basic foundation in bioinformatics for nuclear genome projects. In Chapter 3, I use various bioinformatics software and pipelines to assemble and annotate the nuclear genome of the green algae UWO241 using DNA-Seq and RNA-Seq. The assembled nuclear DNA contigs, alongside transcriptomic data, are fed into a customized annotation pipeline based on the most up-to-date eukaryotic bioinformatics gene-profiling software. In chapter 4, I carry out comparative genomic analysis of UWO241 with its close green algal relatives and other cold-adapted algae, with the ultimate goal of better understanding psychrophily. In short, the draft genome of UWO241 from the Chapter 3 is compared to the genomes of the mesophilic green algae *Chlamydomonas reinhardtii*, *Volvox carteri*, and *Dunaliella salina*, among others, allowing me to interpret some of the crucial hallmarks of the UWO241 genome. I argue that highly conserved duplicate genes are associated with environmental survival in an extreme environment. I show that UWO241 has acquired numerous ice-binding proteins (IBPs) via horizontal gene transfer (HGT), and that gene families for

reverse transcriptases, ribosomal proteins, and antenna proteins are noticeably expanded as compared to its mesophilic algal relatives. Presumably, the existence of common mechanisms underlying cold adaptation can be attributed to these unique genomic features. Lastly, in Chapter 5, I briefly review the challenges and opportunities for bioinformatics researchers.

1.1 Life at the Edge: Psychrophiles and Photopsychrophiles

1.1.1 Psychrophiles

The Earth is a cold place (~80% of it is permanently below 5 °C) (Russell 1990). This is largely due to 70% of the Earth surface is formed by oceans, among which 90% of water is at 5 °C or lower (Golomb 1993). Although the remaining 30% Earth surface is for land, 10% of which is covered with glacial ice, containing ice caps, glaciers, and the ice sheets of polar regions (Kwok *et al.* 2020). These polar regions are not limited to Antarctica, but parts of North America and Europe that are within the Arctic circle (Anisimov *et al.* 2001). Apart from that, mountainous regions such as Alps, Himalayas and Rocky Mountains are also contribute to the cold environment of earth's surface (Margesin and Miteva 2011).

Despite the frigid conditions, these cold realms are teeming with life—microbial life. In 2006, D'Amico *et al.* reviewed the availability of several bacteria and archaea genomes living in the cold (D'Amico *et al.* 2006). Two years later, researchers have discovered novel groups of cyanobacteria, fungi, and viruses adapted to cold (Margesin *et al.* 2008). In 2010, Horikoshi and colleagues (2010) published the extremophiles handbook with a special focus in Chapter 6 reviewing those microorganisms living in the cold. Despite the excitement around these discoveries, many questions remained about the role of these organisms in the cold environment, such as the definition of the cold-adapted species.

Cold adapted organisms are termed as psychrophiles or psychrotrophs (psychrotolerant) (Morita 1975). In 1975, Morita (Morita 1975) first defined the terms psychrophiles meaning organisms are able to grow optimally at temperatures lower than 15 °C and cannot tolerate above 20 °C, which differs from psychrotrophs, organisms that are capable of growth at temperatures lower than 15 °C, but also are able to grow and survive at

temperatures above 25 °C. Psychrophilic and psychrotolerant representatives can be found in all three domains of life, including Bacteria, Archaea, and Eukaryota. Bacteria are usually abundant in frozen environments such as lake ice, glaciers, and sea ice; however, archaea can survive many permanently cold environments such as polar marine water and deep oceans, which are thought to be associated with extremophiles (Mikucki *et al.* 2011; Siddiqui *et al.* 2013). Notably, an extremophile is an organism that can survive in extreme environmental conditions with optimal growth (e.g., high salt concentration, high pressure, low temperature etc.), which differs from a typical carbon-based life form using water as a solvent to survive (Rothschild and Mancinelli 2001). Also, many psychrophiles have evolved strategies to withstand stresses apart from coldness, such as low or high light, excessive ultraviolet (UV) radiation, high or low pH, high osmotic pressure and low nutrients (Rodrigues and Tiedje 2008)

In recent years, the application of metagenomics and associated meta-functional approaches (metaproteomics and metatranscriptomics) has deepened the insights into the molecular mechanisms of cold adaptation (De Maayer *et al.* 2014; Lyon and Mock 2014; Åqvist *et al.* 2017). There are a wide range of strategies for coping with extreme cold, including maintaining functional cold-adapted enzymes (Åqvist *et al.* 2017), having high levels of polyunsaturated fatty acids (PUFAs), which can increase cellular membrane fluidity in a cold climate (Becker *et al.* 2011; Lyon and Mock 2014), as well as employing a mixotrophic lifestyle, forming cysts, storing carbohydrates, and/or altering the photosynthetic machinery (Lyon and Mock 2014).

There is diversity of microbial life harboured in various cold environments such as lake, sea-ice and deep-sea, representing a broad range of physicochemical conditions in those areas (D'Amico *et al.* 2006). Especially, many of them are psychrophiles whose enzymes are sensitive to temperature change (De Maayer *et al.* 2014). The percentage of Arctic sea ice has experienced a dramatic decline linked to global warming in recent years (Obbard *et al.* 2014; Yadav *et al.* 2020). Thus, the structural and functional assessment of cold-active enzymes of psychrophilic species may provide insights into the impacts of climate change on the microbes present in these extreme environments (Siddiqui *et al.* 2013).

1.1.2 Photosynthetic Psychrophiles

Our understanding of psychrophily is largely shaped by studies on bacteria and archaea, and the field as a whole is still in its infancy (D'Amico *et al.* 2006; De Maayer *et al.* 2014), especially regarding the comparative genomics of psychrophiles with those from closely related mesophiles which grow best in moderate temperature ranging from 20 to 40 °C (Mock *et al.* 2017; Zhang *et al.* 2020). Several psychrophilic bacteria and archaea have been compared in recent publications. For example, the comparison of *Alteromonas* sp. SN2 genome with its two close mesophilic strains suggested the presence of 15 genomic islands in strain SN2 likely confer ecological fitness traits (especially membrane transport and fatty acid biosynthesis) (Math *et al.* 2012). Moreover, the halophilic archaeon *Halorubrum lacusprofundi*, isolated from Antarctica, was compared to 12 mesophilic Haloarchaea, indicating the type of amino acid substitutions are consistent with structural flexibility and protein function at low temperature (DasSarma *et al.* 2013).

Furthermore, the psychrophilic and psychrotrophic fungi have also brought great attention in recent years. This is partly due to their potential applications in biotechnology and pharmaceuticals (Yadav *et al.* 2019). As reviewed by Hassan *et al.*, fungi are found to be able to survive at low temperature as well as some of the extreme environments in polar regions, such as high UV, frequent freeze and thaw cycles and low nutrient availability (Hassan *et al.* 2016). Many adaptation features of fungi are identified to tolerate extreme environments, such as the production of bioactive metabolites and cold-active enzymes (Robinson 2001).

One of the coldest regions on Earth is the Antarctic (Feller and Gerday 2003). One consequence of the apparent inhospitable environment is the lack of the diversity of endemic terrestrial plant species found on the Antarctic continent. The Antarctic is limited to two angiosperms: *Deschampsia antarctica* and *Colobanthus quitensis* (Santiago *et al.* 2017). In contrast, Antarctica is teeming with microbial life despite its frigid conditions. This includes heterotrophic microbial diversity and an abundance localized to endolithic rock surfaces on ice-free Antarctic land surfaces (Coleine *et al.* 2020). Eukaryotic algae and cyanobacteria are the dominant microbial aquatic life forms in the Antarctic (Morgan-Kiss *et al.* 2016). Many of these algae are bona fide psychrophiles, among the best studied

psychrophilic eukaryotic algae is the diatom *Fragilariopsis cylindrus* (Appendix A: Table S1). Usually, the seawater and sea ice of both the Arctic and Antarctic Oceans harbor this diatom, which tends to have higher silicate concentrations (Mock and Hoch 2005). Recently, Mock and colleagues (2017) reported the complete sequenced and assembled genome of this species; furthermore, this team provided particular insights into this cold-adapted diatom from the Southern Ocean, discovering that the diploid *F. cylindrus* genome harboured around 24.7% genetic loci with highly divergent alleles, suggesting that divergent alleles might be involved in adaptation to environmental fluctuations.

The psychrotolerant green alga *Coccomyxa subellipsoidea* C-169 also has a complete genome sequence, which was completed and assembled earlier in 2012 (Blanc *et al.* 2012). It was the first psychrotolerant green alga from a polar environment to be sequenced. Several gene families in *C. subellipsoidae*, such as those involved in lipid metabolism, transporters, cellulose synthases and short alcohol dehydrogenases (Blanc *et al.* 2012). Notably, psychrotolerant species are the same as mesophiles to grow at 20-40 °C but are able to tolerate lower temperatures, albeit with slower growth rates. This psychrotolerant green alga does not meet the strict definition of psychrophily, but it strengthens the knowledge of the adaptations associated with low temperature.

1.1.3 Psychrophilic Chlamydomonadaleans

Psychrophilic chlamydomonadalean algae can be found in diverse (and sometimes strange) environments. Remarkably, almost one-third of known photopsychrophiles (i.e., photosynthetic psychrophiles) belong to the green algal order Chlamydomonadales , which is found in the Chlorophycean class of the Chlorophyta (Cvetkovska *et al.* 2017). Moreover, many chlamydomonadalean algae inhabiting polar and alpine environments are drought resistant, and can tolerate high levels of UV radiation and low-nutrient stresses (Quesada and Vincent 2012; Umen and Olson 2012), making them ideal models for studying not only psychrophily but adaptations to extreme environments in general. Some species that can withstand freezing have been intensively studied, including *Chlamydomonas nivalis* (Brown *et al.* 2015)*, Chlamydomonas* sp. UWO241 (Morgan-Kiss *et al.* 2006; Cvetkovska *et al.* 2018; Cvetkovska *et al.* 2019), and *Chlamydomonas* sp. ICE-L (Zhang *et al.* 2020).

Unfortunately, little is known about how psychrophilic algae evolved from their respective mesophilic ancestors by adapting to particular cold environments. However, it is presumed that the present origin of the species distribution in Antarctica and the Southern Ocean is a consequence of the separation of Antarctica, Australia, Africa and South America from the former supercontinent, Gondwana, approximately 165M years ago (Florindo and Siegert 2008; Wright *et al.* 2020). The subsequent climate change from warm, 'greenhouse conditions' to an 'ice-house condition' 46M years ago established the permanent, cold Antarctic environment about 35M years ago (Séranne 1999; Montañez and Poulsen 2013). Some interesting hypotheses are available. *Chlamydomonas nivalis*, for example, can thrive in alpine and polar snowfields (Remias *et al.* 2010) where it can withstand high light levels (up to 5000 μmol photons m^{-2} s^{-1}), intense UV radiation, and, of course, low temperatures (Williams *et al.* 2003). Moreover, *C. nivalis* is commonly identified in mixtures of bacteria and fungi and these large assemblages of bacteria and fungi can develop symbiotic or parasitic relationships with the alga (Brown *et al.* 2015). It is tempting to link psychrophily to the unique lifestyles of this snow alga. Researchers further recognized that *C. nivalis* can enter into a dormant diploid zygotic stage, withstand winter freezing and later switch back to an active stage during the warmer summer season, suggesting a potential strategy for surviving harsh environments (Remias *et al.* 2010).

Coincidentally, researchers are also unravelling the mechanisms by which psychrophiles survive such harsh environments as floating ice. *Chlamydomonas* sp. ICE-L (a.k.a. ICE-L) was first isolated from floating marine ice near the Antarctic coast (Liu *et al.* 2006). Researchers discovered that the primary nitrogen metabolism of ICE-L is consistent with light exposure. This might be a strategy for adapting to continuous light in summer and sustained darkness in winter (Wang *et al.* 2015). Recently, Zhang *et al.* presented the genome of ICE-L that provided evidence of its adaptation to its extreme Antarctic environment via expanded repertoire of genes for diverse metabolic processes and genes gained by horizontal gene transfer (Zhang *et al.* 2020).

Psychrophilic green algae from the Chlamydomonadales are of our particular interest. This order harbors not only some of the best-studied cold-adapted algae to date, such as UWO241, *C. nivalis*, and ICE-L, but also a myriad of model mesophilic species. These

species include but are not limited to *C. reinhardtii*, *V. carteri*, *G. pectorale* and *D. salina*. *C. reinhardtii*, for example, is an excellent comparison target for the investigations of psychrophilic chlamydomonads (Cvetkovska *et al.* 2017), not only because its immense volume of scientific literature regarding the molecular biology and physiology of photosynthesis, but also the availability of genomic data that can be used to perform several critical comparisons with psychrophilic genomes, such as the comparisons of gene family expansion and contraction, pathway loss and gain, the comparison of substitution rates at synonymous and nonsynonymous sites of protein-coding genes (i.e., calculating dN/dS).

There are a wide range of strategies for coping with extreme cold, including having high levels of polyunsaturated fatty acids (PUFAs), which can increase cellular membrane fluidity in a cold climate (Becker *et al.* 2011; Lyon and Mock 2014), as well as employing a mixotrophic lifestyle, forming cysts, storing carbohydrates, and/or altering the photosynthetic machinery (Lyon and Mock 2014). Thus, it appears that the next logical step is to explore the lineage-specific genes and gene families via the genome content. Unfortunately, there are very few complete nuclear genome sequences from psychrophilic green algae. Having such sequence data could greatly improve our understanding of cold-adaptation, and extremophily in general. In contrast, there are a wide range of genome sequences from mesophilic green algae, including various chlamydomonadaleans (Appendix A: Table S2). The first completely sequenced green algal nuclear genome was that of the prasinophyte *Ostreococcus tauri*—a feat carried out by the whole genome shotgun Sanger sequencing and aided by the extremely small genome size of *O. tauri* (12.5 Mb) (Derelle *et al.* 2006). Soon thereafter, scientists began decoding (again, using a Sanger-based approach) much larger nuclear DNAs (nucDNAs) from green algae, including those of the unicellular chlamydomonadalean *C. reinhardtii* (~120 Mb) (Merchant *et al.* 2007) and its close multicellular relative *V. carteri* (138 Mb) (Prochnik *et al.* 2010). More recently, researchers have used NGS to assemble entire nuclear DNA (nucDNA) from chlamydomonadaleans, such as the ~150 Mb nuclear genome of the colonial green alga *G. pectorale* (Hanschen *et al.* 2016) and the recently completed genome of the acidophile *C. eustigma* (~130 Mb) (Hirooka *et al.* 2017). These various genomic data sets have provided important insights into green algal evolution, including the origins of multicellularity (Prochnik *et al.* 2010; Hanschen *et al.* 2016). Thus, it is my hope to gain

a better understanding of cold adaptation by decoding the nucDNA from the psychrophilic green alga UWO241.

1.1.4 *Chlamydomonas* sp. UWO241

Figure 1: The geography of year-round lake ice in McMurdo Dry Valleys.

(A) Site of isolation of UWO241, Lake Bonney in Taylor Valley. (B) and (C) Research work in Lake Bonney and Lake Fryxell, Antarctica. Pictured: Luke Winslow (University of Wisconsin, Madison), Kyle Cronin and Dr. Peter Doran (University of Illinois, Chicago). (D) Don Jon Pond (the saltiest body of water on Earth) in Wright Valley, Antarctica. The images are reworked with the credit from original author Hilary Dugan and the images source can be found via EGUblogs (https://blogs.egu.eu/network/geosphere/2014/01/13/).

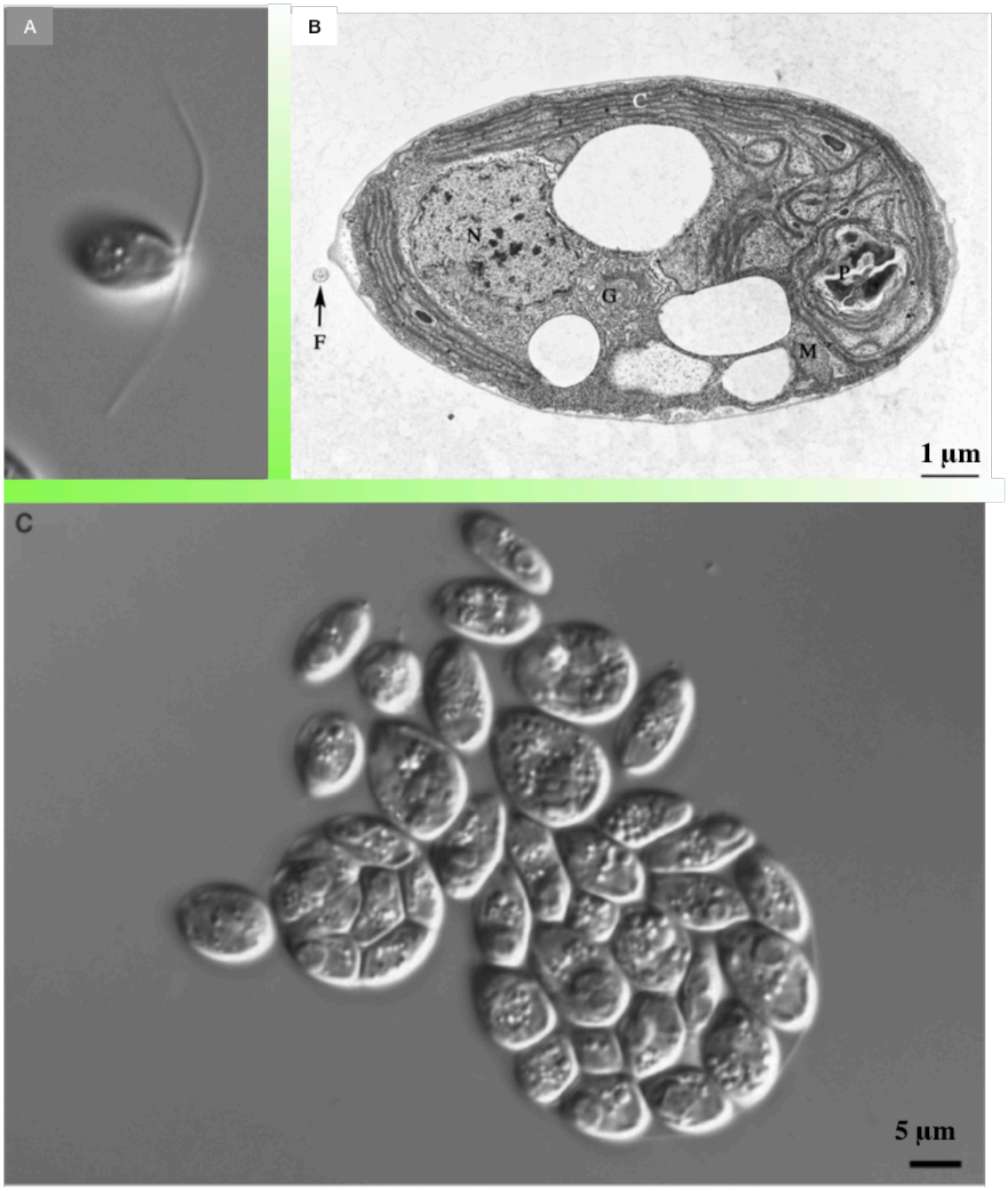


Figure 2: The UWO241 images under light microscope and electron microscope.

(A) The single cell of UWO241. (B) The electron micrographs of UWO241 grown under laboratory-controlled conditions (8 °C/20 µmol photons m^{-2} s^{-1}). The letters in the image indicating different organelles (C: chloroplast; N: nucleus; G: golgi apparatus; M: mitochondrion; F: flagellum). (C) A colony of UWO241 was observed ruptured under the light microscope. Reproduced from (Pocock *et al.* 2004) with permission.

The psychrophile *Chlamydomonas* sp. UWO241, which was isolated 17 m below the bottom of the permanent ice surface of Lake Bonney (Figure 1) in the McMurdo Dry Valleys of Victoria Land, Antarctica (Neale and Priscu 1995), is emerging as a model for studying cold-adaptation. This unicellular biflagellate is surprisingly resilient, persisting in an environment that not only is a perpetually cold environment but also has a high saline content (700 mM) and low irradiance transitions (Morgan-Kiss *et al.* 2006). UWO241 possesses an unusual photosynthetic apparatus (working best at 8 °C), but it presents rates of photosynthesis relative to those of *C. reinhardtii* at 25–35 °C (Cvetkovska *et al.* 2017) (Figure 2). In addition to withstanding constant low temperatures at approximately 5 °C year round, UWO241 is exposed to perpetual shading (5 μmol photons m^{-2} s^{-1} during midday in summer) and seasonal extremes in photoperiod (e.g., 24 h of light during the peak summer), which are represented by the blue-green spectrum (450–550 nm) (Dolhi *et al.* 2013). Lake Bonney is also phosphorus limited and contains high levels of dissolved oxygen (200% saturation) and high salinity (0.7 M) (Bowman *et al.* 2016). In UWO241, many unique cellular and physiological features have evolved to handle with these extreme conditions of Lake Bonney, such as high PSI cyclic electron transport, the inability to grow under red light and a lack of state transitions which balance the energy distribution between photosystem I (PSI) and photosystem II (PSII) (Morgan-Kiss *et al.* 2006).

Over the past 25 years, several important aspects of UWO241, including its physiology and molecular biology of photosynthesis, have been studied in detail (Morgan-Kiss *et al.* 2006; Cvetkovska *et al.* 2017). Previous findings have already indicated that UWO241 has two near-identical copies of the ferredoxin gene and accumulates large amounts of functional ferredoxin protein maintaining high activity and increased structural flexibility at low temperature, suggesting an adaptation to cold environments (Cvetkovska *et al.* 2018). In the same year, Possmayer and his colleagues (2018) investigated UWO241 transcriptome data, revealing the absence of upregulation of genes encoding heat-shock proteins (HSPs) under high growth temperature stress and heat shock. One year after, Cvetkovska *et al.* reported a functional chlorophyll biosynthesis pathway lacking of light-independent protochlorophyllide oxidoreductase (DPOR) in UWO241, and this pathway is solely dependent on light-dependent protochlorophyllide oxidoreductase (LPOR) for the

enzymatic reduction of protochlorophyllide (Cvetkovska *et al.* 2019). Recently, UWO241 was discovered to express comparable levels of the Stt7 protein kinase to *C. reinhardtii* (Stt7 protein kinase is important in the regulation of energy distribution between PSII and PSI through state transitions), but exhibited a distinct low temperature-dependent phosphorylation pattern in the absence of a classical state transition (Szyszka-Mroz *et al.* 2019).

Given all the previous assessments, UWO241, a psychrophilic alga, has been widely explored and has generated particular interest in photosynthetic adaptation associated with cold adaptation. However, its naming has experienced a complex journey. Initially, it was identified as *Chlamydomonas subcaudata* via cell morphology (Neale and Priscu 1995); however, sequencing patterns suggest that it is more likely to be a psychrophilic strain of mesophilic *Chlamydomonas raudensis* (Pocock *et al.* 2004). Until recently, studies have considered it to be a unique lineage within the Moewusinia clade of the Chlamydomonadales but not a strain of *Chlamydomonas raudensis* (Possmayer *et al.* 2016) (Figure 3). In addition to UWO241, some of the psychrophiles have been highlighted in the phylogeny (Figure 3), such as ICE-L and *Chlamydomonas* sp. ICE-MDV both fit in Monadinia clade, and *C. nivalis* occupies the Chloromonadinia clade; however, not all of these psychrophiles have been completely sequenced. Fortunately, there are many mesophilic algal species in the order Chlamydomonadales. For instance, *C. reinhardtii* is in the Reinhardtinia clade and is an excellent comparison target for the investigations of psychrophilic chlamydomonads. In addition, various model green algal genomes are available in *V. carteri*, *G. pectorale*, and *D. salina* across the clades of Reinhardtinia and Dunaliellinia (Figure 3).

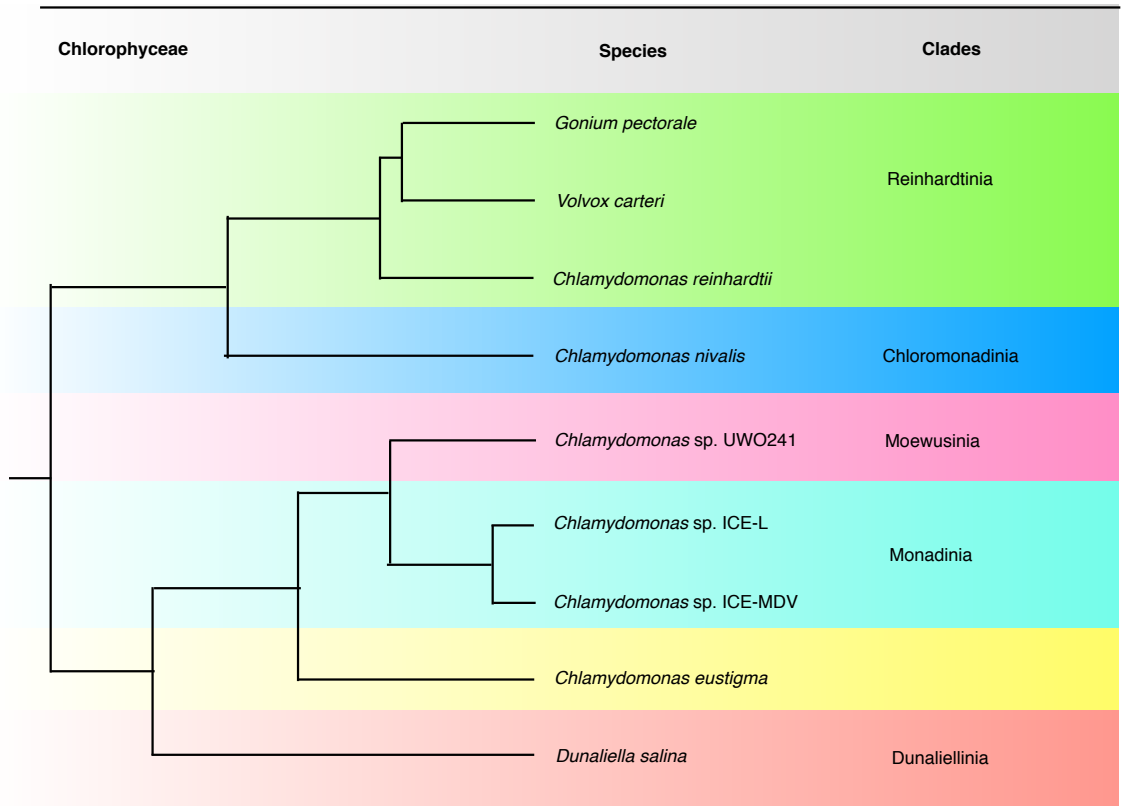


Figure 3: Phylogenetic relationship of the green algae in the order Chlamydomonadales.

Representative psychrophiles (blue, pink, cyan-blue) and other known model green algae (green, red, yellow) are highlighted in different colors. Adapted from (Zhang *et al.* 2020) with permission.

1.2 A Brief History of DNA Sequencing

DNA sequencing technologies have advanced at an impressive rate over the past 40 years (Liu *et al.* 2012) (Figure 4). In the late 1970s, two different "first-generation" DNA-Seq technologies were developed: Maxam-Gilbert sequencing and Sanger sequencing (Sanger and Coulson 1975; Maxam and Gilbert 1977). It was the latter, however, that was adopted by most researchers, due to its relatively high efficiency and low radioactivity. The early forms of Sanger sequencing were labor intensive, but in the late 1980s and early 1990s, a multitude of innovations in reagents and instruments were developed to support high-throughput Sanger sequencing, which in turn spurred the initiation, and the eventual

completion, of the Human Genome Project (Lander *et al.* 2001; Venter *et al.* 2001). Today, automated Sanger sequencing and the associated bioinformatics software have been widely applied to the genomes from diverse species throughout the tree of life, spurring the massive field of comparative genomics (Mardis 2008; Pop and Salzberg 2008; Schuster 2008).

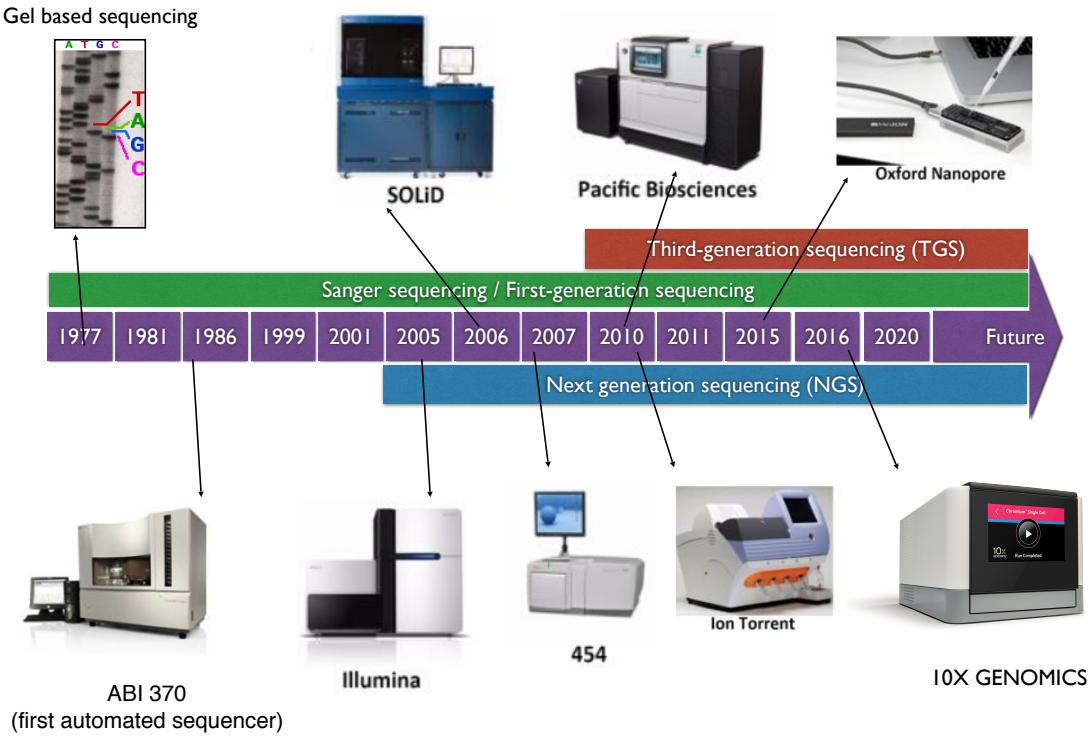


Figure 4: Timeline of DNA sequencing technology and representative DNA sequencers.

Pictures of the figure are utilized mainly from Wikimedia Commons.

By the mid 2000s, NGS technologies started to appear (Heather and Chain 2016). These new forms of DNA sequencing, such as 454 pyrosequencing and SOLiD sequencing, were faster, cheaper, and had much greater throughput than their Sanger predecessor, but they were also more error-prone (~0.1-15%) and gave much shorter read lengths (35-700 nt) (Goodwin *et al.* 2016). In 2007, the NGS technology Solexa was purchased by Illumina (Balasubramanian 2015), rebranded as Illumina sequencing, and quickly became the leading DNA sequencing technology, and arguably still is today. Indeed, Illumina has

decoded more genomes than other kinds of DNA sequencing, but its biggest drawback is the short length of its reads (50-250 nt).

To overcome the read-length limitations of NGS, long-read third-generation sequencing (TGS) sequencing technologies have been developed. In 2010, Pacific Biosciences (PacBio) released the Single Molecule Real Time (SMRT) sequencing system, which can yield reads that are thousands of nucleotides long. At the same time, Oxford Nanopore designed a portable DNA sequencing device (MinION), which was even tested in space (Rainey 2017). The long reads of these TGS technologies are great for resolving large, structurally complex genomes, but can be expensive, have a very large error rate (5-15%), and lower throughput than their NGS counterparts (Goodwin *et al.* 2016).

1.3 Green Algal Genomics

As DNA sequencing technologies have improved so has our ability to assemble entire genomes, especially large, complex eukaryotic genomes (Figure 5), including those of green algae. The first complete green algal nuclear genomes to be sequenced (those of *O. tauri* and *C. reinhardtii*) were completed using solely Sanger sequencing via the whole genome shotgun method (Derelle *et al.* 2006; Merchant *et al.* 2007). These projects also involved teams of hundreds of researchers, took many years to finish, and usually contained thousands of gaps. For example, the previous published assembly of the ~120 Mb *C. reinhardtii* nucDNA comprised 1500 repeat-rich scaffolds, 15,143 intron-dense genes, and was about 95% complete (Merchant *et al.* 2007). The last green algal nucDNA to be sequenced using an entirely Sanger-based approach was that of *V. carteri* (131 Mb), and was carried out by a team of approximately 20 researchers (Prochnik *et al.* 2010). Soon thereafter, scientists started using NGS, or a combination of NGS and Sanger sequencing, to obtain green algal nucDNA sequences. In 2014, the massive draft nuclear genome (>340 Mb) of the chlamydomonadalean alga *D. salina* was sequenced using a primarily Illumina-based approach (Polle *et al.* 2017). More recently, the genomes of the chlamydomonadaleans *G. pectorale* (~150 Mb) and *C. eustigma* (~130 Mb) were sequenced using a combination of 454 and Illumina sequencing (Hanschen *et al.* 2016; Hirooka *et al.* 2017). Today, small teams of researchers are also resequencing some of the early Sanger-based green algal genomes using NGS, and to great effect—such an approach

has helped reduce the *C. reinhardtii* genome assembly to 37 scaffolds, representing 99.5% of the genome.

Given the availability of sequenced green algal genomes, without a comprehensive genomic framework , some of the key findings (e.g., evolution of multicellularity and environmental adaptation) to the green algal genomics are severely impeded (Blaby-Haas and Merchant 2019). Comparison of *C. reinhardtii* and *V. carteri* has already revealed the evolution of multicellularity and cellular differentiation in volvocine algae (Leliaert *et al.* 2012). Specifically, Prochnik *et al.* discovered that the organismal complexity is highly associated with the lineage-specific protein modifications in the multicellular green alga *V. carteri* (Prochnik *et al.* 2010). Six years later, Hanschen and colleagues (2016) furtherly explored the colonial alga *G. pectorale* and emphasized that the early co-option of cell cycle regulation for group-level life cycle and reproduction are key step in the evolution of multicellularity. Until recently, the adaptation of green algae to some extreme environments have been explained by the genome availability of acidophilic green alga *C. eustigma*, halophilic green alga *D. salina* and psychrophilic green alga ICE-L. Although details have not been clarified in *D. salina* about the adaptive strategies in sea salt fields, it can alleviate the stresses via accumulating glycerol and β-carotene in response to high salinity and intense UV light (Polle *et al.* 2017). Furthermore, the existence of common mechanisms in the adaptation to extreme environments have been observed in *C. eustigma* and ICE-L. Hirooka *et al.* revealed that the energy shuttle and buffering system and arsenic detoxification genes were acquired via HGT in *C. eustigma* to survive in acidic environment (Hirooka *et al.* 2017). Similarly, multiple IBPs genes originated from bacteria were assumed to contribute to the origin of the psychrophilic lifestyle in ICE-L (Zhang *et al.* 2020). The rapidly increasing availability of genomic data can provide a window into understanding the complexity of algal genomics. Comparative genomics will become a very effective tool allowing us to answer some of the critical questions, such as how the evolutionary transition of green algae from unicellular to multicellular occurred, how the acidophilic green algae evolved from their respective neutrophilic ancestors, and what difference is between psychrophilic green algae and their close mesophilic counterparts.

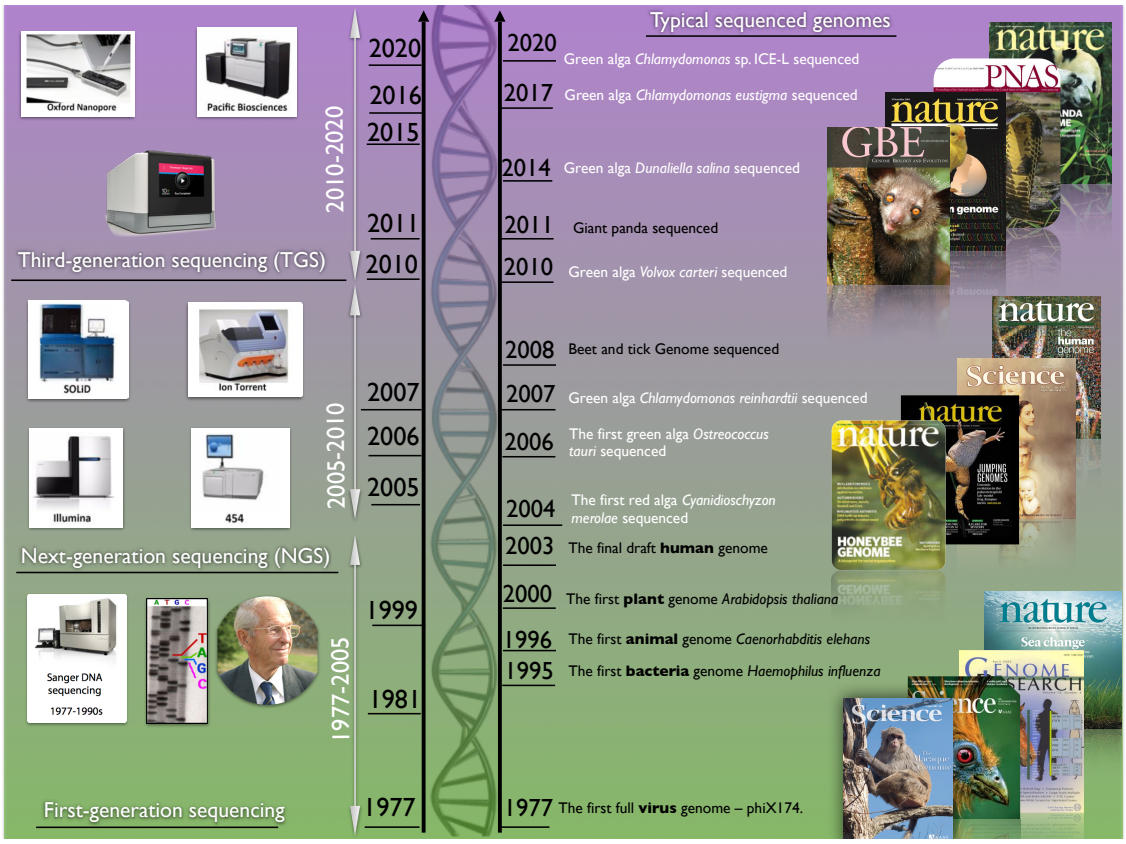


Figure 5: The DNA sequencing technologies and representative genomes.

1.4 Genome Assembly and Annotation

Why is the eukaryotic genome assembly so challenging? There are many reasons, including the large size of nucDNAs, high densities of repeats, heterozygosity, low read coverage, biased sequencing, high error rates, chimeric reads, sequencing adapters in the reads, and sample contamination, among many others (Li *et al.* 2010). Some of these challenges (e.g., chimeric reads) can be handled via a robust quality control process. Others, such as low read coverage and a high error rate, can still be overcome by employing different types and greater amounts of sequencing (Lee *et al.* 2016). However, for repeats (i.e., long terminal repeats (LTRs) and terminal inverted repeats (TIRs)) interfering the exon and intron boundaries, manual efforts usually have to be utilized. Although repeats can be alleviated through the state-of-the-art long-read sequencing, such as Oxford Nanopore Technologies (ONT) or SMRT Pacific Biosciences (PacBio) sequencing platforms, repeat finding can be incorporated as a critical step in the genome assembly pipeline (Haridas *et al.* 2011), which

is a series of computational steps that input raw sequencing reads and ultimately output an assembled draft genome.

While a genome can be annotated without a highly contiguous assembly, some of the key annotation information might also be missed, such as incomplete genes (losing stop codons or start codons) being mistakenly treated as pseudogenes and repeat regions being falsely regarded as coding genes. To facilitate some aspects of the downstream analysis (e.g., novel genes and gene family identification, HGT detection and duplicate gene exploration), every step of the genome pipeline must undoubtedly be followed carefully to obtain a well-annotated genome, given that it is a daunting task. Eukaryotic genome annotation entails many different steps, but usually begins with repeat masking, whereby all repetitive regions are masked to not confuse the annotation algorithms. This is then followed by the identification of open reading frames and the structural prediction of all coding regions, including exon and intron boundary prediction, and finally, functional annotations are assigned to these regions. Once complete, genome annotation allows for detailed comparative genomic analyses, from gene content and order comparisons to phylogenetic analyses.

Although sequencing more algal genomes can help better understand the diversity of algal biology, high-quality genome assembly and structural annotations are necessary to facilitate protein identification (Blaby-Haas and Merchant 2019). Fortunately, some green algae (Chlorophyta) genomes are haploid and represent relatively small genome size, such as the smallest free-living eukaryote *O. tauri* (12.5 Mb) (Derelle *et al.* 2006). Additionally, the *C. reinhardtii* has been updated with high-confidence gene models (JGI v5.6), which provide an excellent reference system to explore the biological functions of other green algae (Merchant *et al.* 2007; Blaby *et al.* 2014). However, for those non-model organisms without an available reference genome such as UWO241, the challenges of green algal nuclear genomics are not limited to the relatively huge genome size (~ 230 Mb) but the highly repetitive regions. The number and distribution of repeats can greatly influence the genome assembly and genome annotation, because sequencing reads from these regions are very similar which will confuse the assembly tools to extend the contigs at these regions (Walker *et al.* 2014). Moreover, some of the long repetitive sequences such as LTR

retrotransposons (5~9 kb) and LINEs (5~8 kb) are even longer than sequencing reads, especially for Illumina reads (250 bp) (Lerat 2010). It is reported that a total of 63.78% of the ICE-L genome assembly lengths (345.23 Mb) were identified as repeat regions, among them approximately 40.67% are transposable elements (TEs). Long terminal repeat retrotransposons (LTR-RTs) were the most dominant type of TEs, representing 23.32% of the assembly (Zhang *et al.* 2020). Generally, to alleviate the likely confusion and misassemblies from repeats during genome assembly, long-read technologies (e.g., PacBio or Nanopore) are selected to generate hybrid assemblies, because they stretch repetitive regions and thus provide more contiguous reconstructions of the genome (De Maio *et al.* 2019).

Organellar DNAs can interfere with nuclear assemblies. Thus, it is good to assemble these genomes first during the assembly and annotation process. Organellar DNAs were first completely sequenced from human and mouse mitochondria in 1981 (Anderson *et al.* 1981; Bibb *et al.* 1981). Five years later (1986), plastid genomes were unraveled in *Marchantia polymorpha* (Ohyama *et al.* 1986) and tobacco (Shinozaki *et al.* 1986). With the efforts of researchers worldwide, thousands of other organellar genomes have been sequenced and published. As of Oct. 2020, there were ~17,000 complete mitochondrial DNA (mtDNA) and plastid DNA (ptDNA) sequences in GenBank, making organellar genomes the most highly sequenced types of genomes.

Although not the main focus of this thesis, it should be noted that organellar genomes are usually filtered to acquire a pure nuclear genome assembly. Otherwise, the organellar DNA will create confusion during the nuclear genome assembly and annotation. The UWO241 mtDNA and ptDNA are both available and can be found in the publication (Cvetkovska *et al.* 2019).

