

**ISOLATION, CHARACTERIZATION AND
MAPPING OF EXPRESSED SEQUENCE TAGS
(ESTs) FROM PINEAPPLE FRUIT cDNA
LIBRARY**

ONG WEN DEE



UIMS
PERPUSTAKAAN
UNIVERSITI MALAYSIA SABAH
UNIVERSITI MALAYSIA SABAH

**THESIS SUBMITTED IN FULFILLMENT FOR
THE DEGREE OF MASTER OF SCIENCE**

**BIOTECHNOLOGY RESEARCH INSTITUTE
UNIVERSITI MALAYSIA SABAH
2011**

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS@

JUDUL: ISOLATION, CHARACTERIZATION, MAPPING OF THE EXPRESSED SEQUENCE TAGS FROM PINEAPPLE CDNA LIBRARY

IJAZAH: DEGREE OF MASTER OF SCIENCE

SAYA ONG WEN DEE

SESI PENGAJIAN: 2011

Mengaku membenarkan tesis Sarjana ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:-

1. Tesis adalah hakmilik 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 institutsi pengajian tinggi.
4. Sila tandakan (/)

SULIT

(Mengandungi maklumat yang berdarjah keselamatan atau Kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

TERHAD

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

TIDAK TERHAD

(TANDATANGAN PENULIS)

PERPUSTAKAAN
UNIVERSITI MALAYSIA SABAH

Disahkan Oleh

(TANDATANGAN PERPUSTAKAAN)

Prof. Madya Dr. Vijay Kumar
Nama Penyelia

Tarikh: 28 Julai 2011

CATATAN:-*Potong yang tidak berkenaan.

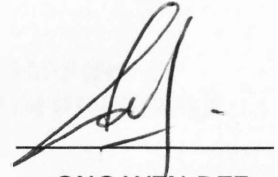
**Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai SULIT dan TERHAD.

@ Tesis ini dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah dan Sarjana secara penyelidikan atau sertai bagi pengajian secara kerja kursus dan Laporan Projek Sarjana Muda (LPSM).

DECLARATION

I hereby declare that this dissertation is the result of my own research except for quotations and citations which have been duly acknowledged.

28 July 2011



ONG WEN DEE

PB2007-8432



UMS
UNIVERSITI MALAYSIA SABAH

CERTIFICATION

NAME : **ONG WEN DEE**

MATRIC NO. : **PB2007-8432**

TITLE : **ISOLATION, CHARACTERIZATION AND MAPPING OF
EXPRESSED SEQUENCE TAGS (ESTs) FROM PINEAPPLE
FRUIT cDNA LIBRARY**

DEGREE : **MASTER DEGREE OF SCIENCE**

VIVA DATE : **4 JULY 2011**

DECLARED BY

1. SUPERVISOR

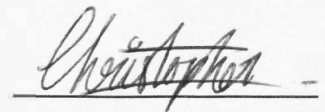
Associate Professor Dr. Vijay Kumar

Signature



2. CO-SUPERVISOR

Dr. Christopher Voo Luk Yung



ACKNOWLEDGEMENT

First of all I thank God for His love and encouragement. To my family, thank you for giving me full support, love and understanding all this while. I am also very thankful to those who have supported and encouraged me throughout the duration of my MSc project.

I sincerely thank my supportive supervisor Assoc. Prof. Dr. Vijay Kumar for his guidance and also for providing me an opportunity to learn more in this field. I truly thank him for not only spending time reading the thesis but for the all valuable comments throughout these years. Most importantly, I would thank him for the encouragements and advice he gave me throughout the duration of this project.

I thank Mr. Ahmad Kamal Bin Ghazali from Science Vision for his important contribution to the success of my work. I thank him for spending his precious time teaching and guiding me on the analysis of my NGS data during my attachment at the Malaysia Genome Institute (MGI). His opinion and suggestions have truly contributed to a better and correct analysis and writing of my results. I also like to thank all the staff members at MGI for allowing me to access their facility and for their comments and technical troubleshooting during the assembly of the NGS data.

