

Sequencing, assembly and annotation of the mitochondrial and plastid genomes of *Gelidium pristoides* (Turner) Kützing from Kenton-on-Sea, South Africa



University of Fort Hare
Together in Excellence

A DISSERTATION



Submitted in fulfilment of the requirements for the degree of

University of Fort Hare
Together in Excellence
MASTER OF SCIENCE (MSc)

IN

(BIOCHEMISTRY)

in the Faculty of Science and Agriculture at the University of Fort Hare

DEPARTMENT OF BIOCHEMISTRY AND MICROBIOLOGY

BY

MANGALI SANDISIWE

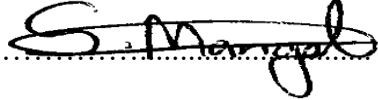
JANUARY 2019

SUPERVISOR: PROF. GRAEME BRADLEY

DECLARATIONS

I, Mangali Sandisiwe (201314918), hereby declare that this dissertation titled "Sequencing, assembly and annotation of the mitochondrial and plastid genomes of *Gelidium pristoides* (Turner) Kützing from Kenton-on-Sea, South Africa" submitted in fulfilment of the Master of Science degree in Biochemistry at the University of Fort Hare is my own work, and has not been submitted for any other degree.

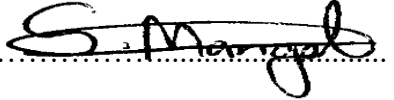
Date...14th April 2019.....

Signature.....

I, Mangali Sandisiwe (201314918), hereby declare that I am fully aware of the University of Fort Hare policy on plagiarism and I have taken every necessary precaution to comply with the regulations of the university.

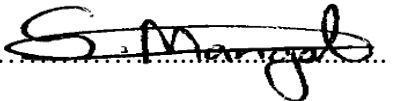
Date...14th April 2019.....



Signature.....

I, Mangali Sandisiwe (201314918), hereby declare that I am fully aware of the University of Fort Hare's policy on research ethics and I have taken every precaution to comply with the regulations. I confirm that my research constitutes an exemption to Rule G17.6.10.5 and an ethical certificate with a reference number is not required.

Date...14th April 2019.....

Signature.....

DEDICATION

I dedicate this dissertation to my Lord and Saviour, to my mom, Nanahomba Angelinah Mangali and also to my siblings, Linda Christopher, Lindiwe Christina, Nosicelo Lilian and Nosithembele Leticia Mangali. I further dedicate this hard work to my daughter, Ngomsolwethu Lulonke Mangali as an encouragement for academic desires and achievements.



University of Fort Hare
Together in Excellence

ACKNOWLEDGEMENTS

I thank the Almighty God for strength and courage. You offered me to make this work possible. Without you, Lord, none of this would have been possible.

I offer my sincere gratitude to my supervisor, Prof G. Bradley for being an excellent supervisor and a mentor towards my project. For being a parent in both happy and sad times. Thank you, Prof, for your guidance and constructive professional advice. The Lord will reward you perfectly.

I thank my mom, Nanahomba Angelinah Mangali for both her emotional and financial support. Thank you, mom, for being there as my engine and motivator. I also thank my siblings for their support and motivation.

I am also very grateful to Mr Carel Van Heerden from Stellenbosch University for his excellent expertise in the genome sequencing part of Biology.

I am very grateful to Dr Nqumla for her support and guidance towards my project.

I am very thankful to Gogela Yanga for his support and assistance during the stressful moments and ‘second-hand’-demanding experiments of my project. May God bless you.

I thank my colleagues and departmental staff for their support towards this difficult journey.

I thank you, Dr Vuyani Moses, from Rhodes University and Phumelela from the University of Fort Hare for your assistance towards the installation of Linux computer programs.

I offer a very special thanks to the South African Institute of Aquaculture and Biodiversity (SAIAB), ACEP Phuhlisa, the National Research Foundation (NRF) and the Govern Mbeki Research and Development Centre (GMRDC) for development and financial support.

I am also very grateful to the University of Fort Hare for affording me a platform for conducting this research.

Lastly, I would like to thank everyone that has ever had a contributing role towards my project.



University of Fort Hare
Together in Excellence

ABSTRACT

The genome is the complete set of an organism's hereditary information that contains all the information necessary for the functioning of that organism. Complete nuclear, mitochondrial and plastid DNA constitute the three main types of genomes which play interconnected roles in an organism. Genome sequencing enables researchers to understand the regulation and expression of the various genes and the proteins they encode. It allows researchers to extract and analyse genes of interests for a variety of studies including molecular, biotechnological, bioinformatics and conservation and evolutionary studies. Genome sequencing of Rhodophyta has received little attention. To date, no published studies are focusing on both whole genome sequencing and sequencing of the organellar genomes of Rhodophyta species found in along the South African coastline.

This study focused on genome sequencing, assembly and annotation mitochondrial and plastid genomes of *Gelidium pristoides*. *Gelidium pristoides* was collected from Kenton-on-Sea and was morphologically identified at Rhodes University. Its genomic DNA was extracted using the Nucleospin® Plant II kit and quantified using Qubit 2.0, Nanodrop and 1% agarose gel electrophoresis. The Ion Plus Fragment Library kit was used for the preparation of a 600 bp library, which was sequenced in two separate runs through the Ion S5 platform. The produced reads were quality-controlled through the Ion Torrent server version 5.6. and assessed using the FASTQC program. The SPAdes version 3.11.1 assembler was used to assemble the quality-controlled reads, and the resultant genome assembly was quality-assessed using the QUAST 4.1 software. The mitochondrial genome was selected from the produced *Gelidium pristoides* draft genome using mitochondrial genomes of other *Gelidiales* as search queries on the local BLAST algorithm of the BioEdit software. Contigs matching the organellar genomes were ordered according to the mitochondrial genomes of other *Gelidiales* using the trial version of Geneious R11.12 software. The plastid genome was also selected following the same approach but using plastid genomes of *Gelidium elegans* and *Gelidium vagum* as search queries. Gaps observed in the organellar genomes were closed by amplification of the relevant gap using polymerase chain reaction with newly designed primers and Sanger sequencing. Open reading frames for both organellar genomes were annotated using the NCBI ORF-Finder and alignments obtained from BlastN and BlastX searches from the NCBI database, while the tRNAs and rRNAs were identified using the tRNAscan-SE1.21

and the RNAmmer 1.2 servers. The circular physical map of the mitochondrial genome was constructed using the CGView server. Lastly, *in silico* analysis of cytochrome c oxidase 3 and Heat Shock Protein 70 was performed using the PRIMO and the SWISS-MODEL pipelines respectively. Their phylogenies were analysed through Clustal omega and the trees viewed on TreeView 1.6.6 software.

Qubit and Nanodrop genomic DNA qualification revealed A_{260}/A_{280} and A_{230}/A_{260} ratios of 1.81 and 1.52 respectively. The 1% agarose gel electrophoresis further confirmed the good quality of the genomic DNA used for library preparation and sequencing. Pre-assembly quality control of reads resulted in a total of 30 792 074 high-quality reads which were assembled into a total of 94140 contigs, making up an estimated genome length of 217.06 Mb. The largest contig covered up to 13.17 kb of the draft genome, and an N50 statistic value of 3.17 kb was obtained.

The *G.pristoides* mitochondrial genome mapped into a circular molecule of 25012 bp, with an overall GC content of 31.04% and a total of 45 genes distributed into 20 tRNA-coding, 2 rRNA-coding genes and 23 protein-coding genes, mostly adopting the modified genetic code of Rhodophyta. The *SecY* and *rps12* genes overlapped by 41 bp. This study presents a partial plastid genome composed of 89 (38%) fully annotated genes, of which 71 are protein-coding, and 18 are distributed among 15 tRNA-coding, 2 rRNA-coding and 1 RNaseP RNA-coding genes. Sixty-one (26%) partial protein-coding genes were predicted, while approximately 84 (36%) genes are not yet predicted. *In silico* analysis of the cytochrome c oxidase and heat shock protein 70 showed that the gene sequences obtained in this study and the resultant transcribed protein have sequences and structures that are similar to those from several other different species, thus validating the integrity of the genome sequences. This study provides genomic data necessary for understanding the genomic constituent of *G.pristoides* and serve as a foundation for studies of individual genes and for resolving evolutionary relationships.

TABLE OF CONTENTS

DECLARATIONS	ii
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
ABSTRACT	v
TABLE OF CONTENTS	vii
LIST OF ABBREVIATIONS	x
LIST OF FIGURES	xiii
LIST OF TABLES	xvi
BRIEF SYNOPSIS OF CHAPTERS	xviii
CHAPTER ONE	xviii
CHAPTER TWO	xviii
CHAPTER THREE	xviii
CHAPTER FOUR	xviii
CHAPTER FIVE	xviii
1 CHAPTER ONE: LITERATURE REVIEW	1
1.1 General introduction to red algae species (Rhodophytes)	1
1.2 Economic and Medical benefits of Red algae	2
1.3 Sequenced <i>Rhodophyta species</i> <i>Together in Excellence</i>	2
1.4 Sequenced genomes of Rhodophytes in South Africa	3
1.5 Geographical distribution of marine Rhodophyta in South Africa	5
1.6 Current Rhodophyta studies in South Africa	6
1.7 The Choice of <i>Gelidium pristoides</i> for this study	6
1.8 <i>Gelidium pristoides</i>	7
1.9 Genome types	8
1.10 Significance of genome sequences	11
1.11 Genome sequencing	12
1.12 Genome sequencing approaches	12
1.12.1 The Hierarchical shotgun sequencing and whole genome shotgun (WGS) sequencing	12
1.12.2 Sanger sequencing	14
1.12.3 The Next Generation Sequencing	15
1.13 The choice of sequencing technology	16
1.13.1 The Principle of Ion Torrent	17
1.13.2 The Ion S5 sequencing platform	19

1.14	Genome assembly.....	19
1.14.1	Reference-based assembly	20
1.14.2	De novo genome assembly	21
1.14.3	Greedy assemblers	21
1.14.4	The OLC Assemblers.....	22
1.14.5	The DBG assemblers	24
1.15	Genome annotation.....	25
1.16	Mapping.....	26
1.16.1	Computational gene prediction	26
1.17	Problem statement	26
1.18	Hypothesis	27
1.19	Aim.....	27
1.20	Specific objectives.....	27
2	CHAPTER TWO: MATERIALS AND METHODS	29
2.1	Sampling and storage of <i>Gelidium pristoides</i>	29
2.2	Sample preparation.....	30
2.3	Genomic DNA (gDNA) extraction and qualification (quantity and quality).....	30
2.4	Library preparation and quantification.....	30
2.5	Template preparation, Enrichment and Sequencing.....	31
2.6	Pre-assembly quality assessment.....	32
2.7	<i>De novo</i> genome assembly	32
2.8	Post-assembly quality assessment	32
2.9	Identification and selection of <i>Gelidium pristoides</i> organellar (mitochondrial and plastid) genomes.....	33
2.10	Organellar genome gap filling through Sanger sequencing chemistry.....	33
2.10.1	Amplification of the organellar gaps and purification of PCR products.....	33
2.11	Quantitation and cycle sequencing of the organellar genomes.....	34
2.12	Ethanol-EDTA precipitation and Capillary electrophoresis of products of cycle sequencing	35
2.13	Annotation of organellar genomes and construction of circular mitochondrial	36
2.14	Structural analysis of the mitochondrial Cytochrome c oxidase subunit 3 (Cox3) and plastid heat shock protein 70 (HSP70)	36
2.15	Phylogeny analysis of the <i>Gelidium pristoides</i> Cox3 and HSP70.....	37
3	CHAPTER THREE: RESULTS AND DISCUSSION.....	40
3.1	gDNA extraction and qualification (quantity and quality).....	40
3.2	Library preparation and quantification	41



3.3	Genome sequencing and <i>de novo</i> genome assembly analysis	42
3.3.1	Pre-assembly quality assessment	42
3.4	Post-genome assembly assessment.....	48
3.4.1	Estimated length of <i>Gelidium pristoides</i> genome assembly	48
3.4.2	The N50 genomic statistic.....	48
3.4.3	The GC Content	48
3.5	Identification and selection of <i>Gelidium pristoides</i> organellar genomes	49
3.6	Organellar genome gap filling.....	52
3.6.1	Mitochondrial gap amplification.....	52
3.6.2	Plastid gap amplification.....	54
3.7	Annotation of <i>Gelidium pristoides</i> organellar genomes	57
3.7.1	The <i>Gelidium pristoides</i> mitochondrial genome.....	57
3.7.2	The <i>Gelidium pristoides</i> plastid genome.....	64
3.8	Construction of 3D model structure for the mitochondrial Cox3 and plastid HSP70	71
3.9	Evaluation of the Cox3 and HSP70 3D model quality	74
3.10	Phylogeny analysis of the mitochondrial Cox3 and Plastid HSP70 proteins	76
4	CHAPTER FOUR: GENERAL DISCUSSION, CONCLUSION AND FUTURE PROSPECTS	84
4.1	General discussion.....	84
4.1.1	Introduction.....	84
4.1.2	Rhodophyta mitochondrial genomes.....	84
4.1.3	Rhodophyta plastid genomes	86
4.2	Conclusion.....	87
4.3	Future prospects.....	87
5	CHAPTER FIVE: REFERENCES OF THE STUDY.....	89
A	APPENDICES: Referral Information	104
A.1	APPENDIX A: Primers used for closing the organellar genome gaps.....	104
A.2	APPENDIX B: Mitochondrial and plastid genome gap filling	111
A.3	APPENDIX C: Mitochondrial Annotation File.....	118
A.4	APPENDIX D: Plastid Annotation File	132




LIST OF ABBREVIATIONS

-A	Adenine
-ATP	Adenosine triphosphate
-AMOS	A Modular Open-Source
-ABYSS	Assembly By Short Sequences
-ABI	Applied Biosystems
-Å	Ångström
-BAC	Bacterial artificial chromosome
-BLAST	Basic Local Alignment Tools
-BlastN	Basic Local Alignment Search Tool N
-BlastX	Basic Local Alignment Search Tool X
-bp	Base pairs
-C	Cytosine
-CNV	Copy Number Variation
-CAP	Contig assembler program
-CABOG	Celera Assembler with the Best Overlap Graph
-CAF	Central Analytical Facilities
-CDF	Computational Fluid Dynamics
-Cox3	Cytochrome c oxidase 3
-CTAB	Cetyltrimethylammonium Bromide
-DNA	Deoxyribonucleic acid
-dNTPs	Deoxynucleotides
-DBG	De Bruijn graph
-dsDNA	Double-stranded DNA
-emPCR	Emulsification Polymerase Chain Reaction
-EST	Expressed sequence tag
-EDTA	Ethylene Diamine Triacetic Acid
-ETC	Electron Transport Chain

-FastAP	Fiscal and Strategic Technical Assistance Program
-FASTA	FAST-All
-G	Guanine
-GC	Guanine Cytosine
-gDNA	Genomic DNA
-HS	High sensitivity
-HSP70	Heat Shock Protein 70
-HGPO	Human Genome Project Organization
-HGP	Human Genome Project
-Inc.	Incorporated
- kb	Kilobases
-KEGG	Kyoto Encyclopedia of Genes and Genomes
- Mb	Megabases
- mRNA	Messenger Ribonucleic acid
- MIRA	Mimicking Intelligent Read Assembly
-mtDNA	Mitochondrial DNA
-NCBI ORF-Finder	National Center for Biotechnology Information Open Reading Frame Finder
-NCBI	National Center for Biotechnology Information
-NADH	Nicotinamide adenine dinucleotide
-NGS	Next Generation Sequencing
-NHGRI	National Human Genome Research Institute
-ng	Nanograms
-ORF	Open Reading Frame
-ORFs	Open Reading Frames
-OH	Hydroxyl group
- OLC	Overlap layout Consensus



University of Fort Hare
Together in Excellence

- PGM	Personal Genome Machine
-PCR	Polymerase Chain Reaction
-PRIMO	Protein Interactive Modeling
-pM	Picomolar
-PHRAP	Phragment assembly program
-qPCR	Quantitative Polymerase Chain Reaction
-QUAST	Quality Assessment Tool for Genome Assemblies
-RNA	Ribonucleic acid
-rRNAs	Ribosomal ribonucleic acids
- <i>rnpB</i> RNA	RNaseP RNA
-RFLP	Restriction Fragment Length Polymorphism
-SSR	Simple sequence repeats or Short Tandem Repeats
-SNP	Single nucleotide polymorphism
- SSAKE	Short sequence assembly by progressive K-mer
-SHARCGS	Short-read Assembler based on Robust Contig-extension for Genomic Sequencing  University of Fort Hare <i>Together in Excellence</i>
-SOAPdenovo	Short Oligonucleotide Analysis Package <i>de novo</i>
-SAIAB	South African Institute for Aquatic Biodiversity
-T	Thymine
-T-COFFEE	Tree based Consistency Objective Function For alignment Evaluation
-TE	Tris Ethylene Diamine Triacetic Acid
-tRNAs	Transfer ribonucleic acids
-TIGR	The Institute for Genomic Research
-USA	United States of America
-VCAKE	Verified consensus assembly by K-mer extension
-WGS	Whole genome shotgun
-3D	Three dimensional
- μ l	Microliters

LIST OF FIGURES

CHAPTER ONE

Figure 1. 1: Nine South African marine bioregions as defined by Lombard (2004).	5
Figure 1. 2: Photograph of <i>Gelidium pristoides</i> (Turner) Kützing.	8
Figure 1. 3: Illustration of the location of the different types of genomes.	10
Figure 1. 4: Schematic representation of the Genome sequencing approaches.	14
Figure 1. 5: Graphical illustration of the reduction in sequencing costs since the introduction of NGS platforms from 2005 to 2015 and future predictions of the sequencing costs.	16
Figure 1. 6: The general workflow and sequencing chemistry of Ion Torrent sequencing.	18
Figure 1. 7: Illustration of the genome assembly approaches.	20
Figure 1. 8: The general strategy adopted by OLC-based de novo assemblers.	23
Figure 1. 9: The general strategy adopted by DBG-based de novo assemblers.	24
Figure 1. 10: Sequence annotation levels.	25

CHAPTER TWO

Figure 2. 1: Geographical location of the sampling site on the Kenton-On-Sea region in the Eastern Cape Province.	29
Figure 2. 2: Protein sequence of Cox3 protein (MG_0000000001.1) from <i>G.pristoides</i> .	36
Figure 2. 3: Protein sequence of HSP70 protein (MG_111111111.1) from <i>G.pristoides</i> .	37



University of Port Harcourt
Together in Excellence

CHAPTER THREE

Figure 3. 1: Agarose gel electrophoresis (1% agarose ran with 0.5M TAE buffer) of gDNA extracted from <i>Gelidium pristoides</i> .	41
Figure 3. 2: Per Base Sequence Quality plot of reads generated with Ion S5 sequencer system.	44
Figure 3. 3: Per Sequence Quality Scores plot of reads generated with Ion S5 sequencer system	45
Figure 3. 4: Per Sequence GC content plot of reads generated with Ion S5 sequencer system.	46
Figure 3. 5: Illustration of Per Base N Content of reads generated with Ion S5 sequencer system	46
Figure 3. 6: Illustration of Sequence Length Distribution of reads generated with Ion S5 sequencer system	47
Figure 3. 7: GC content distribution curve of the <i>Gelidium pristoides</i> genome estimated by the QCAST 4.1 software operating under default parameter settings on Contigs of <i>Gelidium pristoides</i> .	49
Figure 3. 8: A portion of the mitochondrial genome gaps after mapping of the quality-controlled contigs of <i>G.pristoides</i> against the mitochondrial genomes of other Gelidales.	50

Figure 3. 9: A portion of the <i>G.pristoides</i> plastid genome gaps after mapping of the quality-controlled contigs of <i>G.pristoides</i> against the plastid genomes of <i>G.vagum</i> and <i>G.elegans</i>	51
Figure 3. 10: Agarose gel electrophoresis (1% agarose) of PCR products of <i>Gelidium pristoides</i> mitochondrial genome.	52
Figure 3. 11: Agarose gel electrophoresis (1% agarose) of PCR products of <i>Gelidium pristoides</i> mitochondrial genome.	53
Figure 3. 12: Agarose gel electrophoresis (1% agarose) of PCR products of <i>Gelidium pristoides</i> mitochondrial genome.	53
Figure 3. 13: Agarose gel electrophoresis (1% agarose) of PCR products of <i>Gelidium pristoides</i> plastid genome.	54
Figure 3. 14: Agarose gel electrophoresis (1% agarose) of PCR products of <i>Gelidium pristoides</i> plastid genome.	55
Figure 3. 15: Agarose gel electrophoresis (1% agarose) of PCR products of <i>Gelidium pristoides</i> plastid genome.	56
Figure 3. 16: Agarose gel electrophoresis (1% agarose) of PCR products of <i>Gelidium pristoides</i> plastid genome.	57
Figure 3. 17: Physical map of the circular <i>Gelidium pristoides</i> mitochondrial genome.	58
Figure 3. 18: Distribution of the GC content and nucleotides in the mitochondrial genome of <i>Gelidium pristoides</i>	59
Figure 3. 19: The current status of the <i>Gelidium pristoides</i> plastid genome.	64
Figure 3. 20: Three-dimensional model structure of the Cox3 protein standing alone (a) and fitted into the <i>Bos taurus</i> Cox3 protein (b). The 3D model structure of <i>Gelidium pristoides</i> Cox3 protein was constructed with the PRIMO webservice pipeline.	73
Figure 3. 21: Three-dimensional model structure of the HSP70 (a) and alignment of the HSP70 constructed model with <i>Escherichia coli</i> HSP70 protein (b). The 3D model structure of <i>Gelidium pristoides</i> HSP70 was constructed with the SWISS-MODEL webservice pipeline.	74
Figure 3. 22: Ramachandran plot obtained from PROCHECK evaluation of the PRIMO constructed <i>Gelidium pristoides</i> Cox3 3D model structure.	75
Figure 3. 23: Ramachandran plot obtained from the evaluation of the SWISS-MODEL constructed <i>Gelidium pristoides</i> HSP70 3D model structure.	76
Figure 3. 24: Portion of sequence alignment of the Cox3 protein from eleven Rhodophyta species.	77
Figure 3. 25: Portion of sequence alignment of the HSP70 protein from eleven Rhodophyta species.	78
Figure 3. 26: Cladogram tree of the Cox3 protein from eleven Rhodophyta species.	81
Figure 3. 27: Cladogram tree of the HSP70 from eleven Rhodophyta species.	82

APPENDIX B

Figure referral 1. 1: Agarose gels obtained from PCR amplification of the <i>Gelidium pristoides</i> mitochondrial genome	111
--	-----

Figure referral 1. 2: Agarose gel electrophoresis (1% agarose) of PCR products of Gelidium pristoides mitochondrial genome.....	112
Figure referral 1. 3: Agarose gel electrophoresis (1% agarose) of gel purified PCR products of Gelidium pristoides mitochondrial genome.....	112
Figure referral 1. 4: Agarose gels obtained from PCR amplification of the plastid genome...	115
Figure referral 1. 5: Agarose gel electrophoresis (1% agarose) of PCR products of Gelidium pristoides plastid genome.....	116
Figure referral 1. 6: Agarose gel electrophoresis (1% agarose) of PCR products of Gelidium pristoides plastid genome.....	116
Figure referral 1. 7: Agarose gel electrophoresis (1% agarose) of PCR products of Gelidium pristoides plastid genome.....	117
Figure referral 1. 8: Agarose gel electrophoresis (1% agarose) of PCR products of Gelidium pristoides plastid genome.....	117



University of Fort Hare
Together in Excellence

LIST OF TABLES

CHAPTER ONE

Table 1. 1: Collection sites and Institutes for the genome sequencing of the specific Rhodophyta species	4
---	---

CHAPTER TWO

Table 2. 1: Thermal cycling conditions for purification of PCR products through the Thermo Scientific ExoSAP protocol	34
Table 2. 2: Concentrations of PCR products recommended for BigDye Terminator v3.1 cycle sequencing.....	35
Table 2. 3: Thermal cycling conditions optimised for the Veriti 96 well-thermal cycler for Cycle Sequencing.....	35

CHAPTER THREE

Table 3. 1: Quantification and quality assessment results of gDNA extracted from Gelidium pristoides	40
Table 3. 2: The qPCR library quantification results for different gDNA amounts.....	42
Table 3. 3: The basic statistics of the quality controlled genomic data of Gelidium pristoides generated by the Ion S5 sequencer for the concatenated bam file of sequencing reads	43
Table 3. 4: Genome assembly statistics generated by QUAST based on Contigs.....	48
Table 3. 5: Genes predicted from the mitochondrial genome of Gelidium pristoides.....	60
Table 3. 6: Comparison of tRNAs in the mitochondrial genomes of some Gelidiales.....	63
Table 3. 7: The complete genes predicted in the Gelidium pristoides plastid genome.....	66
Table 3. 8: The partial genes predicted in the Gelidium pristoides plastid genome	69
Table 3. 9: Percentage identity matrix created by Clustal 12.1 during alignment of the Cox3 proteins from different species.....	79
Table 3. 10: Percentage identity matrix created by Clustal 12.1 during alignment of the HSP70 proteins from different species.....	80

BRIEF SYNOPSIS OF CHAPTERS



University of Fort Hare
Together in Excellence

BRIEF SYNOPSIS OF CHAPTERS

CHAPTER ONE

Chapter One reviews relevant published literature and provides insight regarding the progress that has been made not only in the development of next-generation sequencing platforms but also the information regarding the increments made in the number of sequencing projects as portrayed by different bioinformatics databases. This chapter also gives insight into the economic and scientific significance of the *Gelidium pristoides* species, as well as the influential factors which led to the choice of the species under study. Chapter One ends with identifying the knowledge gap that this study hopes to address, with the specific objectives that the current research study proposes to achieve to accomplish the overall aim and ultimately test the hypothesis on which it is premised.

CHAPTER TWO

Chapter Two provides a detailed methodology used in this study. It starts by providing the source of the *Gelidium pristoides* species and the conditions under which it was stored to ensure long term use. It further gives details of the methods used for gDNA extraction, library preparation, sequencing and annotation of the two organellar genomes.

CHAPTER THREE

Chapter Three presents and discusses the results obtained during this study and provides conclusions based on comparisons to other published results related to this type of research. It also gives insights into the unique aspects of the current study, thus revealing novelty.

CHAPTER FOUR

This chapter presents the general discussion; the overall conclusion is drawn from the study and proposes future studies based on these findings. The chapter also presents new opportunities to the scientific community and possible future applications and knowledge that the study brings to the economic and scientific communities.

CHAPTER FIVE

Chapter Five provides the sources, including books, journal articles and websites, cited in this study. This chapter, therefore, provides evidence of assertions and claims made in this study. It is a chapter that allowed me as a researcher to accredit and acknowledge the contribution of other writers and researchers to my research. This chapter further allowed me to acknowledge the



University of Fort Hare
The heart of the nation

intellectual property rights of different researchers, as many ideas for the current study are drawn from the cited sources.



University of Fort Hare
Together in Excellence

CHAPTER ONE

Literature Review



University of Fort Hare
Together in Excellence

1 CHAPTER ONE: LITERATURE REVIEW

1.1 General introduction to red algae species (Rhodophytes)

Seaweed can be categorised as red algae, brown algae and green algae, based on their pigments (McHugh, 2003). Red algae species are a specialised group of seaweeds (Cole and Sheath 1990) classified under phylum Rhodophyta, which is currently classified under kingdom Plantae, also referred to as the kingdom Archaeplastida (Seckbach and Chapman, 2010; Guiry and Guiry, 2018). Phylum Rhodophyta is comprised of two subphyla, the Cyanidophytina and the Rhodophytina subphylum. The Cyanidophytina subphylum mothers the Cyanidiophyceae class while the Rhodophytina subphylum mothers a variety of classes, namely; Bangiophyceae, Compsopgonophyceae, Stylonematophyceae, Rhodellophyceae, Porphyridiophyceae and *Florideophyceae*, containing approximately 7124 species (Guiry and Guiry, 2018; Nan *et al.*, 2017).

Rhodophytes comprise a monophyletic eukaryotic lineage with a fossil record that extends back 1.2 billion years (Butterfield, 2000; Qiu *et al.*, 2015). These species occupy a particular position that serves as a link between primary and secondary endosymbiosis on the tree of life. They are the bearers of primary endoplasmic plastids (DePriest *et al.*, 2014; Delwiche, and Palmer, 1997) and contain phycobilin pigments (Dixon, 1973). They are named for their red colour which arises as a result of the dominating phycoerythrin pigment which masks their photosynthetic antenna (Chlorophyll and carotenoids) (Seckbach and Chapman, 2010). This property is not observed in other red algae species such as *Cyanidioschyzon merolae* (*C. merolae*) and *Galdieria sulphuraria* (*G. sulphuraria*), as these tend to appear blue or greenish due to decreased levels of phycoerythrin and increased levels of the phycocyanin pigment (Hoek *et al.*, 1995). The fact that Rhodophytes lack centrioles, flagella, external endoplasmic reticula and have unstacked thylakoids and floridean starch further differentiate them from other algae species (Dixon, 1973; Cole and Sheath, 1990).

Ecologically, Rhodophytes are predominantly marine dwellers comprising about 7501 species, with some freshwater species constituting a small portion of these organisms (Guiry and Guiry, 2018). Rhodophyta species are claimed to form the largest group of seaweeds (Garrison, 2012) with a broad spectrum of beneficial applications.

1.2 Economic and Medical benefits of Red algae

Rhodophyta species are recognised as highly economically significant species. Their useful applications include their use as a source of food, antimicrobial agents, production of polyelectrolytes, pharmaceutical products, cosmetics, and biofuels (Bird and Ragan, 2012; Bird and Benson, 1987). Rhodophytes are known as a rich source of bioactive compounds such as proteins, lipids, fibres, pigments, polyphenols, and polysaccharides (agar and carrageenan) responsible for imparting various health benefits (Kumar *et al.*, 2008; Holdt and Kraan, 2011). Rhodophyta pigments such as β -carotene and lutein were reported with anticancer activities (Okai *et al.*, 1996). According to Holdt and Kraan (2011), polysaccharides, peptides, proteins and amino acids from different seaweeds were reported with beneficial activities against diabetes, cancer, AIDS, and vascular diseases. They are a rich source of proteins of medicinal importance; for example, they have phycobiliproteins that possess anti-inflammatory, anti-tumour, antioxidant, antiviral and antiatherosclerosis activities (Su *et al.*, 2010). They are a source of heat shock protein 70 (HSP70) which can be potentially used in Huntington disease therapy and in preventing other neurodegenerative diseases such as Parkinson's disease since HSP70 of some other organisms are reported with such medicinal applications (Kurucz *et al.*, 2002; Padmalayam, 2014). *Pyropia yezoensis* (*P. yezoensis*), *Chondrus crispus* (*C. crispus*) and other species are widely used as the source of food for humans and extraction of carrageenan and agar, the widely used microbiological medium (El Gamal, 2010; Brawley *et al.*, 2017). The cosmetic industry and agriculture also benefit from Rhodophytes as the mineral deposits of *Lithothamnion* spp. are used as agricultural fertilisers (Boenigk *et al.*, 2015).

1.3 Sequenced *Rhodophyta* species

Despite the economic and medicinal significance of Rhodophyta genomes, surprisingly few genomic studies have been conducted on these species, particularly in South Africa where ocean economy has been identified as an essential sector that needs to be developed to sustain economic development. According to published literature, only a few genomes of Rhodophytes have been sequenced from more than 7 thousand species (Brawley *et al.*, 2017). According to the NCBI genome database and realDB (“the genome database for red algae”), the sequenced Rhodophyta genomes are of *Cyanidioschyzon merolae* (*C. merolae*), *Kappaphycus alvarezii* (*K. alvarezii*),

Gracilariopsis chorda (*G. chorda*), *Porphyridium purpureum* (*P. purpureum*) also known as *Porphyridium cruentum* (*P. cruentum*), *Porphyra umbilicalis* (*P. umbilicalis*), *Gracilaria chilensis* (*G. chilensis*), *Pyropia pulchra* (*P. pulchra*), *Gracilariopsis lemaneiformis* (*Gp. lemaneiformis*), *Chondrus crispus* (*C. crispus*), *Galdieria sulphuraria* (*G. sulphuraria*), *Thalassiosira pseudonana* (*T. pseudonana*) and *Galdieria phlegrea* (*G. phlegrea*) (<https://www.ncbi.nlm.nih.gov> accessed 01/12/2018; <http://realdb.algaegenome.org/> Accessed on 01/12/2018). The Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2016) further supports the existence of genomes coming from *C. crispus*, *G. sulphuraria* and *C. merolae* (<https://www.genome.jp/kegg/> Accessed on 01/12/2018).


In addition to these sequenced Rhodophyta genomes, a study conducted by Nakamura *et al.*, (2013) revealed a 43 Mbp draft genome of *P. yezoensis*. The NCBI genome database and published studies report numerous sequenced organellar (mitochondrial and plastid) genomes of Rhodophytes, which play a crucial role in evolutionary studies (Lee, 2018). The list of these sequenced organellar genomes includes the genomes of *Gelidium arborescens*, *Gelidium sinicola*, *Gelidium isabelae*, *Gelidium galapagense*, *Gelidium sclerophyllum*, *Gelidium elegans*, *Gelidium vagum* and *Gelidium crenale*, *Gelidium kathyanniae* (*G. kathyanniae*) and *Gelidium gabrielsonii* (*G. gabrielsonii*) (Boo *et al.*, 2016; Boo and Hughey, 2018; <https://www.ncbi.nlm.nih.gov> Accessed on 01/12/2018), as well as various genomes of other Rhodophyta species (<https://www.ncbi.nlm.nih.gov> Accessed on 01/12/2018).

1.4 Sequenced genomes of Rhodophytes in South Africa

Based on available information there are no published South African studies focusing on genome sequencing of the Rhodophyta species. According to Table 1.1, the sequenced *C. crispus* was collected from Peggy's Cove, Nova Scotia, Canada (44°29'31"N, 63°55'11"W) in 1985 by Juan Correa (Collén *et al.*, 2013) while the *Cyanidioschyzon merolae* study was conducted in Italy, at the Naples University (Matsuzaki *et al.*, 2004). The Rhodophyta species, *Kappaphycus alvarezii*, *Gracilariopsis chorda*, *Porphyridium purpureum*, *Porphyra umbilicalis*, *Gracilaria chilensis*, *Pyropia pulchra*, and *Galdieria sulphuraria* were not collected from South Africa for sequencing as they do not exist in South Africa as per the "AlgaeBase" database (Guiry and Guiry, 2018). Although *Gracilariopsis lemaneiformis* is found in South Africa, it was collected from Zhanshan

Bay, Qingdao, China (36°02' N, 120°20' E) for the genome sequencing (Zhou *et al.*, 2013) and this was also the case for sequencing of the organellar genomes of *Gelidiales*, as they were mostly collected in the United States of America (USA) (<https://www.ncbi.nlm.nih.gov> **Accessed on 01/12/2018**). The keyword searches ‘Red algae complete mitochondrion AND South Africa’ and ‘Red algae complete plastid AND South Africa’ on the NCBI genome, and nucleotide databases do not match any published genomic information (<https://www.ncbi.nlm.nih.gov> accessed 01/12/2018). A thorough search of the NCBI database indicated that the institutions in which the sequenced organellar genomes were submitted are found in countries other than South Africa, including China, Australia, Canada, Malaysia, Taiwan, United Kingdom, USA, Korea, Brazil, Asia and Germany (<https://www.ncbi.nlm.nih.gov> **Accessed on 20/01/2019**).

Table 1. 1: Collection sites and Institutes for the genome sequencing of the specific Rhodophyta species

Sequenced Rhodophyta species	Collection site /Institute for sequencing	References
<i>Chondrus crispus</i>	Peggy’s Cove, Nova Scotia, Canada  University of Fort Hare <i>Together in Excellence</i>	Collén <i>et al.</i> , 2013
<i>Cyanidioschyzon merolae</i>	Italy, Naples university	Matsuzaki <i>et al.</i> , 2004
<i>Galdieria sulphuraria</i>	Mt. Lawu, Indonesia	Gross and Schnarrenberger, 1995
<i>Gracilaria chilensis</i>	Rongcheng, Shandong Province, China	Liu <i>et al.</i> , 2017
<i>Gracilariopsis chorda</i>	Jangheung, Jeonnam, Korea	Yang <i>et al.</i> , 2014
<i>Gracilariopsis lemaneiformis</i>	Zhanshan Bay, Qingdao, China	Zhou <i>et al.</i> , 2013
<i>Kappaphycus alvarezii</i>	Hainan, China	https://www.ncbi.nlm.nih.gov/nucleotide/NADL000000000.2 Accessed on 01/12/2018
<i>Porphyridium purpureum</i>	Eel Pond, Massachusetts	Bhattacharya <i>et al.</i> , 2013
<i>Porphyra umbilicalis</i>	Schoodic Point, Maine	Brawley <i>et al.</i> , 2017
<i>Pyropia pulchra</i>	University Herbarium, University of California at Berkeley (UC)	Lee <i>et al.</i> , 2016

It is evident that South Africa lacks both whole genome sequencing and organellar genome sequencing studies of its own economically and scientifically significant Rhodophyta species. As a result, this study serves as the pioneer study for the genomic composition of a South African Rhodophyta species collected from Shelly beach in the Kenton-On-Sea region, Eastern Cape.

1.5 Geographical distribution of marine Rhodophyta in South Africa

According to the Guide To Marine Life Of Southern Africa (Branch, 2017), the South African marine Rhodophytes are found all along the South African coastline which represents about 3650 km of coast (Maslo and Lockwood, 2014) and five major bioregions; Namaqua, South-Western Cape, Agulhas, Natal and Delagoa Bioregions shown in Figure 1.1 (Lombard, 2004). The Rhodophyta found along this coastline are represented by more than 500 species (Griffiths *et al.*, 2010), including species representing i) Flat Red Algae ii) Foliose Red Algae iii) Membranous Red Algae iv) Balloon and Tongue-like Red Algae v) Tough branching Red Algae vi) Gelatinous Red Algae (this group contain the *Gelidium* genus) vii) Ribboned and feathery Red Algae viii) Spiky and iridescent Red Algae ix) Branching Red Algae x) Elephant Red Algae xi) Epiphytic and fine Red Algae xii) Lightly calcified Red Algae xiii) Upright coralline algae and xiv) Encrusting algae (Branch, 2017).

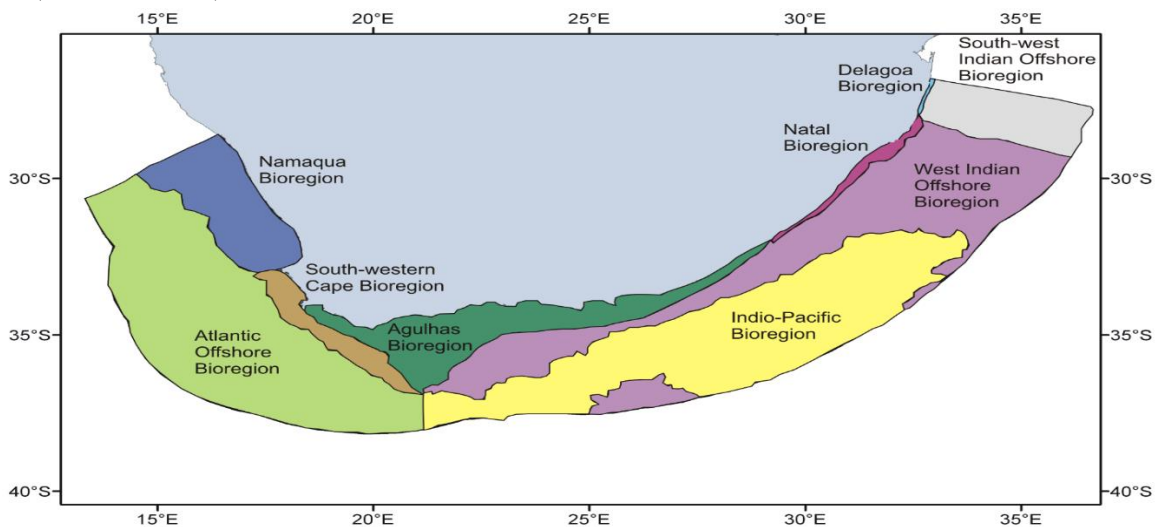


Figure 1. 1: Nine South African marine bioregions as defined by Lombard (2004). South African marine Rhodophytes are found in the coastal bioregions, Namaqua, South-Western Cape, Agulhas, Natal and Delagoa bioregions.

1.6 Current Rhodophyta studies in South Africa

Published studies of the South African Rhodophytes focused on biogeography and taxonomic studies achieved through morphological and DNA barcoding approaches. A study conducted by Kogame *et al.* (2017) has focused on DNA barcoding of geniculate Coralline red algae. There is one study that focused on molecular studies of galactan biosynthesis of *G. pristoides* (Hector, 2013). Some other studies of the South African Rhodophytes focused on the evaluation of the South African Rhodophytes for bioactive compounds with beneficial medicinal applications. A study evaluating the anti-cancer properties of the *Plocamium maxillosum* species was recently published (Knott *et al.*, 2019). The Phuhlisa Marine Research Group at the University of Fort Hare have focused on research studies which extract and characterise enzymes (such as serine proteases, Alcohol dehydrogenases) of medicinal and industrial applications from Kenton-On-Sea Rhodophytes including *G. pristoides*. Other members of our research group have focused on DNA barcoding while others focused on the characterisation of bioactive compounds with anti-diabetic and anti-cancer properties from *G. pristoides*.



1.7 The Choice of *Gelidium pristoides* for this study

University of Fort Hare
Together in Excellence

The choice of *G. pristoides* for this study is based on the following criteria. First, the economic value of *G. pristoides* serves as a motivation for its sequencing. Its genomic sequences can provide answers to the recently reported shortage of agar which has resulted in a drastic increase in price (Callaway, 2015). It can potentially pave a way forward to scientific practices including genome editing, and genetic modification to result in an increased yield of *Gelidium* which can be used for agar production, further boosting the South African economy.

Second, the scientific significance of Rhodophytes regarding pharmaceutical and medicinal applications was also a motivation for sequencing the *G. pristoides* genome. Also, the fact that the Phuhlisa Marine Research group based at the University of Fort Hare have shown interest in the utilization of *G. pristoides* for extraction of genes and other bioactive compounds further motivated the sequencing of *G. pristoides*. Third, this study is novel since the genome of *G. pristoides* has never been sequenced according to the history of genome sequencing projects. This study will serve as the pioneer genomic study for *G. pristoides*, which will provide a genomic reference for

future research projects. Fourth, the abundance of *G. pristoides* in the Kenton-On-Sea region served as a motivation for its genome sequencing to allow further molecular biology and genetic research studies.

1.8 *Gelidium pristoides*

The *Gelidium* species are known for their high standard agar used widely in microbiology, public health facilities and clinical laboratories, for the growth of microbes and used as an incorporative ingredient in different reagents (Callaway, 2015). Consequently, the species of this genus has been harvested from the marine intertidal zones of the Eastern Cape shores since the 1950s. The *Gelidium pristoides* (*G.pristoides*) species (Figure 1.2) has contributed about 80% of this *Gelidium* harvest (Anderson *et al.*, 1991; Lubke and De Moor, 1998) and it has been recently reported as a source of polysaccharides with anti-tumour properties (Hector, 2014 <https://www.all4women.co.za/168665/health/south-african-red-algae-gene-discovery-to-help-treat-tumours> **Accessed on 20/10/2018**). *G. pristoides* is defined as a South African endemic Rhodophyta species that inhabit limpet shells of the intertidal zone in marine environments (Anderson and Rothman, 2013). It is currently classified under *Gelidiales* in the *Florideophyceae* class which is known as the most taxonomically diverse and successful marine-striving Rhodophyta class (Seckbach and Chapman, 2010). It is distributed from the west coast (Sea Point) to the east coast (Port Edward) (Day, 1969) and is abundant in the Eastern Cape (Seagrief, 1967; Simons, 1977).



Figure 1. 2: Photograph of *Gelidium pristoides* (Turner) Kützing. The photograph was taken at the Department of Biochemistry and Microbiology, University of Fort Hare.

1.9 Genome types



University of Fort Hare

The term Genome defined as the complete set of an organism's hereditary information and encoded as either deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) in many viruses, contains all the information that is necessary for the functioning of an organism (Hieter and Boguski, 1997). It is made of both coding and non-coding regions and remains the ultimate determinant of a cell or the organism itself (Joshua and Alexa, 2001). The three main different types of genomes are nuclear, mitochondrial and the plastid genome (Jennings, 2017). Of these genomes, the mitochondrial and the plastid genomes, which are located in the organelles of endosymbiotic origin, are smaller and are usually present in high copy numbers (Hirakawa and Ishida, 2014; Titilade and Olalekan, 2015).

The nuclear genome [Figure 1.3(a)] is the largest and bears almost all functional genes, including some genes responsible for mitochondrial and plastid proteins (Dey and Harborne, 1997; Day, 2004). This eukaryotic linear genome is located in the nucleus, divided among several chromosomes that vary widely across species (Dey and Harborne, 1997).

The mitochondrial genome, also referred to as mitochondrial DNA (mtDNA) [Figure 1.3(b)], is found in the matrix of mitochondria, a double membraned-organelle responsible for energy supply in eukaryotic cells (Henze and Martin, 2003). The size of this genome varies from species to species and is mostly characterised by a compact structure except for the large plant mitochondrial genomes (Gagliard and Gaulberto, 2004). Additionally, mitochondrial genome structures also can vary according to divisions of life. This phenomenon is observed in mitochondrial genomes of some protists as well as in green algae, which possess linear mitochondrial genomes, while other divisions of life, including plants (and Rhodophytes) and animals, possess circular mitochondrial genomes (Gray *et al.*, 1998; Nosek *et al.*, 1998; Burger *et al.*, 2003).

The plastids, which are generally defined as eukaryotic double-membrane bound cytoplasmic organelles responsible for synthesis and storage of cell food, bear their own genome, commonly known as the plastome [Figure 1.3 (c)], which also varies regarding size, architecture and gene content between organisms (Maier and Schmitz-Linneweber, 2004; Chan and Bhattacharya, 2010). All these types of genomes play critical roles in an organism's well-being, development, and adaptation, and the following discussion will examine the significance of the genome sequences, particularly in biology.



University of Fort Hare
Together in Excellence

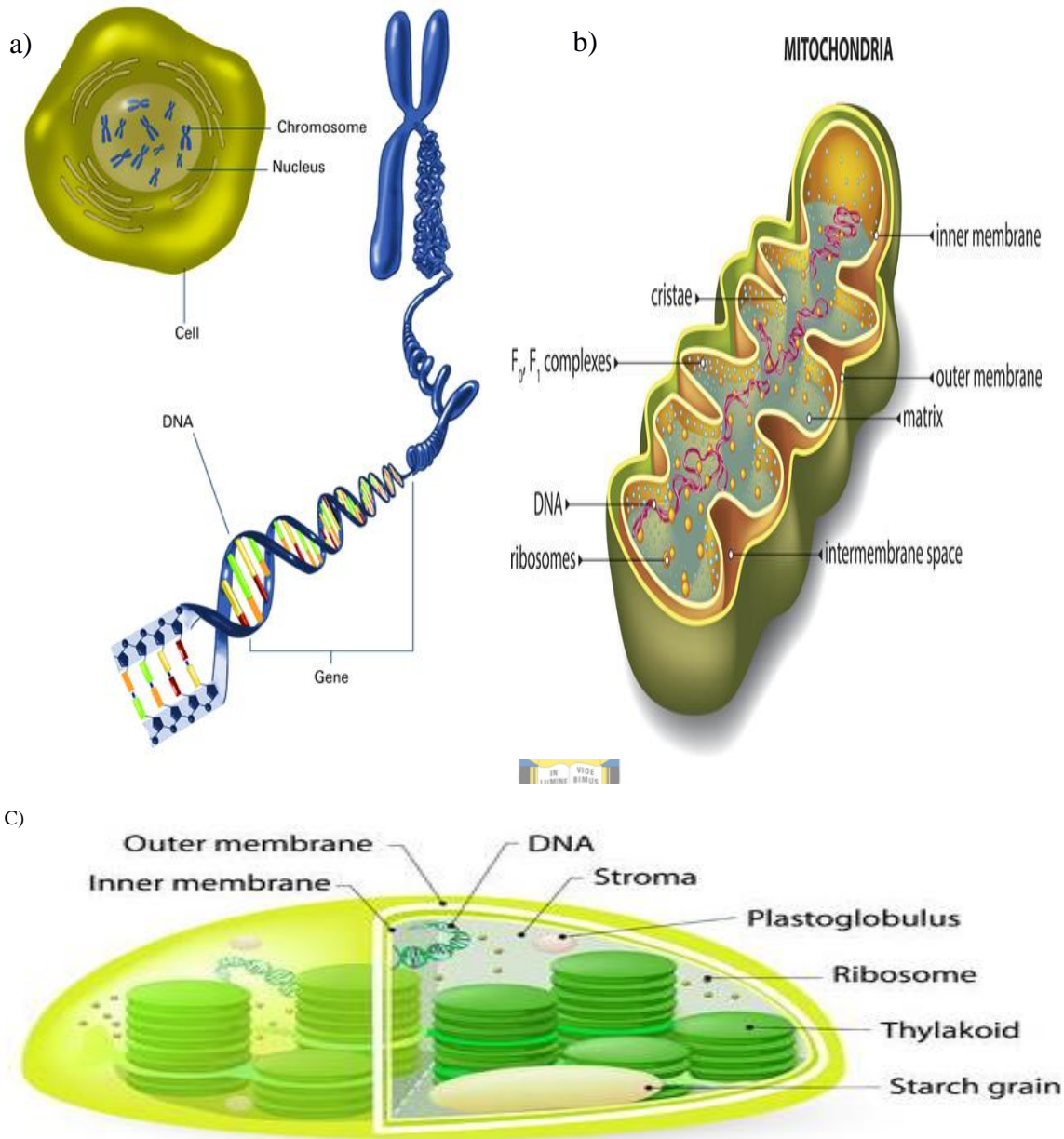


Figure 1. 3: Illustration of the location of the different types of genomes. (a) The nuclear genome from the nucleus; (b) The mitochondrial genome in the mitochondria; (c) The plastid genome located in the chloroplast. Figures adapted from https://www.google.co.za/search?biw=1366&bih=613&tbm=isch&sa=1&ei=8CXnWuewDMPo_gAb336KwBw&q=nuclear+DNA&oq=nuclear+DNA&gs_l=psy-ab.3..017j0i5i30k1j0i24k112.126587.133939.0.134920.12.9.0.3.3.0.630.2050.2-2j1j1j1.5.0...0...1c.1.64.psy-ab..4.8.2079.0..0i67k1.0.EiXp92bjFlg#imgrc=z0MvgHP0f2_tWM; <https://sciencestruck.com/how-is-mitochondrial-dna-used-in-forensics;> https://www.123rf.com/photo_46959995_stock-vector-chloroplast-structure-of-a-typical-higher-plant-chloroplast.html respectively). (Accessed on 25/ 11/2018)

1.10 Significance of genome sequences

As a genome is all of an organism's genetic material, their sequences are known to be an archive for all genetic information (Patnaik *et al.*, 2012) and enable researchers to understand the composition of a genome, regulation of gene expression and also provide information on how the whole genome works. In all, it provides a way of understanding the functioning of genes to direct the growth, development, and maintenance of an organism (Marzillier, 2008). Russell *et al.* (2016) also emphasised that having the complete sequence of a genome enables scientists to study the complete set of genes in organisms, enabling understanding of structures and functions of all the genes which constitute a genome (Russell, 2016). Availability of genomic information allows for direct isolation, cloning of genes (Alberts *et al.*, 2002) and design of DNA agents for molecular-based research studies. For example, plastid genome information can be used for highly efficient plastid transformation vectors, as gene sequences and regulatory elements are the central information for the design of efficient plastid transformation strategies (Rogalski *et al.*, 2015).

Genome sequences play a crucial role in finding a correlation between how the genome relates to the development of diseases (identification of pathogenic genes), susceptibility to certain diseases and drug metabolism and the discovery of agents (such as genes, and proteins) of both economic and pharmaceutical importance (Marzillier, 2008). It enables breeders to identify molecular markers for desired characteristics (Kim *et al.*, 2015; Ramakrishnan *et al.*, 2015) and to better understand the effects of potential modulations of synthetic pathways, including genetic modifications (Holst-Jensen *et al.*, 2016).

Comparative and evolutionary studies also benefit from genomic studies, as they reveal information regarding genomic rearrangements for species' survival in multiple environmental conditions (Liu *et al.*, 2012). For example, genomic sequences serve as a reference for the investigation of other members of the same species or closely related members. A good example of the evolutionary and taxonomic application of genomic sequences is the use of mitochondrial and plastid genomic data (genes, non-coding regions, Restriction Fragment Length Polymorphism (RFLP) and Simple sequence repeats or Short Tandem Repeats (SSR) markers) in the determination of the evolutionary distance from different taxonomy levels (Rogalski *et al.*, 2015). Mitochondrial genomic sequences of Rhodophytes have been used in both systematic and

population studies as molecular markers (Yang *et al.*, 2015). Genomic information plays a significant role in the conservation of biodiversity by providing hints regarding how a species becomes endangered. For example, genomic sequences can provide information regarding deleterious mutations in the genes responsible for brain function, metabolism, immunity and so on. In a nutshell, genomic sequences enable researchers to explore the population structure, genetic variations, and recent demographic events in threatened species (Khan *et al.*, 2016).

1.11 Genome sequencing

Genome sequencing refers to the process of determining or figuring out the exact arrangement of nucleotide bases, namely Cytosine (C), Thymine (T), Adenine (A) and Guanine (G) which makes up DNA molecules (Angeleska *et al.*, 2014). To sequence a genome generally means uncovering the nucleotide base sequences for all the chromosomes constituting the species of interest, providing its genomic map (Ekblom and Wolf, 2014). The process through which the full genome of an organism is sequenced is referred to as whole genome sequencing, or complete sequencing, and involves sequencing of all the types of genomes an organism has (Angeleska *et al.*, 2014; Jonoska and Saito, 2014). Genome sequencing is therefore regarded as a valuable technique as it provides vast amounts of data which assist in the characterisation of biological systems, as well as in the development of major projects (Griffin and Griffin, 1993).

1.12 Genome sequencing approaches

1.12.1 The Hierarchical shotgun sequencing and whole genome shotgun (WGS) sequencing

Hierarchical shotgun sequencing and the whole genome shotgun (WGS) sequencing, shown in Figure 1.4, form the two basic approaches of genome sequencing. Historically, the hierarchical shotgun sequencing, also known as Clone-by-Clone or BAC-to-BAC sequencing, was the first sequencing approach used and was much preferred by the international Human Genome Project Organization (HGPO) during human genome sequencing, which employed both approaches (Trivedi, 2000; Sharma, 2008). Hierarchical shotgun sequencing is based on the principle of first taking genomic DNA (gDNA) isolated from an organism and then fragmenting it into larger fragments of about 150 kb, which are then amplified and arranged through a tilling path for

generation of a physical map. Every fragment after making the map is further fragmented into smaller pieces which are then clonally amplified (through Polymerase Chain reaction [PCR]), sequenced, and assembled (Sharma, 2008). This approach is widely used for sequencing large, complex genomes such as those of human, sugarcane, bread wheat since their genomes are difficult to sequence through the WGS approach (Visendi *et al.*, 2016). Although the traditional Hierarchical shotgun sequencing can produce high-quality assembly, it remains a costly and time-consuming approach, involving making BAC libraries, fingerprinting them, as well as sequencing large numbers of overlapping clones (Chandler and Brendel, 2002; Visendi *et al.*, 2016).