1.5 Thesis Objectives

There are three major objectives in my thesis: (1) To generate a high-quality nuclear genome assembly of UWO241 via both NGS and TGS sequencing reads, (2) To accurately and thoroughly annotated this genome, and (3) To use these data in a comparative framework for a better understanding of the evolution of psychrophily. More specifically,

the nuclear genome assembly and gene annotation pipelines will be carried out using the most appropriate available bioinformatics software and algorithms. Second, gene content and genomic architecture will be compared to the well-annotated chlamydomonadalean genomes. Little is known about the genome; it is tempting to deepen our understanding in these questions. For example, does the UWO241 genome harbor large numbers of duplicate genes? Has it acquired any genes via HGT, such as IBP genes? Does UWO241 contain unique gene families compared to close related relatives? This thesis will examine the basis of the psychrophily in UWO241 and hopefully provide insights into what allows UWO241 to survive in such an extreme environment.

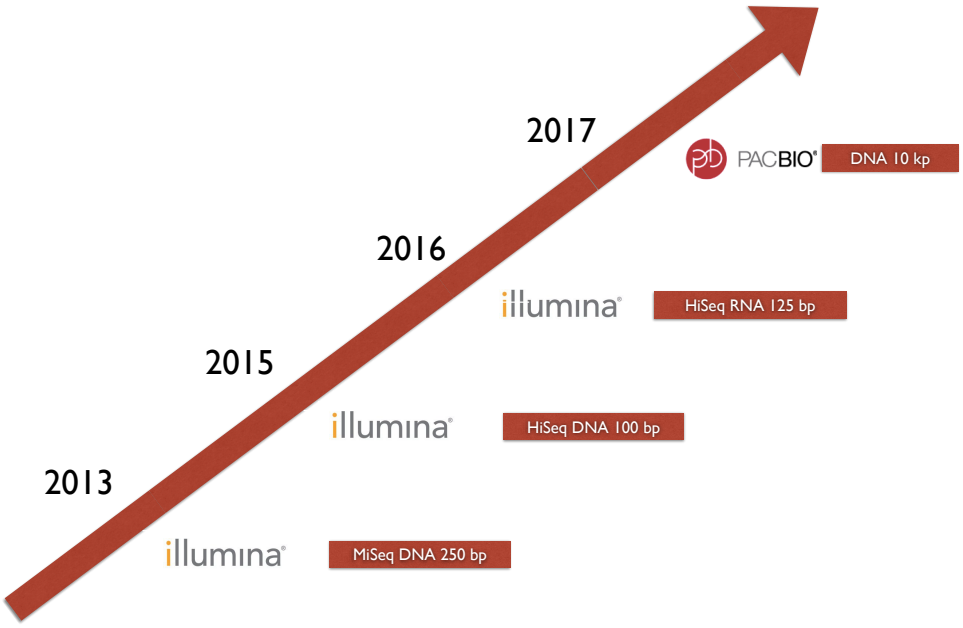


Figure 6: The timeline of sequencing data acquired from UWO241 genome.

Over the past seven years, the Hüner and Smith laboratories have carried out various NGS (Illumina) and TGS (PacBio) sequencing data for UWO24 (Figure 6 and Table 1), almost all of which remain largely unexplored and unannotated. It is the key objective of my thesis to employ these data to assemble and annotate UWO241 nucDNA. It is worth of mention that although it took only a few weeks or days to generate these genomic data sets, assembling them into a draft nuclear genome is not a trivial undertaking, explaining why they remained unanalyzed after years of being available. Indeed, constructing a nuclear

genome assembly is a lengthy and computationally intensive process. Fortunately, many other teams have assembled green algal nuclear genomes and performed painstaking and pioneering bioinformatics work to help guide me through the process.

Table 1: NGS (Illumina) and TGS (PacBio) sequencing data from UWO241.

	Year of sequencing	# Number of reads	Average read length (bp)	Average genome coverage
Illumina MiSeq paired-end DNA	2013	17,071,586	~250	~17x
Illumina HiSeq paired-end DNA	2015	193,716,744	~100	~77x
PacBio SMRTcell-DNA	2017	1,649,659	~10,000	~66x
Illumina HiSeq-RNA	2016	37,748,239	~125	~19x

1.6 References

Anderson, S., A. T. Bankier, B. G. Barrell, M. H. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe and F. Sanger (1981). Sequence and organization of the human mitochondrial genome. Nature 290: 457-465.

Anisimov, O., B. Fitzharris, J. Hagen, R. Jefferies, H. Marchant, F. Nelson, T. Prowse and D. Vaughan (2001). Polar regions (arctic and antarctic). Climate Change: 801-841.

Åqvist, J., G. V. Isaksen and B. O. Brandsdal (2017). Computation of enzyme cold adaptation. Nature Reviews Chemistry 1: 1-14.

Balasubramanian, S. (2015). Solexa sequencing: Decoding genomes on a population scale. Clinical Chemistry 61: 21-24.

Becker, S., M. L. Quartino, G. L. Campana, P. Bucolo, C. Wiencke and K. Bischof (2011). The biology of an Antarctic rhodophyte, *Palmaria decipiens*: recent advances. Antarctic Science 23: 419-430.

Bibb, M. J., R. A. Van Etten, C. T. Wright, M. W. Walberg and D. A. Clayton (1981). Sequence and gene organization of mouse mitochondrial DNA. Cell 26: 167-180.

Blaby, I. K., C. E. Blaby-Haas, N. Tourasse, E. F. Hom, D. Lopez, M. Aksoy, A. Grossman, J. Umen, S. Dutcher and M. Porter (2014). The *Chlamydomonas* genome project: a decade on. Trends in Plant Science 19: 672-680.

Blaby-Haas, C. E. and S. S. Merchant (2019). Comparative and functional algal genomics. Annual Review of Plant Biology 70: 605-638.

Blanc, G., I. Agarkova, J. Grimwood, A. Kuo, A. Brueggeman, D. D. Dunigan, J. Gurnon, I. Ladunga, E. Lindquist and S. Lucas (2012). The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. Genome Biology 13: 1-12.

Bowman, J. S., T. J. Vick-Majors, R. Morgan-Kiss, C. Takacs-Vesbach, H. W. Ducklow and J. C. Priscu (2016). Microbial community dynamics in two polar extremes: The lakes of the McMurdo Dry Valleys and the West Antarctic Peninsula marine ecosystem. BioScience 66: 829-847.

Brown, S. P., B. J. Olson and A. Jumpponen (2015). Fungi and algae co-occur in snow: an issue of shared habitat or algal facilitation of heterotrophs? Arctic, Antarctic, and Alpine Research 47: 729-749.

Coleine, C., N. Pombubpa, L. Zucconi, S. Onofri, J. E. Stajich and L. Selbmann (2020). Endolithic fungal species markers for harshest conditions in the McMurdo Dry valleys, Antarctica. Life 10: 1-12.

Cvetkovska, M., N. P. A. Huner and D. R. Smith (2017). Chilling out: the evolution and diversification of psychrophilic algae with a focus on Chlamydomonadales. Polar Biology 40: 1169-1184.

Cvetkovska, M., S. Orgnero, N. P. Hüner and D. R. Smith (2019). The enigmatic loss of light-independent chlorophyll biosynthesis from an Antarctic green alga in a light-limited environment. New Phytologist 222: 651-656.

Cvetkovska, M., B. Szyszka-Mroz, M. Possmayer, P. Pittock, G. Lajoie, D. R. Smith and N. P. Hüner (2018). Characterization of photosynthetic ferredoxin from the Antarctic alga *Chlamydomonas* sp. UWO241 reveals novel features of cold adaptation. New Phytologist 219: 588-604.

D'Amico, S., T. Collins, J. C. Marx, G. Feller, C. Gerday and C. Gerday (2006). Psychrophilic microorganisms: challenges for life. EMBO Reports 7: 385-389.

DasSarma, S., M. D. Capes, R. Karan and P. DasSarma (2013). Amino acid substitutions in cold-adapted proteins from *Halorubrum lacusprofundi*, an extremely halophilic microbe from Antarctica. PLoS One 8: e58587.

De Maayer, P., D. Anderson, C. Cary and D. A. Cowan (2014). Some like it cold: understanding the survival strategies of psychrophiles. EMBO Reports: e201338170.

De Maio, N., L. P. Shaw, A. Hubbard, S. George, N. D. Sanderson, J. Swann, R. Wick, M. AbuOun, E. Stubberfield and S. J. Hoosdally (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microbial Genomics 5: e000294.

Derelle, E., C. Ferraz, S. Rombauts, P. Rouzé, A. Z. Worden, S. Robbens, F. Partensky, S. Degroeve, S. Echeynié and R. Cooke (2006). Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proceedings of the National Academy of Sciences 103: 11647-11652.

Dolhi, J. M., D. P. Maxwell and R. M. Morgan-Kiss (2013). The Antarctic *Chlamydomonas raudensis*: an emerging model for cold adaptation of photosynthesis. Extremophiles 17: 711-722.

Feller, G. and C. Gerday (2003). Psychrophilic enzymes: hot topics in cold adaptation. Nature Reviews Microbiology 1: 200-208.

Florindo, F. and M. Siegert (2008). A History of Antarctic Cenozoic Glaciation–View from the Margin. Antarctic Climate Evolution 8: 33.

Golomb, D. (1993). Ocean disposal of CO2: Feasibility, economics and effects. Energy Conversion and Management 34: 967-976.

Goodwin, S., J. D. McPherson and W. R. McCombie (2016). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 17: 333-351.

Hanschen, E. R., T. N. Marriage, P. J. Ferris, T. Hamaji, A. Toyoda, A. Fujiyama, R. Neme, H. Noguchi, Y. Minakuchi and M. Suzuki (2016). The *Gonium pectorale* genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. Nature Communications 7: 1-10.

Haridas, S., C. Breuill, J. Bohlmann and T. Hsiang (2011). A biologist's guide to de novo genome assembly using next-generation sequence data: a test with fungal genomes. Journal of Microbiological Methods 86: 368-375.

Hassan, N., M. Rafiq, M. Hayat, A. A. Shah and F. Hasan (2016). Psychrophilic and psychrotrophic fungi: a comprehensive review. Reviews in Environmental Science and Bio/Technology 15: 147-172.

Heather, J. M. and B. Chain (2016). The sequence of sequencers: the history of sequencing DNA. Genomics 107: 1-8.

Hirooka, S., Y. Hirose, Y. Kanesaki, S. Higuchi, T. Fujiwara, R. Onuma, A. Era, R. Ohbayashi, A. Uzuka and H. Nozaki (2017). Acidophilic green algal genome provides insights into adaptation to an acidic environment. Proceedings of the National Academy of Sciences 114: 8304-8313.

Horikoshi, K., G. Antranikian, A. T. Bull, F. T. Robb and K. O. Stetter (2010). Extremophiles: Psychrophiles. Extremophiles handbook. Tokyo/Dordrecht/Heidelberg/London/New York, Springer Science & Business Media 2: 755-891.

Kwok, R., S. Kacimi, M. Webster, N. Kurtz and A. Petty (2020). Arctic Snow Depth and Sea Ice Thickness From ICESat-2 and CryoSat-2 Freeboards: A First Examination. Journal of Geophysical Research: Oceans 125: e2019JC016008.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle and W. FitzHugh (2001). Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Lee, H., J. Gurtowski, S. Yoo, M. Nattestad, S. Marcus, S. Goodwin, W. R. McCombie and M. Schatz (2016). Third-generation sequencing and the future of genomics. BioRxiv: 048603.

Leliaert, F., D. R. Smith, H. Moreau, M. D. Herron, H. Verbruggen, C. F. Delwiche and O. De Clerck (2012). Phylogeny and molecular evolution of the green algae. Critical Reviews in Plant Sciences 31: 1-46.

Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity 104: 520-533.

Li, R., W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang, Q. Cai, B. Li and Y. Bai (2010). The sequence and *de novo* assembly of the giant panda genome. Nature 463: 311-317.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu and M. Law (2012). Comparison of next-generation sequencing systems. Journal of Biomedicine and Biotechnology 2012: 1-11.

Liu, S., C. Liu, X. Huang, Y. Chai and B. Cong (2006). Optimization of parameters for isolation of protoplasts from the Antarctic sea ice alga *Chlamydomonas* sp. ICE-L. Journal of Applied Phycology 18: 783-786.

Lyon, B. R. and T. Mock (2014). Polar microalgae: new approaches towards understanding adaptations to an extreme and changing environment. Biology 3: 56-80.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. Trends in Genetics 24: 133-141.

Margesin, R. and V. Miteva (2011). Diversity and ecology of psychrophilic microorganisms. Research in Microbiology 162: 346-361.

Margesin, R., F. Schinner, J.-C. Marx and C. Gerday (2008). Psychrophiles: from biodiversity to biotechnology, Springer Verlag, Berlin Heidelberg 1:1-685.

Math, R. K., H. M. Jin, J. M. Kim, Y. Hahn, W. Park, E. L. Madsen and C. O. Jeon (2012). Comparative genomics reveals adaptation by *Alteromonas* sp. SN2 to marine tidal-flat conditions: cold tolerance and aromatic hydrocarbon metabolism. PLoS One 7: e35784.

Maxam, A. M. and W. Gilbert (1977). A new method for sequencing DNA. Proceedings of the National Academy of Sciences 74: 560-564.

Merchant, S. S., S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz, G. B. Witman, A. Terry, A. Salamov, L. K. Fritz-Laylin and L. Maréchal-Drouard (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318: 245-250.

Mikucki, J. A., S. Han and B. D. Lanoil (2011). Ecology of psychrophiles: subglacial and permafrost environments. Extremophiles Handbook. Tokyo/Dordrecht/Heidelberg/London/New York, Springer 1: 755-775.

Mock, T. and N. Hoch (2005). Long-term temperature acclimation of photosynthesis in steady-state cultures of the polar diatom *Fragilariopsis cylindrus*. Photosynthesis Research 85: 307-317.

Mock, T., R. P. Otillar, J. Strauss, M. McMullan, P. Paajanen, J. Schmutz, A. Salamov, R. Sanges, A. Toseland and B. J. Ward (2017). Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. Nature 541: 536-540.

Montañez, I. P. and C. J. Poulsen (2013). The Late Paleozoic ice age: an evolving paradigm. Annual Review of Earth and Planetary Sciences 41: 629-656.

Morgan-Kiss, R., M. Lizotte, W. Kong and J. Priscu (2016). Photoadaptation to the polar night by phytoplankton in a permanently ice-covered Antarctic lake. Limnology and Oceanography 61: 3-13.

Morgan-Kiss, R. M., J. C. Priscu, T. Pocock, L. Gudynaite-Savitch and N. P. Huner (2006). Adaptation and acclimation of photosynthetic microorganisms to permanently cold environments. Microbiology and Molecular Biology Reviews 70: 222-252.

Morita, R. Y. (1975). Psychrophilic bacteria. Bacteriological Reviews 44: 983-1015.

Neale, P. J. and J. C. Priscu (1995). The photosynthetic apparatus of phytoplankton from a perennially ice-covered Antarctic lake: acclimation to an extreme shade environment. Plant and Cell Physiology 36: 253-263.

Obbard, R. W., S. Sadri, Y. Q. Wong, A. A. Khitun, I. Baker and R. C. Thompson (2014). Global warming releases microplastic legacy frozen in Arctic Sea ice. Earth's Future 2: 315-320.

Ohyama, K., H. Fukuzawa, T. Kohchi, H. Shirai, T. Sano, S. Sano, K. Umesono, Y. Shiki, M. Takeuchi and Z. Chang (1986). Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature 322: 572-574.

Pocock, T., M. A. Lachance, T. Pröschold, J. C. Priscu, S. S. Kim and N. P. Huner (2004). Identification of a psychrophilic green alga from Lake Bonney Antarctica: *Chlamydomonas raudensis* ETTL. (UWO241) Chlorophyceae. Journal of Phycology 40: 1138-1148.

Polle, J. E., K. Barry, J. Cushman, J. Schmutz, D. Tran, L. T. Hathwaik, W. C. Yim, J. Jenkins, Z. McKie-Krisberg and S. Prochnik (2017). Draft nuclear genome sequence of the halophilic and beta-carotene-accumulating green alga *Dunaliella salina* strain CCAP19/18. Genome Announcements 5: e01105-17.

Pop, M. and S. L. Salzberg (2008). Bioinformatics challenges of new sequencing technology. Trends in Genetics 24: 142-149.

Possmayer, M. (2018). Phylogeny, Heat-Stress and Enzymatic Heat-Sensitivity in the Antarctic Psychrophile, *Chlamydomonas* sp. UWO241. Electronic Thesis and Dissertation Repository 5737. Retrieved from https://ir.lib.uwo.ca/etd/5737.

Possmayer, M., R. K. Gupta, B. Szyszka‐Mroz, D. P. Maxwell, M. A. Lachance, N. P. Hüner and D. R. Smith (2016). Resolving the phylogenetic relationship between *Chlamydomonas* sp. UWO 241 and *Chlamydomonas raudensis* SAG 49.72 (Chlorophyceae) with nuclear and plastid DNA sequences. Journal of Phycology 52: 305-310.

Prochnik, S. E., J. Umen, A. M. Nedelcu, A. Hallmann, S. M. Miller, I. Nishii, P. Ferris, A. Kuo, T. Mitros and L. K. Fritz-Laylin (2010). Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science 329: 223-226.

Quesada, A. and W. F. Vincent (2012). Cyanobacteria in the cryosphere: snow, ice and extreme cold. Ecology of Cyanobacteria II. Spain/Canada, Springer 14: 387-399.

Rainey, K. (2017). First DNA Sequencing in Space a Game Changer. NASA. Retrieved from https://www.nasa.gov/mission_pages/station/research/news/dna_sequencing.

Remias, D., U. Karsten, C. Lütz and T. Leya (2010). Physiological and morphological processes in the Alpine snow alga *Chloromonas nivalis* (Chlorophyceae) during cyst formation. Protoplasma 243: 73-86.

Robinson, C. H. (2001). Cold adaptation in Arctic and Antarctic fungi. New Phytologist 151: 341-353.

Rodrigues, D. F. and J. M. Tiedje (2008). Coping with our cold planet. Applied and Environmental Microbiology 74: 1677-1686.

Rothschild, L. J. and R. L. Mancinelli (2001). Life in extreme environments. Nature 409: 1092-1101.

Russell, N. (1990). Cold adaptation of microorganisms. Philosophical Transactions of the Royal Society of London. B, Biological Sciences 326: 595-611.

Sanger, F. and A. R. Coulson (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology 94: 441-448.

Santiago, I. F., C. A. Rosa and L. H. Rosa (2017). Endophytic symbiont yeasts associated with the Antarctic angiosperms Deschampsia antarctica and Colobanthus quitensis. Polar Biology 40: 177-183.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. Nature Methods 5: 16-18.

Séranne, M. (1999). Early Oligocene stratigraphic turnover on the west Africa continental margin: a signature of the Tertiary greenhouse-to-icehouse transition? Terra Nova-Oxford 11: 135-140.

Shinozaki, K., M. Ohme, M. Tanaka, T. Wakasugi, N. Hayashida, T. Matsubayashi, N. Zaita, J. Chunwongse, J. Obokata and K. Yamaguchi ‐ Shinozaki (1986). The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. The EMBO Journal 5: 2043-2049.

Siddiqui, K. S., T. J. Williams, D. Wilkins, S. Yau, M. A. Allen, M. V. Brown, F. M. Lauro and R. Cavicchioli (2013). Psychrophiles. Annual Review of Earth and Planetary Sciences 41: 87-115.

Szyszka-Mroz, B., M. Cvetkovska, A. G. Ivanov, D. R. Smith, M. Possmayer, D. P. Maxwell and N. P. Hüner (2019). Cold-adapted protein kinases and thylakoid remodeling impact energy distribution in an Antarctic psychrophile. Plant Physiology 180: 1291-1309.

Umen, J. G. and B. J. Olson (2012). Genomics of volvocine algae. Advances in Botanical Research 64: 185-243.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans and R. A. Holt (2001). The sequence of the human genome. Science 291: 1304-1351.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman and S. K. Young (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One 9: e112963.

Wang, D. S., D. Xu, Y. T. Wang, X. Fan, H. Y. Nai, W. Q. Wang, X. W. Zhang, S. L. Mou and Z. Guan (2015). Adaptation involved in nitrogen metabolism in sea ice alga *Chlamydomonas* sp. ICE-L to Antarctic extreme environments. Journal of Applied Phycology 27: 787-796.

Williams, W. E., H. L. Gorton and T. C. Vogelmann (2003). Surface gas-exchange processes of snow algae. Proceedings of the National Academy of Sciences 100: 562-566.

Wright, N. M., M. Seton, S. E. Williams, J. M. Whittaker and R. D. Müller (2020). Sea level fluctuations driven by changes in global ocean basin volume following supercontinent break-up. Earth-Science Reviews: 103293.

Yadav, A. N., S. Mishra, S. Singh and A. Gupta (2019). Recent advancement in white biotechnology through fungi: volume 1: diversity and enzymes perspectives. Switzerland, Springer International Publishing 1: 1-62.

Yadav, J., A. Kumar and R. Mohan (2020). Dramatic decline of Arctic sea ice linked to global warming. Natural Hazards 1: 1-5.

Zhang, Z., C. Qu, K. Zhang, Y. He, X. Zhao, L. Yang, Z. Zheng, X. Ma, X. Wang and W. Wang (2020). Adaptation to extreme Antarctic environments revealed by the genome of a sea ice green alga. Current Biology 30: 1-12.

Chapter 2

2 Step-by-Step User Guide in Characterizing the Assembly and Annotation of the Eukaryotic Genomes via High-throughput Sequencing Analysis

2.1 Introduction

The sequencing costs have fallen so dramatically that even a single laboratory can now afford to sequence large eukaryotic genomes (Lee *et al.* 2016); however, it remains a challenging task for the genome project especially in genome annotation (Yandell and Ence 2012). This is in part due to the many barriers in genome projects, the published literature simplifies details in methods sections or omits some tedious bioinformatics steps, which should be part of supplementary materials. Consequently, these factors can create great difficulty in understanding and following for biologists with little to no background in high-throughput sequencing analysis (i.e., DNA-Seq and RNA-Seq). Moreover, there is no objectively 'correct' way of performing genome projects. Given that the commercial software suites of today are being developed powerfully with user-friendly graphical interfaces and 'one click' analysis workflows, such software bundles, unfortunately, often includes expensive, proprietary (closed-source) programs, which are constrained to narrowly defined selections of the most popular analyses (De Wit *et al.* 2012; Del Angel *et al.* 2018). Alternatively, most bioinformatic software requires considerable knowledge of programming. Taking the genome project of green alga *Chlamydomonas* sp. UWO241 as an example, data files were obtained and managed within a UNIX-like environment; scripting languages, such as Python and Perl, were utilized to manipulate and clean the data; and the processed outputs were analyzed and visualized using language such as R script. It is no exaggeration to say that there is a widespread and exponentially growing demand for bioinformatic skills, and this is particularly in line with the concomitant expansion of guidance in such skills.

The pipelines and algorithms described in this chapter were used to assemble and annotate the UWO241 genome and these methods were used to form a step-by-step user guide. Here, I present a comprehensive bioinformatics foundation for genome projects specifically for

those researchers who have diverse backgrounds but no prior experience in programming. First and foremost, the assembly pipeline was developed to process DNA-Seq reads into genomic contigs. Taken together with these contigs, RNA-sequencing data were fed into an annotation pipeline, which selected the most up-to-date eukaryotic bioinformatic gene-profiling software. Finally, computational analyses were carried out on an in-house computer and supercomputing network, which is a great computing resource for computationally intensive bioinformatics work. Additionally, a small set of comparative genomic analyses were carried out as an example across the green algae from the order Chlamydomonadales.

2.2 Genome Assembly

As DNA sequencing technologies have improved, so has our ability to assemble entire genomes, especially large, complex eukaryotic genomes (Henson *et al.* 2012). However, high-quality genome assembly and annotation are still major issues (Simão *et al.* 2015). Researchers have to devote considerable time, computing resources and storage resources to perform their genome projects. For example, it could take fairly few resources and little time for small genomes, such as those of bacteria or archaea, but it will take months or even years for eukaryotic genomes, especially those of non-model organisms without an available reference genome (Del Angel *et al.* 2018). Therefore, it is important to understand the goal of the project before proceeding, such as to what extent the genome assembly and annotation will be able to address the respective biological questions. In the case of a draft genome being needed, financial and computational resources are important to consider. This is because sequence coverage relies on the amount of DNA to be sequenced, and the number of computing hours highly depends on the computing cluster performances (Haas *et al.* 2013). Presumably, the Illumina sequencing will need a more than 60x sequence depth, which means that the total number of nucleotides in the reads must be at least 60 times the number of nucleotides in the genome. Therefore, the importance of evaluating the genome size beforehand should not be underestimated. Although utilizing the flow cytometry could be an option (measuring the amount of DNA in a nucleus), the genome size can also be roughly estimated by *k*-mer (Genome Size Estimation Tutorial; https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/) and the comparison to

the genomes of closely related species (Pflug *et al.* 2020), but closely related species can have very different genome sizes (Pellicer *et al.* 2018).

Table 2: The representative genome assemblers being used in genome projects.

Data types	Assemblers	Remarks
Illumina reads	SPAdes	SPAdes has been successfully applied on some eukaryotic genomes.
PacBio reads	Canu	Canu is designed for long reads from PacBio or Nanopore.
Illumina and PacBio reads	MaSuRCA	MaSuRCA builds mega-reads for hybrid PacBio and Illumina to do *de novo* assembly.
	Pilon	Using the Illumina data to polish the long-read assemblies, which can lower consensus errors and mismatches.

It is tempting to decipher genome assembly pipelines in part due to their imperative role in genome projects. Genome assembly pipelines usually contain the following necessary steps: read quality control (QC), genome assembly, contig scaffolding, and gap filling. First and foremost, QC is the step involving the removal of sequencing adapters and the screening of low-quality reads. Various bioinformatics programs have been developed, such as FastQC, which is a user-friendly toolbox (Andrews 2010).

Importantly, without a comparative understanding of the assembly mechanisms and tools, it is impossible to obtain a highly contiguous genome assembly. There are two major genome assembly mechanisms. The overlap-layout-consensus (OLC) assembly approach is specialized for long reads (10-15 kb) from Pacific Biosciences Single-Molecule Real-Time (PacBio SMRT) and Oxford Nanopore sequencing technologies, while the de Bruijn Graph (DBG) approach is designed for short NGS reads. Via these algorithms, a wide variety of assemblers have been developed, which can be grouped into three straightforward categories: short-read assemblers, long-read assemblers, and hybrid assemblers (Table 2). Short-read assemblers, such as Abyss (Simpson *et al.* 2009), Spades (Bankevich *et al.* 2012), and SOAPdenovo2 (Luo *et al.* 2012), are excellent for high-coverage, repeat-poor genomes, whereas long-read assemblers, such as Canu (Koren *et al.*

2017), are able to work with low-coverage data sets and navigate through complex repeats. Alternatively, hybrid assemblers, represented by MaSuRCA (Zimin *et al.* 2013; Zimin *et al.* 2017) and Spades (Bankevich *et al.* 2012), can combine the efficiency of the DBG approach with the benefits of the OLC approach. Therefore, a mixture of short and long reads can be assembled together. Spades is designed for assembling small genomes (i.e., bacterial genomes), whereas MaSuRCA has been applied to some of the largest genomes on record such as human-sized genomes (Callaway 2017; Zimin *et al.* 2017).

Furthermore, to facilitate genome identification, a draft genome assembly should be polished beforehand. Gap filling and scaffolding are two strategies that should not be underestimated. Many bioinformatic tools are available for scaffolding and gap filling, including SSPACE (Boetzer *et al.* 2010), which can scaffold contigs using paired-end (PE) and/or mate-pair (MP) libraries, as well as PBJelly (English *et al.* 2012), which is an automated pipeline for aligning PacBio reads to draft assemblies. It is worth noting that long reads (i.e., PacBio reads) are used primarily for contig construction, while the short reads (i.e., Illumina PE reads) are employed for polishing (Figure 7). For example, Pilon (Walker *et al.* 2014) is a tool used to improve genome assembly accuracy and resolve misassemblies with either short or long reads.

Although uniform standards are lacking, the quality of any genome assembly is critically assessed using the following three factors: contiguity, completeness, and accuracy (Lee *et al.* 2016). Longer contigs are always meaningful in terms of contiguity, but for completeness, the assembled contigs should take into account most of the genome. Moreover, misassemblies and consensus errors should be alleviated to increase accuracy (Li *et al.* 2010; Lee *et al.* 2016). Nonetheless, without the common metrics used to indicate the quality of genome assembly, the progress of genome projects will be greatly impeded. Fortunately, some quality assessment tools have been developed to visualize the quality of a genome assembly. For instance, Quast (Gurevich *et al.* 2013) takes advantage of genome assemblies by computing various metrics, including N50 (i.e., the length for which the collection of all contigs of that length or longer covers at least 50% of the assembly length) and L50 (i.e., the number of contigs whose length are no shorter than N50). Alternatively, BUSCO v3 (Simão *et al.* 2015) provides quantitative measures for the assessment of

genome assembly, gene set, and transcriptome completeness based on evolutionarily informed expectations of gene content (Zdobnov *et al.* 2017).

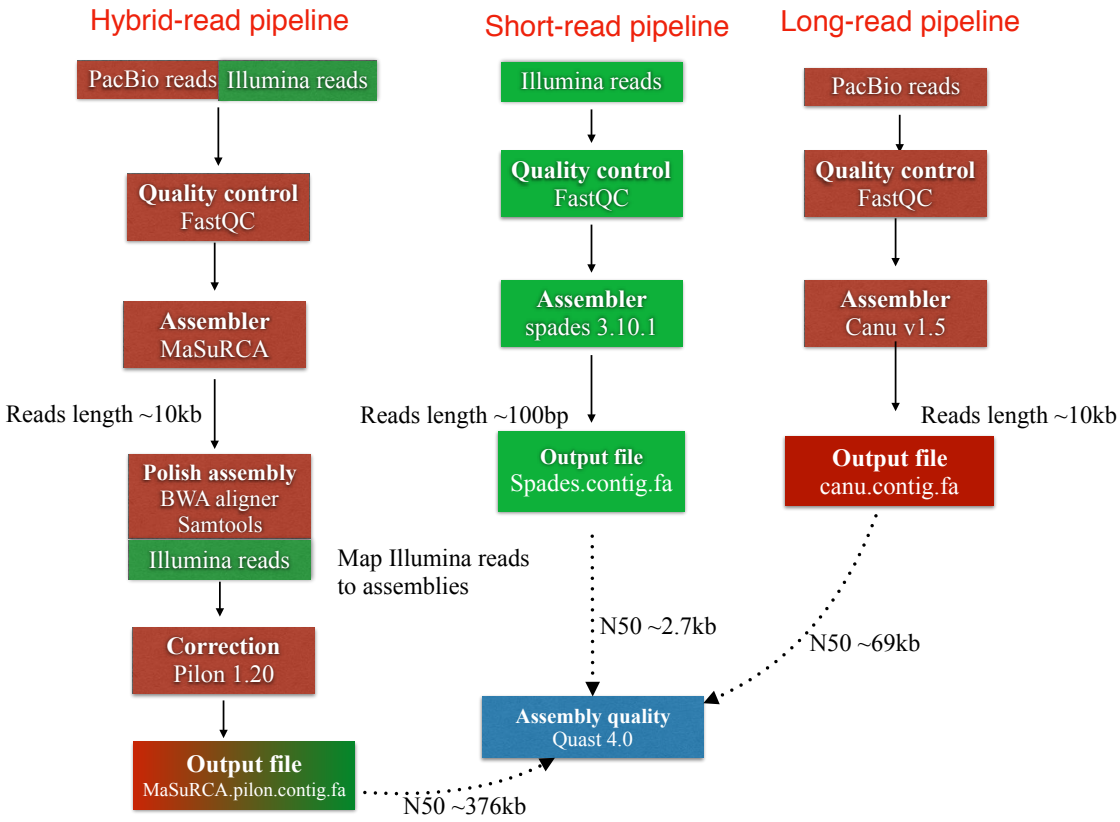


Figure 7: The genome assembly pipelines for assembling the UWO241 genome.

Illumina reads (short-read pipeline in green), PacBio reads (long-read pipeline in red) and hybrid reads (hybrid-read pipeline is highlighted in red and green, respectively). The module in blue indicates the assessment of genome assembly qualities. The colored boxes represent sequencing files, algorithms and assembled results in respective pipelines.

2.3 Genome Annotation

Annotating a eukaryotic genome is a daunting task, partly because many eukaryotic genomes are repeat rich and contain thousands of genes and introns (Yandell and Ence 2012). Eukaryotic genome annotation entails many different steps (Figure 8) but usually involves inferring the structure and function of assembled sequences. Protein-coding sequences are often explored first, and other noncoding sequences, such as noncoding RNA

(e.g., tRNA and rRNA), regulatory or repetitive sequences (e.g., enhancers, promoters, short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs)), can also be interpreted as well. Additionally, prior to acquiring nuclear genome annotation, contigs containing organellar DNA sequences (mitochondrial and/or chloroplastic DNA) should be filtered. Otherwise, the genome assembler will be confused and cost more computing resources when assembling the nuclear DNA sequences.

2.3.1 Structural Annotation

Structural annotation begins with repeat masking, whereby all repetitive regions are masked as not to disturb the annotation algorithms. Why are repeats so annoying? The number and distribution of the repeats can greatly influence the genome assembly and genome annotation results because sequencing reads from these repeat regions are very similar (Del Angel *et al.* 2018). Additionally, a high repeat content can contribute to a fragmented assembly, in part because the assembly tools cannot distinguish the correct assembly from these zones (Tørresen *et al.* 2019). Even worse, contigs will stop extending and will be bordered by repeats. Fortunately, brilliant tools have been developed to detect and identify these low-complexity regions, including transposable elements by making the nucleotide sequences lower case letters to distinguish from other regions, which are kept in upper case letters (i.e., soft masking method). RepeatMasker (Tarailo-Graovac and Chen 2009) and RepeatModeler (Smit and Hubley 2008) are two reputable repeat detection tools. RepeatModeler is a *de novo* repeat family identification and modeling package integrated with two *de novo* repeat finding programs (RECON (Haas *et al.* 2013) and RepeatScout (Price *et al.* 2005)). RepeatMasker harnesses nhmmer, cross_match, ABBlast/WUBlast, RMBlast and Decypher as search engines and utilizes curated libraries of repeats such as Dfam (profile HMM library) and Repbase (Bao *et al.* 2015) (Table 3). Given these repeat detection tools, there is an underestimation of the disturbance by repeats. Because partial sequencing reads, especially for Illumina reads (~250 bp), are shorter than some long repetitive sequences, such as LTR retrotransposons (5~9 kb) and LINEs (5~8 kb), confusion and misassemblies are very likely during genome assembly. Generally, to alleviate such issues, long-read technologies (e.g., PacBio or Nanopore) are selected to

generate hybrid assemblies, because they can often stretch past the entire length of repetitive regions and thus provide more contiguous reconstructions of the genome.

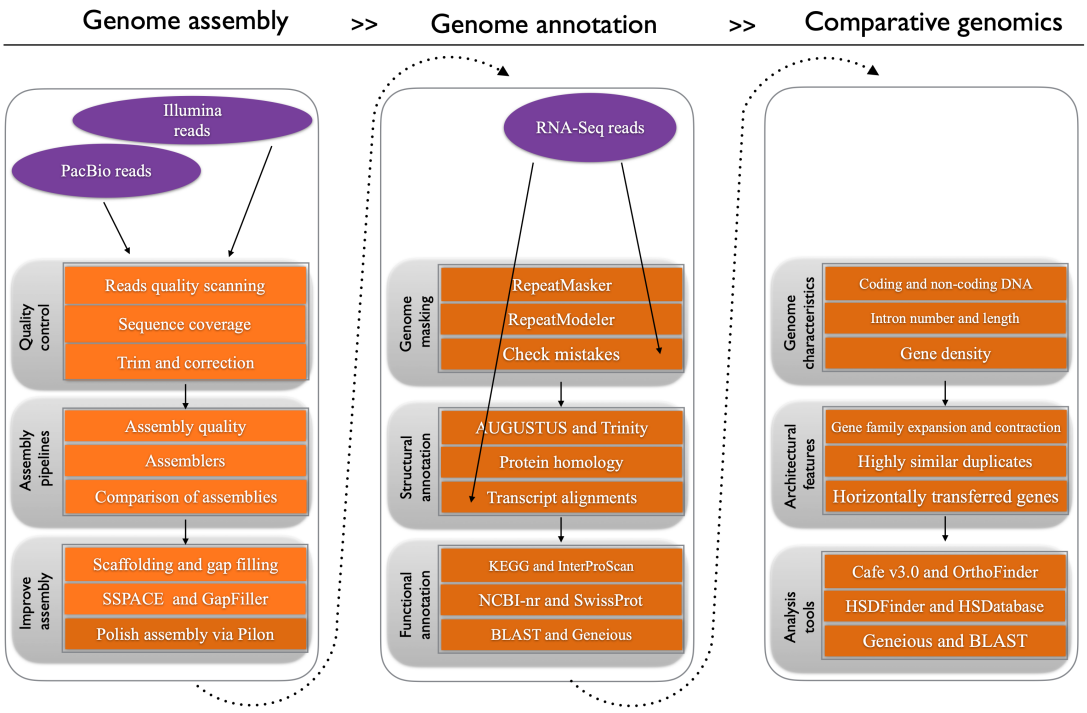


Figure 8: The typical workflow of a nuclear genome assembly and annotation. The purple color indicates the raw DNA-Seq and RNA-Seq reads. The orange modules present the detailed steps in the genome project.

Given a fully masked genome, genome annotation is advanced by deciphering open reading frames and coding region structures. This step includes but is not limited to exon and intron boundary prediction (Figure 8). In total, there are three main scenarios for predicting genes in a genome: intrinsic (*ab initio*), extrinsic, and the combiners. The *ab initio* method targets information that can be extracted from the genomic sequence itself, such as coding potential and splice site prediction. AUGUSTUS (Stanke *et al.* 2006) is one of the most representative tools using the conditional random field (generalization of HMM) method to predict eukaryotic genome genes via structural signals such as intron and coding sequencing (CDS) evidence. Thus, it appears that the intrinsic method is able to predict non-model organisms and their species-specific genes without external information. However, there is likely an underestimation of intensive labor, such as that needing to

manually create the training set which is a file with thousands of genes in standard formats (GenBank or GFF3) using for predicting the genome structure. Additionally, the respective software such as AUGUSTUS (Stanke *et al.* 2006) should be trained and optimized due to species differences. However, effort can be partially saved by retrieving the training gene sets from third party bioinformatic software, such as BUSCO (Simão *et al.* 2015), which automatically generate genome annotation training sets. It is noteworthy that plants usually need confident training sets to predict the genomes (Foissac *et al.* 2008). Moreover, plant genomes contain a large number of pseudogenes as well as novel protein-coding and noncoding genes, and these patterns of gene structure differ among organisms.

Although intrinsic methods are associated with information from the genome alone, it is once again very difficult to accurately interpret genome structure without external evidence such as transcripts and/or polypeptide sequencing data. Many pipelines and tools have been designed to utilize external information, such as BRAKER1 (Hoff *et al.* 2016) and MAKER (Cantarel *et al.* 2008). On the one hand, important external evidence includes transcripts, which can provide accurate gene coding information for correcting gene structure. Representatively, Trinity (Haas *et al.* 2013) was developed to reconstruct transcriptomes *de novo* from RNA-Seq data. On the other hand, protein homology evidence can indicate the presence and location of genes. This is partly because polypeptide sequences are more conserved and can be aligned even among distantly related species. Nevertheless, it should be noted that protein homology evidence greatly facilitates determining the presence of gene loci, but it is not always effective in outlining the exact structure of a gene. Some protein evidence detection tools are listed as follows. It is known that BLASTX (Kent 2002) can search the nucleotide query against the protein database by comparing protein sequences to the six translation-frames of the nucleotide sequences. However, when proceeding with large-scale pairwise alignment between protein data sets and whole genome sequences, Exonerate is deemed much more efficient (Slater and Birney 2005), allowing the alignment of sequences using a multiple alignment model.

In addition to the previously described structural annotation methods, researchers have developed a combined method that integrates *ab initio* draft prediction with extrinsic information. For example, EVidenceModeler (aka EVM) (Haas *et al.* 2008) software

integrates *ab initio* gene predictions, protein homology and transcript alignments into weighted consensus gene structures. Specifically, protein homology evidence and transcript alignment evidence are acquired from Exonerate (Slater and Birney 2005) and PASA (Haas *et al.* 2003), respectively. Program to Assemble Spliced Alignments (PASA) is a eukaryotic genome annotation tool that exploits spliced alignments of expressed transcript sequences to automatically model gene structures. Notably, balancing the weight value of the combined method is a tricky and subjective process. Many researchers unwillingly fall into a trap by consistently rerunning the weight value or the metrics, aiming for "perfect" data. However, it can easily takes months to iterate these gene prediction processes, and carefully proceeding to the next step is recommended as long as the structural annotation can help answer the current biological question (Del Angel *et al.* 2018)

Table 3: The summary of reputable software and algorithms in genome projects.

Software and Algorithms		
Genome assembly	Canu v1.6	https://github.com/marbl/canu
Transcriptome assembly	Trinity v2.4.0	https://github.com/trinityrnaseq/trinityrnaseq
Genome assembly	MaSuRCA v3.2.3	https://github.com/alekseyzimin/masurca
Assembly polishing	Pilon v1.20	https://github.com/broadinstitute/pilon
Contig scaffolding	SSPACE v3.0	https://github.com/nsoranzo/sspace_basic
Repetitive DNA-motif masking	RepeatMasker v4.0.7	https://github.com/rmhubley/RepeatMasker
Repetitive DNA-motif identification	RepeatModeler v1.0.8	https://github.com/rmhubley/RepeatModeler
Genome completeness	BUSCO v3.0.2	https://gitlab.com/ezlab/busco
Protein alignment	Diamond v0.9.18	https://github.com/bbuchfink/diamond
Gene prediction	AUGUSTUS	http://bioinf.uni-greifswald.de/augustus/
Gene prediction	EVidenceModeler	https://evidencemodeler.github.io/
Gene prediction	Exonerate v2.2.0	https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate
Gene prediction	PASA	https://github.com/PASApipeline/PASApipeline
Functional annotation	InterProScan v5.27	https://www.ebi.ac.uk/interpro/
Gene family prediction	OrthoFinder v2.1.2	https://github.com/davidemms/OrthoFinder
Gene family prediction	OrthoMCL	http://orthomcl.org/orthomcl/
Maximum likelihood tree calculation	RAxML v8.2.4	https://sco.h-its.org/exelixis/web/software/raxml/index.html
tRNA identification	tRNAscan-SE v1.31	http://lowelab.ucsc.edu/tRNAscan-SE/

2.3.2 Functional Annotation

If the purpose of structural annotation is to understand where the genes are located and what do they look like, then functional annotation aims to match coding and noncoding sequences with relevant biological information. Furthermore, the functions of coding gene models can be inferred by amino acid sequence similarity between the genome of interest and genomes in public sequence repositories. There are a wide variety of publicly available searchable databases, such as GenBank's nonredundant protein database (NR) (Pruitt *et al.* 2005), the manually annotated and curated Uniprot Swiss-Prot database (Apweiler *et al.* 2004) and the automatically annotated TrEMBL (Boeckmann *et al.* 2003) protein database. Alternatively, there are many tools available for searching the protein sequence similarity, starting with BLASTP (Kent 2002), which searches for a protein query against the protein database. Diamond is another sequence aligner for protein and translated DNA searches, designed for high-performance analysis of large sequence data sets (Buchfink *et al.* 2015). The significant matches from those aligners maintain information such as the gene name, a general description and the gene ID, among others. However, not all the matches from the aligner are considered significant, and the quality of a match depends on the length of the alignment and the percentage similarity. In addition, the E-value is often utilized as the criterion when screening outstanding sequence hits. The E-value describes the number of hits one can expect to see by chance when searching against a database of a particular size. Briefly, the lower the E-value, the more "significant" a match to a database sequence is (i.e., there is a smaller probability of finding a match just by chance).

