GENOME SEQUENCING AND ANALYSIS OF *ACROPORA GEMMIFERA* AND *PLEURACTIS GRANULOSA* FROM SABAH COASTAL WATERS

DEXTER LEE JIUNN HERNG

BIOTECHNOLOGY RESEARCH INSTITUTE
UNIVERSITI MALAYSIA SABAH
2018

GENOME SEQUENCING AND ANALYSIS OF *ACROPORA GEMMIFERA* AND *PLEURACTIS GRANULOSA* FROM SABAH COASTAL WATERS

DEXTER LEE JIUNN HERNG

THESIS SUBMITTED IN FULFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE

BIOTECHNOLOGY RESEARCH INSTITUTE
UNIVERSITI MALAYSIA SABAH
2018

PUMS 99:1

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN TESIS

JUDUL : _____
_____
_____

IJAZAH : _____
_____

SAYA : _____ **SESI PENGAJIAN** : _____
(HURUF BESAR)

Mengaku membenarkan tesis *(LPSM/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:-

1. Tesis adalah hak milik Universiti Malaysia Sabah.
2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan (/)

☐ SULIT (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di AKTA RAHSIA RASMI 1972)

☐ TERHAD (Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

☐ TIDAK TERHAD

Disahkan oleh:

_____ _____
(TANDATANGAN PENULIS) (TANDATANGAN PUSTAKAWAN)

Alamat Tetap: _____
_____
_____
_____

_____
(NAMA PENYELIA)

TARIKH: _____ TARIKH: _____

Catatan:
*Potong yang tidak berkenaan.
*Jika tesis ini SULIT dan TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai SULIT dan TERHAD.
*Tesis dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah dan Sarjana Secara Penyelidikan atau disertai bagi pengajian secara kerja kursus dan Laporan Projek Sarjana Muda (LPSM).

DECLARATION

I hereby declare that the materials in this thesis are my own except for excerpts, equations, summaries, and reference, which have been duly acknowledged.

15 September 2017 .....

Dexter Lee Jiunn Herng
MZ1411023T

CERTIFICATION

NAME : **DEXTER LEE JIUNN HERNG**

MATRIC NO. : **MZ1411023T**

TITLE : **GENOME SEQUENCING AND ANALYSIS OF *ACROPORA GEMMIFERA* AND *PLEURACTIS GRANULOSA* FROM SABAH COASTAL WATERS**

DEGREE : **MASTERS OF SCIENCE (BIOTECHNOLOGY)**

VIVA DATE : **13 FEBRUARY 2018**

CERTIFIED BY;

SUPERVISOR SIGNATURE

Dr. Christopher Voo Lok Yung .....

ACKNOWLEDGEMENTS

I would like to express my most sincere appreciation and thanks to my supervisor, Dr. Christopher Voo, for his unending advice, suggestions and support throughout the duration of this study. I would also like to extend my thanks to the investigators and other parties involved in the project from Institut Penyelidikan Bioteknologi (Dr. Kenneth Francis Rodrigues and Dr. Vijay Kumar), Sabah State Government (Sabah Biodiversity Centre, SaBC), and Protein Technologies Limited (PTL) for their efforts in making this study, especially in terms of (but not limited to) paperwork (permit for sample collection), funding (Grant ID GL00110), as well as for their inputs and technical advice.

At this juncture, I would like to thank my parents for their unyielding financial and moral support, as well as the Ministry of Higher Education (MoHE) for their financial support in the form of the MyMasters (MyBrain 15) scholarship. Although the outcome of this study is the end goal (as with any other studies), reaching the said goal alone and without any form of support would be an almost impossible task. As such, I would like to thank Mr. Rolando Roberto, fellow colleagues (past and present), the institute's lab assistants (past and present), suppliers, and service providers for making it possible to achieve the outcomes of this study.