Here I would also like to take this opportunity to acknowledge my gratefulness to all the lab assistants in BRI for assisting me in the laboratory work and for providing me ideas to solve my technical difficulties and for not hesitating to assist me in using some sophisticated instruments. Last but not least, I would like to express my greatest gratitude to Biotechnology Research Institute, UMS for providing me a great workplace and environment for the success of the this research.

Ong Wen Dee

28 July 2011

ABSTRACT

ISOLATION, CHARACTERIZATION AND MAPPING OF EXPRESSED SEQUENCE TAGS (ESTs) FROM PINEAPPLE FRUIT cDNA LIBRARY

Pineapple (*Ananas comosus* var. *comosus*), is an important tropical non-climacteric fruit with high commercial potential. Understanding the phenomena behind fruit ripening with a focus on improving fruit quality traits such as flavor, texture, appearance and sweetness may be possible through gene expression profiling of pineapple fruit transcriptome. As such, the objectives of this project are to, firstly, construct and sequenced mature green pineapple cDNA and *de novo* assembly of paired-end Solexa reads. Secondly, to characterize and functionally annotate the transcripts through similarity search and mapping against Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) database respectively. Finally, to develop a database of Expressed Sequence Tags containing Simple Sequence Repeats (EST-SSRs) using the newly obtained transcripts and/or through pineapple ESTs that are available in GenBank. The results show that both the unique transcripts (UT) assembled pineapple sequences and contigs from *de novo* assembly generated a total of 28,896 transcripts being generated with length ranges from 100 bp to 3.8 kb. A search for sequence similarity with NCBI's non-redundant database identified about 17,049 transcripts which were found to be associated with primary metabolisms, amino acid synthesis and processing, membrane and transport, cell division, cytoskeleton, cell wall and metabolism, RNA related gene expression, signal transduction, defense and stress related protein and also secondary metabolisms. Out of these transcripts, 71% returned GO terms with the distribution among the ontologies given as such: 35.8% in molecular function, 33.5% in cellular component and 30.7% in biological process. Annotation against the KEGG database pathways on the other hand, enabled the assignment of 542 enzyme commissions to 13,598 transcripts. The enzymes were further categorized into a total of 126 pathways with 122 pathways being involved in pineapple metabolism. The metabolic and cellular processes points out that there are tremendous changes in metabolic activities during pineapple fruit maturation as seen by the large numbers of the annotated transcripts. Data mining of the pineapple transcripts EST-SSRs showed that only 4% of the pineapple transcripts contained SSRs. Dinucleotide SSR (49.5%) was the most abundant followed by trinucleotide SSR (46.8%). The least abundant was tetranucleotide SSR (3.7%). Out of these, about 40% of the pineapple transcripts were found to have suitable flanking sites to enable the design of the upstream and downstream primers for future PCR amplification. This research cataloged the first pineapple fruit transcriptome. The transcripts will be subsequently useful to develop microarray chips for future gene expression studies among different plant tissue and development stages of the fruit. Further validation and/or relevant use of the EST-SSRs found will be useful in comparative mapping and genome mapping and gene tagging in pineapple.

