

***IN SILICO* BASED GENE PREDICTION AND
ANNOTATION OF A LOCAL PATHOGENIC
GANODERMA SP. ISOLATED FROM AN
OIL PALM PLANTATION**

HUDA BINTI MOHAMED DARUL



PERPUSTAKAAN
UNIVERSITI MALAYSIA SABAH

UMS

UNIVERSITI MALAYSIA SABAH

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

**BIOTECHNOLOGY RESEARCH INSITUTE
UNIVERSITI MALAYSIA SABAH
2013**

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN TESIS

JUDUL : _____

IJAZAH : _____

SAYA : _____ SESI PENGAJIAN : _____

(HURUF BESAR)

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

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

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

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

TIDAK TERHAD

Disahkan oleh:

(TANDATANGAN PENULIS)

(TANDATANGAN PUSTAKAWAN)

Alamat Tetap: _____

(NAMA PENYELIA)

TARIKH: _____

TARIKH: _____

Catatan:

*Potong yang tidak berkenaan.

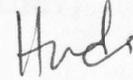
*Jika tesis ini SULIT dan TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai SULIT dan TERHAD.

*Tesis dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah dan Sarjana Secara Penyelidikan atau disertai bagi pengajian secara kerja kursus dan Laporan Projek Sarjana Muda (LPSM).

DECLARATION

I hereby declare that this thesis is my own work and effort and has not been previously submitted for any other award or degree at University Malaysia Sabah as well as other institution. Where other sources of information have been used, they have been acknowledged.

26th April 2013



Huda Binti Mohamed Darul
PB20118203



UMS
UNIVERSITI MALAYSIA SABAH

CERTIFICATION

NAME : HUDA BINTI MOHAMED DARUL
MATRIC NO. : PB20118203
**TITLE : IN SILICO BASED GENE PREDICTION AND ANNOTATION OF
A LOCAL PATHOGENIC GANODERMA SP. ISOLATED FROM
AN OIL PALM PLANTATION**
DEGREE : MASTER OF SCIENCE (MOLECULAR BIOLOGY)
VIVA DATE : 13 JUNE 2013

DECLARED BY

1. SUPERVISOR

Dr. Christopher Voo Lok Yung



UMS

UNIVERSITI MALAYSIA SABAH

Signature

Chris

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to my supervisor Dr. Christopher Voo Luk Yung for his excellent guidance, useful comments, remarks and engagement through the learning process of this master thesis. Thanks for letting me experience this new area of research.

I would like to express the deepest appreciation to Biotechnology Research Institute of University Malaysia Sabah, lecturers, staff and friends which has continually conveyed a spirit of adventure in regard to my master study and research. I've been surrounded by lots of brilliant and impressive people which always motivate me in various ways.

My special appreciation goes to my course mates Desi Eka Sapta Ahmad, Stevell Lumbasi, Ivawati Yunus, Noor Haniza Amit and Siti Norasmah Salam, not only for the encouraging, sharing knowledge, data and helpful discussions; but also for the happy and funny moments we've spent together for this last two years.

The success of this project depends largely on the development various tools. Therefore, special thanks goes to all the developers of a whole bunch of useful packages, all their contributions provided me with the tools with which I have performed my analyses.

Finally, I would like to thank my parents, Mohamed Darul Haji Kamit and Hamidah Jikrun, and for the rest of my family Dono Sukarno, Done Wani, Dorah, Mohd. Adha and Nurkaiyisah for the love and full supports.