2.4 Comparative Genomics

Via the assembly of highly contiguous and well-annotated genomes, researchers usually hope to deepen their investigations via comparative genomic analyses of factors, such as genome characteristics, metabolic pathways and phylogenetic relationships across closely related species. One of the major successes of comparative genomics is the dramatic increase in genome projects over the last decade, but without reputable bioinformatics websites and tools grounding the basis of analysis, it is impossible to smoothly interpret the findings. For example, the InterProScan database is commonly utilized to assess gene loss and gain in a genome of interest relative to the genomes of closely related species

(Quevillon *et al.* 2005), which integrates predictive information about protein function from a number of partner resources, giving an overview of the families to which a protein belongs and the domains and sites it contains. Moreover, the Kyoto Encyclopedia of Gene and Genomes (KEGG) (Kanehisa and Goto 2000) is a database specialized for categorizing the metabolic pathways according to KEGG Orthology (KO) identifier, and it is useful for pathway loss and gain analysis across species. Remarkably, the detection of orthologs is becoming much more important with the rapid progress in genome sequencing. OrthoFinder is a fast, accurate and comprehensive platform for comparative genomics (Emms and Kelly 2015). It mainly identifies orthogroup which is the set of genes that are descended from a single gene in the last common ancestor of all the species being considered, but there are also options for inferring a rooted species tree of the species being analyzed and mapping gene duplication events from gene trees to branches in the species tree. Alternatively, OrthoMCL (Li *et al.* 2003) is a genome-scale algorithm for grouping orthologous protein sequences. It provides not only groups shared by two or more species/genomes but also groups representing species-specific expanded gene families.

2.5 Perspectives

NGS and TGS technologies have made it quite easy to obtain large quantities of DNA-Seq data from green algae. Therefore, it is tempting to have pipelines detailing the installation of tools, databases and comparative genomics frameworks in large-scale genome projects. Here, the genome assembly protocol and annotation pipelines for the UWO241 genome were described in a step-by-step user-guide-like manner. This protocol definitely cannot cover everything, but it can introduce the bioinformatic methods used in eukaryotic nuclear genomics, enabling a user to gain familiarity with the basic analysis steps. The chapter summarized the necessary steps, which is also publicly available at GitHub website (https://github.com/zx0223winner/Eukaryotic-genome-project). The link detailed bioinformatic tools for data sets processing as well as some custom-made scripts and command lines used in Python and Unix platforms. Remarkably, steps included sample collection, reads quality correction, *de novo* assembly, gap closing, scaffolding, genome assembly assessment, transcriptome assembly, genome masking, structural annotation, gene models training, BLAST annotation, functional annotation, genome annotation

assessment and comparative genomic analysis. Although the technical aspects of genome tools are evolving very quickly, it is my hope that this user guide will provide a comprehensive bioinformatics foundation for future genome projects specifically for those researchers who have diverse backgrounds but no prior experience in programming.

2.6 References

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics. Retrieved from https://www.bioinformatics.babraham.ac.uk.

Apweiler, R., A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez and M. Magrane (2004). UniProt: the universal protein knowledgebase. Nucleic Acids Research 32: D115-D119.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham and A. D. Prjibelski (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19: 455-477.

Bao, W., K. K. Kojima and O. Kohany (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6: 11.

Boeckmann, B., A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'donovan and I. Phan (2003). The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research 31: 365-370.

Boetzer, M., C. V. Henkel, H. J. Jansen, D. Butler and W. Pirovano (2010). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27: 578-579.

Buchfink, B., C. Xie and D. H. Huson (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods 12: 59-60.

Callaway, E. (2017). Small group scoops international effort to sequence huge wheat genome. Nature News. Retrieved from https://www.nature.com/news/small-group-scoops-international-effort-to-sequence-huge-wheat-genome-1.22924.

Cantarel, B. L., I. Korf, S. M. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. S. Alvarado and M. Yandell (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research 18: 188-196.

De Wit, P., M. H. Pespeni, J. T. Ladner, D. J. Barshis, F. Seneca, H. Jaris, N. O. Therkildsen, M. Morikawa and S. R. Palumbi (2012). The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. Molecular Ecology Resources 12: 1058-1067.

Del Angel, V. D., E. Hjerde, L. Sterck, S. Capella-Gutierrez, C. Notredame, O. V. Pettersson, J. Amselem, L. Bouri, S. Bocs and C. Klopp (2018). Ten steps to get started in Genome Assembly and Annotation. F1000Research 7: 1-25.

Emms, D. M. and S. Kelly (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16: 1-14.

English, A. C., S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, D. M. Muzny, J. G. Reid and K. C. Worley (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PloS One 7: e47768.

Foissac, S., J. Gouzy, S. Rombauts, C. Mathé, J. Amselem, L. Sterck, Y. V. de Peer, P. Rouzé and T. Schiex (2008). Genome annotation in plants and fungi: EuGene as a model platform. Current Bioinformatics 3: 87-97.

Gurevich, A., V. Saveliev, N. Vyahhi and G. Tesler (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072-1075.

Haas, B. J., A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith Jr, L. I. Hannick, R. Maiti, C. M. Ronning, D. B. Rusch and C. D. Town (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research 31: 5654-5666.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li and M. Lieber (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8: 1494-1512.

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

Henson, J., G. Tischler and Z. Ning (2012). Next-generation sequencing and large genome assemblies. Pharmacogenomics 13: 901-915.

Hoff, K. J., S. Lange, A. Lomsadze, M. Borodovsky and M. Stanke (2016). BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32: 767-769.

Kanehisa, M. and S. Goto (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28: 27-30.

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. Genome Research 12: 656-664.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman and A. M. Phillippy (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Research 27: 722-736.

Lee, H., J. Gurtowski, S. Yoo, M. Nattestad, S. Marcus, S. Goodwin, W. R. McCombie and M. Schatz (2016). Third-generation sequencing and the future of genomics. BioRxiv: 048603.

Li, L., C. J. Stoeckert and D. S. Roos (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 13: 2178-2189.

Li, R., W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang, Q. Cai, B. Li and Y. Bai (2010). The sequence and *de novo* assembly of the giant panda genome. Nature 463: 311-317.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan and Y. Liu (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience 1: 1-6.

Pellicer, J., O. Hidalgo, S. Dodsworth and I. J. Leitch (2018). Genome size diversity and its impact on the evolution of land plants. Genes 9: 1-14.

Pflug, J. M., V. R. Holmes, C. Burrus, J. S. Johnston and D. R. Maddison (2020). Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). G3: Genes, Genomes, Genetics 10: 3047-3060.

Price, A. L., N. C. Jones and P. A. Pevzner (2005). *De novo* identification of repeat families in large genomes. Bioinformatics 21: 351-358.

Pruitt, K. D., T. Tatusova and D. R. Maglott (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 33: D501-D504.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler and R. Lopez (2005). InterProScan: protein domains identifier. Nucleic Acids Research 33: 116-120.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva and E. M. Zdobnov (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210-3212.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones and I. Birol (2009). ABySS: a parallel assembler for short read sequence data. Genome Research 19: 1117-1123.

Slater, G. S. C. and E. Birney (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6: 1-11.

Smit, A. F. and R. Hubley (2008). RepeatModeler Open-1.0. Retrieved from http://www.repeatmasker.org.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack and B. Morgenstern (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Research 34: W435-W439.

Tarailo-Graovac, M. and N. Chen (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics 4: 10-14.

Tørresen, O. K., B. Star, P. Mier, M. A. Andrade-Navarro, A. Bateman, P. Jarnot, A. Gruca, M. Grynberg, A. V. Kajava and V. J. Promponas (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Research 47: 10994-11006.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman and S. K. Young (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One 9: e112963.

Yandell, M. and D. Ence (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics 13: 329-342.

Zdobnov, E. M., F. Tegenfeldt, D. Kuznetsov, R. M. Waterhouse, F. A. Simao, P. Ioannidis, M. Seppey, A. Loetscher and E. V. Kriventseva (2017). OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Research 45: D744-D749.

Zimin, A. V., G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg and J. A. Yorke (2013). The MaSuRCA genome assembler. Bioinformatics 29: 2669-2677.

Zimin, A. V., D. Puiu, M.-C. Luo, T. Zhu, S. Koren, G. Marçais, J. A. Yorke, J. Dvořák and S. L. Salzberg (2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Research 27: 787-792.

Chapter 3

3 The Nuclear Draft Genome of the Antarctic Psychrophilic Green Alga *Chlamydomonas* sp. UWO241

This chapter was adapted from the publication entitled "Draft genome sequence of the Antarctic green alga *Chlamydomonas* sp. UWO241" published on iScience in 2021 by X. Zhang, M. Cvetkovska, R. Morgan-Kiss, N. P. A. Hüner and D. R. Smith (Zhang *et al.* 2021).

3.1 Introduction

The permanently ice-covered lake (Lake Bonney) in the McMurdo Dry Valleys of Victoria Land, Antarctica (Neale and Priscu 1995), harbors the psychrophile *Chlamydomonas* sp. UWO241 (hereafter UWO241). Molecular and genetic analyses of UWO241 have already revealed some peculiar features, including its apparent inability to perform traditional photosynthetic state transitions or grow under red light (Morgan-Kiss *et al.* 2006). Furthermore, UWO241 has recently found to have a functional chlorophyll biosynthesis pathway that lost light-independent protochlorophyllide reductase (DPOR) and is solely dependent on light-dependent protochlorophyllide oxidoreductase (LPOR) for the enzymatic reduction of protochlorophyllide (Cvetkovska *et al.* 2019). Moreover, investigations of the UWO241 transcriptome suggest the absence of the upregulation of genes encoding heat-shock proteins (HSPs) (Possmayer 2018). Notably, it appears that UWO241 has two nearly identical copies of the ferredoxin gene, and accumulates large amounts of functional ferredoxin protein, which reveals an adaptation to cold environments (Cvetkovska *et al.* 2018). Given all the previous assessments, UWO241, a psychrophilic alga, has been widely explored and has generated particular interest with respect to the psychrophilic and mesophilic species in its order.

Without a comprehensive genomic framework, the broader application of UWO241 as a model system for cold adaptation research is severely impeded. Fortunately, there are many mesophilic algal species and few psychrophilic algae in the order Chlamydomonadales. For instance, *Chlamydomonas reinhardtii* and *Chlamydomonas* sp. ICE-L are excellent

comparison targets for the investigations of psychrophilic chlamydomonads (Cvetkovska *et al.* 2017; Zhang *et al.* 2020).

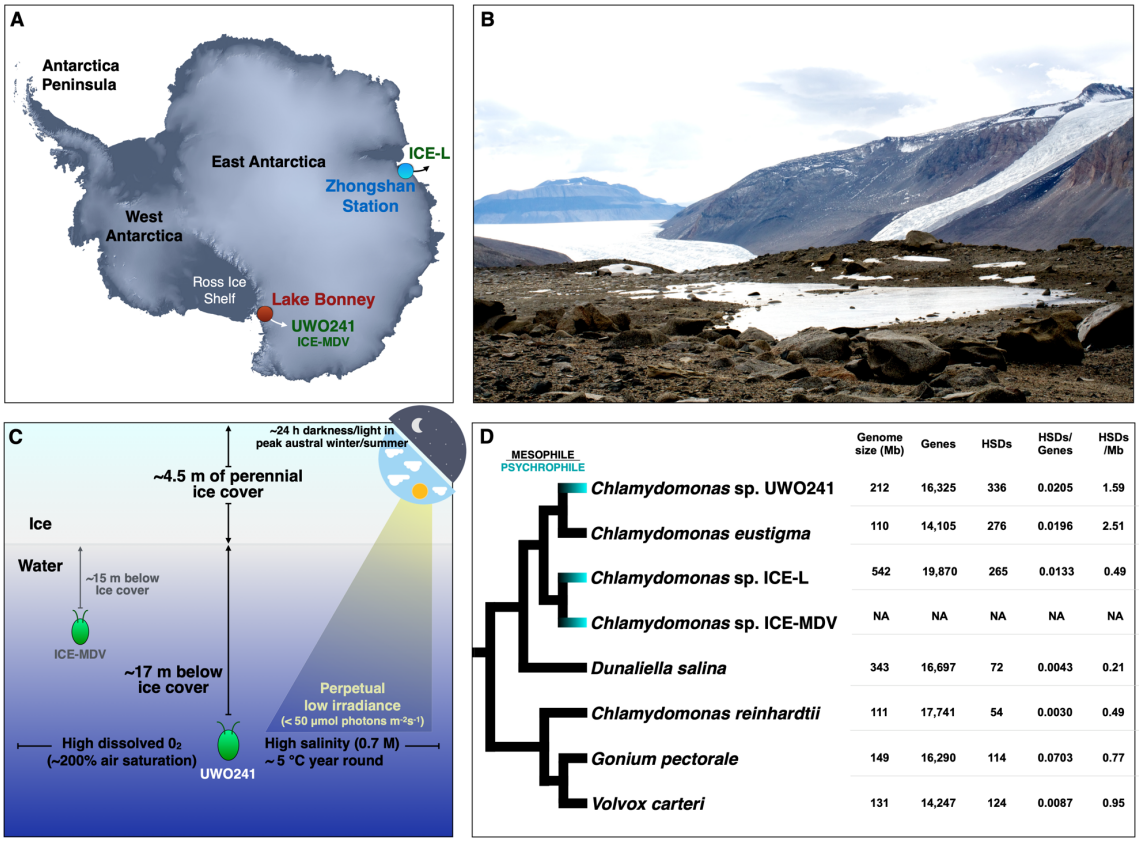
3.2 Results and Discussions

3.2.1 Habitat, Taxonomic Position, and Physiological Features of the Psychrophilic Green Alga UWO241.

The past decade has brought draft nuclear genomes for >25 different green algal species, with especially strong sampling from the order Chlamydomonadales (Chlorophyceae) (Figure 3 and Figure 9D). The psychrophile *Chlamydomonas* sp. UWO241, which was isolated 17 m below the bottom of the permanent ice surface of Lake Bonney in the McMurdo Dry Valleys of Victoria Land, Antarctica (Neale and Priscu 1995) (Figure 9A, B, C), is emerging as a model for studying cold-adaptation. Until recently, UWO241 was considered to be a lineage within the Moewusinia clade of the Chlamydomonadales (Possmayer *et al.* 2016) (Figure 3 and Figure 9D). The phylogeny has also highlighted two other psychrophiles, *Chlamydomonas* sp. ICE-L and *Chlamydomonas nivalis*; however, only the ICE-L genome has been completely sequenced recently (Zhang *et al.* 2020). Remarkably, almost one-third of known photopsychrophiles belong to the green algal order Chlamydomonadales, which is found in the Chlorophycean class of Chlorophyta (Cvetkovska *et al.* 2017). Indeed, many chlamydomonadalean algae inhabiting polar and alpine environments are drought resistant, and they can tolerate high levels of UV radiation and low-nutrient stress (Quesada and Vincent 2012; Umen and Olson 2012), which makes them ideal models for studying adaptation to extreme environments. What immediately stands out for the UWO241 genome as compared to other available green algal nuclear DNAs (nucDNAs) is its relatively large size (twice that of *C. reinhardtii*), record-setting intron density, and high repeat content, outdone only by that of ICE-L (~64% repeats) (Zhang *et al.* 2020). However, close inspection of the UWO241 coding regions uncovered something very unique: widespread gene duplication to a degree unmatched in any chlorophyte studied to date.

Although the green algae ICE-L and UWO241 are closely related, they originate from very different Antarctic environments. ICE-L was isolated from open sea ice off of Zhongshan

Station whereas UWO241 is from Lake Bonney in the McMurdo Dry Valleys, which is ~2000 km away from Zhongshan Station (Zhang *et al.* 2020). Lake Bonney is permanently covered in ~5 m of ice and UWO241 lives ~17 m below the ice where the temperature is around 5 °C year-round (Neale and Priscu 1995). Additionally, UWO241 is surprisingly resilient, persisting in an environment that not only is a perpetually cold environment but also has a high saline content (700 mM) and low irradiance (Figure 2). UWO241 possesses an unusual photosynthetic apparatus , tailored to work best at 8 °C, but it presents rates of photosynthesis relatively similar to those of *C. reinhardtii* at 25-35 °C (Cvetkovska *et al.* 2017). In addition to withstanding constant low temperatures of approximately 5 °C year-round, UWO241 is exposed to perpetual shading (5 μmol photons m^{-2} s^{-1} during midday in summer) and seasonal extremes in photoperiod (e.g., 24 h of light during the peak summer), which is enriched in the blue-green wavelengths of the visible spectrum (450-550 nm). Lake Bonney is also phosphorus limited and contains high levels of dissolved oxygen (200% saturation). In UWO241, many unique cellular and physiological features have been evolved to handle with the extreme conditions of Lake Bonney, such as high PSI cyclic electron transport, the inability to grow under red light and a lack of state transitions (Morgan-Kiss *et al.* 2006; Kalra *et al.* 2020).



	Genome size (Mb)	Genes	HSDs	HSDs/ Genes	HSDs /Mb
Chlamydomonas sp. UWO241	212	16,325	336	0.0205	1.59
Chlamydomonas eustigma	110	14,105	276	0.0196	2.51
Chlamydomonas sp. ICE-L	542	19,870	265	0.0133	0.49
Chlamydomonas sp. ICE-MDV	NA	NA	NA	NA	NA
Dunaliella salina	343	16,697	72	0.0043	0.21
Chlamydomonas reinhardtii	111	17,741	54	0.0030	0.49
Gonium pectorale	149	16,290	114	0.0703	0.77
Volvox carteri	131	14,247	124	0.0087	0.95

Figure 9: *Chlamydomonas* sp. UWO241.

(A) Origins of isolation of UWO241 and ICE-MDV (from Lake Bonney) as well as ICE-L (from sea ice off of Zhongshan Station); image from NASA Earth Observatory. (B) Photograph of Lake Bonney (Wikimedia-Commons 2020). (C) Simplified diagram showing underwater conditions of Lake Bonney. (D) Tree of various chlamydomonadalean algae and their nuclear genome statistics; branching order based on previous phylogenetic analyses; HSDs inferring the number of highly similar duplicates (Nakada *et al.* 2008; Possmayer *et al.* 2016; Zhang *et al.* 2020).

3.2.2 Characteristics of *Chlamydomonas* sp. UWO241

The haploid nuclear genome of UWO241 was assembled *de novo* using a combination of long-read PacBio (~16.5 Gb) and short-read Illumina (~40 Gb) data, resulting in 2,458 scaffolds (N50 = 375.9 kb) with an accumulative length of 211.6 Mb (%GC = 60.6) (Figure 9D and Figure 10). This length is consistent with flow cytometry and *k*-mer spectral

analysis of UWO241, which predicted an overall genome size of ~230 Mb (Figure 10A, B). In total, 16,325 protein-coding genes were annotated (all supported by transcriptomic data), capturing ~85% of the Chlorophyte Benchmarking Universal Single-Copy Orthologs (BUSCO) datasets (Figure 10C), indicating a high level of gene-region completeness. The UWO241 genome is rich in functional RNAs (630 tRNAs and 480 rRNAs) as well as noncoding DNA (~87%), having the highest average intron density yet observed from a green alga (~10 introns/gene; avg. intron length 0.9 kb). The intergenic regions abound with repeats, accounting for ~104 Mb (~49%) of the total assembly length, ~70 Mb of which are represented by transposable elements (TEs) (discussed in Chapter 4).

Although utilizing a hybrid of long-read single-molecule, real-time (SMRT) sequencing (Pacific Biosciences) for *de novo* assembly and short-read Illumina HiSeq DNA sequencing (Table 1), I have produced scaffold-level genome assemblies for UWO241. However, multiple approaches have been utilized to improve the genome assembly. As displayed in Table 2, the hybrid-read assembler performs better than the single-read assemblers. By using the Illumina reads and PacBio reads alone, the single-read assemblers yield assembly sizes of only 157 Mb and 150 Mb, accounting for 68.2% and 65.2% of the estimated genome size. However, a hybrid-read assembler taking advantage of both read types yields as much as 212 Mb, which covers 92% of the estimated genome size. Additionally, the contigs assembled with the single-read assembler appear more fragmented than those assembled with the hybrid-read assembler. The contig-level N50/L50 values in Illumina reads and PacBio reads are 3,188 bp/14,804 and 69,116 bp/635, respectively. However, via the hybrid method, the scaffold-level metrices are much more contiguous, with an N50/L50 of 375,862 bp/165. While this model genome could be substantially improved by additional sequencing effort, it is my goal to obtain the best genome assembly to date with the current data available. Therefore, the scaffolds from the hybrid-read assembler are advanced by filling the gaps and polishing the mismatches. Taken together, ~16.5 Gb of PacBio reads and ~40 Gb of Illumina reads are assembled into 2,464 scaffolds (211.6 Mb), covering ~92% of the estimated haploid genome (Table 4). The genome assembly is highly contiguous, with N50 of 375,902 bp and L50 of 165.

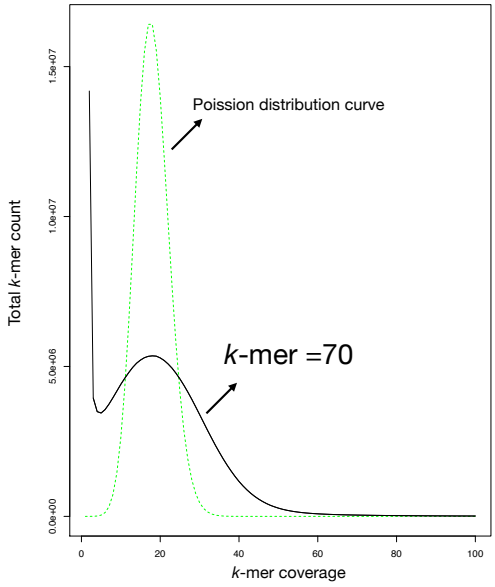
A

Cells measured by Flow cytometry	Nuclear DNA content (1 pg = 978 Mb)	Estimated genome size (Mb)	Standard error of mean genome size (Mb)
Mixed cell culture of UWO241	0.53 pg/2C	259	2
Large cell culture of UWO241	0.53 pg/2C	259	1
Medium cell culture of UWO241	0.49 pg/2C	239	1
Small cell culture of UWO241	0.50 pg/2C	245	2
Mean	0.51 pg/2C	250	1.5

B (a)

k-mer length	65	70	75
Total *k*-mer count	4574846901	3783302136	3039573138
Coverage	23	18	14
Genome size (Mb)	198.9	210.2	217.1

B (b)



C

Genome statistics	***Chlamydomonas*. sp UWO241**
Genome size (Mb)	211.6
Scaffold N50 (Mb)/L50	0.37/165
GC (%)	60.61
Number of protein coding genes	16325
Gene density (genes/Mb)	77.2
mean gene length	8058
Average exon length (bp)	1563
% of genome covered by genes	62.8
% of genome covered by CDS	12.2
Average intron per gene	10.11
Average intron length (bp)	934
Genome completness by BUSCO	85%

Figure 10: Summary of statistics of the UWO241 genome.

(A) The estimated genome size measured via flow cytometry. (B) (a) The estimated genome size via different *k*-mer lengths. (b) The *k*-mer spectrum of *Chlamydomonas* sp. UWO241. The X-axis is the number of times a given *k*-mer was observed in the UWO241 sequencing data. The Y-axis is the total number of *k*-mers with a given *k*-mer coverage. (C) Nuclear genome statistics of UWO241.

The assembled genome size varies across the seven species (Figure 11 and Table 7), ranging from 111.1 Mb in *C. reinhardtii* to 541.8 Mb in ICE-L. Surprisingly, the genome UWO241 is the third largest across the species, as shown in the Table 6, which is nearly double the genome size of *C. reinhardtii*. It is not uncommon for plants surviving in extreme environments to accumulate redundancy, resulting in the novel gene sets and genome size expansion (Qian and Zhang 2014; Panchy *et al.* 2016; Zhang *et al.* 2020). *Dunaliella salina* as a halophile is able to tolerate the high-salt conditions, similar to

UWO241 and ICE-L, which are psychrophiles surviving in both cold and salty environments. Furthermore, it is likely that the differences of metrics (e.g., intron length and intergenic region length) are related to the varied living environment of genomes. As interpreted in Figure 11, the intron length (yellow) and intergenic region length (green) contribute to the majority of the genome size for UWO241, *D. salina* and ICE-L. They exhibit accumulative sizes of 110.69 Mb,158.1 Mb and 209.0 Mb for intron length, and 74.67 Mb, 161.88 Mb and 303.8 Mb for accumulative intergenic region length, respectively. While the other fresh-water algae such as *C. reinhardtii* and *Volvox carteri* have smaller genome sizes and intron lengths. The genome-wide GC content ranged from the highest (64.5%) in *Gonium pectorale* to the lowest (49.1%) in *D. salina.* UWO241 has a GC-rich genome with GC content of 60.6% recorded. It is assumed that the GC content diversity is critical for gene and organismal evolution, and plants tend to evolve the contrasting GC contents to survive in different environments (Šmarda *et al.* 2014). A research team has demonstrated that the nucleotide composition landscapes in monocots are shaped by the GC-biased gene conversion (Singh *et al.* 2016).

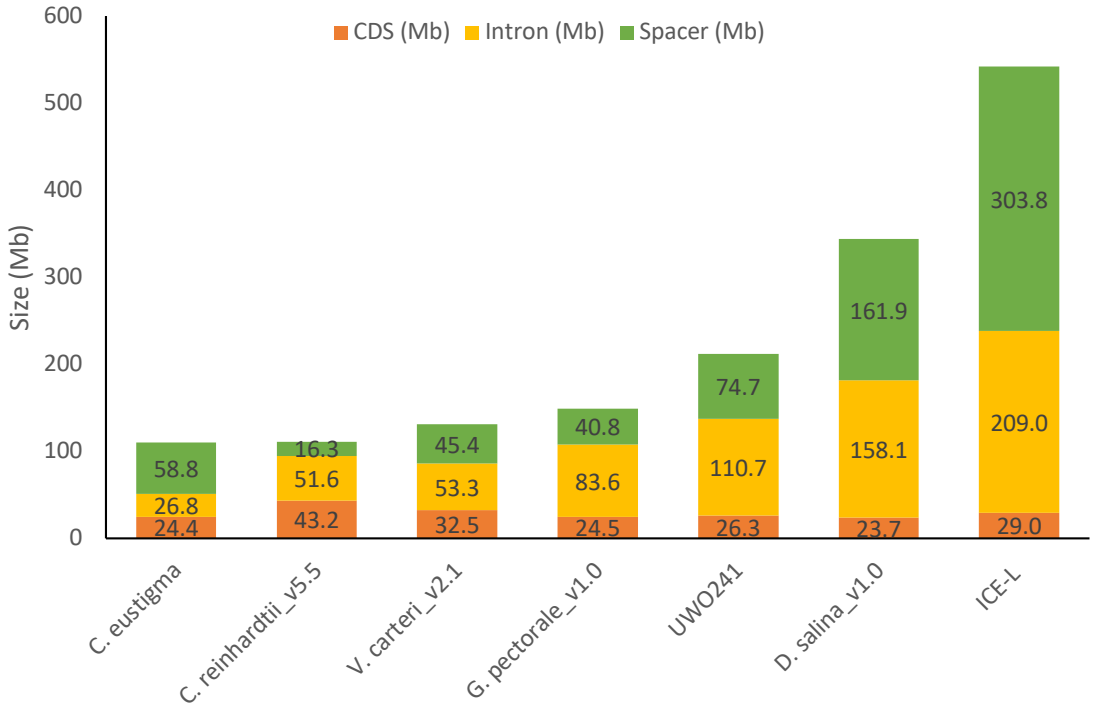


Figure 11: Genome size distribution of UWO241 and its closely green algal relatives.

Recently, it was reported that retrotransposon proliferation resulted in the large genome of ICE-L (Zhang *et al.* 2020). ICE-L was found 63.78% of the ICE-L genome assembly lengths (345.23 Mb) to be repeat regions. Transposable elements (TEs) accounted for 40.67% of the ICE-L genome assembly (220.37 Mb), and long terminal repeat retrotransposons (LTR-RTs) were the most dominant type of TEs, representing 23.32% of the assembly (126.36 Mb) (Zhang *et al.* 2020). Similarly, to decipher the reasons for the large genome size of UWO241, the gene content in the intronic and intergenic regions was explored. As presented in Table 5, UWO241 harbors approximately 104 Mb of repeat regions, accounting for 49.25% of the whole genome. Specifically, these repetitive regions can largely be attributed to the large numbers of LINEs and simple repeats and a myriad of unclassified elements. LINEs are TEs that occupy 20.3 Mb in UWO241, accounting for 9.6% of the genome. Across the unicellular species, the percentage of LINEs in UWO241 and ICE-L are at the top (Supplementary Information: Table 8). It has been reported that RNA-mediated retrotransposons might play an important role in organismal diversity and adaptation (Casola and Betrán 2017). Thus, it is enticing to link many unique features of the UWO241 genome, such as high proportions of gene duplications, with these unique LINE-rich patterns. Simple repeats are also presented at higher levels in UWO241 (Table 5), but this is not uncommon due to the complexity of the genome. Simple repeats are usually defined as the duplications of simple sets of DNA bases (typically 1-5 bp), such as A, CA, and CGG (Smit *et al.* 2015). A larger proportion of unclassified elements are observed in UWO241, partly due to the divergence of the repetitive sequence patterns from those of *C. reinhardtii*. Because a curated repeat library for *C. reinhardtii* directly contributes to masking the repeats in related species, unclassified categories can be minimized in the future via the increasing sequencing number of diverse closely related species. Although not the main focus of the study, future manual curation of the most abundant TE families across species could benefit the repeat masking and genome annotation of related species and shed light on the evolutionary processes shaping genomes (Hubley *et al.* 2016).

Prior to identifying the function of a coding sequence, scaffolds containing organellar DNA are filtered. The remaining 2458 scaffolds from the nuclear genome are used for gene model construction. Finally, I have predicted 16,325 nuclear protein-coding genes, all of

which are supported by RNA-Seq transcripts. The assembly completeness is explored further by the genome-mode BUSCO scores (Supplementary Information: Table 9), the metrics of UWO241(~85%) compare favorably to those of the existing model assemblies from 64.6% to 95.9%. Analyses of genome completeness indicate that ~76% of the conserved Chlorophyta genes (Chlorophyta_odb10) are annotated and complete in the transcriptome data. The BUSCO scores in protein mode suggest the higher gene duplication levels in UWO241 (456, 21.0%) and the ICE-L (240, 11.1%) across the green algal species (Supplementary Information: Table 9). Some of the genes were identified as fragmented relative to the genome assembly mode of BUSCO. UWO241 and ICE-L both have higher levels of missing data, which might be due to the expanded size of their genomes. As shown in the Table 7, consistent with the genome size increases from 111.1 Mb to 541.8 Mb, gene density shows the opposite trend, ranging from 159.7 genes/Mb to 36.7 genes/Mb, with the exception of *Gonium pectorale.* This might result from the larger gene number predicted in *G. pectorale*.

Together with the benefits of comparative genomics, the BLASTP search against the National Center for Biotechnology Information nonredundant (NCBI-nr) database (release 201902) shows that 60% of UWO241 proteins significantly (E-value < 1e-5, ≥ 80% protein length) matched those of Volvocales (*C. reinhardtii*, *G. pectorale*, and *V. carteri*), whereas 21.8% shows no significant similarity to any known proteins.

Table 4: Genome assembly results from different assemblers.

	Single-read assembler		Hybrid-read assembler
Assembler	SPAdes (Illumina reads)	Canu (PacBio reads)	MaSuRCA (Illumina and PacBio reads)
No. of total contigs	70,273	2,858	2,464
No. of contigs (≥ 1000 bp)	49,313	2,858	2,463
Total length (Mb)	157	150	212
N50 (bp)	3,188	69,116	375,902
L50	14,804	635	165
GC (%)	60.3	60.9	60.6

Table 5: Summary of repeats being masked in *Chlamydomonas* sp. UWO241.

Repeats category	No. of elements	Length occupied (bp)	Percentage of sequence
LINEs	40,919	20,307,308	9.60%
LTR elements	548	396,476	0.19%
DNA elements	791	352,579	0.17%
Unclassified	199,177	50,349,480	23.79%
Simple repeats	417,034	31,185,848	14.74%
Low complexity	27,339	1,977,216	0.93%
Total	685,808	104,568,907	49.42%

3.2.3 The General Features of Comparative Genomics Analysis in UWO241 and its Closely Green Algal Relatives

Without comparative genomic analysis, investigations of the unique patterns in UWO241 would have been severely impeded. Given the highly contiguous genome assembly and well-annotated genome annotation, I performed an array of comparative genomic analyses of UWO241 with other sequenced mesophilic and psychrophilic chlamydomonadaleans, including *C. reinhardtii, V. carteri*, *G. pectorale, D. salina*, *C. eustigma* and ICE-L (Table 6). As previouly discussed, the genome size of UWO241 genome is double that of the model alga *C. reinhardtii*. The predicted gene number is roughly the same to other mesophilic algal species, and the number of gene families is slightly lower compared to that in other Chlorophyceae, including the volvocine algae (*Chlamydomonas*, *Gonium*, and *Volvox*). The GC percentages of UWO241, *C. reinhardtii* and *G. pectorale* are above 60%, while those of the ICE-L, *D. salina* and *V. carteri* are 49.2%, 49.1% and 56.1%, respectively. The average intron length of UWO241 (934 bp) and ICE-L (1951.5 bp) are larger than that of the other species (279 bp in *C. reinhardtii*, 399 bp in *V. carteri*, 407 bp in *G. pectorale*).

While UWO241 and ICE-L exhibit a lower gene density, the intron length and the intergenic region length are greater than those of the other chlamydomonadalean species.

Presumably, this is in part due to the highly repetitive elements enriched in these regions, such as simple repeats and TE elements, which both present a larger proportion in the UWO241 (14.74% and 43.67% of the genome, respectively) and ICE-L (8.13% and 45.01% of the genome, respectively). Indeed, it is widely believed that TE elements play a role in shaping the genome by expanding the genome size and gene structure (Casola and Betrán 2017). As displayed in Table 7, there are an average of 10.1 introns per gene with an average intron length of 934 bp in UWO241, which are larger than the corresponding numbers in *C. reinhardtii* (7.4/279.2 bp), *G. pectorale* (6.5/407.0 bp), *V. carteri* (6.3/399.5 bp). Possibly, at least one driving force is attributable to the higher level of introns in the UWO241 genome. For example, the intronless genes originating from bacteria or archaea are acquired by the host via HGT events. However, the horizontally transferred genes such as IBP genes are likely to acquire introns due to selection pressure (Raymond and Kim 2012). As reported in the psychrophilic diatom *Fragilariopsis cylindrus*, there are 11 unique IBP isoforms, most of which have no introns, while a few have single, short introns near the 3' end (Mock *et al.* 2017). The same is observed in UWO241, where there are ≥ 37 IBPs, 27 of which contain introns, suggesting the evolutionary timeline among these horizontally transferred genes. Additionally, the introns could also accumulate because of TE elements, since retrocopies (retrogenes) generated from the RNA-mediated retrotransposition might acquire novel introns throughout intronization from the parental coding sequence (Casola and Betrán 2017). A large number of genes with retrocopies patterns are observed in the UWO241 genome, suggesting the driving force for the enrichment of introns in UWO241 (Appendix A: Table S3). It should be noted that the intron number might be underestimated because retrocopies can undergo erosion and yield retropseudogenes due to a lack of regulatory regions (Kubiak and Makałowska 2017). Therefore, I was very careful when exploring those retrocopies in the UWO241 genome, and only the functional and expressed gene copies were selected. Furthermore, in my attempt to understand whether the phenomenon of large scale retrocopies is unique to UWO241, I most strikingly found that many Pfam domains of UWO241 function as reverse transcriptases (RTs) compared to other algae. Specifically, there are 77 autonomous virus-like LTR retrotransposons and 324 non-LTR retrotransposons (e.g., LINE1) in the UWO241 genome (Appendix A: Table S3). Notably, some RNA-mediated TE elements

are also detected in the intronic and intergenic regions of the UWO241 genome. Preliminary findings show that the photosynthetic ferredoxin gene from UWO241 has a much higher intron content than its *C. reinhardtii* counterparts. Moreover, unlike in *C. reinhardtii*, the UWO241 ferredoxin gene has highly similar duplicates (Cvetkovska *et al.* 2018). Although the ferredoxin gene is not found in the HSDs list of ICE-L, UWO241 (336) and ICE-L (265) both have large size of HSDs candidates (Figure 9D). Many of the HSDs genes from the two psychrophiles encode the same functions such as antenna proteins, ribosomal proteins and histones (Appendix A: Table S6).

Table 6: Species list and genome versions used for annotation and comparative genomic analysis.

Species	**Source**	**References**
C. reinhardtii v5.5	JGI 5.5 (Phytozome 12.1)	(Merchant *et al.* 2007)
V. carteri v2.1	JGI 2.1 (Phtyzome 12.1)	(Prochnik *et al.* 2010)
G. pectorale v1.0	GenBank (GCA_001584585.1)	(Hanschen *et al.* 2016)
D. salina v1.0	JGI 1.0 (Phytozome 12.1)	(Polle *et al.* 2017)
C. eustigma (Acidophile)	GenBank (GCA_002335675.1)	(Hirooka *et al.* 2017)
Chlamydomonas sp. ICE-L (Psychrophile)	GenBank (GCA_013435795.1)	(Zhang *et al.* 2020)

Table 7: Genome characteristics comparison of between UWO241 and closely related green algae.

Genome statistics	*C. reinhardtii*_v5.5	*V. carteri*_v2.1	*G. pectorale*_v1.0	UWO241	*D. salina*_v1.0	*C. eustigma*	ICE-L
Genome size (Mb)	111.1	131.1	148.8	211.6	343.7	110	541.8
Scaffold N50 (Mb)/L50	7.80/7	2.59/15	1.27/30	0.37/165	0.35/310	0.46/519	19.23/946
GC (%)	64.1	56.1	64.5	60.6	49.1	50.6	49.2
Number of protein coding genes	17,741	14,247	16,290/17,984*	16,325**	16,697	14,105	19,870
Gene density (genes/Mb)	159.7	108.6	109.5/120.9*	77.2**	48.5	128.2	36.7
Average intron per gene	7.4	6.3	6.5	10.1	NA	NA	NA
Average intron length (bp)	279.2	399.5	407.0	934.0	NA	259.8	1951.5

***** Although the genome paper of *G. pectorale* reported 17,984 genes they found (Hanschen *et al.* 2016), the genome assembly from NCBI source was detected 16,290.

****** The draft genome of UWO241 was detected 16,325 genes supported by transcriptomic data (Zhang *et al.* 2021), while the NCBI source filtered the dataset to 16,018 genes for downloading.

3.3 Conclusions

Utilizing the highly contiguous nuclear assembly and well-annotated genomes of a psychrophilic green alga, namely, UWO241, I have presented the first nucleotide-level comparative genomic framework for this important model organism. I explored some of key questions, such as the following: How to improve the nuclear genome assembly for UWO241 by using both NGS and TGS sequencing reads? How to optimize the training sets to have this genome been accurately and thoroughly annotated? How to decipher the data in a comparative genomic framework to better understand the evolution of psychrophily? Specifically, I, first, developed an assembly pipeline for processing high-throughput DNA sequencing reads into genomic contigs. These contigs, alongside RNA-

Seq data, are fed into an annotation pipeline, which is designed based on state-of-the-art eukaryotic bioinformatic gene-profiling software. Last but not least, computational analyses are carried out on an in-house computer as well as a supercomputing network, which yielded the draft nuclear genome (~212 Mb, 16,325 protein-coding genes) sequence of the psychrophilic green algae UWO241. This comparative genomic framework across psychrophilic chlamydomonads is able to be conducted via comparison to the mesophilic and psychrophilic relatives *C. reinhardtii*, *V. carteri*, *D. salina* and ICE-L, among others. I hope that this work will aid in studies of other psychrophiles and provide insights into the evolution of psychrophily.

3.4 Methods and Experiments

3.4.1 Strains and Growth Conditions

UWO241 is available from the National Center for Marine Algae and Microbiota (NCMA; strain CCMP 1619). The strain used for genome sequencing is the original isolate, obtained directly from Priscu (Pocock *et al.* 2004). UWO241 was grown axenically in Bold's Basal Medium (BBM) supplemented with 70 mM NaCl. Cultures were grown at 5 °C in 3-layer BD FalconTM Multi Flasks with agitation at a continuous light of 150 μmol photons m^{-2} s^{-1} measured with a quantum sensor attached to a radiometer (Model LI-189; Li-Cor, Lincoln, NE, USA). Cultures were grown to mid-log phase prior to harvesting.

3.4.2 DNA and RNA Extraction and Library Construction

Genomic DNA (gDNA) for Illumina HiSeq2000 sequencing was extracted using the Qiagen Plant DNeasy Maxi Kit (Qiagen) following manufacturer's instructions. UWO241 was harvested by centrifugation (6000 g, 5 min, 4 °C), flash-frozen in liquid nitrogen, and stored at 80 °C. The DNA was purified by ethanol precipitation using standard methods and resuspended in 10 mM Tris, pH 7.5. DNA quality was monitored using wavelength absorbance scan and electrophoresis on a 1% (w/v) TBE agarose gel.

For single-molecule, real-time (SMRT) sequencing (Pacific Biosciences, Menlo Park, CA, USA), gDNA was extracted using a modified CTAB protocol. In short, cell pellets were resuspended in 1 ml of lysis buffer (50 mM Tris-HCl, pH 8.0; 200 mM NaCl, 20 mM

EDTA, 2% (w/v) SDS, 20 mg/ml Proteinase K) and mixed by inversion. Equal volume of pre-heated CTAB buffer (2% (w/v) CTAB, 1.4 M sodium chloride, 20 mM EDTA, 100 mM Tris, 2% (w/v) polyvinylpyrrolidone (M.W. 40000), 1% (v/v) β-mercaptoethanol, pH 8.0) was added, and the cells were incubated at 65 °C for 30 min, followed by centrifugation (14,000 g, 5 min) to remove the insoluble materials. The supernatant was treated with RNase A (100 mg/ml) for 30 min at 37 °C, and nucleic acids were extracted 2x with equal volume of phenol:chloroform:isoamyl alcohol (25:24:1). The extract was centrifuged (16,000 g, 10 min) and nucleic acids were precipitated with 1x volume of ice-cold isopropanol and incubated at -20 °C for 1 hour. The samples were centrifuged (16,000 g, 15 min, 4 °C) and the pellet was washed 3x with ice-cold 70% (v/v) ethanol. DNA was precipitated with 1/10th volume 3M sodium acetate (pH 5.2) and 2 volumes 100% ethanol, samples were incubated at -20 °C for 1hour, centrifuged (16000 g, 4 °C, 30 min), and the resulting DNA pellets washed with 70% (v/v) ethanol. The pellets were air-dried and resuspended in 10 mM Tris (pH 7.8) by incubating them for 24 hours at 4 °C.

