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

ONG WEN DEE

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	

Disahkan Oleh

(TANDATANGAN PENULIS) (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

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

Signature

1. **SUPERVISOR**

   Associate Professor Dr. Vijay Kumar

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 trankriptome ekspresi gene. Dengan itu, objektif kajian adalah pertamanya pembinaan perpustakaan jujukan saling melengkapi DNA dan pengelompokan pasangan hujung ke hujang jujukan Solexa. Keduanya, adalah menjelaskan transkript yang didapati melalui pencarian persamaan dan penentuan fungsi meggunakan 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 dikelompokan 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 metabolisma asas, penghasilan dan pemprosesan asid amino, dinding dan pengangkutan, pembahagian sel, metabolisma dan pembinaan struktur dinding, expresi gen berhubungkait dengan jujukan RNA, transduksi isyarat, protein berkait dengan pertahanan dan tekanan, dan juga metabolisma sekunder. Daripada jumlah ini, 71% mempunyai penanda ontology dengan 35.8% dalam kumpulan fungsi molekular, 33.5% dalam komponen sel dan 30.7% dalam prosses biologi. Penentuan fungsi menggunakan pangkalan data KEGG mendapati sebanyak 13,598 trankript mempunyai fungsi yang sama dengan sejumlah 542 kod enzim yang mana boleh dikelompokan kepada 126 laluan. Daripada jumlah laluan ini 122 didapati berhubung kait dengan metabolism nenas. Penentuan fungsi transcript mendapati kebanyakan transcript terlibat dalam metabolik and proses sel dinding. Ini menunjukan semasa pemasakan buah nenas, aktiviti metabolik giat berlaku. Kajian rangkaian jujukan berulang dalam transkript nenas pula menunjukan 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 kegunaaan 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 expresi 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 pemetaaan komparatif dan genome serta penandaan gene dalam nenas.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

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 phospahte
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-hydroxlase
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	methylerythrito

LIST OF SYMBOLS

%	percentage
>	more than
<	less than
≤	less or equal to
=	equal to
≈	approximately
/	per
A	absorbance
λ	lambda

LIST OF UNITS

bp	basepair
cm	centimeter
kg	kilogram
kb	kilobase
µl	microliter
µg	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
µM	micromolar
°C	degree celcius
w/v	weight per volume

LIST OF EQUATIONS

LIST OF APPENDICES

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.