Dexter Lee Jiunn Herng

15 September 2017

ABSTRACT

Coral are marine invertebrates classified under the Anthozoa class of the Cnidaria phylum. In this study, *Acropora gemmifera* and *Pleuractis granulosa* from Sabah were morphologically identified and found to have green fluorescence (*A. gemmifera* and *P.* granulosa) and red fluorescence (*P. granulosa*) when exposed to blue and green wavelength LED lights respectively. The DNA for both *A. gemmifera* and *P. granulosa*, as well as the mRNA from *A. gemmifera* were extracted and sequenced using Illumina (Solexa) sequencers. The reads generated from the genome sequencing were quality filtered, binned and *de novo* assembled, resulting in assembled genome sizes of 686.45 Mbp (*A. gemmifera*) and 1.23 Gbp (*P. granulosa*) with mean coverages in excess of 80X. The assembled genome sizes may be overinflated due to the possible inclusion of contigs from other contaminating organisms, redundant contigs and mis-assemblies. Nucleotide variant calling and ploidy estimation suggest that both *A. gemmifera* and *P. granulosa* are diploid organisms. The mitochondrial genome assemblies for both *A. gemmifera* and *P. granulosa* resulted in two mitochondrial genomes with a size of approximately 18 kbp per genome (mean coverage in excess of 1,300X), with BLASTN alignment suggesting a high degree of similarity between both mitochondrial genomes with other mitochondrial genomes from the *Acropora* genus. Unsupervised rRNA predictions using the assembled genomes and their subsequent annotation suggests that the assembled genomes may contain contigs from non-coral organisms. A total of 131,730 mRNAs were predicted and assembled (reference-guided assembly) for *A. gemmifera*, while a total of 152,504 mRNAs were predicted for *P. granulosa* with the aid of mRNA sequencing data from *Lobactis scutaria*. These transcripts, as well as the pathways from the KEGG pathway analysis, will serve as a starting point for further investigations and possible commercialisation, though these transcripts and pathways would need to be experimentally validated beforehand.

ABSTRAK

PENJUJUKAN DAN ANALISA GENOM ACROPORA GEMMIFERA DAN PLEURACTIS GRANULOSA DARI PERAIRAN PANTAI SABAH

Batu karang merupakan invertebrat marin yang diklasifikasikan dibawah kelas Anthozoa di dalam filum Cnidaria. Di dalam kajian ini, Acropora gemmifera dan Pleuractis granulosa dari Sabah telah dikenalpasti secara morfologi dan didapati mempunyai kependarfluoran hijau (A. gemmifera dan P. granulosa) dan kependarfluoran merah (P. granulosa) apabila masing-masing didedahkan kepada jarak gelombang cahaya LED biru dan hijau. DNA dari A. gemmifera dan P. granulosa serta mRNA dari A. gemmifera telah diekstrak dan seterusnya dijujuk dengan menggunakan mesin-mesin penjujukan Illumina (Solexa). Hasil jujukan genom telah disaring berdasarkan kualiti jujukan, diasingkan mengikut kumpulan dan dipasang secara de novo, di mana saiz genom-genom yang dipasang adalah sebesar 686.45 Mbp (A. gemmifera) dan 1.23 Gbp (P. granulosa) dan kedua-dua genom tersebut mempunyai liputan min melebihi 80X. Terdapat kemungkinan inflasi dalam saiz kedua-dua genom yang telah dipasang dan ini boleh disebabkan oleh kehadiran kontig-kontig dari organisma lain yang terpasang, kontig-kontig lewah, serta kontig-kontig yang tersalah pasang. Keputusan pangillan varian nukleotida dan anggaran ploidi mencadangkan bahawa A. gemmifera dan P. granulosa merupakan organisma diploid. Pemasangan genom mitokondria untuk A. gemmifera dan P. granulosa telah menghasilkan dua genom mitokondria yang dianggar sebesar 18 kbp satu genom (liputan min melebihi 1,300X), dengan penjajaran menggunakan BLASTN mencadangkan persamaan yang tinggi dengan genom-genom mitokondria lain dari genus Acropora. Jangkaan rRNA tanpa pemantauan menggunakan genom yang dipasang dan anotasi rRNA tersebut mencadangkan bahawa genom-genom yang telah dipasang mengandungi contig dari organisma bukan batu karang. Sebanyak 131,790 mRNA telah dijangka dan dipasang (pemasangan berlandaskan rujukan) untuk A. gemmifera, sementara sebanyak 152,504 mRNA teleh dijangka untuk P. granulosa dengan pengunaan data penjujukan mRNA dari Lobactis scutaria, dan terdapat keberangkalian terlebih jangkaan untuk kedua-kedua jangkaan. Transkrip-transkrip ini, serta laluan-laluan dari analisa laluan KEGG, boleh dijadikan sebagai titik permulaan untuk siasatan lanjut dan kemungkinan pengkomersialan, namun transkrip-transkrip dan laluan-laluan tersebut harus disahkan secara eksperimen terlebih dahulu.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES

LIST OF ABBREVIATIONS

$CaCO_3$	Calcium carbonate
IUCN	International Union for Conservation of Nature
OIST	Okinawa Institute of Science and Technology
DNA	Deoxyribonucleic acid
ds DNA	Double stranded deoxyribonucleic acid
cDNA	Complimentary deoxyribonucleic acid
ds cDNA	Double stranded complimentary deoxyribonucleic acid
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
tRNA	Transfer ribonucleic acid
mRNA	Messenger ribonucleic acid
RIN	RNA integrity number
PE	Paired-end
SE	Single-end / Second edition (context dependent)
Mbp	Mega base pair
Gbp	Giga base pair
kbp	Kilo base pair
NGS	Next generation sequencing
KEGG	Kyoto Encyclopaedia of Genes and Genomes
SRA	Sequence Read Archive
NCBI	National Center for Biotechnology Information
ENA	European Nucleotide Archive