HUDA MOHAMED DARUL

ABSTRACT

***IN SILICO* BASED GENE PREDICTION AND ANNOTATION OF A LOCAL PATHOGENIC *GANODERMA SP.* ISOLATED FROM AN OIL PLAM PLANTATION**

Basal stem rot (BSR) which is caused by *Ganoderma* spp. is currently the most prevalent disease in Malaysian oil palm plantations which can causes infection of up to 80% of oil palms in an infected area within 15 years of plantation. As the main producer of palm oil and has the largest planted area of 1.43 million hectares in Malaysia, Sabah oil palm plantation is not excluded from the risk of infection by *Ganoderma* spp. Understanding the molecular mechanism underlying the pathogenesis of BSR is of primary importance to control the disease but there is limited number of genomic resources found at public databases. This present study aims to perform *in silico* based gene prediction and annotation of a local pathogenic *Ganoderma* sp. UMS isolated from infected palm oil tree at Langkon Oil Palm Estate in Sabah. This involved the annotation of the possible genes and transcripts in *Ganoderma* sp. UMS set of contigs using two approaches, sequence similarity search and ab initio. Draft genome, EST, protein, RNA-seq data and gene model of the closely related species of *Ganoderma lucidium* which recently deposited to NCBI and JGI retrieved as the reference sequences. The 50.8-megabases assembled genome with an N50 of 6.24 kb and maximum contig length of 129.8kb predicted to encode about 15,624 genes by using AUGUSTUS and 12,250 of the predicted genes synthesized from evidence alignments using MAKER. Further assessment on the longest contig 140 showed total 38 genes was predicted with the average score of AED was 0.13. Blast2GO analysis provided functional annotation for 9226 genes. The knowledge gained from this *Ganoderma* sp. UMS draft genome will aid in the future functional genomic research to elucidate the host-pathogen interaction in BSR disease.

ABSTRAK

Penyakit reput pangkal batang (BSR) yang disebabkan oleh sejenis kulat *Ganoderma* spp. adalah merupakan penyakit yang paling sinonim dengan aktiviti perladangan kelapa sawit di Malaysia sehingga berpotensi menyebabkan jangkitan kepada hampir 80% kawasan perladangan dalam tempoh masa 15 tahun. Sebagai pengeluar utama minyak sawit dan mempunyai kawasan penanaman yang terbesar di Malaysia, Sabah tidak terkecuali daripada risiko jangkitan *Ganoderma* spp. ini. Memahami mekanisme interaksi penyakit BSR diperingkat molekul adalah amat penting untuk mengawal penyakit ini. Namun, data genomic yang boleh didapati dalam pangkalan data awam berkaitan genome *Ganoderma* spp. adalah amat terhad. Kajian ini bertujuan untuk melaksanakan ramalan gene dan menganotasi genom secara *in silico* terhadap pathogen *Ganoderma* sp. UMS yang diperolehi daripada ladang kelapa sawit Langkon Sabah. Ini melibatkan kaedah ramalan gen secara ab initio dan homologi melalui program AUGUSTUS dan MAKER dengan menggunakan data EST, RNA-seq dan protein daripada spesis terdekat, *G. lucidum*. Draft genom yang bersaiz 50.8-megabes dengan N50 bernilai 6.24 kb dan contig terpanjang berukuran 129.8kb diramalkan mengkod sebanyak 15,624 jumlah gen melalui program AUGUSTUS dan 12,250 daripada keseluruhan gen tersebut disintesis hasil daripada penjajaran bukti daripada data *G. lucidum* menggunakan program MAKER. Penilaian lanjut ke atas contig 140 yang terpanjang menunjukkan 38 jumlah gen yang dikodkan mempunyai skor purata AED sebanyak 0.13. Analisis menggunakan program Blast2GO telah memberikan anotasi fungsi terhadap 9226 gen daripada keseluruhan gen yang diramalkan. Analisis perbandingan yang dijalankan melibatkan *Ganoderma* sp. UMS dan *G. lucidum* yang dimuat turun sebagai rujukan menunjukkan potensi sejumlah 442 gen adalah berkemungkinan unik bagi sampel yang dikaji. Data yang diperolehi daripada kajian ini dapat membantu dalam kajian di dalam bidang fungsi genomik di masa akan datang dalam menjelaskan mekanisme interaksi penyakit BSR ini.