In contrast, the widely preferred WGS sequencing approach, which was invented by J. Craig Venter in 1996 (Trivedi, 2000), is based on the same principle but skips the time-consuming map-generation step employed in hierarchical shotgun sequencing (Sharma, 2008). This means that the gDNA is directly fragmented into shorter fragments which are amplified and subjected to sequencing through sequencing technologies. This strategy is used for sequencing both prokaryotic and eukaryotic genomes and was first used by Fleischman and colleagues (1995) for sequencing *Haemophilus influenza* (*H. influenza*) Rd, and then used in sequencing of other microbial organisms including *Mycoplasma genitalium*, *Saccharomyces cerevisiae* and *Escherichia coli* and later for sequencing the *Drosophila* genome (Sharma, 2008). The journey for this approach did not end with sequencing the species mentioned but has expanded across different divisions of life, and it is still widely used today.



University of Fort Hare
Together in Excellence

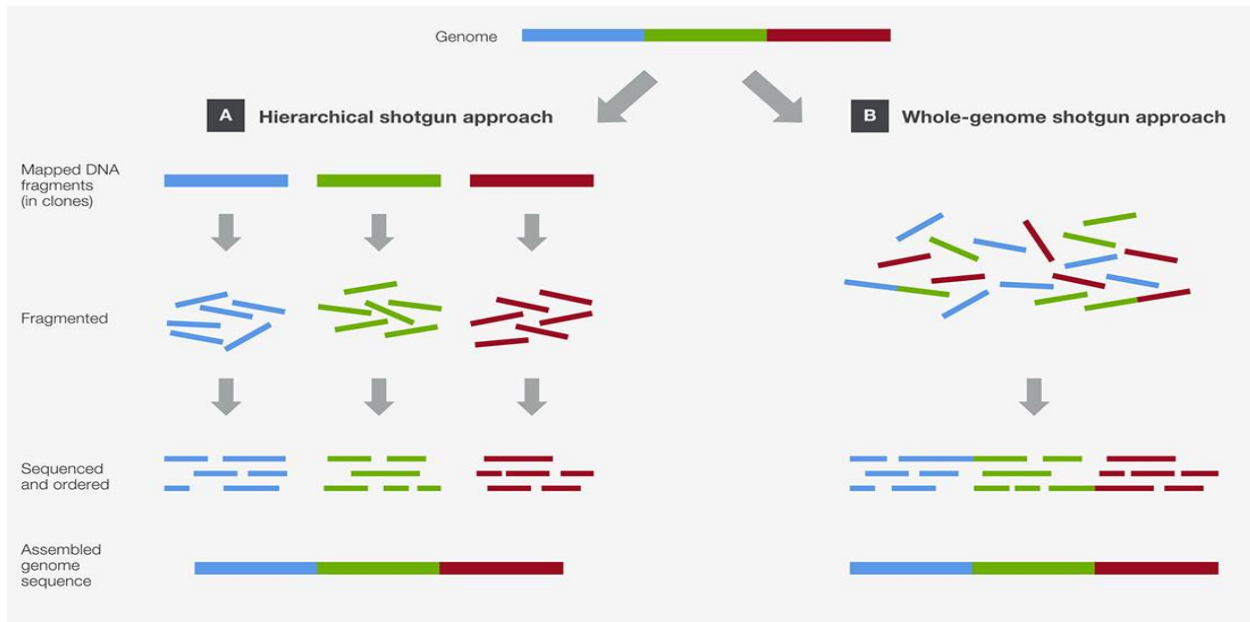


Figure 1. 4: Schematic representation of the Genome sequencing approaches. A. Hierarchical shotgun sequencing. B. The whole genome shotgun sequencing approaches. Figure adapted from https://www.google.co.za/search?q=Hierarchical+shotgun+sequencing.+B:+the+whole+genome+shotgun+sequencing+strategies&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiO__CAgorXAhUkKMAKHRjNCUsQ_AUICigB&biw=1366&bih=662#imgsrc=twDzZMAdWj3TM (Accessed on 25/ 11 2018).



University of Fort Hare
Together in Excellence

1.12.2 Sanger sequencing

Before the mid-1970s, there were no existing methods by which DNA could be directly sequenced until Fred Sanger and colleagues (1977) developed the first-generation sequencing method, known as the Sanger method. This method is also known as a dideoxy method and is based on utilisation of chain-terminating deoxynucleotide analogues called dideoxynucleotides (ddNTPs) for base-specific termination of primed DNA synthesis. During Sanger sequencing, the target DNA is denatured into single strands, used as a template strand to which the primer is annealed for DNA synthesis. The traditional Sanger sequencing technology requires four reaction tubes in which the DNA template, DNA polymerase, primer, all four normal dNTPs and one ddNTP are added. During the DNA synthesis, incorporation of the ddNTPs terminates incorporation of normal dNTPs due to their lack of the 3' hydroxyl group (OH), which allows for phosphodiester bond formation. Chain termination results in fragments of varying sizes that are visualised in agarose gel where small fragments migrate faster than the larger ones, allowing determination of the

sequence of the target DNA molecule (Sanger *et al.*, 1977; Obenrader, 2003). According to the history of genome sequencing, the Sanger sequencing method was developed to sequence the first full genome of bacteriophage ϕ X174 in 1977 (Sanger *et al.*, 1997). The Sanger method was then automated, and only one reaction tube was required. This automation reduced not only the sequencing costs of traditional Sanger sequencing but also provided a relatively improved sequencing speed that enabled sequencing of the first human genome in the Human Genome Project (HGP) to be completed (<https://www.healio.com/hematology-oncology/learn-genomics/whole-genome-sequencing/whole-genome-sequencing-methods> accessed 25/ 11 2018).

1.12.3 The Next Generation Sequencing

The Next Generation Sequencing (NGS) technologies refer to the high-throughput or massively parallel sequencing methods, which were developed in 2005 (Egholm *et al.*, 2005; Kim *et al.*, 2014). Published literature confirms that the NGS sequencing technologies are all based on the principles of the shotgun approach explained in this chapter. However, the sequencing chemistry, throughput, sequencing duration, read length, error rate, and sequencing costs are different between the various NGS sequencing platforms (Choudhry and Wu, 2013; Türктаş *et al.*, 2015). The literature further states that these differences influence the choice of sequencing technology and its platform. The most important advantage of NGS technologies is their positive impact on both sequencing duration and sequencing costs. More specifically, NGS technologies have significantly reduced the sequencing time and costs of sequencing the human genome (https://www.google.co.za/search?q=nhgri+sequencing+costs&source=lnms&tbm=isch&sa=X&ved=0ahUKEwj78ZDlq9TbAhWCHsAKHYjLCH8Q_AUICigB&biw=1242&bih=557#imgrc=nw_L_WMKdoBAeM:accessed 21/09/2018). The NGS sequencing technologies offer a bright future of an even further reduced sequencing cost (Figure 1.5), making possible personalised sequencing for individualised treatment of diseases, as it is predicted that the sequencing of the human genome will be reduced to only \$1 sometime between 2020 and 2025 (National Human Genome Research Institute [NHGRI]). Moreover, highly reduced sequencing costs will progressively produce more genome sequences from non-model organisms. Examples of NGS sequencing technologies include Roche 454, Illumina Solexa and Sequencing by Oligonucleotide Ligation and Detection (SOLiD), PacBio, Ion Torrent (Verma *et al.*, 2013).

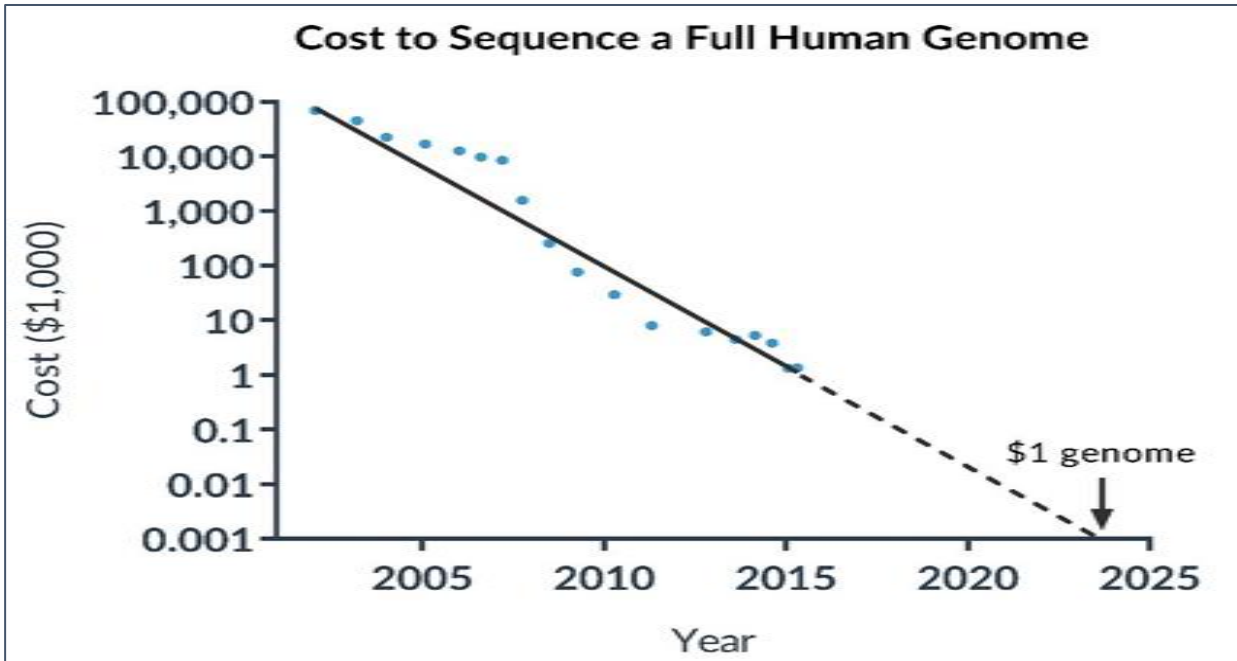
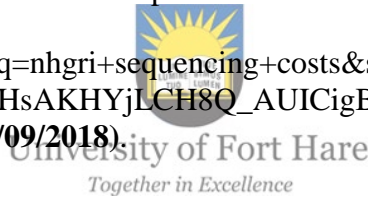


Figure 1. 5: Graphical illustration of the reduction in sequencing costs since the introduction of NGS platforms from 2005 to 2015 and future predictions of the sequencing costs. Figure adapted from

https://www.google.co.za/search?q=nhgri+sequencing+costs&source=lnms&tbn=isch&sa=X&ved=0ahUKEwj78ZDlq9TbAhWCHsAKHYjLCH8Q_AUICigB&biw=1242&bih=557#imgrc=hiCEPEHLhTPH5M: (Accessed 21/09/2018).



1.13 The choice of sequencing technology

Ion Torrent was chosen for this study based on its availability, and the fact that it is a more affordable platform with a reasonable read length of up to 600 bp and uses good quality assurance protocols. Simplicity, accuracy and the faster sequencing rate offered by Ion Torrent platforms (ThermoFisher Scientific) also contributed to the choice of this platform for the study at hand. The application history of Ion Torrent platforms in genome sequencing includes many plant genomes, which also contributed to selecting the Ion Torrent platform for this study. The complete genomes of *Sedum album*, commonly known as white stonecrop, and *Hordeum vulgare* (Barley) (Mascher *et al.*, 2013), are examples of plant genomes that were sequenced using the Ion Torrent platform, in particular the Ion Personal Genome Machine (PGM) (Al-Khayri *et al.*, 2015).

1.13.1 The Principle of Ion Torrent

Ion Torrent (Semiconductor) sequencing was introduced into the market in 2010 and works by detecting hydrogen ions (protons) released as nucleotides are incorporated by DNA polymerase during DNA synthesis (Rothberg *et al.*, 2011). As shown in Figure 1.6, the Ion Torrent adopts the shotgun sequencing approach as an initial step of library preparation, in which the extracted DNA is first fragmented with either enzymatic or mechanical shearing. Before ligation of adaptors serving as primer-binding sites, the DNA fragments are end-repaired to form blunt ends which allow adaptor ligation, followed by size selection achieved with either agarose gel or paramagnetic beads (Knierim *et al.*, 2011). The prepared library is clonally amplified through Polymerase Chain Reaction (emulsification Polymerase Chain Reaction (emPCR) and isothermal PCR) conducted on ion sphere particles (beads), and the template-bearing beads are selected through a magnetic-based enrichment process (Rothberg *et al.*, 2011; Quail *et al.*, 2012). The primers and DNA polymerase necessary for the sequencing process are bound to the template-bearing beads before they are inserted into a sequencing chip that is then installed into a sequencing platform where each of the four nucleotides is introduced sequentially. On incorporation of each nucleotide base during DNA synthesis, hydrogen ions are released, which induce a drop in pH directly measured by the ion-sensitive layer (Rothberg *et al.*, 2011; Verma and Singh, 2013; Reuter *et al.*, 2015). A signal-processing software changes the raw voltages produced during incorporation of each base into base calls. Low-quality reads are excluded, as reads are passed through two-signal based filters (Rothberg *et al.*, 2011).

Ion Torrent sequencing is fast as it does not involve the time-consuming optics (Ari and Arikan, 2016). Its sequencing chemistry is suitable for various applications, including *de novo* sequencing of any whole genome, targeted DNA sequencing, targeted RNA sequencing, exome sequencing, whole transcriptome sequencing, viral typing, bacterial typing, aneuploidy and Copy Number Variation (CNV) analysis, small RNA and miRNA sequencing and chromatin immunoprecipitation sequencing (<https://www.thermofisher.com/za/en/home/life-science/sequencing/next-generation-sequencing/ion-Torrent-next-generation-sequencing-applications.html> accessed on 21/ 09/2018).

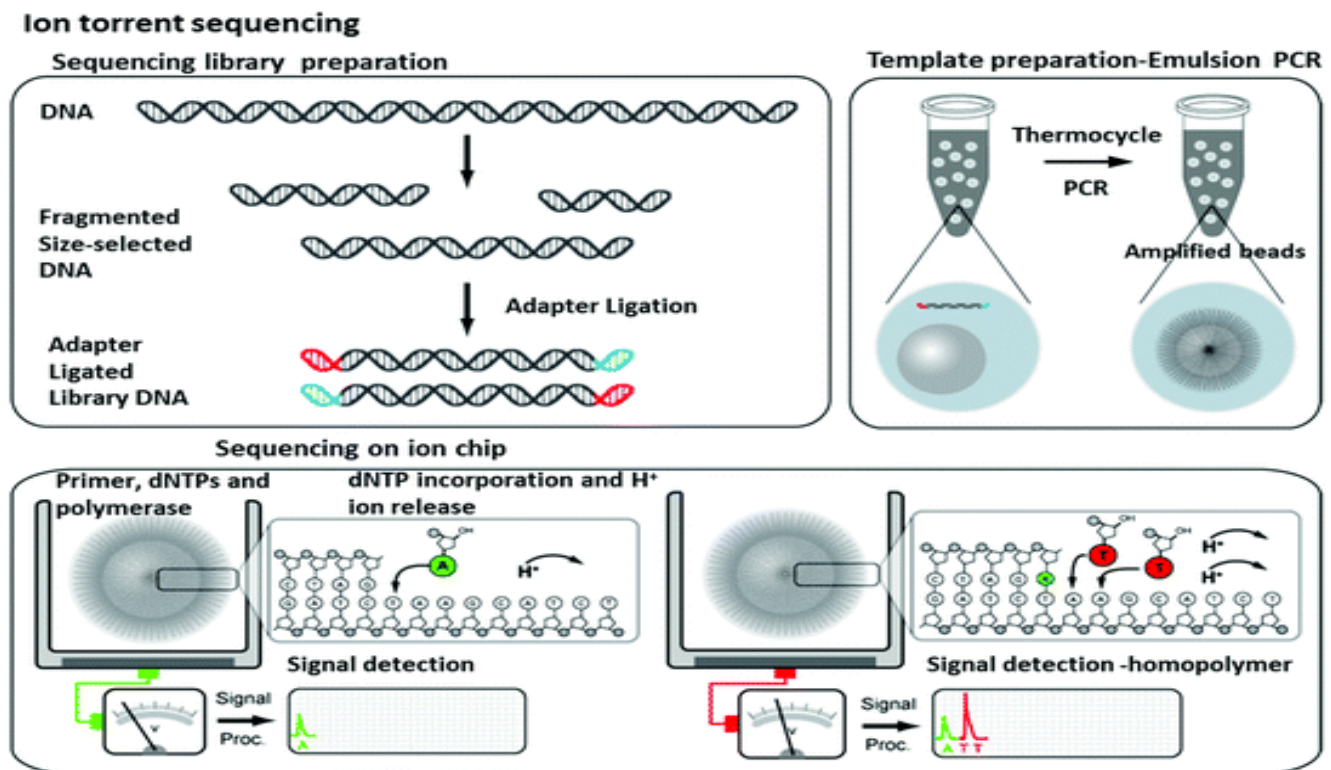


Figure 1. 6: The general workflow and sequencing chemistry of Ion Torrent sequencing. Figure adapted from https://www.google.co.za/search?q=ion+Torrent+principle&rlz=1C1GNAM_enZA679ZA679&source=lnms&tbm=isch&sa=X&ved=0ahUKEwifmK2F5JzbAhUHAsAKHeRTDpc_Q_AUICigB&biw=1280&bih=869#imgdii=7CaL2aZJTW6N2M:&imgcr=Fhdo40aEphKfYM (Accessed on 21/09/2018).

The simple principle of Ion Torrent sequencing has allowed improvements in both the production of sequencing chips and many sequenced genomes, as well as in sequencing accuracy. The ThermoFisher Scientific webpage confirms the availability of Ion PGM, Ion Proton System, Ion S5 and Ion S5 XL Systems and other platforms (<https://www.thermofisher.com/za/en/home/brands/ion-Torrent.html>, Accessed on 22/09/2018). However, although the Ion Torrent sequencing technology is a cheap NGS sequencing platform, it becomes costly to try all the different Ion Torrent sequencing platforms for one project. Hence, the Ion S5 sequencing platform was chosen for sequencing the genome of the *Gelidium pristoides* (*G. pristoides*) from the Kenton-On-Sea region.

1.13.2 The Ion S5 sequencing platform

Thermo Fisher Scientific introduced the Ion S5 sequencing platform in September 2015 as the latest Ion Torrent innovation (<http://www.bio-itworld.com/2015/9/1/thermo-fisher-clarifies-vision-sequencing-release-ion-s5-instruments.html>, accessed 22/11/2018). It comes with three sequencing chips, including the Ion 520 chip, Ion 530 chip and the Ion 540 chip, which produce about 3 to 5, 15–20 and 60 to 80 million reads per run respectively, with read lengths ranging from 200–600 bp and a maximum output of between 10–100 Gb (<https://www.thermofisher.com/za/en/home/life-science/sequencing/next-generation-sequencing/ion-Torrent-next-generation-sequencing-workflow/ion-Torrent-next-generation-sequencing-run-sequence/ion-s5-ngs-targeted-sequencing/ion-s5-specifications.html>, accessed 22/11/2018). Internal validation of the Ion S5 sequencing chips has resulted in a raw read accuracy of 99.1%, 99.2% and 98.4% for the 520, 530, and 540 Chips respectively (Felton, 2015). The Ion 530 chip was used for this study based on the report of its raw read accuracy by Felton (2015). As with other Ion Torrent platforms, applications of the Ion S5 sequencing platform include whole genome sequencing, exome sequencing, RNA-sequencing, mRNA-sequencing, targeted sequencing, Amplicon sequencing, Chip-sequencing, metagenomics, microbial sequencing, mitochondrial sequencing, epigenetic sequencing, single nucleotide polymorphism (SNP) Genotyping and variant detection (<https://www.thermofisher.com/za/en/home/life-science/sequencing/dna-sequencing.html>, accessed 26/11/2018).

1.14 Genome assembly

As the shotgun sequencing approach employed by sequencing platforms involves fragmentation of DNA into fragments of desired sizes, it presents the challenging task of merging the reads into contigs and scaffolds for generation of the overall genomic sequence, through a process generally known as genome assembly. Genome assembly is made possible by ‘*de novo* genome assembly’ and reference-based genome assembly, as illustrated in Figure 1.7 (Pop, 2009; Miller *et al.*, 2010; Alkan *et al.*, 2010; Chiu, 2015).

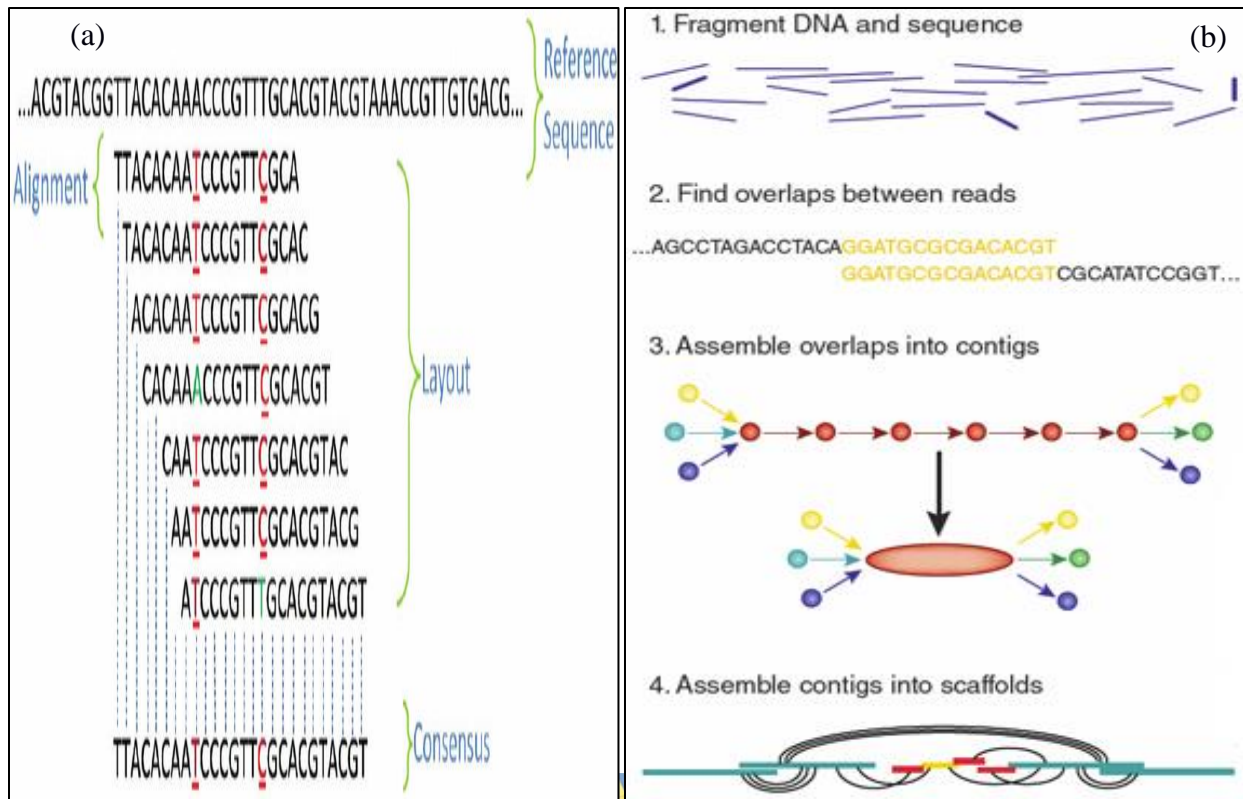


Figure 1. 7: Illustration of the genome assembly approaches. (a) Reference-based genome assembly, (b) *De novo* genome assembly. *Figures adapted from (Baker, 2012):*

https://www.google.com/search?tbm=isch&q=Referencebased+genome+assembly&chips=q:reference+based+genome+assembly,online_chips:alignment&sa=X&ved=0ahUKEWimfrXqsLdAhUHAcAKHTfdACoQ4IYILygI&biw=1280&bih=913&dpr=1#imgrc=5PUPCRJI9acNEM
(Accessed on 26/11/2018).

1.14.1 Reference-based assembly

Reference-based genome assembly involves sequencing and assembling a genome of an organism whose reference genome has been previously established. This approach makes use of the related genomes made available by improvements in genome sequencing platforms for easing and facilitating the assembly process of a similar genome (Pop *et al.*, 2004; Ng and Kirkness, 2010) through reference-based assemblers. These assemblers build a consensus sequence by mapping every sequenced read against a genome of known sequence (Ng and Kirkness, 2010), as illustrated in Figure 1.7a. Examples of such assemblers include Newbler, AMOS and MIRA (Chevreux *et al.*, 1999; Chevreux *et al.*, 2004 Wong, 2016). Although these assemblers are not memory

expensive and tend to facilitate the genome assembly process, novel regions of the newly sequenced genome need to be assembled with the *de novo* approach, as they do not have a reference sequence (Ng and Kirkness, 2010).

1.14.2 De novo genome assembly

De novo genome assembly refers to assembling a genome of an organism which has never been sequenced before. Using this approach, the sequenced reads are merged using *de novo* assemblers which detect overlapping regions of the reads to generate an overall genome sequence (Figure 1.7b) for the particular organism of interest. The consensus sequence generated through these assemblers serves as a reference point for re-sequencing and other genomic studies (Chiu, 2015).

De novo assembly is widely achieved through utilisation of the overlap and k-mer-based approaches (Constantinescu, 2015). The de novo assemblers based on the overlap approach are the Overlap layout Consensus (OLC) and Greedy assemblers, while the k-mer-based approach is used in the De Bruijn graph (DBG) assemblers. The overlap-based assemblers rely on different ways of measuring the similarity between reads for detecting overlaps. “Levenstein distance, defined as the minimum number of insertions, deletions or substitutions required to transform one string, or read, into another” (Levenshtein, 1966) forms one of the ways of measuring similarity. The Levenstein distance is inversely proportional to the similarity between reads (Constantinescu, 2015). The overlap-based assemblers also use the overlap score as defined by Lander and Waterman for detecting similarity between overlapping regions of reads (Lander and Waterman, 1988). “This overlap score is defined as the number of similar base pairs between similar regions (prefix or suffix areas) of two reads and are required to be above some minimum overlap threshold value defined as

- T = Minimum amount of base pairs needed to detect an overlap
- $\phi_{\min} = \frac{T}{L}$ = Minimum read fraction needed to detect an overlap” (Constantinescu, 2015).

1.14.3 Greedy assemblers

The name given to these assemblers implies that their goal is achieved by application of greedy rules as they rely on algorithms that select the best overlapping reads for construction of contigs.

Two reads or fragments are said to be overlapping only if the tail of one read is aligned to the head of another read (Kasahara and Morishita, 2006) thus priority during contig construction is given to reads with better quality and best overlapping reads (Jonoska and Saito, 2014). The ‘greedy’ approach adopted by these algorithms, instructs the greedy assemblers to add one or more reads or contigs with the largest overlap at each step of the process (Angelini *et al.*, 2016). In achieving the basic operation rule of greedy assemblers, some greedy algorithms begin by iteratively merging reads with optimal overlaps to form multiple contigs, while others tend to extend a given read at the 5’ and 3’ ends into a contig by attaching a read that best overlaps with the preceding read. The primary advantage of these assemblers is that they apply in both the genome assembly of Sanger reads and in assembling NGS reads, as they were first used in Sanger assemblers including TIGR and CAP3, and in NGS assemblers such as SSAKE, VCAKE and SHARCGS (Jonoska and Saito, 2014). However, their major drawback is that they are most suitable for assembling relatively small genomes, as they are very slow and require a lot of computer memory (Angelini *et al.*, 2016).

1.14.4 The OLC Assemblers



The name of this intuitive approach suggests the three-step strategy it adopts. For example, overlap (O) refers to the detection of overlapping fragments whereas, layout (L) refers to graph construction based on the overlapping fragments followed by consensus (C), referring to determination of the consensus sequence based the constructed graph (Li *et al.*, 2011; Rodríguez-Ezpeleta *et al.*, 2011; Wang and Kole, 2015). During the graph construction phase, the OLC assemblers first create alignments called all-against-all pairwise alignments that are used for building a graph which represents reads as nodes and overlaps between reads as edges. As illustrated in Figure 1.8, the constructed graph is used for determination of the consensus sequence by following a Hamiltonian path (Angelini *et al.*, 2016), a path that visits each vertex once. The consensus sequence is a result of choosing which nucleotide is represented by the majority of overlapping reads for every sequence position. Assemblers such as ARACHNE, PHRAP, CAP, TIGR, CELERA and Newbler all designed based on this approach, are optimised for assembling large genomes and relatively short reads of 100 bp (Rodríguez-Ezpeleta *et al.*, 2011).

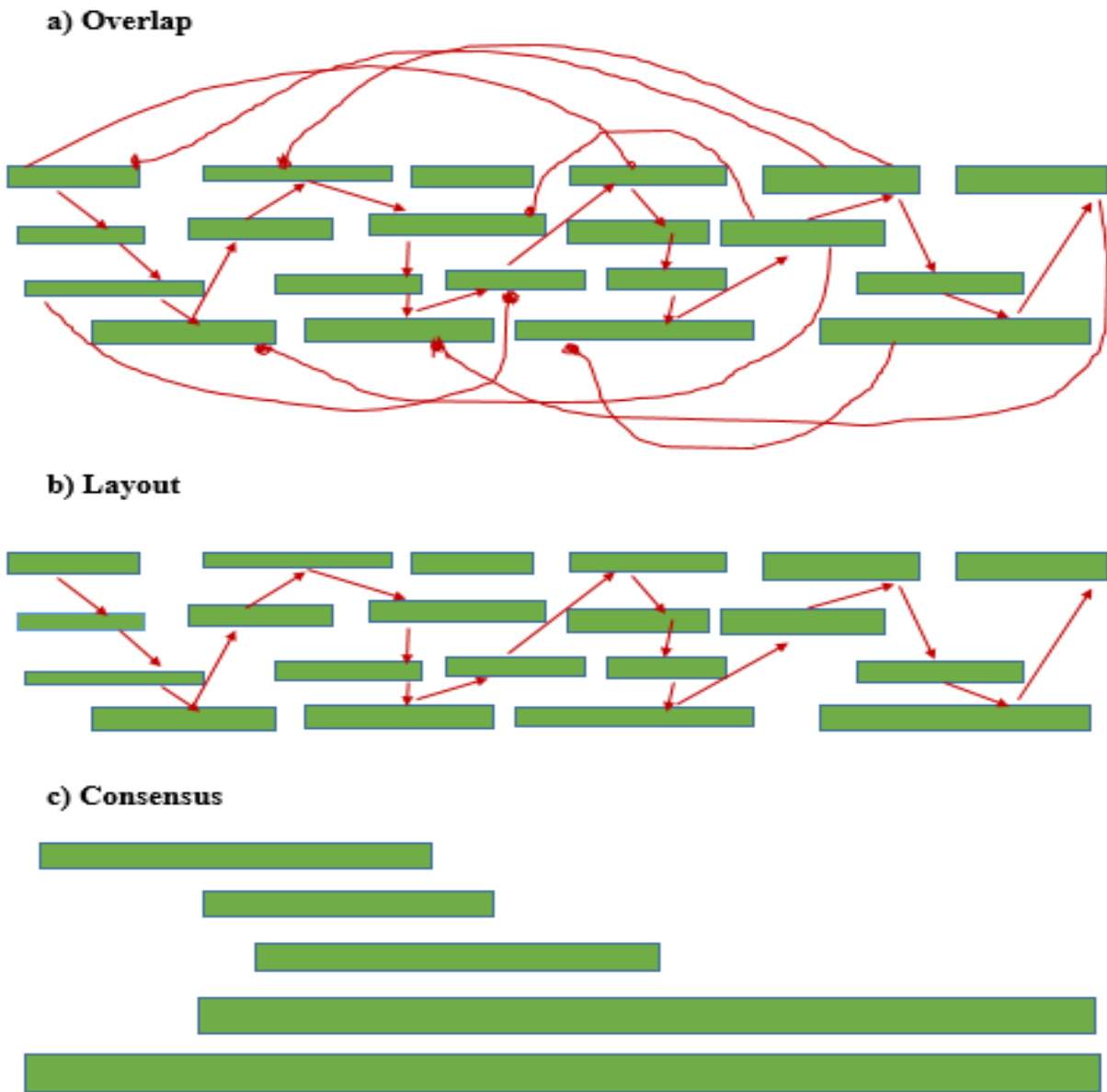


Figure 1. 8: The general strategy adopted by OLC-based *de novo* assemblers. a). **Overlap:** The OLC-based algorithm finds overlapping reads by aligning the sequenced reads. b). **Layout:** Built based on which reads align to which read. c). **Consensus:** A consensus sequence is built by merging the sequenced overlapping reads. The figure was redrawn from https://www.google.com/search?q=The+OLC+Assemblers&source=lnms&tbn=isch&sa=X&ved=0ahUKEwjnv6KjpcLdAhXLLcAKHcpEAtsQ_AUIDygC&biw=1280&bih=913#imgrc=CjkkkpChRZX6nM (Accessed 26/11/2018).

1.14.5 The DBG assemblers

The DBG algorithms, developed in 1995, were first designed for assembling small genomes such as bacterial genomes, but advancements made in genome sequencing which allow for sequencing of large genomes such as plants and animal genomes were a driving force for the improvement of these assemblers to accommodate even large genomes (Idury and Waterman, 1995; Li *et al.*, 2011; Wang and Kole, 2015). For example, the CABOG assembler, initially developed for assembling Sanger reads, was modified to accommodate both 454 Pyrosequencing and Illumina (Solexa) reads, while the Masurca assembler can generate assemblies of Illumina (Solexa), 454 Pyrosequencing, and Sanger reads (Angelini *et al.*, 2016). The DBG assemblers came into full usage in 2005 during the introduction of the Illumina (Solexa) sequencing platforms. The strategy adopted by DBG assemblers such involves chopping of reads into even shorter sequences called k-mers, for removal of possible sequencing errors before construction of a DBG. The DBG is then used for the construction of the overall genome sequence of an organism (Li *et al.*, 2012; Wang and Kole, 2015). The k-mers are used as edges while the overlaps between them serve as nodes during the DBG construction. Rather than using the Hamiltonian path followed by OLC assemblers, these assemblers tend to follow a Eulerian path for finding the consensus sequence, as illustrated in Figure 1.9 (Angelini *et al.*, 2016). Assemblers including, Euler- SR, Velvet, ABySS, ALLPATH- LG, SOAPdenovo and SPAdes are examples of DBG-based assemblers (Elloumi, 2017).

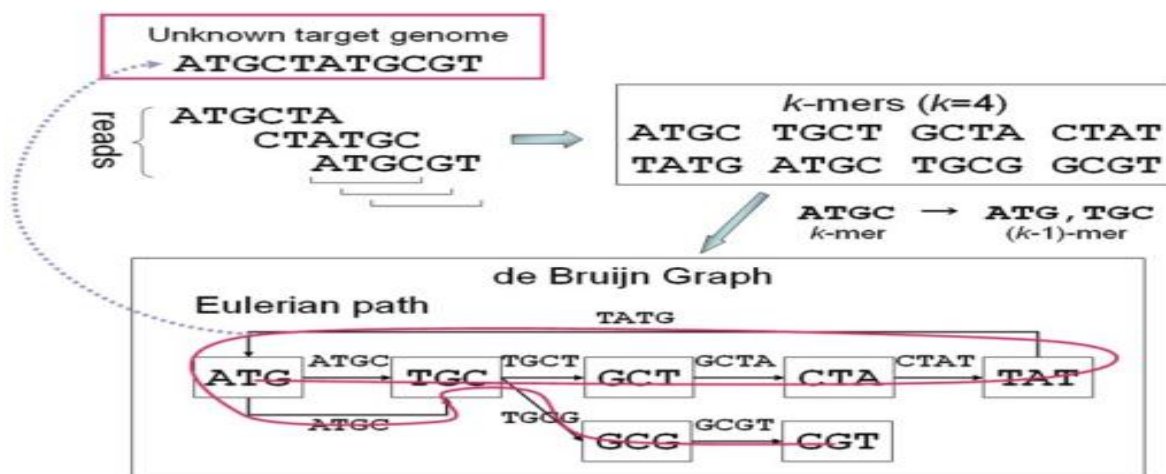


Figure 1. 9: The general strategy adopted by DBG-based *de novo* assemblers. Figure adapted from Namiki *et al.* (2012).

1.15 Genome annotation

Genome annotation is a critical process which happens immediately after the genome assembly process (DeSalle and Rosenfeld, 2012). Genome annotation analyses the newly assembled genome and assigns its biological functions (Stein, 2001). Precisely, genome annotation provides a platform for the scientific community to understand the genomic data. As an example, biologists presented with a large FASTA file of unannotated genomic sequence would find the data meaningless; they might ask themselves in this situation, ‘So what now?’ ‘Where’ and ‘How’? To answer these questions, the raw genomic sequence must be annotated for its constituents (structural annotation) and their relative functions (functional annotation) through nucleotide, protein, and process-level annotation, as shown in Figure 1.10 (Stein, 2001). Genome annotation is divided into a structural section which answers the ‘Where’ questions, and functional annotation which answers the ‘What’ and ‘How’ questions.

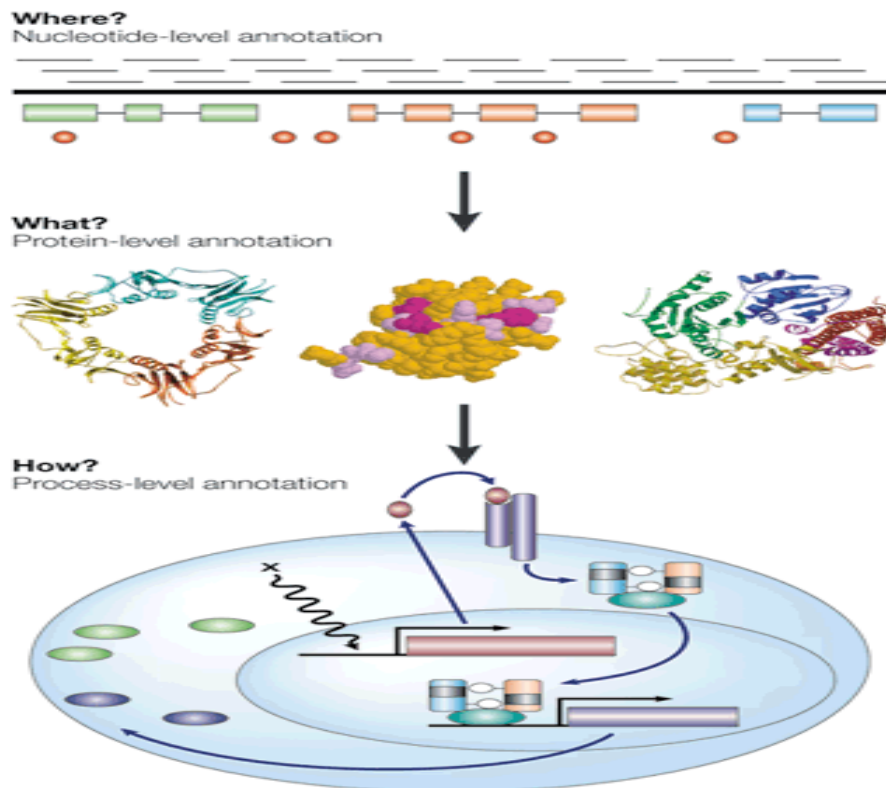


Figure 1. 10: Sequence annotation levels. Figure adapted from Stein (2001).

1.16 Mapping

The part of genome annotation which seeks to identify the location of genomic constituents is a process called ‘mapping’ (Stein, 2001). It is the first step in the genome annotation process that seeks to answer the ‘where’ part of the genome annotation process and forms structural annotation (Stein, 2001). According to Stein (2001), genes (both protein-coding and RNA-coding genes), genetic markers, repetitive elements, endpoints of the putative duplicated regions are identified during the mapping process.

1.16.1 Computational gene prediction

Gene prediction forms the most visible division of the mapping phase of genome annotation (Stein, 2001). It refers to a process that utilises computational tools for detecting genes (essential DNA structures in a genome which are transcribed into RNA, which can be directly functional or carry genetic information essential for the translation process, which then results in the synthesis of proteins of various functions) (Hvidsten, 2004). This process predicts protein-coding genes and RNA-coding genes but can also predict other functional elements such as regulatory regions (Kahl and Meksem, 2008). Gene prediction is the key to genomics, as genes play a considerable role in the functioning of an organism and bear a wide variety of applications in different scientific disciplines, including gene therapy, agriculture, biotechnology, and molecular biology. The main approaches of computational gene prediction are the intrinsic, extrinsic and the comparative approach (DeSalle and Rosenfeld, 2012). According to Kahl and Meksem (2008), the ultimate goal of gene prediction is to identify gene components such as coding and regulatory regions. Computer programs such as NCBI-ORF finder predicts open reading frames in a genomic sequence while tRNA and rRNA-coding genes are predicted with programs including tRNAScan-SE (Lowe and Eddy, 1997) and RNAmmer (Lagesen *et al.*, 2007).

1.17 Problem statement

Despite the economic and medicinal significance of Rhodophyta genomes, surprisingly few genomic studies have been conducted on these species, particularly in South Africa where ocean economy has been identified as an essential sector that needs to be developed to sustain economic development. To the best of my knowledge, there is no other genome sequencing study of any of

the South African Rhodophytes possessing a broad array of applications in various scientific disciplines. As a result, this study serves as the pioneer study to sequence the genome of the South African endemic, economically and scientifically significant *G. pristoides*. Its genome sequence will not only enable understanding of its genomic composition but also serve as a foundation for future research projects, including extraction of genes and novel enzymes with a broad spectrum of scientific applications. As a result, it is necessary for genomes of Rhodophytes occurring on the South African coastline to be sequenced to gain scientific insights and sustain biodiversity through conservation genomics.

1.18 Hypothesis

Genome sequencing, assembly, and annotation of *G. pristoides* from the Kenton-On Sea region will provide biological information that can be used in different scientific disciplines including Biotechnology, Molecular, Bioinformatics and Evolutionary studies.

1.19 Aim

To sequence, assemble and annotate the organellar genomes of *G. pristoides* obtained from the Kenton-On Sea region in South Africa.



University of Fort Hare
Together in Excellence

1.20 Specific objectives

1. To collect *G. pristoides* from the Kenton-On-Sea region.
2. To extract genomic DNA, construct DNA libraries and sequence the red algae organellar genomes using the ThermoFisher Scientific Ion S5™ sequencer system.
3. To assemble the Ion Torrent reads and analyse the genomic sequence using appropriate software programs.
4. To use bioinformatics tools to predict and annotate open reading frames from the sequenced *G. pristoides* genomes mitochondrial and plastid genomes.
5. To recombinantly express, purify and confirm function/activity of a selected protein predicted in Objective 4.

CHAPTER TWO

Materials and Methods



University of Fort Hare
Together in Excellence

2 CHAPTER TWO: MATERIALS AND METHODS

2.1 Sampling and storage of *Gelidium pristoides*

The red algal species was collected by hand-picking on the 14 of July 2017 from the rocky shores of the Shelly beach located in the Kenton-On-Sea region, Eastern Cape, South Africa (Figure 2.1). Upon sampling, the red algal material was cleaned in seawater for removal of mud, sand and epiphytes, and then with distilled water, as the final wash step. The red algal material was then captured and sent for identification at Rhodes University, South Africa, where it was morphologically identified as the *G. pristoides* species. The *G. pristoides* was then stored at -80°C at the Department of Biochemistry and Microbiology at the University of Fort Hare, South Africa.

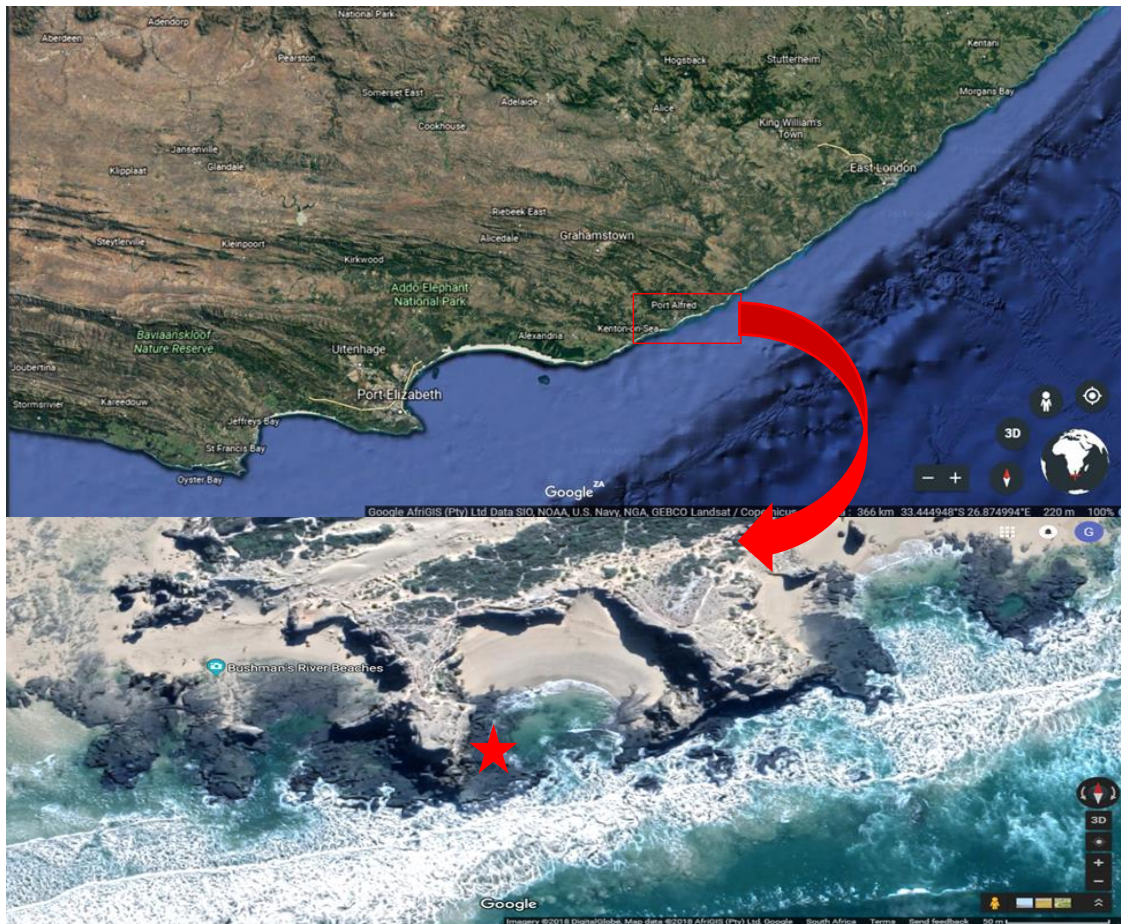


Figure 2. 1: Geographical location of the sampling site on the Kenton-On-Sea region in the Eastern Cape Province. Adapted from Google Earth (<https://earth.google.com/web/@-33.68146685,26.66828445,64.15513905a,9672.3130936d,35y,0h,0t,0r> (Accessed on 5/12/2018).

2.2 Sample preparation

Wet laboratory sample preparations and sequencing experiments were conducted at the CAF molecular laboratory in Stellenbosch University (November 2018) by the candidate and Ms A. Vorster.

2.3 Genomic DNA (gDNA) extraction and qualification (quantity and quality)

The Nucleospin® Plant II kit (Machery-Nagel, Düren, Germany) was used for gDNA extraction from *G. pristoides*, following the manufacturer's instructions based on the CTAB lysis method. However, minor modifications of a double volume of lysis Buffer PL1 were made to ensure sample recovery after the homogenization process. Briefly, 100 mg of wet algal material was mixed with 800 µl of lysis Buffer PL1 and mechanically homogenised into a fine paste using pestle and mortar. The recovered lysate was then treated precisely according to the manufacturer's instructions, and the gDNA was eluted in a low salt TE buffer using double elution, as per the Nucleospin® Plant II kit protocol. For quality control purposes, the quantity of the eluted DNA was measured on the Qubit 2.0 fluorometer (Invitrogen) using the dsDNA HS assay kit, following the protocol described in MAN0002326 REVA.0 (<https://www.thermofisher.com>; **Accessed on 25/11/2017**). The gDNA was further quantified and assessed for its purity with spectrophotometry performed on the Nanodrop® ND- 1000 (ThermoFisher Scientific) instrument which was blanked with the low salt TE buffer. Finally, the extracted gDNA was visualised on 1% agarose gel run with 0.5M TAE buffer.

2.4 Library preparation and quantification

The Ion Plus Fragment Library Kit (ThermoFisher Scientific) was used following the protocol MAN0009847 Rev F.O (ThermoFisher Scientific). However, some minor modifications optimised for a 600 bp library by the Stellenbosch CAF DNA sequencing unit were made. Briefly, two separate 600 bp gDNA libraries were prepared from different input amounts of 250 ng and 500 ng of the previously extracted gDNA. The increment in gDNA input amount was done as an attempt to obtain a high yield genomic library. The 250 ng input was used for the sequencing run randomly named S5-59, while the 500 ng input was used for the so-named S5_60 sequencing run. The following steps describe the library preparation for one sequencing run (S5-59), as precisely the same steps were followed for the library preparation of the second run (S5_60). During the library

preparation process, 130 µl of the gDNA from different tubes was fragmented separately on the Covaris S2 ultrasonicator system (Covaris, Inc.; Woburn, MA, USA) which was operated at minor modifications of 3% intensity and 30 seconds treatment time. The fragmented gDNA contained in different tubes was then purified with 234 µl of the Agencourt™ AMPure™ XP Reagent (Beckman Coulter) and the gDNA bound to the Agencourt™ AMPure™ XP magnetic beads was eluted in 80 µl of low TE buffer.

Seventy-nine microliters of the purified gDNA were end-repaired into blunt ends and subsequently purified using the AMPure™ XP Reagent. The gDNA library bound to the Agencourt™ AMPure™ XP Reagent was eluted in 25 µl of low salt TE buffer before ligation of the IonCode™ Barcode Adapters. The adapter-ligated library was purified with 100 µl of Agencourt™ AMPure™ XP Reagent (Beckman Coulter) and the gDNA library bound to the Agencourt™ AMPure™ XP magnetic beads was eluted in 20 µl low TE buffer.

To achieve the targeted 600 bp library, the Pippin Prep (Sage Science Inc., Beverly, MA, USA) and the Pippin Prep™ Kit CDF 2010 were used for tight selection of 675 bp fragments on a 2% dye-free agarose gel, which was loaded with a 40 µl mixture of 20 µl of gDNA library, 10 µl of low TE, and 10 µl of marker L. The size-selected gDNA library was purified with 42 µl of Agencourt™ AMPure™ XP Reagent (Beckman Coulter) and the gDNA library bound to the Agencourt™ AMPure™ XP magnetic beads was eluted in 25 µl low TE buffer. Twenty-five microliters of the purified gDNA library were first amplified and then quantified on the StepOnePlus™ Real-time system following the Ion Library TaqMan® Quantitation Kit (ThermoFisher Scientific) user guide MAN0015802 Rev A.O (<https://www.thermofisher.com> **Accessed on 25/11/2017**). After this, 1:100 dilution of the library for the S59-run was prepared and used for further quantification and fragment size distribution assessment of the library on the Agilent 2100 Bioanalyzer, using the Agilent™ High Sensitivity DNA Kit (Cat No.1724) and chips based on the protocol G2938-90322 REV. C (<https://www.thermofisher.com> **Accessed on 25/11/2017**).

2.5 Template preparation, Enrichment and Sequencing

The Ion 520™ & Ion 530™ Ext Chef Kit (Thermo Fisher Scientific) was used for template preparation on the Ion chef system (ThermoFisher Scientific) according to the protocol

MAN0015805, REV C.0 (<https://www.thermofisher.com> Accessed on 25/11/2017), with a few modifications optimised by the CAF DNA sequencing unit, Stellenbosch. Briefly, the barcoded-gDNA library quantified by qPCR was first diluted to a final concentration of 12 pM of 50 µl. The diluted barcoded-gDNA library plus a 4 µl Ion S5™ Ext calibration standard were loaded on the Ion Chef liquid handler using Reagents, Solutions and Supplies. After this, the enriched, template-positive ion sphere particles were loaded onto an Ion 530™ Chip (ThermoFisher Scientific). The loaded Ion 530™ Chip was then inserted into the Ion S5™ (ThermoFisher Scientific) platform for massively parallel sequencing using the Ion S5™ Ext Sequencing Solutions and Sequencing Reagents Kits (ThermoFisher Scientific) according to the protocol MAN0015805 REV C.0. (<https://www.thermofisher.com> Accessed on 25/11/2017).

2.6 Pre-assembly quality assessment

Raw data analysis was performed using Torrent Suite™ version 5.6 software (<https://www.thermofisher.com> Accessed on 25/11/2017) with default parameter settings. Briefly, read calibration using a calibration standard was enabled and reads were quality trimmed using a Phred-scale quality score-cutoff of 17 (Q17) over a 30 bp window within the Torrent Suite™ version 5.6 software. Adaptor sequences remaining on the 3' end of reads were also removed during this analysis. The FastQC program (Andrews, 2010) within the Torrent Suite™ was then used for quality assessment of the raw genomic data before assembling the pre-processed reads.

2.7 De novo genome assembly

De novo genome assembly of the Ion S5™ reads was performed using the SPAdes assembler version 3.11.1 (Nurk *et al.*, 2013), operating on default parameter settings. Briefly, the reads from the two different runs were concatenated into one BAM file and were assembled into contigs using the SPAdes version 3.11.1 programme, operating at default parameter settings.

2.8 Post-assembly quality assessment

Post-assembly quality assessment was performed for the *G. pristoides* genome assembly using default parameter settings of the QUAST 4.1 software (Mikheenko *et al.*, 2016) within the Torrent Suite™ Software 5.6. Briefly, the estimated length, N50, and GC content of the *G. pristoides*

genome were estimated through the QCAST 4.1 software, considering only contigs greater than 500 bp.

2.9 Identification and selection of *Gelidium pristoides* organellar (mitochondrial and plastid) genomes

The quality-controlled contigs were scanned for the mitochondrial genome using local BLAST (Altschul *et al.*, 1998) through the BioEdit software (Hall, 1999) with eight mitochondrial genomes of *Gelidiales*, namely; *G. vagum* (Yang, *et al.*, 2014), *G. elegans* (Yang, *et al.*, 2015), *G. sinicola*, *G. sclerophyllum*, *G. crinale f. luxurians*, *G. galapaganse*, *G. isabelae* and *G. arbasence* (Boo *et al.*, 2016) as search queries. After this, the obtained *G. pristoides* mitochondrial contigs were then ordered (arranged) according to the previously published genome sequences of the eight *Gelidiales* using the ‘map against reference’ algorithm of the Geneious R11.12 software trial version (Kearse *et al.*, 2012). The *G. pristoides* plastid genome sequence was also obtained from the quality-controlled contigs as described previously. However, only the available *Gelidium elegans* and *Gelidium vagum* (Lee *et al.*, 2016) plastid genomes were used as search queries for BLAST and as reference genomes for mapping against with the Geneious R11.12 software trial version. The produced *G. pristoides* mitochondrial, and plastid genomes had some gaps which were closed as described later (section 2.10, 2.11 and 2.12).

