

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

**RAIMI BINTI MOHAMED REDWAN**



**UMS**  
UNIVERSITI MALAYSIA SABAH

**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 Doktor 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

UNIVERSITI MALAYSIA SABAH

Disahkan oleh,

**NURULAIN BINTI ISMAIL**

LIBRARIAN

UNIVERSITI MALAYSIA SABAH

(Tandatangan Pustakawan)

*Raimi*

**RAIMI MOHAMED REDWAN**  
**PZ1211017T**

Tarikh: 07 September 2017

*Vijay Kumar*

(Assoc. Prof. Dr. Vijay Kumar)

Penyelia

*Christopher Voo Lok Yung*

(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



UMS  
UNIVERSITI MALAYSIA SABAH

## 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**

**1. MAIN SUPERVISOR**



Assoc. Prof. Dr. Vijay Kumar

**CERTIFIED BY;**

**UMS**  
UNIVERSITI MALAYSIA SABAH

Signature

A handwritten signature in black ink, appearing to read 'Vijay Kumar', written over a horizontal line.

**2. CO-SUPERVISOR**

Dr. Christopher Voo Lok Yung

A handwritten signature in black ink, appearing to read 'Christopher', written over a horizontal line.

## 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**

### ***Penjjukan 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 penambahbaikan nenas dalam penghasilan kultivar elit dan bagi membantu pemahaman mekanisma molekul. 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 spesifiknya 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 penolakan. 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 kepentingan 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 draf ini akan membantu aplikasi hiliran yang sebelum ini terhalang disebabkan oleh kekangan maklumat genetik.*



## TABLE OF CONTENTS

	Page
<b>TITLE</b>	i
<b>DECLARATION</b>	ii
<b>CERTIFICATION</b>	iii
<b>ACKNOWLEDGEMENT</b>	iv
<b>ABSTRACT</b>	v
<b><i>ABSTRAK</i></b>	vi
<b>TABLE OF CONTENTS</b>	vii
<b>LIST OF TABLES</b>	xii
<b>LIST OF FIGURES</b>	xv
<b>LIST OF ABBREVIATIONS</b>	xx
<b>LIST OF APPENDICES</b>	xxii
<b>CHAPTER 1: INTRODUCTION</b>	
1.1 Research Background	1
1.2 Problem Statement	3
1.3 Research Questions	6
1.4 Hypothesis	7
1.5 Research Objectives	8
1.6 Scope of the Study	8
1.7 Research Contribution	11
1.8 Structure of Thesis	12
<b>CHAPTER 2: LITERATURE REVIEW</b>	
2.1 Pineapple	13
2.1.1 Origin of Distribution	13
2.1.2 Taxonomic Classification	14
2.1.3 Morphology	15
2.1.4 Genetic Diversity	17

2.1.5	Domestication in Pineapple	19
2.1.6	Pineapple Cultivars	21
2.2	Pineapple in Biotechnology	23
2.2.1	Cytogenetics	23
2.2.2	Genetic Map of Pineapple	23
2.2.3	Gene, Transcript and Genomic Sequences	25
2.3	Whole Genome Sequencing	29
2.3.1	History of Plant Whole Genome Sequencing	29
2.3.2	Whole Genome Sequencing Using NGS	30
2.3.3	Improvement in Whole Genome Sequencing	34
2.3.4	Assembling Heterozygous Genomes	37
2.4	Fruit Ripening	39
2.4.1	Climacteric Pattern of Ripening in Fruits	40
2.4.2	Biosynthesis of Ethylene	41
2.4.3	Regulation of Ethylene Production	43
2.4.4	Ethylene Signal Transduction Process	45
2.4.5	Sugar Accumulation During Ripening	48
<b>CHAPTER 3: <i>DE NOVO</i> ASSEMBLY OF MD-2 PINEAPPLE GENOME AND ITS ANNOTATION</b>		
3.1	Introduction	50
3.2	Materials and Methods	52
3.2.1	Sample Materials	52
3.2.2	Illumina Library Preparation and Sequencing	52
3.2.3	PacBio Library Preparation and Sequencing	54
3.2.4	Genome Survey	56
3.2.5	Genome Assembly	56
3.2.6	Repeat Annotation	62
3.2.7	Gene Annotation	64
3.2.8	Chloroplast Genome	65
3.2.9	Mitochondrial Genome	67
3.3	Results and Discussion	67

