DE NOVO SEQUENCING AND ASSEMBLY OF THE PINEAPPLE GENOME AND COMPARATIVE TRANSCRIPTOMICS OF TWO DEVELOPMENTAL STAGES OF THE FRUIT

RAIMI BINTI MOHAMED REDWAN

THESIS SUBMITTED IN FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

BIOTECHNOLOGY RESEARCH INSTITUTE
UNIVERSITI MALAYSIA SABAH
2017

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS

JUDUL: ***DE NOVO* SEQUENCING AND ASSEMBLY OF THE PINEAPPLE GENOME AND COMPARATIVE TRANSCRIPTOMICS OF TWO DEVELOPMENTAL STAGES OF THE FRUIT**

IJAZAH: **DOKTOR FALSAFAH (BIOTEKNOLOGI)**

Saya, **RAIMI MOHAMED REDWAN**, Sesi **2012-2017**, mengaku membenarkan tesis Doktoral ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:

1. Tesis ini 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 dalam AKTA RAHSIA 1972)
☐	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
☑	TIDAK TERHAD	

Disahkan oleh,

NURULAIN BINTI ISMAIL
LIBRARIAN
UNIVERSITI MALAYSIA SABAH

Raimi

RAIMI MOHAMED REDWAN
PZ1211017T

(Tandatangan Pustakawan)

Tarikh: 07 September 2017

(Assoc. Prof. Dr. Vijay Kumar)
Penyelia

(Dr. Christopher Voo Lok Yung)
~~Penyelia~~ Bersama

DECLARATION

I hereby declare that the materials written in this thesis are original and the experimental data is the result of my own independent work. This thesis has not been submitted previously for a higher degree in any university.

In addition, I also declare that the material in this thesis is my own except for quotations, excerpts, equations, summaries, and references, which have been duly acknowledged.

24 August 2017

Raimi

Raimi Mohamed Redwan
PZ1211017T

CERTIFICATION

NAME : **RAIMI MOHAMED REDWAN**

MATRIC NO. : **PZ1211017T**

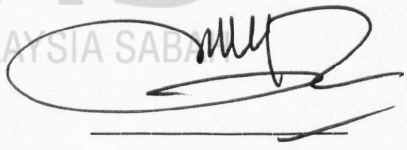
TITLE : ***DE NOVO* SEQUENCING AND ASSEMBLY OF THE PINEAPPLE GENOME AND COMPARATIVE TRANSCRIPTOMICS OF TWO DEVELOPMENTAL STAGES OF THE FRUIT**

DEGREE : **DOCTOR OF PHILOSOPHY (BIOTECHNOLOGY)**

VIVA DATE : **25 JULAI 2017**

CERTIFIED BY;

1. MAIN SUPERVISOR Signature



Assoc. Prof. Dr. Vijay Kumar

2. CO-SUPERVISOR

Dr. Christopher Voo Lok Yung

ACKNOWLEDGEMENTS

First and foremost, my gratitude is to Allah S.W.T. with His compassion and blessing, for giving me the strength and opportunity to complete this thesis. I wish to extend my deep appreciation to my main supervisor, Associate Professor Dr. Vijay Kumar for his guidance throughout my study. I also wish to thank my co-supervisor, Dr. Christopher Voo Lok Yung for his advice and assistance.

A special word of gratitude to my parents and siblings for their constant support and for the cheers especially during the challenging times at every stage of my research. In addition, I would like to thank my colleagues who have been with me throughout the ups and downs during this study. Special thanks go out to Chee Fong Tyng, Nur Afizah Nuin, Aswini Leela, Sally Venda Law, Elizabeth James and other member of BRI's family for support and company during my research. Special appreciation also goes to my husband, Zaidi Jakaria who endlessly pushed me through at the last moment of this journey.

Big thanks to the staff members of Novocraft Technology Sdn. Bhd. especially to Akzam Saidin who had guided and helped me tremendously during my research. This thesis would not come to accomplishment without their help and support. Also, I would like to express my gratitude to the staff members of PacBio Asia, TreeCode and ScienceVisions for their assistant in sequencing, especially to Caroline Chan, and Dana Chow. I would also like to extend my appreciation to the faculty member of BRI for their constructive comments during my study and to the technical staff of BRI for their help during my research. My appreciation goes to Ministry of Education and the Ministry of Science, Technology and Innovation, Malaysia, for the fund through Fundamental Research Grant Scheme (FRG0319-SG-2013) and Science Fund (SCF0087- BIO-2013), respectively. Lastly, I would like to extend my gratitude to Ministry of Higher Education and Universiti Malaysia Kelantan for their financial support during my study.

Raimi Mohamed Redwan

24 August 2017

ACKNOWLEDGEMENTS

First and foremost, my gratitude is to Allah S.W.T. with His compassion and blessing, for giving me the strength and opportunity to complete this thesis. I wish to extend my deep appreciation to my main supervisor, Associate Professor Dr. Vijay Kumar for his guidance throughout my study. I also wish to thank my co-supervisor, Dr. Christopher Voo Lok Yung for his advice and assistance.