2.10 Organellar genome gap filling through Sanger sequencing chemistry

2.10.1 Amplification of the organellar gaps and purification of PCR products

For closing the gaps observed in the resultant consensus sequences of mitochondrial and plastid genome, primer pairs (see Appendix A) were designed across the gaps of the mitochondrial and plastid genomes respectively. The Mitochondrial and Plastid primer pairs were designed following the conditions of primers of 21–30 bp, with GC content ranging between 40–60% and a T_m of between 60°C and 68°C. The designed primers were then used for amplification of the gDNA, constituting the observed gaps using the standard PCR method following the manufacturer’s instructions of the Promega technical manual for GoTaq® Long PCR Master Mix. The PCR products with primer dimers and multiple bands were gel purified following the manufacturer’s instructions on the QIAquick® Gel Extraction kit and were then viewed on 1% agarose gels. The

PCR products without primer dimers or other contaminating bands were purified of the unincorporated nucleotides and excess primers using Thermo Scientific ExoSAP purification protocol. Briefly, different reaction mixtures each containing 5 μ l of PCR product, Exonuclease I and FastAP were prepared and incubated in the thermal cycler under the conditions described in Table 2.1.

Table 2. 1: Thermal cycling conditions for purification of PCR products through the Thermo Scientific ExoSAP protocol

Stage	Number of cycles	Temperature ($^{\circ}$ C)	Time (min)
1	1	37	15
2	1	85	15
3	1	16	∞

2.11 Quantitation and cycle sequencing of the organellar genomes



Spectrophotometric quantification of the purified PCR products was done using the Nanodrop (ThermoFisher Scientific) instrument before sequencing at the South African Institute of Aquaculture and Biodiversity (SAIAB) in Grahamstown, South Africa. As per the BigDye Terminator v3.1 cycle sequencing protocol, varying amounts of the pure PCR products were incorporated into 10 μ l cycle sequencing reaction mixtures comprised of 1 μ l BigDye Terminator v3.1 Ready Reaction Mix, 1.5 μ l 5X BigDye Sequencing Buffer, 1 μ l of 3.2 μ M primer stock, and varying amounts of Molecular Grade water following the quantity requirement guide described in Table 2.2. Cycle sequencing was then performed in the Applied Biosystems Veriti TM 96-well thermal cycler operating at conditions described in Table 2.3.

Table 2. 2: Concentrations of PCR products recommended for BigDye Terminator v3.1 cycle sequencing

Size of PCR product	Concentration for cycle sequencing
100–200bp	1–3 ng/μl
200–500bp	3–10 ng/μl
500–1000bp	5–20 ng/μl
1000–2000bp	10–40 ng/μl
>2000bp	40–100 ng/μl

Table 2. 3: Thermal cycling conditions optimised for the Veriti 96 well-thermal cycler for Cycle Sequencing

Stage	Number of Cycles	Step	Temperature	Time
1	1	Denaturation	96	1 min
2	25	Denaturation	96	10 s
		Annealing	50	5 s
		Extension	60	4 min
3	1	Hold	16	∞

2.12 Ethanol-EDTA precipitation and Capillary electrophoresis of products of cycle sequencing

Unincorporated dye-labelled ddNTPs were removed from the sequenced PCR products using the ethanol-EDTA precipitation. The cycle sequencing products were briefly centrifuged and transferred into 0.6 ml tubes containing 2.5 μl EDTA (125 mM), pH 8.0 and 30 μl of 99.9% ethanol. The resulting reaction mixture was briefly vortexed and incubated at room temperature in the dark for 1 hour. The mixture was then centrifuged at 13000 rpm in a microcentrifuge for precisely 20 minutes, and the supernatant was aspirated and discarded from each tube. The pellet was suspended in 30 μl of ice-cold ethanol and centrifuged at 13000 rpm for 15 minutes. The supernatant was then aspirated and discarded from each tube before the samples were dried at 60°C for 5 minutes. The dried samples were then suspended in 12 μl of sequencing buffer before they

were subjected to capillary electrophoresis performed on the (Applied Biosystems) ABI 3500xL Genetic Analyzer.

2.13 Annotation of organellar genomes and construction of circular mitochondrial

Annotation of the mitochondrial and the plastid genome of *G. pristoides* was performed according to the method described by Boo *et al.*, (2016). Briefly, the open reading frames (ORFs) of genes for both the mitochondrial and the plastid genome of *G. pristoides* were annotated using NCBI ORF-Finder (<https://www.ncbi.nlm.nih.gov/orffinder> Accessed on 01/10/2018) and alignments obtained from BlastN and BlastX searches from the NCBI database. The tRNAs and rRNAs constituting these organellar genomes were identified using the tRNAscan-SE1.21 server (Lowe and Eddy, 1997; Lowe and Chan, 2016) and the RNAmmer 1.2 server (Lagesen *et al.*, 2007). The CGView server (Grant and Stothard, 2008) was used to construct the physical map of the *G. pristoides* mitochondrial genome.

2.14 Structural analysis of the mitochondrial Cytochrome c oxidase subunit 3 (Cox3) and plastid heat shock protein 70 (HSP70)

The mitochondrial *Cox3* (MG_000000001.1) and the plastid HSP70 (MG_111111111.1) proteins were chosen for *in silico* analysis for this study. The PRIMO webserver pipeline (Hatherley, 2016) was used for the construction of the 3D model structure of the *Cox3* protein while that of HSP70 was constructed through the SWISS-MODEL webserver pipeline (Waterhouse *et al.*, 2018; <https://swissmodel.expasy.org> Accessed on 6/12/2018). Briefly, the protein sequences of *G. pristoides Cox3* (Figure 2.2) and HSP70 (Figure 2.3) were used separately as query search sequences in the PRIMO and SWISS-MODEL webserver pipelines operating with default parameter settings.

```
>Cytochrome c oxidase subunit 3_cox3_ G. pristoides
MTLLSQISKSVQRHPFHLVDPSPPWPFVSLAAAFSCAVSGVMYMHAFKRGGFSLLSIFISLLIIMFVWWRDVIREATF
EGHHTGIVQQGLRYGIILFIISEILVFFAFFWFAFFHSSLSPPGVEIGSIWPPKGISVIDPWEIFPLNTLILLLSGCTV
TWSHHAIVANLRFQALLSLFLTIVILAVIFTILQAYEYTLADFRSLSDGIYGSTFYMATGFHGFHVFIGTVSLLICFIR
LNQHQLTQQHHFGFESAAWYWHFVDVVWVLFVLSIYWWGGL
```

Figure 2. 2: Protein sequence of *Cox3* protein (MG_000000001.1) from *G.pristoides*


```

>Heat Shock Protein 70 Chaperone_HPS70_G. pristoides
MSKVVGIDLGTNSVAVMEGGKPTVIPNKEGLRTPSVVAYTKKQDKLVGQIAKRQAVMNPENTFYSVKRFGRKQ
EEVGSESKQSSYSVKTDANLNLIKLPALGKDFAPEEISAQVLRKLVEDASTYLGQPVTQAVITVPAYFNDSQRQAT
KDAGQIAGLDVLRINEPTAASLSYGLDKKENETILVFDLGGGTFDVSILEVGDGVFEVLSTSGDTHLGGDDFDSTI
VQWLIKEFYNDQIGIDLAQDRQALQRLTEAAEKAKMELSSLSQTDINLFPFITSTDTGPKHLEKTIITRAQFEQLCNLI
DRCQIPVTNALKDAQLESSNIDEIVLVGGSTRIPAIQDLVKRIIGKDPNQSVNPDEVVAIGA AVQAGVLAGEVKDIL
LLDVTPLSLG VETLGGVMTKII PRNNTTVPTKKSEVFSTAVDNQPNVEIHILQGEREFTKDNKSLGTFRLD GIMPAPR
GVPQIEVTFDIDANGILSVNAKDKGTGKEQSITITGASTLPKDEVEKLVQEAERNSDLKQKREQVDLKNQADSLCY
QSENQLKDLEDKIDNQDKQQANSLIGDLK KLMQTEDYDQIKKVQSE LQQLMMAIGKKVYNNSAPQPDDNNTKDTVID
TESKETN

```

Figure 2. 3: Protein sequence of HSP70 protein (MG_11111111.1) from *G.pristoides*

During the template identification step of homology modelling in the PRIMO web server, the 1v54 (*Bos taurus*) protein with a resolution of 1.8 Å, 95 % coverage and 56% sequence identity to the query sequence was chosen as the template for the *Cox3* protein. Target-template alignments were generated using the T-COFFEE (Di Tommaso *et al.*, 2011) sequence aligner operating in a 3D mode within the PRIMO web server. The 3D model of the *Cox3* protein was constructed using very slow refinement, and its quality was evaluated using PROCHECK (Laskowski, 1993) within the PRIMO web server. The 2kho protein (*Escherichia coli*) with 94% coverage and 57.39% sequence identity was chosen as the template for construction of the HSP70 protein 3D model in the SWISS-MODEL webserver pipeline. The quality of the generated model of HSP70 was evaluated within the SWISS-MODEL pipeline.

2.15 Phylogeny analysis of the *Gelidium pristoides* Cox3 and HSP70

The protein sequence of the *Cox3* protein was BLAST searched against the NCBI non-redundant (nr) database and the top ten blast hits of *G. arbarescens*, *G. sclerophyllum*, *G. kathyanniae*, *G. vagum*, *G. sinicola*, *G. gabrielsonii*, *G. galapaganse*, *G. elegans*, *G. isabelae* and *Grateloupia filicina* (*G. filicina*) (Li *et al.*, 2018) were downloaded and the cladogram tree was then constructed by aligning the downloaded protein sequences to the *G. pristoides* *Cox3* protein using Clustal Omega. The cladogram tree for the HSP70 protein was constructed following the same approach but using the top ten blast hits of *G. vagum*, *G. gabrielsonii*, *G. kathyanniae*, *G. elegans*, *Sebdenia flabellata* (*S. flabellata*) (Lee *et al.*, 2016), *Gracilariopsis tenuifrons* (*G. tenuifrons*) (Iha *et al.* 2018) and *Gracilariopsis longissima* (*G. longissima*)

([https://www.ncbi.nlm.nih.gov/protein/YP_009511292.1?report=genbank&log\\$=protop&blast_](https://www.ncbi.nlm.nih.gov/protein/YP_009511292.1?report=genbank&log$=protop&blast_)

rank=9&RID=05YRS061015 **Accessed on 01/12/2018**), *Schimmelmannia schousboei* (*S. schousboei* (Lee *et al.*, 2016), *Grateloupia taiwanensis* (*G. taiwanensis*) (DePriest *et al.*, 2013) and *G. filicina* [https://www.ncbi.nlm.nih.gov/protein/YP_009488724.1?report=genbank&log\\$=protop&blast_rank=8&RID=05YRS061015](https://www.ncbi.nlm.nih.gov/protein/YP_009488724.1?report=genbank&log$=protop&blast_rank=8&RID=05YRS061015) **Accessed on 01/12/2018**. The cladogram trees were viewed using the TreeView 1.6.6 software (Page, 1996).



University of Fort Hare
Together in Excellence

CHAPTER THREE

Results and Discussion



University of Fort Hare
Together in Excellence


3 CHAPTER THREE: RESULTS AND DISCUSSION

3.1 gDNA extraction and qualification (quantity and quality)

Despite the advancements made in genome sequencing, extraction of sufficient gDNA remains a prerequisite for any sequencing platform as nucleic acids cannot be efficiently sequenced *in situ* (Gansauge *et al.*, 2017). This study extracted, quantified and evaluated the quality of the gDNA from *G. pristoides* with Qubit 2.0 (Life technologies), Nanodrop (Thermo Fisher) and agarose gel electrophoresis. As shown in Table 3.1, the Qubit 2.0 and Nanodrop quantification resulted in a final concentration of 28 ng/μl and 35.5 ng/μl respectively.

Table 3. 1: Quantification and quality assessment results of gDNA extracted from *Gelidium pristoides*

Sample ID	gDNA Qualification		gDNA quality assessment	
	Qubit 2.0 ng/μl	Nanodrop ng/μl	A280/A260	A260/A230
<i>G. pristoides</i>	28	35.5	1.81	1.52


 University of Fort Hare
 Together in Excellence

The primary reason for a higher concentration in Nanodrop quantification is because it also quantifies single-stranded DNA, while the Qubit 2.0 quantification method utilises a dye which specifically binds to double-stranded (dsDNA) only (Simbolo *et al.*, 2013). The quantity used for downstream analysis was 28ng/μl, as it represented the exact amount of dsDNA, which is the type of DNA used for ligation of double-stranded adapters during library preparation. The data are shown in Table 3.1 also show that the gDNA extracted from *G. pristoides* passes the purity test as it fell within the expected range of 1.8–2.0 for pure DNA (Gallagher, 1992). Moreover, the second measure of DNA purity, the A₂₆₀/A₂₃₀ ratio of 1.52 further confirmed that the extracted gDNA falls within the expected guideline of pure DNA, which is higher than 1.5 and close to 1.8 (<https://www.ogt.com/resources/literature/483> Accessed on 21/11/2018).

Figure 3.1 containing the *G. pristoides* gDNA in lane 2, 4, and 6, further illustrates that the gDNA used in this study was of high integrity and was free of contaminating agents, as only a single band

of high molecular weight DNA can be seen in different lanes. Therefore, the gDNA extracted for this study is of relatively good quality.

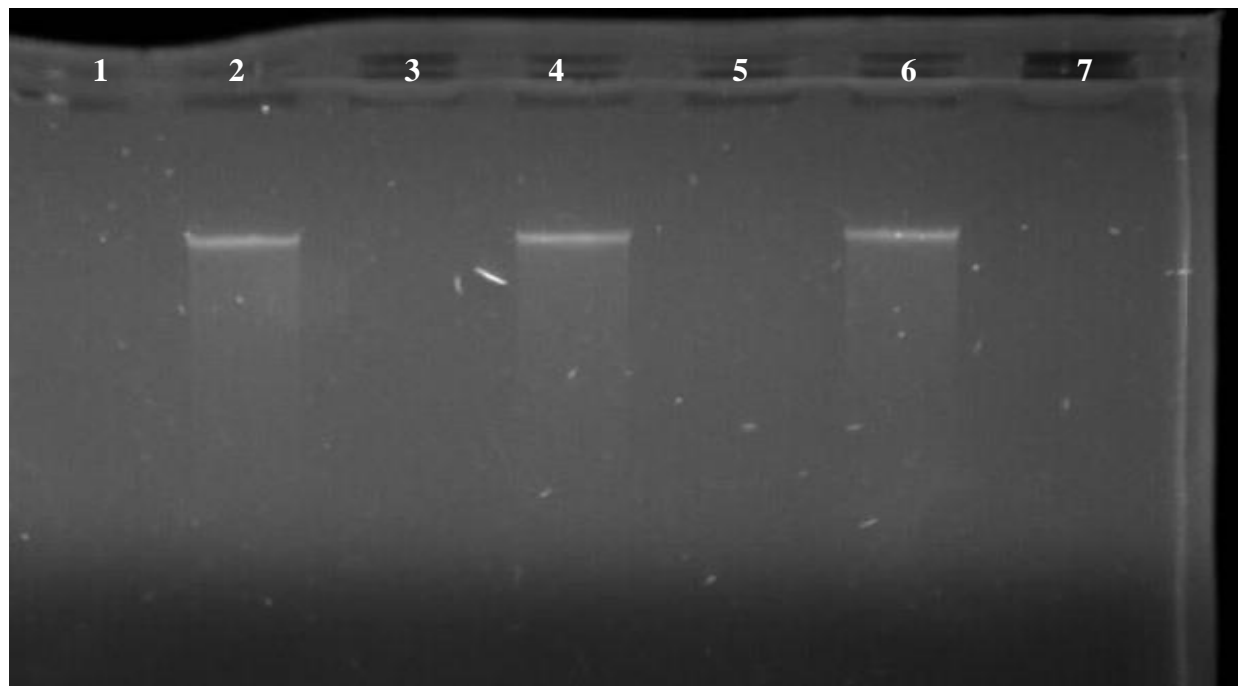


Figure 3. 1: Agarose gel electrophoresis (1% agarose run with 0.5M TAE buffer) of gDNA extracted from *Gelidium pristiodes*. Lane 2, 4, 6: gDNA samples from *G. pristoides*.

3.2 Library preparation and quantification

The whole genome sequencing approach requires preparation of the gDNA library through fragmentation of gDNA into small fragments that are quantified before ligation of double-stranded adaptors which allow the attachment of sequencing primers. Library quantification is the critical step in library preparation as the quality of the sequencing data strongly relies on a sufficient amount of the start library

(<https://www.thermofisher.com/search/results?query=4484177&persona=DocSupport&type=Citations+%26+References> Accessed on 21/11/2017). Library quantity has a considerable effect on the success of the sequencing run and sequencing-based experiment. The NGS sequencing chemistries, for example, require an optimal quantity of adaptor-ligated library to be loaded on the sequencing chip. Compromising the library quantity usually results in a bad sequencing runs; too little DNA library leads to sparsely populated flow cell while too much DNA library leads to an over-populated flow cell with clusters that are close together, making it difficult to interpret the

sequencing data because of poor resolution (<https://www.neb.com/tools-and-resources/feature-articles/the-quantitation-question-how-does-accurate-library-quantitation-influence-sequencing> Accessed on 22/11/2017).

The qPCR used for library quantification accurately quantifies the adaptor-ligated library as required by the NGS sequencing platforms (Aigrain *et al.*, 2016). This study obtained an average library quantity of 32 pM for both DNA input amounts as shown by the qPCR results in Table 3.2. This suggests that the increment in gDNA input amount did not affect the library quantity, and sufficient gDNA library quantity was obtained for template preparation as only 12 pM concentrated library was used for templating.

Table 3. 2: The qPCR library quantification results for different gDNA amounts

Sample ID	Input DNA amount	Average library quantity (pM)	Run ID
<i>G. pristoides</i>	250 ng	32	S5_59
<i>G. pristoides</i>	500 ng	32	S5_60



University of Fort Hare
Together in Excellence

3.3 Genome sequencing and *de novo* genome assembly analysis

3.3.1 Pre-assembly quality assessment

The NGS sequencing platforms are not capable of outputting the entire genomic data as one long continuous stretch. Therefore the Ion S5 used in this study produced numerous sequencing reads that required pre-assembly quality assessment, which is a general requirement before assembly. This is the step that reduces bias problems in genomic data that may lead to erroneous conclusions (<https://insidedna.io/tutorials/view/fastqc-quality-control-and-filtering-of> Accessed on 23/11/2017). For quality control assessment, the FASTQC program which generates multiple results of quality control assessment was used in this study. According to Table 3.3, the total number of read sequences obtained after trimming of adapters and other low-quality sequences

with Q17 added up to 30792074, with zero sequences flagged as poor quality. This indicated that trimming of such sequences with a quality score of 17 was sufficient to produce a genomic data of good quality with an overall GC content of 39%.

Table 3. 3: The basic statistics of the quality controlled genomic data of *Gelidium pristoides* generated by the Ion S5 sequencer for the concatenated bam file of sequencing reads

Genomic Parameter	Value
Total Sequences	30792074
Sequences flagged as poor quality	0
Percentage GC	39

The FASTQC program further generated a Per Base Sequence quality plot (Figure 3.2) showing the probability of each base being correctly called (probability of base calls). The higher the quality score, the better is the base call, as higher quality scores represent a smaller probability of error and vice versa (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> **Accessed on 23/11/2017**). Based on the plot below, most of the quality scores were of good quality, although some were of reasonable quality. Table 3.3, depicting the basic statistics of the raw genomic data is in conjunction with this observation as it shows that there were zero poor quality sequences. This means that most of the bases were correctly called as they fell in the green and the orange regions of Figure 3.2. As per published literature, quality scores that are greater or equal to 20 are reliable base calls as they represent a base call accuracy of above 99% (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality> **Accessed on 23/11/2017**).

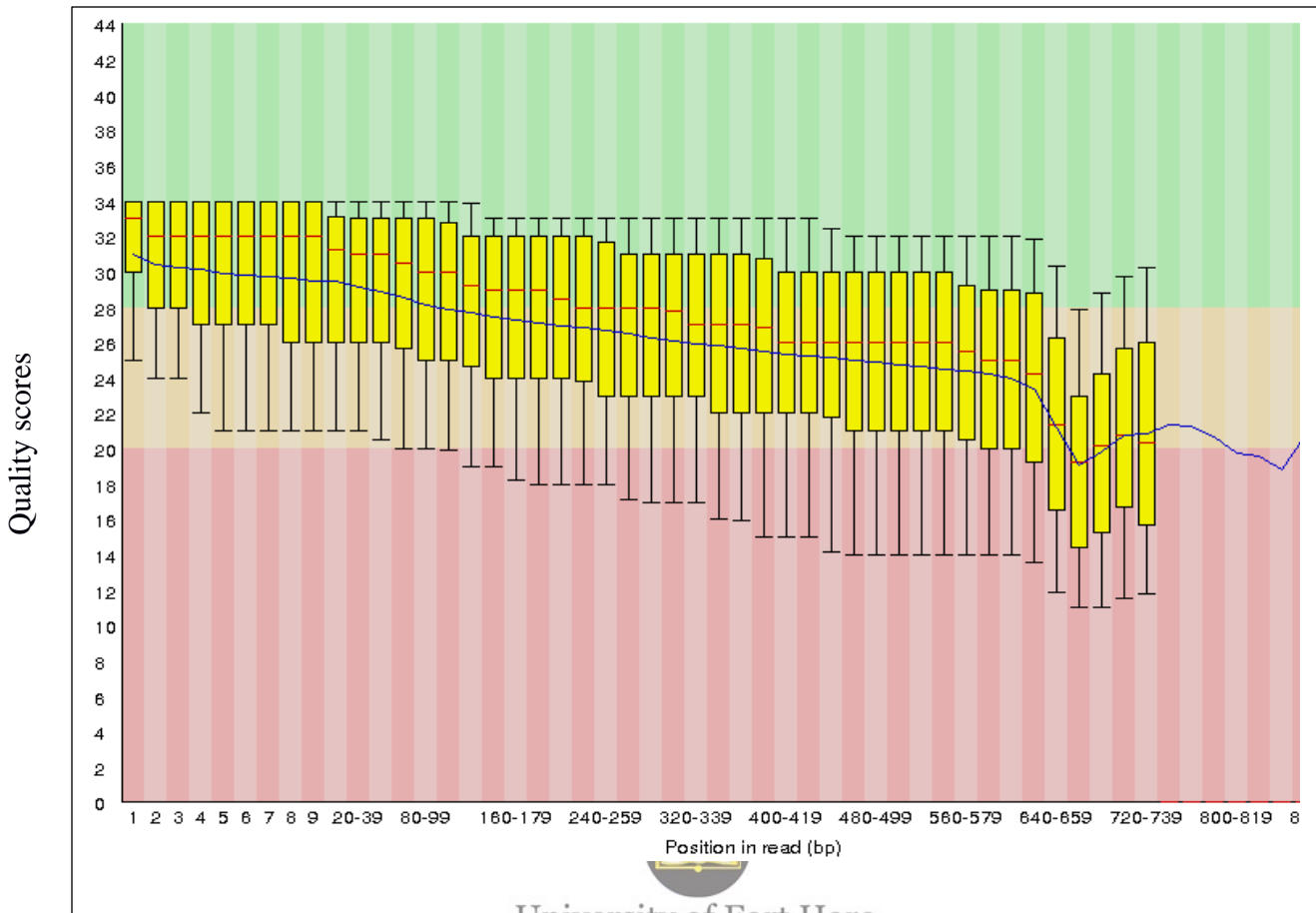


Figure 3. 2: Per Base Sequence Quality plot of reads generated with Ion S5 sequencer system. Green: very good quality base calls Orange: Reasonable quality base calls Red: base calls of poor quality

The per sequence quality score plot further illustrated the good sequence quality of the raw genomic data of *G. pristoides*. Figure 3.3 shows that the generated reads had an average score of 28 which is an indication of a good quality sequence, as average quality scores below 27 and 20 indicate 0.2% and 1% error rates respectively <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/3%20Per%20Sequence%20Quality%20Scores.html> Accessed on 23/11/2018.

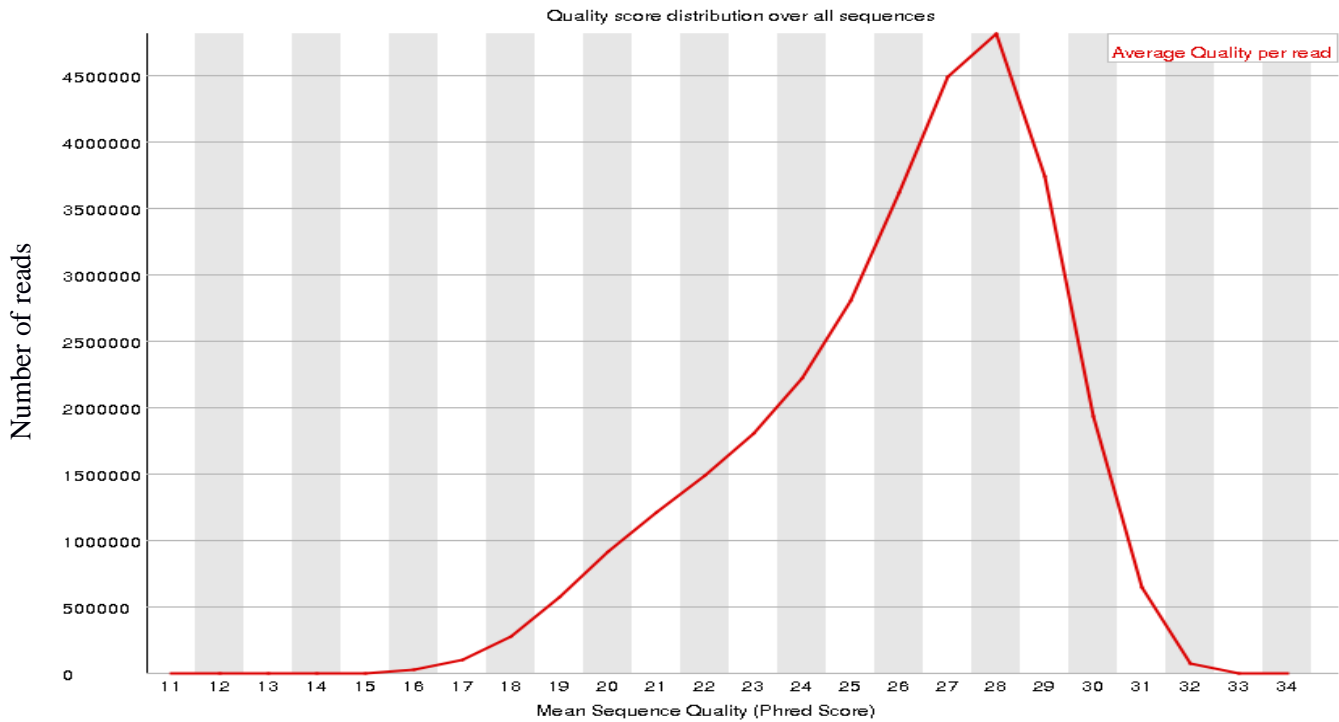


Figure 3. 3: Per Sequence Quality Scores plot of reads generated with Ion S5 sequencer system



As per published literature, the per-base GC content assumes a normal distribution curve, as indicated by the 'Theoretical Distribution curve.'

(<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/5%20Per%20Sequence%20GC%20Content.html> Accessed on 23/11/2017) in Figure 3.4. The per sequence GC content curve (Figure 3.4: red curve) obtained for the *G. pristoides* raw genomic data followed the normal distribution pattern with an overall average GC content of 39%. This suggested that the genomic data obtained for this study was not contaminated, as an unusually distributed curve for per sequence GC content analysis is an indication of a contaminated library or other kinds of sequencing errors

(<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/5%20Per%20Sequence%20GC%20Content.html> Accessed on 23/11/2017).

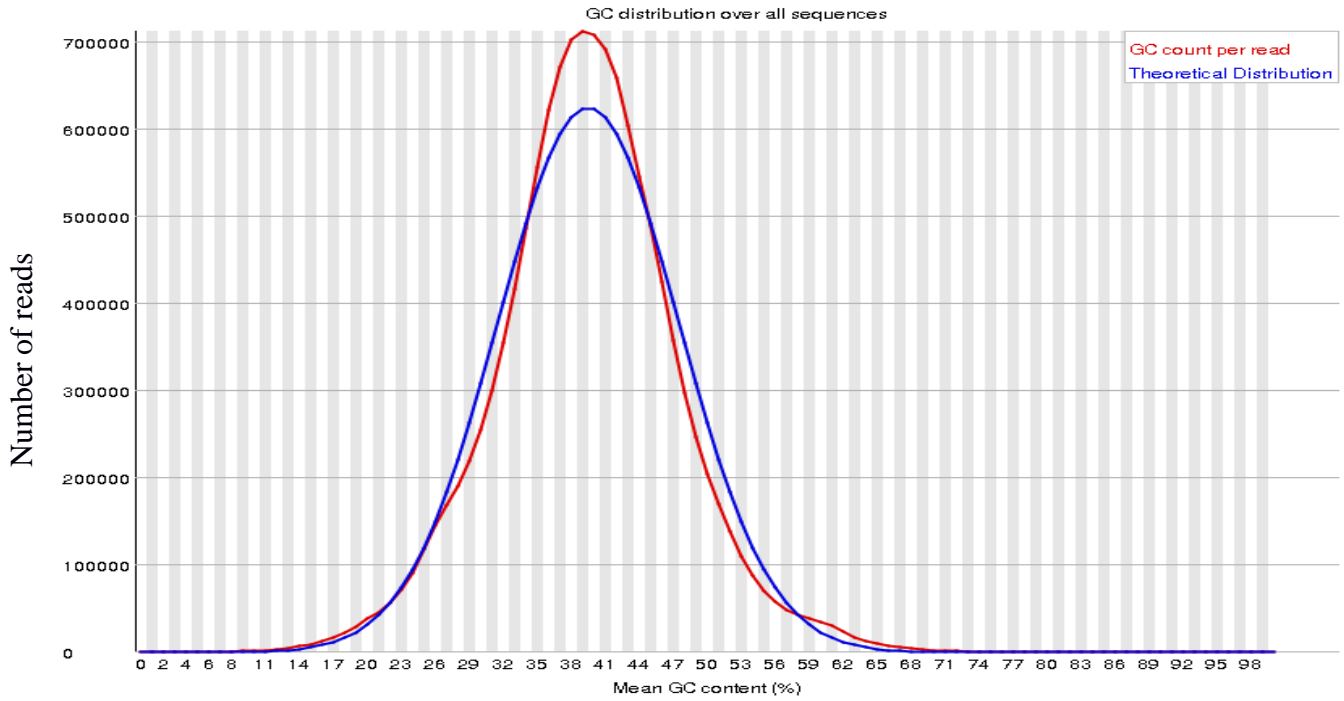


Figure 3. 4: Per Sequence GC content plot of reads generated with Ion S5 sequencer system



University of Fort Hare
Together in excellence

Figure 3.5 indicates that the IonS5 sequencer called all the bases without difficulty as there were no ambiguity bases (usually indicated by Ns in a sequence) inserted in the *G. pristoides* genomic data.

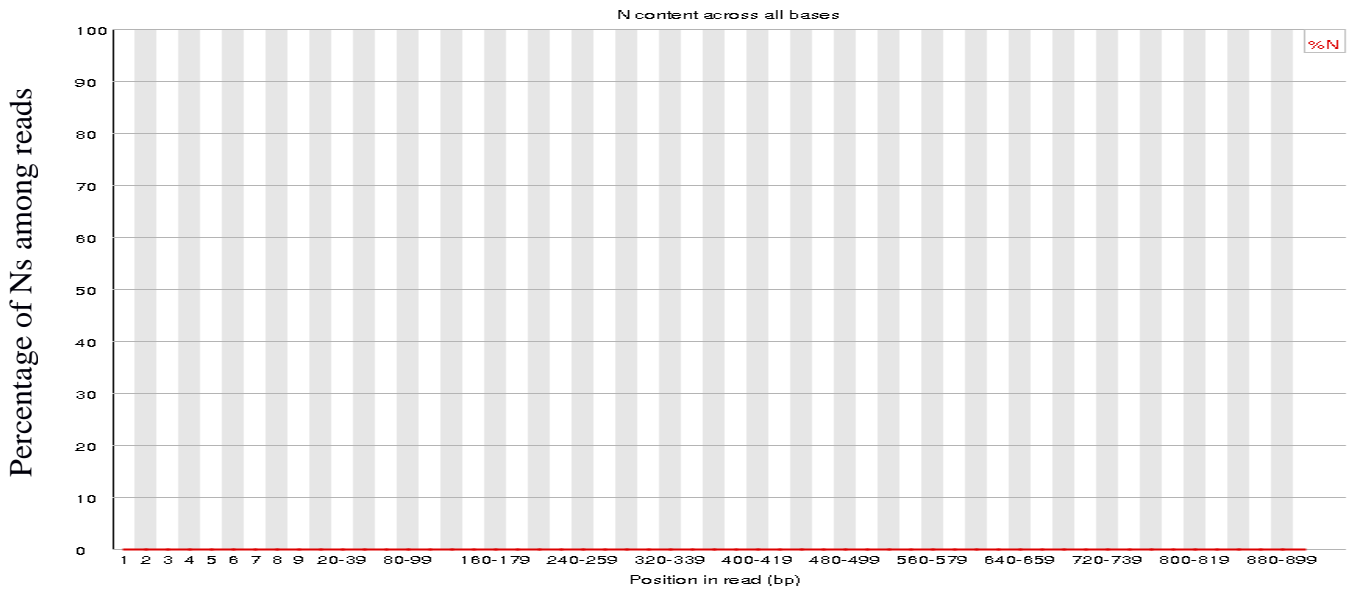


Figure 3. 5: Illustration of Per Base N Content of reads generated with Ion S5 sequencer system

Even if a sequencer can generate reads of uniform length, the quality control process, which involves trimming and filtering of poor quality base calls, results in reads of varying lengths (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/7%20Sequence%20Length%20Distribution.html> Accessed on 23/11/2017). This phenomenon was experienced in this study as reads of varying lengths ranging from 25 bp to 640 bp were obtained as indicated in Figure 3.6. The peak falling between 559 and 620 indicates that the majority of the reads fell within the desired library size of 600bp.

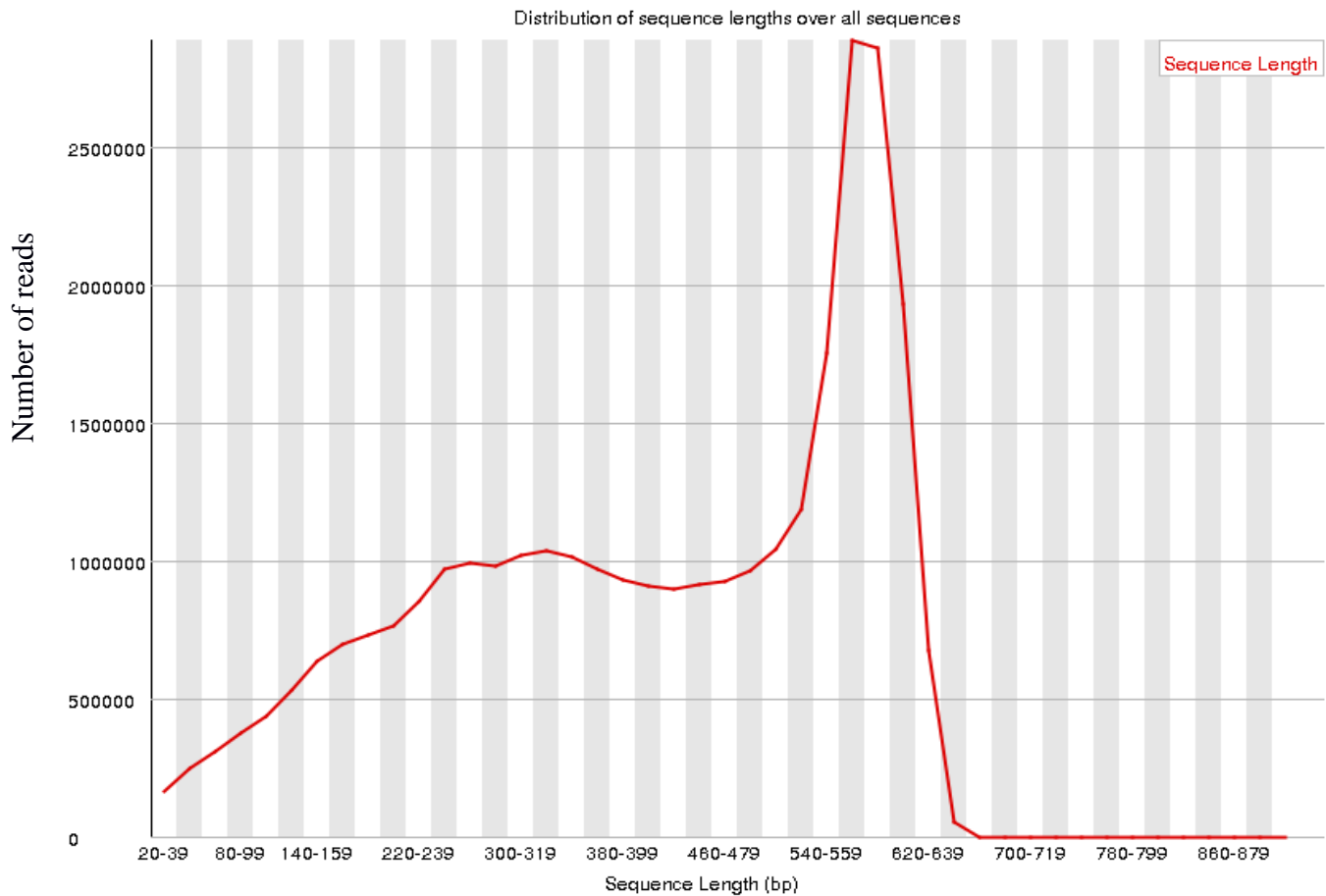


Figure 3. 6: Illustration of Sequence Length Distribution of reads generated with Ion S5 sequencer system

3.4 Post-genome assembly assessment

3.4.1 Estimated length of *Gelidium pristoides* genome assembly

According to the QUAST report statistics shown in Table 3.4, a total length of 217.06 megabases (Mb) sequence was generated from a total of 94140 contigs. The largest contig for *G. pristoides* assembly covered up to 13.17 kilobases (kb) of the total sequence length. This substantial length of the *G. pristoides* genome sequence is due to the presence of the three different types of genomes, the nuclear, plastid and the mitochondrial genome. As mentioned in Chapter One, these types of genomes do not only have different genes but also differing lengths, implying that each genome in the total of 217.06 Mb contribute a different proportion.

Table 3. 4: Genome assembly statistics generated by QUAST based on Contigs

Parameter	Size
Total length	217.06 Mb
Largest contig	13.17 kb
N50	3.17 kb

University of Fort Hare
Together in Excellence

3.4.2 The N50 genomic statistic

The N50 statistic, defined as the minimum length of contig where 50% of the base pairs of the genome assembly are represented, is a widely used metric to evaluate the continuity of the genome assembly (Gurevich *et al.*, 2013; Alhakami, 2017). This study obtained an N50 statistic value of 3.17 kb (Table 3.4) indicating that half of the assembled genome consisted of contigs greater than or equal to 3.17 kilobases. The N50 statistic value is variable amongst genomes of different organisms, for example, that of *Gp. lemaneiformis* was obtained as 3.64 kb (Zhou *et al.*, 2013) while that of *P. yezoensis* was reported as 1.66 kb (Nakamura *et al.*, 2013) and that of *C. crispus* was 64 kb (Collén *et al.*, 2013).

3.4.3 The GC Content

The GC content distribution plot (Figure 3.7) follows a Gaussian distribution as expected for a non-contaminated assembly (Bohlin *et al.*, 2010). It indicated an average GC content of 41.72%

which is slightly higher than that of 39% observed in the FASTQC report. This GC content variation is primarily because the QUAST report was generated using the QUAST software under the default parameter setting which only considers contigs greater than 500 bp. This suggested that the contigs less than 500 bp were GC poor. Hence there was an increment in the GC content. The GC content of 41.72% correlated with the overall genome GC content of *Gp. lemaneiformis* of 48% (Zhou *et al.*, 2013). As revealed by other studies, GC content varies among different species. For example, *C.merolae*, *C.crispus*, *P.purpureum*, and *P.yezoensis* were reported with GC contents of 55.0%, 52.86%, 55.5%, and 63.6% respectively (Matsuzaki *et al.*, 2004; Collén *et al.*, 2013; Bhattacharya *et al.*, 2013; Nakamura *et al.*, 2013).

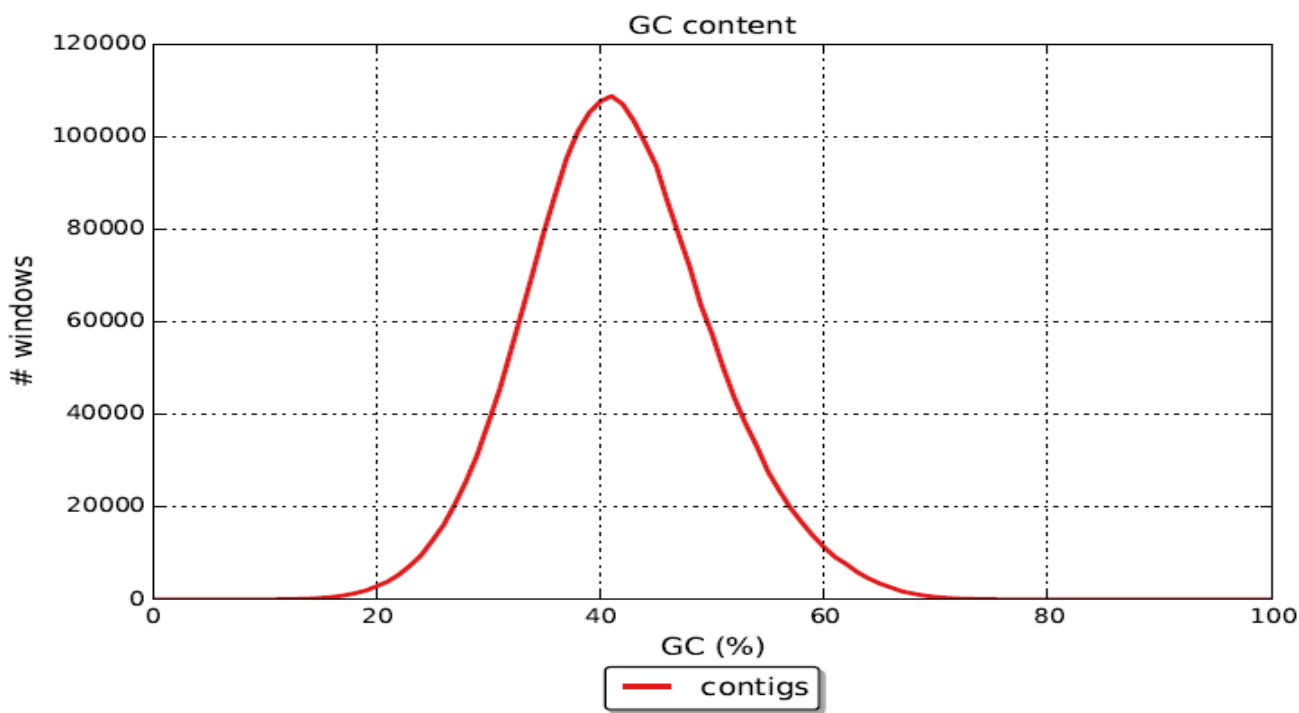


Figure 3. 7: GC content distribution curve of the *Gelidium pristoides* genome estimated by the QUAST 4.1 software operating under default parameter settings on Contigs of *Gelidium pristoides*

3.5 Identification and selection of *Gelidium pristoides* organellar genomes

The mitochondrial and plastid genomes of *G. pristoides* produced by mapping the quality-controlled contigs against the organellar genomes of *Gelidales* had some gaps, as shown in Figure 3.8 and Figure 3.9 respectively. According to the conserved architecture of the mitochondrial and plastid genomes of *Gelidales*, these gaps observed in the organellar genomes of *G.pristoides* are

expected to be a mix of RNA-coding genes, protein-coding genes and non-coding regions (<https://www.ncbi.nlm.nih.gov> Accessed 25/03/2019). A study by Yang et al. (2015) and Wang et al. (2013) also produced draft mitochondrial and plastid genomes with gaps which were closed by PCR and Sanger sequencing as in this study. These gaps were observed in organellar genomes of different Rhodophyta organellar genomes sequenced with different NGS platforms (Ion torrent and Roche 454) in studies of Yang et al. (2015) and Wang et al. (2013). Generally, these gaps are just an indication of incomplete sequencing. According to The Institute for Genomic Research (TIGR), these gaps can be due to genomic data that is difficult to clone (http://rice.plantbiology.msu.edu/training/Sequencing_assembly.pdf Accessed on 22/09/2018).

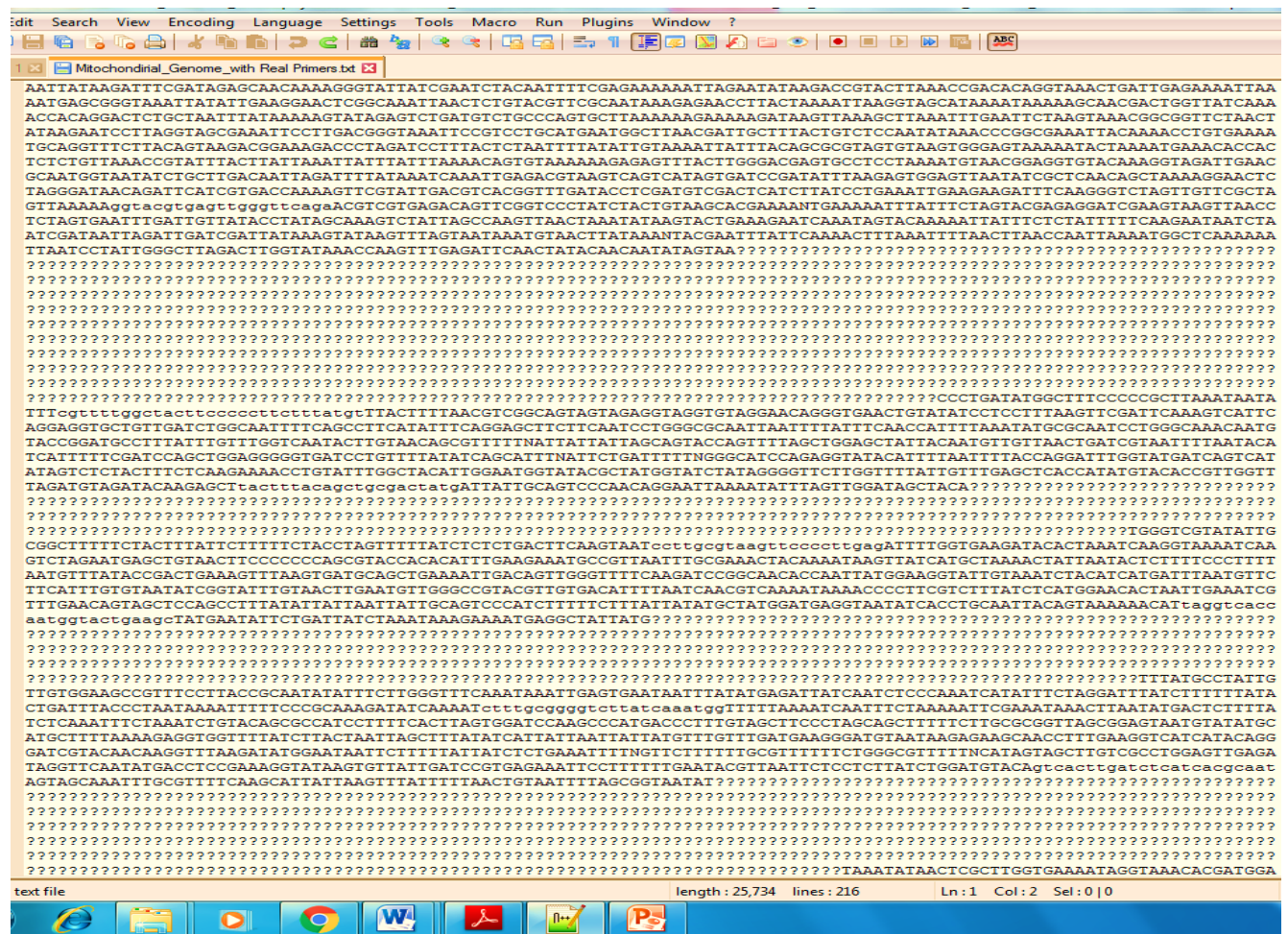


Figure 3. 8: A portion of the mitochondrial genome gaps after mapping of the quality-controlled contigs of *G.pristoides* against the mitochondrial genomes of other *Gelidales*. **Question marks (?)**: The *G.pristoides* mitochondrial gaps.

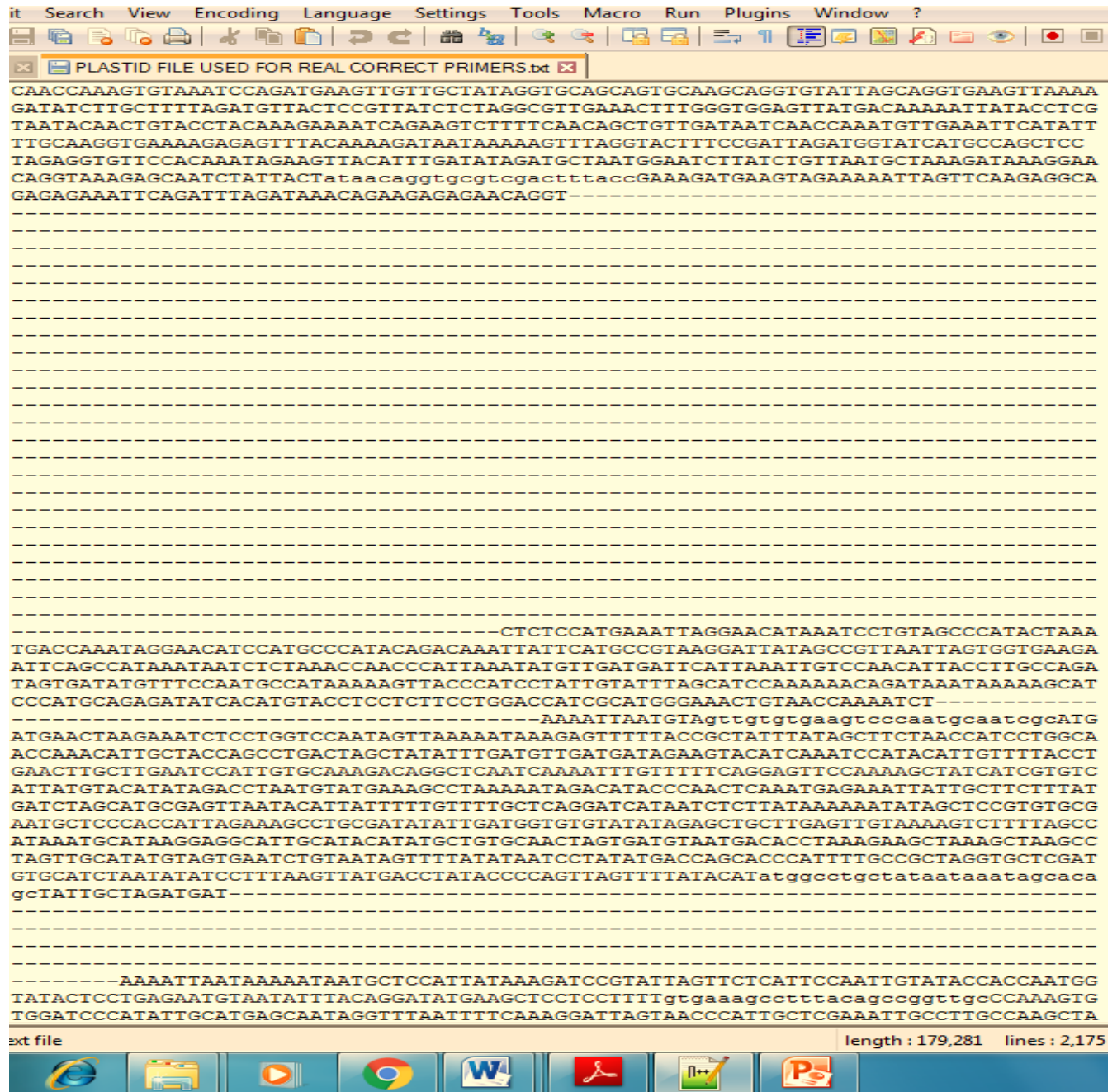


Figure 3. 9: A portion of the *G.pristoides* plastid genome gaps after mapping of the quality-controlled contigs of *G.pristoides* against the plastid genomes of *G.vagum* and *G.elegans*. Dashes (-): The *G.pristoides* plastid gaps.

3.6 Organellar genome gap filling

3.6.1 Mitochondrial gap amplification

All the gaps observed in the mitochondrial genome of *G. pristoides* were PCR amplified before being sequenced to produce a complete circular mitochondrial genome. Figures 3.10, 3.11 and 3.12 indicate the amplified gDNA across the mitochondrial gaps. Figure 3.10 contains gap amplicons from 1–11 while Figure 3.11 contains gap amplicon 12. Figure 3.12 contains gap amplicons 1.1, 7.1, 8.1, 9.1, 11.1 and *yml* that were amplified from mitochondrial gaps that did not close after incorporation of the sequences of the gap amplicons 1, 7, 8, 9 and 11, due to their sizes. The overall molecular weight range of the amplified mitochondrial gaps ranged from about 400 bp to 3000 bp as indicated in the gel figures.

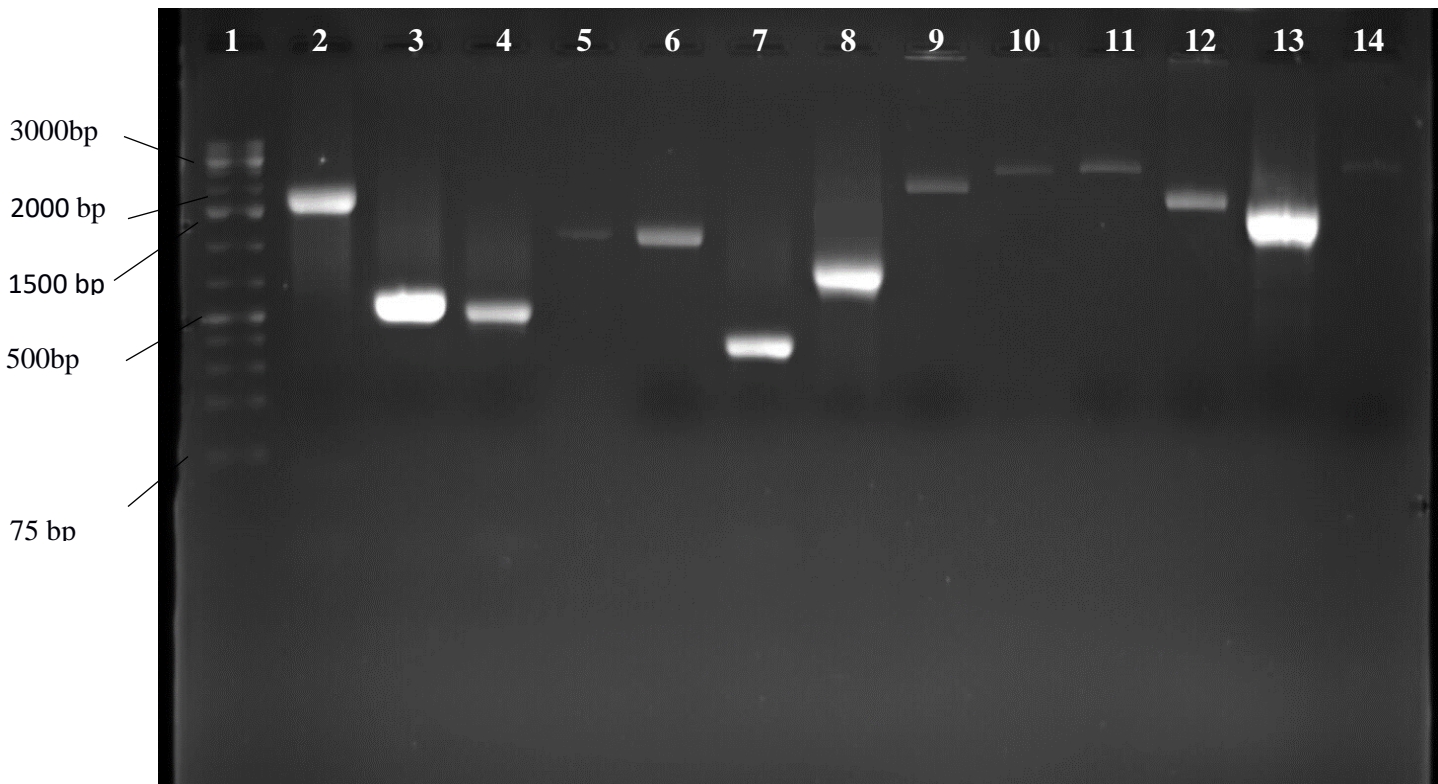


Figure 3. 10: Agarose gel electrophoresis (1% agarose) of PCR products of *Gelidium pristoides* mitochondrial genome. **Lane 1:** 1 kb GeneRuler™ Plus DNA size marker **Lane 2–4:** gap 1-3 amplicons **Lane 5–6:** gap 4 amplicons **Lane 7–14-:** gap 5–11 amplicons.