TABLE OF CONTENTS

| | Page |
|--|----------------|
| TITLE | |
| DECLARATION | ii |
| APPROVAL | ii |
| ACKNOWLEDGEMENT | iii |
| ABSTRACT | iv |
| ABSTRAK | v |
| LIST OF CONTENTS | vi-viii |
| LIST OF TABLES | ix |
| LIST OF FIGURES | x |
| LIST OF APPENDIX | xi |
| CHAPTER 1: INTRODUCTION | 1-3 |
| CHAPTER 2: LITERATURE REVIEW | 4-29 |
| 2.1 Basal Stem Rot (BSR) of Oil Palm in Malaysia | 4 |
| 2.2 The genus Ganoderma | 5 |
| 2.2.1 The Ganoderma genome and current status of genomic researches | 6 |
| 2.2.2 Phanerochaete chrysosporium as the model organism of basidiomycetes | 10 |
| 2.3 Next generation sequencing technology and genome assembly | 11 |
| 2.4 The concept of gene and transcriptome structure | 12 |
| 2.5 Eukaryote genome annotation | 15 |
| 2.5.1 Repeat identification | 15 |
| 2.5.2 <i>Ab initio</i> gene prediction | 16 |
| 2.5.3 Gene prediction program training | 17 |
| 2.5.4 The accuracy of gene prediction program | 17 |
| 2.5.5 Homology based prediction | 18 |
| 2.6 Building complete transcriptome map using RNA-seq | 19 |
| 2.7 Bioinformatic tools and algorithm for genome assembly, prediction and annotation | 21 |
| 2.7.1 Commercial and open-source software | 22 |
| 2.7.2 Automated annotation pipeline | 22 |
| 2.7.3 BLAST | 26 |

| | |
|--|--------------|
| CHAPTER 3: MATERIALS AND METHODS | 29-34 |
| 3.1 Data set used for annotation | 29 |
| 3.2 Data sets used for references and evidences file | 30 |
| 3.3 Gene prediction using AUGUSTUS | 31 |
| 3.4 Genome annotation using MAKER automatic annotation pipeline | 32 |
| 3.4.1 Basic input files | 32 |
| 3.4.2 Repeat identification in MAKER | 32 |
| 3.4.3 EST, Protein and RNA-seq data alignment | 32 |
| 3.5 Functional annotation using Blast2go | 35 |
| 3.6 Assessment of the annotation result on contig 140 using CLC Genomics Workbench | 35 |
| 3.7 Comparative analysis between <i>Ganoderma</i> sp. UMS, <i>Ganoderma</i> sp. 10597 v1.0 and <i>G.lucidum</i> 260125-1 | 36 |
| CHAPTER 4: RESULT | 37-52 |
| 4.1 Gene prediction and genome annotation | 37 |
| 4.2 Repeat content identification | 40 |
| 4.3 An overview of the annotation result on contig 140 using CLC genomic workbench | 42 |
| 4.3.1 Detailed features coordinate of ganobriumscontig140.g356.t1 | 42 |
| 4.5 Functional annotation of <i>Ganoderma</i> sp. predicted gene model | 50 |
| 4.6 Comparative analysis between <i>Ganoderma</i> sp. UMS, <i>Ganoderma</i> sp. 10597 v1.0 and <i>G.lucidum</i> 260125-1 | 50 |
| CHAPTER 5: DISCUSSION | 53-62 |
| 5.1 <i>Ganoderma</i> sp. UMS genome sequencing and assembly | 54 |
| 5.2 <i>Ganoderma</i> sp. UMS gene prediction and genome annotation | 55 |
| 5.2.1 Repeatmasking | 58 |
| 5.2.2 The challenge in annotation | 59 |
| 5.3 Comparative analysis between <i>Ganoderma</i> sp. UMS and <i>G. lucidium</i> | 60 |
| 5.4 Data processing, management and integration | 60 |
| 5.5 Contribution of RNA-seq data to genome annotation | 61 |