5.3.2	Assessment of Multiple Transcriptome Assemblies and Its Annotation	173
5.3.3	Differential Expression of Transcriptome Analysis Using <i>De Novo</i> Transcriptome Assembly as Reference	181
5.3.4	Differential Expression of Transcriptome Analysis Using the Pineapple Genome as Reference	188
5.3.5	Comparison of De-Novo and Reference-Based Transcriptome Analysis	189
5.3.6	Ripening of the Pineapple Fruit	192
<b>CHAPTER 6: GENERAL DISCUSSION</b>		
6.1	Genome Assembly	208
6.1.1	Sequencing Technology	208
6.1.2	The MD-2 Genome	212
6.2	The Genome of Pineapple	213
6.2.1	Reduced Gene Number in Pineapple Genome	215
6.2.2	Pineapple as the Earliest Divergent Clade in Poales	216
6.3	The Babagon Pineapple Variety through the Lens of Genome Information	217
6.3.1	Genetic Variation of Babagon Pineapple Variety	217
6.3.2	High Density of Variants at the Non-Coding Region	218
6.3.3	Variants within the Coding Regions	219
6.4	The Process of Ripening in Non-Climacteric Pineapple Fruit	220
6.4.1	The Synthesis of Sucrose and Aromatic Compound in Ripening Pineapple Fruit	221
6.4.2	Lignification Process and Vascular System Development during Ripening Process of Pineapple	222
6.4.3	Ethylene Hormone in the Non-Climacteric Fruit	223
6.5	Limitation of the Study	224
<b>CHAPTER 7: CONCLUSION</b>		227

**REFERENCES**

231

**APPENDICES**

271



**UMS**  
UNIVERSITI MALAYSIA SABAH

## LIST OF TABLES

	Page
Table 2.1: The quality of plant draft genome assemblies using next generation sequencing.	32
Table 3.1: Quality and quantity of DNA as measured by NanoVue spectrophotometer and Qubit for pineapple leaf extracted using Carlier <i>et al.</i> (2004) protocol.	68
Table 3.2: Number of Illumina's sequencing reads from three different libraries, before and after trimming.	73
Table 3.3: Quality and quantity of DNA as measured by NanoDrop spectrophotometer and Qubit for pineapple leaf extracted using Dellaporta <i>et al.</i> (1983) protocol.	76
Table 3.4: The number of PacBio reads before and after novoCleaveLR processing to remove adapters and duplicated reads.	78
Table 3.5: Summary of assembly metrics across three different pineapple draft genomes produced using the respective assembly software.	84
Table 3.6: Summary of the assembly metrics of the Platanus's assembly before and after processing using PBJelly for gap-filling and scaffolding.	86
Table 3.7: Gap fill statistics for Platanus assembly after PBJelly.	87
Table 3.8: Number of pineapple transcripts mapped to pineapple draft genome assembled using Platanus and PBJelly.	89
Table 3.9: CEGMA assessment of the pineapple draft genome assembled using Platanus and PBJelly.	90
Table 3.10: Assembly metrics of contigs from pineapple draft genome assembled using DBG2OLC.	91
Table 3.11: Assembly metrics of assembly from pineapple draft genome assembled using DBG2OLC at contigs and scaffolds level.	92
Table 3.12: The number of short reads mapped to the DBG2OLC draft assembly.	93

Table 3.13:	Number of transcripts mapped to the draft genome assembled by DBG2OLC.	93
Table 3.14:	CEGMA assessment of the pineapple draft genome assembled using DBG2OLC.	94
Table 3.15:	Assembly metrics of contigs from pineapple draft genome assembled from error-corrected PacBio long reads using Celera.	95
Table 3.16:	The number of unmapped transcripts in the draft genome of Celera at contig level.	96
Table 3.17:	The number of short reads mapped to the contigs from Celera draft assembly.	97
Table 3.18:	The number of short reads mapped to the contigs from Celera draft assembly after redundancy removal.	98
Table 3.19:	Summary of assembly statistics of Celera assembly of error-corrected PacBio reads before and after improvements.	101
Table 3.20:	Number of pineapple transcripts mapped to pineapple draft genome assembled using Celera.	102
Table 3.21:	CEGMA assessment of the pineapple draft genome assembled using Celera.	102
Table 3.22:	Comparison of the assembly metrics of the available draft genomes of plant species.	106
Table 3.23:	COMPASS metrics for CELERA assembly using the F153 assembly as reference.	109
Table 3.24:	Summary of the repeated elements identified in the pineapple draft genome.	112
Table 3.25:	Number of predicted genes within the sequenced genomes of Commelinids.	116
Table 3.26:	List of genes in the chloroplast genome of pineapple, divided based on their process's functionality and groups of genes.	126
Table 4.1:	Characteristics of pineapple variety of MD-2 and the Babagon pineapple.	134