Complementary DNA of UWO241 was performed using 125 bp paired-end (PE) reads on an Illumina HiSeq 2500 v4 sequencing platform. Three biological replicate cultures of UWO241 were grown at 15 °C. Algal cells were harvested by centrifugation (6,000 g, 5 min, 4 °C), flash frozen in liquid nitrogen, and stored at -80 °C. RNA was isolated using a modified CTAB protocol (Possmayer *et al.* 2016) and sequenced at the Génome Québec Innovation Centre (Montreal, QC, Canada). Total RNA was quantified using a NanoDrop Spectrophotometer ND-1000 (NanoDrop Technologies, Inc.) and its integrity was assessed using a 2100 Bioanalyzer (Agilent Technologies). Libraries were generated from 250 ng of total RNA using the TruSeq stranded mRNA Sample Preparation Kit (Illumina), as per manufacturer's recommendations. Libraries were quantified using the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems). Average size fragment was determined using a LabChip GX (PerkinElmer) instrument.

3.4.3 Genome Sequencing

Genomic HiSeq 2000 sequencing was performed at the Princess Margaret Genomics Centre (Toronto, ON, Canada), using 101-cycle PE reads at 100x coverage. DNA was fragmented using a Covaris M220 Focused-Ultrasonicator (Covaris Inc., Woburn, MA,

USA) and libraries were constructed with the TruSeq DNA HT Sample Preparation Kit (FC-121-2003; Illumina, San Diego, CA, USA). PacBio SMRT sequencing was performed by Génome Québec on an RSII instrument, using 19 cells at 81x coverage. 7.5 µg of high-molecular-weight gDNA was sheared using the Covaris g-TUBES (Covaris Inc.). DNA libraries were prepared using the SMRTbell Template Prep Kit 1.0 reagents (Pacific Biosciences). The DNA library was size-selected on a BluePippin system (Sage Science Inc., Beverly, MA, USA) using a cut-off range of 10-50 kb. Complementary DNA of UWO241 for Illumina HiSeq 2500 was sequenced by Génome Québec.

3.4.4 Estimation of Genome Size

The nuclear genome size of UWO241 was estimated using *k*-mer analysis and flow cytometry. Approximately ~30 Gb of high-quality, short-insert reads (250 bp) were used to estimate genome size via the *k*-mer analysis tool Jellyfish (Arumuganathan and Earle 1991). The *k*-mer frequency followed a Poisson distribution. The *k*-mer depth (i.e., mean coverage) was divided by the total *k*-mer number, giving a genome-size estimate of 210Mb (± 10Mb; mean ± standard error) when using a default *k*-mer size 65, 70 and 75. The genome size estimation via *k*-mer is followed through the tutorial with the link (https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/).

Flow cytometry predicted the UWO241 genome size to be 250 Mb (± 2Mb; mean ± standard error), following a modified protocol by Arumuganathan and Earle (Arumuganathan and Earle 1991). Briefly, intact nuclei were suspended in $MgSO_4$ buffer mixed with DNA standards and stained with propidium iodide (PI) in a solution containing DNAase-free RNAase (Arumuganathan and Earle 1991). Fluorescence intensities of the stained nuclei were measured by a flow cytometer. Values for nuclear DNA content were estimated by comparing fluorescence intensities of the nuclei of UWO241 with those of various internal DNA standards, including nuclei from *C. reinhardtii* (0.35 pg/2C), mixed cell culture of UWO241 (0.53 pg/2C), large cell culture of UWO241 (0.53 pg/2C), medium cell culture of UWO241 (0.49 pg/2C) and small cell culture of UWO241 (0.51 pg/2C). Specifically, for flow cytometric analysis, one mL of UWO241 was placed in microfuge tubes and centrifuged for 5 sec. The pellet was suspended by vortexing vigorously in 0.5 mL solution containing 10 mM $MgSO_4.7H_2O$, 50mM KCl, 5 mM Hepes, pH 8.0, 3 mM

dithiothreitol, 0.1 mg / mL propidium iodide, 1.5 mg / mL DNAse free RNAse (Rhoche, Indionapolis, IN) and 0.25% Triton X-100. The suspended nuclei were withdrawn using a pipettor, filtered through 30-µm nylon mesh, and incubated at 37 °C for 30 min before flow-cytometric analysis. Suspensions of sample nuclei was spiked with suspension of standard nuclei (prepared in above solution) and analyzed with a FACScalibur flow cytometer (Becton-Dickinson, San Jose, CA). For each measurement, the propidium iodide fluorescence area signals (FL2-A) from 1000 nuclei were collected and analyzed by CellQuest software (Becton-Dickinson, San Jose, CA) on a Macintosh computer (Dickinson and Dickinson 1998). The mean position of the G0/G1 (nuclei) peak of the sample and the internal standard were determined by CellQuest software. The mean nuclear DNA content of each plant sample, measured in picograms, was based on 1000 scanned nuclei.

3.4.5 Nuclear Genome Assembly

The nuclear genome of UWO241 was assembled *de novo* using Illumina and PacBio SMRT sequencing reads. The Illumina read quality was evaluated using FastQC v0.11.8 (Andrews 2010), and the PacBio sequencing reads were assessed via the error-correction step of Canu v1.7.1 (Koren *et al.* 2017). The hybrid *de novo* assembly was carried out with MaSuRCA v3.3.2 (Zimin *et al.* 2017), using an automatically determined *k*-mer size (i.e., GRAPH_KMER_SIZE = auto), which computes the optimal size based on the read data and GC content; a cgwErrorRate of 0.15; and a KMER_COUNT_THRESHOLD of 1. Scaffolding and gap-filling algorithms were then applied to all hybrid-assembled contigs to extend the length of the assembly and to minimize mismatches. SSPACE v3.0 (Boetzer *et al.* 2010) was used to extend and scaffold pre-assembled contigs by using Illumina PE libraries. GapFiller v2.1.1 (Boetzer and Pirovano 2012) was used to close the gaps (’N’) in the scaffolds by mapping with long PacBio reads. The genome assembly was further polished with highly accurate Illumina reads via Pilon v1.22 (Walker *et al.* 2014). Assemblies of the plastid and mitochondrial genomes were produced independently (Cvetkovska *et al.* 2019). The Illumina HiSeq transcriptomic data were *de novo* assembled via Trinity v2.8.4 (Haas *et al.* 2013). Adapters and low-quality bases were trimmed from

each RNA-seq dataset using Trimmomatic v0.38 (Bolger *et al.* 2014). Genome assembly metrics were generated using QUAST v5.0.0 (Gurevich *et al.* 2013).

3.4.6 *De novo* Repeat Finding and Repeat Masking

A *de novo* repeat library was created with RepeatModeler v1.0.8 (Smit and Hubley 2008), RepeatScout v1.0.5 (Price *et al.* 2005), LTR_FINDER (Xu and Wang 2007), and LTR_retriever (Ou and Jiang 2018) using default parameters. Unknown elements were screened with BLASTX (Altschul *et al.* 1997) (E-value < 1e-5) against UniRef90 database (Suzek *et al.* 2015) (subset Viridiplantae) and removed from the repeat library if necessary. The repeat library of UWO241 was used by RepeatMasker (4.0.7) (rmblastn version 2.2.27+) (Tarailo-Graovac and Chen 2009) to mask the repetitive elements in the assembly, which resulted in 104 Mb (~49 %) of the UWO241 genome being masked. The masked regions were further inspected for overlaps with UWO241 RNA-Seq transcripts via GENEIOUS v10.1 (Biomatters Ltd, Auckland, New Zealand) (Kearse *et al.* 2012). Considering some genes such as TE-related can partially overlap with repeat regions, it is not uncommon to have some "noise" when inspecting the masked regions. RepeatMasker (Tarailo-Graovac and Chen 2009) allows for a soft-masked genome to help prevent over-masking.

3.4.7 Gene Prediction

Coding regions were annotated by incorporating RNA-seq data with the *ab initio* gene prediction tool AUGUSTUS v3.0.3 (Stanke *et al.* 2008). RNA-Seq transcripts were fed into the pipeline of AUGUSTUS as hints using the "--UTR=on" and "--alternatives-from-evidence=true" options. UTR flag was set to "on" to perform untranslated region annotations. The alternative-evidence flag was set to "true" to predict alternative splicing. The training sets of AUGUSTUS were acquired from the first run of EVidenceModeler (aka EVM) (Haas *et al.* 2008) gene models. The extrinsic evidence for EVM were acquired from transcript alignments and homolog-based predictions. The RNA-Seq data were first used to reconstruct the transcripts via Trinity v2.8.4 (Haas *et al.* 2013), then the transcripts alignments for EVM were created using PASA v2.3.3 (Haas *et al.* 2003). To create the evidence of homolog-based predictions, the protein sequences of closely related species

(*C. reinhardtii* (Merchant *et al.* 2007), *G. pectorale* (Hanschen *et al.* 2016), *C. eustigma* (Hirooka *et al.* 2017), *D. salina* (Polle et al. 2017) and *V. carteri* (Prochnik *et al.* 2010)) were downloaded from JGI (https://phytozome.jgi.doe.gov/pz/portal.html) or NCBI (https://www.ncbi.nlm.nih.gov) database. Then the evidence of protein alignments for EVM were created with Exonerate (Slater and Birney 2005), seeded by Diamond (Buchfink *et al.* 2015). The list of numeric weight values was set to default for each type of "evidence" for EVM.

Functional annotation of protein-coding genes was obtained from the best blast hit by BLASTP (E-value < 1e-5) against SwissProt (Boutet *et al.* 2007), TrEMBL (Boeckmann *et al.* 2003), and NCBI NR databases (non-redundant protein sequence database with entries from GenPept, SwissProt, PIR, PDF, PDB, and RefSeq). I developed a tool called NoBadWordsCombiner v1.0 (Zhang *et al.* 2020), which can automatically merge the BLAST results from the databases of SwissProt (Boutet *et al.* 2007), TrEMBL (Boeckmann *et al.* 2003) and NCBI NR databases. More importantly, it can strengthen the gene definition by filtering those protein function descriptions containing 'bad words', such as hypothetical and uncharacterized proteins. GENEIOUS v10.1 (Biomatters Ltd, Auckland, New Zealand) was used to visualize the gene models and manually trim short gene models. The gene models were manually filtered if genes contained internal stop codons, deduced protein sequences less than 35 amino acids, or coding regions with > 70% of elements from low complexity regions and simple repeats. Pfam domains were annotated by using InterProScan (v4.7) (Zdobnov and Apweiler 2001), which integrates predictive information about protein function from a number of partner resources, such as the InterPro (Quevillon *et al.* 2005) and Pfam (Finn *et al.* 2014) databases. Gene Ontology (GO) terms (Ashburner *et al.* 2000) for each gene were retrieved from the corresponding InterPro or Pfam descriptions. Gene sets were mapped to a KEGG (Kanehisa and Goto 2000) pathways to identify the best match classification for each gene. Genome annotation quality was evaluated by BUSCO (Simão *et al.* 2015), which gave a quantitative measures for single-copy orthologous genes from the dataset Chlorophyta odb10 (Zdobnov *et al.* 2017).

The tRNA genes were predicted by tRNAscan-SE v1.3.1 (Lowe and Eddy 1997) using default parameters for eukaryotes. The miRNA and snRNA fragments were identified by INFERNAL (Nawrocki *et al.* 2009) software against the Rfam (release 12.0) database (Griffiths-Jones *et al.* 2003). Homology-based rRNA fragments were annotated by mapping algal rRNAs to the UWO241 genome using BLASTN with parameters (E-value < 1e-5). Transcription factors (TF) and transcriptional regulators (TR) were annotated by first screening the proteins for domains and then applying a domain-based rule set (Lang *et al.* 2010; Wilhelmsson *et al.* 2017).

3.5 Data Availability

The assembled genome sequences and the raw sequencing data of UWO241 were deposited at US National Center for Biotechnology Information (NCBI) database under BioProject accession PRJNA547753 and BioSample accessions SAMN11975472 and SAMN11975511.

3.6 References

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics. Retrieved from https://www.bioinformatics.babraham.ac.uk.

Arumuganathan, K. and E. Earle (1991). Estimation of nuclear DNA content of plants by flow cytometry. Plant Molecular Biology Reporter 9: 229-241.

Arumuganathan, K. and E. Earle (1991). Nuclear DNA content of some important plant species. Plant Molecular Biology Reporter 9: 208-218.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight and J. T. Eppig (2000). Gene ontology: tool for the unification of biology. Nature Genetics 25: 25-29.

Boeckmann, B., A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan and I. Phan (2003). The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research 31: 365-370.

Boetzer, M., C. V. Henkel, H. J. Jansen, D. Butler and W. Pirovano (2010). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27: 578-579.

Boetzer, M. and W. Pirovano (2012). Toward almost closed genomes with GapFiller. Genome Biology 13: R56.

Bolger, A. M., M. Lohse and B. Usadel (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

Boutet, E., D. Lieberherr, M. Tognolli, M. Schneider and A. Bairoch (2007). UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. Plant Bioinformatics: Methods and Protocols: 89-112.

Buchfink, B., C. Xie and D. H. Huson (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods 12: 59-60.

Casola, C. and E. Betrán (2017). The genomic impact of gene retrocopies: what have we learned from comparative genomics, population genomics, and transcriptomic analyses? Genome Biology and Evolution 9: 1351-1373.

Cvetkovska, M., N. P. A. Huner and D. R. Smith (2017). Chilling out: the evolution and diversification of psychrophilic algae with a focus on Chlamydomonadales. Polar Biology 40: 1169-1184.

Cvetkovska, M., S. Orgnero, N. P. Hüner and D. R. Smith (2019). The enigmatic loss of light-independent chlorophyll biosynthesis from an Antarctic green alga in a light-limited environment. New Phytologist 222: 651-656.

Cvetkovska, M., B. Szyszka-Mroz, M. Possmayer, P. Pittock, G. Lajoie, D. R. Smith and N. P. Hüner (2018). Characterization of photosynthetic ferredoxin from the Antarctic alga *Chlamydomonas* sp. UWO241 reveals novel features of cold adaptation. New Phytologist 219: 588-604.

Dickinson, N. B. and B. Dickinson (1998). CellQuest Software Reference Manual. Becton Dickinson Immunocytometry Systems, San Jose: 1-227.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm and J. Mistry (2014). Pfam: the protein families database. Nucleic Acids Research 42: 222-230.

Griffiths-Jones, S., A. Bateman, M. Marshall, A. Khanna and S. R. Eddy (2003). Rfam: an RNA family database. Nucleic Acids Research 31: 439-441.

Gurevich, A., V. Saveliev, N. Vyahhi and G. Tesler (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072-1075.

Haas, B. J., A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith Jr, L. I. Hannick, R. Maiti, C. M. Ronning, D. B. Rusch and C. D. Town (2003). Improving the *Arabidopsis*

genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research 31: 5654-5666.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li and M. Lieber (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8: 1494-1512.

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

Hanschen, E. R., T. N. Marriage, P. J. Ferris, T. Hamaji, A. Toyoda, A. Fujiyama, R. Neme, H. Noguchi, Y. Minakuchi and M. Suzuki (2016). The *Gonium pectorale* genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. Nature Communications 7: 1-10.

Hirooka, S., Y. Hirose, Y. Kanesaki, S. Higuchi, T. Fujiwara, R. Onuma, A. Era, R. Ohbayashi, A. Uzuka and H. Nozaki (2017). Acidophilic green algal genome provides insights into adaptation to an acidic environment. Proceedings of the National Academy of Sciences 114: 8304-8313.

Hubley, R., R. D. Finn, J. Clements, S. R. Eddy, T. A. Jones, W. Bao, A. F. Smit and T. J. Wheeler (2016). The Dfam database of repetitive DNA families. Nucleic Acids Research 44: D81-D89.

Kalra, I., X. Wang, M. Cvetkovska, J. Jeong, W. McHargue, R. Zhang, N. Hüner, J. S. Yuan and R. Morgan-Kiss (2020). Chlamydomonas sp. UWO 241 exhibits high cyclic electron flow and rewired metabolism under high salinity. Plant Physiology 183: 588-601.

Kanehisa, M. and S. Goto (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28: 27-30.

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz and C. Duran (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647-1649.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman and A. M. Phillippy (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Research 27: 722-736.

Kubiak, M. R. and I. Makałowska (2017). Protein-coding genes' retrocopies and their functions. Viruses 9: 1-27.

Lang, D., B. Weiche, G. Timmerhaus, S. Richardt, D. M. Riaño-Pachón, L. G. Corrêa, R. Reski, B. Mueller-Roeber and S. A. Rensing (2010). Genome-wide phylogenetic

comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. Genome Biology and Evolution 2: 488-503.

Lowe, T. M. and S. R. Eddy (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research 25: 955-964.

Merchant, S. S., S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz, G. B. Witman, A. Terry, A. Salamov, L. K. Fritz-Laylin and L. Maréchal-Drouard (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318: 245-250.

Mock, T., R. P. Otillar, J. Strauss, M. McMullan, P. Paajanen, J. Schmutz, A. Salamov, R. Sanges, A. Toseland and B. J. Ward (2017). Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. Nature 541: 536-540.

Morgan-Kiss, R. M., J. C. Priscu, T. Pocock, L. Gudynaite-Savitch and N. P. Huner (2006). Adaptation and acclimation of photosynthetic microorganisms to permanently cold environments. Microbiology and Molecular Biology Reviews 70: 222-252.

Nakada, T., K. Misawa and H. Nozaki (2008). Molecular systematics of Volvocales (Chlorophyceae, Chlorophyta) based on exhaustive 18S rRNA phylogenetic analyses. Molecular Phylogenetics and Evolution 48: 281-291.

Nawrocki, E. P., D. L. Kolbe and S. R. Eddy (2009). Infernal 1.0: inference of RNA alignments. Bioinformatics 25: 1335-1337.

Neale, P. J. and J. C. Priscu (1995). The photosynthetic apparatus of phytoplankton from a perennially ice-covered Antarctic lake: acclimation to an extreme shade environment. Plant and Cell Physiology 36: 253-263.

Ou, S. and N. Jiang (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiology 176: 1410-1422.

Panchy, N., M. Lehti-Shiu and S.-H. Shiu (2016). Evolution of gene duplication in plants. Plant Physiology 171: 2294-2316.

Pocock, T., M. A. Lachance, T. Pröschold, J. C. Priscu, S. S. Kim and N. P. Huner (2004). Identification of a psychrophilic green alga from Lake Bonney Antarctica: *Chlamydomonas raudensis* ETTL. (UWO241) Chlorophyceae. Journal of Phycology 40: 1138-1148.

Polle, J. E., K. Barry, J. Cushman, J. Schmutz, D. Tran, L. T. Hathwaik, W. C. Yim, J. Jenkins, Z. McKie-Krisberg and S. Prochnik (2017). Draft nuclear genome sequence of the halophilic and beta-carotene-accumulating green alga *Dunaliella salina* strain CCAP19/18. Genome Announcements 5: 01105-01117.

Possmayer, M. (2018). Phylogeny, Heat-Stress and Enzymatic Heat-Sensitivity in the Antarctic Psychrophile, *Chlamydomonas* sp. UWO241. Electronic Thesis and Dissertation Repository 5737. Retrieved from https://ir.lib.uwo.ca/etd/5737.

Possmayer, M., R. K. Gupta, B. Szyszka‐Mroz, D. P. Maxwell, M. A. Lachance, N. P. Hüner and D. R. Smith (2016). Resolving the phylogenetic relationship between *Chlamydomonas* sp. UWO 241 and *Chlamydomonas raudensis* SAG 49.72 (Chlorophyceae) with nuclear and plastid DNA sequences. Journal of Phycology 52: 305-310.

Price, A. L., N. C. Jones and P. A. Pevzner (2005). *De novo* identification of repeat families in large genomes. Bioinformatics 21: 351-358.

Prochnik, S. E., J. Umen, A. M. Nedelcu, A. Hallmann, S. M. Miller, I. Nishii, P. Ferris, A. Kuo, T. Mitros and L. K. Fritz-Laylin (2010). Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science 329: 223-226.

Qian, W. and J. Zhang (2014). Genomic evidence for adaptation by gene duplication. Genome Research 24: 1356-1362.

Quesada, A. and W. F. Vincent (2012). Cyanobacteria in the cryosphere: snow, ice and extreme cold. Ecology of Cyanobacteria II. Spain/Canada, Springer 14: 387-399.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler and R. Lopez (2005). InterProScan: protein domains identifier. Nucleic Acids Research 33: 116-120.

Raymond, J. A. and H. J. Kim (2012). Possible role of horizontal gene transfer in the colonization of sea ice by algae. PloS One 7: e35968.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva and E. M. Zdobnov (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210-3212.

Singh, R., R. Ming and Q. Yu (2016). Comparative analysis of GC content variations in plant genomes. Tropical Plant Biology 9: 136-149.

Slater, G. S. C. and E. Birney (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6: 1-11.

Šmarda, P., P. Bureš, L. Horová, I. J. Leitch, L. Mucina, E. Pacini, L. Tichý, V. Grulich and O. Rotreklová (2014). Ecological and evolutionary significance of genomic GC content diversity in monocots. Proceedings of the National Academy of Sciences 111: E4096-E4102.

Smit, A. and R. Hubley (2008). RepeatModeler Open-1.0. Retrieved from http://www.repeatmasker.org.

Smit, A., R. Hubley and P. Green (2015). RepeatMasker Open-4.0. Retrieved from http://www.repeatmasker.org.

Stanke, M., M. Diekhans, R. Baertsch and D. Haussler (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics 24: 637-644.

Suzek, B. E., Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu and U. Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31: 926-932.

Tarailo-Graovac, M. and N. Chen (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics: 4.10. 11-14.10. 14.

Umen, J. G. and B. J. Olson (2012). Genomics of volvocine algae. Advances in Botanical Research 64: 185-243.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman and S. K. Young (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One 9: e112963.

Wikimedia-Commons (2020). Wikimedia commons, the free media repository. Retrieved from https://commons.wikimedia.org/wiki/Main_Page.

Wilhelmsson, P. K., C. Mühlich, K. K. Ullrich and S. A. Rensing (2017). Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte algae. Genome Biology and Evolution 9: 3384-3397.

Xu, Z. and H. Wang (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research 35: 265-268.

Zdobnov, E. M. and R. Apweiler (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847-848.

Zdobnov, E. M., F. Tegenfeldt, D. Kuznetsov, R. M. Waterhouse, F. A. Simao, P. Ioannidis, M. Seppey, A. Loetscher and E. V. Kriventseva (2017). OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Research 45: D744-D749.

Zhang, X., Cvetkovska, M., Morgan-Kiss, R., Hüner, N.P., and Smith, D.R. (2021). Draft genome sequence of the Antarctic green alga *Chlamydomonas* sp. UWO241. iScience, 102084.

Zhang, X., Y. Hu and D. R. Smith (2021). NoBadWordsCombiner-a tool to integrate the gene function information together without 'bad words' from Nr-NCBI, UniProtKB/Swiss-Prot, KEGG, Pfam databases. . Retrieved from https://github.com/zx0223winner/HSDFinder/blob/master/NoBadWordsCombiner.py.

Zhang, Z., C. Qu, K. Zhang, Y. He, X. Zhao, L. Yang, Z. Zheng, X. Ma, X. Wang and W. Wang (2020). Adaptation to extreme Antarctic environments revealed by the genome of a sea ice green alga. Current Biology 30: 1-12.

Zimin, A. V., D. Puiu, M.-C. Luo, T. Zhu, S. Koren, G. Marçais, J. A. Yorke, J. Dvořák and S. L. Salzberg (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Research 27: 787-792.

3.7 Supplementary Information

Table 8: Comparison of repeats in the genome of selected green algae.

Species	*Chlamydomonas* sp. UWO241		*Chlamydomonas reinhardtii*		*Dunaliella salina*		*Gonium pectorale*		*Volvox carteri*		*Chlamydomonas* sp. ICE-L	
Repeats category	Numbers	Length(bp)/ percentage	Numbers	length(bp)/ percentage	Numbers	length(bp)/ percentage	Numbers	length(bp)/ percentage	Numbers	length(bp)/ percentage	Numbers	length(bp)/ percentage
SINEs	1966	286,815/0.14 %	288	95,696/0.09 %	-	-	919	68,293/0.05 %	-	-	9,325	1,408,737/0.26%
LINEs	72,637	25,259,586/11.94%	14,905	5,451,206/4.91%	115,742	47,084,016/13.70%	6251	1,827,383/1.23%	-	-	50,041	25,334,170/4.68%
LTR elements	4549	1,358,609/0.64 %	1358	514,334/0.46%	14,425	5,819,049/1.69%	3583	1,008,765/0.68%	-	-	52,043	46,458,114/8.57%
DNA elements	34,763	5,064,215/2.39 %	16,948	3,587,707/3.23%	10,102	2,226,438/0.65%	1365	191,356/0.13%	-	-	29,364	10,823,438/2.0%
Unclassified	282,664	60,449,198/28.56%	30,781	4,334,103/3.90%	298,034	62,119,743/18.07%	46,926	8,545,222/5.74%	102,273	27,262,548/20.79%	751,814	159,882,493/29.51%
Total transposable elements	396,579	92,418,423/43.67%	64,280	13,983,046/12.59%	438,303	117,249,246/34.11%	59,004	11,641,019/7.82%	102,273	27,262,548/20.79%	892,587	243,906,952/45.01%
Simple repeats/Satellites	417,034	31,185,848/14.74%	129,144	8,278,494/7.45%	147,030	9,850,665/2.86%	87,793	4,997,301/3.36%	147,623	7,485,528/5.71%	515,710	44,023,155/8.13%
Low complexity	23,564	1,538,790/0.73 %	12,368	774,505/0.70%	4288	268,195/0.08 %	12,993	741,080/0.50%	-	-	53,522	4,971,406/0.92%
Total repeats	1,233,756	112,319,356/53.07%	205,792	23,527,046/21.18%	1,028,754	126,979,512/36.94%	218,857	17,374,479/11.68%	352,169	33,773,474/25.75%	1,461,819	293,495,403/54.16

Table 9: Statistics of BUSCO assessment of the green algae genome assembly and genome annotation.

BUSCO mode: genome assembly (No. of genes/Percentage)						
Species	No. of total BUSCOs	Complete BUSCOs	Complete and single-copy BUSCOs	Complete and duplicated BUSCOs	Fragmented BUSCOs	Missing BUSCOs
Chlamydomonas sp. UWO241	2168	1840/84.9%	1706/78.7%	134/6.2%	77/3.6%	251/11.5%
Chlamydomonas reinhardtii	2168	2079/95.9%	2069/95.4%	10/0.5%	47/2.2%	42/1.9%
Dunaliella salina	2168	1400/64.6%	1377/63.5%	23/1.1%	224/10.3%	544/25.1%
Gonium pectorale	2168	1844/85.0%	1826/84.2%	18/0.8%	117/5.4%	207/9.6%
Volvox carteri	2168	2061/95.0%	2045/94.3%	16/0.7%	68/3.1%	39/1.9%
Chlamydomonas sp. ICE-L	2168	1684/77.7%	1519/70.1%	165/7.6%	142/6.5%	342/15.8%
BUSCO mode: genome annotation (No. of genes/Percentage)						
Species	No. of total BUSCOs	Complete BUSCOs	Complete and single-copy BUSCOs	Complete and duplicated BUSCOs	Fragmented BUSCOs	Missing BUSCOs
Chlamydomonas sp. UWO241	2168	1652/76.2%	1196/55.2%	456/21.0%	113/5.2%	403/18.6%
Chlamydomonas reinhardtii	2168	2105/97.1%	1964/90.6%	141/6.5%	53/2.4%	10/0.5%
Dunaliella salina	2168	1319/60.9%	1229/56.7%	90/4.2%	333/15.4%	516/23.7%
Gonium pectorale	2168	1640/75.6%	1618/74.6%	22/1.0%	245/11.3%	283/13.1%
Volvox carteri	2168	2087/96.2%	1898/87.5%	189/8.7%	47/2.2%	34/1.6%
Chlamydomonas sp. ICE-L	2168	1656/76.4%	1416/65.3%	240/11.1%	175/8.1%	337/15.5%

Chapter 4

4 Comparative Genomic Analysis of the Antarctic Psychrophilic Green Alga *Chlamydomonas* sp. UWO241 Provides Insights into Gene Duplication Driving Cold Adaptation

This chapter was adapted from the publication entitled "Draft genome sequence of the Antarctic green alga *Chlamydomonas* sp. UWO241" published on iScience in 2021 by X. Zhang, M. Cvetkovska, R. Morgan-Kiss, N. P. A. Hüner and D. R. Smith (Zhang *et al.* 2021).

The introduction of this chapter was adapted in part from the publication entitled "HSDFinder: an integrated tool for predicting highly similar duplicates in eukaryotic genomes" in 2021 by X. Zhang, Y. Hu and D. R. Smith (Appendix C).

4.1 Introduction

What is the role of gene duplicates?

It is often disadvantageous to retain highly similar expressed sequences; therefore, it should be rare to have duplicates encoding the same functions maintained in the genome (Kubiak and Makałowska 2017). However, Zhang suggested that the generation of large-scale duplicates was possible only if they were genes in high demand, such as gene for rRNAs and histones (Zhang 2003). Thereafter, Libuda and Winston discovered that the appearance of pairs of adjacent paralogous proteins arose from a compensatory mechanism restoring normal dosage when one locus was deleted (Libuda and Winston 2006). Recently, the controversy has been in whether the evolution of duplicate genes affects fitness (Innan and Kondrashov 2010). Some duplication models assume that the fixation of the duplicate copy is a neutral process, while others support the gene dosage hypothesis, where if an increase in the dosage of a particular gene is beneficial, then a duplication of this gene may be fixed by positive selection (Qian and Zhang 2008). Nevertheless, mechanisms that do not require the evolution of new functions (e.g., dosage balance) may play an important role in the initial retention of duplicate genes (Panchy *et al.* 2016). Indeed, many examples have

accumulated in the literature suggesting that stress response genes, sensory genes, transport genes and genes that have a metabolism-related function are likely to be fixed as duplicate copies under certain environmental conditions (Kondrashov 2012). In addition, genes encoding the protein products requiring large doses, such as ribosomal or histone genes, are also maintained in the genome (Innan and Kondrashov 2010). In *Chlamydomonas* sp. UWO241, many ribosomal protein duplicates were detected, which might benefit gene expression. However, the gene dosage hypothesis could be further tested by determining whether retrogene-parental gene pairs with overlapping expression show a higher combined transcription level than parental genes in multiple closely related outgroup species lacking those retrogenes (Casola and Betrán 2017).

How do gene duplicates arise?

The next key question is how these duplicates arise. There are five main broad classes of duplication events in genomes: whole-genome duplication (WGD), tandem duplication, transposon-mediated duplication, segmental duplication and retroduplication (Panchy *et al.* 2016). Polyploidization or WGD, is a straightforward gene duplication mechanism that increases both genome size and entire gene sets. However, it is not the only mechanism that generates duplicate genes. A cluster of two to many paralogous sequences with no or few intervening gene sequences is a pattern of tandem (or local) duplication that results from unequal crossing-over of chromosomes or transposable-element-(TE)-mediated duplication. Furthermore, transposon-mediated duplication usually contains the hallmarks of two terminal inverted repeats (TIRs) less than 5 kb long. Segmental duplication usually arises from non-LTR (long terminal repeats) retrotransposons, such as LINEs (intact LINE1s are up to 6 kb in length and contain internal promoters). Retroduplication refers to retrogenes generated via 5~9 kb LTR-retrotransposons, such as *gypsy* LTR elements (Panchy *et al.* 2016). Notably, if a gene is duplicated via reverse transcription of mRNA and then inserts into the genome, it is referred to as retrocopy, and the original gene is referred to as the parental gene. Although a retrocopy can arise from both LTR and non-LTR retrotransposable elements (e.g., LINE1), the expression of the retrocopy is largely dependent on the regulatory region (i.e., promoters, binding sites for the RNA polymerase, and/or enhancers) (Kubiak and Makałowska 2017).

What are TE-generated gene duplicates?

Given the multiple mechanisms of duplicate/retrocopy generation, WGDs and tandem duplications usually account for the majority of plant duplicates. However, TE-based mechanisms (i.e., retroduplication and transposon-mediated duplication) also generate a significant number of duplicates (Lisch 2013; Zhang *et al.* 2013; Tan *et al.* 2016; Casola and Betrán 2017; Cerbin and Jiang 2018). TEs are categorized into two classes. Class I TEs (retrotransposons) are RNA-mediated and operate via a copy-and-paste transposition mechanism, while class II TEs (DNA transposons) use a DNA-mediated mechanism with a cut-and-paste process (Wicker *et al.* 2007; Del Angel *et al.* 2018). Based on the appearance of LTRs, class I TEs are further classified as LTR retrotransposons, including the superfamily of *copia* and *gypsy* retrotransposons, and non-LTR retrotransposons containing elements such as SINEs and LINEs (Han 2010). DNA transposon could result in vast amounts of duplication and reshuffling the surrounding host sequences via the high frequent cut-and-paste process (Bourque *et al.* 2018). In the rice genome, Jiang *et al.* identified over 3,000 DNA transposons (Pack-MULEs) containing fragments derived from more than 1,000 cellular genes (Jiang *et al.* 2004). In humans, most retrotransposons are non-LTRs, but in plants, the genome size is expanded significantly due to the large size and number of LTR retrotransposons. For instance, retrotransposons contribute to approximately 75% of the size of the maize (*Zea mays*) genome (Schnable *et al.* 2009). Indeed, the redundancy of the duplicate genes (i.e., retrocopies) in the genome is largely attributed to the retrotransposition. Because the regulatory regions (e.g., promoters) cannot be duplicated together with coding regions via retrotransposition, most retrocopies lack expression, resulting in extreme redundancy (Kubiak and Makałowska 2017). However, some retrogenes have successfully acquired regulatory regions (e.g., promoters, enhancers, and binding sites for the RNA polymerase) in different ways. For example, in mice and humans, the majority (86%) of retrogenes appear to be transcribed from newly evolved regulatory regions, while only 3% of retrogenes inherited regulatory regions from their parental genes, and 11% are transcribed from bidirectional regulatory regions of upstream genes in head-to-head orientation (Carelli *et al.* 2016). Consistent with the ability to acquire the regulatory regions, the fate of the duplicates varies dramatically. There are generally three potential outcomes for gene duplicates. Gene duplication most often results in a

nonfunctional duplicate gene copy (nonfunctionalization). Some duplicate genes, however, undergo functional divergence. For example, one of the gene copies evolves a new beneficial function while the parental copy retains the original function (neofunctionalization) or both the original and the duplicate genes evolve to fulfill complementary functions previously performed by the original gene (subfunctionalization) (Conrad and Antonarakis 2007).

Recently, it was shown that UWO241, unlike other surveyed algae, produces two near-identical copies of photosynthetic ferredoxin (PETF), resulting from a duplication of the nuclear *petf* gene (Cvetkovska *et al.* 2018). The retention and expression of this duplicate gene is hypothesized to be an adaptation to the cold, leading to higher protein accumulation (i.e., gene dosage); indeed, UWO241 accumulates greater amounts of PETF than its mesophilic close relative *Chlamydomonas reinhardtii* (Merchant *et al.* 2007; Cvetkovska *et al.* 2018). Similarly, UWO241 expresses three isoforms of an unusual bidomain enzyme, allowing it to produce high levels of osmoprotectant glycerol (>400 mM) (Kalra *et al.* 2020). If gene dosage is contributing to psychrophily in UWO241, one might expect other genes to be duplicated.

Comparative genomic analysis across species has been widely used to identify new genes and functional coding sequences, and for a long period of time (Nobrega and Pennacchio 2004). These analyses have undoubtedly made important contributions in understanding differences in gene content, such as intron length and abundance as well as the numbers and types of repeats. Nonetheless, the further comparisons of gene families, pathways and conserved domains are limited due to the lack of an appropriate comprehensive genomic framework. Fortunately, UWO241 is nested together with numerous mesophilic algal species in the order Chlamydomonadales, including *C. reinhardtii*, which is an excellent comparison target for the investigations of psychrophilic chlamydomonads (Cvetkovska *et al.* 2017). Many comparative genomic analyses, such as comparisons of gene family expansion and contraction, pathway loss and gain, and substitution rates at synonymous and nonsynonymous sites of protein-coding genes can help further understand the role of UWO241 as a psychrophile.

In this chapter, the UWO241 genome is compared with those of other model green algae, including *C. reinhardtii*, *Volvox carteri*, *Dunaliella salina, Chlamydomonas eustigma, Chlamydomonas* sp. ICE-L and *Gonium pectorale*. Some of the questions I try to address are as follows: Does the UWO241 genome harbor large numbers of duplicate genes? Has it acquired any genes via HGT, such as ice-binding proteins (IBPs) genes? Does UWO241 contains unique or expanded/contracted gene families compared to its close relatives? Preliminary findings show that the genomic architecture of UWO241 is very different from that of *C. reinhardtii*. For example, the genome size of the UWO241 is double that of the model alga *C. reinhardtii*. The predicted gene numbers are roughly the same as that in the other species, and the number of gene families is slightly lower than that in other Chlorophyceae, including the volvocine algae (*Chlamydomonas*, *Gonium*, and *Volvox*). Given all the previous assessments, UWO241, as a psychrophilic alga, has been widely explored and has generated wide interest. Here, genome sequencing of UWO241 exposed hundreds of gene duplicates for crucial cellular pathways and dozens of genes encoding IBPs. These findings for UWO241 (isolated from a constantly cold but non-freezing environment) mirror many of those from the recent genomic analysis of the psychrophiles, *Chlamydomonas* sp. ICE-L (Zhang *et al.* 2020), which originates from a cold but fluctuating Antarctic sea ice environment, and enhance our understanding of photopsychrophily and the evolutionary dynamics within Antarctic lakes.

4.2 Results and Discussion

4.2.1 Gene Duplication Analysis Across Species

Generally, genome or gene duplication is widely considered to facilitate environmental adaptation because redundancy allows the evolution of novel beneficial gene functions (Kondrashov 2012). Plant genomes are thought to be rich in gene duplicates due to ancient duplication events (Panchy *et al.* 2016). As previously reported (Cvetkovska *et al.* 2018), the photosynthetic ferredoxin gene (Fd) 1A and 1B in the UWO241 genome are quite similar to each other, with 91% identity in coding regions (both have a length of 1,114 bp, 3 introns, and 4 exons); however, the Fd gene of *C. reinhardtii* is only 593 bp in length (1 intron and 2 exons). By exploring the gene content, it is not difficult to detect the higher abundance of noncoding regions attributed to differences in the two species. It is likely that

the UWO241 Fd gene has undergone a gene duplication event, and the coding DNA sequences remaining similar due to selective pressure operating at the protein level. More importantly, Cvetkovska *et al.* further discovered that the two ferredoxin proteins in UWO241 were novel class of cold-adapted enzymes, which were shown to have the unusual feature of both high activity at low temperatures and high stability at moderate temperatures compared to its mesophilic orthologue (Cvetkovska *et al.* 2018).

Given the previous assessment of the duplicate genes in UWO241, I hypothesized that genes that are crucial for the extremophilic lifestyle of UWO241 are likely present as highly conserved copies. Indeed, the identification of these gene copies has improved our understanding of gene duplications as a mechanism of adaptation. Functional annotation of the 16,325 RNA-supported gene models revealed the standard cohort of proteins typically encoded in green algal nuclear genomes (Appendix A: Table S4) as well as many hypothetical proteins (21.8%), paralleling the trends from other available chlamydomonadalean nuclear gene sets, which are generally 20-30% hypothetical. There were no obvious signs of contamination in the annotations and, with one conspicuous exception (discussed below), little evidence of horizontal gene transfer (HGT). Examining the annotations in detail, it became obvious that many were represented two or more times within the genome. To explore the validity of these multi-copy genes, I performed a series of BLAST-based analyses with strict downstream filtering. Specifically, to count the number of gene duplicates in UWO241 genome, a protein BLAST of the UWO241 gene models against themselves (E-value < 1e-5) detected 901 putative duplicates (encompassing 2,012 gene copies) all with pairwise amino acid identities ≥80%. I filtered this gene set to only those with near-identical protein lengths (within 10 amino acids) and ≥90% pairwise identities, giving a pared-down list of 336 highly similar duplicates (HSDs), totaling 1,339 gene copies (Table 10 and Appendix A: Table S5). By setting such a strict cut-off, I have undoubtedly removed some genuine duplicates from this list, but I would rather be conservative in our approach, ensuring that the gene pairs in question are bona fide duplicates rather than spurious ones. The protein sequences of the HSDs were searched against the KEGG and Pfam databases, providing a functional breakdown (Table 10 and Appendix A: Table S5). HSDs in UWO241 are involved in various cellular pathways,

including gene expression, cell growth, membrane transport, and energy metabolism (Table 10 and Appendix A: Table S5), but also include hypothetical proteins (~37%) and reverse transcriptases (11%). HSDs for protein translation, DNA packaging, and photosynthesis were particularly prevalent, with 19 duplications of genes for ribosomal proteins, 10 for histones, and 7 for proteins of the chlorophyll *a*/*b* binding light harvesting complex (LHCB) (Table 10). As with the previously described *petf* duplication (Cvetkovska *et al.* 2018), many of these HSDs are virtually indistinguishable from each other at the amino acid level, and 65 are identical across their nucleotide coding regions (Appendix A: Table S5).

Surprisingly, these large gene duplicate numbers are quite unusual compared to the numbers in their close mesophilic green algal relatives. Subsequently, I followed the similar gene duplicate detection protocol and obtained duplicates in other closely related algal species. However, the number of gene duplicates was not nearly as large as that in UWO241. Although the other close relatives also contain duplicates, such as glycolysis genes involved in sugar metabolism, genes encoding antenna proteins important for photosynthesis, and genes for purine relative to nucleotide metabolism, the duplication level is not nearly as high as that in UWO241. Noticeably, the duplicate genes could be involved in all fundamental pathways of the cell, many of which might be linked to how this organism survives its harsh environment. It is currently not immediately obvious if these genes are linked to cold adaptation, and connection between growth rate and the expression of cold adapted enzymes are required to be verified by wet laboratory approaches, such as over-expression or knock-out experiments. Nonetheless, there are some examples of gene duplicates worth exploring further, such as previously discussed photosynthetic ferredoxin proteins, which have been related to cold adaptation experimentally (Cvetkovska *et al.* 2018). There are a few other gene duplicates encoding important functions, for example, antenna proteins involved in the photosynthetic light harvesting system (Dolhi *et al.* 2013), the histones that package DNA (Tariq and Paszkowski 2004), the transporter involved in nutrient uptake that might be necessary for extreme environments (Saier 2000), and even the ribosomal proteins involved in DNA translation (McIntosh and Bonham-Smith 2006). As displayed in Table 10, UWO241 was

identified as having many duplicates in energy metabolism (10 HSDs), lipid metabolism (3 HSDs) and translation (27 HSDs).

Table 10: Summary statistic of highly similar duplicate genes (HSDs) in UWO241.