ABSTRAK

Nenas, (Ananas comosus var. comosus) merupakan buah tropika yang mempunyai nilai komersial yang tinggi. Memahami fenomena disebalik pemasakan buah dengan tumpuan untuk memperbaharui nilai buah dari segi rasa, struktur, rupa dan manis buah boleh dicapai melalui analisa transkriptome ekspresi gene. Dengan itu, objektif kajian adalah pertamanya pembinaan perpustakaan jujukan saling melengkapi DNA dan pengelompokan pasangan hujung ke hujung jujukan Solexa. Keduanya, adalah menjelaskan transkript yang didapati melalui pencarian persamaan dan penentuan fungsi menggunakan pangkalan data ontologi serta 'Kyoto Encyclopedia of Genes and Genomes' (KEGG). Akhir sekali kajian ini akan membina satu pangkalan data transkript yang wujud dalam kawasan gen berkod dengan menggunakan transkript yang dihasilkan dan juga jujukan saling melengkapi DNA yang sedia ada dalam GenBank. Kesemua transcript unik (UT) dan contigs yang dihasilkan dapat dikelompokkan dalam lebih kurang 30, 000 transkript dengan panjang antara 100 bp ke 3.8 kb. Pencarian persamaan dengan pangkalan data "non-redundant" NCBI mengenalpasti sejumlah 17,049 transkript dengan penglibatan dalam metabolisme asas, penghasilan dan pemprosesan asid amino, dinding dan pengangkutan, pembahagian sel, metabolisme dan pembinaan struktur dinding, ekspresi gen berhubungkait dengan jujukan RNA, transduksi isyarat, protein berkait dengan pertahanan dan tekanan, dan juga metabolisme sekunder. Daripada jumlah ini, 71% mempunyai penanda ontology dengan 35.8% dalam kumpulan fungsi molekular, 33.5% dalam komponen sel dan 30.7% dalam proses biologi. Penentuan fungsi menggunakan pangkalan data KEGG mendapati sebanyak 13,598 transkript mempunyai fungsi yang sama dengan sejumlah 542 kod enzim yang mana boleh dikelompokkan kepada 126 laluan. Daripada jumlah laluan ini 122 didapati berhubung kait dengan metabolisme nenas. Penentuan fungsi transcript mendapati kebanyakan transcript terlibat dalam metabolik and proses sel dinding. Ini menunjukkan semasa pemasakan buah nenas, aktiviti metabolik giat berlaku. Kajian rangkaian jujukan berulang dalam transkript nenas pula menunjukkan sebanyak 4% daripadanya mempunyai rangkaian jujukan berulang. Dua-nukleotid paling banyak dijumpai dengan sebanyak 676 (49.5%) jujukan penanda terungkap mengandungi rangkaian jujukan berulang. Ini diikuti dengan tiga-nukleotid dan empat-nukleotid dengan masing-masing sebanyak 639 (46.8%) dan 51 (3.7%). Daripada jumlah ini, 40% daripadanya dikenalpasti mempunyai rusuk yang sesuai untuk pencorakan "primers" bahagian depan dan belakang bagi kegunaan amplifikasi PCR pada masa akan datang. Kajian ini menghasilkan transkriptome buah nenas yang pertama. Jujukan penanda terungkap ini berguna untuk penghasilan cip microarray bagi kajian ekspresi gen dalam pelbagai tisu dan peringkat pembentukan buah. Analisa yang lebih terperinci dan/atau penggunaan jujukan penanda terungkap mengandungi rangkaian jujukan berulang boleh diaplikasi dalam pemetaan komparatif dan genome serta penandaan gene dalam nenas.

LIST OF CONTENTS

| | Page |
|---|-------|
| TITLE | i |
| DECLARATION | ii |
| CERTIFICATION | iii |
| ACKNOWLEDGEMENT | iv |
| ABSTRACT | v |
| ABSTRAK | vi |
| LIST OF CONTENTS | vii |
| LIST OF TABLES | xii |
| LIST OF FIGURES | xiii |
| LIST OF ABBREVIATIONS | xvi |
| LIST OF SYMBOLS | xviii |
| LIST OF UNITS | xix |
| LIST OF EQUATIONS | xx |
| LIST OF APPENDICES | xxi |
| CHAPTER 1: INTRODUCTION | |
| 1.0 Introduction | 1 |
| 1.1 The Objectives of the Study | 3 |
| CHAPTER 2: LITERATURE REVIEW | |
| 2.1 Pineapple | 4 |
| 2.2 Pineapple Fruit Maturity and Ripening | 5 |
| 2.3 Uses of Pineapple | 7 |
| 2.3.1 Pineapple Fruit Processing | 7 |
| 2.3.2 Application in the Meat Industry | 7 |
| 2.3.3 Therapeutic Application | 8 |
| 2.3.4 By-products | 9 |
| 2.4 Problems in the Pineapple Industry | 9 |
| 2.5 Application of Biotechnology for Crop Improvement | 10 |
| 2.6 Transcriptomic Studies in Plant | 12 |
| 2.7 Sanger Sequencing | 13 |
| 2.8 Next Generation Sequencing (NGS) | 15 |

| | | |
|------|--|----|
| 2.9 | Expressed Sequence Tags | 18 |
| 2.10 | Application of Expressed Sequence Tags | 21 |
| | 2.10.1 Hybridization Experiments | 21 |
| | 2.10.2 Gene Discovery | 22 |
| | 2.10.3 Simple Sequence Repeats (SSRs) Marker Development | 24 |
| 2.11 | ESTs in Non-climacteric Fruits | 26 |