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION AND OBJECTIVES

1.1 Introduction

Corals are marine invertebrates, currently classified under the Anthozoa class in the Cnidaria phylum. Corals can be found residing in tropical and sub-tropical waters, with temperatures ranging from 18 °C to 30 °C (although they can survive at temperatures of up to 40 °C for a short period of time [US Department of Commerce, National Oceanic and Atmospheric Administration, n.d.]) and water salinity between 32 to 42 parts per thousand at shallow depths (such that adequate sunlight can reach the corals for the coral-associated symbionts to carry out photosynthesis) (National Oceanic and Atmospheric Administration, n.d.). There are, however, exceptions as corals (deep water corals) had been discovered at depths where sunlight is not available, and at temperatures well below 18 °C (Roberts *et. al.,* 2009).

Coral reefs are underwater structures (mostly $CaCO_3$) formed by corals which serves a habitat for both the corals (involved in the reef building) as well as to an estimated 25 % of other marine organisms (National Oceanic and Atmospheric Administration Office of Habitat Conservation, n.d.). Although most of the reef-building corals are hard corals, not all hard corals are reef-building corals. Only colonial hard corals, secreting sufficient amounts of $CaCO_3$, are able to form coral reefs. The growth rate of the coral reef (which is generally slow) is dependent on the amount of $CaCO_3$, which in turn, is dependent on the coral species as well as the amount of nutrients available for the coral to use for growth. Most of this nutrient is

obtained from the combination of sheer amounts of coral-associated zooxanthellae, optimal conditions for photosynthesis as well as efficient nutrient recycling (Barnes and Hughes, 1999, p. 141). Coral reefs are under threat, both directly and indirectly, as a result of human activities. These include, but are not limited to, climate change, destructive fishing practices (e.g. fish bombing), and pollution. The increase in the average sea temperature surrounding the coral reefs beyond 30 °C will cause the coral polyps to expel the zooxanthellae within them, thus depriving the corals polyps of a significant chunk of their nutrient source as well as discolouring the polyps, thus leading to the corals being bleached. Should this condition persist for an extended period of time, the coral polyps will be unable to reacquire their zooxanthellae and will ultimately die, with the affected parts of the coral reef becoming permanently bleached (Hovland, 2008, p. 10) as in the case of the bleaching of parts of the Great Barrier Reef in 2016. This, combined with destructive fishing practices and pollution, not only destroyed the habitat of marine organisms, it has also pushed some of the corals species to the "critically endangered" category of the IUCN Red List, with *Tubastraea floreana* and *Rhizopsammia wellingtoni* being the first two coral species to enter the list in 2007 (International Union for Conservation of Nature, 2009). Coral reefs had been a source of income, with the main focus on fishing and ecotourism. However, given the aquatic nature of corals, their bio-resource potential has remained relatively untapped compared to terrestrial organisms. Genome and transcriptome sequencing can be used to aid in the identification of potential genes (and their respective gene products) and pathways with potential commercial applications before more destructive forms (i.e. removing larger chunks of corals) of testing is carried out, in addition to preserving the genetic information of coral species (especially those which are threaten with extinction).

The first coral genome (*Acropora digitifera*) was sequenced by the Marine Genomics Unit of OIST in 2011 (Shinzato, *et. al.*, 2011). The DNA was obtained from the sperm originating from a single *A. digitifera* colony, and sequenced using both the Illumina Genome Analyser IIx and Roche 454 GS-FLX sequencers with a combined sequencing coverage of approximately 150 X. *De novo* assembly was done using a combination of one *de novo* assembler, two scaffolders and one gap filler, resulting in a *de novo* assembled genome with 4,765 scaffolds and roughly 420 Mbp.

The total RNA (from various stages of coral development) was also sequenced and assembled, and subsequent analysis was focused on the genes involved in the coral's response towards changes in the environment. Other coral genomes had been sequenced in the following years by various institutions, groups and individuals, mostly using second generation sequencers. Some of these genomes were made public, while others were announced but were not released to the public.

1.2 Objectives

The problem statements for this study are:

1. The isolation and maintenance of coral culture for the sole purpose of obtaining the coral genome and transcriptome would not be feasible given the time and financial constraints. The observed phenotype of the coral sample in a laboratory setting forms the basis of further targeted coral sample harvesting from the environment, though there is a possibility that the observed phenotype could be originating from the coral symbiotes.
2. The sequencing data (passing relevant quality control filters) obtained from the sequencing of environmental samples will be a mixture of reads contributed from corals as well as other non-coral organisms. This will increase the time and computing resources required for genome and transcriptome assembly (where applicable).
3. As the coral species were chosen due to their observed phenotype and novelty (i.e. the genome and transcriptome of the selected coral species were not reported as being sequenced), comprehensive reference-based binning may be difficult due to the limited number of coral genomes that had been sequenced and are publicly available in public databases available since 2011.
4. Any isolated protein products of interest may not be directly reflected in the predicted mRNA transcripts and transcriptome assembly.