Table 4.2:	DNA quantification of genomic DNA extracted from pineapple leaves of Babagon pineapple.	141
Table 4.3:	Mapping quality of sequencing reads of pineapple from Babagon onto ACMD2 reference genome.	143
Table 4.4:	Number of variant identified from variant calling of Babagon pineapple.	144
Table 4.5:	Distribution of the number of effects categorized based on its impact caused on the protein coding genes.	149
Table 4.6:	Singular enrichment analysis by using Fisher Test to determine enrichment of the GO functional annotation of gene set that contain variants with 'high' impact effect.	150
Table 4.7:	Number of protein coding genes within the delimited values of Ka/Ks.	153
Table 5.1:	FastQC results for the sequencing data before and after trimming and filter were performed for both mature green and mature yellow fruits.	170
Table 5.2:	Assembly metric of transcriptome assembly constructed using Trinity, Oases-M and CLC Genomic Workbench.	174
Table 5.3:	Summary of similarity searched of known sequence to the transcripts assembled.	176
Table 5.4:	The top 10 transcripts with up-regulated in yellow mature pineapple fruit transcriptome as compared to green mature fruit.	186
Table 5.5:	The top 10 transcripts with up-regulated in yellow mature pineapple fruit transcriptome as compared to green mature fruit.	187

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



## LIST OF FIGURES

	Page
Figure 1.1:	10
Figure 2.1:	17
Figure 2.2:	47
Figure 3.1:	57
Figure 3.2:	61
Figure 3.3:	68
Figure 3.4:	70
Figure 3.5:	71
Figure 3.6:	74
Figure 3.7:	76
Figure 3.8:	77
Figure 3.9:	79
Figure 3.10:	81

Figure 3.11:	Coverage distribution of kmers counts coloured based on their characteristics and tabulated is the number of kmers within specific range of coverage based on its characteristic as derived from its respective coverage distribution.	82
Figure 3.12:	Comparison of the assembly metrics of three different short reads assembly using de-bruijn based method.	85
Figure 3.13:	Pie chart of gap improvement performed by PBJelly.	88
Figure 3.14:	Methods of scaffolding and polishing the Celera assembly with respective milestone of assembly improvement after each process.	100
Figure 3.15:	Plot to compare the assembly metrics between drafts produced using three different strategies.	104
Figure 3.16:	Plot representing the comparison of the number of mapped transcripts (above) and CEGMA (below) among the draft assembled by the three strategies.	105
Figure 3.17:	Distribution of coverage for mapping of the MD-2 scaffolds onto F153 pineapple draft genome. For all of the linkages, the mapping covered throughout the genome, but may not be visible in the plot as the mapping value was undersized by the high coverage value.	108
Figure 3.18:	The plot showed the distribution of coverage of Illumina short reads and the MD-2 scaffolds mapping to the F153 pineapple genome assembly.	110
Figure 3.19:	Summary of repeat elements identified within the draft genome of pineapple. The elements were sorted based on their abundance in the genome.	114
Figure 3.20:	Comparison of the features of genes predicted across different genomes from the members of Commelinids and <i>Arabidopsis thaliana</i> . (a) CDS Length, (b) Exon length, (c) Gene length and (d) exon number.	117
Figure 3.21:	Syntenic dotplot of nucmer alignment between contig of pineapple chloroplast genome and <i>Typha latifolia</i> chloroplast genome.	120
Figure 3.22:	Mapping back of the error-corrected long read (pink line), short reads (green line) and uncorrected long reads (blue	122

line) on the modified pineapple contig.