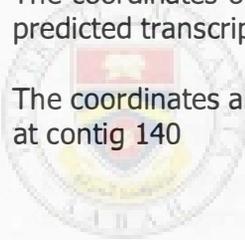
| | |
|------------------------------|--------------|
| CHAPTER 6: CONCLUSION | 56 |
| REFERENCES | 64-69 |
| APPENDIX A | 70 |
| APPENDIX B | 71-86 |
| APPENDIX C | 87 |



UMS
UNIVERSITI MALAYSIA SABAH

LIST OF TABLES

| | Page | |
|-----------|---|----|
| Table 2.1 | Summary of JGI <i>Ganoderma</i> sp. 10597 SS1 v1.0 draft genome | 8 |
| Table 2.2 | <i>G. lucidum</i> 260125-1 genome information | 8 |
| Table 2.3 | Taxonomy comparison between <i>Ganoderma</i> sp and <i>P. chrysosporium</i> | 10 |
| Table 2.4 | <i>Ab initio</i> gene prediction program | 17 |
| Table 2.5 | Different variants of BLAST | 26 |
| Table 2.6 | GFF3 sequence annotation format | 27 |
| Table 3.1 | The preassembled draft genome of <i>Ganoderma</i> sp. UMS | 29 |
| Table 4.1 | The summary of the predicted gene | 39 |
| Table 4.2 | Genome comparison of <i>Ganoderma</i> sp and <i>G. lucidum</i> | 39 |
| Table 4.3 | Summary of repeat masking output | 40 |
| Table 4.4 | Summary of repeat content in <i>Ganoderma</i> sp. genome | 41 |
| Table 4.5 | The coordinates of ganobriumscontig140.g356.t1 predicted transcript | 43 |
| Table 4.6 | The coordinates and AED score of the predicted transcript at contig 140 | 48 |



UMS
UNIVERSITI MALAYSIA SABAH

LIST OF FIGURES

| | | Page |
|------------|---|------|
| Figure 2.1 | Summary of MPOB <i>Ganoderma</i> sequencing project | 9 |
| Figure 2.2 | Eukaryotic gene structure | 14 |
| Figure 2.3 | Evaluation of prediction accuracy at exon level | 18 |
| Figure 2.4 | Different approaches to annotate a genome which compared based on relative time, effort and the degree that they dependent on external evidence like RNA-seq data | 20 |
| Figure 2.5 | Automatic annotation pipeline | 23 |
| Figure 2.6 | Different tools and analysis involved in JGI annotation pipeline | 24 |
| Figure 3.1 | Data sets for references from <i>G. lucidum</i> database | 30 |
| Figure 3.2 | AUGUSTUS training process | 31 |
| Figure 3.3 | MAKER web annotation service(MWAS) | 34 |
| Figure 3.4 | Tract tools function in CLCGW used to align the annotated features of <i>Ganoderma</i> sp. | 35 |
| Figure 3.5 | <i>Ganoderma</i> sp. UMS annotation workflow | 36 |
| Figure 4.1 | Ab initio gene prediction workflow using AUGUSTUS | 38 |
| Figure 4.2 | Annotation of ganobriumscontig.g356.t1 gene after all available evidences synthesized | 44 |
| Figure 4.3 | Annotation of ganobriumscontig.g357.t1 gene after all available evidences synthesized | 45 |
| Figure 4.4 | Annotation of ganobriumscontig.g358.t1 gene after synthesizing all evidence | 46 |
| Figure 4.5 | Annotation of ganobriumscontig.g359.t1 gene with low coverage RNA-seq mapping | 47 |
| Figure 4.6 | Blast2Go functional annotation results overview | 51 |
| Figure 4.7 | Top-Hit species distribution | 51 |