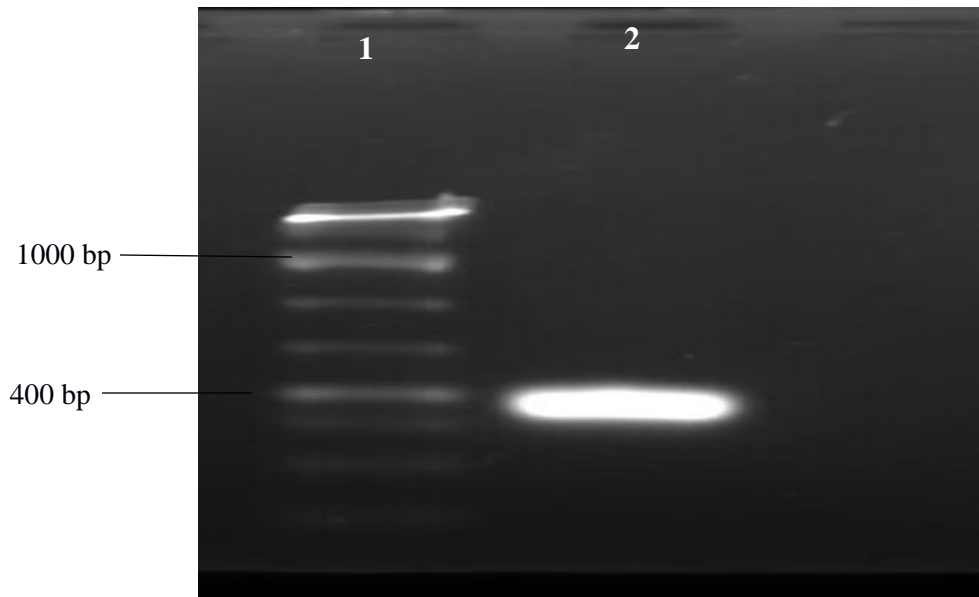


Figure 3. 11: Agarose gel electrophoresis (1% agarose) of PCR products of *Gelidium pristoides* mitochondrial genome. **Lane 1:** 1 kb GeneRuler™ Plus DNA size marker **Lane 2:** gap 12 amplicon.

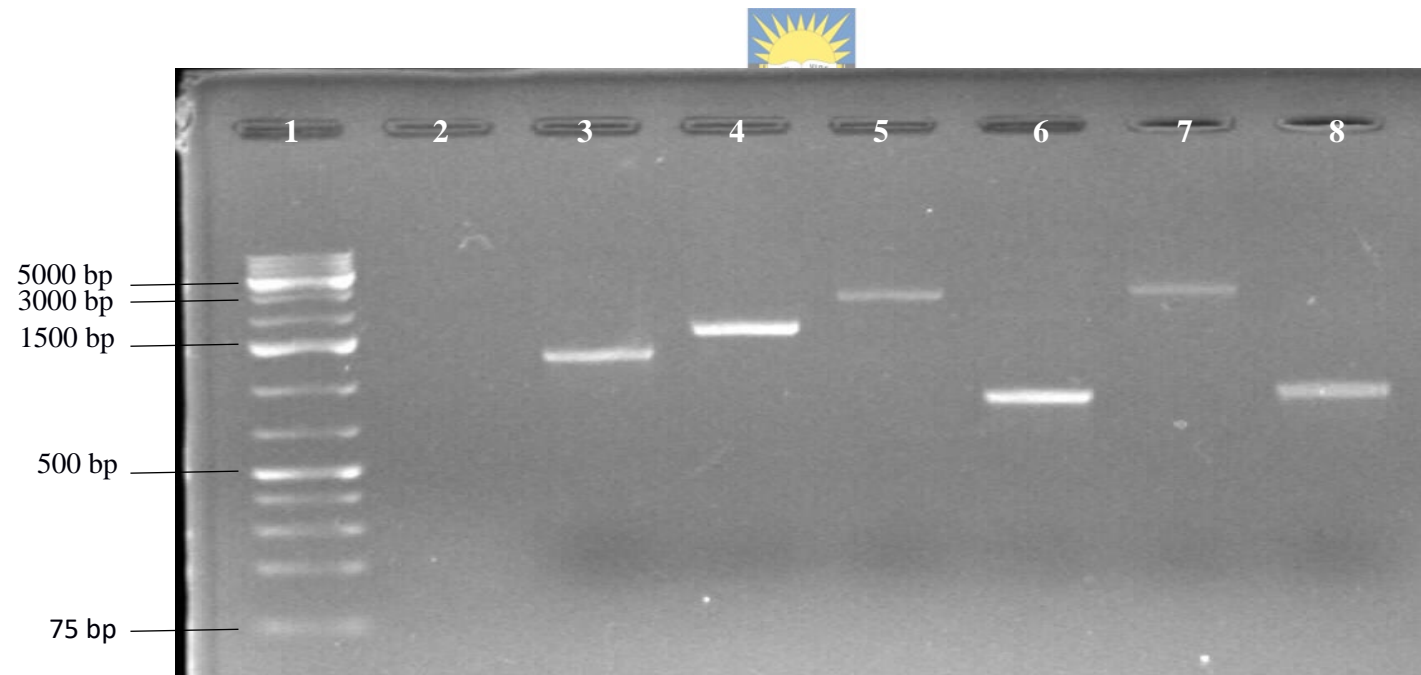


Figure 3. 12: Agarose gel electrophoresis (1% agarose) of PCR products of *Gelidium pristoides* mitochondrial genome. **Lane 1:** 1 kb GeneRuler™ Plus DNA size marker **Lane 3:** gap 1.1 amplicon **Lane 4:** gap 7.1 **Lane 5:** gap 8.1 amplicon **Lane 6:** gap 9.1 amplicon **Lane 7:** gap 11.1 amplicon **Lane 8:** gap ymf amplicon.

3.6.2 Plastid gap amplification

Attempted gap filling of the plastid genome gaps did not amplify all the gaps observed in the plastid genome. As a result, the plastid genome represented in this study is partial. These unamplified plastid gaps form the future of this study as the primer design of these gaps needs to be revised since they did not produce amplicons despite optimisation attempts. Of the 67 plastid gaps, 36 samples were able to amplify, as indicated in Figures 3.13, 3.14, 3.15 and 3.16. Figure 3.13 contains gap amplicons from 1–14, except for 12 which did not amplify while Figure 3.14 contains gap amplicons from 17 to 31 except 15,16, 23, 24, and 26 which also did not amplify. Figure 3.15 contains gap amplicons from 34–67 except for 32, 33, 35, 36, 37, 39, 40, 41, 43, 44, 45, 46, 47, 48, 49, 51, 52, 55, 56, 57, 58, 59, 60, 63 and 64, which did not amplify during PCR amplification of the plastid gaps. Figure 3.16 contains gap amplicon 4, 7, 11 and 66. The molecular weight range of the amplified plastid gaps ranged from about 500 bp to 4000 bp as indicated in these figures.

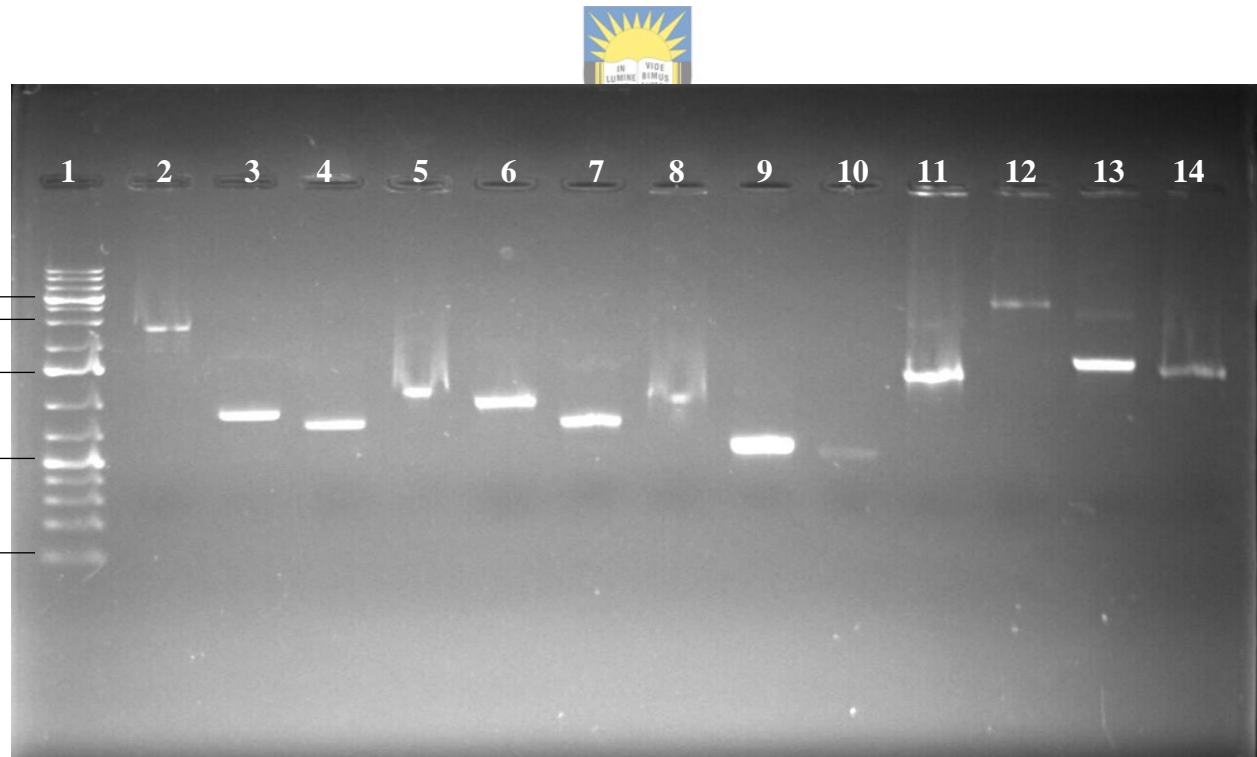


Figure 3. 13: Agarose gel electrophoresis (1% agarose) of PCR products of *Gelidium pristoides* plastid genome. **Lane 1:** 1 kb GeneRuler™ Plus DNA size marker **Lane 2–12:** gap 1–11 **Lane 13:** gap 13 **Lane 14:** gap 14.

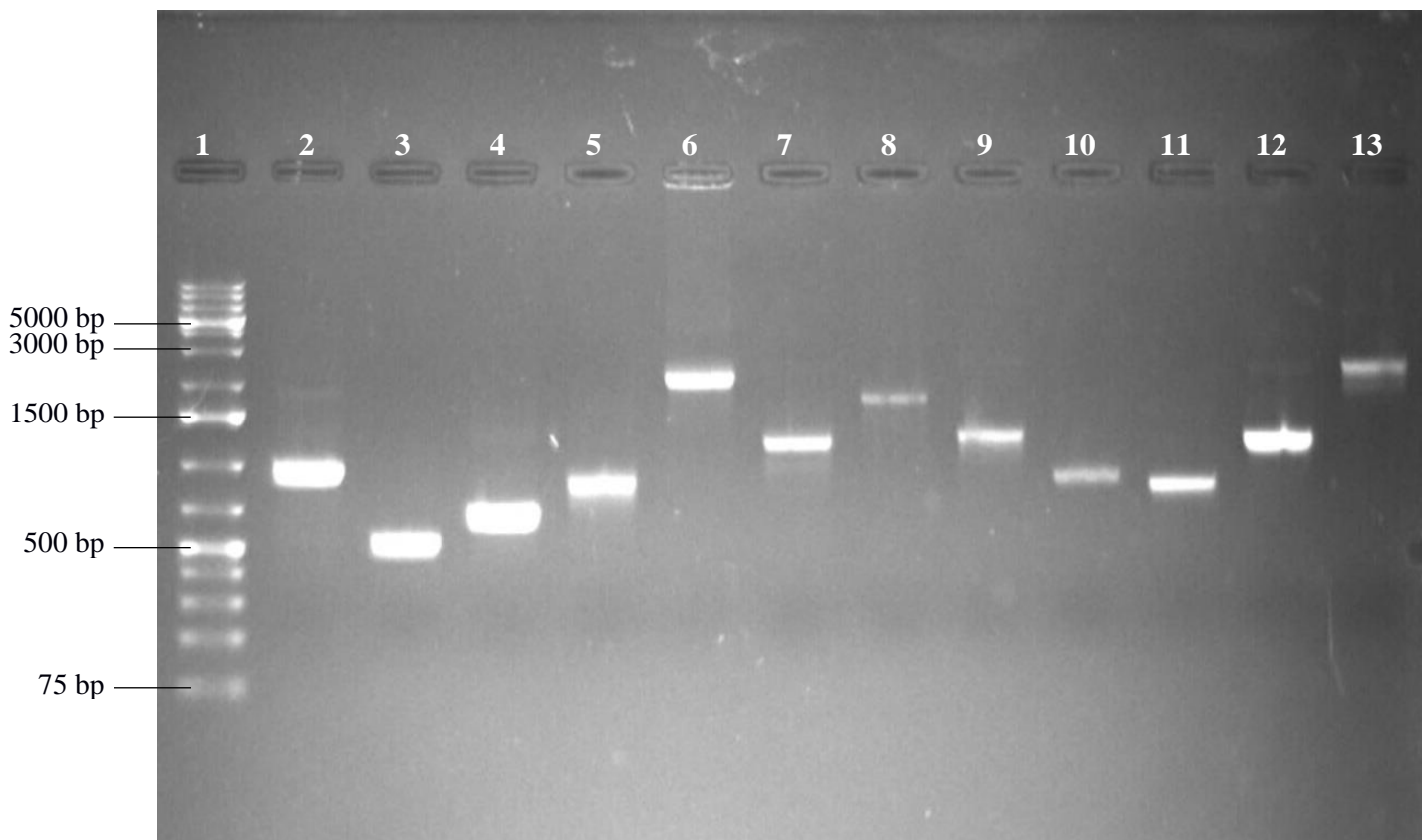


Figure 3. 14: Agarose gel electrophoresis (1% agarose) of PCR products of *Gelidium pristoides* plastid genome. **Lane 1:** 1 kb GeneRuler™ Plus DNA size marker **Lane 2–7:** gap 17–22 amplicons **Lane 8:** gap 25 amplicon **Lane 9–13:** gap 27–31 amplicons.

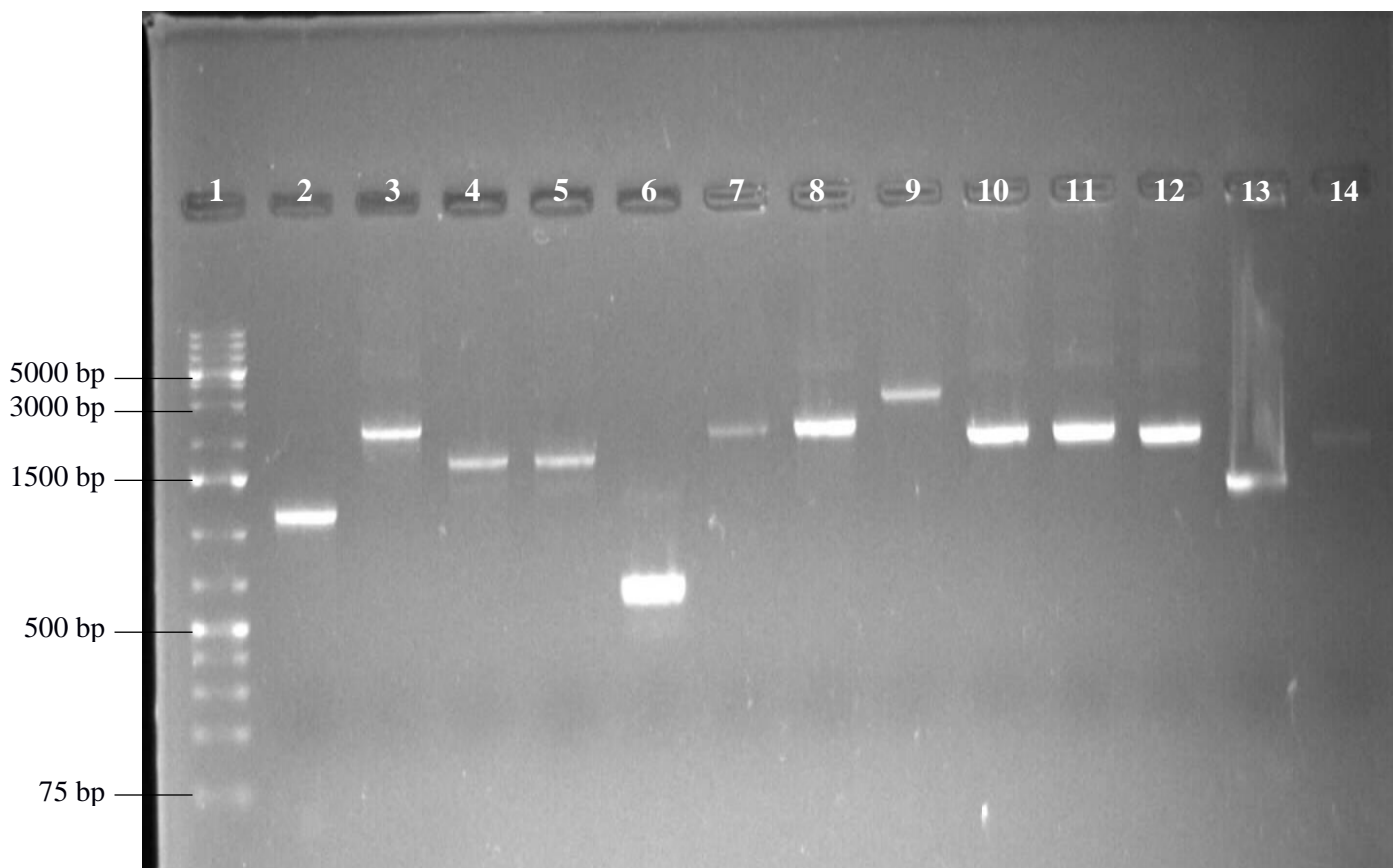


Figure 3. 15: Agarose gel electrophoresis (1% agarose) of PCR products of *Gelidium pristoides* plastid genome. **Lane 1:** 1 kb GeneRuler™ Plus DNA size marker **Lane 2:** gap 34 amplicon **Lane 3:** gap 38 amplicon **Lane 4-5:** gap 42 amplicon **Lane 6:** gap 50 amplicon **Lane 7:** gap 53 amplicon **Lane 8:** gap 54 amplicon **Lane 9:** gap 61 amplicon: **Lane 10-11:** gap 62 amplicon **Lane 12:** gap 65 amplicon **Lane 13:** gap 66 amplicon **Lane 14:** gap 67 amplicon

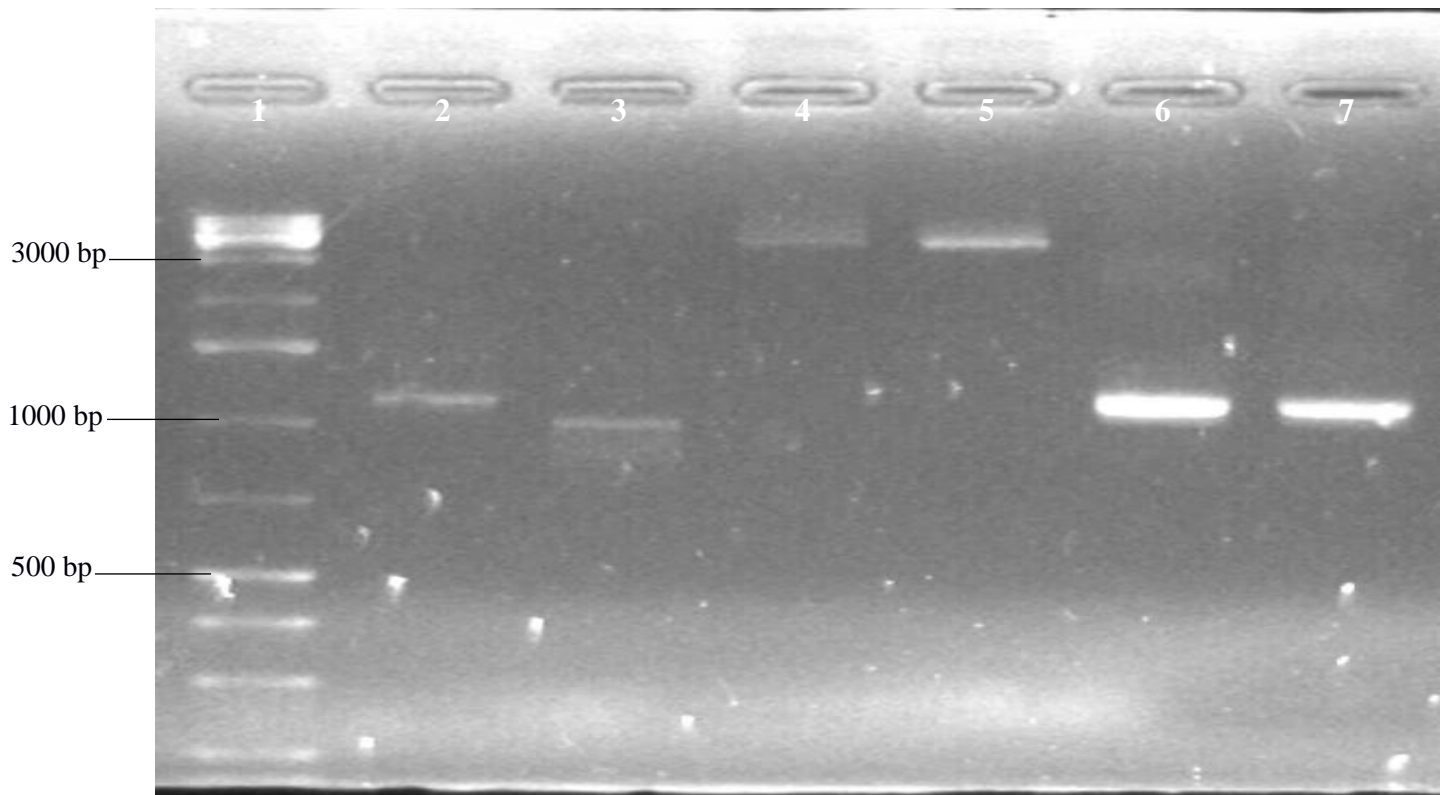
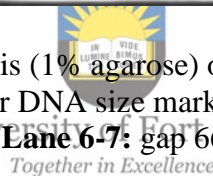


Figure 3. 16: Agarose gel electrophoresis (1% agarose) of PCR products of *Gelidium pristoides* plastid genome. **Lane 1:** 1 kb GeneRuler DNA size marker **Lane 2:** gap 4 amplicon **Lane 3:** gap 7 amplicon **Lane 4–5:** gap 11 amplicon **Lane 6–7:** gap 66 amplicon



3.7 Annotation of *Gelidium pristoides* organellar genomes

3.7.1 The *Gelidium pristoides* mitochondrial genome

The mitochondrial genome of *G. pristoides* mapped as a circular molecule (Figure 3.17) consisting of a total of 25012 bp which is within the size range of mitochondrial genomes from other members of the *Florideophyceae* class. For example, the *Gracilaria textorii* mitochondrial genome mapped into a circular genome of 25743 bp while that of *C. crispus* mapped into 25836 bp. Compared to sequenced mitochondrial genomes of other *Gelidiales*, the mitochondrial genome of *G. pristoides* is the largest. For example, the mitochondrial genome of *G. vagum* is 24901 bp, *G. crinale f. luxurians* is 24910 bp, *G. sclerophyllum* is 24916 bp, *G. elegans* is 24922 bp, *G. arborescens* is 24935 bp, *G. isabellae* is 24937 bp, *G. sinicola* is 24969 bp, *G. galapaganse* is 24970 bp, *G.*

kathyanniae is 24963 bp, and *G. gabrielsonii* is 24964 bp long (Yang *et al.*, 2014; Boo *et al.*, 2016; Boo and Hughey, 2018).

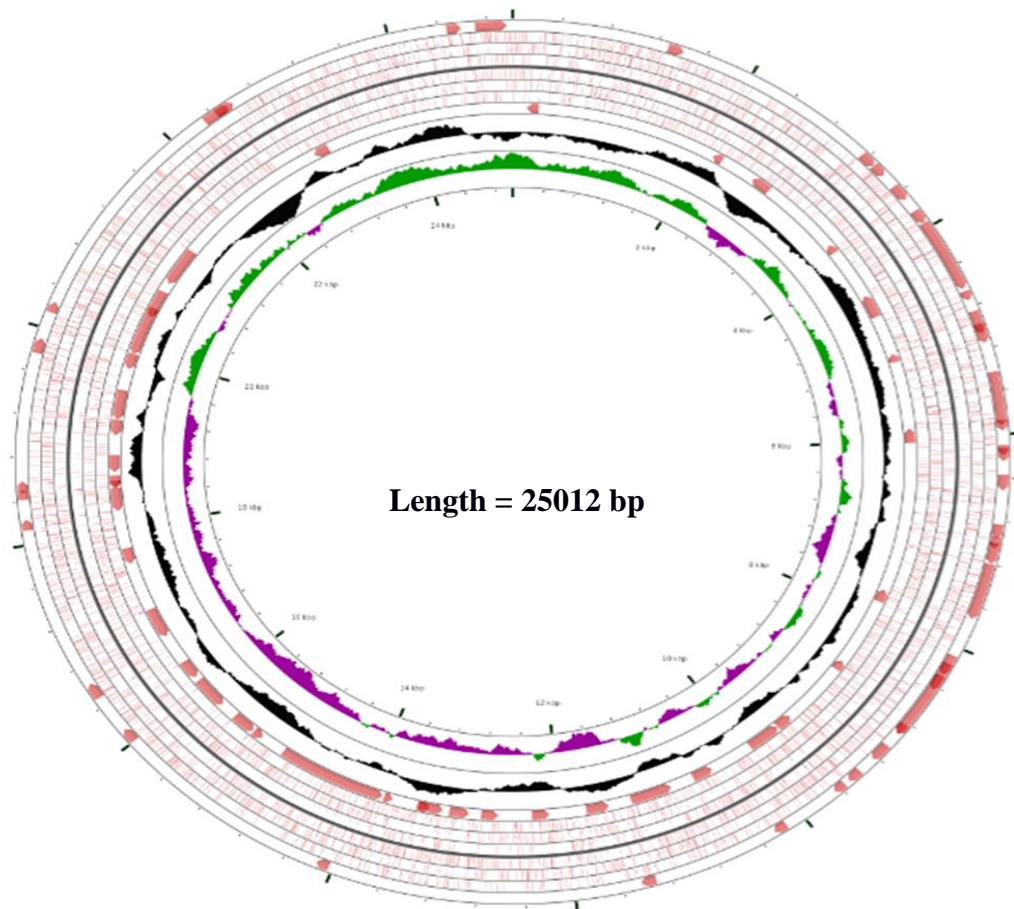
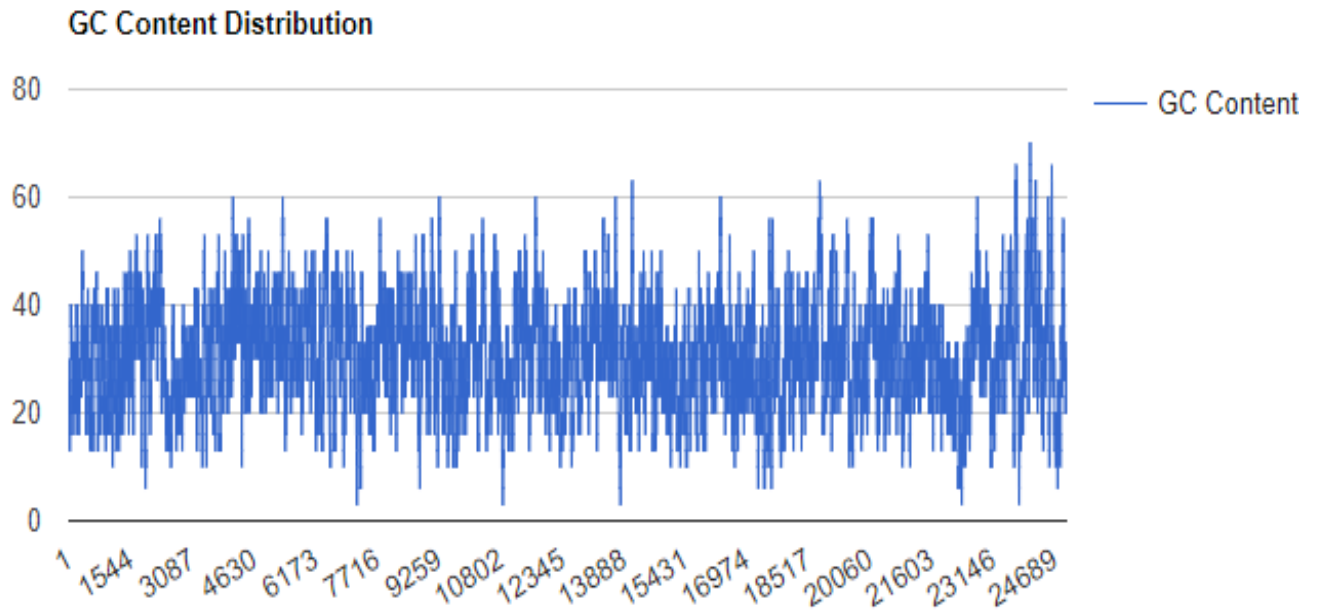


Figure 3. 17: Physical map of the circular *Gelidium pristoides* mitochondrial genome. The physical map was created using the CGView server.

As shown in Figure 3.18, the total length of the *G. pristoides* mitochondrial genome is distributed among 9109 Adenines (A), 8140 Thymines (T), 3946 Guanines (G) and 3817 Cytosines (C) adding up to a GC content of 31.04%, which is comparable to the GC content range of other *Gelidiales* falling between 28.7 to 30.4% as reported by Yang *et al.*, (2014), Boo *et al.*, (2016) and Boo and Hughey (2018).



Summary: Full Length(25012bp) | A(36% 9109) | T(34% 8140) | G(15% 3946) | C(15% 3817)

Figure 3. 18: Distribution of the GC content and nucleotides in the mitochondrial genome of *Gelidium pristoides*. Graphical representation and statistics obtained from <https://www.biologicscorp.com/tools/GCContent/> Accessed on 11/12/2018).

The mitochondrial genome of *G. pristoides* is comprised of a total of 45 genes of different nature and functions as shown in Table 3.5. This is the same as the number of genes reported for the recently published mitochondrial genomes of the *G. kathyanniae* and *G. gabrielsonii* (Boo and Hughey, 2018) while other *Gelidiales* possess 43–44 genes (Yang *et al.*, 2014; Boo *et al.*, 2016). These 45 genes are distributed in both strands of the *G. pristoides* mitochondrial genome, as it is also the case in mitochondrial genomes of other published *Gelidiales* and other members of the *Florideophytes* (Yang *et al.*, 2014; Boo *et al.*, 2016; Boo and Hughey, 2018).

Table 3. 5: Genes predicted from the mitochondrial genome of *Gelidium pristoides*

Gene number	Gene name	Description	Gene positions
1	<i>rrl1</i>	Large subunit ribosomal RNA	16–2570
2	<i>rps3</i>	Ribosomal protein 3	2589–3287
3	<i>rpl16</i>	Ribosomal protein 16	3289–3696
4	<i>trnD</i>	tRNA-Asp	3701–3772
5	<i>Cox1</i>	Cytochrome c oxidase subunit 1	3817–5418
6	<i>Cox2</i>	Cytochrome c oxidase subunit 2	5425–6267
7	<i>Cox3</i>	Cytochrome c oxidase subunit 3	6372–7191
8	<i>Ymf39</i>	ATP Synthase B chain precursor	7195–7737
9	<i>trnG</i>	tRNA-Gly	7739–7811
10	<i>trnQ</i>	tRNA- Gln	7822–7893
11	<i>trnL</i>	tRNA -Lue	7949–8025
12	<i>Cob</i>	apocytochrome b	8082–9233
13	<i>trnL</i>	tRNA-Lue	9271–9358
14	<i>nad6</i>	NADH dehydrogenase subunit 6	9359–9967
15	<i>trnG</i>	tRNA-Gly	9983–10056
16	<i>trnH</i>	tRNA-His	10061–10132
17	<i>sdh2</i>	Succinate: cytochrome c oxidase subunit 2	10132–10878
18	<i>sdh3</i>	Succinate: cytochrome c oxidase subunit 2	10880–11263
19	<i>trnF</i>	tRNA-Phe	11281–11353
20	<i>trnS</i>	tRNA-Ser	11358–11446
21	<i>trnP</i>	tRNA-Pro	11454–11526
22	<i>atp9</i>	ATP synthase FO subunit 9	11537–11767
23	<i>trnC</i>	tRNA-Cys	11804–11874
24	<i>trnM</i>	tRNA-Met	11877–11951
25	<i>rps11</i>	Ribosomal protein 11	11954–12313
26	<i>nad3</i>	NADH dehydrogenase subunit 3	12455–12820
27	<i>nad1</i>	NADH dehydrogenase subunit 1	12832–13815
28	<i>nad2</i>	NADH dehydrogenase subunit 2	13835–15319
29	<i>sdh4</i>	Succinate: cytochrome C oxidoreductase subunit 4	15333–15575
30	<i>nad4</i>	NADH dehydrogenase subunit 4	15576–17051

Gene number	Gene name	Description	Gene positions
31	<i>nad5</i>	NADH dehydrogenase subunit 5	17600–19573
32	<i>atp8</i>	ATP synthase FO subunit 8	19590–19994
33	<i>atp6</i>	ATP synthase FO subunit 6	19994–20755
34	<i>trnSup</i>	tRNA-Suppressor	20772–20845
35	<i>trnA</i>	tRNA-Ala	21305–21377
36	<i>trnN</i>	tRNA-Asn	21471–21543
37	<i>trnV</i>	tRNA-Val	21549–21621
38	<i>trnR</i>	tRNA-Arg	21629–21702
39	<i>trnK</i>	tRNA-Lys	21716–21789
40	<i>SecY</i>	Preprotein translocase subunit <i>secY</i>	21813–22598
41	<i>rps12</i>	Ribosomal protein S12	22558–22923
42	<i>trnE</i>	tRNA-Glu	22934–23006
43	<i>trnM</i>	tRNA-Met	23008–23079
44	<i>rrs</i>	Small ribosomal RNA	23333–24659
45	<i>nad4L</i>	NADH dehydrogenase subunit 4L	24707–25012

Of the 45 genes, 23 encoded for proteins, while 20 encoded for tRNAs and 2 encoded for the large rRNA and the small rRNA subunits which associate themselves with ribosomal proteins to form the ribosomal complex, which is where the small rRNA subunit decodes the mRNA, while the large rRNA subunit catalyzes the process of peptide bond formation during protein synthesis (Ramakrishnan, 2002). The distribution of genes in the *G. pristoides* genome is identical to the published mitochondrial genome of other *Gelidiales* as they also have a conserved compact architecture of protein-coding genes, tRNA-coding genes and rRNA-coding genes (Yang *et al.*, 2014; Boo *et al.*, 2016; Boo and Hughey, 2018). The 23 mitochondrial genes encoded separately for a mixture of enzymes and proteins (Table 3.5 and Appendix C) involved in different biological processes including oxidative phosphorylation and protein synthesis as well as other cellular processes such as protein translocation (Ng *et al.*, 2017). All the *G. pristoides* mitochondrial protein-coding genes initiated with the ATG start codon and terminated with either the TAA or TAG stop codon (Appendix C). A 41 bp overlap between the *SecY* and the *rps 12* genes was observed, and this was also the case in mitochondrial genomes of other *Gelidiales* (Yang *et al.*, 2014; Boo *et al.*, 2016; Boo and Hughey, 2018). Furthermore, most of the mitochondrial protein-coding genes of *G. pristoides* used the modified genetic code of Rhodophyta, where the codon TGA is no longer recognised as a stop codon but as a tryptophan-coding codon (Liu *et al.*, 2017).

The mitochondrial genome of *G. pristoides* bears a total of 20 mitochondrial tRNA-coding genes which are also found in published mitochondrial genomes of other *Gelidiales* (Yang *et al.*, 2014; Boo *et al.*, 2016; Boo and Hughey, 2018). As observed in the studies of Yang *et al.* (2014), Boo *et al.* (2016), Boo and Hughey (2018) and this study, mitochondrial tRNAs are the primary source of the varying number of mitochondrial genes amongst *Gelidiales*. The *G. pristoides*, *G. sclerophyllum*, *G. elegans*, *G. arborescens*, *G. galapaganse*, *G. kathyanniae* and *G. gabrielsonii* mitochondrial genomes consist of the trn-His (anticodon: GTG), which is absent in the *G. isabelae*, *G. vagum*, *G. sinicola* and *G. crinale f. luxurians* mitochondrial genomes (Yang *et al.*, 2014; Boo *et al.*, 2016; Boo and Hughey, 2018) as shown in Table 3.6. In *G. pristoides* the trn-His (GTG) gene overlaps with the *sdh2* gene with one codon, as indicated in Table 3.5. The *G. pristoides* mitochondrial genome bear the tRNA-Gly (anticodon: TCC) also present in the mitochondrial genome of *G. crinale f. luxurians*, *G. kathyanniae* and *G. gabrielsonii* (Boo *et al.*, 2016; Boo and Hughey, 2018) but absent in other *Gelidiales* (Boo *et al.*, 2016). The tRNA-Gly is sandwiched between the ATP synthase β subunit and tRNA-Glu in the *G. pristoides*, *G. crinale f. luxurians*, *G. kathyanniae* and *G. gabrielsonii* (Boo and Hughey, 2018). As with other mitochondrial genomes, the *G. pristoides* lack other tRNAs such as tRNA-Ile, tRNA-Thr, and tRNA-Tyr. This suggested that the mitochondrial genome of *G. pristoides* and others require some tRNA import for a complete translation of its mitochondrial proteins.

Table 3. 6: Comparison of tRNAs in the mitochondrial genomes of some *Gelidiales*.

tRNA	Anticodon	G.pri	G.arb	G.ele	G.vag	G.sin	G.isa	G.scl	G.gal	G.cri	G.kat	G.gab
trn-Asp	GTC	+	+	+	+	+	+	+	+	+	+	+
trn-Gly	TCC	+	-	-	-	-	-	-	-	+	+	+
trn-Gln	TTG	+	+	+	+	+	+	+	+	+	+	+
trn-Lue	TAA	+	+	+	+	+	+	+	+	+	+	+
trn-Lue	TAG	+	+	+	+	+	+	+	+	+	+	+
Trn-Gly	GCC	+	+	+	+	+	+	+	+	+	+	+
trn-His	GTG	+	+	+	-	-	-	+	+	-	+	+
trn-Phe	GAA	+	+	+	+	+	+	+	+	+	+	+
trn-ser	TGA	+	+	+	+	+	+	+	+	+	+	+
trn-Pro	TGG	+	+	+	+	+	+	+	+	+	+	+
trn-Cys	GCA	+	+	+	+	+	+	+	+	+	+	+
trn-Met	CAT	+	+	+	+	+	+	+	+	+	+	+
trn-Sup	TCA	+	+	+	+	+	+	+	+	+	+	+
trn-Ala	TGC	+	+	+	+	+	+	+	+	+	+	+
trn-Asn	GTT	+	+	+	+	+	+	+	+	+	+	+
trn-Val	TAC	+	+	+	+	+	+	+	+	+	+	+
trn-Arg	ACG	+	+	+	+	+	+	+	+	+	+	+
trn-Lys	TTT	+	+	+	+	+	+	+	+	+	+	+
trn-Glu	TTC	+	+	+	+	+	+	+	+	+	+	+
trn-Met	CAT	+	+	+	+	+	+	+	+	+	+	+

A (+) indicates the presence and a (-) indicates the absence of a particular tRNA in a species. G.pri: *G. pristoides*, G.arb: *G. arborescens*, G.ele: *G. elegans*, G.vag: *G. vagum*, G.gal: *G. galapagense*, G.sin: *G. sinicola*, G.isa: *G. isabelae*, G.scl: *G. sclerophyllum*, G.cri: *G. crinale f. luxurians*, G.kat: *G. kathyanniae* and G.gab: *G. gabrielsonii*

3.7.2 The *Gelidium pristoides* plastid genome

The *G. pristoides* genome represented in this study is not yet complete, which is why the *G. elegans* and the *G. vagum* plastid genomes, which were used for ordering the contigs of the *G. pristoides* plastid genome, were used to estimate the total approximate number of expected plastid genes for the *G. pristoides* genome. As Figure 3.19 indicates, the partial plastid genome of *G. pristoides* was composed of approximately 38% complete genes while 26% was a share of partial genes which are due to incomplete sequencing (gaps) of the *G. pristoides* plastid genome. Compared to the *G. elegans* and the *G. vagum* plastid genomes, this study may be expanded to uncover approximately 36% of the genes encoding for a variety of proteins as well as RNAs to make a total of approximately 234 plastid genes, as seen in the *G. elegans* and *G. vagum* plastid genomes (Lee *et al.*, 2016).

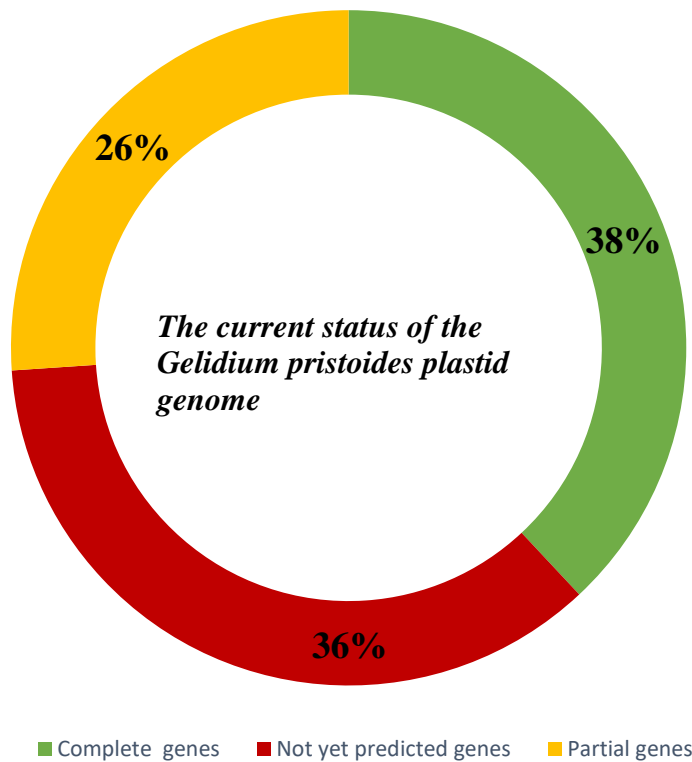


Figure 3. 19: The current status of the *Gelidium pristoides* plastid genome.

The 38% constituting the complete genes shown in Table 3.7 was distributed into 71 protein-coding genes, 18 RNA-coding genes encoding for 15 tRNAs, 2 rRNAs, and an RNaseP RNA (*rnpB* RNA). The 71 protein-coding genes fully annotated in this study are widely conserved and named genes, including the conserved hypothetical genes. A portion of the *G. elegans* and the *G.*

vagum plastid genomes (Lee *et al.*, 2016) are also constituted of the genes shown in Table 3.7. Of the 71 protein-coding genes, two were annotated as conserved ORFs, the ORF12 and the ORF64, representing conserved hypothetical proteins also found in the plastid genomes of other *Gelidiales* including the *G. elegans*, *G. vagum*, *G. kathyanniae* and *G. gabrielsonii* (Lee *et al.*, 2016; Boo and Hughey, 2018). Other hypothetical proteins annotated in this study were encoded by genes named *ycf45*, *ycf34*, *ycf63*, *ycf19*, and *ycf65*, although their function is not yet known (which is why they are termed hypothetical proteins). These hypothetical proteins are conserved as they are also found in the plastid genomes of other *Gelidium* species (Boo *et al.*, 2016; Boo and Hughey, 2018). Most of the annotated protein-coding genes of *G. pristoides* plastid (Appendix D) started with the widely used start codon ATG, while some, including ATP synthase CFO Chain subunit II, Hypothetical protein ORF64, Thioredoxin, Mg-protoporphyrin IX chelatase and the putative ribosomal protein 3 initiated with either leucine (TTG), isoleucine (ATA) or valine (GTG) codons. These alternative start codons, TTG, ATA and GTG, are also used in the plastid genomes of *Gelidiales*. The protein-coding genes, on the other hand, terminated with the three widely used stop codons, TAA, TAG and TGA, with TAA being the most widely used followed by TAG and then TGA.

This study predicted the large subunit ribosomal RNA, the small subunit ribosomal RNA and the *mnpB* RNA, commonly known as a subunit of ribonuclease P enzyme, responsible for generating mature 5' termini of tRNA through cleaving of tRNA precursors (Shevelev *et al.*, 1995). Plastid genomes such as those of *G. elegans*, *G. vagum*, *Gracilaria firma* (*G. firma*), also bear the same types of RNAs in addition to the 5S subunit ribosomal RNA (Lee *et al.*, 2016; Ng *et al.*, 2017), which we anticipate will be found when the plastid genome of *G. pristoides* is fully sequenced. However, the *G. firma* plastid genome was recently reported to bear an additional type of RNA known as the Transfer-messenger RNA (tmRNA), a unique bifunctional RNA bearing properties of both RNA and mRNA (Ng *et al.*, 2017; Keiler and Ramadoss, 2011), which was not reported in the plastid genome of *G. vagum*, *G. elegans*, *G. kathyanniae*, and *G. gabrielsonii* (Lee *et al.*, 2016; Ng *et al.*, 2017; Boo and Hughey, 2018) and is not yet predicted in the *G. pristoides* genome.

As with the plastid genomes of other organisms, the plastid genome of *G. pristoides* bears a variety of tRNAs of different lengths, carrying different anticodons specifying different amino acids, with some specifying the same amino acids. The lengths of the 15 predicted tRNAs ranged from 70–87 bp. This length range is comparable to those found in the *G. elegans*, *G. vagum*, *G. kathyanniae*,

G. gabrielsonii and *G. firma* plastid genomes (Lee *et al.*, 2016; Ng *et al.*, 2017; Boo and Hughey, 2018). Of the 15 plastid tRNAs predicted so far, 2 encoded for Glycine, 1 for Serine, 1 for Methionine, 1 for Threonine, 1 for Cysteine, 1 for Leucine, 1 for Glutamic acid, 1 for lysine, 1 for valine, 2 for Arginine, 1 for Alanine, 1 for Isoleucine, and 1 for Phenylalanine. So far, none of these tRNAs contained introns; this feature has also been observed in similar types of tRNAs found in plastid genomes of other *Gelidiales* as well as in the *G. firma* plastid genome (Lee *et al.*, 2016; Ng *et al.*, 2017; Boo and Hughey, 2018).

Table 3. 7: The complete genes predicted in the *Gelidium pristoides* plastid genome

Gene number	Gene ID	Description
1	<i>dnaK</i>	Heat Shock Protein 70
2	<i>trnG</i>	tRNA-Gly (GCC)
3	<i>Psbz</i>	Photosystem II protein Z
4	<i>psaA</i>	photosystem I P700 chlorophyll an apoprotein A1
5	<i>accB</i>	acetyl-CoA carboxylase biotin carboxyl carrier protein
6	<i>Ycf45</i>	Hypothetical protein
7	<i>acpP</i>	Acyl carrier protein
8	<i>trnS</i>	tRNA-Ser (TGA)
9	<i>psaD</i>	Photosystem I subunit II
10	<i>acsF</i>	Magnesium-protoporphyrin IX monomethyl ester oxidase cyclase
11	<i>petN</i>	Cytochrome b6/f complex subunit III
12	<i>SecG</i>	Preprotein translocase subunit G
13	<i>Ycf36</i>	hypothetical protein
14	<i>trnM</i>	trn-Met (CAT)
15	<i>BsaI</i>	Thiol-specific antioxidant protein
16	<i>pbsA</i>	Heme oxygenase
17	<i>rpl35</i>	Ribosomal protein L35
18	<i>rpl20</i>	ribosomal protein L20
19	<i>apcE</i>	Phycobilisome core- membrane linker protein
20	<i>apcA</i>	Allophycocyanin alpha subunit
21	<i>atpB</i>	ATP synthase CF1 beta subunit
22	<i>rps8</i>	Ribosomal protein S18

Gene number	Gene ID	Description
23	<i>rpl33</i>	Ribosomal protein l33
24	<i>rpoB</i>	RNA polymerase beta subunit
25	<i>rps2</i>	Ribosomal protein S2
26	<i>atp1</i>	ATP synthase CFO A chain subunit IV
27	<i>atpH</i>	ATP synthase CFO C chain, lipid-binding subunit III
28	<i>atpG</i>	ATP synthase CFO B chain Subunit II
29	ORF12	Hypothetical protein
30	<i>trnG</i>	tRNA-Gly (TCC)
31	<i>psaL</i>	Photosystem I subunit XI
32	<i>trnT</i>	tRNA-Thr (TGT)
33	ORF64	Hypothetical protein
34	<i>rpl28</i>	ribosomal protein l28
35	<i>trxA</i>	Thioredoxin
36	<i>Rbcs</i>	ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit
37	<i>trnC</i>	tRNA-Cys (GCA)
38	<i>trnL</i>	tRNA-Lue (TAA)
39	<i>Ycf34</i>	Hypothetical protein
40	<i>ilyH</i>	Acetohydroxyacid synthase small subunit
41	<i>cpeA</i>	R-phycoerythrin class I alpha subunit
42	<i>ccSA</i>	Cytochrome c biogenesis protein
43	<i>trnE</i>	tRNA-Glu (TTC)
44	<i>trpA</i>	Tryptophan synthase alpha subunit
45	<i>trnK</i>	tRNA-Lys (TTT)
46	<i>ptrB</i>	Ferredoxin-thioredoxin reductase beta subunit
47	<i>psaI</i>	Photosystem I subunit VIII
48	<i>psaJ</i>	photosystem II protein J
49	<i>psbL</i>	Photosystem II protein L
50	<i>psbf</i>	cytochrome b559 subunit
51	<i>psbE</i>	cytochrome b559 subunit alpha
52	<i>Ycf63</i>	hypothetical protein
53	<i>trnV</i>	tRNA-val(TAC)
54	<i>trnR</i>	tRNA-Arg (TCT)



Gene number	Gene ID	Description
55	<i>ChlI</i>	Mg-protoporphyrin IX chelatase
56	<i>psaM</i>	Photosystem I subunit XII
57	<i>psbV</i>	photosystem II cytochrome c550
58	<i>psbY</i>	photosystem II protein Y
59	<i>rpl22</i>	Ribosomal protein L32_rpl32
60	<i>thiG</i>	Thiamin biosynthesis protein G
61	<i>Ycf60</i>	Hypothetical protein
62	<i>rrl</i>	23S Large Subunit ribosomal RNA
63	<i>trnA</i>	tRNA-Ala (TGC)
64	<i>trnI</i>	tRNA_Ile (GAT)
65	<i>rrs</i>	16S Small subunit ribosomal RNA
66	<i>Ycf19</i>	Hypothetical protein
67	<i>Ycf65</i>	putative ribosomal protein 3
68	<i>trnR</i>	tRNA-Arg (ACG)
69	<i>psbW</i>	Photosystem II protein W
70	<i>RnpB</i>	<i>RnpB_ncRNA</i>
71	<i>trnF</i>	tRNA-Phe(GAA)
72	<i>Tufa</i>	Elongation factor Tu
73	<i>rps7</i>	Ribosomal protein S7
74	<i>rps12</i>	Ribosomal protein S12
75	<i>rpl31</i>	Ribosomal protein L31
76	<i>rps9</i>	Ribosomal protein S9
77	<i>rpl13</i>	Ribosomal protein L13
78	<i>rps13</i>	Ribosomal protein S13
79	<i>rpl36</i>	Ribosomal protein L36
80	<i>rps8</i>	Ribosomal protein S8
81	<i>rps17</i>	Ribosomal protein S17
82	<i>rpl29</i>	Ribosomal protein L29
83	<i>rpl16</i>	Ribosomal protein L16
84	<i>rps3</i>	Ribosomal protein S3
85	<i>rpl22</i>	Ribosomal protein L22
86	<i>rps19</i>	Ribosomal protein S19
87	<i>rpl2</i>	Ribosomal protein L2
88	<i>rpl4</i>	Ribosomal protein L4
89	<i>rpl3</i>	Ribosomal protein L3



University of Fort Hare
Together in Excellence

The partial genes constituting 25% of the *G. pristoides* genome were comprised of only protein-coding genes (Table 3.8), which are widely distributed across the plastid genomes of other *Gelidiales* (Lee *et al.*, 2016; Boo and Hughey, 2018). To date, the *G. pristoides* plastid gene content is very similar to that of other *Gelidiales*, with different genes involved in different biological pathways including photosynthesis, protein translocation, protein synthesis and regulation, amino acid synthesis, fatty acid synthesis (Ng *et al.*, 2017) as well as other cellular processes such as cell division and response to stresses. The 36% portion of the genome constituted of genes that are not yet predicted is expected to be a mix of RNA and protein-coding genes. This expectation arises from the pattern of gene arrangement and distribution of genes in the *G. pristoides* plastid genome, which is similar to that of the published plastid genomes of *G. vagum* and *G. elegans* (Lee *et al.*, 2016) which were used for ordering the contigs of the *G. pristoides* plastid genome.