Database	Identifiers	Number of HSDs (%)[a]	Number of gene copies (%)[a]
Pfam			
Chlorophyll A-B binding protein	PF00504	4 (1%)	25 (2%)
Ribosomal protein	PF01015; PF01775; PF00828	19 (5%)	42 (3%)
Core histone H2A/H2B/H3/H4	PF00125	5 (1%)	99 (7%)
Ice-binding protein (DUF3494)	PF11999	8 (2%)	21 (2%)
Reverse transcriptases	PF00078	38 (11%)	151 (11%)
KEGG			
09101 Carbohydrate metabolism	K13979 (alcohol dehydrogenase)	12 (4%)	89 (7%)
09102 Energy metabolism	K02639 (ferredoxin); K08913(light-harvesting complex II chlorophyll a/b binding protein 2)	10 (3%)	51 (4%)
09103 Lipid metabolism	K01054 (acylglycerol lipase)	3 (1%)	15 (1%)
09122 Translation	K02868 (large subunit ribosomal protein L11e)	27 (8%)	47 (4%)
Hypothetical Proteins	NA	125 (37%)	357 (27%)

[a] A total of 336 HSDs were identified within the UWO241 genome, encompassing 1,339 gene copies. HSDs share ≥90% pairwise amino acid identity and have lengths within 10 amino acids of each other.

How does RNA-mediated duplication work?

While the current literature has advanced our knowledge of the mechanisms of gene duplication, many of them remain to be determined. In Arabidopsis, 30% of duplicates could not be assigned to any known mechanisms, while the other approximately 70% of the duplicate genes could be attributed to WGD, tandem duplication and segmental duplication, among other processes. Given the complexity of gene duplication mechanisms, gene duplication analysis of the UWO241 genome was performed here. A large number of retrocopies were detected in the UWO241 genome (See Methods section). In total, 77 autonomous virus-like LTR retrotransposons and 324 non-LTR retrotransposons (e.g., LINE1) were detected (Appendix A: Table S4 and Figure 15D). It should be noted that the real number might be higher because the TE elements could be subject to erosion and yield incomplete TE fragments (Kubiak and Makałowska 2017). Indeed, some RNA-mediated TE elements were also detected in the intronic and intergenic regions of the UWO241 genome. Considering these factors, I filtered the retrocopies with the criteria of an aligned length of at least 50 amino acids and a greater than 80% amino

acid length identity to detect more recent retrocopies, because young TE elements do not have sufficient time to accumulate deleterious mutations (Panchy *et al.* 2016). One of the young TE elements maintained intact is a non-LTR retrotransposon (LINE1) adjacent to a group of antenna protein duplicates. As outlined in a simplified graph (Figure 12), six photosynthetic light-harvesting system gene duplicates were lettered from A to F and located on two different contigs. The non-LTR retrotransposon (LINE1) of contig 1 remained intact with the complete structures of a short remnant poly(A) tail at the 3' end, 5-10 bp target site duplications (TSDs), and 2 ORFs containing reverse transcriptase (RT), while the LINE1 on contig 2 was fragmented but retained the partial non-LTR retrotransposon structure of RTs and a poly (A) tail. RNA-mediated transposition is a "copy and paste process" (Tan *et al.* 2016), but after looking closely into the gene contents of the two contigs, I found that A, B and C shared the same number of exons and introns, while D, E and F had similar exon and intron structures. Furthermore, the B and C and the E and F genes were inverted in a head-to-head orientation. This is certainly not something unheard of; for example, in the mice and humans, 11% of retrocopies are transcribed from bidirectional regulatory regions of upstream genes in a head-to-head orientation (Carelli *et al.* 2016). Additionally, on contig 1, the gene length of D was greater, and the distance of D was farther than those of A, B and C, suggesting that shorter tandem duplicate gene clusters are duplicated earlier. Indeed, in the maize genome, a higher than expected proportion of single-exon genes in tandemly duplicate gene clusters was potentially attributed to duplication efficiency (Kono *et al.* 2018). Relative to the retrocopies D, E and F, A, B, and C exhibited shorter introns, suggesting that D, E, and F are more ancient duplicates that accumulated novel introns. Actually, retrocopies may acquire introns via different strategies. First and foremost, they can inherit introns from their parental genes. Second, they may acquire novel introns via *de novo* exons from flanking genomic DNA or intronization of their original coding sequences. Third, retrocopies can acquire novel introns by the formation of fusion (chimeric) transcripts that include exons from nearby genes (Nefedova and Kim 2017). Taken together, these results suggested that this is an ongoing duplication event, with recently duplicated copies A, B and C and the ancient duplicated copies E, F and D.

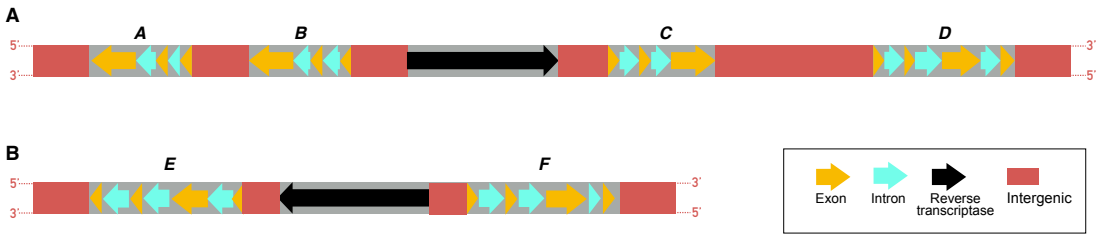


Figure 12: Simplified graph of antenna protein genes in UWO241 genome.

(A) Four distinct copies of *lhcb2* (A: g12385.t1, B: g12386.t1, C: g12388.t1, D: g12389.t1) and a non-LTR retrotransposon (LINE1) (black), all located on scaffold scf7180000014917. (B) Two distinct copies of *lhcb2* (E: g2060.t1, F: g2062.t1) and a non-LTR retrotransposon (LINE1) (black), located on scaffold scf7180000011443.

What is the potential driving force of gene duplication?

The arrangements of the HSDs are informative. Approximately 20% contain gene copies that are situated close to one another, often in a head-to-head or head-to-tail orientation, and have very similar intron numbers and intronic sequences, implying that they result from recent tandem duplication events (Figure 13 and Appendix A: Table S5). A clear example of this is the duplication of the *lhcb2* gene (Figure 13A). The remaining HSDs are generally far apart (most on distinct scaffolds) and, despite their matching coding regions, many (~50%) have un-alignable intronic sequences and differing numbers of introns, suggesting that they derive from more ancient duplication events (Figure 13 and Appendix A: Table S5). This is the case for *petf* (Cvetkovska *et al.* 2018) as well as for *hspa5* (encoding heat shock 70-kDa protein 5), the two copies of which are found in the middle of distinct scaffolds, share 93% coding sequence identity but <12% similarity across their introns (Figure 13B, C).

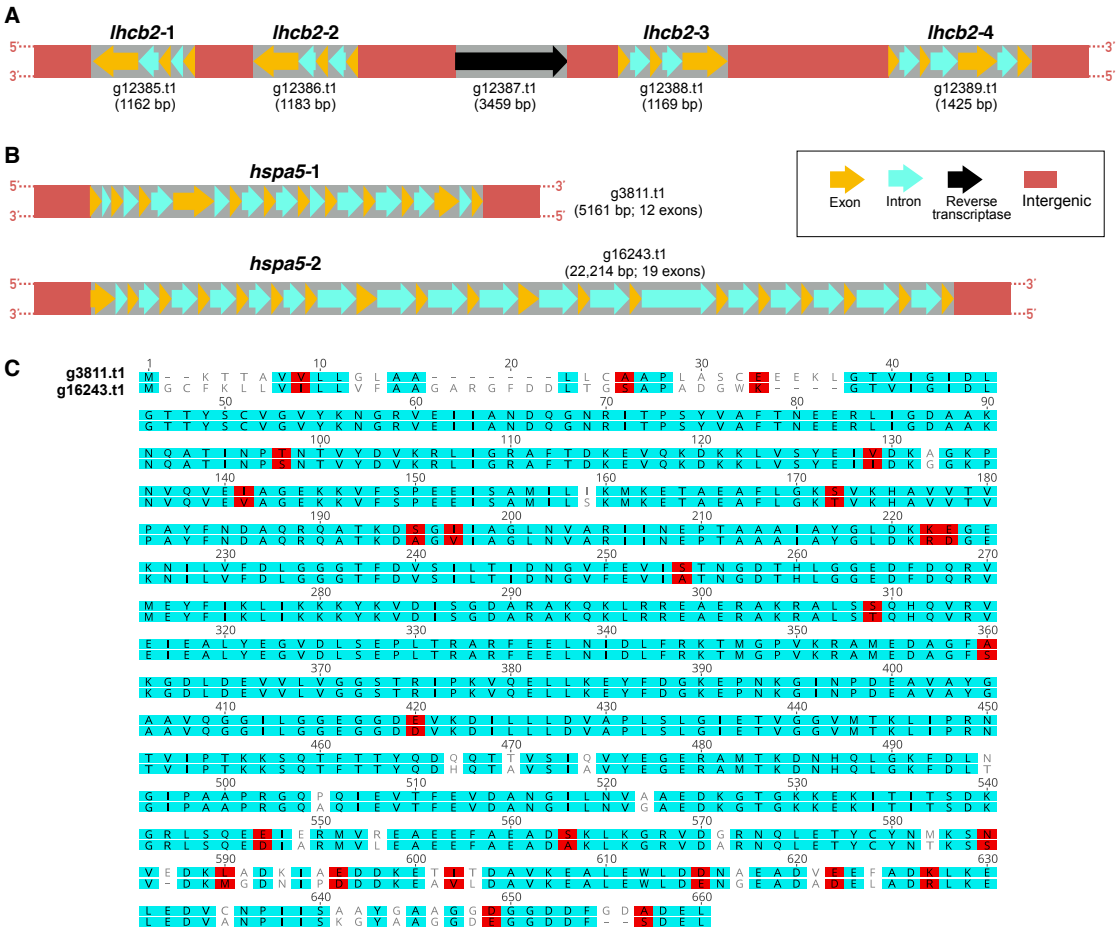


Figure 13: Examples of duplicate genes in *Chlamydomonas* sp. UWO241.

(A) Four distinct copies of *lhcb2*, all located on scaffold scf7180000014917 (B) Two distinct copies of *hspa5*, located on scaffolds scf7180000011611 (*hspa5*-1) and scf7180000015050 (*hspa5*-2). (C) Pairwise alignment of the deduced amino acid sequences of *hspa5*-1 and *hspa5*-2.

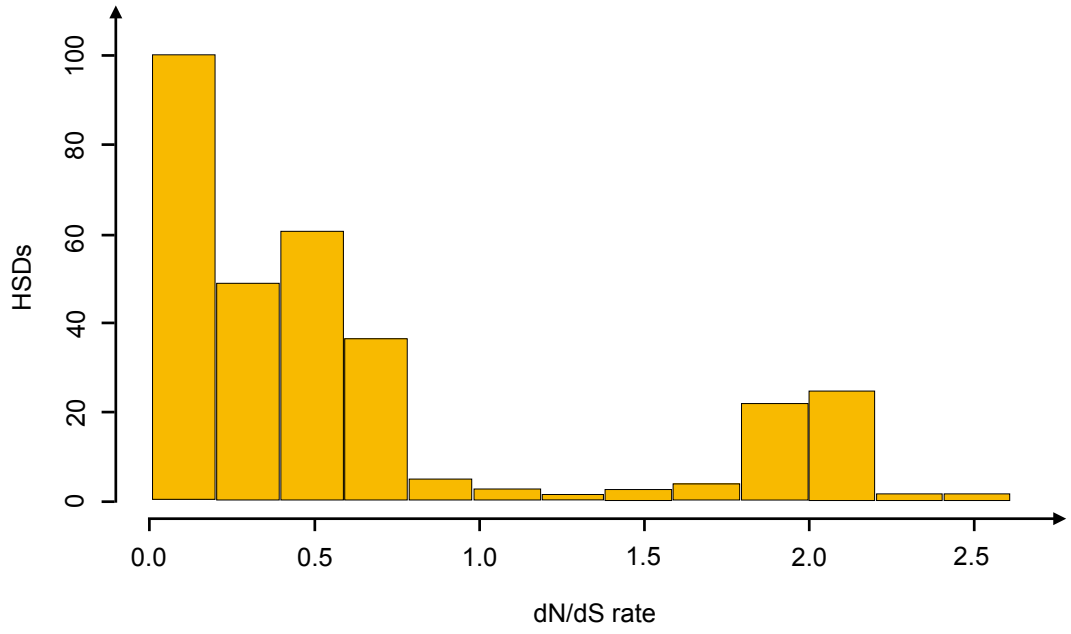


Figure 14: The distribution of nonsynonymous to synonymous substitution rates (dN/dS) among 316 HSDs in UWO241.

Unlike elusive duplicate structure, nonsynonymous (dN) and synonymous (dS) substitution rates (dN/dS) are of great significance for understanding the evolutionary dynamics of protein-coding sequences across closely related and recently diverged species (Fay and Wu, 2003). The dN/dS (ω) ratio can provide a measure of selection pressure at the amino acid level (Yang and Bielawski 2000). Previous studies have explored how ω can be used to determine whether the Fd enzymes of UWO241 are under evolutionary pressure to gain cold adaptation characteristics (Cvetkovska *et al.* 2018). Here, I conducted a selection pressure analysis of duplicated genes in the UWO241 genome. The pairwise model of the PAML 4 package (Yang 2007) was used on the duplicates (approximately 1000 highly similar duplicate genes were selected). As displayed in Figure 14, if the dN/dS rate approaches zero, that is a sign of purifying selection, which refers to the evolutionary force maintaining the same function for a pair of sequences. The exonic sequences of more than half of the HSDs (~190) are under strong purifying selection as evidenced by very low (<<1) nonsynonymous to synonymous substitution rates (dN/dS), ranging from 0-0.5 (avg.

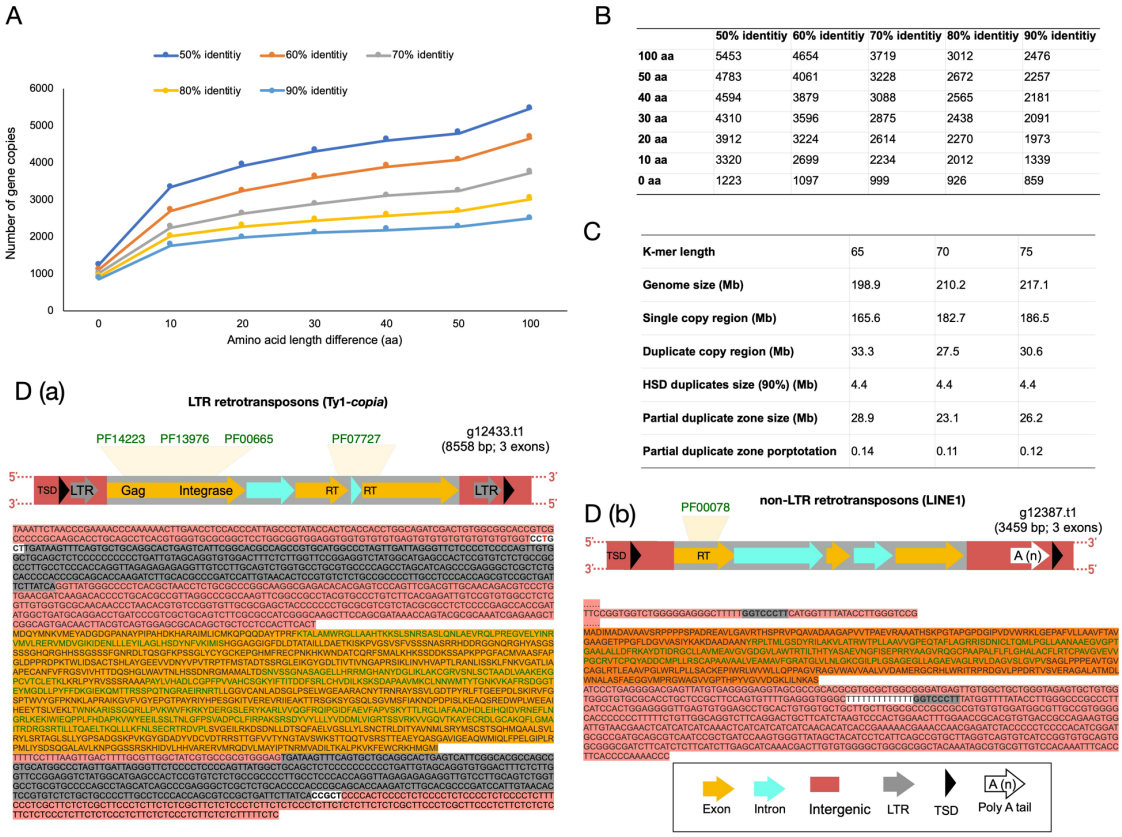
= 0.2) (Figure 14). This leaves open the possibility that natural selection is, in at least some instances, maintaining the expression of similar (if not identical) proteins in UWO241, as it is for PETF, which could aid its survival in Lake Bonney, perhaps due to increased gene dosage, as previously suggested (Innan and Kondrashov 2010; Kondrashov 2012). The HSDs, however, represent only a fraction of duplicated regions within the genome.

Why partial gene duplicates make the genome more complex?

The UWO241 nucDNA contains thousands of partial gene duplicates, characterized by gene fragments and pseudogenes, as well as duplicated segments of intergenic and intronic DNA (Figure 15 and Appendix A: Table S6). These incomplete duplicates range in size from ~100-12,000 bp, can exist in high copy numbers (>6) and, like the HSDs, can be found in tandem or on different scaffolds (Figure 15 and Appendix A: Table S6). But unlike the HSDs, they are in various states of decay, possibly reflecting an ongoing birth-death process, which is supported by the fact that many of the complete and partial duplicates are directly associated with or occur near to retrotransposons (RTs) (Figure 15 and Appendix A: Table S6), as outlined for the duplication of *lhcb2* in Figure 13A.

RT-mediated gene duplication is a recurring theme within nuclear genomes (Qian and Zhang 2014; Panchy *et al.* 2016; Casola and Betrán 2017; Kubiak and Makałowska 2017), including those of green algae (Jąkalski *et al.* 2016), and the UWO241 genome contains the standard hallmarks of such a phenomenon, such as poly(A) tail insertions and target-site duplications (Figure 15D). But this certainly does not rule out the possibility that other processes, such as unequal crossing-over (Zhang 2003), are contributing to gene duplication within UWO241. Do note that 83% of the HSDs contain introns, a characteristic not generally associated with RT-mediated duplications, but not unprecedented (Casola and Betrán 2017; Kubiak and Makałowska 2017). Retrocopies often inherit introns from parental genes, flanking genomic DNA, or the fusion of transcripts (Catania and Lynch 2008; Zhu *et al.* 2009; Szcześniak *et al.* 2011; Kang *et al.* 2012; Zhang *et al.* 2014). Altogether, I identified 401 putatively functional RTs in the nucDNA, including 77 long terminal repeat (LTR) and 324 non-LTR RTs. These numbers do not include retropseudogenes, partial retroelements, or identified RTs with no RNA-seq support, which together account for >10% of the assembly. What's more, there are >480

duplicated regions containing a reverse-transcriptase domain, including ones in noncoding DNA. UWO241 has more retroelements than all other surveyed chlorophytes (4-times that of *C. reinhardtii*) with the exception of ICE-L, for which non-LTR RTs account for a staggering ~23% of the genome (Zhang *et al.* 2020). In addition to RTs, the UWO241 and ICE-L genomes share another atypical feature—genes for IBPs.



	50% identity	60% identity	70% identity	80% identity	90% identity
100 aa	5453	4654	3719	3012	2476
50 aa	4783	4061	3228	2672	2257
40 aa	4594	3879	3088	2565	2181
30 aa	4310	3596	2875	2438	2091
20 aa	3912	3224	2614	2270	1973
10 aa	3320	2699	2234	2012	1339
0 aa	1223	1097	999	926	859

K-mer length	65	70	75
Genome size (Mb)	198.9	210.2	217.1
Single copy region (Mb)	165.6	182.7	186.5
Duplicate copy region (Mb)	33.3	27.5	30.6
HSD duplicates size (90%) (Mb)	4.4	4.4	4.4
Partial duplicate zone size (Mb)	28.9	23.1	26.2
Partial duplicate zone porptotation	0.14	0.11	0.12

Figure 15: Partial gene duplicates, retrogenes, and retrotransposons in UWO241.

(A) The line graph of duplicates set to different thresholds of amino acid pairwise identity and deduced amino acid length. The X-axis indicates the deduced amino acid length (aa) of each duplicate, the Y-axis tells the number of gene copies. (B) The table of total gene copies number at different thresholds of amino acid pairwise identity and deduced amino acid length. (C) The rough gauge of the proportion of partial duplicates in UWO241. (D) (a) The structure example of LTR-retrotransposon (Ty1- *copia*). The LTR retrotransposon is flanked by long terminal repeats (LTRs, grey) and short black triangles indicate target site duplications (TSDs, black). (D) (b) The structure example of non-LTR retrotransposon

(LINE1) which is terminated by a 3' poly(A) tail (A(n), white arrow). The Pfam domains (green) are detailed here (PF14223: gag-polypeptide of LTR *copia*-type; PF13976: gag-pre-integrase domain; PF00665: Integrase core domain; PF07727: Reverse transcriptase; PF00078: Reverse transcriptase).

4.2.2 Acquisition of Ice-Binding Proteins (IBPs) through Horizontal Gene Transfer (HGT)

Environmental adaptation seems to have been facilitated by HGT from various bacteria and archaea (Keeling and Palmer 2008). IBPs usually maintain the unknown functional domain DUF3494 with the Pfam identifier PF11999, which has been detected in more than 170 microorganisms from various habitats (Mock *et al.* 2017). Previous studies reported a common trend regarding the existence of IBPs in cold-adapted algal species. For example, Raymond and Morgan-Kiss detected at least 12 isoforms of IBPs in the Antarctic lake alga *Chlamydomonas* sp. UWO241 (Raymond and Morgan-Kiss 2013). A few years later, as many as 50 isoforms of IBPs were found in another polar alga, *Chlamydomonas* sp. ICE-MDV (Raymond and Morgan-Kiss 2017). More importantly, it is revealed that the IBPs are more closely related to bacterial IBP sequences than other chlorophyte IBP sequences, suggesting that IBP genes were acquired from other microorganisms by HGT (Raymond and Kim 2012; Raymond and Morgan-Kiss 2017).

Here, to verify whether the genes have a bacterial origin, a BLAST search of the UWO241 proteome (BLASTP, E-value < 1e-5 and at least 50 amino acids overlapping) was carried out against the NCBI-nr database. Approximately 100 top hits with a bacterial origin were selected. Alternatively, candidate IBP genes were obtained via BLAST searches against the genome using known UWO241 IBP sequences as the query. The UWO241 genome encodes no fewer than 37 proteins with an ice-binding domain (DUF3494) (Figure 16A), which is among largest number of IBPs ever recorded in a photosynthetic protist. This wealth of IBPs appears to be the consequence of HGT events in combination with gene duplication. Phylogenetic analyses of the IBP genes, which range in size from 483-37,549 bp, show their grouping with psychrophilic bacterial and archaeal IBPs (Figure 16B), which is consistent with previous work (Raymond and Morgan-Kiss 2013). Nuclear genes acquired via recent HGT events from bacteria usually lack introns (Keeling and Palmer

2008), as do 14 of the IBP genes from UWO241; the remaining genes, with 4 exceptions, all have a single, short intron at their 3' ends. The largest IBP gene, however, contains 29 introns. The IBP genes show varying degrees of similarity with each other (Figure 16C), including 8 groupings of almost identical genes, suggesting a complicated history of IBP gene acquisition and duplication within UWO241. The presence of pseudogenes and gene fragments with similarity to IBPs (Appendix A: Table S6) indicates that some previously functional IBP coding regions might have been lost.

These findings add to the growing list of psychrophilic and psychrotolerant algae encoding IBPs (Blanc *et al.* 2012; Raymond and Morgan-Kiss 2013; Mock *et al.* 2017; Raymond and Morgan-Kiss 2017), mirroring the pattern of ice-associated bacteria and fungi (Margesin *et al.* 2008). Genome sequencing of the psychrophilic, polar diatom *Fragilariopsis cylindrus* identified 11 IBPs (Mock *et al.* 2017), almost as many as found in ICE-L (12)(Zhang *et al.* 2020). *Chlamydomonas* sp. ICE-MDV, a close relative of ICE-L and a resident of Lake Bonney (Figure 9A, C, D), currently holds the record for the greatest number of IBP isoforms (50) in a green alga (Raymond and Morgan-Kiss 2017). In all these examples, the IBPs are believed to have been acquired from bacteria via HGT, and their existence is thought to be an adaptation to polar environments (Raymond and Kim 2012). It might seem obvious why a species that lives in the Antarctic would acquire IBPs, which can have ice recrystallization inhibition activities and, thus, protect cells from freezing damage (Davies 2014). However, the potential benefits bestowed upon UWO241 by having these genes is not immediately clear. Unlike ICE-L, UWO241 does not live on ice or snow (Morgan-Kiss *et al.* 2006) but deep within lake water, which remains at ~5 °C year-round.

Given the striking number of IBP genes, however, this is not the only case of HGT in UWO241 genome. Genes encoding for ribosomal proteins were also detected in the HGT list, which is certainly not unheard of (Kondrashov 2012). In addition, stress response-related genes (DnaJ and DnaK) and transporter genes (ABC transporters, sugar transporters and ammonium transporters) appeared at higher frequencies. This suggests that these important biological functions have been enriched by HGT from prokaryotes. The possibility of bacterial contamination can be excluded because most genes of bacterial or

archaeal origin were located on large contigs. Moreover, many genes of bacterial or archaeal origin had acquired introns, further excluding the possibility of contamination.

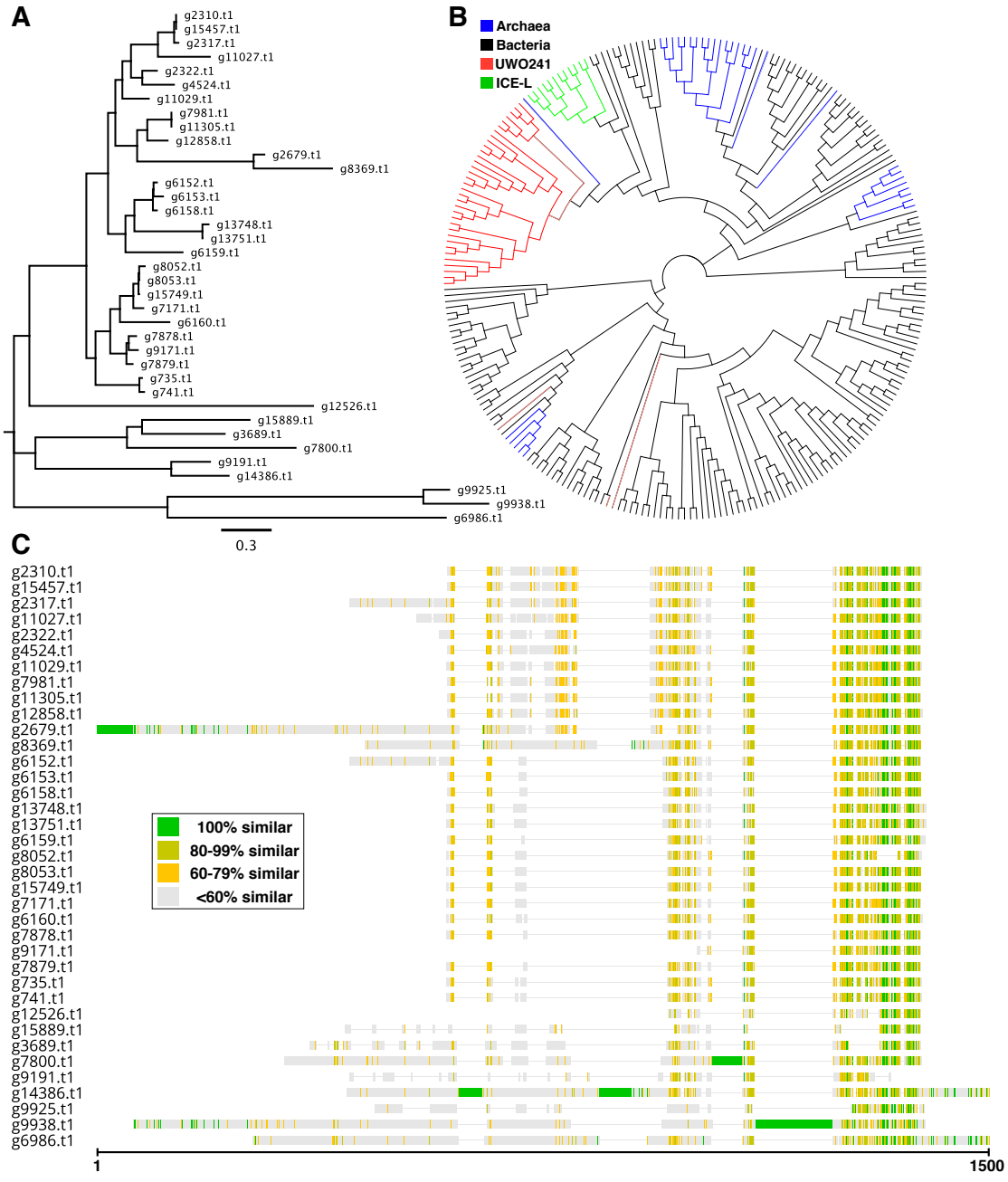


Figure 16: Ice-binding proteins from UWO241.

(A) The Maximum likelihood (ML) phylogenetic trees phylogenetic based on the amino acid alignments of 37 IBPs in UWO241. (B) Phylogenetic relationships of IBPs in

UWO241 (red), ICE-L (green), Archaea (blue) and Bacteria (black). (C) Amino acid alignment of 37 IBPs in UWO241 via Clustal Omega v1.2.4 with default parameters.

To conclude, it is widely believed that IBPs improve survival in sea ice by modifying the structure of the sea ice, trapping water in small brine pockets, and thus preventing the water from draining (Raymond and Kim 2012). This supports the hypothesis that acquisition of IBPs through HGT could improve the survival of UWO241 in cold environments. However, it is interesting to note that despite being localized 17 m below the bottom of the permanent ice surface of Lake Bonney where the water never freezes, UWO241 has retained an impressive number of IBP genes in its genome.

4.2.3 Genome Evolution in a Permanently Ice-covered Antarctic Lake

One must be mindful not to instantly invoke positive selection when trying to explain the evolution of genomic architecture (Lynch 2007; Brunet and Doolittle 2018). It is tempting to propose that pervasive gene duplication within the UWO241 genome is an adaptation to life in Lake Bonney. But one could also reason that these features are neutral (or slightly deleterious) outcomes of random genetic events, such as the whims of selfish elements. As with many aspects of molecular evolution, the truth likely falls somewhere in-between these two extremes.

It is my belief that the underlying mechanisms behind the duplications within the UWO241 nucDNA, be it retrotransposition and/or other processes, are neutral or even maladaptive. Likewise, I contend that most of the observed duplicates in the genome, such as those encoding reverse transcriptases, were fixed through random genetic drift, perhaps exacerbated by the hermetic environment of Lake Bonney. (Unfortunately, there are no data on the effective population size of UWO241 and how it compares to that of other green algae, but it does appear to be rare (Dolhi *et al.* 2015)). But if enough duplicates are generated, eventually one will arise that results in an increase in fitness and, thus, could be maintained through positive selection. For instance, if an increase in dosage of a particular gene is beneficial, then the duplication of this gene could be fixed by positive selection (Innan and Kondrashov 2010; Kondrashov 2012). This is arguably the best explanation for

the existence of the *petf* duplicates (Cvetkovska *et al.* 2018) as well as some of the other HSDs in UWO241, including the IBP genes. It is noteworthy in this context that neither the UWO241 mitochondrial or chloroplast genomes (Cvetkovska *et al.* 2019), contain duplicate genes or retroelement-like sequences. This is different from another chlamydomonadalean green alga *Haematococcus lacustris* with the same repetitive elements spreading throughout the mitochondrial and chloroplast (or plastid) DNA (Zhang *et al.* 2019).

Gene duplication is increasingly being identified as a means for adaptation to extreme environments (Kondrashov 2012; Qian and Zhang 2014). Moreover, duplication events resulting in increased gene dosage are known to play important roles in the initial retention of duplicate genes (Innan and Kondrashov 2010). The data presented here add to this theme. But something neutral can sometimes give rise to something useful. Remarkably, similar evolutionary processes appear to be operating in the ICE-L genome, in which gene duplication, potentially driven by RTs, has led to large expansions in various gene families, including IBP genes (Zhang *et al.* 2020), as well as many HSDs (265 duplicates covering 717 gene copies) (Figure 9D and Appendix A: Table S6). Many of the HSDs in ICE-L have similar functions to those in UWO241 (Appendix A: Table S6). This stands in stark contrast to other green algal nucDNAs, which do not have large numbers of HSDs. Indeed, when the same bioinformatics procedures used to identify and classify HSDs in UWO241 were carried out on available chlamydomonadalean genomes, small to moderate numbers of gene duplications were identified (Figure 9D and Appendix A: Table S6), which is consistent with previous analyses of these genomes and underscores just how unusual the UWO241 and ICE-L genomes are. When comparing the novel impact of HSDs in two psychrophiles UWO241 and ICE-L, the HSDs per Mb are almost 3-fold greater in UWO241 (1.59) than ICE-L (0.49) (Figure 9D). It will be interesting to see if the ICE-MDV genome also harbours expanded gene families and HSDs.

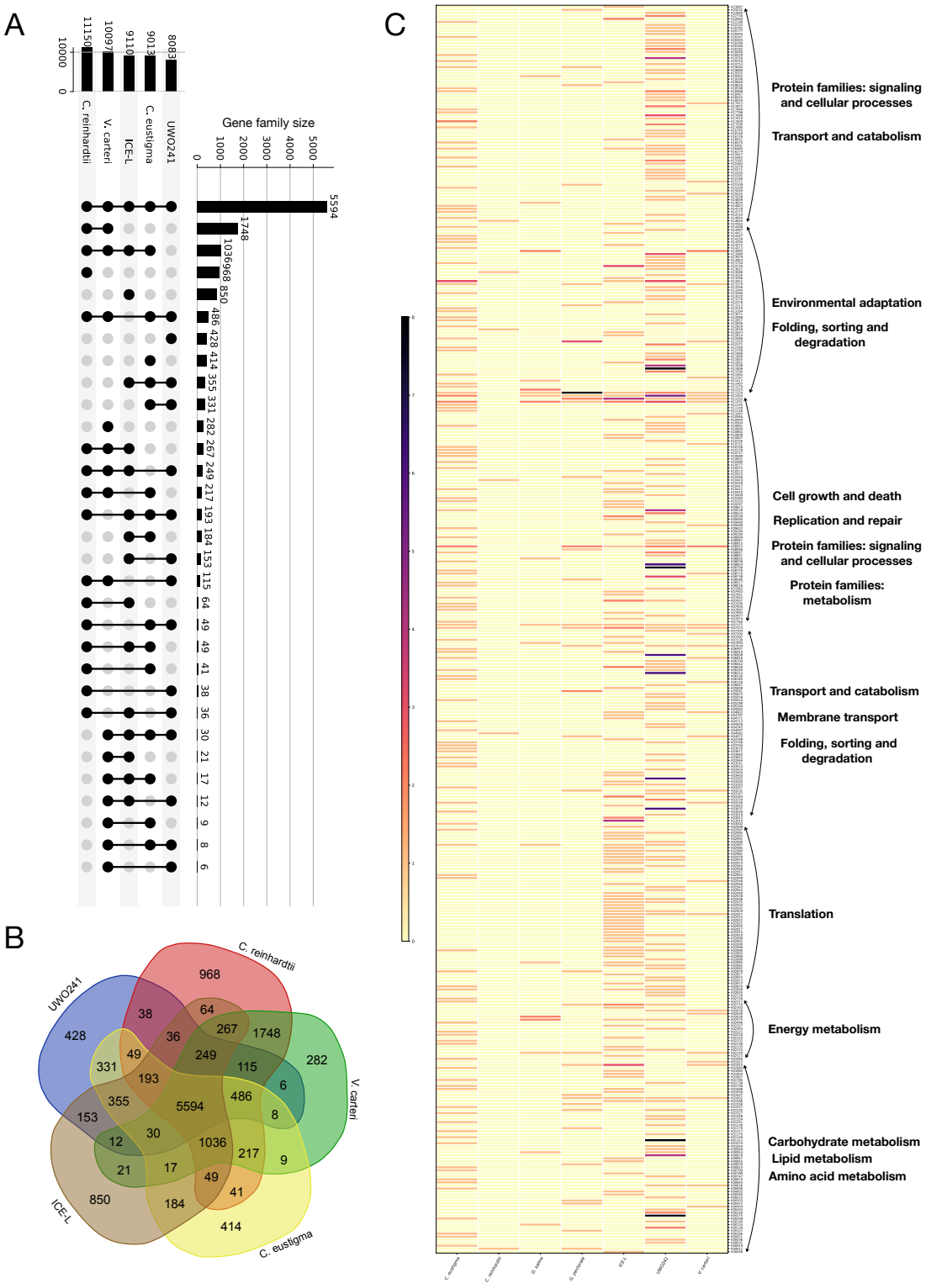


Figure 17: Comparative genomic analysis across algae species.

(A) and (B) The UpSet plot and Venn diagram displaying unique and shared gene families between UWO241 and selected algae species. (C) Number of HSDs across various chlamydomonadalean species grouped based on their KEGG functional category.

Finally, large number of RTs and rampant gene duplication can cause errors during genome assembly (Zimin *et al.* 2017). I performed multiple iterations of the UWO241 assembly, using different protocols and algorithms, and am confident that the available draft genome sequence in GenBank is of good quality. The HSDs, in particular, are supported by RNA-seq, meaning there exists a specific transcript corresponding to each duplicate gene. But given the massive extent of duplications in the UWO241 genome, it is likely that some regions were misassembled, especially segments of duplicated noncoding DNA, and will need to be resolved through subsequent sequencing projects. That said, the overall conclusions presented here should remain the same.

4.2.4 Gene Family Expansions and Contractions Across Species

To explore the gene family evolution in UWO241 and other green algal species, I performed orthologous group analysis among seven species (*Chlamydomonas* sp. UWO241, *C. reinhardtii*, *D. salina* , *V. carteri*, *G. pectorale*, *C. eustigma and Chlamydomonas* sp. ICE-L), which resulted in 14932 orthogroups (Figure 17A, B). The majority of orthogroups (5,594) were shared by all species, with the second most abundant category (1,748) shared by *V. carteri* and *C. reinhardtii*. The orthogroups included sets of genes descended from a common ancestor and encoding the same function in different species. As presented in Figure 17A, considering that UWO241 survives in an environment differing from those of mesophiles, it contains fewer classified orthogroups (8083) with 428 unique categories. Not surprisingly, the previously discussed IBP genes were included within the species-specific orthogroups. More lineage-specific orthogroups and the relationships between the species are illustrated in Venn diagrams (Figure 17B).

Furthermore, to explore the considerable number of genes contributing to gene family expansion, the typical orthogroups associated with functional domains are summarized in the Table 9. Expansions and contractions of orthologous gene families were determined using a birth and death process to model gene gain and loss over a phylogeny (Han *et al.*

2013). UWO241 orthogroups that harbored several domains have expanded in comparison to those in the mesophilic species. For example, reverse transcriptase gene families were expanded in the UWO241 genome; these genes are involved in RNA-mediated transposition, suggesting the availability to generate large numbers of retrocopies. Moreover, the expansion of antenna protein domains in the species increased from lowest (8) in *C. eustigma* to highest (36) in UWO241, but it should be noteworthy that UWO241 is rich of duplicates (as many as six HSDs belong to antenna proteins). The higher number of gene copies might not reflect the real polypeptide level, although 36 genes were classified into families of antenna protein genes with conserved functional domains, which led to the number in the other species. In Figure 17C, the yellow color in the matrix indicates duplicates, and the dark red and purple color indicate the presence of many duplicates. In UWO241, broadscale of red and purple cells are observed.

Table 11: The key expanded gene families in UWO241 genome.

Gene family identifier	UWO241	*C. reinhardtii*	*D. salina*	*C. eustigma*	*G. pectorale*	ICE-L	*V. carteri*	Pfam identifier	Pfam domain description
OG0000021, OG0000040, OG0000168, OG0000461, OG0000742, OG0001396	129	10	2	56	8	2	1	PF00078	Reverse transcriptase
OG0000010, OG0000012	55	59	26	5	53	42	24	PF00125	Core histone H2A/H2B/H3/H4
OG0000218, OG0000026, OG0000080	36	11	10	8	16	28	29	PF00504	Antenna protein
OG0004015, OG0004435, OG0000156, OG0000047, OG0000288, OG0000109, OG0000222	34	12	2	9	31	4	14	PF00069	Protein kinase domain
OG0000103, OG0000121	37	0	0	0	0	12	0	PF11999	Ice-binding proteins (DUF3494)

Although many of the expanded orthogroups were related to functional domains, it is once again very difficult to interpret these results without comparison of the gene expression levels. Nonetheless, the exploration of a comparative framework between UWO241 and

its close mesophilic relatives will greatly aid in the understanding of psychrophily for the future researchers.

4.3 Conclusions

As one of the most comprehensively studied photosynthetic psychrophiles, UWO241 has been studied in detail for 25 years in relation to several important aspects of its biology, including physiology and the molecular biology of photosynthesis. Like its close relative ICE-L, UWO241 encodes a large number (≥37) of ice-binding proteins, putatively originating from horizontal gene transfer. Even more striking, UWO241 harbors hundreds of highly similar duplicate genes involved in diverse cellular processes, some of which I argue are aiding its survival in the Antarctic via gene dosage. Gene and partial gene duplication appear to be an ongoing phenomenon within UWO241, one which might be mediated by retrotransposons. Also, within a comparative genomics framework, UWO241 have the expansion of gene families such as RT, IBPs and antenna protein gene families. Ultimately, I explored how such a process could be associated with adaptation to low temperatures and hypersalinity.

4.4 Methods

4.4.1 Comparative Genomic Analyses

Protein sequences from the nuclear genomes of 7 green algae belonging to the Chlorophyta (*C. reinhardtii* (Merchant *et al.* 2007), *G. pectorale* (Hanschen *et al.* 2016)*, C. eustigma* (Hirooka *et al.* 2017)*, D. salina* (Polle *et al.* 2017), *V. carteri* (Prochnik *et al.* 2010), *Chlamydomonas* sp. ICE-L (Zhang *et al.* 2020) and *Chlamydomonas* sp. UWO241) were used to construct homologous gene clusters (orthogroups) by OrthoFinder v2.1.2 (Emms and Kelly 2015). The longest transcript of each gene was retained to remove redundancy resulting from alternative splicing variations, and genes encoding protein sequences shorter than 50 amino acids were filtered to exclude putative fragmented genes. Orthogroups with single-copy genes shared by all 7 genomes were retained for further analyses. 2123 single-copy genes were retrieved to create a phylogenetic tree. Expansions and contractions of orthologous gene families were determined using CAFÉ v4.1 (Han *et al.* 2013). The

program uses a birth and death process to model gene gain and loss over a phylogeny. Multiple sequence alignments were performed for each orthogroup using Clustal Omega v1.2.4 (Sievers *et al.* 2011) with default parameters. Poorly aligned regions were further trimmed using the trimAl v1.4 (Capella-Gutiérrez *et al.* 2009). Maximum likelihood trees were generated using RAxML v7.0.4 (Stamatakis *et al.* 2004) with the PROTCATJTT model.