LIST OF APPENDIX

| | Page | |
|--------------|--|----|
| Appendix A | Supplementary notes | 70 |
| Appendix A.1 | Origin of the <i>Ganoderma</i> sp. UMS sample and sequencing | 70 |
| Appendix A.2 | Assembly process of <i>Ganoderma</i> sp. UMS genome | 70 |
| Appendix B | Supplementary data | 71 |
| Appendix B.1 | List of 38 gene predicted in contig 140 by MAKER | 71 |
| Appendix C | Supplementary files provided in DVD-ROM | 87 |



UMS
UNIVERSITI MALAYSIA SABAH

CHAPTER 1

INTRODUCTION

Oil palm (*Elaeis guineensis*) is an economically important crop in Malaysia since the country has become the world's largest exporter of palm oil, contributing ~65% of the global palm oil demand. The production of palm oil in Malaysia reached up to 18.9 million tonnes in 2011 which is in line with the increase in the total oil palm planted area to 5 million hectares (Malaysian Palm Oil Statistics, 2011). As the main producer of palm oil of all states in Malaysia, palm oil industry is also a significant business sector in Sabah, since it has the largest planted area of 1.43 million hectares.

Despite this rapid expansion of oil palm industry in Malaysia, the oil palm plantations face major constraint of the incidence of Basal Stem Rot (BSR) disease caused by *Ganoderma* spp. *Ganoderma* is one of the important wood-decaying types of fungi. There were 15 species of *Ganoderma* that have been reported associated with BSR, but in Malaysia it is predominantly caused by pathogenic strain of *G. boninense* (Utomo *et al.*, 2005).

BSR can be spread through direct contacts with adjacent infected root or by spreading the fungal spores, thus complicate the disease control. BSR process involving the degradation of lignin and cellulose and followed by the growth of the *Ganoderma* within the oil palm tissue. The symptoms of BSR are normally comprise the failure of young leaves to open, leaves turn yellow and the appearance of *Ganoderma* fruiting body on the trunks or roots. However, the symptoms of BSR can only be observed when at least half of the basal tissues have been affected by *Ganoderma*.

BSR has been proved to cause economic loss of oil palm due to total reduction in weight and number of fruit bunches. Eventually, BSR will shorten the productive life of oil palm plantation. This is not just occurring in Malaysia but other oil palm growing countries as well such as Indonesia (Ariffin *et al.*, 2000) which make it as presently the most prevalent and devastating disease in oil palm plantations.

As an economically important fungal pathogen, understanding the molecular mechanism underlying the interaction between host (*Elaeis Guineensis*) and pathogen (*Ganoderma* spp.) is of primary importance in developing strategies to control the disease. However, the genomic resources of *Ganoderma* spp. are still limited and the genetic basis of their mechanism remains poorly defined. This is based on the small numbers of genomic resources such as genome and transcriptome sequence data can be found at any publicly accessible databases. Therefore, further development of *Ganoderma* spp. genomic resources is very much needed to facilitate detailed studies of its system biology. Sequencing data that can be readily produced by next-generation sequencing platforms which currently become active research areas can provide valuable genomic information of *Ganoderma* spp.

To date, most research on *Ganoderma* has focused more on the species of *G. lucidum*, which widely known of its medicinal properties. Study done by Chen *et al.* (2012) was successfully sequenced the 43.3 Mbps genome of *G. lucidum* at chromosome level. Other than that, the available genomic resources of *Ganoderma* sp. still at contigs and scaffolds level, such as the one produced at Joint Genome Institute (JGI). Even though there are already few studies have been done dealing with sequencing of the *Ganoderma* genome but studies had shown that dominant *Ganoderma* spp for BSR can differ based on locality (Wong *et al.*, 2012). Therefore, this present study focused on the local pathogenic *Ganoderma* isolated from Langkon Oil Palm Estate in Sabah Malaysia, as Sabah is the largest oil palm planted state. The objectives of this present study are:

- a. To predict local pathogenic *Ganoderma* sp. UMS's gene using *ab initio* method in AUGUSTUS program.
- b. To annotate the genome of local pathogenic *Ganoderma* sp. UMS structurally and functionally using MAKER and Blast2GO program by integrating RNA-seq, EST and protein of publicly available data of *Ganoderma* spp.
- c. To do comparative analysis between *Ganoderma* sp. UMS and *G. lucidum* draft genome.