The hypotheses for this study are as follow:

1. Non-coral reads are to be separated as much as possible so as to remove it from the assembly process.
2. The genomes size *de novo* assembled coral genomes are likely to be overinflated.
3. Some observed coral phenotypes may not be reflected in the gene prediction results and vice versa, possibly due to, but not limited to, misassemblies, post translational modification and the expressed gene does not belong to the genome of the target coral.

As such, the objectives of this study are as follow:

1. To sequence the genomes of selected coral species with potential commercial applications.
2. To classify sequenced reads (passing relevant quality control filters) using an existing genome database.
3. To *de novo* assemble the genomes (genomic and mitochondrial) of the selected coral species.
4. To annotate the *de novo* assembled draft genomes.

CHAPTER 2

LITERATURE REVIEW

2.1 Hard corals and the Scleractinia order

Corals are marine invertebrates classified under the Anthozoa class of the Cnidaria phylum. Depending on the number of axes of symmetry present in their polyps, corals can be further classified into either the Hexacorallia subclass (6 axes of symmetry) or Octacorallia subclass (8 axes of symmetry). Hard corals are generally classified under the Hexacorallia subclass, while soft corals are generally classified under the Octacorallia subclass. Corals which are able to secrete aragonite (a form of $CaCO_3$), with the exception of *Heliopora coerulea*, are classified under the Scleractina order.

The basic unit of any mature coral is a polyp (Dai and Horng, 2009, p. 4). The anatomy of these polyps is similar to that of sea anemones (Sheppard, Davy, and Pilling, 2009, p. 36), with the exception that these polyps are able to product $CaCO_3$, which are secreted as aragonite. The secreted aragonite is then deposited at the bottom polyp in such a way that it encircles the bottom portion of the polyp, thus forming a corallite structure in which the said polyp resides. This corallite structure will grow upwards and outwards as more aragonite is deposited, thus increasing the size of the overall coral structure with the coral polyps residing on the outermost layer of the said coral structure. As a member of the Cnidaria phylum, hard corals possess cnidocytes, which can be used for defensive and feeding purposes (Roberts *et. al.*, 2009, p. 68). These cnidocytes may be used offensively (as part of the

sweeper tentacles and sweeper polyps) against other corals when competing for space (Sheppard, Davy, and Pilling, 2009, p. 44).

As with other members of the Anthozoa class, corals do not have a medusa stage. Instead, the adult stage in the coral's life cycle take the form of a polyp. Reproduction of hard corals can either take place sexually or asexually. Asexual reproduction in hard corals mainly proceed via fragmentation, with subsequent dispersal of the said fragments by sea currents. This is particularly common in branching hard coral species (Sheppard, Davy, and Pilling, 2009, p. 42). In sexual reproduction, fertilisation of the ova can either take place externally or internally, depending on the coral species in question. For external fertilisation to occur, both sperm and ova would need to be synchronously released during mass spawning events (which may be influenced by, but not limited to, daylight length, lunar cycle stage, water temperatures, and sunset) (Sheppard, Davy, and Pilling, 2009, p. 41). Upon successful external fertilisation, the fertilised ovum develops into a planulae, which is the larvae stage of the coral's lifecycle. The planulae exhibit limited mobility, and is dependent on the sea currents for dispersal away from the parent corals. However, the planulae can exhibit certain degrees of chemotaxis and phototaxis (Sheppard, Davy, and Pilling, 2009, p. 40), as well as being able to select (to a certain degree) settlement sites (Baird and Morse, 2004). Upon settlement on a suitable substrate, the planuale matures into a polyp, after which colony formation will take place. In contrast to external fertilisation, internal fertilisation occurs with the release of sperm into the surrounding waters while the ovum is witheld in the polyp body. As such, the ovum is fertilised and developed internally into a planulae. These planula will be released after sufficient development (Sheppard, Davy, and Pilling, 2009, p. 42), though they may end up settling within a close proximity to their respective female parent coral. In both external and internal fertilisation, there is a possibility of hybrid larvae being formed.

Regardless of the method of reproduction, the now independent polyps (with respect to the parent corals) would need to multiply in terms of numbers so as to increase in size. This is achieved by either intratentacular budding or extratentacular budding. In intratentacular budding, the daughter polyps (still attached to the parent polyp) are formed within the perimeter encircled by the the polyp's ring of tentacles.