CHAPTER 3: MATERIALS AND METHODS

| | | |
|------|--|----|
| 3.1 | Overview of Methodology | 29 |
| 3.2 | Plant Materials and Total RNA Extraction | 29 |
| 3.3 | First Strand cDNA Synthesis | 30 |
| 3.4 | Confirmation of Successful of Reverse Transcription | 31 |
| 3.5 | Determination of cDNA Amplification Cycles | 31 |
| 3.6 | Normalization of cDNA Library | 33 |
| | 3.6.1 Hybridization | 33 |
| | 3.6.2 Duplex-specific Nuclease Treatment | 33 |
| 3.7 | Determination of Optimal Number of PCR Cycles for Normalized cDNA | 34 |
| 3.8 | Amplification of Normalized cDNA | 35 |
| 3.9 | Proteinase K and <i>Sfi</i> I Restriction Treatment | 35 |
| 3.10 | Size Fractionation | 36 |
| 3.11 | Ligation of cDNA to Vector and cDNA Library Generation | 37 |
| 3.12 | Screening of cDNA Library | 38 |
| | 3.12.1 Preparation of Host Bacteria | 38 |
| | 3.12.2 Mass Excision of λ TripIEx2 to pTripIEx2 | 38 |
| 3.13 | Colony PCR | 38 |
| 3.14 | Plasmid Isolation by Alkaline Lysis | 39 |
| 3.15 | DNA Sequencing | 40 |
| 3.16 | Bioinformatic Analysis | 40 |
| | 3.16.1 Sequence Characterization and Gene Ontology (GO) Annotation of Transcripts | 40 |
| | 3.16.2 Functional Classification by Kyoto Encyclopedia of Genes and Genomes (KEGG) | 41 |
| | 3.16.3 Identification of EST-SSR Motifs and Flanking Primers | 41 |

| | | |
|--------|--|----|
| 3.17 | Solexa Sequencing | 42 |
| 3.18 | <i>De novo</i> Assembly of Solexa Sequencing Reads | 42 |
| 3.18.1 | Running Velvet | 43 |
| 3.18.2 | Running Velvetg | 44 |
| 3.19 | Counting of Contigs Size | 44 |

CHAPTER 4: RESULTS

| | | |
|---------|---|----|
| 4.1 | Overview of Result Presentation | 47 |
| 4.2 | Total RNA Extraction | 47 |
| 4.3 | Construction of cDNA Library | 48 |
| 4.3.1 | First and Second Strand cDNA Synthesis | 48 |
| 4.3.2 | Determination of Optimal PCR Cycles for Non-normalized and Normalized cDNA | 48 |
| 4.3.3 | <i>Sfi</i> I Enzyme Treatment, Size Fractionation and Vector Ligation of Normalized cDNA | 51 |
| 4.3.4 | Titrating of cDNA Library, Amplification of Insert and Plasmid Extraction of Positives Clones | 52 |
| 4.3.5 | Single Pass Sequencing of Partial cDNA Clones | 54 |
| 4.3.6 | Characterization of ESTs from Mature Green Fruit cDNA Library | 56 |
| 4.4 | Solexa Sequencing | 61 |
| 4.4.1 | Solexa Reads and <i>De novo</i> Assembly using Velvet Software | 61 |
| 4.4.2 | Distribution of Velvet <i>K</i> -mer 47 Contigs Size and Coverage (Expression Level) | 64 |
| 4.4.3 | Characterization of Velvet <i>K</i> -mer 47 Contigs | 65 |
| 4.4.4 | High Coverage Contigs | 68 |
| 4.5 | Functional Annotation | 68 |
| 4.5.1 | Gene Ontology (GO) Annotation | 68 |
| 4.5.2 | Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Assessment | 73 |
| 4.5.3 | Gene Encoding Important Traits in Pineapple Fruit | 78 |
| 4.5.3.1 | Fruit Ripening | 78 |
| 4.5.3.2 | Fragrance Biosynthesis | 80 |
| | a) Terpene Biosynthesis | 80 |