A special word of gratitude to my parents and siblings for their constant support and for the cheers especially during the challenging times at every stage of my research. In addition, I would like to thank my colleagues who have been with me throughout the ups and downs during this study. Special thanks go out to Chee Fong Tyng, Nur Afizah Nuin, Aswini Leela, Sally Venda Law, Elizabeth James and other member of BRI's family for support and company during my research. Special appreciation also goes to my husband, Zaidi Jakaria who endlessly pushed me through at the last moment of this journey.

Big thanks to the staff members of Novocraft Technology Sdn. Bhd. especially to Akzam Saidin who had guided and helped me tremendously during my research. This thesis would not come to accomplishment without their help and support. Also, I would like to express my gratitude to the staff members of PacBio Asia, TreeCode and ScienceVisions for their assistant in sequencing, especially to Caroline Chan, and Dana Chow. I would also like to extend my appreciation to the faculty member of BRI for their constructive comments during my study and to the technical staff of BRI for their help during my research. My appreciation goes to Ministry of Education and the Ministry of Science, Technology and Innovation, Malaysia, for the fund through Fundamental Research Grant Scheme (FRG0319-SG-2013) and Science Fund (SCF0087- BIO-2013), respectively. Lastly, I would like to extend my gratitude to Ministry of Higher Education and Universiti Malaysia Kelantan for their financial support during my study.

Raimi Mohamed Redwan

24 August 2017

ABSTRACT

Pineapple (*Ananas comosus* var. *comosus*) is the third most important fruit globally after banana and citrus. Genetic information of the species will help expedite pineapple improvement program in producing elite cultivar and to facilitate understanding of its molecular mechanism. As such, this project aims to *de novo* sequence, assemble and annotate the genome of the commercially important MD-2 pineapple. The draft genome was then used as a reference to identify genetic variations in the Babagon pineapple (which is a domesticated local Sabah variety) and for comparative genomic study among the sequenced member of the sub-class Commelinidae. Furthermore, gene expression profiling of two developmental stages of the ripening fruit, specifically the mature green and mature yellow fruits, were performed using in-house available transcriptomic data. The genome was sequenced using two leading-edge sequencing technologies i.e. the highly accurate short Illumina reads and the ultra-long PacBio reads. A total of 110 Gbp reads were obtained which constitute 209X coverage of the pineapple genome. The final assembly of the MD-2 pineapple genome achieved an N50 scaffold of 153,084. Approximately, 27,017 protein-coding genes were predicted with 45.21% of the genome were identified as repetitive elements. Analyses of the Babagon variety showed one variant in every 108 bases with 86.6% of the variants composed of single-nucleotide variant (SNVs) and the remaining were insertion or deletion. The Ka/Ks analysis revealed that 48 genes in the Babagon pineapple differ significantly in comparison to MD-2. Among them were genes that are involved in the synthesis of terpene and plant defence system. Transcriptome analysis at the fruiting stage of the Babagon pineapple revealed several key genes related to the production of 4-hydroxy-2,5-dimethyl-3(2H)-furanone (HDMF), which is known to contribute to the flavour of pineapple. Furthermore, the genome-assisted-transcriptomic analysis suggests the important role ethylene plays in non-climacteric fruit, especially at the early stage of ripening and not throughout the ripening process as observed in climacteric fruit. The draft genome of the MD-2 pineapple has facilitated genomic analysis of pineapple as shown in the study and will allow further downstream applications that may have been hindered previously due to the lack of genomic information.

ABSTRAK

Penjujukan de novo dan penyusunan jujukan genom nenas dan perbandingan transkriptom di antara dua tahap kematangan buah