Table 3. 8: The partial genes predicted in the *Gelidium pristoides* plastid genome

Gene number	Gene ID	Encoded protein
1	<i>psaB</i>	Photosystem I P700 chlorophyll an apoprotein A2
2	<i>preA</i>	Prenyl transferase
3	<i>odpB</i>	Pyruvate dehydrogenase E1 component beta subunit
4	<i>apcB</i>	Allophycocyanin beta subunit
5	<i>rpoCI</i>	RNA polymerase beta' subunit
6	<i>rpoC2</i>	RNA polymerase beta'' subunit
7	<i>tsf</i>	elongation factor Ts
8	<i>atpF</i>	ATP synthase CFO B chain Subunit I
9	<i>atpD</i>	ATP synthase CF1 delta subunit
10	<i>atpA</i>	ATP synthase CF1 alpha subunit
11	<i>sufC</i>	Iron-sulfur cluster formation ABC transporter ATP-binding subunit
12	<i>sufB</i>	Cysteine desulfurase activator complex subunit
13	<i>Ycf39</i>	hypothetical protein Gele_053
14	<i>cemA</i>	Chloroplast envelope membrane

Gene number	Gene ID	Encoded protein
15	<i>rbcL</i>	ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit
16	<i>cbby</i>	Putative rubisco expression protein
17	<i>apcF</i>	allophycocyanin beta 18 subunit
18	<i>infC</i>	Translation initiation factor 3
19	<i>Ycf20</i>	Hypothetical protein Gele_082
20	<i>cpcG</i>	Phycobilisome rod-core linker protein
21	<i>nblA</i>	Phycobilisome degradation protein
22	<i>cpeB</i>	R-phycoerythrin class I beta subunit
23	<i>gltB</i>	Ferredoxin-dependent glutamate synthase
24	<i>rpoZ</i>	DNA-directed RNA polymerase omega chain
25	<i>cpcB</i>	Phycocyanin beta subunit
26	<i>SecA</i>	preprotein translocase subunit
27	<i>trls</i>	tRNA-lysine synthase
28	<i>Ycf26</i>	hypothetical protein Gele_112
29	<i>Ycf46</i>	Hypothetical protein Gele_118
30	<i>fabH</i>	3-oxoacyl-acyl-carrier-protein synthase
31	<i>accD</i>	Acetyl-CoA carboxylase beta subunit
32	<i>petJ</i>	Cytochrome c553
33	<i>CarA</i>	Carbamoyl-phosphate synthase arginine-specific small subunit
34	<i>Ycf55</i>	hypothetical protein Gele_133
35	<i>Ycf56</i>	hypothetical protein
36	<i>rne</i>	Ribonuclease E
37	<i>psbA</i>	Photosystem II protein D1
38	<i>ccdA</i>	cytochrome c biogenesis protein transmembrane
39	<i>ccSI</i>	C-type cytochrome biogenesis protein

Gene number	Gene ID	Encoded protein
40	<i>thiG</i>	Thiamin biosynthesis protein G
41	<i>psbD</i>	Photosystem II protein D2
42	<i>psbC</i>	photosystem II CP43 protein
43	<i>rps16</i>	Ribosomal protein S16
44	<i>groEL</i>	Chaperonin
45	<i>syh</i>	Histidine-tRNA synthetase
46	<i>rps1</i>	Ribosomal protein S1
47	<i>petB</i>	Cytochrome b6
48	<i>petD</i>	Cytochrome b6, Cytochrome b6/f complex subunit IV
49	<i>rpl1</i>	Ribosomal protein L1
50	<i>rpl11</i>	Ribosomal protein L11
51	<i>dnaB</i>	Replication helicase subunit
52	<i>clpC</i>	Clp protease ATP binding subunit
53	<i>ilvB</i>	acetoxyacid synthase large subunit
54	<i>argB</i>	Acetylglutamate kinase
55	<i>ftsH</i>	Cell division protein
56	<i>psbB</i>	Photosystem II CP47 protein
57	<i>Ycf38</i>	ABC-2-type-transporter
58	<i>rps10</i>	Ribosomal protein S10
59	<i>rpoA</i>	DNA-directed RNA polymerase alpha subunit
60	<i>secY</i>	Preprotein translocase subunit <i>SecY</i>
61	<i>rpl14</i>	Ribosomal protein L14

3.8 Construction of 3D model structure for the mitochondrial Cox3 and plastid HSP70

The 3D structure of a protein is of undeniable significance in a variety of studies, including protein function, protein dynamics, and protein-ligand interaction as well as protein-protein interaction studies (<https://proteinstructures.com/Modeling/homology-modeling.html>, accessed 05/11/2018). Each protein's three-dimensional (3D) structure is claimed to be a reflection of its function (Cao *et al.*, 2005). Techniques dealing with construction and analysis of 3D structures are of paramount significance. Time-consuming experimental methods such as X-ray crystallography, NMR spectroscopy and electron microscopy have been used in the past for solving 3D structures of proteins. The incremental increase of protein sequences in genomic databases demanded faster and

more accurate methods focusing on solving as many 3D structures of proteins as possible, which led to the homology modelling method (also known as comparative modelling) (Floudas, 2007; Bishop *et al.*, 2008). Homology modelling enables the computational prediction of protein structures by utilising the existing experimental-determined protein structures (templates) deposited in protein databases such as the Protein Data Bank database. Technically, homology modelling predicts the 3D protein structure based on the similarity of the protein sequence of the protein of interest (target) to existing proteins of known structures in the databases (Bishop *et al.*, 2008).

In this study, the homology modelling approach was used for prediction of the 3D model structures of the mitochondrial *Cox3* and the plastid HSP70. *Cox3*, together with *CoxI* and *Cox2*, forms the catalytic core of the Cytochrome c oxidase, a biogenomic complex that catalyses the rate-limiting step of the energy-generating mitochondrial electron transport chain (ETC) (Srinivasan and Avadhani, 2012). The *Cox3* enzymatic subunit plays a crucial role in the assembly and stabilisation of the Cytochrome c oxidase complex (Wilson and Prochaska, 1990). It is also involved in proton translocation in the ETC process (Wu *et al.*, 1995). Heat Shock Proteins (HSPs), including HSP70 protein, play a crucial role in folding and unfolding of proteins, assembly of multiprotein complexes, cell-cycle control and signalling and apoptosis. They also protect cells against stress originating from a broad range of unfavourable biotic and abiotic conditions, including rising temperatures causing protein misfolding and protein aggregation leading to loss of biological functions (Morimoto, 1998; Beere, 2004; Kumar *et al.*, 2016). HSP70 can keep the cellular environment clean as it degrades proteins that are no longer needed by the cell (Bercovich *et al.*, 1997; Chiang *et al.*, 1989). The HSP70 has a wide range of medical benefits, including its role in Huntington disease therapeutics and preventing neurodegenerative diseases such as Parkinson's disease and Alzheimer's disease. HSP70 has also been reported to protect the heart and assist in the prevention of diabetes (Kurucz *et al.*, 2002; Padmalayam, 2014).

Conservation of 3D structures of proteins makes good indicators of functional similarity between proteins. As seen in Figure 3.20, the constructed 3D model structure of the *Cox3* protein is similar to that of the *Cox3* 3D protein structure of *Bos taurus* used as a template. The Plastid HSP70 3D model structure represented in Figure 3.21 is also similar to the HSP70 from the *Escherichia coli*, used as the template during homology modelling. The similarity between the templates and the

targets (*G. pristoides Cox3* and HSP70) suggested that these proteins perform the same functions in different organisms. It also indicated the conservation of *Cox3* and the HSP70 proteins in different species and further gave further credence that the predicted protein sequences are *Cox3* and HSP70.

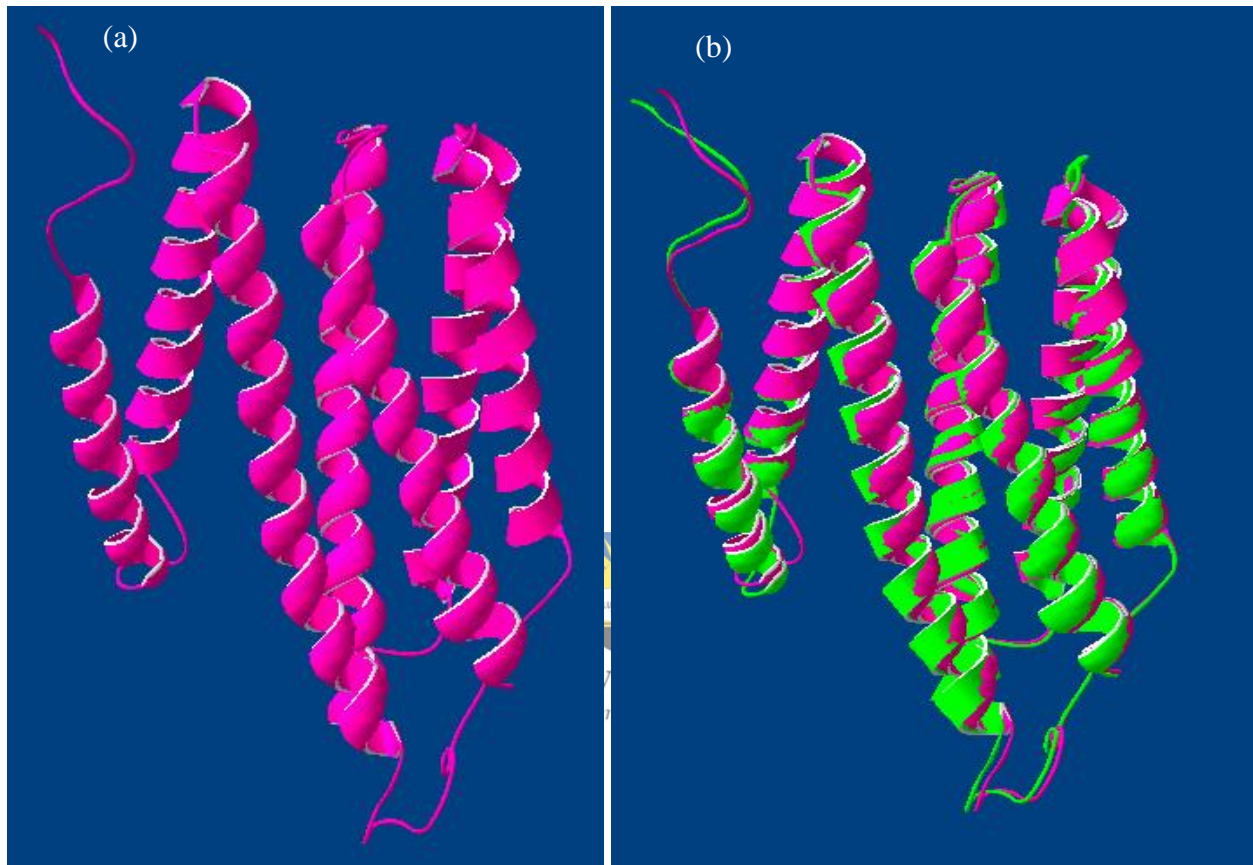


Figure 3. 20: Three-dimensional model structure of the *Cox3* protein standing alone (a) and fitted into the *Bos taurus Cox3* protein (b). The 3D model structure of *Gelidium pristoides Cox3* protein was constructed with the PRIMO webserver pipeline.

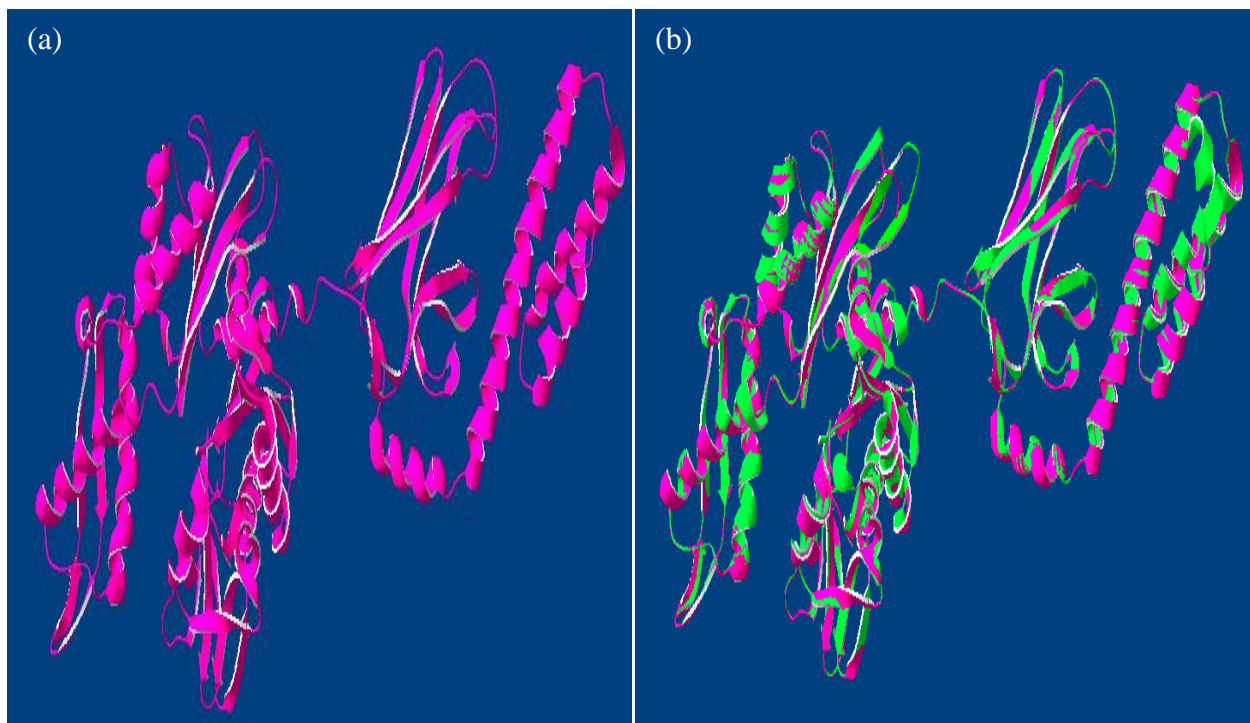


Figure 3. 21: Three-dimensional model structure of the HSP70 (a) and alignment of the HSP70 constructed a model with *Escherichia coli* HSP70 protein (b). The 3D model structure of *Gelidium pristoides* HSP70 was constructed with the SWISS-MODEL webserver pipeline.



University of Fort Hare
Together in Excellence

3.9 Evaluation of the Cox3 and HSP70 3D model quality

The accuracy of the constructed 3D model structure of a protein evaluated through analysis of the model quality is vital in homology modelling, as the models are not experimentally determined structures, but rather predictions obtained from experimentally functional assigned proteins (Kryshtafovych and Fidelis, 2009). According to published literature, for a model to be of good quality, at least 90% of the residues must fall within the most favoured region in the Ramachandran plot (Laskowski *et al.*, 1993; Hatherley *et al.*, 2016). The constructed 3D model structures of both the *Cox3* and the HSP70 proteins were of good quality, as 94% and 95.03% of their amino acid residues respectively fell within the most favoured regions, indicated by the red and green shades in the Ramachandran plots in Figures 3.22 and 3.23 respectively.

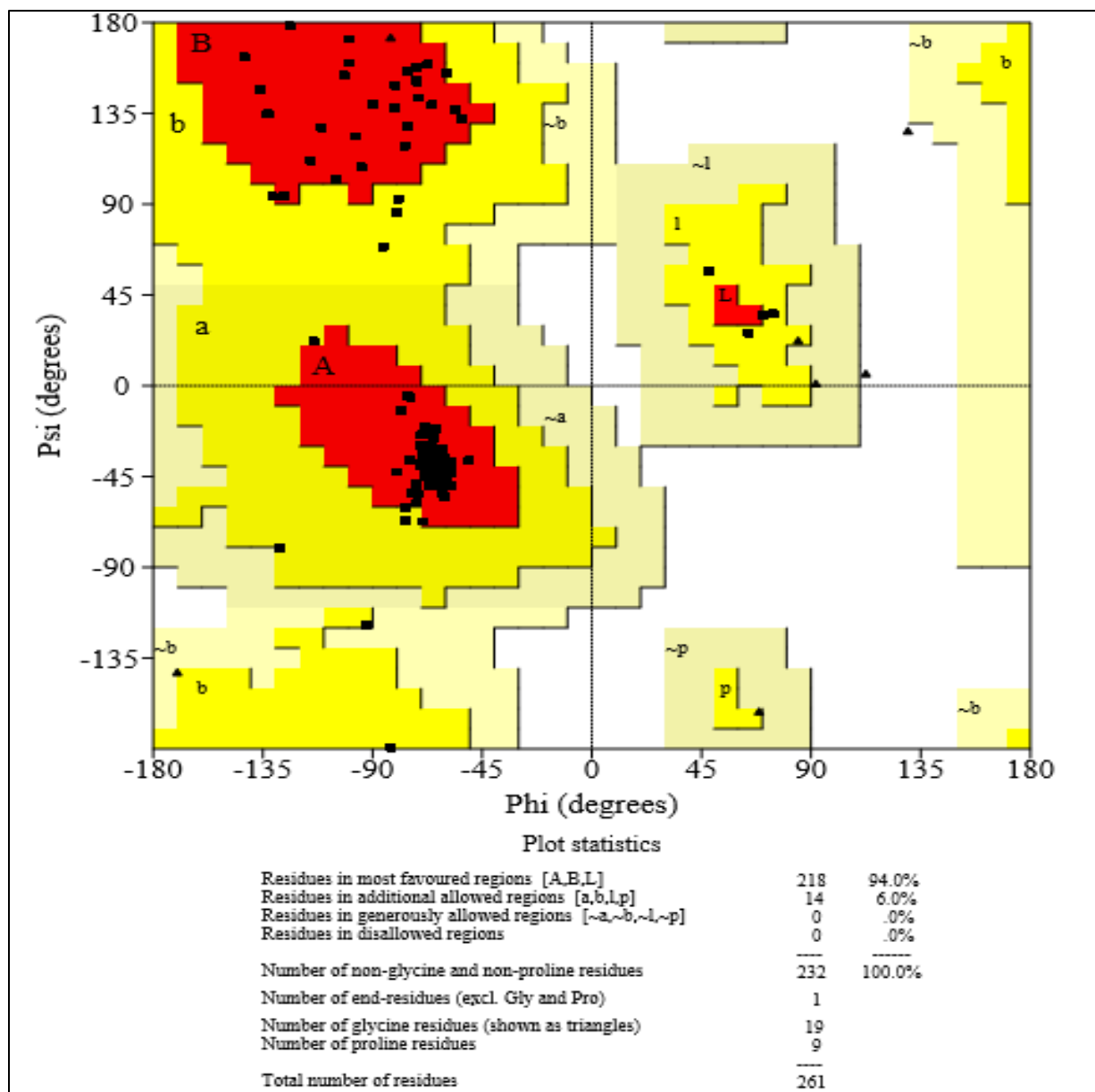


Figure 3. 22: Ramachandran plot obtained from PROCHECK evaluation of the PRIMO constructed *Gelidium pristoides* Cox3 3D model structure.

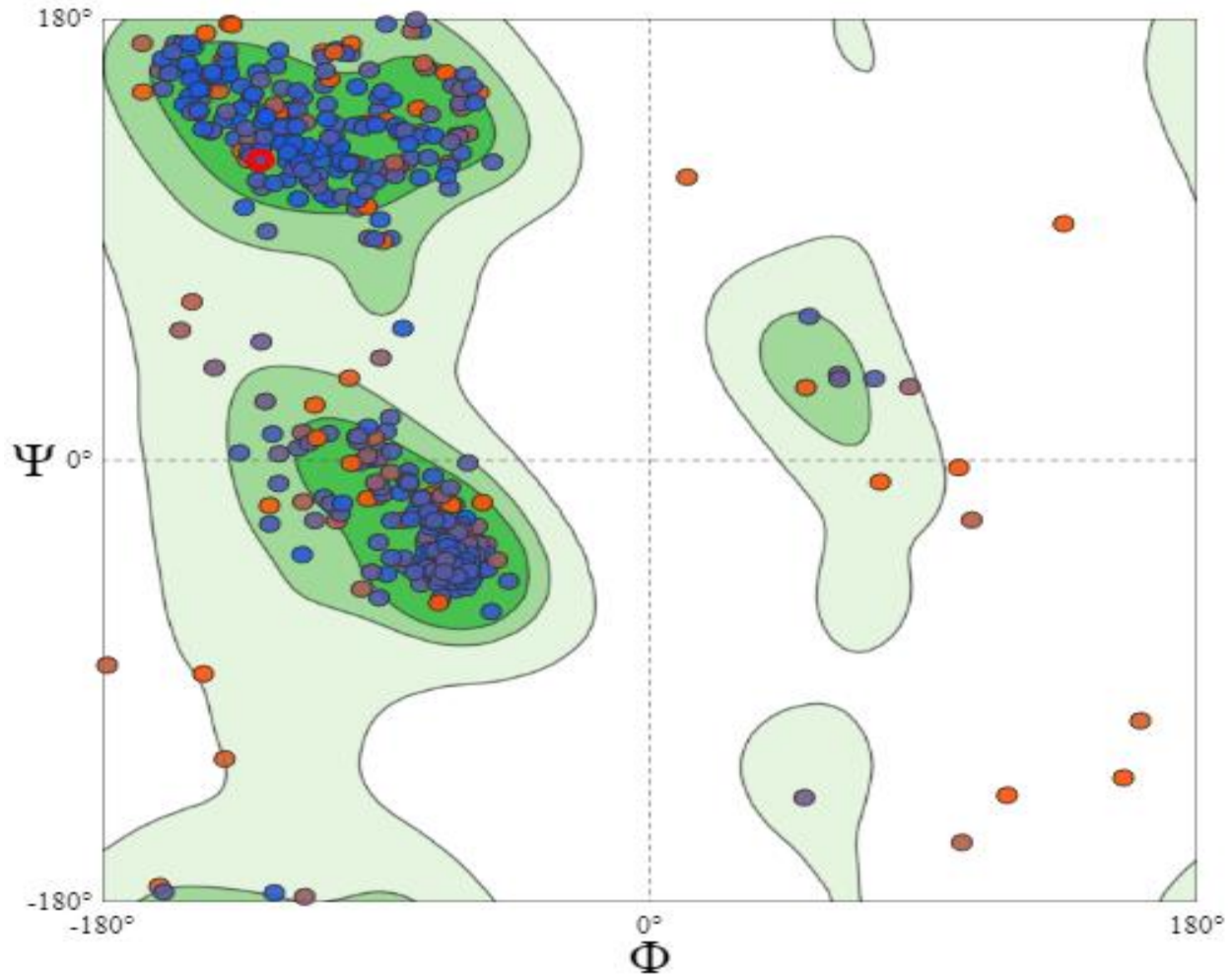


Figure 3. 23: Ramachandran plot obtained from the evaluation of the SWISS-MODEL constructed *Gelidium pristoides* HSP70 3D model structure.

3.10 Phylogeny analysis of the mitochondrial Cox3 and Plastid HSP70 proteins

Multiple sequence alignment during phylogenetic analysis is crucial, as it aims to model changes occurring over time and thus derive the evolutionary relationships between sequences (Ebenezer and Priyakumari, 2017). Comparison of the protein sequence through multiple alignments affords us a clear view of conserved protein regions between different species (Thompson, 2004). As indicated by the asterisks in Figures 3.24 and 3.25, the *Cox3* and the HSP70 are highly conserved in the different species used in this study. The conserved regions of the protein sequence are known to be most resistant to evolutionary changes, as they are significant for the structure and function of the protein (Sitbon and Pietrovski, 2007).



Figure 3. 25: Portion of the sequence alignment of the HSP70 protein from eleven Rhodophyta species. *G. pristoides* (MG_111111111.1), *G. vagum* (YP_009244224.1), *G. gabrielsonii*(AYO27611.1), *G. kathyanniae* (AYO27834.1), *G. elegans* (YP_009244024.1), *S. flabellate* (YP_009296153.1), *G. tenuifrons* (AXF36120.1), *G. longissima* (YP_009511292.1), *Schimmelmannia schousboei* (*S. schousboei*) (YP_009295747.1), *G. taiwanensis* (YP_008144853.1) and *G. filicina* (YP_009488724.1) produced by the Clustal Omega software. “*” = positions have a single, fully conserved region, “:” = conserved region between groups of strongly similar properties, “.” The conserved region between groups of weakly similar properties

Based on the percentage identity matrix in Table 3.9, the similarity between the *Cox3* protein sequences ranged from 84% to 100%. This suggested that there were fewer than 20% differences between the sequences. According to this matrix, the *G. pristoides Cox3* protein (MG_0000000001.1) is 84.19% similar to *G. filicina*, 84.19% similar to *G. taiwanensis*, 87.50% similar to *P. musciformis*, 90.81% similar to *G. vagum*, 93.75% similar to *G. sinicola*, 92.65% similar to *G. galapaganse*, 94.85% similar to *G.arborescens*, 92.65% similar to *G. isabelae*, 92.28% similar to *G. sclerophyllum* and 93.75% similar to *G. elegans Cox3* protein sequences. Based on this statistical analysis, the *G. pristoides Cox3* is most similar to the *G.arborescens Cox3* protein.

Table 3. 9: Percentage identity matrix created by Clustal 12.1 during alignment of the *Cox3* proteins from different species

# Percent Identity Matrix - created by Clustal2.1												
1:	YP_009488809.1	100.00	98.53	84.56	85.29	85.66	85.66	84.19	84.93	86.03	86.40	86.76
2:	AIY34289.1	98.53	100.00	84.93	85.29	85.29	85.66	84.19	84.93	85.66	86.03	86.40
3:	YP_009317628.1	84.56	84.93	100.00	87.87	88.97	88.97	87.50	86.76	86.03	85.66	86.40
4:	YP_008963198.1	85.29	85.29	87.87	100.00	90.81	90.44	90.81	90.81	90.07	90.07	90.81
5:	YP_009317559.1	85.66	85.29	88.97	90.81	100.00	95.96	93.75	94.49	93.38	92.65	93.75
6:	YP_009317490.1	85.66	85.66	88.97	90.44	95.96	100.00	92.65	93.75	93.01	93.01	94.12
7:	MG_0000000001.1	84.19	84.19	87.50	90.81	93.75	92.65	100.00	94.85	92.65	92.28	93.75
8:	YP_009317467.1	84.93	84.93	86.76	90.81	94.49	93.75	94.85	100.00	95.59	95.22	97.43
9:	YP_009317513.1	86.03	85.66	86.03	90.07	93.38	93.01	92.65	95.59	100.00	96.32	97.79
10:	YP_009317536.1	86.40	86.03	85.66	90.07	92.65	93.01	92.28	95.22	96.32	100.00	97.79
11:	YP_009114019.1	86.76	86.40	86.40	90.81	93.75	94.12	93.75	97.43	97.79	97.79	100.00

Highlight the *Cox3* of *G.pristoides*

With regard to the HSP70, the percentage identity matrix (Table 3.10) indicated that the HSP70 from *G. pristoides* is 94% identical to *G. vagum*, 95.51% similar *G. gabrielsonii*, 95.51% similar to *G. kathyanniae*, 96.63% similar to *G. elegans*, 87.12% similar to *S. flabellate*, 87.22% similar to *G. tenuifrons*, 86.89% similar to *G. longissima*, 86.41% similar to *S. schousboei*, 86.15% similar to *G. taiwanensis* and 86.15% similar to *G. filicina* HSP70 protein sequences. Based on this statistical analysis, the *G. pristoides* HSP70 is most similar to the *G. elegans* HSP70.

Table 3. 10: Percentage identity matrix created by Clustal 12.1 during alignment of the HSP70 proteins from different species

```
# Percent Identity Matrix - created by Clustal2.1
##
##
1: YP_009244224.1 100.00 96.63 97.11 94.70 95.35 87.44 86.41 86.08 87.54 87.28 87.28
2: AY027611.1 96.63 100.00 98.56 95.51 95.51 87.44 86.73 86.73 88.03 87.12 86.96
3: AY027834.1 97.11 98.56 100.00 95.51 95.67 87.76 87.22 86.89 88.03 87.44 87.44
4: MG_111111111.1 94.70 95.51 95.51 100.00 96.63 87.12 87.22 86.89 86.41 86.31 86.15
5: YP_009244024.1 95.35 95.51 95.67 96.63 100.00 87.60 87.22 86.89 86.73 86.80 86.63
6: YP_009296153.1 87.44 87.44 87.76 87.12 87.60 100.00 88.06 87.74 88.35 87.64 87.96
7: AXF36120.1 86.41 86.73 87.22 87.22 87.22 88.06 100.00 98.87 89.48 87.74 87.58
8: YP_009511292.1 86.08 86.73 86.89 86.89 86.89 87.74 98.87 100.00 89.48 87.58 87.42
9: YP_009295747.1 87.54 88.03 88.03 86.41 86.73 88.35 89.48 89.48 100.00 89.97 89.32
10: YP_008144853.1 87.28 87.12 87.44 86.31 86.80 87.64 87.74 87.58 89.97 100.00 98.39
11: YP_009488724.1 87.28 86.96 87.44 86.15 86.63 87.96 87.58 87.42 89.32 98.39 100.00
```

Highlight the HSP70 of *G.pristoides*

The cladogram tree (Figure 3.26), which was constructed from the 11 *Cox3* species of *G. isabelae*, *G. arbarescens*, *G. pristoides*, *G. vagum*, *P. musciformis*, *G. filicina*, *G. sinicola*, *G. galapaganse*, *G. taiwanensis*, *G. sclerophyllum*, and *G. elegans*, indicated that the *Cox3* protein from *G. pristoides* (MG_0000000001.1) is most closely related to the *G. arbarescens* *Cox3* protein (YP_009317467.1). This observation is per the statistical analysis represented by the percentage identify matrix in Table 3.9.

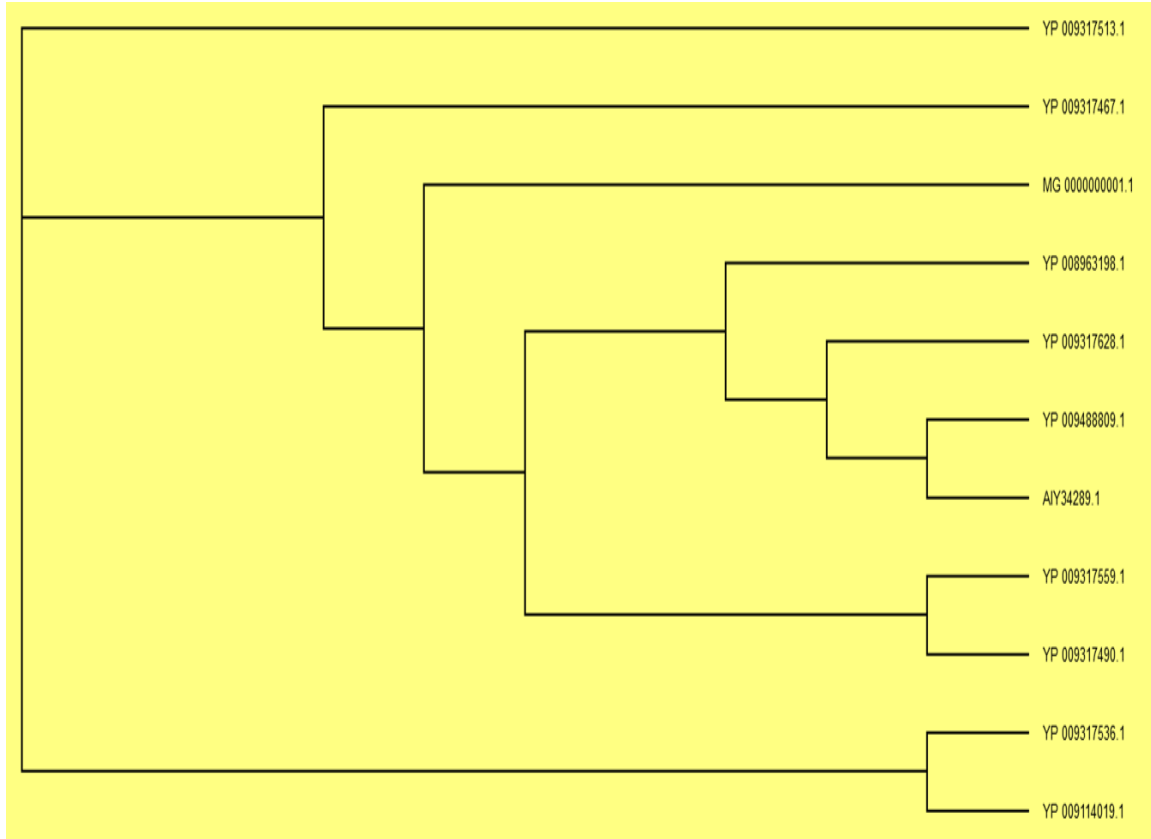


Figure 3. 26: Cladogram tree of the *Cox3* protein from eleven Rhodophyta species. *G. isabelae* (YP_009317513.1), *G. arbarescens* (YP_009317467.1), *G. pristoides* (MG_0000000001.1), *G. vagum* (YP_008963198.1), *P. musciformis* (YP_009317628.1), *G. filicina* (YP_009488809.1), *G. sinicola* (YP_009317559.1), *G. galapaganse* (YP_009317490.1), *G. taiwanensis* (AIY34289.1), *G. sclerophyllum* (YP_009317536.1), and *G. elegans* (YP_009114019.1).

The cladogram tree (Figure 3. 27) of the plastid HSP70 constructed from the 11 HSP70 protein sequences of *G. pristoides*, *G. vagum*, *G. gabrielsonii*, *G. kathyanniae*, *G. elegans*, *S. flabellate*, *G. tenuifrons*, *G. longissima*, *S. schousboei*, *G. taiwanensis* and *G. filicina* indicated that the *G. pristoides* HSP70 is most closely related to the *G. elegans* HSP70 (YP_009244024.1), as they are sister descendants. This observation is per the statistical analysis represented by the percentage identity matrix (Table 3.10).

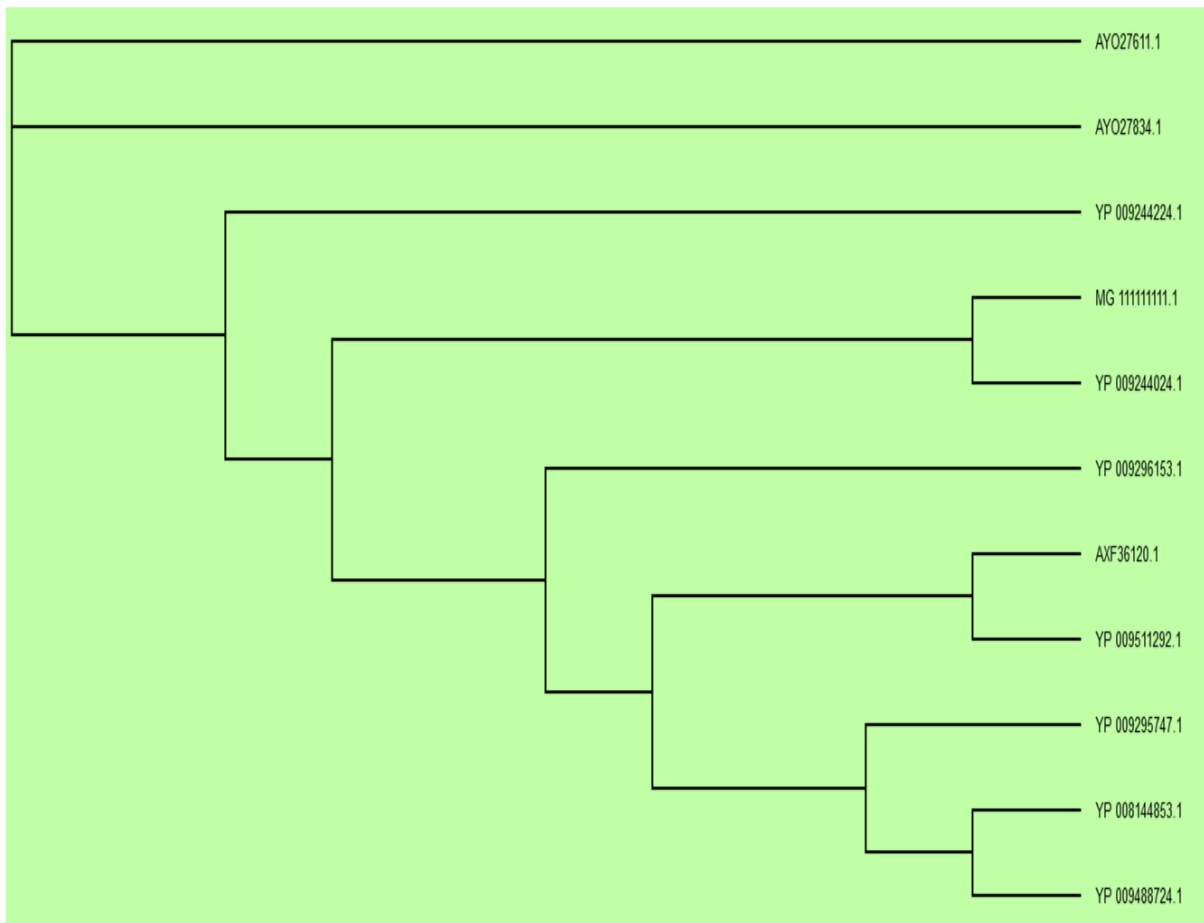


Figure 3. 27: Cladogram tree of the HSP70 from eleven Rhodophyta species. *G. pristoides*, *G. vagum*, *G. gabrielsonii*, *G. kathyanniae*, *G. elegans* (YP_009244024.1), *S. flabellate*, *G. tenuifrons*, *G. longissima*, *S. schousboei*, *G. taiwanensis* and *G. filicina*

CHAPTER FOUR

General Discussion, Conclusion and Future Prospects



University of Fort Hare
Together in Excellence

4 CHAPTER FOUR: GENERAL DISCUSSION, CONCLUSION AND FUTURE PROSPECTS

4.1 General discussion

4.1.1 Introduction

Sequencing is a widely used approach to uncover the constituents of genomes (Angeleska *et al.*, 2014). It is an approach that is not limited to sequencing specific organisms but has broad application to any species that triggers genomic interest to a researcher. A thorough literature search indicated that, in South Africa, genome sequencing of the economically significant Rhodophyta species had been overlooked as no published studies are focusing on either whole genome sequencing or the sequencing of organellar genomes of such species. This study, therefore, sequenced the genomes of the economically significant *G. pristoides* species endemic to South Africa, as genomic information plays a crucial role in a broad array of applications, including maintenance of biodiversity, identification of genes of biotechnological and medicinal significance as well as evolutionary studies (Costa *et al.*, 2016). This study affords the scientific community a sound foundation for future research projects ranging from extraction of genes of interest, manipulation of genes to fit relevant aspects of life such as drug design, as well as a reference point for sequencing of other genomes. Currently, this study presents a completely annotated mitochondrial genome and a partially annotated plastid genome of *G. pristoides*, which are potential reservoirs of information that can contribute to insights across different scientific applications, including eukaryotic evolution and molecular biology studies to express recombinant proteins of medical or industrial importance.

4.1.2 Rhodophyta mitochondrial genomes

Mitochondrial genomes of Rhodophyta species are generally known for their small compact sizes ranging from 24–40 kb (Leblanc *et al.*, 1997; Seckbach and Chapman 2010; Yang *et al.*, 2014; Boo *et al.*, 2016, Boo and Hughey 2018). This study and those of Boo *et al.*, (2016) and Boo and Hughey (2018) indicate that the sequenced mitochondrial genomes of *Gelidiales* sit in the lower limit of the size range of the Rhodophyta mitochondrial genome sizes, as they range from 24–25 kb. Despite the varying sizes amongst the mitochondrial genomes of *Gelidiales*, a conserved

architecture of a mixture of protein-coding, tRNA-encoding and rRNA-coding genes is maintained. The variations observed in the number of genes constituting this conserved architecture in *Gelidiales* are primarily due to tRNA genes which are present and absent from other species. Also, the length of the mitochondrial genome does not necessarily correspond to the total number of genes that it bears. For example, *G. galapaganse* has the most extended mitochondrial genome amongst the mitochondrial genomes published by Boo *et al.*, (2016) but bears only 43 genes. On the other hand, *G. crinale* and *f. luxurians* have the second smallest mitochondrial genome but bear 44 genes (Boo *et al.*, 2016). As a result, one can conclude that the differences in lengths are primarily due to varying sizes of intergenic regions as well as varying sizes of the genes.

Studies by Boo *et al.*, (2016), and Boo and Hughey (2018), as well as this study further, indicate that mitochondrial genomes of species of the *Gelidium* genus are constituted of proteins of known functions since no hypothetical proteins have been reported so far. These proteins are involved in different biological pathways including protein translocation, protein synthesis and regulation, as well as ETC/oxidative phosphorylation which is the energy-producing process, rendering mitochondria the powerhouse of the cell (Ng *et al.*, 2017). NADH dehydrogenases, Succinate dehydrogenase, Cytochrome C oxidases proteins and ATP synthase proteins are involved in ETC/oxidative phosphorylation (Ng *et al.*, 2017) while the preprotein translocase (*SecY*) is involved in protein translocation (Kanehisa *et al.*, 2016) and the ribosomal proteins together with rRNAs constitute the ribosomal complex (Beck, 2018).

It is evident that the mitochondria of *Gelidiales* require import of some cytosolic tRNA (Schneider and Maréchal-Drouard, 2000; Schneider, 2011) in order for them to completely translate their proteins, as they are short of a variable number of tRNAs whose amino acids are found in translated proteins. Studies by Schneider and Maréchal-Drouard, (2000) and Schneider, (2011), suggested that this phenomenon of missing tRNAs is broadly observed in mitochondrial genomes of eukaryotes and is compensated by cytosolic tRNA import. This phenomenon points to an evolutionary change that resulted in the loss of some mitochondrial tRNA (Wallin, 1927; Chuan, 2014) as mitochondrial genomes originate from free-living bacteria where there is no tRNA import. The mitochondrial genomes of some land plants, Chlorophytes, Heterokatophytes and Rhodophytes adopt a modified genetic code except for *C. merolae* and *Cyanidium caldarium* (Ohta

et al., 1998). Osawa *et al.* (1992) suggested that this modification of stop codon to tryptophan is generally correlated with A+T pressure in these mitochondrial genomes, which was also observed in the *G. pristoides* mitochondrial genome.

4.1.3 Rhodophyta plastid genomes

Studies of Rhodophyta plastid genomes have revealed that these species possess plastid genomes that are larger than those of other seaweeds. The *C. merolae* bear a plastid genome of 149987 nucleotides (Ohta *et al.*, 2003) while that of *Porphyridium sordidum* is as large as 259000 nucleotides (Lee *et al.*, 2016). The largest plastid genome reported so far consists of 1127000 nucleotides (Muñoz-Gómez *et al.* 2017). Plastid genomes of the *Florideophyceae* Rhodophyta class have a reported size range of 175000–194000 nucleotides (Lee *et al.* 2016; Boo and Hughey, 2018), while those of the currently sequenced *Gelidiales* range from 174748 in *G. elegans* to 179853 in *G. vagum* (Lee *et al.*, 2016). Rhodophyta plastid genomes are generally known for their high gene content and highly conserved architectures (Janouškovec *et al.*, 2013) with genes coding for proteins, tRNAs and rRNAs genes (Lee *et al.* 2016). Primarily, the gene-rich characteristic of Rhodophyta plastid genomes is due to their short intergenic spacers, large gene clusters, complete or partial lack of introns, and overlapping genes (Giardi and Piletska, 2006). The total number of genes within Rhodophyta plastid genomes varies even amongst species of the same taxonomic class. For instance, *G. firma* has a total of 252 genes while the *G. elegans* and *G. vagum* share a total of 234 plastid genes (Yang *et al.*, 2014; Ng *et al.*, 2017; 2014, Boo *et al.*, 2016).

Although the plastid genome presented in this study is not yet complete, the pattern of gene arrangement and distribution observed suggested that the *G. pristoides* genome possesses the typical conserved architecture of plastid genomes of Rhodophytes. Based on the genes annotated so far, the plastid genome of *G. pristoides* bear genes that are widely found in other sequenced plastid genomes of species of the *Gelidium* genus. Moreover, the Rhodophyta plastid genomes, including one of *G. pristoides*, contain some conserved hypothetical proteins whose functions are not yet known. This implies that the Rhodophyta plastid genomes afford researchers future studies in predicting the exact functions of the hypothetical proteins. Previous studies suggested that the genes shared amongst the Rhodophyta plastid genomes are ones believed to be essential for the basic functioning of Rhodophyta plastid genomes and offer a good reference point for the

annotation of other plastid genomes (DePriest *et al.*, 2013). A study by Janouškovec *et al.*, (2013) demonstrated that Rhodophyta plastid genomes are the potential source for resolving Rhodophyta relationships. They contain a range of biological markers which are useful for species barcoding (Janouškovec *et al.*, 2013). They are a rich source of proteins with medicinal and molecular applications. Proteins such as ribonuclease are used in molecular biology applications while the HSP70 protein plays a crucial role in the prevention and treatment of neurodegenerative and cardiovascular diseases and diabetes (Kurucz *et al.*, 2002; Padmalayam, 2014).

4.2 Conclusion

Based on the available published literature, this dissertation presents the first study focusing on genome sequencing, assembly and annotation of *G. pristoides*, an endemic South African economically important red algae species. Currently, the genomic information of *G. pristoides* from Kenton-On-Sea is represented in the form of a completely annotated mitochondrial and a partial plastid genome. The unsequenced portions (gaps) of the partial plastid genome constituted of genes that are not yet predicted is expected to be a mix of RNA, protein-coding genes and non-coding regions. The study of the organellar genomes of *G. pristoides* provides broad information about their constituents and their arrangement. The current findings and discussions laid out in this study imply that the information gathered in the study of *G. pristoides* organellar genomes can be used in a broad array of scientific disciplines. It provides an essential foundation for future research projects in Biotechnology, Molecular biology and Bioinformatics. This study is a reference point for the development of future projects.

4.3 Future prospects

Future studies will focus on the production of the completely annotated plastid and nuclear genomes of *G. pristoides*. Objective Five of this study (Chapter 1), which was to recombinantly express, purify and confirm the function/activity of a selected protein, as predicted in Objective Four (Chapter 1), will also be done in the future. This study allows for future genome-genome comparisons of the completely annotated mitochondrial genomes. Since this study serves as the foundation study, it opens opportunities for many research projects in Biotechnology, Molecular Biology, Genetics, Genomics and Bioinformatics of the Rhodophyta family.

CHAPTER FIVE

References of the study




University of Fort Hare
Together in Excellence

5 CHAPTER FIVE: REFERENCES OF THE STUDY

- Aigrain, L., Gu, Y. and Quail, M.A., 2016. Quantitation of next generation sequencing library preparation protocol efficiencies using droplet digital PCR assays—a systematic comparison of DNA library preparation kits for Illumina sequencing. *BMC genomics*, 17(1), p.458.
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P., 2002. Studying gene expression and function. In *Molecular Biology of the Cell. 4th edition*. Garland Science.
- Alhakami, H., Mirebrahim, H. and Lonardi, S., 2017. A comparative evaluation of genome assembly reconciliation tools. *Genome biology*, 18(1), p.93.
- Alkan, C., Sajjadian, S. and Eichler, E.E., 2010. Limitations of next-generation genome sequence assembly. *Nature methods*, 8(1), p.61.
- Al-Khayri, J.M., Jain, S.M. and Johnson, D.V. eds., 2015. *Advances in plant breeding strategies: Breeding, biotechnology and molecular tools*. Springer International Publishing
- Altschul, S.F. and Koonin, E.V., 1998. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends in biochemical sciences*, 23(11), pp.444–447.
- Anderson, R.J. and Rothman M., 2013. *Description of the commercial seaweed sector*. https://www.nda.agric.za/doiDev/sideMenu/fisheries/03_areasofwork/Resources%20Research/Seaweed%20main%20web%20page.pdf
- Anderson, R.J., Simons, R.H., Jarman, N.G. and Levitt, G.J., 1991. *Gelidium pristoides* in South Africa. In *International Workshop on Gelidium* (pp. 55–66). Springer, Dordrecht.
- Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. *Babraham Bioinformatics. Online [Mar 2016]*.
- Angeleska, A., Kleessen, S. and Nikoloski, Z., 2014. The Sequence Reconstruction Problem. In *Discrete and Topological Models in Molecular Biology* (pp. 23-43). Springer, Berlin, Heidelberg.
- Angelini, C., Rancoita, P.M. and Rovetta, S. eds., 2016. *Computational Intelligence Methods for Bioinformatics and Biostatistics: 12th International Meeting, CIBB 2015, Naples, Italy, September 10-12, 2015, Revised Selected Papers* (Vol. 9874). Springer.
- Ari, Ş. and Arikan, M., 2016. Next-generation sequencing: advantages, disadvantages, and future. In *Plant Omics: Trends and Applications* (pp. 109-135). Springer, Cham.
- Baker, M., 2012. *De novo* genome assembly: what every biologist should know. *Nature America*, 9(4), pp.333-337.
- Beck K., 2018. The Structure and Function of Ribosomes in Eukaryotes & Prokaryotes. <https://sciencing.com/structure-function-ribosomes-eukaryotes-prokaryotes-20173.html> accessed 24/03/2019
- Beere, H.M., 2004. The stress of dying': the role of heat shock proteins in the regulation of apoptosis. *Journal of cell science*, 117(13), pp.2641–2651.

- Bercovich, B., Stancovski, I., Mayer, A., Blumenfeld, N., Laszlo, A., Schwartz, A.L. and Ciechanover, A., 1997. Ubiquitin-dependent degradation of certain protein substrates in vitro requires the molecular chaperone Hsc70. *Journal of Biological Chemistry*, 272(14), pp.9002–9010.
- Bhattacharya, D., Price, D.C., Chan, C.X., Qiu, H., Rose, N., Ball, S., Weber, A.P., Arias, M.C., Henrissat, B., Coutinho, P.M. and Krishnan, A., 2013. Genome of the red alga *Porphyridium purpureum*. *Nature Communications*, 4, p.1941.
- Bird, C.J. and Ragan, M.A. eds., 2012. *Eleventh International Seaweed Symposium: Proceedings of the Eleventh International Seaweed Symposium, held in Qingdao, People's Republic of China, June 19–25, 1983 (Vol. 22)*. Springer Science & Business Media.
- Bird, K.T. and Benson, P.H., 1987. Seaweed cultivation for renewable resources.
- Bishop, A., De Beer, T.A. and Joubert, F., 2008. Protein homology modelling and its use in South Africa. *South African Journal of Science*, 104(1-2), pp.2–6.
- Boenigk, J., Wodniok, S. and Glücksman, E., 2015. *Biodiversity and earth history*. Springer.
- Bohlin, J., Snipen, L., Hardy, S.P., Kristoffersen, A.B., Lagesen, K., Dønsvik, T., Skjerve, E. and Ussery, D.W., 2010. Analysis of intra-genomic GC content homogeneity within prokaryotes. *BMC Genomics*, 11(1), p.464.
- Boo, G.H. and Hughey, J.R., 2018. Phylogenomics and multigene phylogenies decipher two new cryptic marine algae from California, *Gelidium gabrielsonii* and *G. kathyanniae* (*Gelidiales*, *Rhodophyta*). *Journal of Phycology*.
- Boo, G.H., Hughey, J.R., Miller, K.A. and Boo, S.M., 2016. Mitogenomes from type specimens, a genotyping tool for morphologically simple species: ten genomes of agar-producing red algae. *Scientific reports*, 6, p.35337.
- Branch, G., 2017. *Two Oceans: a guide to the marine life of southern Africa*. Penguin Random House South Africa.
- Brawley, S.H., Blouin, N.A., Ficko-Blean, E., Wheeler, G.L., Lohr, M., Goodson, H.V., Jenkins, J.W., Blaby-Haas, C.E., Helliwell, K.E., Chan, C.X. and Marriage, T.N., 2017. Insights into the red algae and eukaryotic evolution from the genome of *Porphyra umbilicalis* (*Bangiophyceae*, *Rhodophyta*). *Proceedings of the National Academy of Sciences*, 114(31), pp. E6361–E6370.
- Burger, G., Gray, M.W. and Lang, B.F., 2003. Mitochondrial genomes: anything goes. *Trends in genetics*, 19(12), pp.709–716
- Butterfield, N.J., 2000. *Bangiomorpha pubescens* n. gen., n. Sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology*, 26(3), pp.386–404.
- Callaway, E., 2015. Lab staple agar runs low: dwindling seaweed harvest imperils reagent essential for culturing microbes. *Nature*, 528(7581), pp.171–173.

- Cao, J., Yang, L.T., Guo, M. and Lau, F. eds., 2005. Parallel and Distributed Processing and Applications: Second International Symposium, ISPA 2004, Hong Kong, China, December 13-15, 2004, Proceedings (Vol. 3358). Springer.
- Chandler, V.L. and Brendel, V., 2002. The maize genome sequencing project. *Plant Physiology*, 130(4), pp.1594–1597.
- Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Müller, W.E., Wetter, T. and Suhai, S., 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome research*, 14(6), pp.1147-1159.
- Chevreux, B., Wetter, T. and Suhai, S., 1999, October. Genome sequence assembly using trace signals and additional sequence information. In *German conference on bioinformatics* (Vol. 99, No. 1, pp. 45-56).
- Chiang, H.L., Plant, C.P. and Dice, J.F., 1989. A role for a 70-kilodalton heat shock protein in lysosomal degradation of intracellular proteins. *Science*, 246(4928), pp.382–385.
- Chiu, K.P., 2015. *Next-generation sequencing and sequence data analysis*. Bentham Science Publishers.
- Choudhry, H. and Wu, W. eds., 2013. Next Generation Sequencing in Cancer Research: Volume 1: Decoding the Cancer Genome. Springer New York.
- Chuan, Z., 2014. Endosymbiotic theory for organelle origins. *Current opinion in microbiology*. 
- Cole, K.M. and Sheath, R.G. eds., 1990. *Biology of the red algae*. Cambridge University Press.
- Collén, J., Porcel, B., Carré, W., Ball, S.G., Chaparro, C., Tonon, T., Barbeyron, T., Michel, G., Noel, B., Valentin, K. and Elias, M., 2013. Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. *Proceedings of the National Academy of Sciences*, 110(13), pp.5247–5252.
- Constantinescu, R., 2015. A Machine Learning Approach to DNA Shotgun Sequence Assembly. Master of Science in Engineering. University of Witwatersrand. South Africa, pp.24.
- Costa, J.F., Lin, S.M., Macaya, E.C., Fernández-García, C. and Verbruggen, H., 2016. Chloroplast genomes as a tool to resolve red algal phylogenies: a case study in the Nemaliales. *BMC evolutionary biology*, 16(1), p.205.
- Day, D.A., 2004. Mitochondrial structure and function in plants. In *Plant Mitochondria: From Genome to Function* (pp. 1–11). Springer, Dordrecht
- Delwiche, C.F. and Palmer, J.D., 1997. The origin of plastids and their spread via secondary symbiosis. In *Origins of algae and their plastids* (pp. 53-86). Springer, Vienna.
- DePriest, M.S., Bhattacharya, D. and López-Bautista, J.M., 2013. The plastid genome of the red macroalga *Grateloupia taiwanensis* (Halymeniaceae). *PLoS One*, 8(7), p.e68246.
- DeSalle, R. and Rosenfeld, J.A., 2012. *Phylogenomics*. Garland Science.
- Dey, P.M. and Harborne, J.B. eds., 1997. *Plant biochemistry*. Elsevier.

Di Tommaso, P., Moretti, S., Xenarios, I., Orobittg, M., Montanyola, A., Chang, J.M., Taly, J.F. and Notredame, C., 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic acids research*, 39(suppl_2), pp.W13–W17.

Dixon, P.S., 1973. *Biology of the Rhodophyta* (Vol. 4). Edinburgh: Oliver & Boyd.

Ebenezer N. S., Priyakumari C. J., 2017. Analysis of phylogenetic relationship among carangoides species using mega 6. *International Research Journal of Engineering and Technology* Pp.1804–1808.

Egholm, M., Margulies, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. and Dewell, S.B., 2005. Genome sequencing in open microfabricated high density picoliter reactors. *Nature*, 437, pp.376-380.

Eklblom, R. and Wolf, J.B., 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, 7(9), pp.1026–1042.

El Gamal, A.A., 2010. Biological importance of marine algae. *Saudi Pharmaceutical Journal*, 18(1), pp.1–25.