4.4.2 Highly Similar Duplicate Genes (HSDs) Predictions

A protein BLAST (Altschul *et al.* 1997) of the UWO241 gene models against themselves (E-value < 1e-5) was filtered to only those with near-identical protein lengths (within 10 amino acids) and ≥90% pairwise identities. This gave a list of highly similar duplicates (HSDs). The deduced amino acid sequences of the HSDs were searched against the KEGG (Kanehisa and Goto 2000) and Pfam databases (Finn *et al.* 2014), providing a functional breakdown. To extensively identify HSDs with high accuracy and reliability, I developed a web-based tool HSDFinder (http://hsdfinder.com) (Zhang *et al.* 2021), which I also used to predict HSDs in other chlorophyte algae. The predicted results are documented in the database of HSDatabase (http://hsdfinder.com/database/) (Zhang *et al.* 2021), which contain total of 28,214 HSDs in fifteen eukaryotes so far. Using HSDFinder, users have the option to employ different parameters (from 50% to 100% identity and from within 0-100 aa variances) for identifying HSDs.

4.4.3 Substitution Rate Analysis of Highly Similar Duplicate Genes (HSDs)

The protein sequences of each HSD gene copy were aligned using Clustal Omega v1.2.4 (Sievers *et al.* 2011); and poorly aligned regions were trimmed with trimAl v1.4 (Capella-Gutiérrez *et al.* 2009). Nonsynonymous (dN) and synonymous (dS) substitution rates were calculated for each HSD group by reverse-translating the amino acid alignments to the corresponding codon-based nucleotide alignments using PAL2NAL (Suyama *et al.* 2006). Maximum likelihood (ML) phylogenetic trees were inferred based on protein and codon alignments using FastTree v2.1 (Price *et al.* 2010) with default parameters. I then applied

the one-ratio model in the codeml program of PAML v4.9 (Yang 2007) to estimate the dN/dS substitution rates (ω value) with the parameters “runmode = 0” and “model =0”.

4.4.4 Horizontal Gene Transfer (Ice-Binding Proteins)

Preliminary BLAST analyses (BLASTP, E-value < 1e-5) showed that a small proportion of genes in the UWO241genome had a top hit to sequences from non-green algae sources, suggesting that these genes might have been acquired through horizontal gene transfer (HGT). Several steps were taken to estimate the overall reliability of HGT. I checked all annotated genes based on their non-redundant annotations, and extracted genes with non-plant annotations (i.e., those matching to fungi, bacteria, archaea and virus) as candidate for further analyses. The BLAST protein databases labeled as fungi, bacteria, archaea, and viruses were downloaded from UniProt (https://www.uniprot.org/downloads) and used to perform BLASTP searches with an E-value < 1e-5. The bit-score of the top ten BLAST hits were extracted as the candidate HGT genes for further analysis. The Clustal Omega v1.2.4 (Capella-Gutiérrez *et al.* 2009) was used to align the candidate HGT genes. Each alignment was trimmed to exclude regions where only one of the sequences was present, and maximum likelihood phylogenetic trees were built using FastTree v2.1 (Price *et al.* 2010) from amino-acids sequences using a WAG+G model (1,000 replicates). The genes for which gene tree supported a sister grouping between UWO241 and a non-plant with support value ≥ 80 were retained as candidate HGT genes.

4.4.5 Reverse Transcriptase Identification (RT)

The standard hallmarks of LTR retrotransposons and non-LTR retrotransposons (e.g., LINE1), such as poly(A) tail insertions and target-site duplications were manually identified in GENEIOUS v10.1 (Kearse *et al.* 2012) based on the sequence alignments and Pfam domains patterns (PF0078 and PF07727) for reverse transcriptase.

4.5 References

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402.

Blanc, G., I. Agarkova, J. Grimwood, A. Kuo, A. Brueggeman, D. D. Dunigan, J. Gurnon, I. Ladunga, E. Lindquist and S. Lucas (2012). The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. Genome Biology 13: 1-12.

Bourque, G., K. H. Burns, M. Gehring, V. Gorbunova, A. Seluanov, M. Hammell, M. Imbeault, Z. Izsvák, H. L. Levin and T. S. Macfarlan (2018). Ten things you should know about transposable elements. Genome Biology 19: 1-12.

Brunet, T. and W. F. Doolittle (2018). The generality of constructive neutral evolution. Biology & Philosophy 33: 1-25.

Capella-Gutiérrez, S., J. M. Silla-Martínez and T. Gabaldón (2009). TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972-1973.

Carelli, F. N., T. Hayakawa, Y. Go, H. Imai, M. Warnefors and H. Kaessmann (2016). The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Research 26: 301-314.

Casola, C. and E. Betrán (2017). The genomic impact of gene retrocopies: what have we learned from comparative genomics, population genomics, and transcriptomic analyses? Genome Biology and Evolution 9: 1351-1373.

Catania, F. and M. Lynch (2008). Where do introns come from? PLoS Biology 6: e283.

Cerbin, S. and N. Jiang (2018). Duplication of host genes by transposable elements. Current Opinion in Genetics & Development 49: 63-69.

Conrad, B. and S. E. Antonarakis (2007). Gene duplication: a drive for phenotypic diversity and cause of human disease. Annual Review of Genomics and Human Genetics 8: 17-35.

Cvetkovska, M., N. P. A. Huner and D. R. Smith (2017). Chilling out: the evolution and diversification of psychrophilic algae with a focus on Chlamydomonadales. Polar Biology 40: 1169-1184.

Cvetkovska, M., S. Orgnero, N. P. Hüner and D. R. Smith (2019). The enigmatic loss of light-independent chlorophyll biosynthesis from an Antarctic green alga in a light-limited environment. New Phytologist 222: 651-656.

Cvetkovska, M., B. Szyszka-Mroz, M. Possmayer, P. Pittock, G. Lajoie, D. R. Smith and N. P. Hüner (2018). Characterization of photosynthetic ferredoxin from the Antarctic alga *Chlamydomonas* sp. UWO241 reveals novel features of cold adaptation. New Phytologist 219: 588-604.

Davies, P. L. (2014). Ice-binding proteins: a remarkable diversity of structures for stopping and starting ice growth. Trends in Biochemical Sciences 39: 548-555.

Del Angel, V. D., E. Hjerde, L. Sterck, S. Capella-Gutierrez, C. Notredame, O. V. Pettersson, J. Amselem, L. Bouri, S. Bocs and C. Klopp (2018). Ten steps to get started in Genome Assembly and Annotation. F1000Research 7: 1-25.

Dolhi, J. M., D. P. Maxwell and R. M. Morgan-Kiss (2013). The Antarctic *Chlamydomonas raudensis*: an emerging model for cold adaptation of photosynthesis. Extremophiles 17: 711-722.

Dolhi, J. M., A. G. Teufel, W. Kong and R. M. Morgan‐Kiss (2015). Diversity and spatial distribution of autotrophic communities within and between ice‐covered A ntarctic lakes (McMurdo Dry Valleys). Limnology and Oceanography 60: 977-991.

Emms, D. M. and S. Kelly (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16: 1-14.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm and J. Mistry (2014). Pfam: the protein families database. Nucleic Acids Research 42: 222-230.

Han, J. S. (2010). Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. Mobile DNA 1: 1-12.

Han, M. V., G. W. Thomas, J. Lugo-Martinez and M. W. Hahn (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Molecular Biology and Evolution 30: 1987-1997.

Hanschen, E. R., T. N. Marriage, P. J. Ferris, T. Hamaji, A. Toyoda, A. Fujiyama, R. Neme, H. Noguchi, Y. Minakuchi and M. Suzuki (2016). The *Gonium pectorale* genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. Nature Communications 7: 1-10.

Hirooka, S., Y. Hirose, Y. Kanesaki, S. Higuchi, T. Fujiwara, R. Onuma, A. Era, R. Ohbayashi, A. Uzuka and H. Nozaki (2017). Acidophilic green algal genome provides insights into adaptation to an acidic environment. Proceedings of the National Academy of Sciences 114: 8304-8313.

Innan, H. and F. Kondrashov (2010). The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics 11: 97-108.

Jąkalski, M., K. Takeshita, M. Deblieck, K. O. Koyanagi, I. Makałowska, H. Watanabe and W. Makałowski (2016). Comparative genomic analysis of retrogene repertoire in two green algae *Volvox carteri* and *Chlamydomonas reinhardtii*. Biology Direct 11: 1-12.

Jiang, N., Z. Bao, X. Zhang, S. R. Eddy and S. R. Wessler (2004). Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569-573.

Kalra, I., X. Wang, M. Cvetkovska, J. Jeong, W. McHargue, R. Zhang, N. Hüner, J. S. Yuan and R. Morgan-Kiss (2020). *Chlamydomonas* sp. UWO 241 exhibits high cyclic electron flow and rewired metabolism under high salinity. Plant Physiology 183: 588-601.

Kanehisa, M. and S. Goto (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28: 27-30.

Kang, L.-F., Z.-L. Zhu, Q. Zhao, L.-Y. Chen and Z. Zhang (2012). Newly evolved introns in human retrogenes provide novel insights into their evolutionary roles. BMC Evolutionary Biology 12: 1-10.

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz and C. Duran (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647-1649.

Keeling, P. J. and J. D. Palmer (2008). Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics 9: 605-618.

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. Proceedings of the Royal Society B: Biological Sciences 279: 5048-5057.

Kono, T. J., A. B. Brohammer, S. E. McGaugh and C. N. Hirsch (2018). Tandem duplicate genes in maize are abundant and date to two distinct periods of time. G3: Genes, Genomes, Genetics 8: 3049-3058.

Kubiak, M. R. and I. Makałowska (2017). Protein-coding genes' retrocopies and their functions. Viruses 9: 1-27.

Libuda, D. E. and F. Winston (2006). Amplification of histone genes by circular chromosome formation in *Saccharomyces cerevisiae*. Nature 443: 1003-1007.

Lisch, D. (2013). How important are transposons for plant evolution? Nature Reviews Genetics 14: 49-61.

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. Proceedings of the National Academy of Sciences 104: 8597-8604.

Margesin, R., F. Schinner, J.-C. Marx and C. Gerday (2008). Psychrophiles: from biodiversity to biotechnology, Springer Verlag, Berlin Heidelberg 1:1-685.

McIntosh, K. B. and P. C. Bonham-Smith (2006). Ribosomal protein gene regulation: what about plants? Botany 84: 342-362.

Merchant, S. S., S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz, G. B. Witman, A. Terry, A. Salamov, L. K. Fritz-Laylin and L. Maréchal-Drouard (2007). The

Chlamydomonas genome reveals the evolution of key animal and plant functions. Science 318: 245-250.

Mock, T., R. P. Otillar, J. Strauss, M. McMullan, P. Paajanen, J. Schmutz, A. Salamov, R. Sanges, A. Toseland and B. J. Ward (2017). Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. Nature 541: 536-540.

Morgan-Kiss, R. M., J. C. Priscu, T. Pocock, L. Gudynaite-Savitch and N. P. Huner (2006). Adaptation and acclimation of photosynthetic microorganisms to permanently cold environments. Microbiology and Molecular Biology Reviews 70: 222-252.

Nefedova, L. and A. Kim (2017). Mechanisms of LTR‐Retroelement Transposition: Lessons from *Drosophila melanogaster*. Viruses 9: 1-11.

Nobrega, M. A. and L. A. Pennacchio (2004). Comparative genomic analysis as a tool for biological discovery. The Journal of Physiology 554: 31-39.

Panchy, N., M. Lehti-Shiu and S.-H. Shiu (2016). Evolution of gene duplication in plants. Plant Physiology 171: 2294-2316.

Polle, J. E., K. Barry, J. Cushman, J. Schmutz, D. Tran, L. T. Hathwaik, W. C. Yim, J. Jenkins, Z. McKie-Krisberg and S. Prochnik (2017). Draft nuclear genome sequence of the halophilic and beta-carotene-accumulating green alga *Dunaliella salina* strain CCAP19/18. Genome Announcements 5: 01105-01117.

Price, M. N., P. S. Dehal and A. P. Arkin (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PloS One 5: e9490.

Prochnik, S. E., J. Umen, A. M. Nedelcu, A. Hallmann, S. M. Miller, I. Nishii, P. Ferris, A. Kuo, T. Mitros and L. K. Fritz-Laylin (2010). Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science 329: 223-226.

Qian, W. and J. Zhang (2008). Gene dosage and gene duplicability. Genetics 179: 2319-2324.

Qian, W. and J. Zhang (2014). Genomic evidence for adaptation by gene duplication. Genome Research 24: 1356-1362.

Raymond, J. A. and H. J. Kim (2012). Possible role of horizontal gene transfer in the colonization of sea ice by algae. PloS One 7: e35968.

Raymond, J. A. and R. Morgan-Kiss (2013). Separate origins of ice-binding proteins in Antarctic *Chlamydomonas* species. PLoS One 8: e59186.

Raymond, J. A. and R. Morgan-Kiss (2017). Multiple ice‐binding proteins of probable prokaryotic origin in an Antarctic lake alga, *Chlamydomonas* sp. ICE‐MDV (Chlorophyceae). Journal of Phycology 53: 848-854.

Saier, M. H. (2000). A functional-phylogenetic classification system for transmembrane solute transporters. Microbiology and Molecular Biology Reviews 64: 354-411.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton and T. A. Graves (2009). The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112-1115.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert and J. Söding (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7: 764-770.

Stamatakis, A., T. Ludwig and H. Meier (2004). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456-463.

Suyama, M., D. Torrents and P. Bork (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research 34: 609-612.

Szcześniak, M. W., J. Ciomborowska, W. Nowak, I. B. Rogozin and I. Makałowska (2011). Primate and rodent specific intron gains and the origin of retrogenes with splice variants. Molecular Biology and Evolution 28: 33-37.

Tan, S., M. Cardoso-Moreira, W. Shi, D. Zhang, J. Huang, Y. Mao, H. Jia, Y. Zhang, C. Chen and Y. Shao (2016). LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. Genome Research 26: 1663-1675.

Tariq, M. and J. Paszkowski (2004). DNA and histone methylation in plants. Trends in Genetics 20: 244-251.

Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante and O. Panaud (2007). A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973-982.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586-1591.

Yang, Z. and J. P. Bielawski (2000). Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution 15: 496-503.

Zhang, C., A. R. Gschwend, Y. Ouyang and M. Long (2014). Evolution of gene structural complexity: an alternative-splicing-based model accounts for intron-containing retrogenes. Plant Physiology 165: 412-423.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends in Ecology & Evolution 18: 292-298.

Zhang, J., T. Zuo and T. Peterson (2013). Generation of tandem direct duplications by reversed-ends transposition of maize Ac elements. PLoS Genetics 9: e1003691.

Zhang, X., Bauman, N., Brown, R., Richardson, T.H., Akella, S., Hann, E., Morey, R., and Smith, D.R. (2019). The mitochondrial and chloroplast genomes of the green alga Haematococcus are made up of nearly identical repetitive sequences. Current Biology 29, R736-R737.

Zhang, X., Cvetkovska, M., Morgan-Kiss, R., Hüner, N.P., and Smith, D.R. (2021). Draft genome sequence of the Antarctic green alga *Chlamydomonas* sp. UWO241. iScience, 102084.

Zhang, X., Y. Hu and D. R. Smith (2021). HSDFinder- an integrated tool to predict highly similar duplicates in eukaryotic genomes. Retrieved from https://github.com/zx0223winner/HSDFinder.

Zhang, X., Hu, Y., and Smith, D.R. (2021). HSDatabase - a database of highly similar duplicate genes in eukaryotic genomes. Retrieved from http://hsdfinder.com/database/.

Zhang, Z., C. Qu, K. Zhang, Y. He, X. Zhao, L. Yang, Z. Zheng, X. Ma, X. Wang and W. Wang (2020). Adaptation to extreme Antarctic environments revealed by the genome of a sea ice green alga. Current Biology 30: 1-12.

Zhu, Z., Y. Zhang and M. Long (2009). Extensive structural renovation of retrogenes in the evolution of the *Populus* genome. Plant Physiology 151: 1943-1951.

Zimin, A. V., D. Puiu, M.-C. Luo, T. Zhu, S. Koren, G. Marçais, J. A. Yorke, J. Dvořák and S. L. Salzberg (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Research 27: 787-792.

Chapter 5

5 Conclusions and Perspectives

Previous findings mainly focused on the physiology and molecular biology of photosynthetic acclimation and adaptation of *Chlamydomonas* sp. UWO241 to low temperature. However, new insights into what makes UWO241 a psychrophile at the genome level, especially nuclear genome, required genome sequencing and gene annotation. I explored the following questions: Does the UWO241 nuclear genome harbor large numbers of duplicate genes? Has it acquired any genes via HGT, such as IBP genes? Does UWO241 contain expanded gene families compared to its close relatives? To answer these questions, I accomplished the following: (1) I acquired a high-quality nuclear genome assembly for UWO241 using next generation sequencing (NGS) and third generation sequencing (TGS) data and (2) I accurately and thoroughly annotated this genome. Therefore, the genome assembly and gene annotation pipelines fed with the best software and algorithms were applied. The nuclear draft genome of approximately 212Mb and as many as 16,325 protein-coding genes were determined. With these data in hand, a comparative genomics framework was established to better understand the evolution of psychrophily. This included a comparison of gene content, such as coding and noncoding DNA, as well as other major genomic architectural features, including duplicate genes, KEGG pathways, Pfam domains and gene families.

I performed a wide range of comparative genomic analyses of the UWO241 genome with those of other model green algae, including *Chlamydomonas reinhardtii*, *Volvox carteri*, *Dunaliella salina*, *Gonium pectorale*, *Chlamydomonas eustigma and Chlamydomonas* sp. ICE-L. My novel findings turn out to be very impressive: (1) UWO241 harbors hundreds of highly similar duplicate genes involved in diverse cellular processes, some of which I argue may aid in the survival of UWO241 in the Antarctic via gene dosage; (2) UWO241 encodes a large number (≥37) of ice-binding proteins, putatively originating from horizontal gene transfer; (3) UWO241 exhibits expanded orthologous gene families of reverse transcriptases, IBPs and antenna proteins. These features suggest the existence of common mechanisms in the adaptation to cold environments and will also help to guide

future investigations of UWO241 and related species. For example, the existence of multiple gene copies encoding cold adapted enzymes can potentially increase the amount of gene product, which might aid the survival availability of UWO241 in cold temperatures. Large scale of IBPs seems to protect the cells from freezing damage due to the ice recrystallization inhibition activities, but it is not immediately clear how the potential benefits are bestowed upon UWO241 by having these genes. Since UWO241 does not live on ice or snow but deep within lake water, which remains at ~5 °C year-round. The expanded gene families of reverse transcriptases potentially contribute to an ongoing phenomenon of partial gene duplication. Most of these reverse transcriptases might be fixed through random genetic drift and perhaps exacerbated by the hermetic environment of Lake Bonney. Notably, UWO241 lives in an environment tolerating multiple stresses: low or high light, excessive ultraviolet (UV) radiation, high or low pH, high osmotic pressure and low nutrients. The expansion of specific gene families might not be associated with single stress. The future comparison of expanded gene families with differentially expressed genes (via experiments under different stresses) can provide more about how gene family expansion might be associated with the adaptation to different stresses of UWO241. In this final chapter, I explored the challenges and opportunities for bioinformatics researchers.

5.1 The Challenges of a Bioinformatics Project

5.1.1 Self-teaching Resources

In Chapter 2, I developed a step-by-step user guide providing a basic bioinformatics foundation in a genome project. It definitely cannot cover everything, but an introduction to the bioinformatic methods used in eukaryotic genome assembly and annotation, enabling a user to gain familiarity with basic analysis steps. Although the technical aspects of genome tools are changing very quickly, it is my hope that this user guide will provide a comprehensive bioinformatics foundation for genome projects specifically for those researchers who have diverse backgrounds but no prior experience in programming.

Nevertheless, there are many bioinformatics workshops operated annually targeting for the ambitious researchers involved in different genome projects. During my Ph.D., I have been

honing my bioinformatics skills by taking graduate courses on the topic as well as two multi-day, hands-on bioinformatics workshops. During the workshop, I was able to work one-on-one with faculty to help draft an appropriate assembly pipeline for the genome project. There are a series of bioinformatic workshops run by the Canadian Bioinformatics Association (CBA) (https://bioinformatics.ca), such as the ones I have attended: "Informatics on High throughput sequencing data" and "Informatics for RNA-Seq analysis".

5.1.2 Intense Computing Clusters

Figure 18: The multicore server of Smith Laboratory server ("in-house" genomics workstation).

Given the confident user guide on a genome project, the importance of computing clusters should not be underestimated. It is commonly known that the sequence coverage relies on the amount of DNA to be sequenced, and the computing hours highly depend on the computing cluster performance. Although smaller data sets can be processed in computing environments with reduced memory resources, such as on a Mac OS X laptop with 8 GB of RAM, it is not enough for a green algal genome with ~230 Mb genome size yielding ~1.6 million PacBio reads (~20 GB compressed document size) and ~193 million Illumina reads (~40 GB compressed document size). Therefore, it is recommended to have ~1 GB of RAM per 1 million paired-end reads. A typical configuration is a multicore server with

256 GB to 1 TB of RAM, and such systems have become more affordable in recent years ($15,000 to $40,000) (Haas *et al.* 2013).

As for the UWO241 genome project, a multicore server ("in-house" genomics workstation) was built with 32 cores, 384 GB RAM and 1TB solid-state drive (Figure 18). Furthermore, there are commercial clusters for researchers who need the required computing resources (e.g., the Amazon cloud http://aws.amazon.com/ec2/). Besides, many universities have the supercomputing clusters available in-house as well, such as SHARCNET (https://www.sharcnet.ca) which I accessed via Ontario of Compute Canada.

5.1.3 Genome Project Pipelines

There are various reputable pipelines and software for the genome project. For instance, MAKER (Cantarel *et al.* 2008) is a portable and easily configurable genome annotation pipeline. MAKER identifies repeats, aligns ESTs and proteins to a genome, produces *ab initio* gene predictions and automatically synthesizes these data into gene annotations having evidence-based quality values. BRAKER2 (Hoff *et al.* 2015) mainly features semi-unsupervised, extrinsic evidence data (RNA-Seq and/or protein spliced alignment information) supported training of GeneMark (Besemer and Borodovsky 2005) and subsequent training of AUGUSTUS (Stanke *et al.* 2006) with integration of extrinsic evidence in the final gene prediction step. The detailed pipelines of the genome projects are always summarized before or after in the supplementary documents. Pipelines might vary among different genome projects due to the differences in software and procedures. It is the responsibility of the author to detail each step, tools and even the parameters to ensure that, other researchers are able to repeat the results and follow the steps properly.

However, researchers easily fall into the trap of chasing the "perfect" data via consistently rerunning the software rather than trying another tool or an additional setting which might produce better results. Any changes to a genome assembly will unfortunately restart the genome annotation from scratch. Therefore, most researchers wish to assemble as completely as possible (frozen assembly) before moving on to genome annotation. To conclude, one stops when the draft genome assembly or/and annotation are able to answer

the biological questions posed. The updated versions of genome can be released subsequently.

5.2 Bonus Pay for Bioinformatics Project

Dealing with bioinformatics projects can produce many challenges. Overcoming these challenges means progress. And surely, there is bonus pay throughout this process. When I explored the duplicates in UWO241 genome, it was challenging to classify the protein BLAST result of the UWO241 gene models against themselves. Especially I wanted to filter to only those duplicates with near-identical protein lengths (within certain amino acids) and certain pairwise identities. Therefore, to extensively identify highly similar duplicates (HSDs) with high accuracy and reliability, I developed a web-based tool HSDFinder (http://hsdfinder.com) (Zhang *et al.* 2021). Using HSDFinder, users have the option to employ different parameters (from 30% to 100% identity and from within 0-100 aa variances) for identifying HSDs. What's more, I also used the tool to predict HSDs in other chlorophyte algae. The predicted results are documented in the database of HSDatabase (Zhang *et al.* 2021), which contain a total of 28,214 HSDs in fifteen eukaryotes so far (http://hsdfinder.com/database/).

Functional annotations of protein-coding genes can be annoying when obtaining the best BLAST hits from some non-redundant protein sequence database such as NCBI NR databases, SwissProt (Consortium 2019) and TrEMBL (Boeckmann *et al.* 2003), because of the hypothetical and uncharacterized proteins might pop up at the top list. I developed a tool called NoBadWordsCombiner v1.0 (http://hsdfinder.com/combiner/) (Zhang *et al.* 2021), which can automatically merge the BLAST results from the databases of SwissProt (Consortium 2019), TrEMBL (Boeckmann *et al.* 2003) and NCBI NR databases. More importantly, it can strengthen the gene definition by filtering those protein function descriptions containing 'bad words', such as hypothetical and uncharacterized proteins.

5.3 Bioinformatics as A Career

Researchers with bioinformatics expertise are thought to be an asset in today's job market, especially with the increasing demand for the large NGS and TGS datasets analysis. A

bioinformatics job often requires candidates with certain backgrounds or skills. This is usually not limited to a degree in Life Science, but a quantitative discipline, such as Bioinformatics, Computer Science, Statistics or Molecular Biology. If this job is targeted for specialized projects such as precision medicine and regulatory genomics in cancer and COVID-19, the candidates better hope that they are equipped with the knowledge of Cancer Genomics or Virology. If this is a senior position, the prospective employee has to prove certain minimum years' experience of bioinformatics projects.

The responsibility of a bioinformatics job might vary between the different positions. The investigations of the bioinformatics job market have given me some common insights into these issues, such as being able to write or assist manuscripts for publication, present the project in lab meeting or scientific conferences, and collaborate with local or international wet-lab researchers. If the job needs you to be comfortable with the coding environment, you have to be proficient in programming (e.g., Python, Perl, and C++) and have experience working in a Unix/Linux computing environment with large datasets. Apart from the technical levels, employers usually prefer candidates who exhibit independent thinking and involvement in the design of future research projects, have a fellowship or have applied for and successfully obtained a fellowship, or even experience in a supervisory role.

The qualification of a bioinformatics job is for candidates to evaluate themselves whether meeting the employer's requirements or not. This is also why candidates are eager to improve their curriculum vitae (CV). Firstly, a degree in related field is usually needed (e.g., Biology, Computer Science, Statistics, Bioinformatics etc.). Then, the previous experience of published work in peer-reviewed journals (e.g., the number of first author publication(s) in good journals). Lastly, the strong background or minimum years' experience of programming related projects. Whether you believe or not, the area of bioinformatics is booming and continuing to grow with high demand and excellent salaries in the coming years.

5.4 References

Besemer, J. and M. Borodovsky (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Research 33: W451-W454.

Boeckmann, B., A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'donovan and I. Phan (2003). The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31: 365-370.

Cantarel, B. L., I. Korf, S. M. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. S. Alvarado and M. Yandell (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research 18: 188-196.

Consortium, U. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47: D506-D515.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li and M. Lieber (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8: 1494-1512.

Hoff, K. J., S. Lange, A. Lomsadze, M. Borodovsky and M. Stanke (2015). BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32: 767-769.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack and B. Morgenstern (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Research 34: W435-W439.

Zhang, X., Y. Hu and D. R. Smith (2021). HSDFinder- an integrated tool to predict highly similar duplicates in eukaryotic genomes. Retrieved from https://github.com/zx0223winner/HSDFinder.

Zhang, X., Hu, Y., and Smith, D.R. (2021). HSDatabase - a database of highly similar duplicate genes in eukaryotic genomes. Retrieved from http://hsdfinder.com/database/.

Zhang, X., Y. Hu and D. R. Smith (2021). NoBadWordsCombiner-a tool to integrate the gene function information together without 'bad words' from Nr-NCBI, UniProtKB/Swiss-Prot, KEGG, Pfam databases. . Retrieved from https://github.com/zx0223winner/HSDFinder/blob/master/NoBadWordsCombiner.py.

Appendices

Appendix A: List of supplementary tables for each chapter.

Table S1: A list of selected bacterial, archaeal and eukaryotic psychrophiles and psychrotrophs (Chapter 1).

Species	Temperature Range	Environment	Phylum	Class	Order	GOLD Organism ID*	NCBI Taxonomy ID**	References
Prokaryotes								
Cenarchaeum symbiosum A	Psychrophile	Marine	Thaumarchaeota	unclassified Thaumarchaeota	Cenarchaeales	Go0000220	414004	(Hallam *et al.* 2006)
Flavobacterium psychrophilum ATCC 49418	Psychrotolerant	Unclassified	Bacteroidetes	Flavobacteriia	Flavobacteriales	Go0095220	96345	(Wu *et al.* 2015)
Flavobacterium psychrophilum JIP02/86	Psychrophile	Excretory system	Bacteroidetes	Flavobacteriia	Flavobacteriales	Go0000127	402612	(Duchaud *et al.* 2007)
Glaciecola sp. HTCC 2999	Psychrophile	Unclassified	Proteobacteria	Gammaproteobacteria	Alteromonadales	Go0001342	455436	(Beier *et al.* 2015)
Lacinutrix jangbogonensis PAMC 27137	Psychrophile	Unclassified	Bacteroidetes	Flavobacteriia	Flavobacteriales	Go0109128	1469557	(Lee *et al.* 2014)
Methanococcoides burtonii DSM 6242	Psychrophile	Freshwater	Euryarchaeota	Methanomicrobia	Methanosarcinales	Go0000367	259564	(Byrne-Steele *et al.* 2009)
Methanogenium frigidum Ace-2	Psychrophile	Unclassified	Euryarchaeota	Methanomicrobia	Methanomicrobiales	Go0002320	313587	(Franzmann *et al.* 1997)
Polaribacter filamentus	Psychrophile	Unclassified	Bacteroidetes	Flavobacteriia	Flavobacteriales	Go0001780	53483	(Yoon *et al.* 2006)
Pseudoalteromonas haloplanktis TAC125	Psychrophile	Marine	Proteobacteria	Gammaproteobacteria	Alteromonadales	Go0000456	326442	(Médigue *et al.* 2005)

Pseudomonas psychrophila HA-4	Psychrotolerant	Activated Sludge	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Go0024001	1211112	(Jiang *et al.* 2012)
Psychrobacter arcticus 273-4	Psychrophile	Soil	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Go0000467	259536	(Ayala-del-Río *et al.* 2010)
Psychrobacter phenylpyruvicus	Psychrotolerant	Circulatory system	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Go0023726	1123034	(Deschaght *et al.* 2012)
Psychrobacter sp. PAMC 21119	Psychrophile	Soil	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Go0017610	1112209	(Kim *et al.* 2012)
Psychroflexus torquis ATCC 700755	Psychrophile	Marine	Bacteroidetes	Flavobacteriia	Flavobacteriales	Go0001803	313595	(Bowman *et al.* 2006)
Psychromonas ingrahamii 37	Psychrophile	Marine	Proteobacteria	Gammaproteobacteria	Alteromonadales	Go0000238	357804	(Riley *et al.* 2008)
Rhodoferax antarcticus ANT.BR	Psychrotolerant	Unclassified	Proteobacteria	Betaproteobacteria	Burkholderiales	Go0006338	1111071	(Zhao 2011)
Rhodonellum psychrophilum GCM71, DSM 17998	Psychrophile	Geologic	Bacteroidetes	Cytophagia	Cytophagales	Go0013220	1123057	(Schmidt *et al.* 2006)
Eukaryotes								
Chlamydomonas sp. ICE-L	Psychrophile	Marine	Chlorophyta	Chlorophyceae	Chlamydomonadales	NA	309537	(Zhang *et al.* 2020)
Coccomyxa subellipsoidea C-169	Psychrotolerant	Marine	Chlorophyta	Trebouxiophyceae	Trebouxiophyceae incertae sedis	Gs0000070	574566	(Blanc *et al.* 2012)
Fragilariopsis cylindrus	Psychrophile	Marine	Ochrophyta	Bacillariophyceae	Bacillariales	Gs0014619	186039	(Mock *et al.* 2017)
Chlamydomonas sp. UWO241	Psychrophile	Freshwater	Chlorophyta	Chlorophyceae	Chlamydomonadales	NA	1653778	(Pocock *et al.* 2004)
Chlamydomonas nivalis	Psychrophile	Freshwater	Chlorophyta	Chlorophyceae	Chlamydomonadales	NA	47906	(Remias *et al.* 2005)

* GOLD Organism ID is from the US Genomes OnLine database (GOLD), which collects the information of sequencing genome projects.

** NCBI Taxonomy ID is the identifier for a taxon in the Taxonomy Database by the US National Center for Biotechnology Information (NCBI)

Table S2: A list of published green algal genomes predicted from whole-genome sequencing projects (as of Oct. 2020). Accession numbers are from the US National Center for Biotechnology Information (NCBI) (adapted from (Blaby-Haas and Merchant 2019)) (Chapter 1).

Taxonomy				General features				Source	
Phylum	Class	Order	Organism*	Genome size (Mb)	Morphology	Environment	Mesophiles vs Extremophiles	Accession number**	References
Chlorophyta	Chlorophyceae	Chlamydomonadales	*Chlamydomonas reinhardtii*	111.1	Unicellular	Fresh water	Mesophile	V5.6 (Phytozome)	(Merchant *et al.* 2007)
			Volvox carteri	131.2	Multicellular	Fresh water	Mesophile	V2.1 (Phytozome)	(Prochnik *et al.* 2010)
			Chlamydomonas eustigma	66.6	Unicellular	Acidic	Acidophile	GCA_002335675.1	(Hirooka *et al.* 2017)
			Dunaliella salina	343.7	Unicellular	Salt water	Halophile	GCA_002284615.1	(Polle *et al.* 2017)
			Gonium pectorale	148.8	Colonial	Fresh water	Mesophile	GCA_001584585.1	(Hanschen *et al.* 2016)
			Chlamydomonas sp. ICE-L	541.8	Unicellular	Salt water	Psychrophile	GCA_013435795.1	(Zhang *et al.* 2020)
			Tetrabaena socialis	135.7	Colonial	Fresh water	Mesophile	GCA_002891735.1	(Featherston *et al.* 2016)
		Sphaeropleales	*Monoraphidium neglectum*	69.7	Unicellular	Fresh water	Mesophile	GCA_000611645.1	(Bogen *et al.* 2013)
			Chromochloris zofingiensis	60.1	Unicellular	Soil	Mesophile	V5.2.3.2 (Phytozome)	(Roth *et al.* 2017)
			Raphidocelis subcapitata	51.2	Unicellular	Fresh water	Mesophile	GCA_003203535.1	(Suzuki *et al.* 2018)
	Mamiellophyceae	Mamiellales	*Bathycoccus prasinos*	15.1	Unicellular	Salt water	Halophile	GCA_002220235.1	(Moreau *et al.* 2012)

			Micromonas sp. RCC299	21.1	Unicellular	Salt water	Halophile	GCA_000090985.2	(Worden *et al.* 2009)
			Micromonas sp. CCMP1545	21.9	Unicellular	Salt water	Halophile	GCA_000151265.1	(Worden *et al.* 2009)
			Micromonas sp. ASP10–01a	19.6	Unicellular	Salt water	Halophile	GCA_001430725.1	(Benites *et al.* 2019)
			Ostreococcus lucimarinus	13.2	Unicellular	Salt water	Halophile	GCA_000092065.1	(Palenik *et al.* 2007)
			Ostreococcus tauri RCC4221	13.0	Unicellular	Salt water	Halophile	GCF_000214015.3	(Blanc-Mathieu *et al.* 2014)
			Ostreococcus tauri RCC1115	14.8	Unicellular	Salt water	Halophile	GCA_002158475.1	(Clerissi *et al.* 2012)
	Trebouxiophyceae	Chlorellales	*Chlorella variabilis*	46.2	Unicellular	Fresh water	Mesophile	GCA_000147415.1	(Blanc *et al.* 2010)
			Auxenochlorella protothecoides	22.9	Unicellular	Soil	Mesophile	GCA_000733215.1	(Gao *et al.* 2014)
			Chlorella sorokiniana	59.6	Unicellular	Fresh water	Mesophile	GCA_002245835.2	(Arriola *et al.* 2018)
			Micractinium conductrix	61.0	Unicellular	Fresh water	Mesophile	GCA_002245815.1	(Arriola *et al.* 2018)
			Helicosporidium sp. ATCC 50920	12.4	Unicellular	Insect larva	Mesophile	GCA_000690575.1	(Pombert *et al.* 2014)
		Trebouxiophyceae incertae sedis	*Chloroidium* sp. JM	60.4	Unicellular	Fresh water	Mesophile	GCA_004335615.1	(Nelson *et al.* 2019)
			Chloroidium sp. CF	54.3	Unicellular	Fresh water	Mesophile	GCA_004335625.1	(Nelson *et al.* 2019)
			Coccomyxa subellipsoidea C-169	48.8	Unicellular	Fresh water	Psychrotroph	GCA_000258705.1	(Blanc *et al.* 2012)
			Picochlorum SENEW3	13.4	Unicellular	Salt water	Halophile	GCA_000876415.1	(Foflonker *et al.* 2015)

			Picochlorum soloecismus	15.2	Unicellular	Salt water	Halophile	GCA_002818215.1	(Huesemann *et al.* 2017)

* The published green algae genomes are collected upon the day of written (Oct. 2020).

** Accession numbers are from the US National Center for Biotechnology Information (NCBI) GenBank assembly accession numbers or the US Department of Energy's Joint Genome Institute Phytozome assembly version numbers.