Figure 3.23:	The chloroplast genome of pineapple with its annotation.	124
Figure 3.24:	Syntenic dot-plot from the comparison of the first assembly's contigs and the second-round assembly's contigs.	129
Figure 3.25:	From top left (clockwise) are the syntenic dot-plot from the comparison of the pineapple mitochondrial genome with the mitochondrial genome of <i>Phoenix dactylifera</i> , <i>Zea mays</i> and <i>Oryza sativa</i> , respectively.	131
Figure 4.1:	Gel electrophoresis of genomic DNA extracted from leaves sample of Babagon pineapple using Dellaporta <i>et al.</i> (1983). M represents the 1 kb DNA ladder, 1 and 2 are the replicate samples of the Babagon total DNA extraction.	140
Figure 4.2:	Quality score profile of the Illumina short sequence reads of the Babagon pineapple. From leaf is the forward and reverse reads from the paired-end sequencing output.	142
Figure 4.3:	Distribution of number of variants across different scaffolds of Babagon pineapple in relative to ACMD2 reference genome.	146
Figure 4.4:	Distribution of location of variations based on the ACMD2 genome annotation.	147
Figure 4.5:	GO level 3 enrichment of the genes that had variants with 'high' impact effects in comparison with the overall functional annotation of the pineapple genome.	151
Figure 4.6:	Distribution of the Ka/Ks value of the protein coding genes from Babagon and MD-2 pineapple varieties.	152
Figure 4.7:	Bar chart of the distribution of the PFAM domains of the forty-eight genes that had Ka/Ks ratio of above 1.	154
Figure 4.8:	Orthology analysis of protein from <i>A. comosus</i> , <i>Oryza sativa</i> , <i>Brachypodium distachyon</i> , <i>Musa acuminata</i> , and <i>Elaeis guineensis</i> .	157
Figure 4.9:	Phylogenetic tree constructed using 409 single-copy-genes from pineapple, six species from sub-class commelinids, <i>A. thaliana</i> and <i>Am. trichopoda</i> .	159
Figure 4.10:	Timetree analysis using Maximum Likelihood method. The	161

timetree shown was generated using the RelTime method.

Figure 4.11:	Phylogenetic tree of the seven members of the sub-class Commelinids, <i>A. thaliana</i> and <i>Am. trichopoda</i> .	163
Figure 5.1:	Per base sequence content bias that were observed in all four reads.	172
Figure 5.2:	Number of alignment similarity between the reference assembly to the known protein database of (a) pineapple and (b) rice protein.	177
Figure 5.3:	Species distribution of top Blastx result of pineapple fruit transcriptome reference assembly to the plant NR database.	179
Figure 5.4:	Pie chart of the E-value distribution of BLAST hits of the <i>de novo</i> assembled transcripts to the plant protein database (i.e. nr) with e-value cutoff of 1e-6.	180
Figure 5.5:	Similarity percentage distribution of BLAST hits of the <i>de novo</i> assembled transcripts to the plant protein database (i.e. nr) with e-value cutoff of 1e-6.	180
Figure 5.6:	Boxplot of expression value of mature green and mature yellow transcriptome before (left) and after (right) normalization process.	182
Figure 5.7:	Heat map of the 1485 transcripts identified as differentially expressed between green mature and yellow mature pineapple fruit between 12 weeks and 16 weeks after flowering, respectively.	183
Figure 5.8:	GO annotations of the differentially expressed transcripts during pineapple fruit ripening.	184
Figure 5.9:	WEGO plot of the GO annotation of the differentially expressed genes between green and yellow mature fruits transcriptomic analysis.	188
Figure 5.10:	Snippet from the genome browser showing the RNASeq mapping of the green mature unripe fruit and yellow ripe fruit of pineapple on one of the gene involve in biosynthesis of ethylene.	191
Figure 5.11:	Heat-map showing the expression level of the quinone oxidoreductase-like protein found among the transcripts in reference transcriptome assembly.	195

- Figure 5.12: Expression profile of pineapple fruit development for phenylpropanoid pathway. 197
- Figure 5.13: Differentially expressed transcripts commonly present between differential expression studies of ripening fruits of strawberry and pineapple. 200
- Figure 5.14: Biosynthesis of ethylene and expression of its two rate-limiting enzymes, *ACC Synthase* (ACS) and *ACC Oxydase* (ACO). 204
- Figure 5.15: Regulation of ethylene production, from left is class I, auto-inhibition and class II auto-catalytic and the bottom is the heatmap represents the level of expression of ethylene receptors during repering of pineapple fruit. 206



UMS  
UNIVERSITI MALAYSIA SABAH

## LIST OF ABBREVIATIONS

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

## LIST OF APPENDICES

	Page
Appendix 1: Parameter for CELERA assembly.	271
Appendix 2: Perl script of pslscore.pl.	273
Appendix 3: Final annotation list in gff3 format.	275
Appendix 4: Resequencing of the chloroplast for validation.	276
Appendix 5: List of Babagon scaffolds with no variants.	277
Appendix 6: Number of gene's contraction, expansion and unchanged.	280
Appendix 7: List of differentially expressed transcripts from using <i>de novo</i> reference assembly as reference.	281
Appendix 8: List of differentially expressed transcripts from using draft genome as reference.	284
Appendix 9: List of transcripts that were in homology to invertase.	286
Appendix 10: List of publications.	288

# 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