Elloumi, M., 2017. *Algorithms for next-generation sequencing data*. Springer

Felton, A., 2015. State of the art of Ion Torrent technology

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A. and Merrick, J.M., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223), pp.496-512.

Floudas, C.A., 2007. Computational methods in protein structure prediction. *Biotechnology and Bioengineering*, 97(2), pp.207–213.

Gallagher, S.R. and Desjardins, P.R., 2006. Quantitation of DNA and RNA with absorption and fluorescence spectroscopy. *Current protocols in molecular biology*, 76(1), pp.A-3D.

Gansauge, M.T., Gerber, T., Glocke, I., Korlević, P., Lippik, L., Nagel, S., Riehl, L.M., Schmidt, A. and Meyer, M., 2017. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic acids research*, 45(10), pp.e79–e79.


Garrison, T.S., 2012. *Essentials of Oceanography*. Cengage Learning.

Giardi, M.T. and Piletska, E.V., 2006. *Biotechnological applications of photosynthetic proteins: biochips, biosensors and biodevices*. Landes Bioscience and Springer Science+ Business Media, LLC.

Grant, J.R. and Stothard, P., 2008. The CGView Server: a comparative genomics tool for circular genomes. *Nucleic acids research*, 36(suppl_2), pp.W181–W184.

Gray, M.W., Lang, B.F., Cedergren, R., Golding, G.B., Lemieux, C., Sankoff, D., Turmel, M., Brossard, N., Delage, E., Littlejohn, T.G. and Plante, I., 1998. Genome structure and gene content in protist mitochondrial DNAs. *Nucleic acids research*, 26(4), pp.865–878.

Griffin, H.G., and Griffin, A.M., 1993. DNA sequencing. *DNA Sequencing Protocols*, pp.1–8.

- Griffiths, C.L., Robinson, T.B., Lange, L. and Mead, A., 2010. Marine biodiversity in South Africa: an evaluation of current states of knowledge. *PloS one*, 5(8), p.e12008.
- Gross, W. and Schnarrenberger, C., 1995. Heterotrophic growth of two strains of the acidothermophilic red alga *Galdieria sulphuraria*. *Plant and Cell Physiology*, 36(4), pp.633–638.
- Guiry, M.D. & Guiry, G.M. 2018. AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. <http://www.algaebase.org>; searched on 29 November 2018.
- Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G., 2013. QAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), pp.1072–1075.
- Hall, T.A., 1999, January. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In *Nucleic acids symposium series* (Vol. 41, No. 41, pp. 95-98). [London]: Information Retrieval Ltd., c1979–c2000.
- Hatherley, R., Brown, D.K., Glenister, M. and Bishop, Ö.T., 2016. PRIMO: An interactive homology modeling pipeline. *PloS one*, 11(11), p.e0166698.
- Hector, S., 2014. South African red algae gene discovery to help treat tumours. <https://www.all4women.co.za/168665/health/south-african-red-algae-gene-discovery-to-help-treat-tumours>
- Hector, S.B.E., 2013. Molecular studies of galactan biosynthesis in red algae.
- Henze, K., Martin, W. and Martin, W., 2003. Darwinian bioscience: essence of midichloria. *Nature*, 436(6963), pp.137–8. 
- Hieter, P. and Boguski, M., 1997. Functional genomics: it's all how you read it. *Science*, 278(5338), pp.601–602. *Together in Excellence*
- Hirakawa, Y. and Ishida, K.I., 2014. Polyploidy of endosymbiotically derived genomes in complex algae. *Genome biology and evolution*, 6(4), pp.974–980.
- Hoek, C., Van den Hoek, H., Mann, D. and Jahns, H.M., 1995. *Algae: an introduction to phycology*. Cambridge university press.
- Holdt, S.L. and Kraan, S., 2011. Bioactive compounds in seaweed: functional food applications and legislation. *Journal of applied phycology*, 23(3), pp.543–597.
- Holst-Jensen, A., Spilsberg, B., Arulandhu, A.J., Kok, E., Shi, J. and Zel, J., 2016. Application of whole genome shotgun sequencing for detection and characterisation of genetically modified organisms and derived products. *Analytical and bioanalytical chemistry*, 408(17), pp.4595–4614
- Hvidsten, T.R., 2004. *Predicting function of genes and proteins from sequence, structure and expression data* (Doctoral dissertation, Acta Universitatis Upsaliensis).
- Idury, R.M. and Waterman, M.S., 1995. A new algorithm for DNA sequence assembly. *Journal of computational biology*, 2(2), pp.291–306.
- Iha, C., Grassa, C.J., Lyra, G.D.M., Davis, C.C., Verbruggen, H. and Oliveira, M.C., 2018. Organellar genomics: a useful tool to study evolutionary relationships and molecular evolution in Gracilariaceae (Rhodophyta). *Journal of Phycology*.

Janouškovec, J., Liu, S.L., Martone, P.T., Carré, W., Leblanc, C., Collén, J. and Keeling, P.J., 2013. Evolution of red algal plastid genomes: ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers. *PLoS One*, 8(3), p.e59001.

Jennings, W.B., 2017. Future of Phylogenomic Data Acquisition. In *Phylogenomic Data Acquisition: Principles and Practice* (pp. 221-224). CRC Press

Kahl, G. and Meksem, K. eds., 2008. The handbook of plant functional genomics: concepts and protocols. John Wiley & Sons.

Kanehisa, M. and Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research*, 28(1), pp.27-30.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K., 2016. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1), pp.D353–D361

Kasahara, M. and Morishita, S., 2006. *Large-scale genome sequence processing*. Imperial College Press.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C. and Thierer, T., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), pp.1647-1649.



Keiler, K.C. and Ramadoss, N.S., 2011. Bifunctional transfer-messenger RNA. *Biochimie*, 93(11), pp.1993-1997.

Khan, S., Nabi, G., Ullah, M.W., Yousaf, M., Manan, S., Siddique, R. and Hou, H., 2016. Overview on the role of advance genomics in conservation biology of endangered species. *International journal of genomics*, 2016

Kim, K.M., Park, J.H., Bhattacharya, D. and Yoon, H.S., 2014. Applications of next-generation sequencing to unravelling the evolutionary history of algae. *International journal of systematic and evolutionary microbiology*, 64(2), pp.333–345.

Kim, S.K., Nair, R.M., Lee, J. and Lee, S.H., 2015. Genomic resources in mungbean for future breeding programs. *Frontiers in plant science*, 6, p.626

Knierim, E., Lucke, B., Schwarz, J.M., Schuelke, M. and Seelow, D., 2011. Systematic comparison of three methods for fragmentation of long-range PCR products for next-generation sequencing. *PloS one*, 6(11), p.e28240

Knott, M.G., de la Mare, J.A., Edkins, A.L., Zhang, A., Stillman, M.J., Bolton, J.J., Antunes, E.M. and Beukes, D.R., 2019. Plaxenone A and B: Cytotoxic halogenated monoterpenes from the South African red seaweed *Plocamium maxillosum*. *Phytochemistry Letters*, 29, pp.182-185.

Kogame, K., Uwai, S., Anderson, R.J., Choi, H.G. and Bolton, J.J., 2017. DNA barcoding of South African geniculate coralline red algae (Corallinales, Rhodophyta). *South African journal of botany*, 108, pp.337-341.

Kryshtafovych, A. and Fidelis, K., 2009. Protein structure prediction and model quality assessment. *Drug discovery today*, 14(7-8), pp.386–393.

Kumar, C.S., Ganesan, P., Suresh, P.V. and Bhaskar, N., 2008. Seaweeds as a source of nutritionally beneficial compounds-a review. *Journal of Food Science and Technology*, 45(1), pp.1–13.

Kumar, R., Kumari, B. and Kumar, M., 2016. PredHSP: sequence-based proteome-wide heat shock protein prediction and classification tool to unlock the stress biology. *PloS one*, 11(5), p.e0155872.

Kurucz, I., Morva, A., Vaag, A., Eriksson, K.F., Huang, X., Groop, L. and Koranyi, L., 2002. Decreased expression of heat shock protein 72 in skeletal muscle of patients with type 2 diabetes correlates with insulin resistance. *Diabetes*, 51(4), pp.1102–1109.

Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.H., Rognes, T. and Ussery, D.W., 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research*, 35(9), pp.3100–3108.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography*, 26(2), pp.283-291.

Leblanc, C., Richard, O., Kloareg, B., Viehmann, S., Zetsche, K. and Boyen, C., 1997. Origin and evolution of mitochondria: what have we learnt from red algae?. *Current genetics*, 31(3), pp.193-207.



Lee, J., Cho, C.H., Park, S.I., Choi, J.W., Song, H.S., West, J.A., Bhattacharya, D. and Yoon, H.S., 2016. Parallel evolution of highly conserved plastid genome architecture in red seaweeds and seed plants. *BMC biology*, 14(1), p.75.


Lee, J., Kim, K.M., Yang, E.C., Miller, K.A., Boo, S.M., Bhattacharya, D. and Yoon, H.S., 2016. Reconstructing the complex evolutionary history of mobile plasmids in red algal genomes. *Scientific reports*, 6, p.23744.

Lee, J.M., Song, H.J., Park, S.I., Lee, Y.M., Jeong, S.Y., Cho, T.O., Kim, J.H., Choi, H.G., Choi, C.G., Nelson, W.A. and Fredericq, S., 2018. Mitochondrial and plastid genomes from coralline red algae provide insights into the incongruent evolutionary histories of organelles. *Genome biology and evolution*, 10(11), pp.2961-2972.

Li, Y., Chen, C.Y., Kaye, A.M. and Wasserman, W.W., 2015. The identification of cis-regulatory elements: A review from a machine learning perspective. *Biosystems*, 138, pp.6–17.

Li, Y., Meinita, M.D.N., Liu, T., Chi, S. and Yin, H., 2018. Complete sequences of the mitochondrial DNA of the *Grateloupia filicina* (Rhodophyta). *Mitochondrial DNA Part B*, 3(1), pp.76–77.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B. and Yang, B., 2012. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in functional genomics*, 11(1), pp.25–37.

- Lipman, D.J. and Pearson, W.R., 1985. Rapid and sensitive protein similarity searches. *Science*, 227(4693), pp.1435–1441.
- Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M., 2012. Comparison of next-generation sequencing systems. *BioMed Research International*, 2012.
- Liu, N., Wang, G., Li, Y., Zhang, L., Meinita, M.D.N., Chen, W., Liu, T. and Chi, S., 2017. The complete mitochondrial genome of the economic red alga, *Gracilaria chilensis*. *Mitochondrial DNA Part B*, 2(2), pp.716–717.
- Liu, Z.J., 2017. *Bioinformatics in Aquaculture: Principles and Methods*. John Wiley & Sons.
- Lombard, A.T., 2004. Marine component of the National Spatial Biodiversity Assessment for the development of South Africa's National Biodiversity Strategic and Action Plan. p.101
- Lowe, T.M. and Chan, P.P., 2016. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic acids research*, 44(W1), pp.W54–W57.
- Lubke, R. and De Moor, I.J. eds., 1998. *Field guide to the Eastern & Southern Cape Coasts*. Juta and Company Ltd.
- Maier, R.M. and Schmitz-Linneweber, C., 2004. Plastid genomes. In *Molecular biology and biotechnology of plant organelles* (pp. 115–150). Springer, Dordrecht.
- Marzillier, J., 2008. Genome Sequencing. Biological Sciences. Lehigh University. Biological Sciences
- 
- Mascher, M., Wu, S., Amand, P.S., Stein, N. and Poland, J., 2013. Application of genotyping-by-sequencing on semiconductor sequencing platforms: a comparison of genetic and reference-based marker ordering in barley. *PLoS one*, 8(10), p.e76925.
- Maslo, B. and Lockwood, J.L. eds., 2014. Coastal conservation (Vol. 19). Cambridge University Press.
- Matsuzaki, M., Misumi, O., Shin-i, T., Maruyama, S., Takahara, M., Miyagishima, S.Y., Mori, T., Nishida, K., Yagisawa, F., Nishida, K. and Yoshida, Y., 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature*, 428(6983), p.653.
- McHugh, D.J., 2003. A guide to the seaweed industry FAO Fisheries Technical Paper 441. *Food and Agriculture Organization of the United Nations, Rome*.
- Mikheenko, A., Valin, G., Prjibelski, A., Saveliev, V. and Gurevich, A., 2016. Icarus: visualizer for *de novo* assembly evaluation. *Bioinformatics*, 32(21), pp.3321–3323.
- Miller, J.R., Koren, S. and Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6), pp.315–327.
- Morimoto, R.I., 1998. Regulation of the heat shock transcriptional response: cross talk between a family of heat shock factors, molecular chaperones, and negative regulators. *Genes & development*, 12(24), pp.3788–3796.

- Muñoz-Gómez, S.A., Mejía-Franco, F.G., Durnin, K., Colp, M., Grisdale, C.J., Archibald, J.M. and Slamovits, C.H., 2017. The new red algal subphylum Proteorhodophytina comprises the largest and most divergent plastid genomes known. *Current Biology*, 27(11), pp.1677-1684.
- Nakamura, Y., Sasaki, N., Kobayashi, M., Ojima, N., Yasuike, M., Shigenobu, Y., Satomi, M., Fukuma, Y., Shiwaku, K., Tsujimoto, A. and Kobayashi, T., 2013. The first symbiont-free genome sequence of marine red alga, Susabi-nori (*Pyropia yezoensis*). *PloS one*, 8(3), p.e57122.
- Namiki, T., Hachiya, T., Tanaka, H. and Sakakibara, Y., 2012. MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic acids research*, 40(20), pp.e155–e155.
- Nan, F., Feng, J., Lv, J., Liu, Q., Fang, K., Gong, C. and Xie, S., 2017. Origin and evolutionary history of freshwater Rhodophyta: further insights based on phylogenomic evidence. *Scientific reports*, 7(1), p.2934.
- Ng, P.C. and Kirkness, E.F., 2010. Whole genome sequencing. In *Genetic variation* (pp. 215-226). Humana Press, Totowa, NJ.
- Ng, P.K., Lin, S.M., Lim, P.E., Liu, L.C., Chen, C.M. and Pai, T.W., 2017. Complete chloroplast genome of *Gracilaria firma* (Gracilariaceae, Rhodophyta), with discussion on the use of chloroplast phylogenomics in the subclass Rhodymeniophycidae. *BMC Genomics*, 18(1), p.40.
- Nosek, J., Fukuhara, H., Suyama, Y. and Kováč, L., 1998. Linear mitochondrial genomes: 30 years down the line. *Trends in Genetics*, 14(5), pp.184–188.
- Nurk, S., Bankevich, A., Antipov, D., Gurevich, A.A., Korobeynikov, A., Lapidus, A., Prjibelski, A.D., Pyshkin, A., Sirotkin, A., Sirotkin, Y. and Stepanauskas, R., 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *Journal of Computational Biology*, 20(10), pp.714–737.
- Ohta, N., Matsuzaki, M., Misumi, O., Miyagishima, S.Y., Nozaki, H., Tanaka, K., Shin-i, T., Kohara, Y. and Kuroiwa, T., 2003. Complete sequence and analysis of the plastid genome of the unicellular red alga *Cyanidioschyzon merolae*. *DNA research*, 10(2), pp.67–77.
- Ohta, N., Sato, N. and Kuroiwa, T., 1998. Structure and organisation of the mitochondrial genome of the unicellular red alga *Cyanidioschyzon merolae* deduced from the complete nucleotide sequence. *Nucleic acids research*, 26(22), pp.5190–5198.
- Osawa, S., Jukes, T.H., Watanabe, K. and Muto, A., 1992. Recent evidence for evolution of the genetic code. *Microbiological reviews*, 56(1), pp.229–264.
- Padmalayam, I., 2014. The heat shock response: its role in pathogenesis of type 2 diabetes and its complications, and implications for therapeutic intervention. *Discov Med*, 18(97), pp.29–39.
- Page, R.D.M., 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in Biosciences* 12, 357–358.
- Patnaik B.K., Kara T.C., and Ghosh S.N., 2012. Textbook of Biotechnology: Chapter 26, Genomics and Proteomics

- Pop, M., 2009. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, 10(4), pp.354–366.
- Pop, M., Phillippy, A., Delcher, A.L. and Salzberg, S.L., 2004. Comparative genome assembly. *Briefings in bioinformatics*, 5(3), pp.237–248.
- Qiu, H., Price, D.C., Yang, E.C., Yoon, H.S. and Bhattacharya, D., 2015. Evidence of ancient genome reduction in red algae (Rhodophyta). *Journal of Phycology*, 51(4), pp.624–636.
- Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. and Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13(1), p.341.
- Ramakrishnan, A.P., Ritland, C.E., Sevillano, R.H.B. and Riseman, A., 2015. Review of potato molecular markers to enhance trait selection. *American journal of potato research*, 92(4), pp.455–472.
- Ramakrishnan, V., 2002. Ribosome structure and the mechanism of translation. *Cell*, 108(4), pp.557–572.
- Reuter, J.A., Spacek, D.V. and Snyder, M.P., 2015. High-throughput sequencing technologies. *Molecular Cell*, 58(4), pp.586–597.
- Rodríguez-Ezpeleta, N., Hackenberg, M. and Aransay, A.M. eds., 2011. *Bioinformatics for high throughput sequencing*. Springer Science & Business Media.
- Rogalski, M., do Nascimento Vieira, L., Fraga, H.P. and Guerra, M.P., 2015. Plastid genomics in horticultural species: importance and applications for plant population genetics, evolution, and biotechnology. *Frontiers in plant science*, 6, p.586.
- Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M. and Hoon, J., 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356), p.348
- Russell Peter J., Hertz Paul E., and McMillan B. 2016. *Biology: The dynamics of Science*
- Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison III, C.A., Slocombe, P.M. and Smith, M., 1977. Nucleotide sequence of bacteriophage ϕ X174 DNA. *Nature*, 265(5596), p.687.
- Sanger, F., Nicklen, S. and Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12), pp.5463–5467.
- Schneider, A. and Maréchal-Drouard, L., 2000. Mitochondrial tRNA import: are there distinct mechanisms?. *Trends in cell biology*, 10(12), pp.509–513.
- Schneider, A., 2011. Mitochondrial tRNA import and its consequences for mitochondrial translation. *Annual review of biochemistry*, 80, pp.1033–1053.
- Seagrief, S.C., 1967. *The Seaweeds of the Tsitsikama Coastal National Park: Die Seewiere van die Tsitsikama-Seekus Nasionale Park*. National Parks Board.

Seckbach, J. and Chapman, D.J. eds., 2010. Red algae in the genomic age (Vol. 13). Springer Science & Business Media.

Sharma, V., 2008. *Textbook of bioinformatics*. Rastogi Publications.

Simbolo, M., Gottardi, M., Corbo, V., Fassan, M., Mafficini, A., Malpeli, G., Lawlor, R.T. and Scarpa, A., 2013. DNA qualification workflow for next-generation sequencing of histopathological samples. *PloS one*, 8(6), p.e62692.

Simons, R.H., 1977. [1976] Seaweeds of southern Africa: guidelines for their study and identification. *Fish. Bull. S. Afr.*, 7, pp.1–113.

Sitbon, E. and Pietrokovski, S., 2007. Occurrence of protein structure elements in conserved sequence regions. *BMC structural biology*, 7(1), p.3.

Srinivasan, S. and Avadhani, N.G., 2012. Cytochrome c oxidase dysfunction in oxidative stress. *Free Radical Biology and Medicine*, 53(6), pp.1252–1263.

Stein, L., 2001. Genome annotation: from sequence to biology. *Nature reviews genetics*, 2(7), p.493.

Thompson, S.M., 2004. Multiple Sequence Alignment and Analysis: Part I—An Introduction to the Theory and Application of Multiple Sequence Analysis. *Computational Genomics: Theory and Application*.



Titilade, P.R. and Olalekan, E.I., 2015. The importance of marine genomics to life. *Journal of Ocean Research*, 3(1), pp.1–13.

Trivedi, B., 2000. Sequencing the genome. *Genome News Network*.

Türktaş, M., Kurtoğlu, K.Y., Dorado, G., Zhang, B., Hernandez, P. and ÜNVER, T., 2015. Sequencing of plant genomes? A review. *Turkish Journal of Agriculture and Forestry*, 39(3), pp.361–376.

Verma, A. and Singh, A. eds., 2013. *Animal biotechnology: models in discovery and translation*. Academic Press.

Visendi, P., Berkman, P.J., Hayashi, S., Golicz, A.A., Bayer, P.E., Ruperao, P., Hurgobin, B., Montenegro, J., Chan, C.K.K., Staňková, H. and Batley, J., 2016. An efficient approach to BAC-based assembly of complex genomes. *Plant Methods*, 12(1), p.2.

Wallin, I.E., 1927. *Symbiogenesis and the Origin of Species*. Рипол Классик.

Wang, L., Mao, Y., Kong, F., Li, G., Ma, F., Zhang, B., Sun, P., Bi, G., Zhang, F., Xue, H. and Cao, M., 2013. Complete sequence and analysis of plastid genomes of two economically important red algae: *Pyropia haitanensis* and *Pyropia yezoensis*. *PloS one*, 8(5), p.e65902.

Wang, X. and Kole, C., 2015. *The Brassica Rapa Genome* (Vol. 4). Springer.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L. and Lepore, R., 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*.

Wilson, K.S. and Prochaska, L.J., 1990. Phospholipid vesicles containing bovine heart mitochondrial cytochrome c oxidase and subunit III-deficient enzyme: analysis of respiratory control and proton translocating activities. *Archives of biochemistry and Biophysics*, 282(2), pp.413–420.

Wong, K.C., 2016. *Computational biology and bioinformatics: Gene regulation*. CRC Press.

Wu, S., Moreno-Sanchez, R. and Rottenberg, H., 1995. Involvement of cytochrome c oxidase subunit III in energy coupling. *Biochemistry*, 34(50), pp.16298–16305.

Yang, E.C., Kim, K.M., Boo, G.H., Lee, J.H., Boo, S.M. and Yoon, H.S., 2014. Complete mitochondrial genome of the agarophyte red alga *Gelidium vagum* (*Gelidiales*).

Yang, E.C., Kim, K.M., Kim, S.Y., Lee, J., Boo, G.H., Lee, J.H., Nelson, W.A., Yi, G., Schmidt, W.E., Fredericq, S. and Boo, S.M., 2015. Highly conserved mitochondrial genomes among multicellular red algae of the Florideophyceae. *Genome biology and evolution*, 7(8), pp.2394–2406.

Yuan, X., 2018. The complete mitochondrial genome of *Gracilaria textorii* (Gracilariales, Florideophyceae). *Mitochondrial DNA Part B*, 3(1), pp.438-439.

Zhou, W., Hu, Y., Sui, Z., Fu, F., Wang, J., Chang, L., Guo, W. and Li, B., 2013. Genome survey sequencing and genetic background characterisation of *Gracilariopsis lemaneiformis* (Rhodophyta) based on next-generation sequencing. *PLoS One*, 8(7), p.e69909.

WEBLINKS USED

http://cfb.unh.edu/phycokey/Choices/Rhodophyceae/Macroreds/GRACILARIA/Gracilaria_image page.htm Accessed 09/09/2018

<http://realdb.algaegenome.org/> Accessed 01/12/2018

http://rice.plantbiology.msu.edu/training/Sequencing_assembly.pdf Accessed 22/09/2018

<http://www.bio-itworld.com/2015/9/1/thermo-fisher-clarifies-vision-sequencing-release-ion-s5-instruments.html> Accessed 22/11/2018

<https://earth.google.com/web/@33.68146685,26.66828445,64.15513905a,9672.3130936d,35y,0h,0t,0r> Accessed on 5/12/2018

<https://insidedna.io/tutorials/view/fastqc-quality-control-and-filtering-of> Accessed 23/11/2017

<https://proteinstrucures.com/Modeling/homology-modeling.html> Accessed 05/11/2018

<https://sciencestruck.com/how-is-mitochondrial-dna-used-in-forensics>; Accessed 25/ 11/2018

<https://swissmodel.expasy.org> Accessed 6/12/2018

https://www.123rf.com/photo_46959995_stock-vector-chloroplast-structure-of-a-typical-higher-plant-chloroplast.html respectively. Accessed 25/ 11/2018

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> Accessed 23/11/2017



<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality> **Accessed 23/11/2017**

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/3%20Per%20Sequence%20Quality%20Scores.html> **Accessed 23/11/2018**

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/5%20Per%20Sequence%20GC%20Content.html> **accessed 23/11/2017**

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/7%20Sequence%20Length%20Distribution.html> **Accessed 23/11/2017**

<https://www.biologicscorp.com/tools/GCContent/> **Accessed 11/12/2018**

<https://www.genome.jp/kegg/> **Accessed on 01/12/2018**

<https://www.genome.jp/kegg/> **Accessed on 01/12/2018**

https://www.google.co.za/search?biw=1366&bih=608&tbm=isch&sa=1&ei=cyYMXPrKKvuE1fAP4qSeoA0&q=G.sulphuraria&oq=G.sulphuraria&gs_l=img.12...56943.59923..62754...0.0..0.286.567.2-2.....1....1j2..gws-wiz-img.AT13Pzgabmk#imgrc=_AOnL0lLeOhrjM: **Accessed 09/09/2018**

https://www.google.co.za/search?biw=1366&bih=608&tbm=isch&sa=1&ei=BU MMXOGbPIqL1fAPnO60CA&q=Gp.+lemaniformis+red+algae&oq=Gp.+lemaniformis+red+algae&gs_l=img.3...9175.11214..11610...0.0..0.413.3206.2-2j7j1.....1....1..gws-wiz-img.X4dLBFhAcs0#imgdii=8G-AQZVDo9tHtM:&imgrc=uRvEek-1477QeM: **Accessed 09/09/2018**

University of Fort Hare

https://www.google.co.za/search?biw=1366&bih=613&tbm=isch&sa=1&ei=8CXnWuewDMPoGAb336KwBw&q=nuclear+DNA&oq=nuclear+DNA&gs_l=psy-ab.3..017j0i5i30k1j0i24k112.126587.133939.0.134920.12.9.0.3.3.0.630.2050.2-2j1j1j1.5.0....0...1c.1.64.psy-ab..4.8.2079.0..0i67k1.0.EiXp92bjFlg#imgrc=z0MvgHP0f2_tWM; **Accessed 25/ 11/2018**

https://www.google.co.za/search?q=1.+C.+merolae&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiwgDKnhpHfAhXCCewKHcOdAnMQ_AUIDigB&biw=1366&bih=608#imgrc=Pwh_xByJFF5JGM: **Accessed 09/09/2018**

https://www.google.co.za/search?q=Hierarchical+shotgun+sequencing.+B:+the+whole+genome+shotgun+sequencing+strategies&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiO__CAgorXAhUkKMAKHRjNCUsQ_AUICigB&biw=1366&bih=662#imgrc=twDzZMdAdWj3TM **Accessed 25/ 11 2018**

https://www.google.co.za/search?q=ion+Torrent+principle&rlz=1C1GNAM_enZA679ZA679&source=lnms&tbm=isch&sa=X&ved=0ahUKEwifmK2F5JzbAhUHAsAKHeRTDPcQ_AUICigB&biw=1280&bih=869#imgdii=7CaL2aZJTW6N2M:&imgrc=Fhdo40aEphKfYM: **Accessed 21/09/2018**

https://www.google.co.za/search?q=nhgri+sequencing+costs&source=lnms&tbm=isch&sa=X&ved=0ahUKEwj78ZDlq9TbAhWCHsAKHYjLCH8Q_AUICigB&biw=1242&bih=557#imgrc=nw_L_WMKdoBAeM: **Accessed 21/09/2018**

https://www.google.com/search?q=The+OLC+Assemblers&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjnv6KjpcLdAhXLLcAKHcpEAtsQ_AUIDygC&biw=1280&bih=913#imgrc=CjkkkpChRZX6nM **Accessed 26/11/2018**

https://www.google.com/search?tbm=isch&q=Referencebased+genome+assembly&chips=q:reference+based+genome+assembly,online_chips:alignment&sa=X&ved=0ahUKEwimfrXqsLdAhUHAcAKHTfdACoQ4IYILygI&biw=1280&bih=913&dpr=1#imgrc=5PUPCRJI9acNEM **Accessed 26/11/2018**

<https://www.healio.com/hematology-oncology/learn-genomics/whole-genomesequencing/whole-genome-sequencing-methods> **Accessed 25/ 11 2018**

<https://www.ncbi.nlm.nih.gov> **Accessed 01/12/2018**

<https://www.ncbi.nlm.nih.gov> **Accessed 25/03/2019**

<https://www.ncbi.nlm.nih.gov/nuccore/NADL00000000.2> **Accessed 01/12/18**

<https://www.ncbi.nlm.nih.gov/orffinder> **Accessed 01/10/2018**

[https://www.ncbi.nlm.nih.gov/protein/YP_009488724.1?report=genbank&log\\$=protop&blast_rank=8&RID=05YRS061015](https://www.ncbi.nlm.nih.gov/protein/YP_009488724.1?report=genbank&log$=protop&blast_rank=8&RID=05YRS061015) **Accessed 01/12/2018.**

[https://www.ncbi.nlm.nih.gov/protein/YP_009511292.1?report=genbank&log\\$=protop&blast_rank=9&RID=05YRS061015](https://www.ncbi.nlm.nih.gov/protein/YP_009511292.1?report=genbank&log$=protop&blast_rank=9&RID=05YRS061015) **Accessed 01/12/2018**

<https://www.neb.com/tools-and-resources/feature-articles/the-quantitation-question-how-does-accurate-library-quantitation-influence-sequencing> **Accessed 22/11/2017.**

<https://www.ogt.com/resources/literature/483> **Accessed 21/11/2018**

<https://www.thermofisher.com> **Accessed 25/11/2017**

<https://www.thermofisher.com/search/results?query=4484177&persona=DocSupport&type=Citations+%26+References> **Accessed 21/11/2017**

<https://www.thermofisher.com/search/results?query=4484177&persona=DocSupport&type=Citations+%26+References> **Accessed 21/11/2017**

<https://www.thermofisher.com/za/en/home/brands/ion-Torrent.html> **Accessed 22/09/2018**

<https://www.thermofisher.com/za/en/home/life-science/sequencing/dna-sequencing.html> **Accessed 26/11/2018**

<https://www.thermofisher.com/za/en/home/life-science/sequencing/next-generation-sequencing/ion-Torrent-next-generation-sequencing-applications.html> **Accessed 21/ 09/2018**

<https://www.thermofisher.com/za/en/home/life-science/sequencing/next-generation-sequencing/ion-Torrent-next-generation-sequencing-workflow/ion-Torrent-next-generation-sequencing-run-sequence/ion-s5-ngs-targeted-sequencing/ion-s5-specifications.html> **Accessed 22/11/2018**

APPENDICES

Referral Information



University of Fort Hare
Together in Excellence

A APPENDICES: Referral Information

A.1 APPENDIX A: Primers used for closing the organellar genome gaps

Mitochondrial primers

PAIR 1	MitFrwd1: GGTACGTGAGTTGGGTTTCAGA MitRvs1: TCCTACACCTACCTCTACTACTG
PAIR 2	MitFrwd2: CACCGTTGGTTTAGATGTAGA MitRvs2: TCAAGGGGAACTTACGCAAGG
PAIR 3	MitFrwd3: TAGGTCACCAATGGTACTGAAGC MitRvs3: CCATTTGATAAGACCCCGCAAAG
PAIR 4	MitFrwd4: GTCACTTGATCTCATCACGCAAT MitRvs4: GGACTTGAACCAATAACCCATT
PAIR 5	MitFrwd5: TATGAGGTGCTACGGTAATAACC MitRvs5: CAGGGGTAACCAATTGGATTAGC
PAIR 6	MitFrwd6: GTTGATTACGGACAGAGGGGTT MitRvs6: TTAAGCTTTCTCCGCCATAGCG
PAIR 7	MitFrwd7: ACAACCCGTCAAGCTCTAACC MitRvs7: CTTTGAAGATGCGAGAACTAC
PAIR 8.	MitFrwd8: CCAGCAACAAAGAATACTCAAGC MitRvs8: AAGCACCAACTGCAGGCTCAG
PAIR 9	MitFrwd9: CATTGGAATTAGCACGCTCTC MitRvs9: ATAGGCTCATTAGCTTTGGTAGG
PAIR 10	MitFrwd10: CTGATTTTCCAACCTGCACCTAC MitRvs10: CACGTTTACCTTGGCCTTATC
PAIR 11	MitFrwd11: AGTTGGCATTAGAGTTCCCTTG MitRvs11: CCAATGTGGCTGTCCATCTTCA
PAIR 12	MitFrwd12: GATGGAATTGCTAGTAATCGTG

MitRvs12: TGCCAGCCCTATTGAAGATTC

Additional Mitochondrial Primers

PAIR 13

MT_7.1: FTGATAATGCCGATTTTCATCG
MT_7.1R: CTTGATCCCTTTCTGAACTC

PAIR 14

MT_8.2F: GACATCGCTAGAATATACAG
MT_8.2R: GTTCTAGCGAGTTTAGGTATG

PAIR 15

MT_9.1F: CAGTTTCCTGTCATACAGCTTC
MT_9.1R: GCACCAAATTGAGGCAGAGCA

PAIR 16

MT_11.1F: CGTTGGCTAGAACTAACTC
MT_11.1R: ACCATTAAACGATAGGAGC

Plastid primers



PAIR 1

PlsdFwd1: ATGCCAGCTCCTAGAGGTGTT
PlsdRvs1: CATTGGGACTTCACAGAACTAC

PAIR 2

PlsdFwd2: ATGCTCCCACCATTAGAAAGCCT
PlsdRvs2: AACCGGCTGTAAAGGCTTTCAC

PAIR 3

PlsdFwd3: CGAAGCTTGCGCCCAAAGGAA
PlsdRvs3: ATTCTGTATAGTAGGGGCAGG

PAIR 4

PlsdFwd4: GCTAACTGTCATTGAGATACTC
PlsdRvs4: TTGCGAATGATCGTGTAAGGT

PAIR 5

PlsdFwd5: TGCCGTAAATTTTGCTCTGATGG
PlsdRvs5: CTA CTGTATCTAGTGAATCTGC

PAIR 6

PlsdFwd6: TCAGACGCAATCAACCAACTCT
PlsdRvs6: ACCTATCTATGGACGTATAAGC

PAIR 7

PlsdFwd7: TGTTGTAGTCTGCGTGATAGTTC
PlsdRvs7: CAGGTTAGCTATAGAATCGACA

PAIR 8 PlsdFwrd8: CAGCAGCCAATTAATGAGTATGC
PlsdRvs8: AAGCAATTATAGCGGGGAATGG

PAIR 9 PlsdFwrd9: GCTTCAAACGGGTCGAGAAGC
PlsdRvs9: AAACCTGACATTCTCCGCCATAC

PAIR 10 PlsdFwrd10: GCCCAAAGTGCTATGCAATTGTC
PlsdRvs10: GAACGAATCATTAGGAAAGCCATC

PAIR 11 PlsdFwrd11: CTTGCCTAATTTCCCCTTCAGC
PlsdRvs11: CCACCATAATGACCAACATCCTC

PAIR 12 PlsdFwrd12: CATCATGTTATGCAAGCAGTCAC
PlsdRvs12: CAGTGGACTTCCGCTACTTGC

PAIR 13 PlsdFwrd13: TACCTGATCAGCACCCCTTATGG
PlsdRvs13: AATACCTTGCTTCTGCATCTGC

PAIR 14 PlsdFwrd14: ATTCACTAGGTGTACCCATTGGT
PlsdRvs14: CTGGTGCAAGAATGCGTGTTGG

PAIR 15 PlsdFwrd15: GCAGCTTTAATCCTTGATAGTGG
PlsdRvs15: GAGGCTTTAAACTGCTAGAATGAC

PAIR 16 PlsdFwrd16: AATGCGCTCGCTGACTGCTCA
PlsdRvs16: ATGAGCATTACGCAAGCTGACC

PAIR 17 PlsdFwrd17: GATAGTACGTAGCCCTGGCATC
PlsdRvs17: ACCAGCATTAGGGCCTTCAGG

PAIR 18 PlsdFwrd18: ATTCTTGCCAGACGGCACTTC
PlsdRvs18: CATTCTACCTTGCAATATCATCCG

PAIR 19 PlsdFwrd19: GGTATGAGTGGAGAGCGCTAG
PlsdRvs19: GTTCTAGCGCCATTTCTTTAGGTA

PAIR 20 PlsdFwrd20: GATTTTGATGGTGACCAAATGGC
PlsdRvs20: ACTACAGTAATTTGACCACGCT

PAIR 21 PlsdFwrd21: GGTATGAGAGGCCTAATGGCA

PlsdRvs21: TCCGTTTGGCGCTATCACAGAAC
 PAIR 22 PlsdFwrd22: TATTTGCCGGGTGAAATGGTTGA
 PlsdRvs22: CGTATAGCATCATCATTTGCAGG
 PAIR 23 PlsdFwrd23: GCAAATAGCTGCTTGTCCCTCT
 PlsdRvs23: TAAGCGCCTCTTGCTGGCTTTC
 PAIR 24 PlsdFwrd24: GCTAATGTTGGTACCGTACTC
 PlsdRvs24: TTTGCAGCAGAAGCAGATCGC
 PAIR 25 PlsdFwrd25: ATTGTCTCCAGTAGGTACCACT
 PlsdRvs25: GCTGGTGG AATTGTTTCTGGTA
 PAIR 26 PlsdFwrd26: TTGATTCAGCATGTAAGGCTGAGA
 PlsdRvs26: TCAAGGGTTTGTTCCTAACCG
 PAIR 27 PlsdFwrd27: CCTTGGAAGCAGATCAATGAACG
 PlsdRvs27: CAGCGACGTTCTATTAAGAATGG
 PAIR 28 PlsdFwrd28: AGTAAAGGCTGACCCACACGTG
 PlsdRvs28: GTTGACAAGTTGCCTACATAAGG
 PAIR 29 PlsdFwrd29: TTACAGCAGGTTTCTTGGTTGGT
 PlsdRvs29: ATACAGTTGTTGGTGCAGTACCT
 PAIR 30 PlsdFwrd30: CTCAATCTGTAGAAGAACGGACAA
 PlsdRvs30: TTGATGCATTTGACCGCAATGG
 PAIR 31 PlsdFwrd31: CTACCAGATTTAACTGACGAAC
 PlsdRvs31: GAGAAGGTTGGTATATCTCGCAT
 PAIR 32 PlsdFwrd32: CAAGAGAAGAAGGCTAAAGAGG
 PlsdRvs32: TTCTCACAGTTGGGCTAATAGG
 PAIR 33 PlsdFwrd33: CCAGCTCAAACAGGTGATTGGT
 PlsdRvs33: GAATTAGTAGGTACTCCAAGAG
 PAIR 34 PlsdFwrd34: TTCTGGTGAAGCCGGTGATAG
 PlsdRvs34: ATTCCTGGGTCTATCTGTGTA

PAIR 35 PlsdFwrd35: GTTATATGGGTTTGCTACTTAGGG
PlsdRvs35: ATGTGGGCAAAGGAATGAATGGT

PAIR 36 PlsdFwrd36: GTGTAAGGTTTTAGACATCTCAGG
PlsdRvs36: TGCCCTCAGCAACCATTTTCGC

PAIR 37 PlsdFwrd37: ATTAGGAACACCAGGAACATCTG
PlsdRvs37: TCGCGTTCTGCAAGATACTCAT

PAIR 38 PlsdFwrd38: ATGCGAACATACCTTGATCTCA
PlsdRvs38: GATGGTATAGTAGTTGGTAGTGC

PAIR 39 PlsdFwrd39: ACTCGAACCCGGAACATAATCGG
PlsdRvs39: CCGATTAGTTCCGGGTTTCGAGT

PAIR 40 PlsdFwrd40: CTCCCGTTGATCCTCCTGACA
PlsdRvs40: TGTTAAGGAGTCATTGCTACCT

PAIR 41 PlsdFwrd41: CACCTAACTCACCAACAGAAAGC
PlsdRvs41: GAACTTAGACCACAGCTACTTGA

PAIR 42 PlsdFwrd42: TCCAACTGATGCCGCTGAATC
PlsdRvs42: TGGCTTACCAGCACCTAAAGG

PAIR 43 PlsdFwrd43: GGTACTTGTATCTCTGGAGCT
PlsdRvs43: CTGCTACTTTCCAGTCCTGACA

PAIR 44 PlsdFwrd44: TTCCGCAGTACCCGAATTATGC
PlsdRvs44: CCATTACTTAGAAGAAGGGTTG

PAIR 45 PlsdFwrd45: GCATCCGTACAATTGGCGCAAC
PlsdRvs45: CATTGGGGCCAGGATAGAGAA

PAIR 46 PlsdFwrd46: CACCTGTTATGGCTGTACCCAA
PlsdRvs46: GGACAATATATCCGTAATCCGGTT

PAIR 47 PlsdFwrd47: AAAGCCTGCAACAATGGGTCTA
PlsdRvs47: GAAACCAATTACAGTGCCATA

PAIR 48 PlsdFwrd48: AGTCATTGATCTGCAACCACTTC

PlsdRvs48: CCAGTAAGGTAATACAGACCCTA
 PAIR 49 PlsdFwrd49: CCAAGAGTCCTATACCTAATCG
 PlsdRvs49: AGTTCACATCGACGGGGAGGT
 PAIR 50 PlsdFwrd50: GTTCACTAAGTTTGCTCTCTTC
 PlsdRvs50: CTATACGGCGGTGAATCCGTT
 PAIR 51 PlsdFwrd51: TCCCCTTCCTAAAGGCAGATTC
 PlsdRvs51: CTTGTGCTTGAGAAGCCTCAGG
 PAIR 52 PlsdFwrd52: ATGGCATGCTGGACGAGCTAG
 PlsdRvs52: TCTATTGCTTTACCGTCTCTTC
 PAIR 53 PlsdFwrd53: GTAACGATGCATGGGAAGAGCTA
 PlsdRvs53: ATGCGGCAGGCGCTAATCCAA
 PAIR 54 PlsdFwrd54: CTTGGCTATTGTAACGCCATCA
 PlsdRvs54: CATCTGGAACA ACTGTCTCAT
 PAIR 55 PlsdFwrd55: AAGTCCCTTAGACTGGGATAG
 PlsdRvs55: TTGCTAGGTAAACCCCTGTTC
 PAIR 56 PlsdFwrd56: GTCGCAGGCACCATTTTCCATT
 PlsdRvs56: CCGTCACCTAAAGCAGGAACAG
 PAIR 57 PlsdFwrd57: TACTATATCAGCACCTGCTTC
 PlsdRvs57: TCAGTTGGTAGAGCAGTGGAC
 PAIR 58 PlsdFwrd58: CGGCATCAATAGCGCCTTCTG
 PlsdRvs58: TATTGATGCTGCTGAGTTGGG
 PAIR 59 PlsdFwrd59: CAAGTTGGTGGCCTTCATCTACA
 PlsdRvs59: CTGTTGGTCCCGAGTTAGGCA
 PAIR 60 PlsdFwrd60: TTTGTCGGCAACTATAATGCCT
 PlsdRvs60: CTGATGATGGTTCTGAGGCAG
 PAIR 61 PlsdFwrd61: CTAAGCCAGTAATTGTAGTCGT
 PlsdRvs61: AGAAGACACTCTGGTTATCCT

PAIR 62 PlsdFwrd62: ATCTGTCAGTACATTAGCTCCCT
 PlsdRvs62: TTCAAGGATTTGGAGCAACATC

PAIR 63 PlsdFwrd63: TGCCCCGTTACCTATTCCATA
 PlsdRvs63: AAGTGCTAGGCATCGTGGACT

PAIR 64 PlsdFwrd64: ACCTTCAACGCAAATAAACTCGTG
 PlsdRvs64: GAAGAGAAGATGGTATGAGCATTC

PAIR 65 PlsdFwrd65: TTACGAGCTCCACTATTATCTG
 PlsdRvs65: CATTCAACCAAGAGCACAAGGT

PAIR 66 PlsdFwrd66: CTTTGTTTGGTGATAAGCGTAG
 PlsdRvs66: ACACCTGGTACACGTAATAGAAC

PAIR 67 PlsdFwrd67: CTCTATTAGTTGCTCCTCTGA
 PlsdRvs67: ACCCTCCATTACAGCAACCAC



University of Fort Hare
Together in Excellence

A.2 APPENDIX B: Mitochondrial and plastid genome gap filling

Mitochondrial genome amplification

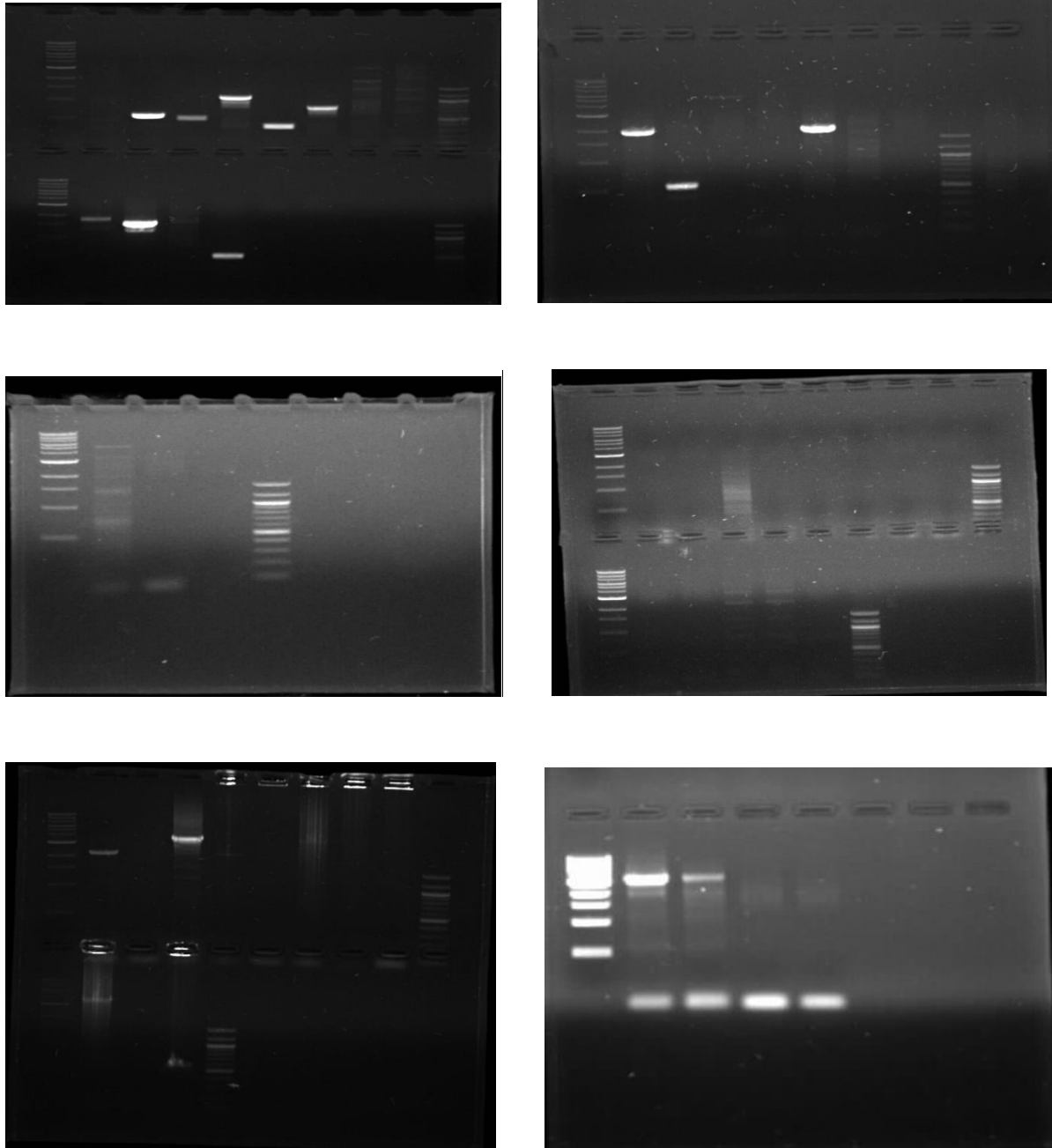


Figure referral 1. 1: Agarose gels obtained from PCR amplification of the *Gelidium pristoides* mitochondrial genome

All positive mitochondrial samples on one gel before gel cutting

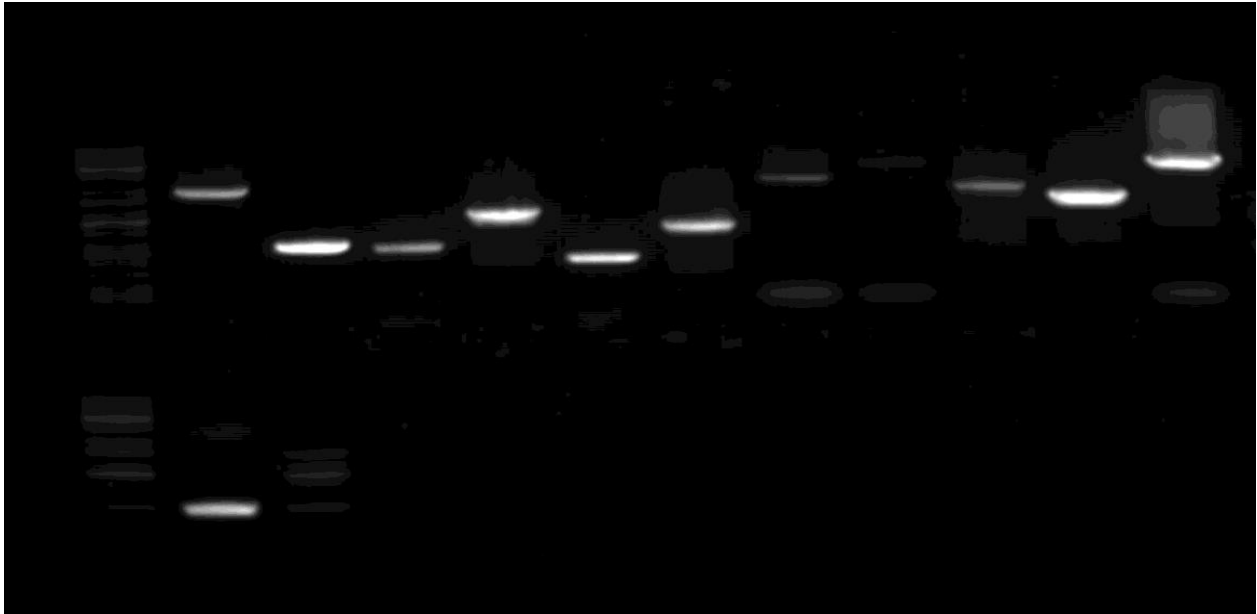


Figure referral 1. 2: Agarose gel electrophoresis (1% agarose) of PCR products of *Gelidium pristoides* mitochondrial genome.