Nenas (Ananas comosus var. comosus) adalah buah ketiga yang paling penting secara global selepas pisang dan buah-buahan citrus. Informasi genetik nenas akan meningkatkan program penambahbaikkan nenas dalam penghasilan kultivar elit dan bagi membantu pemahaman mekanisma molekular. Oleh yang demikian, projek ini bertekad untuk membaca jujukan, menyusun dan menganotasi genom nenas komersial MD-2. Deraf genom ini kemudian digunakan sebagai rujukan untuk mengenal pasti variasi genetik dalam nenas Babagon (nenas tempatan Sabah yang didomestikasikan) dan bagi kajian perbandingan genom di kalangan ahli subkelas Commelinidae yang telah dijujuk. Tambahan lagi, pemprofilan ungkapan gen pada dua tahap perkembangan buah ranum, secara specifiknya pada buah hijau matang dan kuning matang dilakukan dengan menggunakan data transkriptomik sedia ada. Genom telah dijujuk menggunakan dua teknologi jujukan terkemuka i.e. jujukan pendek Illumina yang sangat tepat dan jujukan ultra-panjang PacBio. Sejumlah 110 Gbp jujukan telah diperolehi yang terdapat 209X liputan nenas genom. Deraf terakhir genom nenas MD-2 mencapai kerangka N50 sebanyak 153,084 bp. Lebih kurang, 27,017 gen pengekod protein yang dapat diramalkan bersama dengan 45.21% daripada genom dikenalpasti sebagai elemen berulang. Analisa variasi nenas daripada Babagon menunjukkan satu variasi bagi setiap 108 unit dengan 86.6% daripada variasi tersebut adalah terdiri daripada variasi nukleotida tunggal dan selebihnya adalah daripada penambahan dan penolokan. Analisa Ka/Ks menunjukkan 48 gen mempunyai perbezaan ketara di dalam perbandingan dengan nenas MD2 dan diantara gen tersebut adalah yang terlibat dengan sintesis terpene dan sistem pertahanan tumbuhan. Analisa transkriptom tisu buah tengah masak nenas Babagon menunjukkan beberapa gen kunci kepada penghasilan 4-hydroxy-2,5-dimethyl-3(2H)-furanone (HDMF), yang telah diketahui untuk menyumbang kepada perisa buah nenas. Seterusnya, analisa transkriptomik-dibantu-genom mencadangkan kepentigan peranan etilena di dalam buah tidak berklimaterik, terutamanya di tahap awal kemasakan dan bukan di sepanjang proses kemasakan seperti yang diperhatikan di dalam buah berklimaterik. Draf genom nenas MD-2 telah membantu analisis genom nenas seperti yang ditunjukkan dalam penyelidikan ini dan deraf ini akan membantu aplikasi hiliran yang sebelum ini terhalang disebabkan oleh kekangan maklumat genetik.

TABLE OF CONTENTS

LIST OF TABLES

tRNAs	Transfer RNA
TAGC	Taxon-annotated GC
CCS	Circular Consensus Sequencing
JGI	Joint Genome Institute
LSC	Long single copy
SSC	Short single copy
IR	Inverted region
SEA	Single enrichment analysis
SAR	Systemic acquired resistance
TGICL	TGI Clustering Tool
RSEM	RNASeq by Expectation Maximization
EM	Expectation maximization
TPM	Transcripts Per Million
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
TMM	Trimmed Median Mean
EC	enzyme codes
WEGO	Web Gene Ontology Annotation Plot
PCD	Programmed cell-death
ERS	Ethylene receptor gene
HDMF	4-Hydroxy-2,5-dimethyl-3(2H)-furanone
GWAS	Genome-wide association study
ETP	Economic Transformational Programme
WGS	Whole Genome Sequencing

LIST OF FIGURES

LIST OF ABBREVIATIONS

ncRNA	Non-coding RNA
SNV	Single nucleotide variant
INDEL	Insertion deletion
CAM	Crassulacean acid metabolism
RFLP	Restriction fragment length polymorphism
AFLP	Amplified fragment length polymorphism
SSR	Simple sequence repeat
PRI	Pineapple Research Institute
USDA	US Department of Agriculture
RAPD	Random amplified polymorphic DNA
ISSR	Inter-simple sequence repeat
PPO	Polyphenol oxidase
ACC	1-aminocyclopropane-1-carboxylic acid
GA	Gibberellin
AcAP1	Aspartic acid protease
EST	Expressed sequences tags
RADSeq	Restriction site associated DNA sequencing
BAC	bacterial artificial chromosome
TIGR	The Institute of Genomic Research
MTA	5'-deoxy-5'methylthioadenosine
PPT	Poly purine tract
PBS	Primer binding site
CEG	Core Eukaryotic Genes
CEGMA	Core Eukaryotic Genes Mapping Approach
MITE	Miniature Inverted Repeat Transposable
LTR	Long terminal retrotransposons
QI	Quality Index
AED	Annotation Edit Distance
rRNAs	Ribosomal RNAs

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Research Background

Over the years, demands for fresh pineapple for consumption has increased, especially after the introduction of new pineapple hybrids. Currently, the global pineapple market is dominated by the MD-2 variety, which, since its introduction, has become the leading pineapple variety globally (Vagneron *et al.*, 2009). No other newly introduced hybrids have been able to outperform the MD-2 variety in terms of taste and uniformity in size and ripeness. Even in Malaysia, the MD-2 is now the preferred choice for large-scale cultivation while the production of other major varieties such as the 'Maspine', 'Josapine', and 'Morris' have declined (Syahrin, 2011). However, near complete reliance to a single variety may be detrimental to the pineapple industry as the crop is likely to be susceptible to the same disease and stresses. Moreover, it is crucial that a large gene pool is maintained in any crop for better biodiversity security. As the most successful pineapple hybrid, it is intuitive that the genome of the MD-2 pineapple be decoded to gain better insights into its biology, which may be implicated in the development of new hybrids that can outperform this particular variety.

Genome information opens new gateways to better understand the biology of plants and subsequently, to better manipulate their phenotypic traits. The landscape of research in plant breeding has changed along with the evolution of sequencing technologies from the low-throughput Sanger to massive-throughput sequencing and to the current ultra-long sequencing technology (reviewed by Michael and VanBuren, 2015). Availability of genomes of commercially important crops have enabled genotype-phenotype association studies, discovery of new