| | | |
|---------|---|-----|
| b) | Ester Biosynthesis | 82 |
| 4.5.3.3 | Flavor Biosynthesis | 85 |
| a) | Organic Acid Synthesis | 85 |
| b) | Starch and Sucrose Metabolism | 85 |
| 4.5.3.4 | Texture/structural Biosynthesis –Lignin Biosynthesis | 89 |
| 4.5.3.5 | Health-Related Compound Biosynthesis | 89 |
| a) | Quinate Biosynthesis | 89 |
| b) | Riboflavin Metabolism | 91 |
| c) | Folate Biosynthesis | 92 |
| 4.6 | Detection of Simple Sequence Repeats (SSRs) in Pineapple ESTs | 93 |
| 4.6.1 | Pineapple EST-SSRs Generated from Sanger Sequencing | 93 |
| 4.6.1.1 | Occurrences of Different SSRs | 93 |
| 4.6.1.2 | Sizes of Pineapple SSRs | 95 |
| 4.6.1.3 | Distribution of EST-SSRs | 96 |
| 4.6.1.4 | Determination of Gene Identities of EST-SSRs | 99 |
| 4.6.1.5 | Identification of Flanking Sequences of Type I EST-SSRs | 99 |
| 4.6.2 | Pineapple EST-SSRs from Solexa Sequencing | 117 |
| 4.6.2.1 | Occurrences of Different SSRs in Pineapple Fruit Contigs | 117 |
| 4.6.2.2 | Sizes of Pineapple SSRs in Pineapple Fruit Contigs | 117 |
| 4.6.2.3 | Distribution of SSRs in Pineapple Fruit Contigs | 119 |
| 4.6.2.4 | Determination of Gene Identities of Contigs Containing SSRs | 119 |
| 4.6.2.5 | Identification of Flanking Sequences of Type I EST-SSRs Derived from Contigs | 129 |

CHAPTER 5: DISCUSSION

| | | |
|-----|---|-----|
| 5.1 | Isolation of Total RNA from Pineapple Fruit Tissue | 147 |
| 5.2 | Factors Affecting the Construction of Mature Green Pineapple Fruit Library | 148 |
| 5.3 | Single Pass Sequencing of Green Mature Pineapple Clones | 152 |
| 5.4 | Assembly of Pineapple Transcripts | 154 |
| 5.5 | Characterization through Sequence Similarity Searches | 157 |

| | | |
|---------------------------|---|-----|
| 5.6 | Highly Expressed Transcripts | 158 |
| 5.7 | Functional Annotation of Pineapple Transcripts | 163 |
| 5.8 | Gene Encoding Important Traits in Pineapple Fruit | 165 |
| 5.9 | Pineapple Type I EST-SSRs | 170 |
| 5.9.1 | Distribution and Size | 170 |
| 5.9.2 | Distribution of EST-SSR Motifs | 172 |
| 5.9.3 | Identification of Gene Identities and Flanking Region of EST-SSRs | 174 |
| | | |
| CHAPTER 6: SUMMARY | | |
| 6.1 | Summary and Conclusion | 177 |
| 6.2 | Future Prospects | 179 |
| | | |
| REFERENCES | | 181 |
| APPENDICES | | 199 |



UMS
UNIVERSITI MALAYSIA SABAH

LIST OF TABLES

| | | Page |
|------------|---|------|
| Table 2.1 | Recent transgenic and mutant tomato and grape genotypes with relation to the nutrient or shelf-life change. | 12 |
| Table 2.2 | Abundant transcripts in pineapple fruit identified by Moyle <i>et al.</i> (2005b). | 28 |
| Table 3.1 | PCR cycling parameters for cDNA amplification. | 32 |
| Table 3.2 | Dilution for DSN treatment. | 34 |
| Table 3.3 | The various types of repeat motif screened for in the pineapple transcripts. | 42 |
| Table 4.1 | Blast results of UTs against non-redundant NCBI database. | 60 |
| Table 4.2 | Assembly and characterization of <i>k</i> -mer 47 contigs generated from <i>de novo</i> assembly. | 66 |
| Table 4.3 | Contigs with significant identity and coverage over 500. | 69 |
| Table 4.4 | Summary results of mapping and annotating of pineapple transcripts against Gene Ontology database. | 71 |
| Table 4.5 | List of metabolism pathways in pineapple fruit transcripts. | 74 |
| Table 4.6 | Summary of functional assessment of EST containing SSRs in pineapple ESTs. | 99 |
| Table 4.7 | EST-SSRs with significant identities. | 100 |
| Table 4.8 | EST-SSRs containing flanking primers. | 108 |
| Table 4.9 | Summary of functional assessment of EST containing SSRs in pineapple fruit contigs. | 123 |
| Table 4.10 | EST-SSRs in pineapple fruit contigs with significant identities. | 123 |
| Table 4.11 | Pineapple fruit contigs EST-SSRs with flanking primers. | 130 |

