WHOLE GENOME SEQUENCING AND FUNCTIONAL FEATURES OF A NOVEL BIOSURFACTANT-PRODUCING *Bacillus subtilis* UMX-103

YOUSRI ABDELMUTALAB AHMED ABDELHAFIZ

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

WHOLE GENOME SEQUENCING AND FUNCTIONAL FEATURERS OF A NOVEL BIOSURFACTANT-PRODUCING *Bacillus subtilis* UMX-103

YOUSRI ABDELMUTALAB AHMED ABDELHAFIZ

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: YOUSRI ABDELMUTALAB AHMED ABDELHAFIZ

Matric No: SGR150071

Name of Degree: Master of Science

Title of Thesis: ("Whole Genome Sequencing and Functional Features of A Novel Biosurfactant-Producing *Bacillus subtilis* UMX-103"):

Field of Study: Bioinformatics (Biology and Biochemistry)

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature

Name:

Designation:

Date:

Witness's Signature

Name:

Designation:

Date:

ABSTRACT

The genus *Bacillus* is a Gram-positive, aerobic, endospore-forming, rod-shaped bacterium commonly found in the environment that have important industrial, medical, agriculture and environmental values. This study characterized a novel biosurfactant producing bacterial strain UMX-103 which was isolated from a hydrocarbon contaminated site in Terengganu, Malaysia. An integration of both genomics and biochemical approaches were conducted to analyse the biosurfactant production by the strain UMX-103. Determination of biosurfactant production by the strain was conducted using five different assays including; Hemolytic assay, Oil spreading test, Drop-collapse assay, Emulsification assay and Surface tension measurements. The surface tension test showed that the strain is able to lower surface tension up to (26.4 ± 0.02 mN/m). UMX-103 also showed positive results in the other assays. Whole Genome Sequence analysis revealed the genetic contents and genes involved in biosurfactant production. The whole genome was assembled using a combination of both *de novo* and reference-guided assembly methods. The genome size of UMX-103 is 4,234,627 bp with 4399 genes comprising of 4301 protein-coding genes and 98 RNA genes. The mapping results showed 93.44% of genome similarity with *B. subtilis* strain 168. The functional annotation analysis revealed present of surfactin biosynthetic gene cluster. This gene cluster belongs to the Non-ribosomal Peptide Synthetase (NRPS) family, which is one of the microbial surfactant groups. A total of 25 genes were identified that involved in biosurfactants production. Among these genes, 14 genes were involved in surfactin biosynthesis and while the remaining genes were involved in surfactin regulation. The comparative genomics and Pangenome analysis of UMX-103 with other *Bacillus sp.* highlighted unique features of this strain, especially relating to the biosurfactant gene cluster.

ABSTRAK

Bakteria genus *Bacillus* adalah gram-positif, aerobik, membentuk endospora dan berbentuk rod yang ditemui ditemui di alam sekitar dan mempunyai kepentingan dalam industri, perubatan, pertanian dan alam sekitar. Kajian ini memperincikan bakteria penghasil biosurfaktan UMX-103 yang telah dipencilkan dari laman tercemar hidrokarbon di Terengganu, Malaysia. Integrasi pendekatan genomik dan kimia telah dijalankan untuk menganalisis pengeluaran biosurfaktan oleh UMX-103. Penentuan pengeluaran biosurfaktan dijalankan dengan menggunakan lima asai yang berbeza termasuk; asai hemolitik, asai penyebaran minyak, asai 'drop collapse', asai pengemulsian dan ukuran ketegangan permukaan. Ujian ketegangan permukaan menunjukkan bahawa UMX-103 berupaya menurunkan ketegangan permukaan sehingga (26.4 ± 0.02 mN / m). UMX-103 juga menunjukkan hasil yang positif dalam semua asai-asai lain. Analisis jujukan genom mendedahkan kandungan genetik dan gen yang terlibat dalam pengeluaran biosurfaktan. Jujukan genom telah ditentukan dengan menggunakan gabungan kaedah *de novo* dan himpunan berpandukan rujukan. Saiz genom UMX-103 adalah 4,234,627 bp dengan 4,399 gen yang terdiri daripada 4,301 gen pengekodan protein dan 98 gen RNA. Keputusan pemetaan menunjukkan 93.44% daripada genom UMX-103 mempunyai persamaan dengan *B. subtilis* 168. Analisis anotasi untuk fungsi mendedahkan kehadiran gen biosintetik 'surfactin' berkelompok. Ini adalah kelompok gen yang tergolong sebagai keluarga 'Non –ribosomal Peptide Synthetase' (NRPS), yang merupakan salah satu kumpulan surfaktan mikrob. Sebanyak 25 gen yang telah dikenal pasti terlibat dalam pengeluaran biosurfaktan. Diantara gen ini, 14 gen terlibat dalam biosintesis surfaktin manakala bakinya terlibat dalam pengawalaturan surfaktin. Genomik perbandingan dan analisis pangenom terhadap UMX-103 menonjolkan ciri-ciri unik UMX-103 ini.