Table S3: Gene and partial gene duplicates in UWO241 and their proximity to reverse transcriptase (RT) genes. Download the complete table via the link: https://drive.google.com/file/d/1HTr-E3fFYj8eAkM6kRJLiF16Y80FTF8p/view?usp=sharing (**Chapter 3, 4**)

HSDs or RTs or Partial duplicates	Gene model identifier	Gene copies	Amino acid length of gene copie	Pfam identifier	Pfam Description
Partial duplicates	g12.t1	g12.t1; g15096.t1; g13.t1; g14453.t1	; PF00651; PF00651; PF00651	; BTB/POZ domain; BTB/POZ domain;	; IPR000210; IPR000210; IPR0002
Partial duplicates	g26.t1	g26.t1; g4563.t1			
HSD	g38.t1	g38.t1; g7812.t1; g8958.t1; g9137.t1; g1138	PF00690; ; ; PF02696; PF00580; F	Cation transporter/ATPase, N-terminu	IPR004014; ; ; IPR003846; IPR034
Partial duplicates	g43.t1	g43.t1; g12410.t1	PF12165; PF12165	Alfin; Alfin	IPR021998; IPR021998
Partial duplicates	g56.t1	g56.t1; g3339.t1; g2866.t1			
HSD	g62.t1	g62.t1; g306.t1; g458.t1; g539.t1; g541.t1; g	PF14360; ; ; PF00078; PF00078;	PAP2 superfamily C-terminal; ; ; Reve	IPR025749; ; ; IPR000477; IPR00(
Partial duplicates	g70.t1	g70.t1; g4419.t1; g10898.t1; g136.t1; g494.	PF00067; ; PF13344; PF00962; ; F	Cytochrome P450; ; Haloacid dehalog	IPR001128; ; IPR006357; IPR0013(
Partial duplicates	g71.t1	g71.t1; g5045.t2	PF00211, PF13416; PF01547, PF0	Adenylate and Guanylate cyclase cata	IPR001054, IPR006059; IPR00605
HSD	g79.t1	g79.t1; g13636.t1; g6911.t1; g4525.t1; g135	PF14931, PF00078; PF00078; PF0	Intraflagellar transport complex B, sul	IPR028172, IPR000477; IPR00047
RT	g79.t1		PF00078,PF14931	Reverse transcriptase (RNA-depender	Cre11.g467523.t1.1
HSD	g94.t1	g94.t1; g499.t1	PF00673, PF00281; PF00673, PF0	ribosomal L5P family C-terminus, Ribo	IPR031309, IPR031310; IPR03130
HSD	g95.t1	g95.t1; g500.t1	PF00467, PF08071, PF01479, PF0(	KOW motif, RS4NT (NUC023) domain,	IPR005824, IPR013843, IPR00294
Partial duplicates	g104.t2	g104.t2; g3155.t1			
Partial duplicates	g111.t1	g111.t1; g10839.t1; g2208.t1; g2398.t1; g12	PF00443; PF00034, PF00078; PF0	Ubiquitin carboxyl-terminal hydrolase,	IPR001394; IPR009056, IPR00047
HSD	g113.t1	g113.t1; g3523.t3; g15321.t1; g12102.t1; g1	; PF00009; PF02373; PF01026; ;	; Elongation factor Tu GTP binding do	; IPR000795; IPR003347; IPR0011
Partial duplicates	g118.t1	g118.t1; g1369.t1; g2474.t1; g5130.t1; g732	0.t1; g3337.t1; g6308.t1; g10668.t	1; g1009.t1; g7846.t1; g6082.t1	
Partial duplicates	g122.t1	g122.t1; g7023.t2; g14462.t1	PF08392, PF02797; PF02797, PF0	FAE1/Type III polyketide synthase-like	IPR013601, IPR012328; IPR01232
HSD	g128.t1	g128.t1; g9992.t1; g8278.t2; g3804.t1; g380	; ; PF01073; ; PF00211	; ; 3-beta hydroxysteroid dehydrogena	; ; IPR002225; ; IPR001054
Partial duplicates	g129.t2	g129.t2; g3104.t1	PF04366; PF04366	Las17-binding protein actin regulator;	IPR007461; IPR007461
HSD	g131.t1	g131.t1; g15093.t1	PF00078; PF00078	Reverse transcriptase (RNA-depender	IPR000477; IPR000477
RT	g131.t1		PF00078	Reverse transcriptase (RNA-depender	Cre11.g467523.t1.1
HSD	g132.t1	g132.t1; g3556.t1	PF14775, PF14772; PF03079	Sperm tail C-terminal domain, Sperm t	IPR029440, IPR039505; IPR00431
HSD	g136.t1	g136.t1; g494.t1; g927.t1; g1329.t1; g1694.	PF00962; ; PF00171; PF00078; PI	Adenosine/AMP deaminase; ; Aldehyd	IPR001365; ; IPR015590; IPR0004
HSD	g168.t1	g168.t1; g11892.t1	PF00078; PF00078	Reverse transcriptase (RNA-depender	IPR000477; IPR000477
RT	g168.t1		PF00078	Reverse transcriptase (RNA-depender	Cre11.g467523.t1.1
Partial duplicates	g181.t1	g181.t1; g7140.t1	PF13640; PF13640	2OG-Fe(II) oxygenase superfamily; 2O	IPR005123; IPR005123
HSD	g189.t1	g189.t1; g5647.t1	;	;	;
HSD	g198.t1	g200.t1; g199.t1; g198.t1	PF01716; PF01716; PF01716	Manganese-stabilising protein / photo	IPR002628; IPR002628; IPR00262
Partial duplicates	g198.t1	g200.t1; g199.t1; g198.t1; g7556.t1	PF01716; PF01716; PF01716; PF0	Manganese-stabilising protein / photo	IPR002628; IPR002628; IPR00262

Table S4: Gene models and their functional descriptions in the UWO241 genome. Download the complete table via the link: https://drive.google.com/file/d/1XzjQNYwWoBNYsCBnQd8LBZZ0n8OdMw4N/view?usp=sharing **(Chapter 3, 4)**

Gene model identifier (The lon	Length in amino acids (aa)	BLASTP hit identifier retrieved from N	BLASTP hit description retrieved from NCBI nr database	BLASTP amino acid identity (%)	BLASTP eValue	BLASTP hit identifier retriev
g1.t1	272	gi|1238995578|dbj|GAX75978.1|	hypothetical protein CEUSTIGMA_g3421.t1 [Chlamydomonas	54.26	1.41E-75	
g2.t1	132	gi|1183350135|gb|ORX78377.1|	ankyrin, partial [Anaeromyces robustus]	40.23	3.61E-10	sp|Q05921|RN5A_MOUSE
g3.t1	1188	gi|1238995576|dbj|GAX75976.1|	hypothetical protein CEUSTIGMA_g3419.t1 [Chlamydomonas	38.46	1.15E-39	sp|O04716|MSH6_ARATH
g4.t1	320	gi|1238995575|dbj|GAX75975.1|	hypothetical protein CEUSTIGMA_g3418.t1 [Chlamydomonas	89.51	1.17E-97	sp|P15170|ERF3A_HUMAN
g5.t1	118					
g6.t1	96	gi|1335042461|gb|PNW77074.1|	hypothetical protein CHLRE_10g421079v5 [Chlamydomonas r	58.33	1.66E-18	
g7.t2	2654	gi|1238994727|dbj|GAX76500.1|	hypothetical protein CEUSTIGMA_g3945.t1 [Chlamydomonas	32.68	7.48E-34	
g8.t1	132					
g9.t1	156	gi|1238995573|dbj|GAX75973.1|	hypothetical protein CEUSTIGMA_g3416.t1 [Chlamydomonas	62.12	3.00E-49	sp|P72673|Y729_SYNY3
g10.t1	608	gi|1004134917|gb|KXZ42995.1|	hypothetical protein GPECTOR_108g190 [Gonium pectorale]	78.83	1.18E-103	sp|O94530|SUA5_SCHPO
g11.t1	89					
g12.t1	473	gi|1238987328|dbj|GAX83929.1|	hypothetical protein CEUSTIGMA_g11353.t1 [Chlamydomona	23.44	5.64E-13	
g13.t1	474	gi|1238992126|dbj|GAX79241.1|	hypothetical protein CEUSTIGMA_g6681.t1 [Chlamydomonas	26.91	4.51E-06	
g14.t1	332					
g15.t1	89	gi|545366047|ref|XP_005647946.1|	hypothetical protein COCSUDRAFT_65897 [Coccomyxa subelli	59.32	5.40E-14	sp|A8I6P9|SC61B_CHLRE
g16.t1	298	gi|159487763|ref|XP_001701892.1|	predicted protein, partial [Chlamydomonas reinhardtii]	52.54	7.95E-74	sp|Q8LAN3|P4H4_ARATH
g17.t1	367	gi|1238991300|dbj|GAX80092.1|	hypothetical protein CEUSTIGMA_g7530.t1 [Chlamydomonas	60.00	1.73E-13	sp|Q8L4M6|GATA3_ARATH
g18.t1	334					
g19.t1	1994	gi|1238985607|dbj|GAX85598.1|	hypothetical protein CEUSTIGMA_g13013.t1 [Chlamydomona	43.85	1.90E-54	sp|Q5QD03|SUVH3_CHLRE
g20.t2	1481					
g21.t1	139					
g22.t1	2694	gi|1238996294|dbj|GAX75027.1|	hypothetical protein CEUSTIGMA_g2473.t1 [Chlamydomonas	47.76	0.00E+00	sp|Q00808|HETE1_PODAS
g23.t1	713	gi|1238996007|dbj|GAX75537.1|	hypothetical protein CEUSTIGMA_g2980.t1 [Chlamydomonas	43.45	5.73E-147	
g24.t1	264					
g25.t1	654					
g26.t1	70					
g27.t1	392	gi|1238989081|dbj|GAX82253.1|	hypothetical protein CEUSTIGMA_g9681.t1 [Chlamydomonas	54.89	1.09E-103	sp|Q7TT23|CT194_MOUSE
g28.t1	1225	gi|1238989081|dbj|GAX82253.1|	hypothetical protein CEUSTIGMA_g9681.t1 [Chlamydomonas	51.52	1.01E-111	sp|Q7TT23|CT194_MOUSE
g29.t1	101					
g30.t1	336	gi|1238989080|dbj|GAX82252.1|	hypothetical protein CEUSTIGMA_g9680.t1 [Chlamydomonas	45.09	1.05E-44	sp|Q54KA7|SECG_DICDI
g31.t1	209					
g32.t1	944	gi|1238989077|dbj|GAX82249.1|	hypothetical protein CEUSTIGMA_g9677.t1 [Chlamydomonas	67.78	0.00E+00	sp|P42730|CLPB1_ARATH
g33.t3	312	gi|1238989076|dbj|GAX82248.1|	hypothetical protein CEUSTIGMA_g9676.t1 [Chlamydomonas	60.21	2.56E-70	
g34.t1	264	gi|929742606|ref|XP_014146287.1|	hypothetical protein SARC_15057, partial [Sphaeroforma arcti	54.84	6.17E-35	sp|B1JJB5|KATG_YERPY
g35.t1	1065	gi|1238995997|dbj|GAX75527.1|	hypothetical protein CEUSTIGMA_g2970.t1 [Chlamydomonas	63.94	9.99E-160	sp|Q90640|KIF4_CHICK
g36.t1	435	gi|1238995999|dbj|GAX75529.1|	hypothetical protein CEUSTIGMA_g2972.t1 [Chlamydomonas	77.63	6.93E-118	
g37.t1	313	gi|1238994773|dbj|GAX76546.1|	hypothetical protein CEUSTIGMA_g3992.t1 [Chlamydomonas	41.89	3.03E-50	

Table S5: Highly similar duplicate genes (HSDs) in UWO241. Download the complete table via the link: https://drive.google.com/file/d/1y4I3FVJQNeJ-36e_4521hRy63Dlv3nO0/view?usp=sharing **(Chapter 3, 4)**

Highly Similar Duplicates (HSDs) identifiers	HSDs gene copies (within 10 amino acids, ≥90% pairwise identities)	Amino acid length of HSDs gene copies (aa)	Pfam identifier	Pfam Description
UWO241_HSD_1	g1516.t1; g15297.t1; g3710.t1; g15900.t1; g12375.t1; g8654.t1; g194	228; 228; 228; 228; 228; 229; 233; 233	PF00098; PF00098; PF00098; PF00098;	Zinc knuckle; Zinc knuckle; Zinc knuckle; Zinc knuckle; Zinc knuckle; Zinc knuckle; ;
UWO241_HSD_2	g11310.t1; g11375.t1	1307; 1312	PF00098, PF00665, PF07727, PF13976;	Zinc knuckle, Integrase core domain, Reverse transcriptase (RNA-dependent DNA
UWO241_HSD_3	g807.t1; g4057.t1	464; 469	PF14240; PF14240	YHYH protein; YHYH protein
UWO241_HSD_4	g5701.t1; g9150.t2	884; 885	PF00400;	WD domain, G-beta repeat;
UWO241_HSD_5	g15539.t1; g767.t1	231; 231	PF10260; PF10260	Uncharacterized conserved domain (SAYSvFN); Uncharacterized conserved domai
UWO241_HSD_6	g5920.t1; g5844.t1	256; 256	PF02902; PF02902	Ulp1 protease family, C-terminal catalytic domain; Ulp1 protease family, C-termina
UWO241_HSD_7	g12590.t1; g6100.t1	159; 159	PF00179; PF00179	Ubiquitin-conjugating enzyme; Ubiquitin-conjugating enzyme
UWO241_HSD_8	g3684.t1; g6795.t1	137; 130	PF00240, PF01020; PF00240, PF01020	Ubiquitin family, Ribosomal L40e family; Ubiquitin family, Ribosomal L40e family
UWO241_HSD_9	g5645.t1; g15870.t2	599; 605	PF00443; PF00443	Ubiquitin carboxyl-terminal hydrolase; Ubiquitin carboxyl-terminal hydrolase
UWO241_HSD_10	g2201.t1; g15994.t1; g15997.t1; g15991.t1	442; 442; 442; 442	PF00091, PF03953; PF00091, PF03953;	Tubulin/FtsZ family, GTPase domain, Tubulin C-terminal domain; Tubulin/FtsZ fami
UWO241_HSD_11	g4816.t1; g4805.t1; g4802.t1	450; 450; 450	PF00091, PF03953; PF00091, PF03953;	Tubulin/FtsZ family, GTPase domain, Tubulin C-terminal domain; Tubulin/FtsZ fami
UWO241_HSD_12	g1131.t1; g9728.t1	1744; 1743	PF03151; PF00078	Triose-phosphate Transporter family; Reverse transcriptase (RNA-dependent DNA
UWO241_HSD_13	g9104.t1; g645.t1	196; 196	PF07500; PF07500	Transcription factor S-II (TFIIS), central domain; Transcription factor S-II (TFIIS), ce
UWO241_HSD_14	g15800.t1; g12147.t1	1638; 1643	PF14249, PF00211; PF00069	Tocopherol cyclase, Adenylate and Guanylate cyclase catalytic domain; Protein kin
UWO241_HSD_15	g5257.t1; g13535.t1; g7304.t1; g14487.t1	1296; 1305; 1314; 1317	PF04278; PF00078; PF00069, PF00078;	Tic22-like family; Reverse transcriptase (RNA-dependent DNA polymerase); Protei
UWO241_HSD_16	g8742.t1; g8510.t1	296; 296	PF00082; PF00082	Subtilase family; Subtilase family
UWO241_HSD_17	g13122.t1; g13744.t1; g12836.t1; g4052.t1; g15392.t1; g13707.t1; g13	344; 349; 348; 339; 355; 338; 336; 354; 333	PF00588; PF00588; ; ; ; PF00514; ; ;	SpoU rRNA Methylase family; SpoU rRNA Methylase family; ; ; ; Armadillo/beta-ca
UWO241_HSD_18	g132.t1; g3556.t1	806; 801	PF14775, PF14772; PF03079	Sperm tail C-terminal domain, Sperm tail; ARD/ARD' family
UWO241_HSD_19	g3054.t1; g11238.t1	306; 306	PF16891, PF00149; PF00149, PF16891	Serine-threonine protein phosphatase N-terminal domain, Calcineurin-like phosph
UWO241_HSD_20	g429.t1; g3694.t1	930; 937	PF00530, PF00082; PF00225	Scavenger receptor cysteine-rich domain, Subtilase family; Kinesin motor domain
UWO241_HSD_21	g10399.t1; g10296.t1; g10295.t1; g4237.t1	366; 366; 366; 366	PF13445; PF13445; PF13445; PF13445	RING-type zinc-finger; RING-type zinc-finger; RING-type zinc-finger; RING-type zi
UWO241_HSD_22	g11990.t1; g4365.t1	284; 284	PF13639; PF13639	Ring finger domain; Ring finger domain
UWO241_HSD_23	g3338.t1; g9313.t1	784; 785	PF13639; PF00078	Ring finger domain; Reverse transcriptase (RNA-dependent DNA polymerase)
UWO241_HSD_24	g4681.t1; g5342.t1; g12113.t1; g5638.t1	570; 575; 575; 567	PF00355, PF00078; PF00078; PF00078;	Rieske [2Fe-2S] domain, Reverse transcriptase (RNA-dependent DNA polymerase)
UWO241_HSD_25	g1206.t1; g13528.t1; g10711.t1	171; 167; 168	PF00101; PF00101; PF00101	Ribulose bisphosphate carboxylase, small chain; Ribulose bisphosphate carboxylas
UWO241_HSD_26	g14608.t1; g4489.t1	258; 258	PF01015; PF01015	Ribosomal S3Ae family; Ribosomal S3Ae family
UWO241_HSD_27	g417.t1; g8017.t1	190; 190	PF01775; PF01775	Ribosomal proteins 50S-L18Ae/60S-L20/60S-L18A; Ribosomal proteins 50S-L18Ae/
UWO241_HSD_28	g1892.t1; g15077.t1	147; 147	PF00828; PF00828	Ribosomal proteins 50S-L15, 50S-L18e, 60S-L27A; Ribosomal proteins 50S-L15, 50S
UWO241_HSD_29	g4873.t1; g563.t1	141; 141	PF00380; PF00380	Ribosomal protein S9/S16; Ribosomal protein S9/S16
UWO241_HSD_30	g14344.t1; g14343.t1	204; 204	PF01201; PF01201	Ribosomal protein S8e; Ribosomal protein S8e
UWO241_HSD_31	g5390.t1; g14543.t1; g4216.t1; g853.t1; g6242.t1	221; 231; 213; 215; 220	PF00177; ; ; ;	Ribosomal protein S7p/S5e; ; ; ;
UWO241_HSD_32	g3998.t1; g7778.t1	86; 86	PF01667; PF01667	Ribosomal protein S27; Ribosomal protein S27
UWO241_HSD_33	g6869.t1; g11743.t1	107; 108	PF01283; PF01283	Ribosomal protein S26e; Ribosomal protein S26e
UWO241_HSD_34	g3985.t1; g4643.t1	265; 265	PF01248; PF01248	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family; Ribosomal protein L7Ae/L30e
UWO241_HSD_35	g3676.t1; g2014.t1	160; 160	PF00542, PF16320; PF00542, PF16320	Ribosomal protein L7/L12 C-terminal domain, Ribosomal protein L7/L12 dimerisat
UWO241_HSD_36	g15171.t1; g11536.t1	99; 99	PF00935; PF00935	Ribosomal protein L44; Ribosomal protein L44
UWO241_HSD_37	g6373.t1; g4864.t1	117; 117	PF01198; PF01198	Ribosomal protein L31e; Ribosomal protein L31e
UWO241_HSD_38	g9340.t1; g16479.t1	155; 157	PF01246; PF01246	Ribosomal protein L24e; Ribosomal protein L24e
UWO241_HSD_39	g408.t1; g14240.t1	164; 164	PF01157; PF01157	Ribosomal protein L21e; Ribosomal protein L21e
UWO241_HSD_40	g8280.t1; g8836.t1	136; 134	PF01929; PF01929	Ribosomal protein L14; Ribosomal protein L14
UWO241_HSD_41	g413.t1; g8486.t1	208; 208	PF01294; PF01294	Ribosomal protein L13e; Ribosomal protein L13e
UWO241_HSD_42	g7384.t1; g2186.t1	166; 166	PF00298, PF03946; PF00298, PF03946	Ribosomal protein L11, RNA binding domain, Ribosomal protein L11, N-terminal do
UWO241_HSD_43	g320.t1; g555.t1	187; 187	PF17135; PF17135	Ribosomal protein 60S L18 and 50S L18e; Ribosomal protein 60S L18 and 50S L18e
UWO241_HSD_44	g94.t1; g499.t1	179; 179	PF00673, PF00281; PF00673, PF00281	ribosomal L5P family C-terminus, Ribosomal protein L5; ribosomal L5P family C-ter

Table S6: Highly similar duplicate genes (HSDs) in ICE-L and the similarity to those in UWO241. Download the complete table via the link: https://drive.google.com/file/d/1-ZKHQOTdiEJTzmXs3i0RywFcOPHoaExo/view?usp=sharing **(Chapter 3, 4)**

HSDs identifier	HSDs gene copie names (within 10 amir	Amino acid length of gene copies	Pfam identifier	Pfam Description
ICE-L_HSD_1	225; 10610	188; 188	PF17135; PF17135	Ribosomal protein 60S L18 and 50S L18e; Ribosoma
ICE-L_HSD_2	7092; 7102; 18025	257; 255; 255	PF16974; PF16974; PF16974	High-affinity nitrate transporter accessory; High-af
ICE-L_HSD_3	12236; 13158	161; 163	PF16320, PF00542; PF00542, PF16320	Ribosomal protein L7/L12 dimerisation domain, Rib
ICE-L_HSD_4	203; 13064; 17129; 18645; 15138; 9012;	130; 130; 130; 130; 130; 130; 130; 1	PF16211, PF00125; PF00125, PF16211; PF00	C-terminus of histone H2A, Core histone H2A/H2B/
ICE-L_HSD_5	2650; 18940; 18808; 18654; 18608; 1200	104; 104; 104; 104; 104; 104; 104; 1	PF15511; PF15511; PF15511; PF15511; PF15	Centromere kinetochore component CENP-T histor
ICE-L_HSD_6	12762; 16382	920; 916	PF14214;	Helitron helicase-like domain at N-terminus;
ICE-L_HSD_7	14777; 5152; 8264	304; 304; 302	PF14204, PF17144; PF14204, PF17144; PF14	Ribosomal L18 C-terminal region, Ribosomal large s
ICE-L_HSD_8	3150; 11209	170; 170	PF13499; PF13499	EF-hand domain pair; EF-hand domain pair
ICE-L_HSD_9	4146; 12251	487; 487	PF13499, PF00069, PF13833; PF00069, PF13	EF-hand domain pair, Protein kinase domain, EF-ha
ICE-L_HSD_10	1486; 11159	93; 83	PF12796;	Ankyrin repeats (3 copies);
ICE-L_HSD_11	1250; 1125	186; 186	PF10674; PF10674	Protein of unknown function (DUF2488); Protein of
ICE-L_HSD_12	1480; 10605	252; 254	PF10211; PF10211	Axonemal dynein light chain; Axonemal dynein ligh
ICE-L_HSD_13	2791; 18687	501; 501	PF08707; PF08707	Primase C terminal 2 (PriCT-2); Primase C terminal
ICE-L_HSD_14	996; 19543	822; 827	PF08707, PF08706; PF08707, PF08706	Primase C terminal 2 (PriCT-2), D5 N terminal like; I
ICE-L_HSD_15	8303; 18145	323; 323	PF08241; PF08241	Methyltransferase domain; Methyltransferase don
ICE-L_HSD_16	7110; 19341	199; 199	PF07714; PF07714	Protein tyrosine kinase; Protein tyrosine kinase
ICE-L_HSD_17	7103; 18033; 18038	523; 523; 523	PF07690; PF07690; PF07690	Major Facilitator Superfamily; Major Facilitator Sup
ICE-L_HSD_18	16819; 18415	232; 232	PF07650, PF00189; PF07650, PF00189	KH domain, Ribosomal protein S3, C-terminal doma
ICE-L_HSD_19	12973; 14822	267; 258	PF06026; PF06026	Ribose 5-phosphate isomerase A (phosphoriboisom
ICE-L_HSD_20	8175; 19485	621; 621	PF05787; PF05787	Bacterial protein of unknown function (DUF839); B
ICE-L_HSD_21	9066; 17145; 15221	305; 305; 305	PF05637; PF05637; PF05637	galactosyl transferase GMA12/MNN10 family; gala
ICE-L_HSD_22	16272; 16273	205; 205	PF05615; PF05615	Tho complex subunit 7; Tho complex subunit 7
ICE-L_HSD_23	191; 16327; 16002	191; 191; 192	PF05018; PF05018; PF05018	Protein of unknown function (DUF667); Protein of u
ICE-L_HSD_24	12588; 18898	164; 166	PF04970; PF04970	Lecithin retinol acyltransferase; Lecithin retinol acy
ICE-L_HSD_25	195; 16627	390; 386	PF04851; PF04851	Type III restriction enzyme, res subunit; Type III res
ICE-L_HSD_26	943; 13115	474; 474	PF04851; PF04851	Type III restriction enzyme, res subunit; Type III res
ICE-L_HSD_27	13528; 19704	480; 483	PF04851; PF04851	Type III restriction enzyme, res subunit; Type III res
ICE-L_HSD_28	6438; 19053	83; 83	PF04627; PF04627	Mitochondrial ATP synthase epsilon chain; Mitocho

Supplementary References

Arriola, M. B., N. Velmurugan, Y. Zhang, M. H. Plunkett, H. Hondzo and B. M. Barney (2018). Genome sequences of *Chlorella sorokiniana* UTEX 1602 and *Micractinium conductrix* SAG 241.80: implications to maltose excretion by a green alga. The Plant Journal 93: 566-586.

Ayala-del-Río, H. L., P. S. Chain, J. J. Grzymski, M. A. Ponder, N. Ivanova, P. W. Bergholz, G. Di Bartolo, L. Hauser, M. Land and C. Bakermans (2010). The genome sequence of Psychrobacter arcticus 273-4, a psychroactive Siberian permafrost bacterium, reveals mechanisms for adaptation to low-temperature growth. Applied and environmental microbiology 76: 2304-2312.

Beier, S., A. R. Rivers, M. A. Moran and I. Obernosterer (2015). The transcriptional response of prokaryotes to phytoplankton-derived dissolved organic matter in seawater. Environmental Microbiology 17: 3466-3480.

Benites, L. F., N. Poulton, K. Labadie, M. E. Sieracki, N. Grimsley and G. Piganeau (2019). Single cell ecogenomics reveals mating types of individual cells and ssDNA viral infections in the smallest photosynthetic eukaryotes. Philosophical Transactions of the Royal Society B 374: 20190089.

Blaby-Haas, C. E. and S. S. Merchant (2019). Comparative and functional algal genomics. Annual Review of Plant Biology 70: 605-638.

Blanc, G., I. Agarkova, J. Grimwood, A. Kuo, A. Brueggeman, D. D. Dunigan, J. Gurnon, I. Ladunga, E. Lindquist and S. Lucas (2012). The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. Genome Biology 13: 1-12.

Blanc, G., G. Duncan, I. Agarkova, M. Borodovsky, J. Gurnon, A. Kuo, E. Lindquist, S. Lucas, J. Pangilinan and J. Polle (2010). The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. The Plant Cell 22: 2943-2955.

Blanc-Mathieu, R., B. Verhelst, E. Derelle, S. Rombauts, F.-Y. Bouget, I. Carré, A. Château, A. Eyre-Walker, N. Grimsley and H. Moreau (2014). An improved genome of the model marine alga *Ostreococcus tauri* unfolds by assessing Illumina de novo assemblies. BMC Genomics 15: 1-12.

Bogen, C., A. Al-Dilaimi, A. Albersmeier, J. Wichmann, M. Grundmann, O. Rupp, K. J. Lauersen, O. Blifernez-Klassen, J. Kalinowski and A. Goesmann (2013). Reconstruction of the lipid metabolism for the microalga *Monoraphidium neglectum* from its genome sequence reveals characteristics suitable for biofuel production. BMC Genomics 14: 1-18.

Bowman, J., S. Ferriera, J. Johnson, S. Kravitz, A. Halpern, K. Remington, K. Beeson, B. Tran, Y. Rogers and R. Friedman (2006). Genome sequence analysis reveals unique cold

adaptation features of the extreme psychrophile Psychroflexus torquis ATCC 700755. International Conference on Alpine and Polar Microbiology.

Byrne-Steele, M. L., R. C. Hughes and J. D. Ng (2009). Recombinant production, crystallization and preliminary X-ray analysis of PCNA from the psychrophilic archaeon Methanococcoides burtonii DSM 6242. Acta Crystallographica Section F: Structural Biology and Crystallization Communications 65: 1131-1135.

Clerissi, C., Y. Desdevises and N. Grimsley (2012). Prasinoviruses of the marine green alga *Ostreococcus tauri* are mainly species specific. Journal of Virology 86: 4611-4619.

Deschaght, P., M. Janssens, M. Vaneechoutte and G. Wauters (2012). Psychrobacter isolates of human origin, other than Psychrobacter phenylpyruvicus, are predominantly Psychrobacter faecalis and Psychrobacter pulmonis, with emended description of P. faecalis. International journal of systematic and evolutionary microbiology 62: 671-674.

Duchaud, E., M. Boussaha, V. Loux, J.-F. Bernardet, C. Michel, B. Kerouault, S. Mondot, P. Nicolas, R. Bossy and C. Caron (2007). Complete genome sequence of the fish pathogen Flavobacterium psychrophilum. Nature Biotechnology 25: 763-769.

Featherston, J., Y. Arakaki, H. Nozaki, P. M. Durand and D. R. Smith (2016). Inflated organelle genomes and a circular-mapping mtDNA probably existed at the origin of coloniality in volvocine green algae. European Journal of Phycology 51: 369-377.

Foflonker, F., D. C. Price, H. Qiu, B. Palenik, S. Wang and D. Bhattacharya (2015). Genome of the halotolerant green alga *Picochlorum* sp. reveals strategies for thriving under fluctuating environmental conditions. Environmental Microbiology 17: 412-426.

Franzmann, P. D., Y. Liu, D. L. Balkwill, H. C. Aldrich, E. C. De Macario and D. R. Boone (1997). Methanogenium frigidum sp. nov., a psychrophilic, H2-using methanogen from Ace Lake, Antarctica. International Journal of Systematic and Evolutionary Microbiology 47: 1068-1072.

Gao, C., Y. Wang, Y. Shen, D. Yan, X. He, J. Dai and Q. Wu (2014). Oil accumulation mechanisms of the oleaginous microalga *Chlorella protothecoides* revealed through its genome, transcriptomes, and proteomes. BMC Genomics 15: 1-14.

Hallam, S. J., K. T. Konstantinidis, N. Putnam, C. Schleper, Y.-i. Watanabe, J. Sugahara, C. Preston, J. de la Torre, P. M. Richardson and E. F. DeLong (2006). Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. Proceedings of the National Academy of Sciences 103: 18296-18301.

Hanschen, E. R., T. N. Marriage, P. J. Ferris, T. Hamaji, A. Toyoda, A. Fujiyama, R. Neme, H. Noguchi, Y. Minakuchi and M. Suzuki (2016). The *Gonium pectorale* genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. Nature Communications 7: 1-10.

Hirooka, S., Y. Hirose, Y. Kanesaki, S. Higuchi, T. Fujiwara, R. Onuma, A. Era, R. Ohbayashi, A. Uzuka and H. Nozaki (2017). Acidophilic green algal genome provides insights into adaptation to an acidic environment. Proceedings of the National Academy of Sciences 114: 8304-8313.

Huesemann, M., T. Dale, A. Chavis, B. Crowe, S. Twary, A. Barry, D. Valentine, R. Yoshida, M. Wigmosta and V. Cullinan (2017). Simulation of outdoor pond cultures using indoor LED-lighted and temperature-controlled raceway ponds and Phenometrics photobioreactors. Algal Research 21: 178-190.

Jiang, B., D. Cui, A. Li, Z. Gai, F. Ma, J. Yang and N. Ren (2012). Genome sequence of a cold-adaptable sulfamethoxazole-degrading bacterium, Pseudomonas psychrophila HA-4, Am Soc Microbiol.

Kim, S. J., S. C. Shin, S. G. Hong, Y. M. Lee, I.-G. Choi and H. Park (2012). Genome sequence of a novel member of the genus Psychrobacter isolated from Antarctic soil, Am Soc Microbiol.

Lee, Y. M., C. Y. Hwang, I. Lee, Y.-J. Jung, Y. Cho, K. Baek, S. G. Hong, J.-H. Kim, J. Chun and H. K. Lee (2014). Lacinutrix jangbogonensis sp. nov., a psychrophilic bacterium isolated from Antarctic marine sediment and emended description of the genus Lacinutrix. Antonie Van Leeuwenhoek 106: 527-533.

Médigue, C., E. Krin, G. Pascal, V. Barbe, A. Bernsel, P. N. Bertin, F. Cheung, S. Cruveiller, S. D'Amico and A. Duilio (2005). Coping with cold: the genome of the versatile marine Antarctica bacterium Pseudoalteromonas haloplanktis TAC125. Genome Research 15: 1325-1335.

Merchant, S. S., S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz, G. B. Witman, A. Terry, A. Salamov, L. K. Fritz-Laylin and L. Maréchal-Drouard (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318: 245-250.

Mock, T., R. P. Otillar, J. Strauss, M. McMullan, P. Paajanen, J. Schmutz, A. Salamov, R. Sanges, A. Toseland and B. J. Ward (2017). Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. Nature 541: 536-540.

Moreau, H., B. Verhelst, A. Couloux, E. Derelle, S. Rombauts, N. Grimsley, M. Van Bel, J. Poulain, M. Katinka and M. F. Hohmann-Marriott (2012). Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. Genome Biology 13: R74.

Nelson, D. R., A. Chaiboonchoe, W. Fu, K. M. Hazzouri, Z. Huang, A. Jaiswal, S. Daakour, A. Mystikou, M. Arnoux and M. Sultana (2019). Potential for heightened sulfur-metabolic capacity in coastal subtropical microalgae. Iscience 11: 450-465.

Palenik, B., J. Grimwood, A. Aerts, P. Rouzé, A. Salamov, N. Putnam, C. Dupont, R. Jorgensen, E. Derelle and S. Rombauts (2007). The tiny eukaryote *Ostreococcus* provides

genomic insights into the paradox of plankton speciation. Proceedings of the National Academy of Sciences 104: 7705-7710.

Pocock, T., M. A. Lachance, T. Pröschold, J. C. Priscu, S. S. Kim and N. P. Huner (2004). Identification of a psychrophilic green alga from Lake Bonney Antarctica: *Chlamydomonas raudensis* ETTL. (UWO241) Chlorophyceae. Journal of Phycology 40: 1138-1148.

Polle, J. E., K. Barry, J. Cushman, J. Schmutz, D. Tran, L. T. Hathwaik, W. C. Yim, J. Jenkins, Z. McKie-Krisberg and S. Prochnik (2017). Draft nuclear genome sequence of the halophilic and beta-carotene-accumulating green alga *Dunaliella salina* strain CCAP19/18. Genome Announcements 5: 01105-01117.

Pombert, J.-F., N. A. Blouin, C. Lane, D. Boucias and P. J. Keeling (2014). A lack of parasitic reduction in the obligate parasitic green alga *Helicosporidium*. PLoS Genetics 10: e1004355.

Prochnik, S. E., J. Umen, A. M. Nedelcu, A. Hallmann, S. M. Miller, I. Nishii, P. Ferris, A. Kuo, T. Mitros and L. K. Fritz-Laylin (2010). Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science 329: 223-226.

Remias, D., U. Lütz-Meindl and C. Lütz (2005). Photosynthesis, pigments and ultrastructure of the alpine snow alga Chlamydomonas nivalis. European Journal of Phycology 40: 259-268.

Riley, M., J. T. Staley, A. Danchin, T. Z. Wang, T. S. Brettin, L. J. Hauser, M. L. Land and L. S. Thompson (2008). Genomics of an extreme psychrophile, Psychromonas ingrahamii. BMC Genomics 9: 1-19.

Roth, M. S., S. J. Cokus, S. D. Gallaher, A. Walter, D. Lopez, E. Erickson, B. Endelman, D. Westcott, C. A. Larabell and S. S. Merchant (2017). Chromosome-level genome assembly and transcriptome of the green alga *Chromochloris zofingiensis* illuminates astaxanthin production. Proceedings of the National Academy of Sciences 114: E4296-E4305.

Schmidt, M., A. Prieme and P. Stougaard (2006). Rhodonellum psychrophilum gen. nov., sp. nov., a novel psychrophilic and alkaliphilic bacterium of the phylum Bacteroidetes isolated from Greenland. International Journal of Systematic and Evolutionary Microbiology 56: 2887-2892.

Suzuki, S., H. Yamaguchi, N. Nakajima and M. Kawachi (2018). *Raphidocelis subcapitata* (= *Pseudokirchneriella subcapitata*) provides an insight into genome evolution and environmental adaptations in the Sphaeropleales. Scientific Reports 8: 1-13.

Worden, A. Z., J.-H. Lee, T. Mock, P. Rouzé, M. P. Simmons, A. L. Aerts, A. E. Allen, M. L. Cuvelier, E. Derelle and M. V. Everett (2009). Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. Science 324: 268-272.

Wu, A. K., A. M. Kropinski, J. S. Lumsden, B. Dixon and J. I. MacInnes (2015). Complete genome sequence of the fish pathogen Flavobacterium psychrophilum ATCC 49418 T. Standards in Genomic Sciences 10: 1-19.

Yoon, J.-H., S.-J. Kang and T.-K. Oh (2006). Polaribacter dokdonensis sp. nov., isolated from seawater. International Journal of Systematic and Evolutionary Microbiology 56: 1251-1255.

Zhang, Z., C. Qu, K. Zhang, Y. He, X. Zhao, L. Yang, Z. Zheng, X. Ma, X. Wang and W. Wang (2020). Adaptation to extreme Antarctic environments revealed by the genome of a sea ice green alga. Current Biology 30: 1-12.

Zhao, T. (2011). Genome sequencing and analysis of the psychrophilic anoxygenic phototrophic bacterium Rhodoferax antarcticus sp. ANT. BR, Arizona State University.

Appendix B: Permission for reproduction of scientific articles

Copyright Agreement for the usage of iScience article in Chapter 3 and 4

Copyright Agreement for the usage of image of Figure 1 (Chapter 1)

Re: Credit request for the image usage from EGUblogs

Hilary Dugan
Wed 09/09/20 10:49
To: Xi Zhang

Hi Xi,
You are more than welcome to use the photos. Some of them are also available here: https://dugan.limnology.wisc.edu/antarctica/

Best,
Hilary

On Wed, Sep 9, 2020 at 9:36 AM Xi Zhang

> Hi Prof. Dugan,
>
> I am writing to request your credit for using the photos you used in EGUblogs (https://blogs.egu.eu/network/geosphere/2014/01/13/). Prof. Herod redirect me to ask for your permission to use these photos. Will thank you so much for allowing me to use your four images (attached), these pictures are awesome. I would like to use them in my PhD thesis and personal website background. I will highlight the image source and your credit like this: "The images are reworked with the credit from original author Hilary Dugan and the images source can be found via EGUblogs (https://blogs.egu.eu/network/geosphere/2014/01/13/). For any reuse or distribution, the work is licensed under the Creative Commons Attribution 4.0 International licence (CC BY 4.0)."
>
> Thanks for your time in advance!
> ~Xi
>
> Xi Zhang, MSc
> PhD Candidate
> University of Western Ontario
> David Smith Lab
> Office: BGS 3025

--
Hilary Dugan
Assistant Professor, Center for Limnology
University of Wisconsin - Madison
dugan.limnology.wisc.edu

Copyright Agreement for the image usage of Figure 2 (Chapter 1)

Dear Mr. Xi Zhang,

John Wiley & Sons - Books has approved your recent request. Before you can use this content, you must accept the license fee and terms set by the publisher.

Use this link to accept (or decline) the publisher's fee and terms for this order.

Request Summary:
Submit date: 23-Oct-2020
Request ID: 600026771
Publication: Journal of phycology
Title: IDENTIFICATION OF A PSYCHROPHILIC GREEN ALGA FROM LAKE BONNEY ANTARCTICA: CHLAMYDOMONAS RAUDENSIS ETTL. (UWO 241) CHLOROPHYCEAE1
Type of Use: Republish in other published product

Please do not reply to this message.

Sincerely,

Copyright Clearance Center

Journal of phycology

Article IDENTIFICATION OF A PSYCHROPHILIC GREEN ALGA FROM LAKE BONNEY ANTARCTICA: CHLAMYDOMONAS RAUDENSIS ETTL. (UWO 241) CHLOROPHYCEAE1

GENERAL INFORMATION

Request ID	600026771	**Request Date**	22 Oct 2020
Request Status	Accepted	**Price**	0.00 CAD

ALL DETAILS

ISSN:	1529-8817	**Publisher:**	BLACKWELL PUBLISHING
Type of Use:	Republish in other published product	**Portion:**	Chart/graph/table/figure

LICENSED CONTENT

Publication Title	Journal of phycology	**Rightsholder**	John Wiley & Sons - Books
Article Title	IDENTIFICATION OF A PSYCHROPHILIC ...	**Publication Type**	e-Journal
Author/Editor	Phycological Society of America.	**URL**	http://firstsearch.oclc.org/journal=002...
Date	01/01/1965	**Start Page**	1148
Language	English	**Issue**	6
Country	United Kingdom of Great Britain and N...	**Volume**	40

REQUEST DETAILS

Portion Type	Chart/graph/table/figure	**Distribution**	Worldwide
Number of charts / graphs / tables / figures requested	2	**Translation**	Original language of publication
		Copies for the disabled?	No
Format (select all that apply)	Electronic	**Minor editing privileges?**	No
Who will republish the content?	Academic institution	**Incidental promotional use?**	No
Duration of Use	Life of current edition	**Currency**	CAD
Lifetime Unit Quantity	Up to 999		
Rights Requested	Main product		

NEW WORK DETAILS

Title	Sequencing and assembling the nuclea...	**Produced by**	Western University
Author	Xi Zhang	**Expected publication date**	2021-06-01

ADDITIONAL DETAILS

The requesting person / organization to appear on the license	Tessa Pocock,Norman P. A. Huner

REUSE CONTENT DETAILS

Title, description or numeric reference of the portion(s)	Light microscope images of Chlamydo...	**Title of the article/chapter the portion is from**	IDENTIFICATION OF A PSYCHROPHILIC ...
Editor of portion(s)	Huner, Norman P. A.; Kim, Sam Sulgi; P...	**Author of portion(s)**	Huner, Norman P. A.; Kim, Sam Sulgi; P...
Volume of serial or monograph	40	**Issue, if republishing an article from a serial**	6
Page or page range of portion	1148-1148		
		Publication date of portion	1965-01-01

Appendix C: HSDFinder: an integrated tool for predicting highly similar duplicates in eukaryotic genomes

Xi Zhang, Yining Hu, David Roy Smith

Abstract

Background: Gene duplication as a strategy to adapt to various environmental conditions has been documented in a wide range of species. Zhang *et al.*, for example, argued that hundreds of highly similar duplicate genes (HSDs) are aiding the survival of an Antarctic green alga via gene dosage. However, the numbers of HSDs in other eukaryotic genomes are largely unknown, and computational methods for identifying them can be time-consuming and labor-intensive.

Results: Here, we present an automated online tool (HSDFinder) for identifying HSDs in eukaryotic genomes with high accuracy and reliability annotated with Pfam domains and KEGG pathways. HSDFinder can analyze unannotated genome sequences by integrating data from InterProScan and KEGG databases. The resulting HSDs are displayed in an 8-column spreadsheet. To compare HSDs among different species, we developed an online heatmap plotting option to visualize the results in different KEGG pathway functional categories. The software presented here is the primary selection of HSDs, the manually curation can be done to filter the partial or add the novel HSDs when necessary.

Conclusions: HSDFinder aims to become a useful platform for identification and comprehensive analysis of HSDs in the eukaryotic genomes, which can deepen the insights into how gene duplications can impact adaptation. The web server is freely available at http://hsdfinder.com. The distribution version can be found via the GitHub: https://github.com/zx0223winner/HSDFinder.

Keywords

Next-generation sequencing, green algae, highly similar duplicates, gene copies, KEGG, InterProScan, Pfam

Background

Gene duplication is ubiquitous phenomenon throughout the eukaryotic tree of life [1]. Usually, retaining highly similar expressed sequences is disadvantageous; therefore, it should be rare to have duplicates encoding the same functions maintained in the genome [2]. However, it revealed that the generation of large-scale duplicates was possible if they were highly demanded genes, such as rRNAs and histones [1]. Thereafter, Libuda and Winston [3] discovered that the appearance of pairs of adjacent paralogous proteins arose from a compensatory mechanisms restoring normal dosage when one locus was deleted. There is a controversy in whether the evolution of duplicate genes affects fitness [4]. Some duplication models assume that the fixation of the duplicate copy is a neutral process, while others support the gene dosage hypothesis, where if an increase in the dosage of a particular gene is beneficial, then a duplication of this gene may be fixed by positive selection [5]. Nevertheless, mechanisms that do not require the evolution of new functions (e.g., dosage balance) may play an important role in the initial retention of duplicate genes [6]. Indeed, many examples have been accumulated in the literature suggesting that stress response genes, sensory genes, transport genes and genes that have a metabolism-related function are likely to be fixed as duplicated copies under certain environmental conditions [7]. The large-scale gene amplifications were found in the acidophile *Chlamydomonas eustigma* with ~10 copies of genes encoding arsenate reductase (ArsC) and 20 copies of genes encoding glutaredoxin (Grx), suggesting the adaptations to acidophilic environments [8]. What's more, many gene copies encoding carotene biosynthesis-related protein (CBR) and high intensity light-inducible lhc-like gene (Lhl4) were found in *Chlamydomonas* sp. ICE-L, suggesting the adaptation to the highly variable light conditions in Antarctic sea ice [9]. Just recently, for example, it was suggested that hundreds of highly similar duplicates (HSDs) are aiding the survival of the Antarctic green alga *Chlamydomonas* sp. UWO241 via gene dosage [10]. Although the dosage hypothesis can be further tested by experiments, it is time-consuming and labor-intensive to carry on large-scale comparative analysis.

It is important to clarify the origin of duplicate before setting the threshold to identify them [4]. There are five main broad classes of duplication events in genomes: whole-genome duplication (WGD), tandem duplication, transposon-mediated duplication, segmental

duplication and retroduplication [6]. Polyploidization or WGD, is a straightforward gene duplication mechanism that increases both genome size and the entire gene sets. However, it is not the only mechanism that generates duplicate genes. A cluster of two to many paralogous sequences with no or few intervening gene sequences is a pattern of tandem (or local) duplication that results from unequal crossing-over of chromosomes or TE-mediated duplication. Furthermore, transposon-mediated duplication usually contains the hallmarks of two terminal inverted repeats (TIRs) less than 5 kb long. Segmental duplication usually arises from non-LTR retrotransposons, such as long interspersed nuclear elements (LINEs) (intact LINE1s are up to 6 kb in length and contain internal promoters). Retroduplication refers to retrogenes generated via 5~9 kb LTR-retrotransposons, such as *gypsy* LTR elements [6]. Notably, if a gene is duplicated via reverse transcription of mRNA and then inserts into the genome, it is referred to as retrocopy, and the original gene is referred to as the parental gene. Although a retrocopy can arise from both long LTR and non-LTR retrotransposable elements (e.g., LINE1), the expression of the retrocopy is largely dependent on the regulatory region (i.e., promoters, binding sites for the RNA polymerase, and/or enhancers) [2]. Both gene and partial duplication appear to be an ongoing phenomenon within the eukaryotic genome, one which might be mediated by retrotransposons [10]. For example, to call a retrogene, the aligned sequence must be at least 150 bp long and 50% amino acid identity to parental genes [11].

Many tools and software have been developed for identifying duplications in genomes, some are targeting for specific duplication event and some can handle with the genomes under multiple duplication and rearrangement events [12]. For example, tools such as MCScanX-transposed [13], i-ADHoRe [14] and CYNTENATOR [15] are developed to search for syntenic blocks (mainly for detecting WGD and segmental duplications), which can be defined as two regions of a genome including several homologous genes co-arranged one another [16]. Since orthologs (derived by speciation) and paralogs (derived by duplication) are two types of homologs, which are genes sharing the common ancestry. As for those the tools detecting the duplicated genes via the paralogous relationships, the sequence similarity and gene structure are usually first considered [12]. Alignment tools such as BLAST [17], DIAMOND [18], and nhmmer [19] are commonly chosen to measure the sequence similarity via the metrics such as percentage identity, aligned length

difference and E-value. Notably, due to amino acid substitutions occur less frequently than nucleotide substitutions, the sequence alignments are generally compared among amino acid sequences instead of nucleotides, which allows a greater sensitivity [20]. It is difficult to set the right cut-off for those metrics when detecting the duplicates in a large scale, although lowering the threshold of the metrics might risk of increasing of false positives [12].