LIST OF FIGURES

| | Page | |
|-------------|---|----|
| Figure 2.1 | The changes in the physiochemical properties of pineapple fruit from flowering to senescence. | 6 |
| Figure 2.2 | Schematic workflow of paired-end tags methodology both using cloning based and cloning free procedures. | 19 |
| Figure 3.1 | Perl scripts tracking of contigs size. | 46 |
| Figure 4.1 | Total RNA extracted from pineapple fruit tissue using high salt concentration. | 49 |
| Figure 4.2 | Amplification of metallothionein transcripts. | 49 |
| Figure 4.3 | Amplification of double strand cDNA. | 50 |
| Figure 4.4 | Amplification of cDNA with different number of PCR cycles. | 50 |
| Figure 4.5 | Amplification of normalized cDNA treated with different DSN dilution. | 51 |
| Figure 4.6 | Fractionation of normalized cDNA. | 52 |
| Figure 4.7 | Colony PCR amplification of insert. | 53 |
| Figure 4.8 | Plasmid extraction based on alkaline lysis of different positive clones. | 53 |
| Figure 4.9 | Example of pineapple fruit cDNA sequence. | 54 |
| Figure 4.10 | Chromatogram of single pass sequencing of mature green pineapple cDNA clones. | 55 |
| Figure 4.11 | Contig generated from two pineapple sequences. | 57 |
| Figure 4.12 | Length distribution of UTs assembled from green mature green pineapple sequences. | 58 |
| Figure 4.13 | E-value distribution of mature green pineapple UTs with significant identities. | 58 |
| Figure 4.14 | Blast hit species distribution of UTs of mature green pineapple sequences. | 59 |
| Figure 4.15 | Example of forward and reverse Solexa reads. | 62 |

| | | |
|-------------|--|-----|
| Figure 4.16 | Results of <i>de novo</i> assembly of Solexa reads using different <i>k</i> -mers. | 63 |
| Figure 4.17 | Length distribution of pineapple fruit contigs generated from <i>de novo</i> assembly. | 64 |
| Figure 4.18 | Coverage distribution of contigs from <i>de novo</i> assembly. | 65 |
| Figure 4.19 | E-value distribution of contigs with significant identities. | 67 |
| Figure 4.20 | Blast hit species distribution of pineapple contigs. | 67 |
| Figure 4.21 | Gene Ontology annotation of pineapple fruit transcripts (level 2). | 72 |
| Figure 4.22 | List of biosynthesis involved in the pineapple fruit metabolism. | 77 |
| Figure 4.23 | The ethylene synthesis and signal transduction. | 79 |
| Figure 4.24 | The terpene biosynthesis. | 81 |
| Figure 4.25 | The straight chain ester biosynthesis from fatty acids. | 83 |
| Figure 4.26 | The branched chain ester biosynthesis. | 84 |
| Figure 4.27 | The citrate acid cycle. | 86 |
| Figure 4.28 | The sucrose metabolism. | 87 |
| Figure 4.29 | The starch and sucrose metabolism. | 88 |
| Figure 4.30 | The lignin biosynthesis. | 90 |
| Figure 4.31 | The quinate biosynthesis. | 91 |
| Figure 4.32 | The riboflavin (B2) biosynthesis. | 92 |
| Figure 4.33 | The folate biosynthesis. | 94 |
| Figure 4.34 | Distribution of Type I SSRs of single pass sequencing of pineapple ESTs. | 95 |
| Figure 4.35 | The distribution of single pass sequencing SSRs in categories of repeats unit. | 97 |
| Figure 4.36 | Distribution of single pass sequencing dinucleotide SSRs. | 97 |
| Figure 4.37 | Distribution of single pass sequencing trinucleotide SSRs. | 98 |
| Figure 4.38 | Distribution of single pass sequencing tetranucleotide SSRs. | 98 |
| Figure 4.39 | Distribution of Type I simple sequence repeats in pineapple contigs. | 118 |