Overall, this present study focuses on the annotation of local pathogenic *Ganoderma* sp. UMS isolated from oil palm plantation. Description of all functional elements in the *Ganoderma* sp. system is a prerequisite for conducting future holistic systems approaches to understand the complex pathogenesis of BSR.

CHAPTER 2

LITERATURE REVIEW

2.1 Basal Stem Rot (BSR) of Oil Palm in Malaysia

Oil palm (*Elaeis guineensis* Jacquin) well known to have originated from the Guinea coast of West Africa and has long been introduced in other regions as early as 15th century. Before its commercial value was known, it was often grown as ornamental plants at those areas.

It is an economically important crop in Malaysia since the country has become the world's largest exporter of palm oil, contributing ~65% of the global palm oil demand. The production of palm oil in Malaysia reaches up to 18.9 million tonnes in 2011 which is in line with the increase in the total oil palm planted area to 5 million hectares (Malaysian Palm Oil Statistics, 2011).

Despite this rapid expansion of oil palm industry in Malaysia, the oil palm plantations face major constraint of the incidence of Basal Stem Rot (BSR) disease caused by *Ganoderma* spp. It was first reported in Malaysia in 1931 (Ho & Nawawi, 1985) and the causal agent was identified as *G. lucidum*.

Researches looking into the causal pathogen of BSR later have revealed that this important disease caused predominantly by the subspecies of *Ganoderma boninense* which appears to be restricted to palms oil. *Ganoderma boninense* has been identified as the major pathogen to oil palm in Malaysia and Papua New Guinea (Pillotti, 2001). Chong (2011) showed that the Sabah isolates from Langkon Oil Palm Estate were very similar to virulent *G.boninense* strains with a maximum similarity of 98%.

BSR also has been associated with a number of species out of at least 15 species ranging from *G. boninense*, *G. miniatocinctum*, *G. chalceum*, *G. tornatum*, *G. zonatum* and *G. xylonoides*.

2.2 The genus *Ganoderma*

Back in 1881, a mycologist named Peter Adolf Karsten introduced the genus *Ganoderma* and the species of *G. lucidium* (Seo & Kirk, 2000). The genus originated from the family *Ganodermataceae* from the order of Polyporales that resides in the class of Basidiomycetes. The genus of *Ganoderma* was then further subdivided into *Ganoderma* and *Elfvigia* subgenus differentiated by laccate species based on *G. lucidium* and non-laccate species based on *G. applanatum* respectively.

Ganoderma spp are one of the important wood decaying types of fungi. These fungi can cause white-rot of woods through the process of delignification. The component of plants cell wall mainly comprise of lignin, pectins, cellulose and hemicellulose. During the *Ganoderma* spp. infection, these components need to be degraded first in order to penetrate the plants using polymer-cleaving enzymes (Susanto, 2009). Lignin is a product of phenylpropanoid pathway, the evolution of phenol metabolism (Sabrina *et al.*, 2012).

In the early stages of infection, the infected area will be turned into bleached zones as due to the delignification process. As the infection progresses, the infected area will become softer and the wood eventually loses its strength.

The late stage of infection only can be detected by the appearance of the fruiting bodies or basidiocarps on the infected area. These basidiocarps become the agent of infection by spreading the basidiopores (Sanderson, 2005) in which later incorporated into the wider area, further facilitate the infection to other trees.

2.2.1 The *Ganoderma* genome and current status of genomic researches

The draft genome of *Ganoderma* sp. 10597 SS1 v1.0, a North American isolate has been published by US Department of Energy Joint Genome Institute (<http://www.jgi.doe.gov/>) in collaboration with the user community. The size of 39.52 Mbp genome is assembled into 503 contigs in 156 scaffolds. These data produced from the combination of different sequencing technologies that are Sanger reads, Illumina data and Roche (454) paired end data. The assembler used in this project was Newbler and annotated using the JGI Genome Annotation pipeline.