Upper gel; **Lane 1:** 1 kb New England BioLabs DNA size marker **Lane2-12:** gap amplicon
Lower gel; **Lane 1:** 1 kb New England BioLabs DNA size marker **Lane 2:** gap amplicon 12
Lane 3: 100 bp New England BioLabs DNA size marker

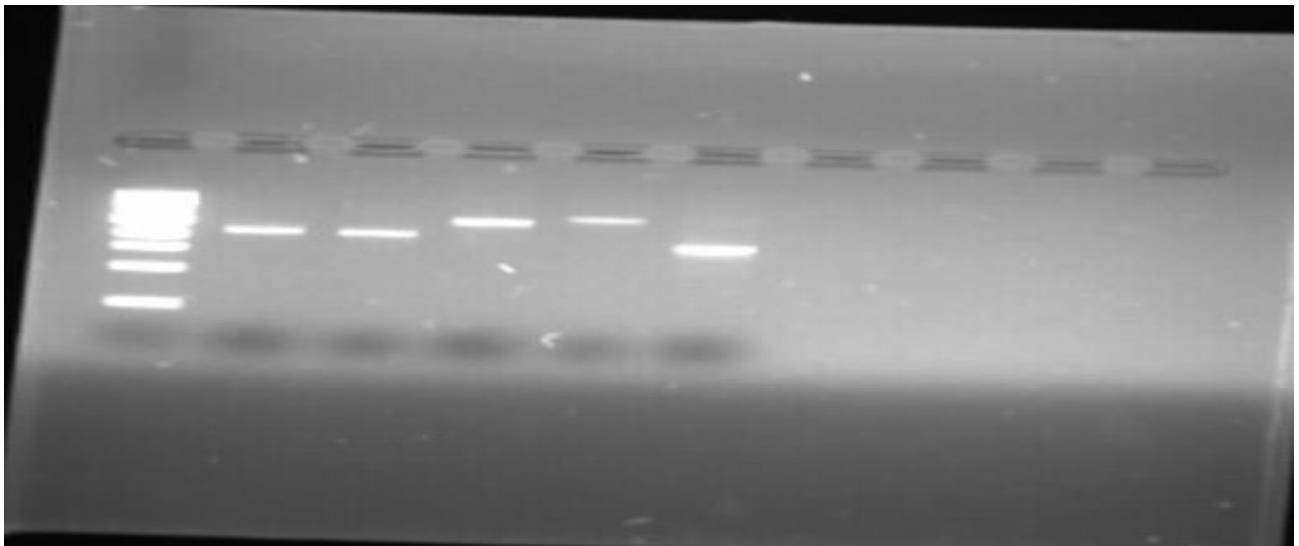
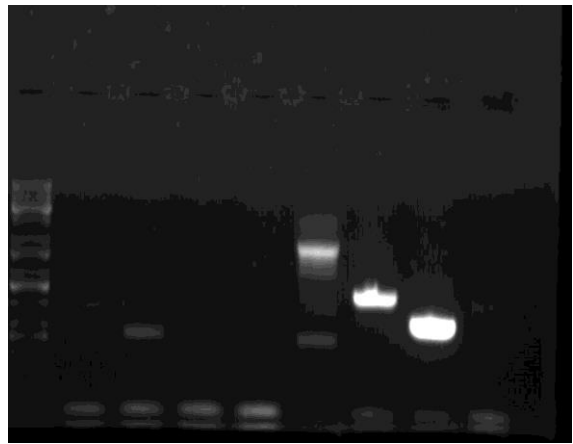
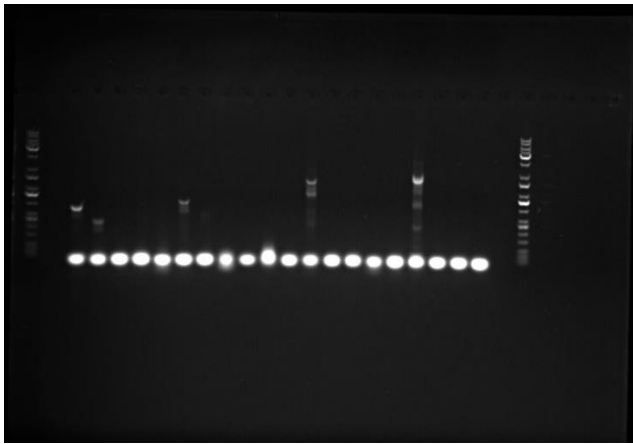
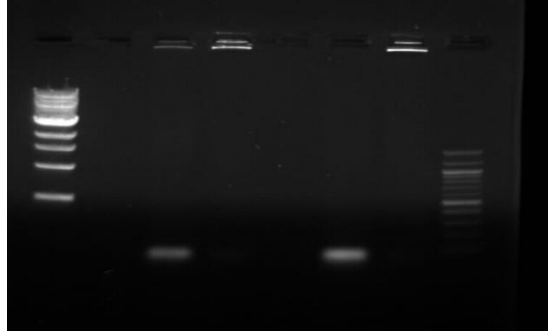
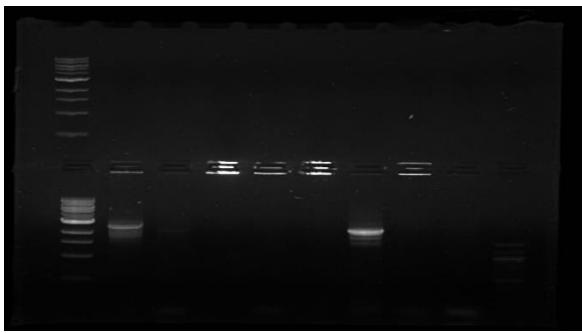
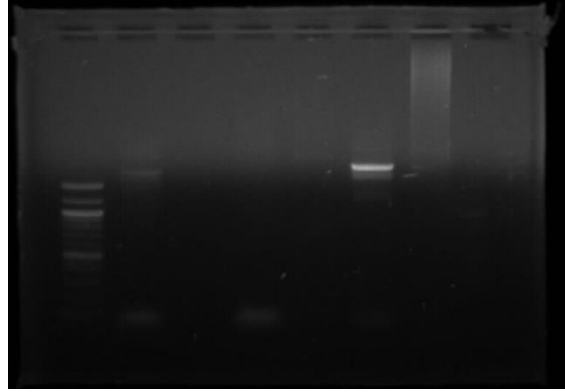
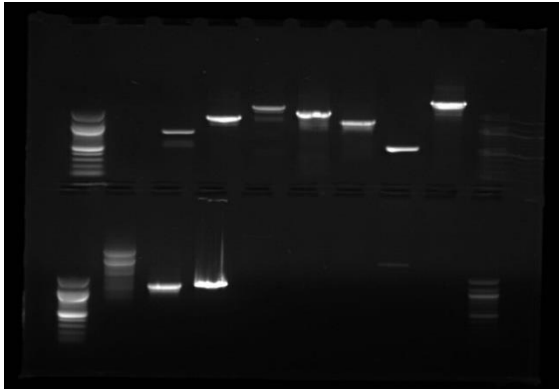
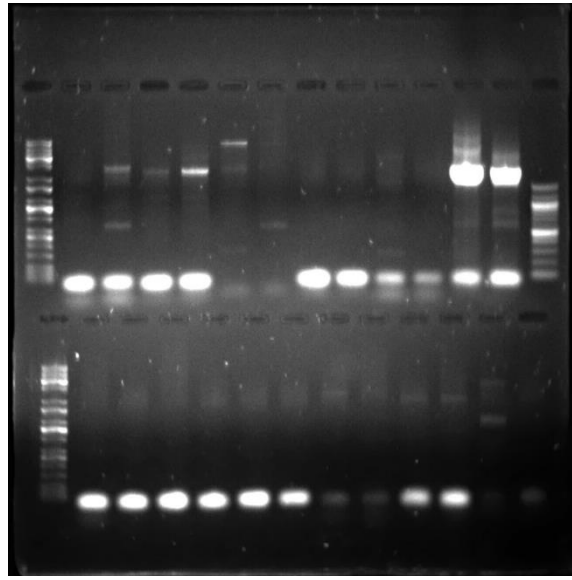
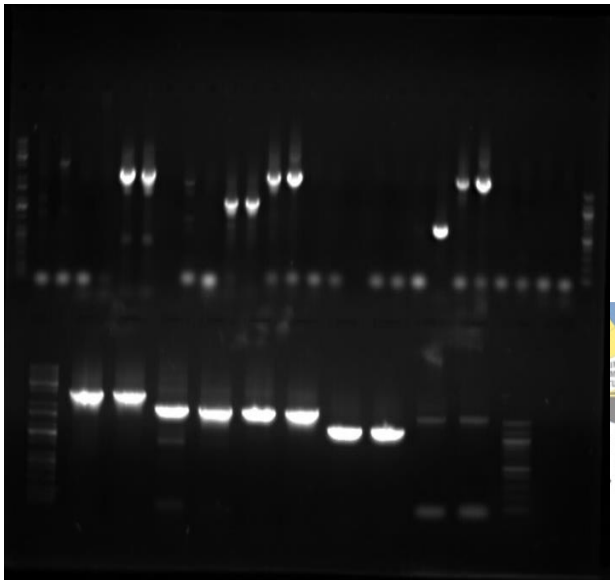
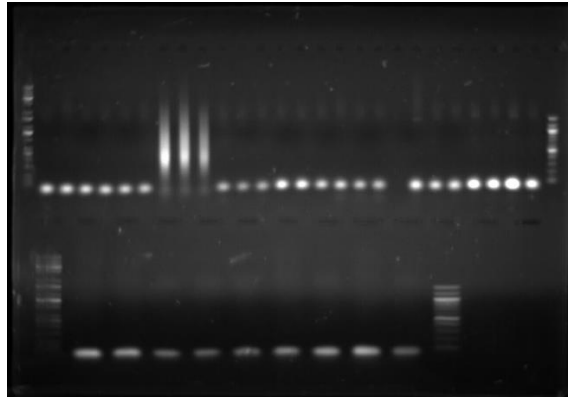
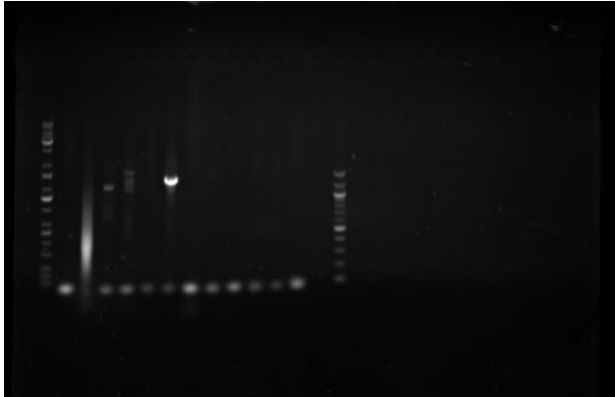


Figure referral 1. 3: Agarose gel electrophoresis (1% agarose) of gel purified PCR products of *Gelidium pristoides* mitochondrial genome. **Lane 1:** 1 kb New England BioLabs DNA size marker **Lane2-3:** gap amplicon 8 **Lane 4-5:** gap amplicon 11 **Lane 6:** gap amplicon 7

Plastid genome gap amplification





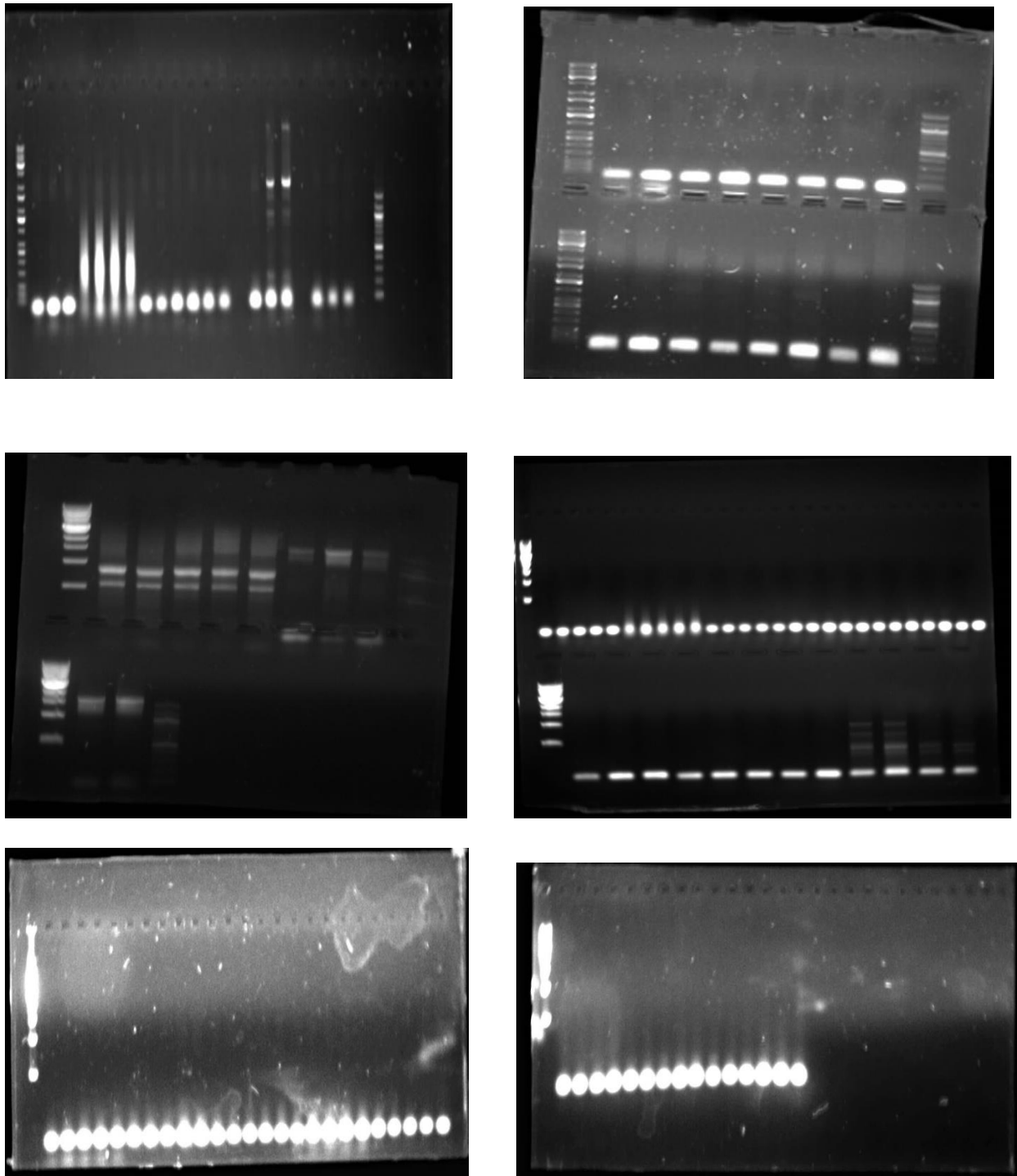


Figure referral 1. 4: Agarose gels obtained from PCR amplification of the plastid genome

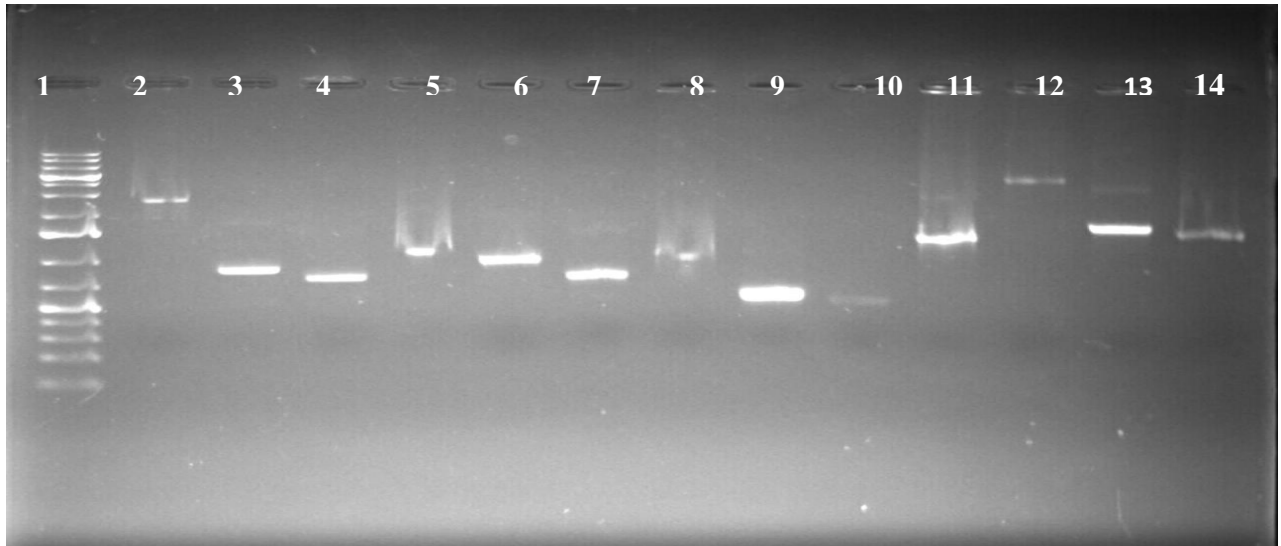


Figure referral 1. 5: Agarose gel electrophoresis (1% agarose) of PCR products of *Gelidium pristoides* plastid genome. **Lane 1:** 1 kb GeneRuler™ Plus DNA size marker **Lane 2-12:** gap 1-11 **Lane 13:** gap 13 **Lane 14:** gap 14

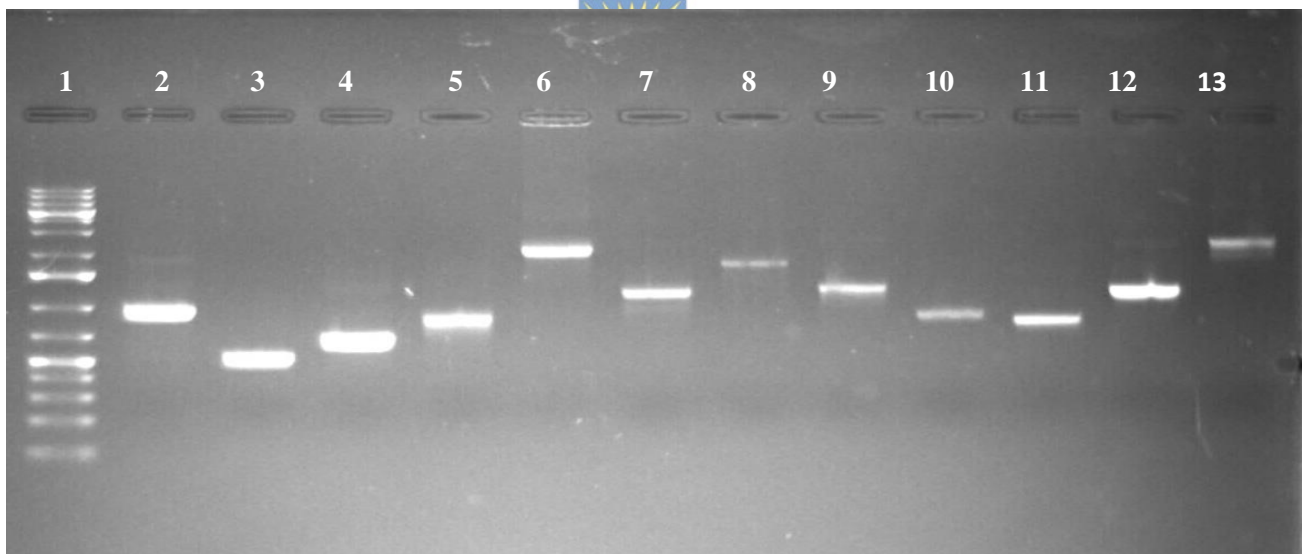


Figure referral 1. 6: Agarose gel electrophoresis (1% agarose) of PCR products of *Gelidium pristoides* plastid genome. **Lane 1:** 1 kb GeneRuler™ Plus DNA size marker **Lane 2-7:** gap 17-22 amplicons **Lane 8:** gap 25 amplicon **Lane 9-13:** gap 27-31 amplicons

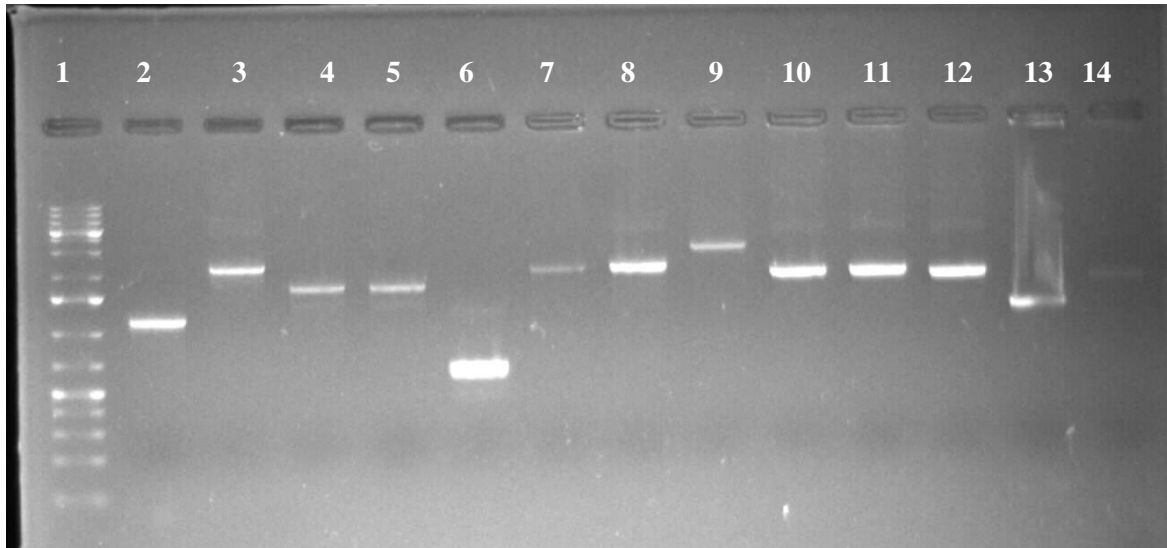


Figure referral 1. 7: Agarose gel electrophoresis (1% agarose) of PCR products of *Gelidium pristoides* plastid genome. **Lane 1:** 1 kb GeneRuler™ Plus DNA size marker **Lane 2:** gap 34 amplicon **Lane 3:** gap 38 amplicon **Lane 4-5:** gap 42 amplicon **Lane 6:** gap 50 amplicon **Lane 7:** gap 53 amplicon **Lane 8:** gap 54 amplicon **Lane 9:** gap 61 amplicon: **Lane 10-11:** gap 62 amplicon **Lane 12:** gap 65 amplicon **Lane 13:** gap 66 amplicon **Lane 14:** gap 67 amplicon.

Sample 4, 7, 11 and 66 were re-loaded in gel Figure referral 1.8.

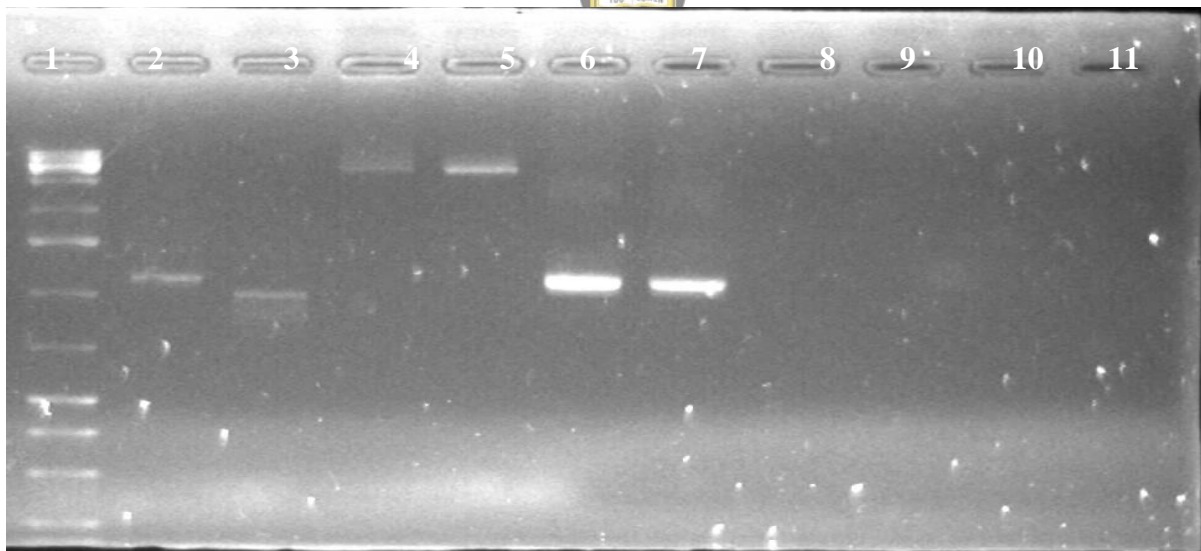


Figure referral 1. 8: Agarose gel electrophoresis (1% agarose) of PCR products of *Gelidium pristoides* plastid genome. **Lane 1:** 1 kb GeneRuler DNA size marker **Lane 2:** gap 4 amplicon **Lane 3:** gap 7 amplicon **Lane 4-5:** gap 11 amplicon **Lane 6-7:** gap 66 amplicon

A.3 APPENDIX C: Mitochondrial Annotation File

25012 DNA circular PLN 07-JAN-2019
DEFINITION *Gelidium pristoides* mtichondria, complete genome.

SOURCE mtichondria *Gelidium pristoides*
ORGANISM *Gelidium pristoides*
Eukaryota; Rhodophyta; Florideophyceae; Rhodymeniophycidae;
Gelidiales; Gelidiaceae; Gelidium.

REFERENCE 1 (bases 1 to 25012)
AUTHORS Mangali, S. and Bradley, G.
TITLE Sequencing, assembly and annotation of the mitochondrial
and plastid genomes of *Gelidium pristoides* (Turner) Kützing
from Kenton-on-Sea, South Africa

JOURNAL Unpublished

REFERENCE 2 (bases 1 to 25012)
AUTHORS Mangali, S. and Bradley, G.
TITLE Direct Submission
JOURNAL Submitted (07-JAN-2019) University of Fort Hare, Alice, 5700
School of Biological and Environmental studies, 2F005, Eastern
Cape, SA

COMMENT ##Assembly-Data-START##
Assembly Method :: SPAdes 3.11.1
Sequencing Technology :: Ion TorrentS5 and Sanger sequencing
##Assembly-Data-END##

FEATURES Location/Qualifiers
source 1.. 25012
/organism="*Gelidium pristoides*"
/organelle="mtichondria"
/mol_type="genomic DNA"
/country="SA: Algoa bay, Kenton-On-Sea, Shelly beach
Eastern Cape
/collection_date="14-July-2017"
/collected_by=" University of Fort Hare Plant stress-
Marine Science Research group"

>large subunit ribosomal RNA_rrl1_Gelidium pristoides_16..2570_2555bp
TTAAAGTCTAAAGAGTATTTAATAAAAAGTCTTGGTAAAAGTTTTAGGACGTCACACTACGAAAATTATAAGTAAGCT
TAGGCTTGTGATCTATAAATTTCTAAGAGAAAACCTCAAAAATTTATTTTAGAGTGAAAATGCGAATTGAATCATCT
TAGTAGCATTAAAAAATCAAACGAGAATTTGCGAGTATTGGCGAAAAAAGCAAATTAAGCCAAAATAATTTACTT
TGTGAAAACAACATTATAAAAGGTAATAATCCTGTAAACAAAGTATATTGGAAAGTAATACAAGTAACACGAGTACTT
TGTGTGAACATGAGGAGAACCACCTTCTAGAGCTTACAAAACCTTTTAACCGATAGTAAACGAGTATTGTGAAAGAAA
GACTAAACTGTCCGTTAGGAATCAAATCAATTGAATATTCTACATATTTTGAAGTTTTTATTAACCACGAGTTGC
CTTTTGCATAATGAATCAGCAAGTTAATACATATTGCAAGTATAAATTACAAAATAAACGTAAGTAATATTTATTA
GACCTGAAACCAAGTGATCTAGTTACAAACAAGCTGAAAATTTATTAATTAATTAAGGCTGTACCCACATATG
TAGCAAAATATGGGGATGATTTGTGATTAGGGGTGAAAGATCAATCAAACCTTGGAGATAGCCGGTTTTCCACGAAAA
GTATTTAAGTACTACATGATTATAAATAAATTATAGTTAGAGTTAACCAATTGAAAAAGGGGGTGTAAAAACTTACT
AAATTTAATAAATACTCCGAATAAATTTAATTAGTATATCATAGCCAGTCTTTAAGCGATAAGGTTTTAAAGACAAA
AGGAAAACAATCCAACCCATATACTAAGATCTTAAAATTATAACTTAGAGAAAAGACTTTTAGAAATTACAAATAAA
TTACAAGGTAAGCTTAGAAGCAGTTACCCATTAGAGAAAAGCGTTTTAGCTCACTATTCTTTTCTTGAAGTTAAAAAT
AATAGAGGCTATAAGTTATATATCGACGTAATGGATTAAGTTAATGGTAGTGGAAACACTCTGTAATTTAAATT
TATTTGAAAAAATTCAGAAAAAATTCAGAAAAGCGAATGCTGATATGAGTAGCGTAAATCTTGTCTAAATCAAGAT
CATTTTATGTTTAAAGGTTTTTAATGAAAATTCAGAAACATTAAGTTAGTCGGCCCTTAAAAACTAGGTGAAAGCTAA
ATTTGATAGGGATTTAATTTTAAATAAAACCATTTGTATGCAAAAACATGTGAAAGAAAATTATAAGATTTTCGATAGA
GCAACAAAAGGGTATTATCGAATCTACAATTTTCGAGAAAAAATTAGAATATAAGACCGTACTTAAACCGACACAGG

TAAACTGATTGAGAAAATTTAAAATGAGCGGGTAAATTATATTGAAGGAACTCGGCAAATTA ACTCTGTACGTTTCGCA
ATAAAGAGAACCTTACTAAAATTAAGGTAGCATAAAAATAAAAAGCAACGACTGGTTATCAAACCACAGGACTCTGC
TAATTTATAAAAAGTATAGAGTCTGATGTCTGCCAGTGCTTAAAAAAGAAAAAGATAAGTTAAAGCTTAAATTTGA
ATTCTAAGTAAACGGCGGTTCTAACTATAAGAATCCTTAGGTAGCGAAATTCCTTGACGGGTAAATCCGTCCTGCA
TGAATGGCTTAACGATTGCTTTACTGTCTCCAATATAAACCCGGCGAAATTACAAAACCTGTGAAAATGCAGGTTTC
TTACAGTAAGACGGAAAGACCCTAGATCCTTTACTCTAATTTTATATTGTAATAATTTTACAGCGCGTAGTGTAAG
TGGGAGTAAAAATACTAAAATGAAACACCCTCTCTGTTAAACCGTATTTACTTATTAAATTTATTTTAAAACAG
TGTA AAAAAGAGAGTTTACTTGGGACGAGTGCCTCCTAAAATGTAACGGAGGTGTACAAAGGTAGATTGAACGCAAT
GGTAATATCTGCTTGACAATTAGATTTTATAAATCAAATTGAGACGTAAGTCAGTCATAGTGATCCGATATTTAAGA
GTGGAGTTAATATCGCTCAACAGCTAAAAGGAACTCTAGGGATAACAGATTCATCGTGACCAAAAAGTTCGTATTGAC
GTCACGGTTTGATACCTCGATGTGACTCATCTTATCTGAAATGAAGAAGATTTCAAGGGTCTAGTTGTTTCGCTA
GTTAAAAGGTCAGTGAGTTGGGTTT CAGAACGTCGTGAGACAGTTTCGGTCCCTATCTACTGTAAGCACGAAAAATGA
AAAATTTATTTCTAGTACGAGAGGATCGAAGTAAGTTAACCTCTAGTGAATTTGATTGTTATACCTATAGCAAAGTC
TATTAGCCAAGTTAACTAAAATATAAGTACTGAAAGAATCAAATAGTACAAAATTTATTTCTCTATTTTTTCAAGAATA
ATCTAATCGATAATTAGATTGATCGATTATAAAGTATAAGTTTAGTAATAAATGTAACCTATAAATACGAATTTATT
CAAACCTTTAAATT

>ribosomal protein S3_rps3_Gelidium pristoides_2589..3287_699bp
ATGGCTCAAAAATTAATCCTATTGGGCTTAGACTTGGTATAAACCAAGTTTGAGATTCAACTATACAACAATATAG
TAAAAGCGTAGTTTACACATAAACTTTTTAAGAAATCAATTTTTTATTTCAGTCATTTTTTTACTCAACATACTAAAT
TTAGAAGCATTAGTTTACCTGTTCAAAAATATTTTATAATGAAGTACGATAATATTACCGTTTATTTTCATGTATAAA
GCTAATTCACTCTTATTAATAATAATTTCTCTAAAAAACTGATGAGAGAAACCCAATCAGTTCTAGAAAAAACAGT
TAAAGCTGAAACAAATATAAGATTGTCTTGTGTTACAAAATCTGTTTTTACTGCAAGTTTACTAGTGACTTACGTCA
ATTGTCTTTTTGAACAAAATACCTCCTTCAAAAAATTTATGTCAAACCTTGATTACGCTCTTAAAATATCAACTCGGA
TCGAAAAGGGTAATTTATTTAAAAAATGCTGCTCAAGAACTTAACTAGTAGGTTTTAAGGTAAAGATTTTCAGGGAG
ATTTGAAAATACTAGAAATAGAATGGCAAAAACCTTACGAACACACAGTAGGCTTTCTCTCTTTAACATGCCTTGATA
ATATTATAGAATTTAATAACCAAGTTATCTACACCAAGCTTGGTACTTGTAACTTTCAAATTTGATTAGTCTACAAA
AGCTAA

Translation=MAQKINPIGLRLGINQVWDSTIQQYSKRSKSLHINFLRNQFLFSHFFTQHTKFRSISLQVQKQYFIM
KYDNIIVYFMYKANSLLLNNNFSKMLMRETQSVLEKTVKAETNIRLSCVTKSVFTASLLVTVYNCLFEQNTSFKKIM
SNLITLLKYQLGSKRVIYLLKNAQELKLVGFVKVKSIGRFENTRNRMAKTYEHTVGFSLTCLDNIIEFNNQVIYTKL
GTCNFQIWLVIYKS

> Ribosomal protein L16_rpl16_Gelidium pristoides_3289..3696_408 bp
ATGAAGCAAGTTTACACTAAAACCTACAATAAATATAAGAGTAAAGGTAAAAAATTTTAAATTTTTTAACTTTAGG
TTCGTTTTGGATTTAAAGCAATTTTCGCGTAGTCGTGTAACAAGGGAGAAGTGAGATTCATTGCAATGAATCTTACGTA
AAAAATTTAAATCAAATTAATACCGAGCAAAAATAAATATGGGGTTTTAGTTGAACTAAATAACTCCTTGACAAAA
TTAACTTAGAATCAAGGATGGGCAAAGGTAAAGGCTTAATATATACTAAAAGTAAGTTTGTTCGTGAAGGGACACT
ATTATTTGAATTCGGCTCATTTCCTGAACAATTTCAAGATGAGATTATAAAATTTTTTGCAAAGTAAACCTTCATTA
AATTA AAAAAGGTAAGGTTTTAA

Translation=MKQVYTKTHNKYKSKGKKNFNFLT LGSFGFKAISRSRV TREKWDLSLQWILRKKFKSKLNTEQNKI
WGLVELNNSLTKLNLESRMGKGLIYTKSKFVREGTLLFEFGSFPEQFQDEI IKFLQSKTSLKLLKQVRF

>tRNA-Asp (GTC) _trnD_Gelidium pristoides_3701..3772_72 bp
GGGGGAGAACTCAAAGGTTGAGTGTTAATCTGTTCGCATTA AAAAGTTGCGGGTTCAAATCCCGTCTCTCTCG

>cytochrome c oxidase subunit 1_Cox1_Gelidium pristoides_3817..5418_1599 bp
ATGCTAACTCTTTCTACCAGTTGAATTTTCCGTTGAATTTTTTCAACAAATCACAAGATATTGGAACCTCTATACTT
AATTTTTGGTGCCTTTTCCGGTATTTTAGGAGGTTGCATGTCCATTTTAAATTAGAATGGAGCTAGCTCAACCAGGAA
ATCATTTATTGTTAGGAAATCATCAAATTTACAACGTGTTAATTACTGCCACGCCTTTTTGATGATTTTTTTTTATG
GTGATGCCTGTATTAATTGGAGTTTTGGTAATTGACTAATACCTATAATGATAGGCAGCCCTGATATGGCTTTCCC
CCGCTTAAATAATATTTTCGTTTTGGCTACTTCCCCCTTCTTTATGTTTACTTTTTAACGTCGGCAGTAGTAGAGGTAG

GTGTAGGAACAGGGTGAACGTATATCCTCCTTTAAGTTCGATTCAAAGTCATTTCAGGAGGTGCTGTTGATCTGGCA
ATTTTCAGCCTTCATATTTTCAGGAGCTTCTTCAATCCTGGGCGCAATTAATTTTATTTCAACCATTTTAAATATGCG
CAATCCTGGGCAAACAATGTACCGGATGCCTTTATTTGTTTGGTCAATACTTGTAAACAGCGTTTTTATTATTATTAG
CAGTACCAGTTTTAGCTGGAGCTATTACAATGTTGTTAACTGATCGTAATTTAATACATCATTTTTTCGATCCAGCT
GGAGGGGGTATCCTGTTTTATATCAGCATTATTCTGATTTTTTGGGCATCCAGAGGTATACATTTTAAATTTTACC
AGGATTTGGTATGATCAGTCATATAGTCTCTACTTTCTCAAGAAAACCTGTATTTGGCTACATTGGAATGGTATACG
CTATGGTATCTATAGGGGTTCTTGGTTTTATTGTTTGGAGCTCACCATATGTACACCGTTGGTTTTAGATGTAGATA
AGAGCTTACTTTACAGCTGCGACTATGATTATTGCAGTCCCAACAGGAATAAAAATATTTAGTTGGATAGCTACAAA
GTGAGAGGGGTCAATTCATTTTAAAACCTCTATGTTATTTGCTATAGGTTTTATCTTTTTTATTCACTATAGGCGGTC
TAACAGGAATTTGTTAGTCAACTCGGGATTAGATATTAGTTTACATGATACTTACTATGTAGTAGCTCATTTCCT
TATGTATTATCTATGGGTGCCGTATTCGCTATTTTTGCTGGCTTCTATTATTGATTTGGAAAAATTTACTGGTGTACA
ATATCCTGAAATATTAGGCCAAATTCATTTTTGATCAACATTCATAGGGGTAAATTTAACTTTTCATGCCTATGCATT
TTTTAGGATTAGCGGGTATGCCTAGAAGAATTCAGATTATCCGGATGCTTATACAGGTTGAAACTTAATAGCGTCT
TATGGGTCGTATATTGCGGCTTTTTCTACTTTATTCTTTTTCTACCTAGTTTTTATCTCTCTGACTTCAAGTAATCC
TTGCGTAAGTTCCCCTTGAGATTTTGGTGAAGATACACAAATACAAGGTAAATCAAGTCTAGAATGAGCTGTAACCT
CCCCCAGCGTACCACACATTTGAAGAAATGCCGTTAATTTGCGAAACTACAAAATA

Translation=MLTLSTSWIFRWIFSTNHKDIGTLYLIFGAFSGILGGCMSILIRMELAQPGNHLLGNHQIYNVL
ITAHAFLMIFFMVPVLIIGGFGNWLIPIMIGSPDMAFPRLNNSIFWLLPPLCLLLTSAVVEVGVGTGWTVPPLSS
IQSHSGGAVDLAIFSLHISGASSILGAINFISTILNMRNPGQTMYRMPFLVWSILVTAFLLLLAVPVLGAIITMLLT
DRNFNTSFFDPAGGGDPVLYQHLFWFFGHPEVYIILILPGFGMISHIVSTFSRKPVFGYIGMVYAMVSIIVLGFIVWA
HHMYTVGLDVDTRAYFTAATMIIAVPTGIKIFSWIATKWEGSIHFKTPMLFAIGFIFLFTIGGLTGIVLANSGLDIS
LHDTYYVVAHFHYVLSMGAVFAIFAGFYWFGKITGVQYPEILGQIHFWSTFIGVNLTFMPMHFLGLAGMPRRIPDY
PDAYTGWNLIASYGSYIAAFSTLFFFYLVFISLTSNPNVSSPWDFGEDTQIQGKSSLEWAVTSPPAYHTFEEMPLI
CETTK

>Cytochrome c oxidase subunit 2_Cox2_ *Gelidium pristoides*_5425..6267_783 bp
ATGATAAACTATTAATACTCTTTCCCTTTAATGTTTATACCGACTGAAAGTTTAAAGTGATGCAGCTGAAAATTG
ACAGTTGGGTTTTCAAGATCCGGCAACACCAATTATCGAAGCTATTGTAAATCTACATCATGATTTAATGTTCTTCA
TTTGTGTAATATCGGTATTTGTAACCTGAATGTTGGGCCGTACGTTGTGACATTTAATCAACGTCAAAAATAAACC
CCTTCGTCTTTATCTCATGGAACACTAATTGAAATCGTTTGAACAGTAGCTCCAGCCTTTATATTATTAATTATTGC
AGTCCCATCTTTTTCTTTATTATATGCTATGGATGAGGTAATATCACCTGCAATTACAGTAAAAACATTAGGTCACC
AATGGTACTGAAGCTATGAATATTCTGATTATCTAAATAAAGAAAATGAGGCTATTATGTATGATAGTTATATGGTT
CCAGAAGAAGACTTACAAACAGGTCAGCTTAGATTACTGGAAGTTGATAATCGCATGGTCGTTCCCCTTAATACTCA
TATAAGGTTGATTGTTTTCAGCTGCTGATGTTTTACATAGTTGAGCAATACCTTCGCTAGGAATTAATGCGATGCCA
TTCCAGGTCGTTTAAACCAAACCTTCTTTTTATAAAAAGAGAAGGTGTTTACTACGGGCAATGCAGTGAAATTTGT
GGTATAAATCATGGGTTTATGCCTATTGTTGTGGAAGCCGTTTCCCTTACCAGCAATATATTTCTTGGGTTTTCAAATA
ATTGAGTGAATAA

Translation=MIKLLILFSLLMFIPTESLSDAENWQLGFQDPATPIMEGIVNLHHDLMFFICVISVFTWMLGR
TLWHFNQRQNKTPSSLSHGTLIEIVWTVAPAFILLIIAVPSFSLLYAMDEVISPAITVKTLGHQWYWSYEYSYDLNK
ENEAIMYDSYMPVEEDLQTGQLRLLLEVDRMVPVNTHIRLIVSAADVLHSWAIPSLGIKCDIIPGRLNQTSLFIKR
EGVYYGQCSEICGINHGFMPIVEAVSLPQYISWVSNKLSE

>Cytochrome c oxidase subunit 3_Cox3_ *Gelidium pristoides*_6372..7191_
819 bp
ATGACTCTTTTATCTCAAATTTCTAAATCTGTACAGCGCCATCCTTTTTCACTTAGTGGATCCAAGCCCATGACCCTT
TGTAGCTTCCCTAGCAGCTTTTTCTTGC GCGGTTAGCGGAGTAATGTATATGCATGCTTTTAAAAGAGGTGGTTTTT
CTTTACTAATTAGCTTTATATCATTATTAATTATTATGTTTGTGTTGATGAAGGGATGTAATAAGAGAAGCAACCTTT
GAAGGTCATCATAACAGGGATCGTACAACAAGGTTTAAAGATATGGAATAATTCTTTTTATTATCTCTGAAATTTTGT
TTTTTTTTGCGTTTTTTTTGGGCGTTTTTTCATAGTAGCTTGTGCGCTGGAGTTGAGATAGGTTCAATATGACCTCCGA
AAGGTATAAGTGTATTGATCCGTGAGAAATTCCTTTTTTGAATACGTTAATTCTCCTCTTATCTGGATGTACAGTC
ACTTGATCTCATCACGCAATAGTAGCAAATTTGCGTTTTCAAGCATTATTAAGTTTTATTTTAACTGTAATTTTAGC
GGTAATATTTACAATTTTACAAGCTTATGAATATACTTTGGCAGATTTTCAGACTTTTCAGATGGGATCTACGGCTCTA

CTTTTTATATGGCTACTGGATTTTCATGGATTTTCATGTTTTTCATTGGCACTGTTTCCTTACTAATTTGTTTTATTTCGC
TTAAATCAACACCAATTAACCTCAGCAACACCATTTTGGCTTTGAATCAGCAGCCTGGTATTGACATTTTCGTAGATGT
TGTATGGTTATTTTTTATTTGTGTCAATTTACTGATGAGGAGGTCTCTAA

Translation=MTLLSQISKSVQRHPFHLVDPSWPWFVASLAAFSCAVSGVMYMHAFKRGGFSLLSIFISLLIIMF
VWWRDVIREATFEGHHTGIVQQGLRYGIILFIISEILVFFAFFWAFHSSLSPGVEIGSIWPPKGISVIDPWEIPFL
NTLILLSSGCTVTWSHHAIVANLRFQALLSFLTVILAVIFTILQAYEYTLADFRLSDGIYGSTFYMATGFHGFHV
IGTVSLLICFIRLNQHQLTQQHHFGFESAAWYWHFVDVWVWVFLFVSIYWWGGL

>ATP synthase B chain precursor_ymf39_ *Gelidium pristoides*_7195..7737_543 bp
ATGCTTAACATTACTATACTTATTTTAATATTATTAATATTAATATCACAAAATATTTTATTATTGAACGAAGAATC
ACTAATTTTGGCTTTGCTTTACAGTGTCTTACTTGATTAATATATAATAAACTTAAATTATTACTTGAGCAAGATTTGA
ACGCCAAAGCACTGTCTATCTACAATTCAATTGAATCTTCATTTAACTCCACAATACTTTTATTAAAGCAAGAATA
GATAATCAACAAGAAGTTGAATCTTTTAAACAGTAGTTTTTGGCTTTAAAGCGACACTTTAGTAAATTAACCTTCTTT
TGCAGTTTTTAACTTAAAACAATTCACGTTAGGCAATCTTCAGAAATCATACAAAAAAATTAACCTTTCACGCAAC
AGCTAGAAAAACAACTTCAAACTGTTAGCTTTTTTACTAGTAAAAAGCTTAATAAAATTGCAAATTTAAATCAG
TACTATTTTAAAGATTTTCCGTTAAATACATGAAGCTGCCTTTATAAAATCTCGTTAAGAGAGTACTTAGAACTAT
TTAG

Translation=MLNITILILILLILISQNILLLNEESLILLCFTVFTWLIYNKLLKLLLEQDLNAKALSIYNSIESS
FNSTILLKQELDNQOEVSFNSSFLALKRHFSLKTSFAVLNLKQFTLGNLQKSYKKKLTFTQOLEKQTSKLLAFLL
VKKLNKIANLNQYYFKDFPLNTWSCLYKISLREYLETI

>tRNA-Gly _trnG_ (TCC) _*Gelidium pristoides*_7739..7811_73 bp
GCGGGTATAGAAAACCTGGTAATTTTCGTTAGTTTTCCAAGCTAAATTTACGGGTTTTCGAATCCCGTTATCCGCT

>tRNA_Gln_trnQ_ (TTG) _*Gelidium pristoides*_7822..7893_72 bp
TGGTGTATAGCCAAAAGGTAAGGCGATGACTTTTGATGTCAATCATGTAAAGGTTTCGAATCCTTTTACACCAG

>tRNA-Lue_trnL_ (TAA) _*Gelidium pristoides*_7949..8025_83 bp
ACTCGCTTGGTGAAAATAGGTAAACACGATGGATTTAAATTCATTCCTAAATGGGTTATTGGTTCAAGTCCAATAG
CGAGTA

>Apocytochrome b_cob_ *Gelidium pristoides*_8082..9233_1152 bp
ATGCGTCTAGCTAAACGCCCATTAATTTCTATTGTCAACAATCACTTAATTGATTACCCGACCCCAATAAATATTCA
TTATGCATGAAATTTTGGATTTCTTTTCAGCAATGTGTTAATTATACAAATTTTAACTGGTATTTTTTTAGCTATGC
ATTACACTCCCCATGTTGATTTAGCTTTTCGCAAGCGTAGAGCATATTATGCGTGATGTCAATTATGGGTGGTTACTT
CGTTACATTGATGCAAATGGCGCGTCTATGTTTTTTGTTGTCGTTTACATTCATATATTTTCGAGGACTTTATTTTGG
ATCTTATACTAAACCACGACATTGAGTTTGGTGTGGGTGTAATTATATTCTTTTTAATGATGGGTACAGCTTTTAA
TGGGTTATATTTTACCTTGAGGTCAAATGAGTTTATGAGGTGCTACGGTAATAACCAATTTAGCATCAGCAGTTCCCT
TTTATCGGGGATTACATTGTAACATGATTGTGAGGAGGTTTCTCTGTAGATAACGCAACTTTAAACCGTTTTTTTAG
CCTTCATTATTTACTACCATTTGTAATTGCTGCTATTACACTATTACATTTAGCAGTTTTACACCAAGATGGCTCAG
GAAACCCATTAGGCATAGAATCAAATGTTGACAAAGTTACCATGTTCCCGTATTTTTATAGTCAAAGATTTTTTGGGA
TTAATTGTTTTTTAATTGTTTTTTCTATTTTTTATTTATTTTTTACCCAATGTTCTTGGCCATGCAGATAATTATAT
TGAAGCTAATCCAATGGTTACCCCTGCGCATATTGTCCCTGAATGATATTTTTTACCCTTCTACGCTATACTAAGGA
GTATACCCCATAAATTAGGAGGTGTTATTGCTATGATTGCCGCAATTTAATTTTAGCTTTTTTACCATGAATACAC
AGTACTGAAATTAGAAGTTCAAGATTTAGACCAATTTATAAATTTCTTTTTTGGACAATGGCTAGTTCTTGCTTTAT
TCTAGGGTGAATTGGCGGAATGCCAGTGGAAGAGCCGTATGTACTAATAGGTCAGGTTGCTAGCGTATATTACTTTT
GCTATTTTTTTATTTGTACTACCTTTACTAGGTAATAATTGAGCGCTTTTTTATTAGAATTCATAACCCAAAAAATAG

Translation=MRLAKRPLISIVNNHLIDYPTPINIHYAWNFGFLSAMCLIIQILTGI FLAMHYTPHVDLAFASVE
HIMRDVNYGWLLRYIHANGASMFFVVVYIHFRLYFGSYTKPRHWVWVWLVGVIIFLMMGTAFMGYILPWGQMSLWG
ATVITNLASAVPFIGDYIVTWLWGGFSDVNDATLNRFFSLHYLLPFVIAAITLLHLAVLHQDGSNPLGIESNVDKVT
MFPYFIVKDFLGLIVFLIVFSIFIFYSPNVLGHADNYIEANPMVTPAHIVPEWYFLPFYAILRSIPHLKGGVIAMIA
AILLILAFLPWIHSTEIRSSFRPIYKFLFWTMASSCFILGWIGGMPVEEYPVVLIGQVASVYVFCYFLVPLLLGKIE
RFLLEFNTQK

>tRNA-Lue_ (TAG) _ *trnL_Gelidium pristoides* _ Complement_9271..9358_88 bp
ACGGATATGATGAAATAGGTaGACGTCTTAGGTTTAGATTCTAAGGGCTTGTTatAAAGCTgTGAGGGTTCGAACCC
CTCTGTCCGTA

Reverse complement

TACGGACAGAGGGGTTTCGAACCCCTCAcAGCTTTatAACAAGCCCTTAGAATCTAAACCTAAGACGTCTACCTATTTTC
ATCATATCCGT

>NADH dehydrogenase subunit 6_ *nad6_Gelidium*

pristoides _ Complement_9359..9967_609 bp

TTACTTTCTTAACTTAATAAAATTTAATGGCCTCTTTTGGATTCTTATAAGTTGTTTTTCAATAAGTTGTTTTTTTAA
CATCAGTTCTTTGATGCATTGTTAAAACATTACCCCAATCATCGCTATTAGTAATATTAACCTAGATAATAGAAAT
ATTAACCTATAATTAGTGTAAGCACATTTCCGATAACTTTAATGTTTGATAAAAAACTTTCTTCCGCTAATCAATG
TACTAAGGTTGGTTGTTGGTGCATTAATCTATACGTATATAGTTCTTTAGTAATTAATTAATTTGGTAAAAAAAC
AACCAAAAATGACTACCCCTAAGGGAACATAAGACAGAGCTTAGATTTAATAAATCAATTTTTTACATTTAACATC
ATTACAACAATAAAAAATAAACTGCAATTGCTCCAACGTAAACAATAATTAATAGGAGCGATAAAAAATTCTGCTCC
TAGTAATAAAAGTAGACTTGCTACATTGCAAAATACTAGAATTAAAAAAGAAGTGAATGTACTGCATTTTTTTAAGG
TTATTACCATTAAAGACGATATCAATGCAAAGAATGAAAAATAGAAAATAGAAAAGTTTCTACAGTCAT

Reverse Complement

ATGACTGTAGAACTTTTTCTATTTTTCTATTTTTCTATTCTTTGCATTGATATCGTCTTTAATGGTAATAACCTTAAA
AAATGCAGTACATTCAGTTCTTTTTTAAATTTCTAGTATTTTGCAATGTAGCAAGTCTACTTTTTATTACTAGGAGCAG
AATTTTTATCGCTCCTATTAATTATTGTTTACGTTGGAGCAATTGCAGTTTTATTTTTATTTGTTGTAATGATGTTA
AATGTAAAAATTGATTTATTAATCTAAGCTCTGTCTCATTAGTTCCCTTAGGGGTAGTCATTTTTGGTTGTTTTTT
TTACCAATTTAATTTAATTAATACTAAAGAATATATACGTATAGATTAATGCACCAACAACCAACCTTAGTACATTGAT
TAGCGGAAGAAAGTTTTTTATCAAACATTAAGTTATCGGAAATGTGCTTTACACTAATTATAGTTTAAATATTTCTA
TTATCTAGTTTAAATATTACTAATAGCGATGATTGGGGTAATAGTTTTAACAATGCATCAAAGAAGTGAATGTAAAAAA
ACAACCTATTGAAAAACAACCTTATAAGAAATCCAAAAGAGGCCATTAATTTTATTAAGTTAAGAAAGTAA

Translation=MTVETFLFSIFSFFALISSLMVITLKNVHVSFLVFLILVFCNVASLLLLLLGAEFLSLLLLIIIVYVGA
IAVLFLFVVMMLNVKIDLLNLSVSLVPLGVVIFGCFYQFNLIKELYTYRLMHQQPTLVHWLAEESFLSNIKVIG
NVLYTNYSLIFLLSSLILLIAMIGVIVLTMHQRTDVKKQLIEKQLIRNPKEAIKFIKLRK

> tRNA-Gly_ (GCC) _ *trnG_Gelidium pristoides* _ Complement_9983..10056_74 bp
GCGAAAATAACTCAATTAGGTAGAGTATAATCTTGCCAAGATTAAAGTTGAGGGTTCGAATCCCTTTTTTCGCT

Reverse complement

AGCGAAAAAGGGATTTCGAACCCCTCAACTTTAATCTTGCAAGATTATACTCTACCTAATTGAGTTATTTTCGC

>tRNA-His_ (GTG) _ *trnH_Gelidium pristoides* _10061..10132_72 bp
GCGGAGAAAGCTTAAAGGTTGAGCGTTAGATTGTGAATCTAGAGGTCATGGGTTCGAATCCCATTCTTCGCC

>Succinate: cytochrome c oxidoreductase subunit

2_ *sdh2* _ Complement_10132..10878_747 bp

CTAAGTAACCAAAAATTGCTTGTTTTATGTTAGTTATAGCCTTACCCGGATTTAAGTTTTTGGGGCAAGTTTTACTAC
AATTCATAATTGTGTGACAACGAAATAGTCGAATTTTTATGATTTAAAAAACTAAGTCTATAATTAGTATTTGAATCT
CTAGAAATCAGCAATCCATCTATAAGATTGCAACAAAATTGCAGGTCCTAAGTAAATATTTTATTCCACCAATAACT
AGGGCAGCTTGCAGAGCAGCAAGCACATAAAATACATTCGTACAACCCGTCAAGCTCTAACCTATCTTTTTTAGATT
GCAGCTGCTCTTTCTTTGGTAATAAATTATTAACATAATCAAGTTTTATTGACTTATATTGACTATAAAAGTTTTGCC
AAATCTACAATTAATCTTTTACAACATACATATGAGGTAAAGGATAAACAGTAATAAATTTAGTTTCTTGATTTAA
TGGCTTTAAACACGCAAGAGAATTAACCCCGTTTTATGTTTCATTGAGCAACTACCACATATACCTTCCCTGCATGAAC
GCCTAAAGCTGAGAGATGAATCTTGCTTATTCTTAATTTGAAACAACGCATCAAGAATCATAGGTCCGCAAGCTTTT
AATGAAATAGGGTATATATTAATCAAGGCTTCTGTTGATAATGCGGATTTTCATCGATATACTCTTAAATAACGATG
ATAATTTGATTCGGTTTTGAGTTTAAATTTTTATGTGAATCTTTAATTAATAAACAT

Reverse Complement

ATGTTTTTAAATTAAGATTACATAAAATTAAGTCAAACCCGAATCAAATTATCATCGTTATTTAAGAGTATATCG
ATGAAATCCGCATTATCAACAGAAGCCTTGATTTAATATATACCCCTATTTTATTAAAAGCTTGCGGACCTATGATTC
TTGATGCGTTGTTTTCAAATTAAGAATAAGCAAGATTATCTCTCAGCTTTAGGCGTTTCATGCAGGGAAGGTATATGT
GGTAGTTGCTCAATGAACATAAACGGGGTTAATTCTCTTGCCTGTTTTAAAGCCATTAAATCAAGAACTAAATTTAT
TACTGTTTTATCCTTTACCTCATATGTATGTTGTAAGATTAAATTGTAGATTTGGCAAACCTTTTATAGTCAATATA
AGTCAATAAAACCTTGATTAGTTAATAATTTATTACCAAAGAAAGAGCAGCTGCAATCTAAAAAAGATAGGTTAGAG
CTTGACGGGTTGTACGAATGTATTTTATGTGCTTGTGCTCTGCAAGCTGCCCTAGTTATTGGTGGAAATCAAAATAT
TTACTTAGGACCTGCAATTTTGTGCAATCTTATAGATGGATTGCTGATTCTAGAGATTCAAATACTAATTATAGAC
TTAGTTTTTTAAATCATAAAATTCGACTATTTTCGTTGTACACAATTATGAATTGTAGTAAAACCTTGCCCCAAAAC
TTAAATCCGGGTAAGGCTATAACTAACATAAAACAAGCAATTTTGGTTACTTAG

Translation=MFLIKDSHKIKLKPESENYHRYLRVYRWNPHYQQKPFWNIYPIISLKACGPMILDALFQIKNKQDSS
LSFRRSREGICGSCSMNINGVNSLACLKPLNQETKFITVYPLPHMYVVKDLIVDLANFYYSQYKSIKPWLVNLLPK
KEQLQSKKDRLELDGLYECILCACCASPCSYWVNQNIYLGPAILLQSYRWIADSRDSNTNYRSLFNHKIRLFRCH
TIMNCSKTCPKNLNPGKAITNIKQAILVT

>Succinate: cytochrome c oxidoreductase subunit 3_sdh3_ *Gelidium pristoides*
Complement 10880..11263_384 bp

TTATATAAATAAATTAGAAAGTAGAAATAGACTTAAAGTAAAACATACACATAAGGTAAAATTAATAGATGGTTTT
GATGTAGCAGAAAACCTAAATCTCAGATTAAATAACGTACTCCGCTAAGAAAATGATATAAAAAAACTGACATAAAT
AAAACCACTAAAAGTTTTTTTTATTTTCGTGAGTTACATTAAAAACCTAAAATAATCTCTTGAATACTACAAGAAAATAT
AGAGTTAGTAAAATTAAGTATAGAACTAAAGAAAATAGTAACTAACTCCTGAAAGGCGATGTCAGATTGATGACA
AGGACGATAGTTGTGGTTTTATATACTGTCAAATGGGGGGATAAAGGACGATTCAAAGTAAAAGTATTGAAGCAT