ACKNOWLEDGEMENTS

In the name of Allah the Most Merciful and the Most Beneficial, I would like to thank his Grace for bestowing me all of these people and the university who played an integral part in my thesis and research:

University of Malaya and Institute of Biological Sciences and CRYSTAL for the facilities.

Prof. Dr. Amir Feisal Merican and Dr. Saharuddin Bin Mohamd for their supervision, knowledge, criticism and patience.

All my friends & colleagues: Dr. Vaani, Ahmad Ali, Kamil Brima, Dr. Anis Al-Maleki, Dr. Hoda, Hamidah and Nava who have given me lots of moral support and advice during my research journey.

My beloved father for his support and guidance to the right path (May Allah Grant him Aljanah).

My mother and family for their patience and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

ANI	:	Average Nucleotide Identity
BAM	:	Binary Alignment Map
bp	:	base pair
BWA	:	Burrows-Wheeler Aligner
ChIP	:	Combination of Chromatin Immunoprecipitation
ChIP-Seq	:	Combination of Chromatin Immunoprecipitation followed by Sequencing
COGs	:	Cluster Orthologues Groups
DDH	:	DNA-DNA Hybridization
DDBJ	:	DNA Data Bank of Japan
DNA	:	Deoxyribose Nucleic Acids
EMBL	:	European Molecular Biology Laboratory
FESEM	:	Field Emission Scanning Electron Microscope
Gbp	:	Giga base pair
Go	:	Gene ontology
GI	:	Genomic Island
GRAS	:	Generally Regarded As Safe
GWAS	:	Genome Wide Association Studies
HGT	:	Horizontal Gene Transfer
HLB	:	Hydrophilic-Lipophilic Balance
HMW	:	High Molecular Weight
KEGG	:	Kyoto Encyclopedia of Genes and Genomes
LMW	:	Low Molecular Weight
MGD	:	Mouse Genome Database

MLST	:	Multilocus Sequence Typing
mN/m	:	milli Newton Per meter
MORE	:	Microbial Enhanced Oil Recovery
NGS	:	Next Generation Sequencing
NRPS	:	Non-ribosomal Peptide Synthetase
OBBO	:	Open Biological and Biomedical Ontologies
PCR	:	Polymerase Chain Reaction
rRNA	:	Ribosomal RNA
SAM	:	Sequence Alignment Map
SGD	:	*Saccharomyces* Genome Database
ST	:	Surface Tension
TSA	:	Tryptone Soya Agar
TSB	:	Tryptone Soya Broth
USFDA	:	United States Food and Drug Administration
WGS	:	Whole Genome Sequencing

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Background

Biosurfactants are amphiphilic molecules produced by microorganisms that are able to lower the surface and interfacial tension between liquid and solid or amongst two liquids. Structurally they comprise both the hydrophobic and hydrophilic moieties which make these surface-active compounds in a high demand for many environmental and industrial applications. Due to environmental issues raised in the past few decades on the impact of synthetic chemical surfactants, biosurfactants have gained attention as alternative surfactants to that of the synthetic chemical surfactants. However, the high cost of production limits their applications. Many studies have been conducted to find ways to reduce the biosurfactant production costs as well as develop cheaper processes using low cost raw materials (Al-Bahry et al., 2013; Gudiña et al., 2015; Mulligan, 2009).

Biosurfactants are environmental friendly due to their low toxicity level, extreme biodegradability, stability to high pH and temperature condition which makes them promising candidate for bioremediation and enhanced oil recovery applications (Mulligan, 2009), food industry (Nitschke & Costa, 2007), pharmaceutical and cosmetics (Banat et al., 2000). Biosurfactant products in the global market was reported to be approximately USD 1735.50 million in 2011, in addition is estimated to reach USD 2210.50 million in 2018 with an average of annual growth rate of 3.5% (Transparency Market Research, 2011).