The *Ganoderma* sp. 10597 SS1 v1.0 gene model was primarily based on EST sequences and protein coding genes in related species. The prediction of gene model was able to incorporate 97.7% of the EST evidence available into 12,910 protein coding gene models. The *Ganoderma* sp. 10597 SS1 v1.0 draft summarized in the Table 2.1 below.

Other than the data published by JGI, the most recent *Ganoderma* genomic information published at June 2012 by Chen *et al.* (2012) which successfully released the 43.3 Mb complete genome sequence of monokaryotic *Ganoderma lucidum* strain 260125-1 up to chromosomes level. *G. lucidum* has been known widely as medicinal macrofungus in traditional Chinese medicine. The genome

analysis revealed 13 numbers of chromosomes which encoding 16,113 predicted genes (Table 2.2). This current gene model provides the most comprehensive gene model to date for *Ganoderma* genus while other data only available at scaffold or contig level.



UMS
UNIVERSITI MALAYSIA SABAH

Table 2.1: Summary of JGI *Ganoderma* sp. 10597 SS1 v1.0 draft genome

| Genome | |
|-----------------------------|-------|
| Assembly Size (Mbp) | 39.52 |
| No. of Contigs | 503 |
| No. of Scaffolds | 156 |
| Scaffold N50 | 6 |
| ESTs | |
| ESTClusters total sequences | 15411 |
| % Mapped to genome | 97.7% |
| Gene Model | |
| No. of gene model | 12910 |
| Mean Gene Length | 1794 |
| Transcript (bp) | 1407 |
| Exon (bp) | 242 |
| Intron (bp) | 82 |
| Protein Length (aa) | 427 |
| Exon per Gene | 5.82 |

Source : (<http://genome.jgi.doe.gov/Gansp1/Gansp1.home.html>)

Table 2.2: *G. lucidium* 260125-1 genome information

| Genome | |
|--------------------------------------|--------|
| Number of chromosomes | 13 |
| Length of genome assembly (Mb) | 43.3 |
| GC content (%) | 55.9 |
| Protein-coding genes | 16,113 |
| Average gene length (bp) | 1,556 |
| Mean exons per gene | 4.7 |
| Mean exon size (bp) | 268 |
| Mean intron size (bp) | 87 |
| Mean size of intergenic regions (bp) | 1206 |

Source : (Chen *et al.*,2012)

In record, there are two major companies in Malaysia that have claimed their involvement in *Ganoderma* sequencing project; Malaysia Palm Oil Board (MPOB), Malaysian Genomics Resource Centre (MGRC) and Asiatic Genome Technology Centre Sdn Bhd (AGTC), a subsidiary of Genting Plantations. This is based on the announcement that has been made in their official website in 2010 at (<http://www.acgt.asia/press/>) and (<http://www.mpob.gov.my>). MPOB has successfully sequenced and assembled three pathogenic species (*G. boninense*, *G. zonatum*, *G. miniatocinctum*) and one non-pathogenic species (*G. tornatum*) with the genome size of *G. boninense* approximately 51Mb in scaffolds and reaches 60Mb in contigs form (Figure 2.1). These invaluable genomic resources together with the rapid advancement of genomic tools will further accelerated the progress in BSR and *Ganoderma* researches.