Reverse complement

ATGCTTCAATACTTTTTACTTTTTGAATCGTCCTTTATCCCCCATTTGACAGTATATAAACCACAACCTATCGTCCTT
GTCATCAATCTGACATCGCCTTTTCAGGAGTTAGTTTACTATTTTTCTTTAGTTCTATACTTTAATTTTACTAACTCTA
TATTTTCTTGTAGTATTCAAGAGATTAGTTTGTAGTTTTAATGTAACCTCACGAAATAAAAAAACTTTTAGTGGTTTTA
TTTTATGTCAGTTTTTTTTATATCATTTTTCTTAGCGGAGTACGTTATTTAATCTGAGATTTAGGTTTTCTGCTACATCA
AAACCATCTATTTAATTTTACCTTATGTGTATGTTTTACTTTAAGTCTATTTCTACTTTCTAATTTATTTATATAA

Translation=MLQYFLLLNRPLSPHLTVYKQQLSSLSSIIWHRLSGVSLLLFSLVLYFNFTNSIFSCSIQEIISLFSN
VTHEIKKLLVFLFMSVFLYHFLSGVRYLIWDLGFLHQNHLFNFTLCVCFTLSLFLLSNLFI

>tRNA-Phe_trnF_(GAA)_*Gelidium pristoides*_Complement_11281..11353_73bp
GTCTAGGTAGCTCAGTAGGTAGAGCATAGAATTGAAGATTCTTGAGtCATGGGTTCGAATCCCATTTCTGGACA

Reverse complement

TGTCCAGAATGGGATTCGAACCCATGACTCAAGAATCTTCAATTCTATGCTCTACCTACTGAGCTACCTAGAC

> tRNA-Ser_trnS_(TGA)_*Gelidium pristoides*_Complement_11358..11446_89 bp
GGGTGAGTGGTTGAGTGGTTGAAAGCACCAATTTTAAAATTGGCATAAACTAAAACGTTTATCGTAGGTTTCAATC
CTACTTCACCTA

Reverse complement

TAGGTGAAGTAGGATTCGAACCTACGATAAACGTTTTAGTTTATGCCAATTTTCAAATTGGTGCTTTcaACCACTC
AACCACTCACCC

>tRNA-Pro_(TGG)_trnP_ *Gelidium pristoides*_Complement_11454..11526_73bp
TAGGGTATAACGAAATAGGTATCGTGTGTTTTGTTTGGGAACAAAAGTTACAGGTTTCAATCCTGTTACCTAA

Reverse complement

TTAGGGTAACAGGATTCGAACCTGTAACTTTTTGTTCACAAAACAAACACGATACCTATTTTCGTTATACCTA

>ATP synthase F0 subunit 9_atp9_Gelidium
pristoides_Complement_11537..11767_231 bp

TTAAGTAAATAAAATTAGAAAAGCCATCATTAAAGCAAATAAAGCTACAGCTTCTGTTAAAGCAAACCTAAAATTG
TGTATCCAAATAATTGTTGTTTTAATGAAGGATTTTCGTGCATACGCCATTACTAATGACCCAAATACAATTCACACC
CCTGCACCAACTCCAGTTAAGCCGATTGTAGCTAAACCTGCACCAATCATTTTTCGCACTTTGAAGGGTAACGTTTCAT

Reverse complement

ATGAACGTTACCCTTCAAAGTGCGAAAATGATTGGTGCAGGTTTAGCTACAATCGGCTTAACTGGAGTTGGTGCAGG
GGTTGGAATTGTATTTGGGTCAATTAGTAATGGCGTATGCACGAAATCCTTCATTAACAACAATTATTTGGATACA
CAATTTTAGGTTTTGCTTTAACAGAAGCTGTAGCTTTATTTGCTTTAATGATGGCTTTTCTAATTTTATTTACTTAA

Translation=MNVTLQSAKMIGAGLATIGLTGVGAGVIVFGSLVMAYARNPSLKQQLFGYTIILGFALTEAVALF
ALMMAFLILFT

>TRNA-Cys_trnC_ (GCA) _Gelidium pristoides_ Complement_11804..11874_71 bp
GGCTAGATGGCAGAATGGTTATGCAAAAGATTGCAAATCTTATTaTATTGGTTTCGATTCCAATTCTAGCTT

Reverse complement

AAGCTAGAATTGGAATCGAACCAATATAATAAGATTTGCAATCTTTTGCATAACCATTCTGCCATCTAGCC

>tRNA-Met_trnM_ (CAT) _Gelidium pristoides_ Complement_11877..11951_75 bp
AGTGAATAGTTCAACTTGGTTAGAACATTGGAATCATAATCTAAAAGTTGCGGGTTCGAATCCTGTTTTCTACTA

Reverse complement

TAGTGA AACAGGATTTCGAACCCGCAACTTTTAGATTATGATTCCAATGTTCTAACCAAGTTGAACTATTCCACT

>Ribosomal protein S11_rps11_Gelidium pristoides_ 11954..12313_60 bp
TTAAATTCTTTTTATTTTTTCGTGCTTTACAACCATTATGGGGTAAAGAGGTTTGATCACAAATAGAAAGAATTTTTTA
CTTTACCTTGTTTGAAAATTTTTAGGATTTGTTTTTATTTTTGCTGAAACCTTTTAATTGAATATGCACGTAATAA
CATTCTAACTCTTTGGTTTTCTTAAGAAGTTGCCTTAAAGCTATTTCTATAGTAGTAGAGGTTATTTTTTTGTACC
TTTTTGTTTTGAGCTCCAACAGAGGTTCAAAAATTAACAGTTCTTTAATATTTGTTAGACATAATAGAATGTTTG
TCGAAGTAAAGAGA ACTTTCAATACAAGTGATTTTGGTTTGGTTAAAAACAT

Reverse complement

ATGTTTTTAACCAAACCAAATCACTTGTATTGAAAGTTCTCTTTACTTCGACAAACATTCTATTATGTCTAACAAA
TATTAAGGAAGTGAATTTTTTGAACCTCTGTTGGAGCTCAAAAACAAAAGGTACAAAAAATAACCTCTACTA
CTATAGAAATAGCTTTAAGGCAACTTCTTAAGAAAACCAAAGAGTTAGAATGTTATTACGTGCATATTCAATTA
GGTTTCAGCAAAAATAAAAACAAATCCTAAAAATTTTCAACAAGGTAAAGTAAAAATTCTTTCTATTTGTGATCA
AACCTCTTTACCCATAATGGTTGTAAAGCAGCAAAAATAAAAAGAATTTAA

Translation=MFLTKPKSLVLKVLFTSTNILLCLTNIKGTVIFWTSVGAQKQKGTKKITSTTIEIALRQLLKKTK
ELECYYVHIQLKGFSGKSKKQILKIFKQKVKILSICDQTSPLPHNGCKARKIKRI

>NADH dehydrogenase subunit 3_nad3_Gelidium pristoides_ Complement
12455..12820_366 bp

TTACTCTCATTCTAAAGCACCTTTACACCATTATATATAAAGCCTATTGTCAAAAACAACCAAGAAAACAACCATAG
TTCAAAGCCAAAAGGGGGTAAATCATTCAAACAAGTGACCAAGGAAAGAGAAAACCTTATTTCTAAATCAAATC
AAAAATAAAATAGCTACTAAATAAAATCTAATATCAAAGTAGTTCTCGCATCTTCAAAGGGATTAAAACCATTC
GTAAGCGCTAACCTTTTCTTGATCAGCTTTTTGCGGAGTAATTATGTAAGAAAGTATAAATATCGCTGAAGATAACA
AGAATGATAGCACATAAAAATTAAGATAACTGAATATTTCACTAAAGATTAACCTTCAT

Reverse complement

ATGAAGTTAATCTTTAGTGAATATTCAGTTATCTTAATTTTTATTGTGCTATCATTCTTGTTATCTTCAGCGATATT
TATACTTTCTTACATAATTACTCCGCAAAAAGCTGATCAAGAAAAGGTTAGCGCTTACGAATGTGGTTTTAATCCCT

TTGAAGATGCGAGAACTACTTTTTGATATTAGATTTTTATTTAGTAGCTATTTTTATTTTTGATTTTTGATTTAGAAATA
AGTTTTCTCTTTCTTTGGTCACTTGTGGTGAATGATTTACCCCTTTTGGCTTTTGAAGTATGGTTGTTTTCTTGGT
TGTTTTGACAATAGGCTTTATATATGAATGGTGTAAAGGTGCTTTAGAATGAGAGTAA

Translation=MKLIFSEYSVILIFIVLSFLLSSAIFILSYIITPQKADQEKVSAYECGFNPFEDARTTFDIRFYLVAILFLIFDLEISFLFPWSLVLNLDLPPFGFWMVFLVVLITIGFIYEWCKGALEWE

>NADH dehydrogenase subunit 1_nad1_Gelidium pristoides_

Complement_12832..13815_984 bp

TTAAAATAGTCAATTGAAACTTAGTAAGGTTCCAGCAACAAAGAATACTCAAGCTAATGCTAAAGGTAAGAGAATTT
TTCAACCAACTCTCATTAAATTTGGTTCATATCGATATCTTGGGAAGGCTGAGCGTACTCAAATAAACCCAAAAAGTAGT
AAAGTAGTTTTTAACCCAAATCAAATTTGGAGGTAATCAATAAAATAATACCAAATTAATAGGGGGTAACCACCC
ACCTAAGAATAGAATAGTTGTTAAGCTACACATTAGTATCATATTTGCATACTCACCCAAAAAGAATAGAGCAAAAC
CCATAGCAGAATACTCAACATTGTAACCTGCCACTAATTCTGCTTCTGCTTCTGGTAAATCAAACGGGGCTCTATTA
GTCTCTGCTAAAATAGAGATGTAAAACATTATTAATAATCGGAAATAGGGGTATCCCATATCAAATAGACTGTTGAGC
TAACACAATTTTCAGTTAAATTCAAAGAACCTGAACACAATAAGACGTTAATTAATAATCAGACCGATAGAACTTCAT
AGGAAACCATTTGAGCAGCTGATCGTAGTGCGCCAAAAAAGCGTATTTTGAATTACTTGACCAACCTGAAATAATA
ATTCCATACACACCTAAAGAAGACATCGCTAGAATATACAGAACCCCAACATTTATATCAGAATACACTAGCCCTTC
ACCTAGGGGTAAGACACATCAAGAAATCAATGCTAATAAGAATGTAAGGATTGGGGCTAATATAAAAATGCTAGTAT
TAGCGCTGGAAGGTAACACTGTTTCTTTTACAAATAGCTTAAGACCATCTGCTAATGGTTGTAATAACCCAAATAAT
CCTACAACATTCGGCCCTTTCTTCTCTGCATGGCTGCCATAACTTTTTCTTTTCTGCTAAAGTCATATAAGCAACCGC
AATAAGTAATGGCAAATTAAGAAAGAGTTTTTAATAAAGTTGTAAGTATAAAAATCAT

Reverse complement

ATGATTTTTTATACTTACAACCTTTATTAAAAACCTCTTTCTTTAATTTTTGCCATTACTTATTGCGGTTGCTTATATGAC
TTTAGCTGAAAGAAAAGTTATGGCAGCCATGCAGAGAAGGAAAGGGCCGAATGTTGTAGGATTATTTGGGTATTAC
AACCATTAGCAGATGGTCTTAAGCTATTTGTAAAAGAAACAGTTTTACCTTCCAGCGCTAATACTAGCATTTTTTATA
TTAGCCCCAATCCTTACATTCTTATTAGCATTGATTTCTTGATGTGTCTTACCCTAGGTGAAGGGCTAGTGTATTC
TGATAAAATGTTGGGGTCTGTATATTCTAGCAGTGTCTTCTTTAGTGTGTATGGAATTTATTTTTCAGTTGGT
CAAGTAATTCAAAATACGCTTTTTTTGGGCGCACTACGATCAGCTGCTCAAATGGTTTTCCCTATGAAGTTCTATCGGT
CTGATTTTTAATTAACGCTTATTGTGTTTCAAGTTCTTTGAATTTAACTGAAATTGTGTTAGCTCAACAGTCTATTTG
ATATGGGATACCCCTATTTCCGATTTTAATAATGTTTTACATCTCTATTTTTAGCAGAGACTAATAGAGCCCCGTTTG
ATTTACCAGAAGCAGAAGCAGAATTAGTGGCAGGTTACAATGTTGAGTATTCTGCTATGGGTTTTGCTCTATTCTTT
TTGGGTGAGTATGCAAATATGATACTAATGTGTAGCTTAACAACCTATTCTATTCTTAGGTGGGTGGTTACCCCTAT
TAATTTGGTATTATTTTATTGATTACCTCCAACAATTTGATTTGGGTAAAAACTACTTTACTACTTTTTGGGTTTA
TTTGAGTACGCTCAGCCTTCCCAAGATATCGATATGACCAATTAATGAGAGTTGGTTGAAAAATTCTCTTACCTTTA
GCATTAGCTTGAGTATTCTTTGTTGCTGGAACCTTACTAAGTTTTCAATTGACTATTTTTAA

Translation=MIFILTTLLKTLNLLPLLIIVAYMTLAERKVMAMQRRKGNVVLGFLQLADGLKLFVKET
VLPSSANTSIFILAPILTFLLALISWCVLPLGEGLVYSDINVGVLYILAMSSLGVYGIISGWSSNSKYAFLGALRS
AAQMVSYEVSIGLILINVLCSGSLNLTEIVLAQQSIWYGIPLFPILIMFYISILAETNRAPFDLPEAEAEELVAGYN
VEYSAMGFALFFLGEYANMILMCSLTILFLGGWLPINLVLVFLYWLPPTIWFGLKTTLLLLFGFIWVRSAPPRYRYDQ
LMRVGWKILLPLALAWVFFVAGTLLSFNWL

>NADH dehydrogenase subunit 2_nad2_Gelidium pristoides_13835..15319_1485 bp

TTATAAAATTGCCATAAAAGCTGAGATAGAAGTTAGCAATTTAGTACTAAGCATAAAACCAATCAACGTTATCGCAG
AAAGTCCTATAGTTAACGAAATTAACCTTATCTGTTTTAAACATAAAATGGGCAATAGATATTTCTTATCAAAGTAAACA
GATTTTTATTAATCGAATATAGTAAAAGCATGCAACACAACCTATAATAATGGCATTACCGCTATCCCAATACTATT
ATTTTGGATAGTTATTAATAATACAAACAACCTTTGAAAAGAAGCCAGCCAATGGGGGGATACCTGCCATGAAAAATA
GAAAGACTAATATAGTAAACGCTAACATAGGATTAGTTACAGAAAGCATTACTAAGTTATTTAAGTAACGAATTTGA
TAAGTTTTTCGGGTAACATAATGCTTTAAGCTCATAATAAAACCAAATGTTCCCTATCATCGTTATTAATAATACTAA
AAGATAAAATAGGGTACTAAAAATACCTAGAGTTGTTCCCTGCCATTAAACCAAGCAACAAAAACCTATGTGATTTA
TTGAGCTATAAGCCATAAAACGCTTCCATTTTTGTTTGGGTAAGCGCACCTAAGCTTCCCTATTATTAATGATAATAAA
ACACAAGGTAAAATAAATTAGACTTCACCAGTCTATGTAATCGTGAAGCAAAAAATAGAAGTCGAAATAGCACTGA
AAAGGCAGTTAGCTTAGGAATAGTTGAAAAAAGGCTGTACTTGTAAATCATTGATCCTTCATATACATCAGGAGCTC
ACATATGAAATGGGCTTGCAAGTAAATTTGAATAATAAAGCTACAATTATAAAAATTGAACTAATCCGTACACCAATC
GAAATAAATGAACCATCAATAAAAAATCCAGTAATAATTTTTGAAAATCATCAAATTACTTACCCCGGTCAAGCC

ATAAATTAGTGAAGATCCAAACAATAAAAAGTGCAGAAGCAAATGCACCTAAAATAAAAATACTTTAAACCTGCCTCTG
TAGAAAACCTCTGAAGTACGCTTAAAACCTCGCTAACACATAAAAATATTTAAAGCTTGTAACCTCAAGGGTTAAATATATG
CTTAATAAATCATACGCTTTTACTATTAGAAGCATAGAACTACAGCTAACAAAATTTAAAATTCAAAACCTCGAAGGA
TGCAACTTTTTGTTGAATAGAATATAAAAAAAGCAAAAATGATCAACTAATTATCACTGACAAAATAATTAATTGAG
ATCCGTGACTAAAATTATCTAAAATTA AAAAGTTATTCCAACCTAACCAATTTATAAATTCCTTGAGAAAGGGCTAAG
TAGAACTAAATAAGAGAATTTGAATTGACAACCTTACTTAAATTTGTGCTTAAAATCGGAAAACCATTTCTAGCAAC
TGAACTAAAGAACACCCCGTAAATTAATAAACGGTTATATTTAACAAAATGTAAGTTTCTATTAGTATTGAGTATA
TATCATATAACGTGTAATTCAT

Reverse complement

ATGAATTACACGTTATATGATATATACTCAATACTAATAGAACTTACATTTTGTAAATATAACCGTTTTTATTAAT
TTACGGGGTGTCTTTAGTTTCAGTTGCTAGAAATGGTTTTCCGATTTTAAAGCACAATTTAAGTAAGTTGTCAATTC
AAATTCCTTATTTAGTTTCTACTTAGCCCTTTCTCAAGAATTTATAAATTTGGTTAGTTGGAATAACTTTTTAATT
TTAGATAATTTTAGTCACGGATCTCAATTAATTAATTTTGTGCGTGATAATTAGTTGATCATTTTTGCTTTTTTTATA
TTCTATTCAACAAAAGTTGCATCCTTCGAGTTTTGAATTTTAAATTTTGTAGCTGTAGTTTCTATGCTTCTAATAG
TAAAAGCGTATGATTTATTAAGCATATATTTAACCTTGAGTTACAAGCTTAAATATTTTATGTGTTAGCGAGTTTT
AAGCGTACTTCAGAGTTTTCTACAGAGGCAGGTTTAAAGTATTTTATTTTAGGTGCATTTGCTTCTGCACTTTTTATT
GTTTGGATCTTCACTAATTTATGGCTTGACCGGGTAAGTAATTTGGATGATTTTTCAAATTTATTTACTGGATTTT
TTATTGATGGTTCATTTATTTTCGATTGGTGTACGGATTAGTTCAATTTTTATAATTGTAGCTTTATTTTCAAAT
ACTGCAAGCCCATTTTCATATGTGAGCTCCTGATGTATATGAAGGATCAATGATTACAAGTACAGCCTTTTTTTCAAC
TATTCCTAAGCTAACTGCCTTTTCAGTGCTATTTTCGACTTCTATTTTTTGTCTTTTACGATTACATAGACTGGTGAA
GTCTAATTAATTTTACCTTGTGTTTTATTATCATTAAATAATAGGAAGCTTAGGTGCGCTTACCCAAACAAAATGGAAG
CGTTTTATGGCTTATAGCTCAATAAATCACATAGGTTTTTTGTTGCTTGGTTAATGGCAGGAACAACCTCTAGGTAT
TTTTAGTACCCTATTTTATCTTTTAGTATATTTAATAACGATGATAGGAACATTTGGTTTTATTATGAGCTTAAAGC
ATTATAGTTACCCGAAAACCTTATCAAATTCGTTACTTAAATAACTTAGTAATGCTTTCTGTAACCTAATCCTATGTTA
GCGTTTACTATATTAGTCTTTCTATTTTCCATGGCAGGTATCCCCCATTTGGCTGGCTTCTTTTCAAAGTTGTTTGT
ATTATTAATAACTATCCAAAATAATAGTATTGGGATAGCCGTAATGCCATTATTATGAGTTGTGTTGCATGCTTTT
ACTATATTCGATTAATAAAAATCTGTTTACTTTGATAAGAATATCTATTGCCCATTTATGTTTTAAACAGATAAGTTA
ATTTTCGTTAACTATAGGACTTTCTGCGATAACGTTGATTGGTTTTATGCTTAGATCTAGAATTGCTAACTTCTATCTC
AGCTTTTATGGCAATTTTATAA

Translation=MNYTLYDIYSILIETYILLNITVLLIYGVFFSSVARNGFPILSTNLSKLSIQILLFSFYLAALSQE
FINLVSWNFLILDNFHSGSQLIILSVIISWSFLLFLYSIQQKVASFEFWILILLAVVSMLLIVKAYDLLSIYLTLE
LQALIFYVLASFKRTSEFSTEAGLKYFILGAFASALLLFGSSLIYGLTGVSNLDDFSKLFTEGFFIDGSFISIGVRIS
SIFIIVALLFKITASPFHMWAPDVYEGSMITSTAFFSTIPKLTAFSVLFRLLFFAFHDYIDWWSLIILPCVLLSLII
GSLGALTQTKWRFMAYSSINHIGFLLLGLMAGTTLGFISTLFYLLVYLITMIGTFGFIMSLKHYSYPKTYQIRYLN
NLVMSVTPMLAFTILVFLFSMAGIPPLAGFFSKLFLVLLITIQNNSIGIAVNAIIMSCVACFYIIRLIKSVYFDKN
IYCPFMFKTDKLISLTIGLSAITLIGLCLDLELLTSISAFMAIL

>Succinate: cytochrome c oxidoreductase subunit 4_sdh4_Gelidium
pristoides_15333..15575_243 bp

TTACATAAACATTTCTACCACACAACGAATTAGTTCAATATAACAAATTTTTATCGCCGATAAAGATACTATATATA
CTTTCTCAGTATGAACGTAATCTTCCACTATCGTTTTTAAACCTAAATTTATATGTAACAAAACAAAACCTTATATTT
AATATTAAGAATTCAATATCAATTACTAAAGCTCCTATAATTACTAAAGCTCCTCAACGCAAACTAAACCAGTTGAA
TTTTTGTAACAT

Reverse Complement

ATGTTACAAAATTCAACTGGTTTAGTTTGCCTTGAGGAGCTTTAGTAATTATAGGAGCTTTAGTAATTGATATTGA
ATTCTTAATATTAATAATAAGTTTTTGTGTTTACATATAAATTTAGGTTTAAAACGATAGTGAAGATTACGTTT
ATACTGAGAAAGTATATATAGTATCTTTATCGGCGATAAAAATTTGTTATATTGAACTAATTCGTTGTGTGGTAGAA
ATGTTTATGTAA

Translation=MLQKFNWFLRWGALVIIGALVIDIEFLILNISFCLLHINLGLKTIVEDYVHTEKVYIVLSLSAIK
ICYIELIRCVVEMFM

> NADH dehydrogenase subunit 4_nad4_Gelidium pristoides_
Complement_15576..17051_1476 bp
TTATATTTTTAGCTGCTCTATTAATACTGATTGTAGAAAAATGAAGCTCTTCTAATAAAGAGTTTGGATGTAAACCTA
CTCATAAAATTAGTAATACAAGAGGGAGTAAAATTCAGAATTCCTCTACGTGAAATATCCTGAAAAAATGAGAAGTAC
TTTGTTTTTAATATTCCAAATATAATTCTGTAAATAATCATATAGAATAGGCCGCTCCAAAAATCATAACCTAACT
CGCTAGAAGTGTACGAGTATACTTGGACTGAAATACTCCGAATAAAACCAAAATTCACCTATAAAATTACCTGTTC
CAGGAAATCCTAGATTAGCAAAAAGAGAAAAATAAAAAAATATTCAAAAATAGGCATCACTTGTATTAACCATTA
TAGTATTTAATAATTTCGTGTCTTATGACGATCATAAAGAATTCCAATGCACAAAAAAGTGCCTAGAAACCAACC
ATGGCTAAGCATAAGCATTAACTACCCTCAATTCCTTGAAACGTAAGTGAATAAATTCCAATTGTTACAAAACCCA
TGTGTGAAACAGATGAGTAAGCTATAATTTTTTCAAATCAACTTGCCTTAACGTAGTTAAAGAAGCATAAAATACA
GCCACTAAGCTTAATGAAAATATAAAAGGGGTAAGAAAAATGACGCTGAGGGAAATAAGGGTAAAGAAAAACGCAA
GAAACCGAACCCCTCCTATTTTCAGTAATACACCAGTAAAATAACTGAGCTGCAGTTGGTGCTTCAGCATGAGCTT
CCGGCAATCAAATATGAAATGGCATCATAGGAATTTTAACTGCAAAGCTCGCAAAGAATGATAACCAAAGAATTAAT
TGCCTAACTCCGAAAATTGAGTTTGTACATAATAATTGAATATCGGTTGTCCCTGTTTGAATAAATGCAAATTAT
GCCTAAAAGCATTAGTAAAGATCCAATTAAGTATACAAAAAAGTTGATACGCTGCTCGTATTTTTCTTAAGCGGG
ACCCTCAAATTCCTATAATTAAGAACATTGGAATTAGCAGCTCTCAAAAAAATATAGAAAAATAGTAGATCTAAA
ACACTAAAGACTTGAATAAGGCAAAATTCCAATAATAAAAAACAAATTAATACTCCCTTACTAGAATGTGAATAGA
ACTCCAATAATTAATGCAGACTATTGTAAGAAAAGTTGTTAAAAGAATAAAAAATAGCGATATACCGTCAACTC
CCACTGTGTAATAAATATTCACCTTTTAACCAATTAGCTGTACAGACAAATTGAAACAAAGAAGTATTTGAATCA
AACCTATTCAAATAAAAAGGGAGCATATAAAGTTAAACAAGAAATATAAAGCGCAATTAGCCGACATAAATATGT
ATTTGTGGAAGAAATAAAACACAAGATTACAACACCTAAAAGGGTGTCAAGGAAGTAAGTAGTAAAGGGTTAATA
ATTCCATAGTCAT

Reverse complement

ATGACTATGGAATTATTTAACCCCTTACTACTTACTTCCTTGACACCCTTTTTAGGTGTTGTAATCTTGTGTTTTAT
TTCTTCCACAAATACATATTTATGTGGCTAATTGCGCTTTATATTTCTTGTTTAACTTTTATATGCTCCCTTTTTA
TTTGAATAGGGTTTGATTCAAATACTTCTTTGTTTCAATTTGTCTGTACAGCTAATTGGTTAAAAGGTTGGAATATT
TATTACACAGTGGGAGTTGACGGTATATCGCTATTTTTTATTCTTTTAAACAACCTTTCTTACAATAGTCTGCATTTT
AATTAGTTGGAGTTCTATTACATTTCTAGTAAGGGAGTATTTAATTTGTTTTTATTATTGGAATTTTGCCTTATTC
AAGCTTTTAGTGTTTTAGACTACTATTTTTCTATATTTTTTTGAGAGCGTGCTAATTTCCAATGTTCTTAATTATA
GGAATTTGAGGGTCCCGCTTAAGAAAAATACGAGCAGCGTATCAACTTTTTTTGTATACTTTAATTGGATCTTTACT
AATGCTTTTAGGCATAATTTGCATTTATTTTCAAACAGGGACAACCGATATTCAATTATTATGACAACTCAATTTT
CGGAGTTTAGGCAATTAATTTCTTTGGTTATCATTCTTTGCGAGCTTTGCGAGTTAAAATTCCTATGATGCCATTTT
ATTTGATTGCCGGAAGCTCATGCTGAAGCACCAACTGCAGGCTCAGTTATTTTAGCTGGTGTATTACTGAAAATAGG
AGGGTTCGGTTTCTTGCCTTTTTCTTTACCCTTATTTCCCTCAGCGTCAATTTTCTTTACCCCTTTTATATTTT
TAAGCTTAGTGGCTGTAATTTATGCTTCTTTAACTACGTTAAGGCAAGTTGATTTGAAAAAATTTATAGCTTACTCA
TCTGTTTACACATGGGTTTTGTAACAATTGGAATTTTTTCACTTACGTTTCAAGGAATTGAGGGTAGTTAATGCT
TATGCTTAGCCATGGTTTTGGTTTTCTAGTGCCTTTTTTTGTGCATTGGAATTTCTTTATGATCGTCATAAGACACGAA
TTATTAATACTATAATGGTTTTAATAACAAGTATGCCTATTTTTGGAATATTTTTTTTTATTTTTCTTTTTGCTAAT
CTAGGATTTTCTGGAACAGGTAATTTTATAGGTGAATTTTTGGTTTTATTTCGGAGTATTTTCAAGTCAAGTATACTCGT
AACAGTTCTAGCGAGTTTAGGTATGATTTTTGGAGCGGCCTATTCTATATGATTATTTAACAGAATTATATTTGGAA
TATTAACAAAGTACTTCTCATTTTTTTTCAGGATATTTTACGTAAGAAATTTCTGAATTTTACTCCCTCTTGTATTA
CTAATTTTATGAGTAGGTTTACATCCAACTCTTTATTAGAAGAGCTTCATTTTTTCTACAATCAGTTTAATAGAGCA
GCTAAAAATATAA

Translation=MTMELFNPLLLTSLTPFLGVVILCFISSTNTYLCRLIALYISCLTFICSLFIWIGFDSNTSLFQF
VCTANWLKGNIIYTVGVDSLFFILLTTFLTIVCILISWSSIHIILVREYLICFLLEFLIQLVFSVLDLFFYIF
FESVLIPMFLIIGIWSRRLKIRAAAYQLFLYTLIGSLMLLGIICIIYFQTGTTDIQLLWQTQFSEFRQLILWLSFFA
SFAVKIPMPFHIWLPEAHAEAPTAGSVILAGVLLKIGGFGLRFSPLFPSASIFFTPFIFSLSLVAVIYASLTTL
RQVDLKKIIAYSSVSHMGFVTIGIFSLTFQGIIEGSLMLMLSHGLVSSALFLCIGILYDRHKTRIIKYNGLIQVMP
FGIFFLFFSFANLGFPGTGNFIFLVLFGVFQSSILVTVLASLGMIFGAAYSIIWLFNRRIIFGILKTKYFSFFQDIS
RREFWILLPLVLLILWVGLHPNSLLEELHFSTISLIEQLKI

>NADH dehydrogenase subunit 5_ *nad5_ Gelidium pristoides_ Complement*
17600..19573_1974

TTAAATAAAAAATATCAACTACTAAATATAAAGTGAGGGGTGGGGAAATAACTGTACCCTGCAAAAAAACCACAACC
CTAAAATTGTAAAAATAGATAGTGGGTTATTTGACCTGTTTGAATTTGTTTTAATGTTATTGCTCAAGATGGAATT
AATTTGGTTAAACCGTAAGGTCCCATTATTTCAATAAAGCCGCGATCTAAATTCTTAAAAGAAACAGAATATCCAAA
ATTTAATATAGGTTTAAATTAATAAATATTATAAATAGAATCCCAATATCACTTTTTTAAATAAAAAAGCTAAGA
AACTTACAATTTTTACCCCTTTATACCATTTATAGATTTTTATATTTAGGAACGACGCCAACGTTACTCCAAAAATA
CTTAAGAGAAAAGGAGCTCATTTAACGTTAATTTCTAAATGCTCTGCCTCAATTTGGTGCGAATGAACAGGTAAATT
AAAAATAGAAGTTTTTCAAAGTCTGTTCCCAAACCTATAAAAAGATCCCTGCAAAAATGCCCTGCAAAAATGCTAC
AAAAACCAAGATAAATAAAGGTATAAGTATAGTTAAATTTGATTACACGGGTAGAATTGATGTTAATTTCTACTCATA
TTACAGTTGTTGAGAAATGTAAGGTATAGTAAACGGAAGAATAAAAAGCTGTAAAGAAGACCCGACAACTCCCAT
TCAACAAGCTAAATTTACTCAGATCATAGATTTAGGTTAGAATTTGATAAATACTTGAGACACTTCTAAAATAAAATCTT
TTGAATAAAATCCGGTTAAGAAAAGGAAATCCTACCAAAGCTAATGAGCCTATCAACATCACTGAATAAGTTATTGGT
AAAAATTTCTTTAAAGAACCATTCTTCTCATGTCTTGCTCATCTGAAATTGCATGGATAACTGAACCTGCGCTCAA
GAAAAGTAAAGCTTTGAAAAAGCATGGTTCGTTAAATGGAATATACTTACATTATAACAAGATAAACCAAGCAA
ATATCATATAACCTAGTTGACTACAAGTCGAATACGCAATAACTCTTTTTAAGTCATTTTGAAAAACCAACCATA
GCGGCAAAAATTGCTGTCAATGATCCAAAAACGAGTAAATACGCAGTACGTAAACTGAGTACTCTATTAATGGAGA
AAATCGAATCAGTAAGAAAACCCCGCTGTTACCATTGTAGCAGCGTGAATTAGGGCCGAAACAGGAGTCGGTCCCT
CCATCGCATCTGGAAGCCAGGTATGTAGACCTAACTGCGCTGATTTTCCAACCTGCACCTACAAATAATAAAAAATCCA
ATTAGTGTAAACTATTAATTTGGAATTGTAAGAAAGAAAAATAGTAATCTTGAAAAACGGCGTTATTGAAAAAT
TGTTAAATACTCAACAGAACCAACGTATAAAAAACAACAATATGGCAAGGCTTAATCCAAAATCACCTACTCTAT
TTACAATCAGCGCTTTTATTGCTGATTGATTTGCCGCAAGCGCGTAAATCAGAAATTTATTAATAAATAAGAAGCC
AATCCAACCCCTTCTCACCCAAGAAACATCTGTAATAATATTATCTGCGGTAACATAAACTAGCATAAAAAAGGTAAA
AATCTCTAAATAAGCCATAAATCTAGGGCAGTGAGGATCGGTTTTCCATATAACTAATTGAGTACAGATGTACCAAC
TTGAAATTGAAGTAATTACAAGTAACATTGTTACCGTTACCGAATCAAATAGAAAACCTCATTTTACCGATAATATT
TCCGCATCAATCAAGATAATAAGTATATGTGACAAACGGTATTACAAAGCCCTATTTTATAAAAAAGCAAGCAGAGA
CAAGAAAAAAGAAGTACCACACAAATTTGTTGAGAAGGTTACTTGCTCCGTAACGACCTAGTCAACGACCTCCAAATC
CAGAGATTGCAGAGCCTACTAAAGGTAATAAAAATTTGTAATATACAT

Rev Complement

ATGTATATTACAATAAATTTTATTACCTTTAGTAGGCTCTGCAATCTCTGCAATTTGGAGGTCGTTGACTAGGTCGTTA
CGGAGCAAGTACCTTCTCAACAATTTGTGTGGTAACTTTCTTTTTCTTGTCTCTGCTTGTCTTTTTATGAAATAGGGC
TTTGTAATACCGTTTGTACATATACTTATTATCTTGAATTGATGCGGAAATATTATCGGTAAAATGAGGTTTTCTA
TTTGATTCCGGTAACGGTAACAATGTTACTTGTAAATACTTCAATTTCAAGTTTGGTACATCTGTACTCAATTAGTTA
TATGGAACCGATCCTCACTGCCCTAGATTTATGGCTTATTTAGAGATTTTACCTTTTTTATGCTAGTTTTAGTTA
CCGCAGATAAATTTTACAGATGTTTCTTGGGTGAGAAGGGGTTGGATTGGCTTCTTATTTATTAATAAATTTCTGA
TTTACGCGCTTGGCGGCAATCAATCAGCAATAAAAGCGCTGATTGTAATAGAGTAGGTGATTTTGGATTAAGCCT
TGCCATATTTGTTGTTTTTATACGTTTGGTCTGTTGAGTATTTAACAATTTTTTCAATAACGCCGTTTTTCCAAG
ATTACTATTTTTCTTTCTTACAATTTCAAATTAATAGTTTAACTAATTTGGATTTTTATTATTTGTAGGTGCAGTT
GGAAAATCAGCGCAGTTAGGTCTACATACCTGGCTTCCAGATGCGATGGAAGGACCGACTCCTGTTTCCGCCCTAAT
TCACGCTGCTACAATGGTAACAGCGGGGGTTTTCTTACTGATTTCGATTTTTCTCCATTAATAGAGTACTCAGTTTACG
TACTGCGTATTTTACTCGTTTTTGGATCATTGACAGCAATTTTTGCCGCTATGGTTGGTGTTTTTCAAATGACTTA
AAAAGAGTTATTGCGTATTGCACTTGTAGTCAACTAGGTTATATGATATTTGCTTGTGGTTTTATCTTGTATAATGT
AAGTATATTCCATTTAACGAACCATGCTTTTTTCAAAGCTTTACTTTTTCTTGAGCGCAGGTTTCAGTTATCCATGCAA
TTTCAGATGAGCAAGACATGAGAAGAATGGGTTCTTTAAAGAAATTTTTACCAATAACTTATTCAGTGATGTTGATA
GGCTCATTAGCTTTGGTAGGATTTCCCTTTCTTAAACCGGATTTTATTCAAAGATTTTTATTGAGAAGTGTCTCAAGT
ATTATCAAATTTCAACCTAATCTATGATCTGAGTAATTTAGCTTGTGAAATGGGAAGTTTGTGCGGCTTTCTTTACAG
CTTTTTTATTCTTTCCGTTTACTATACTTACATTTCTCAACAACGTAAATATGAGTAGAATTAACATCAATTTACTACC
CGTGAATCAAATTTAACTATACTTATACCTTTATTTATCTTGGTTTTTTGTAGCATTTTTTGCAGGGCATTTTTTGCAG
GGATCTTTTTTATAGGTTTGGGAACAGACTTTTTGAAAACTTCTATTTTTAATTTACCTGTTTATTTCGACCAAATTTG
AGGCAGAGCATTTAGAAATTAACGTTAAATGAGCTCCTTTTTCTTAAAGTATTTTTGGAGTAACGTTGGCGTCGTTT
CTAAATATAAAAAATCTATAAATGGTATAAAGGGGTAAAAATTTGTAAGTTTCTTAGCTTTTTTATTAATAAAAAAGTG
ATATTGGGATTCTATTTATAATAAGTTTTTAATTAACCTATATTAATTTTGGATATTCTGTTTCTTTTAAGAATT
TAGATCGCGGCTTTATTGAAATAATGGGACCTTACGGTTTAAACCAATTAATTCATCTTGAGCAATAACATTAATA
CAAATTCAAACAGGTCAAATAACCCACTATCTATTTTTTACAATTTTAGGGTTGTGGTTTTTTTTTGCAGGGTACAGT
TATTTCCCAACCCCTCACTTTATATTTAGTATTGATATTTTTTATTTAA

Translation=MYITIIILLPLVGS AISGFGGRWLG RYGASTFSTICVSSFFLSLLAFYEIGLCNTVCHIYLLSWI
DAEILSVKWFGLFDSVTVTMLLVITSISSLVHLYSISYMETDPHCPRFMAYLEIFTFFMLVLVTADNILQMFLGWEG
VGLASYLLINFWFTRLAANQSAIKALIVNRVGD FGLSLAIFVVVYTFGSVEYLTIFSI T PFFQDY YFSFLQFQINSL
TLIGFLLFVGVGKSAQLGLHTWLPDAMEGPTPV SALIHAATMV TAGVFLLIRFSPLIEYSVYVLRILLVFGSLTAI
FAAMVGVFQNDLKRVIAYSTCSQLGYMIFACGLSCYNVSI FHLTNHAFFKALLFLSAGSVIHAISDEQDMRRMGSLK
KFLPITYSVMLIGSLALVGF PFLTGFYSKDFILEV S QVLSNSNLIYDLSNLACWMGSLSVFFTA FYSFRLLYLTFLN
NCNMSRININSTRESNLTILIP LFILVFCSIFAGHF CRDLFIGLGTDFWKTSIFNL PVHSHQIEAEHLEINVKWAPF
LLSIFGVTLASFLNIKIYKWKYKGVKIVSFLAFLLNKKWY WDSIYNKFLIKPILNFGYSVSFKNLDRGFIEIMGPYGL
TKLIPSWAITLKQIQTGQITHYLF FTILGLWFFLQGTVISPPLTLYLVLIFFI

>ATP synthase F0 subunit 8_ *atp8_Gelidium pristoides*_ Complement
_19590..19994_405 bp

TTAGTTTGGGTAAATTGGTTAACATTAATACAAATAGAATTTAAAACCTTGAAAATTACAAAATAGTATTGAATTTT
TAATTGCGAGGCTCACTAATTCATCAATAGCTTGCGTATCTAAATTTCTAGGATCAGAAATGTTTAGAACTGAGTTT
AAAGAAAAAATTTTTTTATGCTACCCAAATGGTTTTAAAGTATAGACTTTAACTTGGTTTACTCTCAATTGATCT
CTTTGTTATCTCTGAAGATTCTAGCGCATTAAATTTTAAATTATTTCTTTTCGAGATTTTAAAACCTTTCAAAAATTTTG
GAAGAAAAAGTGAATTGACACAATATAAAATGCAGTGAAAGTTGCCAATAGTCAAAAATTTGTGGGAATAAAATT
ATACGATCTAGGTGTGGCAT

Reverse Complement

ATGCCACACCTAGATCGTATAATTTTATTCCCACAAATTTTTTGGACTATTGGCAACTTTCACTGCATTTTATATTGT
GTCAATTCACCTTTTTCTTCCAAAATTTTTGAAAGTTTTAAAATCTCGAAAAGAAATAATTAATAATTAATGCGCTAG
AATCTTCAGAGATAACAAAGAGATCAATTGAGAGTCAAACCAAGTTAAAGTCTATACTTTTAAACCATTTGGGTAGC
ATAAAAAAAGTTTTTTCTTTAACTCAGTTCTAAACATTTCTGATCCTAGAAATTTAGATACGCAAGCTATTGATGA
ATTAGTGAGCCTCGCAATTAATAATCAATACTATTTTGTAAATTTTCAAGTTTTAAATTCTATTTGTATTAATGTTA
AACCAATTTACCCAAACTAA

Translation=MPHLDR IILFPQIFWLLATFTAFYIVSIHFFLPKFLKVLKSRKEIIKINALESSEITKRSIESQT
KLKSILLNHLGSIKKVFSLSNVLNISDPRNLD TQAIDELVSLAIKNSILFCNFQVLNSICINVKPIYPN

>ATP synthase F0 subunit 6_ *atp6_Gelidium pristoides*_ Complement
_19994..20755_762 bp

TTAATGTAAGTCTAAAACATCATTTAAATAATTACAGGTTAGTAGTGTA AAAACATAAGCTTGCAGAGCAGCGATTG
CTAATTCTAACCCAGTTAAAGCTACAAGGAGTCCCAAGGGGATAAGGTGGAAAATTGCTAGTAGGCCGCCAGCTGAC
AACATGGTTCATGCAAATCCAGCGATAATCTTCAATAAAGTATGGCCTGATGTCATGTTAGCAAATAGTCGGATTGA
GAGTGTA AACACTTTAACTGTATAGGAAAGTAATTCAATTGTAATAATTAAGGTACAATAATTAAGGTACTCCTC
GAGGTAAAAATAACGAAAAAATTTTTATACCATGGGTTTGATACCTATGATATTTACACCGATAAAGATTGATAAG
GCCAAGGTAAACGTGAATATGATATGACTTGTA ACTGTAAAGCTATAGGGGACCATTCCAATTAATTACAAAAGAG
TAAAAATAAATGCAAGGTAAATATAAAGGGAAATAAAGTTCCCCTTTAATTCCAAGATTTTCTTGAAGCATGCTTG
CTGTCACCTTCATAAAAATATTCTTTTACTAACTGTCAATTAGTTGGCATTAGAGTTCCCTTGTA AAAACTTAGACTC
ATTCAAAATGAAGCCAGAAATAGCGTTAAAATCATAAAAATAGACGAGTTAGTCAATGACAAATTGACACCTGCAAT
TGACACAGGTAAGATAGTAACAATCTCAA ACTGCTCTAATGGGCTTGGGATAAAAATATACTTGGACAT

Reverse Complement

ATGTCCAAGTATATTTTTATCCCAAGCCATTAGAGCAGTTTGGAGATTGTTACTATCTTACCTGTGTCAATTGCAGG
TGTC AATTTGTCAATTGACTA ACTCGTCTATTTTTATGATTTTAAACGCTATTTCTGGCTTCATTTTGAATGAGTCTAA
GTTTTTACAAGGGA ACTCTAATGCCAACTAATTGACAGTTAGTAAAAGAATATTTTTATGAAGTGACAGCAAGCATG
CTTCAAGAAAATCTTGGAAATTAAGGGGAACTTTATTTCCCTTTTATATTTACCTTGCATTTATTTTTACTCTTTTG
TAATTTAATTGGAATGGTCCCCTATAGCTTTACAGTTACAAGTCATATCATATTCACGTTTACCTTGGCCTTATCAA
TCTTTATCGGTGTAATATCATAGGTATCCAAACCCATGGTATAAAAATTTTTTTCGTTATTTTTACCTCGAGGAGTA
CCTTTAATTATTGTACCTTTAATTATTACAATTGAATTACTTTTCTTATAACAGTTAAAGTGTTTACACTCTCAATCCG
ACTATTTGCTAACATGACATCAGGCCATACTTTATTGAAGATTATCGCTGGATTGTCATGAACCATGTTGTCAGCTG
GCGGCCTACTAGCAATTTTCCACCTTATCCCCTTGGGACTCCTTGTAGCTTTAACTGGGTTAGAATTAGCAATCGCT
GCTCTGCAAGCTTATGTTTTTACACTACTAACCTGTAATTATTTAAATGATGTTTTTAGACTTACATTA

Translation=MSKYIFIPSPLEQFEIVTILPVSIAGVNLSLTNSSIFMILTFLFLASFWMSLSFYKGTLMPTNWQL
VKEYFYEV TASMLQENLG IKGELYFPFI FTLHLFLLFCNLIGMPYSFTVTSHIIFTFTLALSIFIGVNIIGIQTHG

IKFFSLFLPRGVPLIIVPLIITIELLSYTVKVFTLSIRLFANMTSGHTLLKIIAGFAWTMLSAGLLAIFHLIPLGL
LVALTGLELAI AALQAYVFTLLTCNYLNDVLDLH

>trna_Sup_(TCA) _*Gelidium pristoides*_ Complement_20772..20845_74 bp
GGGAGAATAGCTTAACTTGGTAGAGCTCTGGTTTTCAAACCAAGGGTTGAGGGTTCGAATCCTTCTTCTCCTG

Reverse complement
CAGGAGAAGAAGGATTTCGAACCCCTCAACCCTTGGTTTTGAAAACCAGAGCTCTACCAAGTTAAGCTATTCTCCC

>tRNA-Ala_trnA_(TGC) _*Gelidium pristoides*_ Complement_21305..21377_73 bp
GGGGGTATAGTTTAGATGGCAAACATATGTTTTGCATACATAAAAtCATCGGTTTGAATCCGATTACTTCCA

Reverse complement
TGGAAGTAATCGGATTTCGAACCGATGaTTTTATGTATGCAAACATATGTTTTGCCATCTAAACTATACCCCC

>tRNA-Asn_(GTT) _*Gelidium pristoides*_ 21471..21543_73 bp
TTCCTGGTAGCTAAAAGGTTAAAGCGAGTAGCTGTAACTACTAGATTGTAGGTTCAAATCCTACCCAGGGAG

>tRNA-Val_(trnV) _*Gelidium pristoides*_ 21549..21621_73 bp
GGGTAATTAACCTCAATAGGTAGAGTATTTTTATTTACTACTAAAAGGTTAATGGTTTGAATCCATTATTACTCA

>tRNA-Arg_(trnR) _*Gelidium pristoides*_ 21629..21702_74 bp
GTGTTTGTAACTTAACTGGATAAAGTATTAAATTACGAATTTAAAGATTGAAGGTTCAAATCTTTCCAAACACA

>tRNA-Lys_(TTT)_trnK_ *Gelidium pristoides*_ 21716..21789_74 bp
GAGTGTATAGCTAAAAGGTTAAAGCAGTAGACTTTTTAATTTATTGATTTTAGGTTTGAATCCTAATACTCA

>Preprotein translocase subunit *SecY_SecY_Gelidium pristoides*_21813..22598_786 bp
ATGCCTCCAGTTACTATTTTATTCGAACGAGTTTTTTTTATAGGGGTTATTATTACTATATCAGTTCAATATTTATAGT
AGTGATAGTAACAGCTCAATTTGATGCGGTAAATCTAATTAAGTAATACCTCTTATTCACCTTATACAAAAAATTTA
CCATAGTAAAAGTGACTGATTTAATAGAGTTGTTGTGATTTCTTATTTTTTTCAGTTTCCTTTTTTAATTACATGGCCG
TTTTTACTGTTTTCAATTAATGAGTTTTTTAGGGTTAGTTGGTATAAGTATCAATTATACTATGTCAAAATAGTTTG
TATATGTGTTTTTTTTTACAACCTTTTTCGTTTTGAATTTTAACTATTTTAGTTTTTTACCAAATATTCTGCAGTTAC
TAATTGAATGAGATTCAATTAATAGTTACAATAAACTTTACTTAACTAGATGCGCAATTAATCTTTTTAAAGTAC
ACAACCTTGAGTGGTAGAATTTTATTACTTAATTAATTTTATAATGTTAAATATAGTATCTTGTATTGTCATTTCCAA
ATTTTTTTGATTGCTTAATTTTTAATACTATTTAATAAAAAATTACCGTAAGTTAAATATATTTTTGTTTTACGTTTA
ATTTATGTTTTATTTTACCTCCGATATAGTCTTTCAATTTTTAATATTTGCTTTGTTATTTTTATCTTTTGAGCTT
CTTTACCTTTTTGTGTGTTTACGTTTAGTAAATCAAATAACAGAACAAATTTATGCCCACTTTAAAACAATTATTGA
AAAAAAAAGGTTTTAA

Translation=MPPVTIYSNEFFYRGYYYYISSIFIVVIVTAQFDVNLKVIPLIHLYKKFTIVKVTDLIELLWF
LIFSVSFLITWPFLLFQLNEFFRVSWYKYQLYYVKIVCICVFFTTFVWILNYFSFLPNILQLLIEWDSINSYNKTL
LNLDAQLNLLKYTTWVVEFYYLINFIMLNIVSCIVISKFFWLLNFKYYLIKNYRKLNIFCFTFNLCFILPPDIVQF
LIFALLFLSFELLYLFVCLRLVNQITEQIYAHFKTIIIEKKV

>ribosomal protein S12_rps12_ *Gelidium pristoides*_22558..22923_366 bp
ATGCCCACTTTAAAACAATTATTGAAAAAAGGTTTAAAGCTAAAAGTCAAAAAACGCTCAATTGCGTTAGCAAA
AAACCCCAAAAAAAGGGGTTTGTATTCGTGTGTACACAATAAATCCTAAAAACCTAACTCCGCTGAAAGAAAGG
TTGCTAAAGTTCAATTGACTTCTGGCAAGTCTGTAATAGGGTATATTCCGGGCGAAGGTCCTCTTTGCAGGAACAC
TCTTTGGTATTAGTTAGAGGGGGTAGAGTTAAAGATTTACCGGGAGTTTTTTATCATTTTATACGTGGTAAGTATGA
CTTAATACCGGTAAAAACCGAAAATCTGCTAGATCAAATATGGTAAAAAAGCTAG

Translation=MPTLKQLLKKRFRFLKVKKRSIALAKNPQKKGVCIRVYTIINPKKPNSEAKVAKVQLTSGKSVIG
YIPGEGHSLQEHSLLVLRGGRVKDLPGVIFYHFIRGKYDLIPVKNRKSARSKYGKKS

>tRNA-Glu_TTC_trnE_Gelidium pristoides 22934..23006_73 bp
GCTCCTATCGTTTAATGGTTTAAGGCAATACTTTTTTCGTAGTATAAATGTGAGTTCGAATCTCACTAGGAGTA

>tRNA-Met_(CAT)_trnM_Gelidium pristoides_23008..23079_72 bp
AGCGGGATAGAGTAATTGGTAACTCGTTAGGTTTCATGTTCTAAAACCTATAGGTTCAAATCCTGTTCCCGTAA

>small subunit ribosomal RNA_rrs_Gelidium pristoides_23333..24659_1327bp
AGAGTTTGATCCTAGCTCAGAATGAACGTTATTGATTAGCTTAACACATGCAAGTCAAGATTAATTGGCGTACGGGT
GAGTAAGGCGTAGAAAATTGATCTTAGAAAAATTACGAAATGTATGCAAAGATTAATTCATCTAAAAAAGTCTACGT
AAGATTAAGTCGTTGGTGTAAAGATCACCAAGCTAACAACTTTTAGCTGATCTGTGAAGATGGACAGCCACATTG
GAACTGAAAAACGGTCCAGACTTTTATAAAAGGCAGCAGTGAGGAATATTGGGCAATGAGCGAAAGCTTGATCCAGC
TAAATCAAGTATGTGATTAAGGTAGTTAACTGTAAAGCATATTTTAAAGTATAAATAATGATTACTTAAAAAAGTCC
TGGCTAATTTTCGTGCCAGCAGCCGCGGTAACGAGAAGGGCGAACGTTACTCGGATTAATTGGGCGTAAAGTATAT
GTAGGTTGAATATTAATTTTTAATTTAAATCTTAAATTTTCATGTTGAGGAAGATTAATAAACTAATATTCTTGAGTT
TAAATGAAGATTACAGAATTTTAAATGAAACGGTAAAATGTATTGATATTTAAAAGAATCTCAATAGCGAAAGCAGT
AATCTAAATTTAAACTGACACTGAGATATTAAGCATGGGTAGCAAAAGGGATTAGATACCCCTGTAGTCCATGCC
TGAACGATGAGTGTGATTTTATGATTGACGTTATGAAATTAAGTTAAGGCTAAAACACTCCGCCTGGGAACACGG
CCGCAAGGCTAAAACCTCAAAGGAATTGACGGGGACTTACACAAGCAGTGGAGCATGTGGTTTTAAATCGATAATACGC
GCAAAACCTTACCAATTTTTGTATAATTACAGGTGTTGCATGGCTGCCGTGAGTCCGTGCTGTGAAGCGTTTTGGTTT
ATTCCAATAAACGGACAAAACCCCTCATATATTTTAGAATATAAACCGCTGAAGATATATCAAAGGAAGGTGAGGATG
ACGTCAAGTCTTTATGACCCTAATAAATTGGGCTACTTTCATGTGCTACAATGATTTTTTTCAAAGTATAACGTT
GTAAATCAAATTTAACTTTTTAAACAAAATCAAGTGCAGGATTGTTTTCTGAACTTGAAAATATGAAGATGGAATTG
CTAGTAATCGTGAATAAGAACGTCACGGTGAATTTGTTCTTAAGTCTTGTACACACCCGCCGTCACGCTATGGGAAT
TTATTTTCATTGAAAAATTAATAGTTTAAATGGAAAAAGGACTGAAGTGAAGTCGTAACAAGGTAGCCGTAGGGGAA
CCTGCGGCTGGATTAATT

> NADH dehydrogenase subunit 4L_nad4L_Gelidium pristoides_24707..25012_306 bp
ATGTACGAACAAATTAATTATATTAATACATCAACTTTATTATTTTTGATAAGTATTACAGGTATATTTTTAAATCA
AAAAAATATTTTTAGTAATGTTGATGTCTTTAGAAATAATGTTTTTAGCTGTAAGTTTTAATTTTTATTTTTCTCAA
TCTTTTTAGATGACGTTGTAGGTCAAATTTCTCATTGTTAATTTTTGACTGTTGCGGCAGCAGAATCTTCAATAGGG
CTGGCAATTTTTAGTGGTGTATTATAGAATACGCAATGCAATAACAGTTGAATTAATGATTTTTAATGAGCGGTTAA

Translation=MYEQINYINTSTLLFLISITGIFLNQKNILVMLMSLEIMFLAVSFNFIFFSIFLDDVVGQIFSLILITVAAAE
SSIGLAILVVYYRIRNAITVELMILMSG

A.4 APPENDIX D: Plastid Annotation File

- Appendix D is supplied on the accompanying CD.



University of Fort Hare
Together in Excellence