| | | |
|-------------|--|-----|
| Figure 4.40 | The distribution of contigs SSRs in categories of repeats unit. | 118 |
| Figure 4.41 | Distribution of dinucleotide SSRs in pineapple fruit contigs. | 120 |
| Figure 4.42 | Distribution of trinucleotide SSRs in pineapple fruit contigs. | 121 |
| Figure 4.43 | Distribution of tetranucleotide SSRs in pineapple fruit contigs. | 122 |
| Figure 5.1 | Total RNA extracted from pineapple fruit using alkaline phenol:chloroform. | 149 |
| Figure 5.2 | Colony PCR amplification of small insert cDNA clones. | 152 |
| Figure 5.3 | Homopolymer slippage. | 153 |



UMS
UNIVERSITI MALAYSIA SABAH

LIST OF ABBREVIATIONS

| | |
|-------|--|
| ACC | 1-aminocyclopropane-1-carboxylate |
| AFLP | amplified fragment length polymorphism |
| CAD | cinnamyl alcohol dehydrogenase |
| CCD | charge coupled device |
| C-OMT | caffeate O-methyltransferase |
| CTAB | cetyltrimethylammonium bromide |
| DEPC | diethylpyrocarbonate |
| DSN | duplex-specific nuclease |
| DTT | dithiothreitol |
| cDNA | complementary DNA |
| dNTP | deoxynucleotide tri phosphate |
| ddNTP | dideoxynucleotide tri phosphate |
| ddATP | dideoxyadenine tri phosphat |
| ddTTP | dideoxythiamine tri phosphate |
| ddGTP | dideoxyguanine tri phosphate |
| ddCTP | dideoxycytosine tri phosphate |
| EB | extraction buffer |
| EC | enzyme commission |
| EDTA | ethylenediaminetetraacetic acid |
| EMBL | European Molecular Biology Laboratory |
| ESTs | expressed sequence tags |
| EtBr | ethidium bromide |
| EtOH | ethanol |
| FSH | ferulate 5-hydroxylase |
| Gb | gigabase |
| GO | Gene Ontology |
| HCl | hydrochloride |
| IPTG | isopropyl- β -D-thiogalactopyranosid |
| IP | internet protocol |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| LB | luria brutani |
| MEP | methylethylerythritol |

LIST OF SYMBOLS

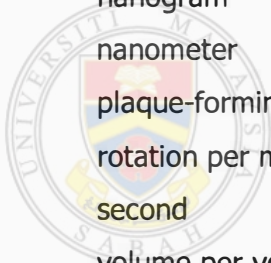
| | |
|-----------|------------------|
| % | percentage |
| > | more than |
| < | less than |
| \leq | less or equal to |
| = | equal to |
| \approx | approximately |
| / | per |
| A | absorbance |
| λ | lambda |



UMS
UNIVERSITI MALAYSIA SABAH

LIST OF UNITS

| | |
|--------------|---------------------|
| bp | basepair |
| cm | centimeter |
| kg | kilogram |
| kb | kilobase |
| μ l | microliter |
| μ g | microgram |
| M | molar |
| m | meter |
| Mbp | megabasepair |
| min | minute |
| ml | mililiter |
| mM | milimolar |
| ng | nanogram |
| nm | nanometer |
| pfu | plaque-forming unit |
| rpm | rotation per minute |
| sec | second |
| v/v | volume per volume |
| μ M | micromolar |
| $^{\circ}$ C | degree celcius |
| w/v | weight per volume |



UMMS
UNIVERSITI MALAYSIA SABAH

LIST OF EQUATIONS

| | | Page |
|------------|--------------------------------|------|
| Equation 1 | $N = X-7$ | 35 |
| Equation 2 | PCR cycles for normalized cDNA | 35 |