Biosurfactants are produced by microorganisms. The isolation and characterization of microbial surfactant producers have been reported from diverse environmental habitats (Al-Bahry et al., 2013). Various biosurfactants producing bacteria were reported such as *P. aeruginosa* N002 (Das et al., 2015), *Achromobacter*

spanius (Alvarez et al., 2015), *Rhodococcus erythroplis*. In addition, many strains from the genus *Bacillus* are biosurfactant producers such as *B. licheniformis*, *B. subtilis* and *B. pumilus* (Płaza et al., 2015). Member of the genus *Bacillus* are genetically diverse and able to grow in various environments (Choudhary & Johri, 2009). *Bacillus spp*. are known to produce a variety of bioactive compounds with potentials for biotechnological applications including antibiotic and enzymes that are widely used in environmental and biomedical applications (Banat et al., 2010).

The revolution of DNA sequencing technology has changed the way scientists think about genetics and genome information (Lasken & McLean, 2014; Zhang et al., 2011). DNA sequencing applications allowed many researchers to sequence the complete genome or a region of the genome, which also highlights the genetics contents and compositions (Kamada et al., 2014; Nishito et al., 2010). The era of bacterial genome sequencing began when Sanger sequencing method was used to sequence two bacterial genomes in 1995 (Land et al., 2015). Two years after the introducing of Sanger sequencing, the first complete genome of *Bacillus subtilis* was published on 20th November 1997 (Kunst et al., 1997). Recently, the introduction of Next Generation Sequencing (NGS) and advanced bioinformatics data analysis pipeline have dramatically changed the face of microbiology. NGS applications are widespread and comprise whole genome sequencing, metagenomic and microbiome analyses, pathogen discovery, transcriptome profiling, and infectious disease diagnosis (Loman & Pallen, 2015). *Mycoplasma genitalium* was the first bacterial genome sequenced using NGS technology (Margulies et al., 2005). Since then, thousands of bacterial genomes were sequenced and analysed. Sequence based analyses not only have delivered unexpected insights into microbial diversity and functions, but also have led to track the spread of infection and aided in designing new drugs and vaccines (Loman & Pallen, 2015).

In this study, the bacterium *B. subtilis* UMX-103 which was isolated from hydrocarbon contaminated soil is characterized using both biochemical and genomics approaches to understand the functional genomics and genetics basis of that involve in biosurfactant production by the bacterium. Furthermore, the whole genome sequence of the bacteria is analysed and compared with other *Bacillus* genomes.

1.2 Research questions

1. Is UMX-103 a biosurfactant producing bacteria?
2. What are the genome sequences of UMX-103?
3. What are the biosurfactant genes that are in the genome of the bacteria?
4. What are the pathways present responsible for biosurfactant production?

1.3 Objectives

1. To identify and characterize the functional features of UMX-103 to produce biosurfactants.
2. To analyze the whole genome sequence of *Bacillus subtilis* UMX-103.
3. To identify pathways responsible for biosurfactant production in UMX-103.

1.4 Thesis organization

This thesis contains seven chapters: Introduction, Literature review, Materials and Methods, three results chapters and the last chapter on general discussion and conclusion. The first chapter includes the background and the objectives of this study. The second chapter covers the literature review of biosurfactants classifications, biosynthesis and their applications. It also covers current DNA sequencing technology and the role of bioinformatics in the analysis of NGS datasets. The third chapter

includes the materials and methodology applied in this study, which is divided into two sections. The first section covers the materials and methods used in screening the biosurfactant produced by the bacteria as well as characterization of the bacteria, while the second part covers the methods used in analysing the whole genome sequence of the bacteria. The fourth chapter contains the results and discussion of the first objective of this study, whereas the fifth and sixth chapters cover the second and third objectives. The last chapter in this thesis contains the general discussion and conclusion of the study.

CHAPTER 2: LITERATURE REVIEW

2.1 Biosurfactants

2.1.1 Biosurfactants properties

Biosurfactants or microbial surfactants are a heterogeneous group of extracellular or surface-active molecules that are produced by microorganisms. These molecules tend to have the capability to reduce surface and interfacial tension between two liquids or between liquid and air, also it could stabilize at high pH and temperatures (Shoeb et al., 2015). In general, biosurfactants are amphiphilic molecules that possess both hydrophilic and hydrophobic regions causing them to composite at interfaces between fluids with different polarities such as water and hydrocarbons. Also, they were found to increase the transport of nutrition across membranes and affect the host-microbe interactions. Biosurfactants have gained several advantages compared to chemical or synthetic surfactants; these advantages include their biodegradability, biocompatability and digestibility (Vijayakumar & Saravanan, 2015).