To help the scientific community flexibly identify and characterize duplicates in eukaryotic genomes, we developed an automated web-based tool called HSDFinder. HSDFinder not only categorizes gene copies together via given thresholds but also annotates the duplicates via protein functional domains and pathway information from the InterProScan and KEGG databases. The results are displayed in an 8-column spreadsheet, which allows for alternative visualization forms, including trendlines and heatmaps. The results are documented in HSDatabase [21] , which allows users to perform large-scale comparative analysis. Although HSDFinder is designed to identify highly similar duplicates, users have the option to employ different parameters (e.g., from 30% to 100% identity and from within 0-100 aa variances). Using HSDFinder, we identified approximately 336 and 265 HSDs in the green algae *Chlamydomonas* sp. UWO241 and *Chlamydomonas* sp. ICE-L [22] , respectively, and employed the software on other chlorophyte algae and model eukaryotic genomes. The predicted results are documented in HSDatabase [21] , which currently contains 28,214 HSDs from fifteen eukaryotes (http://hsdfinder.com/database/) (Table 1).

Implementation

The web server of HSDFinder is implemented on Apache server and the web interface is designed using HTML and Python scripts. The algorithms used to predict HSDs and visualize the correlations using heatmaps are written in Python. There are three steps to implement the software.

Preparing the input files

Before running HSDFinder, two spreadsheets in tab-separated values (tsv.) format need to be prepared as input files (Figure 1A). Note: Example files are provided for guidance as well as frequently asked questions (FAQ) section. A protein BLAST search of the genome

models against themselves (E-value cut-off < 10-5, BLASTP -outfmt 6) will yield the first input file. The BLAST results should be arranged in 12-column spreadsheets, including the key information from the query name to percentage identity, etc. The second spreadsheet is acquired from InterProScan, which can provide the protein signatures, such as Pfam domain. The output file of InterProsScan is tab-separated values (tsv.) format in default.

Running the HSDFinder

The two spreadsheets can be submitted to HSDFinder with some personalized options. The default setting of HSDFinder will identify HSDs with near-identical protein lengths (within 10 amino acids of each other) and ≥ 90% pairwise amino acid identities. Choosing such a strict cut-off will undoubtedly remove many genuine duplicates from the list. Thus, users have the option to employ different parameters for identifying HSDs (e.g., from 50% to 100% pairwise amino acid identity and from within 0-100 amino acid length variances). The output of this step will be an 8-column spreadsheet containing the information of HSD identifier, gene copy number, and Pfam domain. Additionally, the user can conveniently set different values to create a trendline graph of the gene copy numbers under different criteria (Figure 1B).

Visualizing the HSDs across species

For comparative analyses of the HSDs across different species, we developed an online heatmap plotting option to visualize the HSDs results in different KEGG pathway categories. To do so, the user will need to generate HSDs results following the previous steps for the species of interest. The default for plotting the heatmap is at least two species and at least two files are needed to plot the heatmap. Examples are given to guide the appropriate input files (Figure 1C). The first input file is the outputs of your interest species after running HSDFinder; the second file is retrieved from the KEGG database documenting the correlation of KEGG Orthology (KO) accession with each gene model identifier. Since species usually have unique gene model identifiers, we recommend submitting the second KEGG pathway files corresponding to each species. Once the input files have been submitted, the HSDs numbers for each species will be displayed in a heatmap under different KEGG functional categories. On the left side, the color bar indicates a broad category of HSDs who have pathway function matches, such as

carbohydrate metabolism, energy metabolism, and translation. The color for the matrix reflects the number of HSDs across species.

Results and Discussion

The predicted HSDs will be manually curated before submitting to the HSDatabase (Figure 1A). The strict cut-off is to ensure that the gene-pairs in question are functional duplicates rather than spurious ones. Also, the future comparison of substitution rates at synonymous and nonsynonymous sites of protein-coding genes (i.e., calculating dN/dS) analysis in pairwise mode will require appropriate protein alignments in each HSDs. Nevertheless, users always have an option to loosen the cut-off of aligned protein length and percentage identity. But that will increase the chances of generating false positive HSDs. Therefore, users have to find a balance somewhere in-between these criteria.

As displayed in Figure 2A, the HSDFinder results are summarized in an 8-column spreadsheet. The first column is the unique UWO241 gene identifier, which is used to track the HSDs. The second and third column includes different numbers and lengths of gene copies in each HSDs. The Pfam domain identifier as well as the InterPro (IPR) identifier provide more details about the function of each HSDs. Notably, we prefer using the Pfam domain as the functional description. Although the function description of the interested genes can be scanned in NCBI-NR or UniProtKB/Swiss-Prot databases, many hypothetical proteins or 'bad name' proteins may also be included in these databases, which could confuse the interpretation of HSDs results. To address that, we have developed another software NoBadWordsCombiner (http://hsdfinder.com/combiner/) that can integrate the gene function information together without 'bad name' including Nr-NCBI, UniProtKB/Swiss-Prot, KEGG, Pfam and GO etc [23].

Trendline figures can be used to interpret the number of total HSD copies based on different cut-off values (Figure 2B). We provide an example based on the genome analysis of the Antarctic green alga UWO241. The gene sets of the genome are widely explored via employing different parameters for identifying HSDs (e.g., from 50% to 90% pairwise amino acid identity and from within 0-100 amino acid length variances) (Figure 2C). As displayed in Figure 3, comparative analysis of the HSDs across different species can be

carried out using an online heatmap tool to visualize the HSD results in different KEGG pathway categories. One of the 6-column output files has been displayed as an example to indicate the HSDs under the KEGG function categories with matching KO number and descriptions (Figure 3A). The first and second columns are the pathway categories, and the remaining columns describe the correlations of HSDs with unique KO identifiers. The heatmap example based on four species has been presented here (Figure 3C). To create an appropriate heatmap, at least two species are needed. For example, HSDs in the green algal genomes (UWO241, ICE-L and *C. eustigma*) are involved in a diversity of cellular pathways, including gene expression, cell growth, membrane transport, and energy metabolism (Figure 3B and Table 2). The HSDFinder results are categorized into HSDatabase after manual curation. The HSDatabase will be updated timely and the latest version is HSDatabase v1.5, in which a total of 28,214 HSDs in 15 eukaryotic genomes are identified (Table 1). It is our hope to build a comparative analysis framework across species, especially for those extremophiles, to understand the role of gene duplication in different survival environments.

Conclusions

With the decreasing cost of biological analyses (e.g., next-generation sequencing), biologists are dealing with larger and greater amounts of data, and many software analysis suites require considerable knowledge of computer scripting and microprogramming. HSDFinder is designed to fill the demand for custom-made scripts to move from one analysis step to another. HSDFinder is able to efficiently analyze duplicated genes from unannotated genome sequences by integrating the results from InterProScan and KEGG. The result of the predicted HSDs can be visualized in a high resolution heatmap. HSDFinder aims to become a useful platform for the identification and comprehensive analysis of HSDs in the eukaryotic genomes, which deepen our insights into the gene duplication mechanisms driving genome adaptation. In the future, the software will be further improved with the continuous updating by taking into account more scientific discoveries in the field of gene duplication.

Availability and requirements

Project name: HSDFinder

Project home page: http://www.hsdfinder.com

Operating system(s): Platform independent

Programming language: Python

Other requirements: Python 3

License: GNU GPL V3

Any restrictions to use by non-academics: No

Abbreviations

FAQ: frequently asked questions

HSDs: highly similar duplicate genes

ICE-L: *Chlamydomonas* sp. ICE-L

KEGG: Kyoto encyclopedia of genes and genomes

KO: KEGG Orthology

UWO241: *Chlamydomonas* sp. UWO241

Declarations

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Availability of data and materials

The datasets of eukaryotes supporting the conclusions of this article are available from JGI (https://phytozome.jgi.doe.gov/pz/portal.html) or NCBI (https://www.ncbi.nlm.nih.gov) database. The HSDFinder source code has been deposited at https://github.com/zx0223winner/HSDFinder. The web server of HSDFinder is freely available at http://hsdfinder.com. The predicted HSDs of fifteen eukaryotes are documented in HSDatabase, which can be accessed via http://hsdfinder.com/database/.

Competing interests

The authors declare that they have no competing interests.

Funding

XZ and DRS are supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Authors' contributions

The study was conceptualized by XZ and DRS. The data were analyzed by XZ. YNH implemented the HSDFinder website. XZ and DRS drafted the manuscript and all authors commented to produce the manuscript for peer review.

Acknowledgements

Not Applicable.

Figures and tables

Figure 1: The workflow and the file examples of HSDFinder. (A) Two spreadsheets in tab delimited are displayed as examples for the input files of HSDFinder. One is acquired from the BLAST in tabular format (-outfmt 6) and another is the running result in default mode via Interproscan. (B) The output of HSDFinder is an 8-column spreadsheet including information on gene copies to Pfam domain descriptions. Users have a choice to set different cut-off values to acquire potential duplicates. A trendline figures has been used as an example to interpret the number of total gene copies based on different cut-off thresholds. (C) The output file from step B together with a KEGG KO mapper file will be used as the input files to visualize the HSDs distribution across species. To create an appropriate heatmap, at least two species are needed. One of the 6-column output files have been displayed as an example to indicate the HSDs under the KEGG function categories with matching KO number and description. The heatmap example based on four species have been presented here. There is an option for users to download the high resolution heatmap figure and spreadsheet for future analysis.

Figure 2: The interpretation of predicted results from HSDFinder. (A) Screenshot of 8-column spreadsheet example presenting the unique HSDs in each row; the first column is the unique UWO241 gene identifier, which is used to track the HSDs. The second and third column incudes different number and length of gene copies in each HSDs. The Pfam

domain identifier as well as IPR identifier provide more details about the function of each HSDs. (B). A trendline figures has been used as an example to interpret the number of total gene copies with responding to different cut-off of HSDFinder in duplicates rich genome UWO241. (C). The table of gene copy numbers in UWO2421 filtered via different criteria of amino acid length and identity.

Figure 3: The visualization of HSDs results in a heatmap. (A) Example of the 6-column output files highlighting HSDs and the KEGG functional categories with matching KO number and description. The first and second column are the pathway categories, and the remaining columns describe the correlations of HSDs with unique KO identifiers. (B) The heatmap example is based on four species (*Chlamydomonas* sp. UWO241, *Chlamydomonas reinhardtii*, *Chlamydomonas eustigma* and *Chlamydomonas* sp. ICE-L). To create an appropriate heatmap, at least two species are needed.

Table 1: The predicted HSDs in selected eukaryotic genomes.

Domain	Kingdom	Phylum	Class	Order	Species	Accession number*	Ref	HSDs #	Gene copies #
Eucarya	Plantae	Chlorophyta	Chlorophyceae	Chlamydomonadales	*Chlamydomonas* sp. UWO241	GenBank (PRJNA547753)	[24]	336	1339
					Chlamydomonas reinhardtii	JGI 5.5 (Phytozome 12.1)	[25]	54	162
					Volvox carteri	JGI 2.0 (Phtyzome 12.1)	[26]	124	367
					Chlamydomonas eustigma	GCA_002335675.1	[8]	276	560
					Dunaliella salina	JGI 3.0 (Phytozome 12.1)	[27]	72	229
					Gonium pectorale	GCA_001584585.1	[28]	114	325
					Chlamydomonas sp. ICE-L	GCA_013435795.1	[9]	265	717
			Trebouxiophyceae	Trebouxiophyceae incertae sedis	*Coccomyxa subellipsoidea* C-169	GCA_000258705.1	[29]	79	272
		Streptophyta	Brassicaceae	Brassicales	*Arabidopsis thaliana*	GCA_000001735.2	[30]	628	1500
			Poaceae	Poales	*Zea mays* (Maize)	GCA_902167145.1	[31]	2570	6297
	Chromista	Ochrophyta	Bacillariophyceae	Bacillariales	*Fragilariopsis cylindrus* (Diatom)	GCA_001750085.1	[32]	124	317
	Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	*Saccharomyces cerevisiae* (yeast)	GCA_003086655.1	[33]	136	376
	Animalia	Arthropoda	Insecta	Diptera	*Drosophila melanogaster* (Fruit fly)	GCA_000001215.4	[34]	6894	18482
		Tardigrada	Eutardigrada	Parachela	*Hypsibius dujardini* (waterbear)	GCA_002082055.1	[35]	515	1081
		Chordata	Mammalia	Primates	*Homo sapiens* (Human)	GCA_000001405.28	[36]	NA	NA
				Rodentia	*Mus musculus* (Mouse)	GCA_000001635.9	[37]	15993	56802

*Accession numbers are from the US National Center for Biotechnology Information (NCBI) GenBank assembly accession numbers or the US Department of Energy's Joint Genome Institute Phytozome assembly version numbers.

Table 2. Summary statistics of highly similar duplicate genes (HSDs) in selected eukaryotes (UWO241, ICE-L and *C. eustigma*).

Database	Example Identifiers[a]	Number of HSDs (%) / Number of gene copies (%)[b]		
		UWO241	ICE-L	*C. eustigma*
Pfam				
Chlorophyll A-B binding protein	PF00504	4 (1%) / 25 (2%)	5 (2%) / 18 (3%)	3 (1%) / 6 (1%)
Ribosomal protein	PF01015; PF01775; PF00828	19 (5%) / 42 (3%)	41 (15%) / 91(13%)	8 (3%) / 16 (3%)
Core histone H2A/H2B/H3/H4	PF00125	5 (1%) / 99 (7%)	8 (3%) / 93 (13%)	4 (1%) / 13 (2%)
Ice-binding protein (DUF3494)	PF11999	8 (2%) / 21(2%)	NA	NA
Reverse transcriptases	PF00078	38 (11%) / 151(11%)	NA	2 (0.5%) / 3 (0.5%)
KEGG				
09101 Carbohydrate metabolism	K13979 (alcohol dehydrogenase)	12 (4%) / 89 (7%)	9 (3%) / 23(3%)	8 (3%) / 16 (3%)
09102 Energy metabolism	K02639 (ferredoxin); K08913(light-harvesting complex II chlorophyll a/b binding protein 2)	10 (3%) / 51 (4%)	10 (4%) / 20 (3%)	6 (2%) / 15 (3%)
09103 Lipid metabolism	K01054 (acylglycerol lipase)	3 (1%) / 15 (1%)	3 (1%) / 6 (1%)	6 (2%) / 12 (2%)
09122 Translation	K02868 (large subunit ribosomal protein L11e)	27 (8%) / 47 (4%)	44 (16%) / 97 (16%)	16 (6%) / 32 (6%)
Hypothetical Proteins	NA	125 (37%) / 357 (27%)	91 (34%) / 220 (31%)	88 (32%) / 177 (32%)

[a] Not all identifiers are listed.
[b] A total of 336, 265 and 276 HSDs were identified within the eukaryotic genomes of UWO241, ICE-L and *C. eustigma* encompassing 1,339, 717 and 560 gene copies, respectively. HSDs share ≥90% pairwise amino acid identity and have lengths within 10 amino acids of each other.

Table 3. Estimation of the number of duplicated genes and HSDs in different species.

Species	No. of Considered Genes	No. of Estimated Duplicated Genes	% Estimated Duplicated Genes	Methodology	Duplicated Gene Types	References
Arabidopsis thaliana	25,557	11,937	46.7%	All-against-all BLASTN [a]	Not specified, all paralogous pairs were searched	[38]
	27334	1500	5.5%	HSDFinder [b]	All paralogous pairs were searched	
Mus musculus (mouse)	21,305	14,034	65.9%	All-against-all BLASTP [c]	Gene families (tandem duplications searched among families)	[39]
	84985	56802	66.8%	HSDFinder [b]	All paralogous pairs were searched	

[a] All-against-all nucleotide sequence similarity searches using BLASTN among the transcribed sequences. Sequences aligned over >300 bp and showing at least 40% identity were defined as pairs of paralogs.
[b] All-against-all protein sequence similarity search using BLASTP filtered via the criteria within 10 amino acids difference and ≥90% amino acid pairwise identities.
[c] All-against-all protein sequence similarity search using BLASTP with the BLOSUM62 matrix and the SEG filter [40], TribeMCL with the default parameters. Tandem duplications were then searched for among families

Figure 1: The workflow and the file examples of HSDFinder.

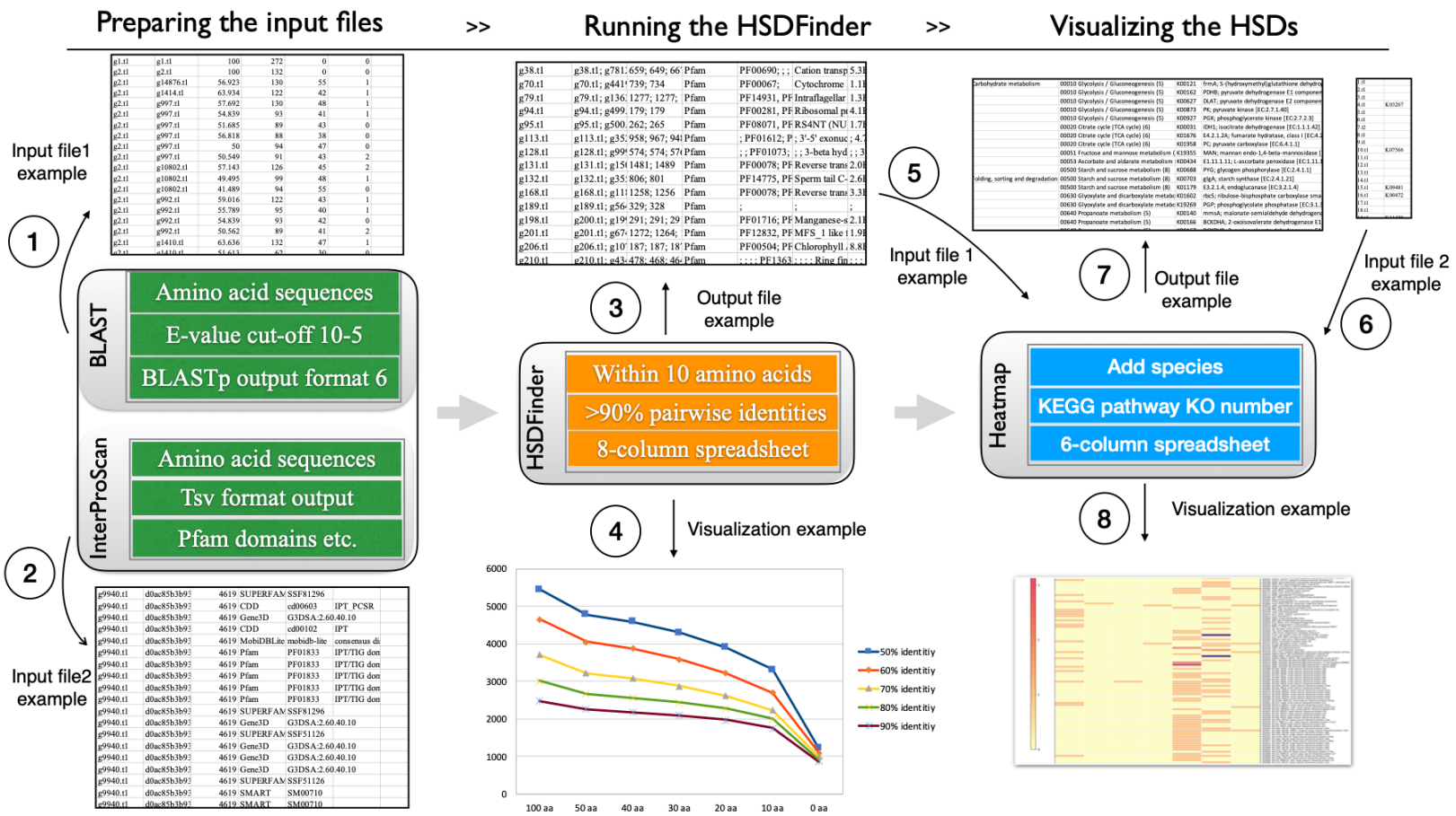
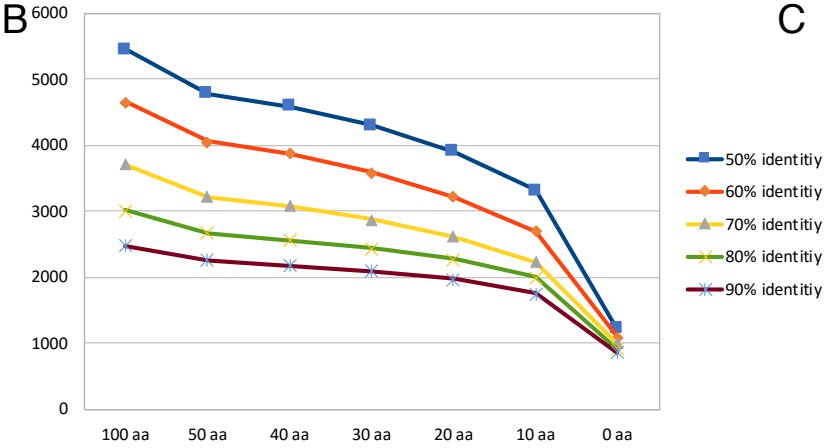


Figure 2: The interpretation of predicted results from HSDFinder.

A

UWO241gene identifers	HSDs gene copies (aa length identity >=90%, within 10aa)	aa length		Pfam identifer	Pfam des	E-value	IPR identifer	IPR des
g1516.t1	g1516.t1; g15297.t1; g3710.t1; g15900.t1; g12375.t1; g865	228; 228; 228; 228; 228; 229; 233; 233	Pfam	PF00098; PF00(	Zinc knuckle; Zinc knuckle; Zinc knuckle; Zinc kr	1.1E-4; 3.8E-5; (	IPR001878; IPR	Zinc finger, CCI
g11310.t1	g11310.t1; g11375.t1	1307; 1312	Pfam	PF00098; PF00(	Zinc knuckle, Integrase core domain, Reverse tran	3.4E-4, 3.0E-16,	IPR001878, IPR	Zinc finger, CCI
g807.t1	g807.t1; g4057.t1	464; 469	Pfam	PF14240; PF14	YHYH protein; YHYH protein	6.8E-9; 3.6E-9	IPR025924; IPR	YHYH domain;
g5701.t1	g5701.t1; g9150.t2	884; 885	Pfam	PF00400;	WD domain, G-beta repeat;	0.0019;	IPR001680;	WD40 repeat;
g767.t1	g15539.t1; g767.t1	231; 231	Pfam	PF10260; PF10	Uncharacterized conserved domain (SAYSvFN);	2.3E-19; 2.3E-1	IPR019387; IPR	Uncharacterised
g5844.t1	g5920.t1; g5844.t1	256; 256	Pfam	PF02902; PF02	Ulp1 protease family, C-terminal catalytic domain	6.2E-6; 6.2E-6	IPR003653; IPR	Ulp1 protease fa
g6100.t1	g12590.t1; g6100.t1	159; 159	Pfam	PF00179; PF00	Ubiquitin-conjugating enzyme; Ubiquitin-conjuga	7.7E-44; 7.7E-4	IPR000608; IPR	Ubiquitin-conjug
g3684.t1	g3684.t1; g6795.t1	137; 130	Pfam	PF00240; PF01(	Ubiquitin family, Ribosomal L40e family; Ubiqui	6.5E-34, 5.1E-2	IPR000626, IPR	Ubiquitin domai
g5645.t1	g5645.t1; g15870.t2	599; 605	Pfam	PF00443; PF00	Ubiquitin carboxyl-terminal hydrolase; Ubiquitin (	4.3E-24; 7.0E-4	IPR001394; IPR	Peptidase C19, u
g2201.t1	g2201.t1; g15994.t1; g15997.t1; g15991.t1	442; 442; 442; 442	Pfam	PF00091, PF03	Tubulin/FtsZ family, GTPase domain, Tubulin C-	5.1E-68, 2.1E-4	IPR003008, IPR	Tubulin/FtsZ, G
g4802.t1	g4816.t1; g4805.t1; g4802.t1	450; 450; 450	Pfam	PF00091, PF03	Tubulin/FtsZ family, GTPase domain, Tubulin C-	7.6E-67, 1.8E-5(	IPR003008, IPR	Tubulin/FtsZ, G
g1131.t1	g1131.t1; g9728.t1	1744; 1743	Pfam	PF03151; PF00(	Triose-phosphate Transporter family; Reverse tran	6.4E-11; 1.3E-2	IPR004853; IPR	Sugar phosphate
g645.t1	g9104.t1; g645.t1	196; 196	Pfam	PF07500; PF07	Transcription factor S-II (TFIIS), central domain;	4.1E-11; 4.1E-1	IPR003618; IPR	Transcription elo
g15800.t1	g15800.t1; g12147.t1	1638; 1643	Pfam	PF14249, PF00	Tocopherol cyclase, Adenylate and Guanylate cyc	1.7E-23, 3.9E-4	IPR025893, IPR	Tocopherol cycla
g5257.t1	g5257.t1; g13535.t1; g7304.t1; g14487.t1	1296; 1305; 1314; 1317	Pfam	PF04278; PF00(	Tic22-like family; Reverse transcriptase (RNA-de	3.3E-39; 5.6E-2	IPR007378; IPR	Tic22-like; Reve
g8510.t1	g8742.t1; g8510.t1	296; 296	Pfam	PF00082; PF00(	Subtilase family; Subtilase family	1.8E-45; 1.8E-4	IPR000209; IPR	Peptidase S8/S5
g13122.t1	g13122.t1; g13744.t1; g12836.t1; g4052.t1; g15392.t1; g13	344; 349; 348; 339; 355; 338; 336; 354;	Pfam	PF00588; PF00	SpoU rRNA Methylase family; SpoU rRNA Met	4.1E-8; 4.2E-10;	IPR001537; IPR	tRNA/rRNA me
g132.t1	g132.t1; g3556.t1	806; 801	Pfam	PF14775, PF14	Sperm tail C-terminal domain, Sperm tail; ARD/A	2.6E-8, 1.3E-19;	IPR029440, IPR	Dynein regulator
g3054.t1	g3054.t1; g11238.t1	306; 306	Pfam	PF16891, PF00	Serine-threonine protein phosphatase N-terminal (	1.6E-19, 2.7E-42	IPR031675, IPR	Serine-threonine
g429.t1	g429.t1; g3694.t1	930; 937	Pfam	PF00530, PF00(	Scavenger receptor cysteine-rich domain, Subtilas	2.7E-8, 7.4E-44;	IPR001190, IPR	SRCR domain, F
g4237.t1	g10399.t1; g10296.t1; g10295.t1; g4237.t1	366; 366; 366; 366	Pfam	PF13445; PF13	RING-type zinc-finger; RING-type zinc-finger; R	3.7E-9; 3.7E-9;	IPR027370; IPR	RING-type zinc-
g4365.t1	g11990.t1; g4365.t1	284; 284	Pfam	PF13639; PF13(	Ring finger domain; Ring finger domain	2.5E-6; 2.5E-6	IPR001841; IPR	Zinc finger, RIN
g3338.t1	g3338.t1; g9313.t1	784; 785	Pfam	PF13639; PF00(	Ring finger domain; Reverse transcriptase (RNA-	2.9E-10; 1.2E-14	IPR001841; IPR	Zinc finger, RIN
g4681.t1	g4681.t1; g5342.t1; g12113.t1; g5638.t1	570; 575; 575; 567	Pfam	PF00355, PF00(	Rieske [2Fe-2S] domain, Reverse transcriptase (R	6.0E-15, 1.4E-25	IPR017941, IPR	Rieske [2Fe-2S]
g1206.t1	g1206.t1; g13528.t1; g10711.t1	171; 167; 168	Pfam	PF00101; PF00	Ribulose bisphosphate carboxylase, small chain; F	3.0E-42; 2.8E-42	IPR000894; IPR	Ribulose bispho
g4489.t1	g14608.t1; g4489.t1	258; 258	Pfam	PF01015; PF01(	Ribosomal S3Ae family; Ribosomal S3Ae family	6.9E-93; 6.9E-93	IPR001593; IPR	Ribosomal prote
g417.t1	g417.t1; g8017.t1	190; 190	Pfam	PF01775; PF01	Ribosomal proteins 50S-L18Ae/60S-L20/60S-L1	1.2E-51; 1.2E-5	IPR023573; IPR	Ribosomal prote
g1892.t1	g1892.t1; g15077.t1	147; 147	Pfam	PF00828; PF00	Ribosomal proteins 50S-L15, 50S-L18e, 60S-L27	1.6E-23; 1.7E-23	IPR021131; IPR	Ribosomal prote

B



C

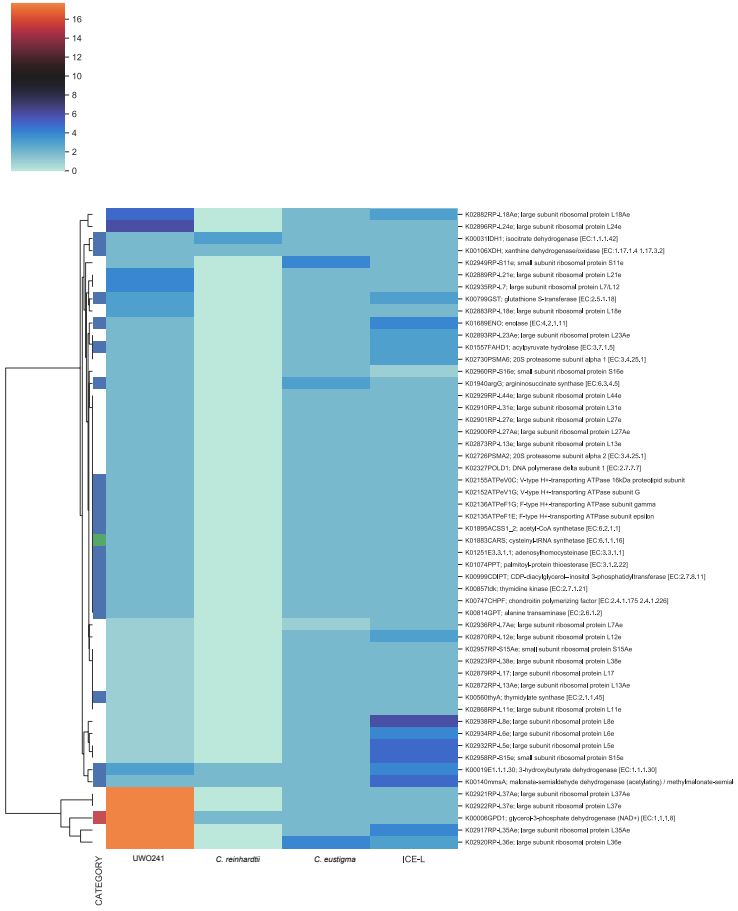
	50% identitiy	60% identitiy	70% identitiy	80% identitiy	90% identitiy
100 aa	5453	4654	3719	3012	2476
50 aa	4783	4061	3228	2672	2257
40 aa	4594	3879	3088	2565	2181
30 aa	4310	3596	2875	2438	2091
20 aa	3912	3224	2614	2270	1973
10 aa	3320	2699	2234	2004	1753
0 aa	1223	1097	999	926	859

Figure 3: The visualization of HSDs results in a heatmap.

A

				uwo241	aa length
Carbohydrate metabolism	00010 Glycolysis / Gluconeogenesis (5)	K00121	frmA; S-(hydroxymethyl)glutathi	g5779.t1, g8291.t1, g10382.t1	340, 380, 391
	00010 Glycolysis / Gluconeogenesis (5)	K00162	PDHB; pyruvate dehydrogenase E1	g5515.t1, g10654.t1	284, 1250
	00010 Glycolysis / Gluconeogenesis (5)	K00627	DLAT; pyruvate dehydrogenase E2	g3352.t1, g10467.t1, g11435.t1	643, 436, 1581
	00010 Glycolysis / Gluconeogenesis (5)	K00873	PK; pyruvate kinase [EC:2.7.1.40]	g4065.t1, g7107.t1	508, 152
	00010 Glycolysis / Gluconeogenesis (5)	K00927	PGK; phosphoglycerate	g5036.t1, g6745.t1	174, 301
	00020 Citrate cycle (TCA cycle) (6)	K00031	IDH1; isocitrate dehydrogenase [EC:	g114.t1, g1446.t1, g1540.t1, g2217.t1,	159, 156, 159, 159, 159, 162, 159, 159,
	00020 Citrate cycle (TCA cycle) (6)	K01676	E4.2.1.2A; fumarate hydratase, class I [EC:	g789.t1, g5752.t1, g13413.t1	738, 598, 93
	00020 Citrate cycle (TCA cycle) (6)	K01958	PC; pyruvate carboxylase [EC:	g4041.t1, g7907.t1	243, 1277
	00051 Fructose and mannose metabolism	K19355	MAN; mannan endo-1,4-beta-	g3766.t1, g8252.t1	459, 459
	00053 Ascorbate and aldarate metabolism	K00434	E1.11.1.11; L-ascorbate peroxidase	g15877.t1, g15878.t1	413, 405
	00500 Starch and sucrose metabolism (8)	K00688	PYG; glycogen phosphorylase [EC:	g4896.t1, g14940.t1	518, 175
Folding, sorting and degradation	00500 Starch and sucrose metabolism (8)	K00703	glgA; starch synthase [EC:2.4.1.21]	g6999.t2, g12919.t1	852, 855
	00500 Starch and sucrose metabolism (8)	K01179	E3.2.1.4; endoglucanase [EC:	g7994.t1, g7995.t1	446, 610
	00630 Glyoxylate and dicarboxylate metabolism (8)	K01602	rbcS; ribulose-bisphosphate carboxylase small	g1206.t1, g10711.t1, g13528.t1	171, 168, 167
	00630 Glyoxylate and dicarboxylate	K19269	PGP; phosphoglycolate	g3281.t1, g9851.t1, g16042.t1	154, 348, 300
	00640 Propanoate metabolism (5)	K00140	mmsA; malonate-semialdehyde	g6608.t1, g6615.t1, g6616.t1, g6617.t1	552, 84, 84, 84
	00640 Propanoate metabolism (5)	K00166	BCKDHA; 2-oxoisovalerate	g6185.t1, g8492.t1	428, 148
	00640 Propanoate metabolism (5)	K00167	BCKDHB; 2-oxoisovalerate	g5907.t1, g13837.t1	394, 85
	00562 Inositol phosphate metabolism	K00999	CDIPT; CDP-diacylglycerol--inositol	g9130.t1, g13028.t1	268, 268
	00562 Inositol phosphate metabolism	K01858	INO1; myo-inositol-1-phosphate synthase	g1123.t1, g11073.t2	1503, 520

B



References

1. Zhang, J. (2003). Evolution by gene duplication: an update. Trends in Ecology & Evolution *18*, 292-298.

2. Kubiak, M.R., and Makałowska, I. (2017). Protein-coding genes' retrocopies and their functions. Viruses *9*, 1-27.

3. Libuda, D.E., and Winston, F. (2006). Amplification of histone genes by circular chromosome formation in *Saccharomyces cerevisiae*. Nature *443*, 1003-1007.

4. Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics *11*, 97-108.

5. Qian, W., and Zhang, J. (2008). Gene dosage and gene duplicability. Genetics *179*, 2319-2324.

6. Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. Plant Physiology *171*, 2294-2316.

7. Kondrashov, F.A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. Proceedings of the Royal Society B: Biological Sciences *279*, 5048-5057.

8. Hirooka, S., Hirose, Y., Kanesaki, Y., Higuchi, S., Fujiwara, T., Onuma, R., Era, A., Ohbayashi, R., Uzuka, A., and Nozaki, H. (2017). Acidophilic green algal genome provides insights into adaptation to an acidic environment. Proceedings of the National Academy of Sciences *114*, 8304-8313.

9. Zhang, Z., Qu, C., Zhang, K., He, Y., Zhao, X., Yang, L., Zheng, Z., Ma, X., Wang, X., and Wang, W. (2020). Adaptation to extreme Antarctic environments revealed by the genome of a sea ice green alga. Current Biology *30*, 1-12.

10. Zhang, X., Cvetkovska, M., Morgan-Kiss, R., Hüner, N.P.A., and Smith, D.R. (2021). Is Gene Duplication Driving Cold Adaptation in the Antarctic Green Alga *Chlamydomonas* sp. UWO241?. Available at SSRN 3732378.

11. Rosikiewicz, W., Kabza, M., Kosiński, J.G., Ciomborowska-Basheer, J., Kubiak, M.R., and Makałowska, I. (2017). RetrogeneDB–a database of plant and animal retrocopies. Database *2017*.

12. Lallemand, T., Leduc, M., Landès, C., Rizzon, C., and Lerat, E. (2020). An overview of duplicated gene detection methods: Why the duplication mechanism has to be accounted for in their choice. Genes *11*, 1046.

13. Wang, Y., Li, J., and Paterson, A.H. (2013). MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. Bioinformatics *29*, 1458-1460.

14. Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y., and Vandepoele, K. (2012). i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. Nucleic Acids Research *40*, e11-e11.

15. Rödelsperger, C., and Dieterich, C. (2010). CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes. PloS One *5*, e8861.

16. Liu, D., Hunt, M., and Tsai, I.J. (2018). Inferring synteny between genome assemblies: a systematic evaluation. BMC Bioinformatics *19*, 1-13.

17. Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. Genome Research *12*, 656-664.

18. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods *12*, 59.

19. Wheeler, T.J., and Eddy, S.R. (2013). nhmmer: DNA homology search with profile HMMs. Bioinformatics *29*, 2487-2489.

20. Koonin, E., and Galperin, M.Y. (2002). Sequence—evolution—function: computational approaches in comparative genomics.

21. Zhang, X., Hu, Y., and Smith, D.R. (2020). HSDatabase - a database of highly similar duplicate genes in eukaryotic genomes. Retrieved from http://hsdfinder.com/database/.

22. Zhang, Z., Qu, C., Zhang, K., He, Y., Zhao, X., Yang, L., Zheng, Z., Ma, X., Wang, X., and Wang, W. (2020). Adaptation to Extreme Antarctic Environments Revealed by the Genome of a Sea Ice Green Alga. Current Biology *30*, 1-12.

23. Zhang, X., Hu, Y., and Smith, D.R. (2021). NoBadWordsCombiner-a tool to integrate the gene function information together without 'bad words' from Nr-NCBI, UniProtKB/Swiss-Prot, KEGG, Pfam databases. Retrieved from https://github.com/zx0223winner/HSDFinder/blob/master/NoBadWordsCombiner.py.

24. Zhang, X., Cvetkovska, M., Morgan-Kiss, R., Hüner, N.P., and Smith, D.R. (2021). Draft genome sequence of the Antarctic green alga *Chlamydomonas* sp. UWO241. iScience, 102084.

25. Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., and Maréchal-Drouard,

L. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science *318*, 245-250.

26. Prochnik, S.E., Umen, J., Nedelcu, A.M., Hallmann, A., Miller, S.M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., and Fritz-Laylin, L.K. (2010). Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science *329*, 223-226.

27. Polle, J.E., Barry, K., Cushman, J., Schmutz, J., Tran, D., Hathwaik, L.T., Yim, W.C., Jenkins, J., McKie-Krisberg, Z., and Prochnik, S. (2017). Draft nuclear genome sequence of the halophilic and beta-carotene-accumulating green alga *Dunaliella salina* strain CCAP19/18. Genome Announcements *5*, 01105-01117.

28. Hanschen, E.R., Marriage, T.N., Ferris, P.J., Hamaji, T., Toyoda, A., Fujiyama, A., Neme, R., Noguchi, H., Minakuchi, Y., and Suzuki, M. (2016). The *Gonium pectorale* genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. Nature Communications *7*, 1-10.

29. Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Brueggeman, A., Dunigan, D.D., Gurnon, J., Ladunga, I., Lindquist, E., and Lucas, S. (2012). The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. Genome Biology *13*, 1-12.

30. Sloan, D.B., Wu, Z., and Sharbrough, J. (2018). Correction of persistent errors in Arabidopsis reference mitochondrial genomes. The Plant Cell *30*, 525-527.

31. Soderlund, C., Descour, A., Kudrna, D., Bomhoff, M., Boyd, L., Currie, J., Angelova, A., Collura, K., Wissotski, M., and Ashley, E. (2009). Sequencing, mapping, and analysis of 27,455 maize full-length cDNAs. PLoS Genetics *5*, e1000740.

32. Mock, T., Otillar, R.P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., and Ward, B.J. (2017). Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. Nature *541*, 536-540.

33. Shao, Y., Lu, N., Wu, Z., Cai, C., Wang, S., Zhang, L.-L., Zhou, F., Xiao, S., Liu, L., and Zeng, X. (2018). Creating a functional single-chromosome yeast. Nature *560*, 331-335.

34. Hoskins, R.A., Carlson, J.W., Wan, K.H., Park, S., Mendez, I., Galle, S.E., Booth, B.W., Pfeiffer, B.D., George, R.A., and Svirskas, R. (2015). The Release 6 reference sequence of the Drosophila melanogaster genome. Genome Research *25*, 445-458.

35. Koutsovoulos, G., Kumar, S., Laetsch, D.R., Stevens, L., Daub, J., Conlon, C., Maroon, H., Thomas, F., Aboobaker, A.A., and Blaxter, M. (2016). No evidence

for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. Proceedings of the National Academy of Sciences *113*, 5053-5058.

36. Mohajeri, K., Cantsilieris, S., Huddleston, J., Nelson, B.J., Coe, B.P., Campbell, C.D., Baker, C., Harshman, L., Munson, K.M., and Kronenberg, Z.N. (2016). Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23. 1 region. Genome Research *26*, 1453-1467.

37. Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W.M., and Ritchie, G.R. (2011). Modernizing reference genome assemblies. PLoS Biology *9*, e1001091.

38. Blanc, G., and Wolfe, K.H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. The Plant Cell *16*, 1667-1678.

39. Shoja, V., and Zhang, L. (2006). A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. Molecular Biology and Evolution *23*, 2134-2141.

40. Wootton, J.C., and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry *17*, 149-163.

Curriculum Vitae

Name: Xi Zhang

Post-secondary Education and Degrees:

The University of Western Ontario
London, Ontario, Canada
2016-2021 Ph.D.

Tianjin University
Tianjin, China
2013-2016 M.Sc.

Tianjin University
Tianjin, China
2008-2012 B.Sc.

Honours and Awards:

Graduate Research Assistant Fellowship
The University of Western Ontario
2016-2020

Spring Travel Award
The University of Western Ontario
2019-2020

Spring Travel Award
The University of Western Ontario
2018-2019

Related Work Experience

Teaching Assistant
University of Western Ontario
2016-2020

Publications:

Zhang, X., Hu, Y., Smith, D. R. (2021). HSDFinder - an integrated tool for predicting highly similar duplicates in eukaryotic genomes. (Manuscript)

Zhang, X., Hu, Y., Smith, D. R. (2021). HSDatabase - a database of highly similar duplicate genes in eukaryotic genomes. (Manuscript)

Zhang, X., Hu, Y., Smith, D. R. (2021). Protocol for HSDFinder to help identify, categorize and annotate duplicate genes in eukaryotic nuclear genomes. STAR Protocols (Manuscript).

Zhang, X., Cvetkovska, M., Morgan-Kiss, R., Hüner, N.P., and Smith, D.R. (2021). Draft genome sequence of the Antarctic green alga *Chlamydomonas* sp. UWO241. iScience, 102084.

Zhang, X., Bauman, N., Brown, R., Richardson, T.H., Akella, S., Hann, E., Morey, R. and Smith, D.R. (2019). The mitochondrial and chloroplast genomes of the green alga Haematococcus are made up of nearly identical repetitive sequences. Current Biology 29: 736-737.

Zhang, X., Peng, C., Zhang, G. and Gao, F. (2015). Comparative analysis of essential genes in prokaryotic genomic islands. Scientific Reports 5: 1-9.

Peng, C., Luo, H., **Zhang, X**. and Gao, F. (2015). Recent advances in the genome-wide study of DNA replication origins in yeast. Frontiers in Microbiology 6: 1-7.