UMS
UNIVERSITI MALAYSIA SABAH

LIST OF APPENDICES

Page

| | | |
|------------|--|-----|
| Appendix A | Sequences from single pass sequencing of green mature pineapple clones. | 199 |
| Appendix B | Blast search of singletons assembled from green mature pineapple sequences with significant identity to non-redundant NCBI database. | 203 |
| Appendix C | Contigs from <i>de novo</i> assembly of ripe yellow paired-end Solexa reads. | 207 |



UMS
UNIVERSITI MALAYSIA SABAH

CHAPTER 1

INTRODUCTION

1.1 Introduction

The pineapple (*Ananas comosus* var. *comosus*), which is a member of the Bromeliaceae, is an economically important tropical fruit. Pineapple together with three other dominant tropical fruits (mango, papaya and avocado) are referred to as "major tropical fruits" as they account for the approximately 75% of global flesh tropical fruit production. The overall productions of pineapples fruit over the past few years has showed an increase and are expected grow in the global demand on the pineapple fruit flesh.

Pineapple fruit is mainly used in the processing industry to make canned pineapple and pineapple juice concentrate. Even though there is a very high demand for the fresh pineapple fruit, the short storage life of pineapple and the occurrence of blackheart disease disorder that is easily induced during storage, has hinder further export of pineapple fruits for direct consumption (Zhou *et al.*, 2003). As such, the export of pineapple is only limited to nearby countries.

Pineapple is a non-climacteric fruit where there is no increase in respiration and ethylene production upon ripening (Moyle *et al.*, 2005b). Therefore, the sweetness of the fruits relies on the time it is harvest. For climacteric fruits such as banana and tomato, the ripening process which follows the ethylene biosynthetic pathway is well characterized (Yang and Hoffman, 1984). In contrast, the mechanism of ripening in non-climacteric fruits such as pineapple, citrus and grape is totally unknown (Giovannoni, 2004).

Expressed Sequence Tags (ESTs) is a powerful tool for gene discovery, gene mapping, and for the analysis of quantitative traits. ESTs are partial sequencing of randomly picked cDNA clones generated by reverse transcription of mRNA. A large number of ESTs collections for various organisms representing

libraries of different tissue and development stages are available in the GenBank EST database, dbEST. As there is a need to sequence large numbers of clones to be able to isolate most if not all the transcripts in an organisms, sequencing of a single library has shifted to large scale sequencing generating EST libraries of more than 10,000 clones. These large scales sequencing has no doubt been able to identify a great number of transcripts but the overall library construction methodology is laborious, time consuming and expensive.

The emergence of next generation sequencing technology has brought molecular study to gain a deeper insight into the mechanisms regulating DNA and RNA level. Instead of a clone-by clone sequencing approach, the massively parallel sequencing, provide a better approach as this sequencing technology greatly reduces the costs, time, labour, errors associated by clone mishandling and also reduces bias associated with the type of vector used in during cloning (Weber, 2007). Aside from the capability to capture large amount of transcripts in a single sequencing reaction, the data generated were able to provide quantitative measurement of the levels of genes expression. This study attempts to both the gene discovery and the identification of up and down-regulated genes by comparison of the transcripts expression.

As of Feb 2011, the pineapple's EST in the publicly available NCBI database only account for approximately 6,000 sequences. Most of the sequences deposited were from pineapple nematode-infected gall cDNA library and root tips cDNA library. Only a small portion of the sequences were generated from pineapple fruit tissue. The limited number of pineapple fruit transcripts available hampers the understanding of the mechanism governing non-climacteric fruit, pineapple. This study applied both the Sanger sequencing and massively parallel sequencing using Solexa paired end sequencing to generate sequence data on the pineapple fruit transcriptome.

1.2 Objectives of the Study

The objectives are;

- a) To identify pineapple mRNA transcripts through the construction of a fruit flesh EST library and assembly of sequences generated from Solexa paired-end sequencing reads.
- b) To characterize and annotate the pineapple transcripts through similarities search against non-redundant NCBI GenBank database and against both GO and KEGG databases respectively.
- c) To identify Type I Simple Sequence Repeats (SSRs) in pineapple fruit transcripts through different motifs searches.



UMS
UNIVERSITI MALAYSIA SABAH