| | <i>G. boninense</i> | | | <i>G. miniatocinctum</i> | | |
|----------------------------------|---------------------|----------------|----------|--------------------------|----------------|----------|
| | # of Objects | Cumulative MBP | N50 (kb) | # of Objects | Cumulative MBP | N50 (kb) |
| Raw Data | | | | | | |
| Reads | 2,369,959 | 863.18 | | 2,676,052 | 950.88 | |
| Effects of Linking | | | | | | |
| Contigs (before linking) | 54,700 | 60.10 | 2.55 | 55,951 | 62.34 | 2.52 |
| Scaffolds (after linking) | 1,871 | 50.57 | 93.13 | 2,374 | 51.88 | 79.46 |
| | <i>G. tornatum</i> | | | <i>G. zonatum</i> | | |
| | # of Objects | Cumulative MBP | N50 (kb) | # of Objects | Cumulative MBP | N50 (kb) |
| Raw Data | | | | | | |
| Reads | 2,398,318 | 787.42 | | 2,425,717 | 966.98 | |
| Effects of Linking | | | | | | |
| Contigs (before linking) | 64,015 | 65.67 | 2.36 | 48,479 | 54.65 | 2.80 |
| Scaffolds (after linking) | 2,269 | 51.72 | 82.62 | 3,415 | 42.98 | 31.96 |

Figure 2.1: Summary of MPOB *Ganoderma* sequencing project

2.2.2 *Phanerochaete chrysosporium* as the model organism of basidiomycetes

P. chrysosporium is the most intensively studied white rot basidiomycete which differ with *Ganoderma* at family level. It has the selective ability to degrade the abundant polymer of lignin because it can releases extracellular enzymes to break down lignin. Only a small group of fungi has this ability which leaves behind crystalline cellulose that referred to as "white rot". *P. Chrysosporium* is the first member of the Basidiomycetes to have its complete genome sequenced which consists of approximately 29.6 million base pairs in ten chromosome (Martinex *et al.*, 2004). As a model of organism, it has been used in gene prediction software to predict gene structures and exon/intron junction such as FGENESH (Softberry, Mount Kisco, NY) and AUGUSTUS (Stanke & Waack, 2003). The used of *P. Chrysosporium* in the gene prediction analysis of *Ganoderma* is feasible as they are the same member of Basidiomycetes (Table 2.3).

Table 2.3: Taxonomy comparison between *Ganoderma* sp and *P. chrysosporium*

| <i>Ganoderma</i> sp. | Taxonomy | <i>Phanerochaete chrysosporium</i> |
|-----------------------------|-----------------|---|
| Fungi | Kingdom | Fungi |
| Basidiomycota | Phylum | Basidiomycota |
| Basidiomycetes | Class | Basidiomycetes |
| Agaricomycetidae | Subclass | Agaricomycetidae |
| Polyporales | Order | Polyporales |
| <i>Ganodermatacea</i> | Family | Phanerochaetaceae |
| <i>Ganoderma</i> | Genus | <i>Phanerochaete</i> |

2.3 Next generation sequencing technology and genome assembly

The objective of doing genome sequencing is to reveal the ordered sequence of DNA molecule which made up by nucleic acids. This can be achieved by randomly breaking the long strand of DNA into smaller and overlapping fragments which is called 'reads'. The advent of second generation sequencing technologies developed by several companies such as Solexa (Illumina), Roche (454), SOLID (ABI) and Helicos, has facilitated the molecular biology researches, particularly the study of functional genomics. Year 2004 is a remarkable year that has revolutionized the field of molecular biology with the advent of massively parallel sequencing technology or the Next Generation Sequencing (NGS) which seen Roche (454) Genome Sequencer become the pioneer of NGS instruments. The newest Roche (454) technology able to produce greater than 1 million reads up to 400bp. In 2006, Illumina come out with Genome analyser (GA) that capable to generate ten millions of 32bp reads (Costa *et al.*, 2010).

The differences of data produced by these various technologies revolve around sample preparation, chemistry, type and volume of raw data (McPherson, 2009). The 454 technology may produce errors regarding the insertion and deletion during the sequencing of the homopolymer regions due to the longer reads. The 454 platform has the ability to generate fewer but longer sequences compared to Illumina platform which producing shorter reads in large amounts. Genome sequencing project able to produce three different types of data which are raw data or 'reads', contigs and scaffolds.