2.1.2 Biosynthesis

The biosynthesis of biosurfactants is based on hydrophilic and hydrophobic moieties. The hydrophilic moiety structural part may be made up of molecules such as carbohydrate, carboxylic acid, phosphate, amino acid, cyclic peptide, or alcohol. The hydrophobic moiety is either made up of a long-chain fatty acid, a hydroxy fatty acid, or a alkylb-hydroxy fatty acid. Two metabolic pathways are proposed for biosurfactants synthesis: the carbohydrate and hydrocarbon pathways (Desai & Banat, 1997). Four possibilities for biosurfactants moieties synthesis and their association were reported. First possibility is that the hydrophobic and hydrophilic moieties are synthesized *de novo* by two independent pathways; second possibility is that the synthesis of the hydrophobic moiety is induced by substrate while the hydrophilic moiety is synthesized

de novo; third possibility is that the hydrophobic moiety is synthesized *de novo*, while the synthesis of the hydrophilic moiety is substrate dependent and the fourth possibility is the synthesis of both the hydrophilic and hydrophobic moieties depend on substrate (Syldatk & Wagner, 1987).

2.1.3 Biosurfactants classification

Biosurfactants are classified according to their microbial origin, molecular weight and their chemical composition. According to their molecular weight they are classified into two categories which are Low Molecular Weight (LMW) and High Molecular Weight (HMW) biosurfactants. LMW biosurfactants includes glycolipids, lipopeptides and phospholipids, while HMW biosurfactants are usually comprise of a mixture of biopolymers such as polysaccharides, lipopolysaccharides and lipoproteins (Pacwa-Płociniczak, Płaza, Piotrowska-Seget, & Cameotra, 2011). There are five main categories of biosurfactants which are; glycolipids, lipopeptides and lipoproteins, phospholipids and fatty acids, polymeric biosurfactants and particulate biosurfactants (Table 2.1). The majority of biosurfactants are glycolipids and amongst the glycolipids group, rhamnolipids, sophorolipids and trehalolipids are the best known (Kuyukina et al., 2015; Soberón-Chávez et al., 2005; Van Bogaert et al., 2007). Lipopeptides are cyclic peptides which are acylated with a fatty acid. In general, there are three classes of lipopeptides which are classified as surfactins, fenqycins and iturins that are mostly produced by *Bacillus* species (Meena & Kanwar, 2015; Stein, 2005). Certain bacteria and yeast produce amount of fatty acids and phospholipids when growing on n-alkanes. *Rhodococcus erythropolis* DSM 43215 produced phosphatidylethanolamine when growing on n-alkane, causing extreme reduction of the interfacial tension between hexadecane and water up to 1 mN/m (Desai & Banat, 1997).

Table 2.1: Biosurfactants classification and their origin

Biosurfactants			
Group	**Class**	**Microorganisms**	**References**
Glycolipids	Rhamnolipids, sophorolipids, trehaloselipids,	*Pseudomonas spp*., *Mycobacterium, Arthrobacter, Actinomycetes, Rhodococcus spp*	(Soberón-Chávez et al., 2005), (Gautam & Tyagi, 2006), (Van Bogaert et al., 2007), (Kuyukina et al., 2015)
Lipopeptides and lipoproteins	Surfactins, Fenqycins, Iturins, daptomycin	*Bacillus subtilis, Bacillus* amyloliquefaciens, *Pseudomonas fluorescens, Bacillus licheniformis, Serratia marcescens*	(Arima et al., 1968), (Davis et al., 2001), (Meena & Kanwar, 2015), (Baltz et al., 2005)
Phospholipids and fatty acids	Phosphatidylethanolamine, corynomycol acids	*Acinetobacter sp.* HO1-N, *Aspergillus* strains, *Arthrobacter* Ak-19, *Rhodococcus erythropolis*	(Johansson & Svensson, 2001), (Kosaric & Sukan, 2014), (Wayman et al., 1984), (Käppeli & Finnerty, 1979), (Desai & Banat, 1997)
Polymeric biosurfactants	Liposan, emulsan, alasan, lipomanan	*Candida lypolytica*, *Acinetobacter calcoaceticus* RAG-1,	(Cirigliano & Carman, 1985), (Amaral et al., 2006), (Gurjar et al., 1995), (Rosenberg et al., 1979)
Particulate biosurfactants	Vesicles, whole microbial cell	*Acinetobacter sp* strain HO1-N, *Acinetobacter calcoaceticus*, *pseudomonas meningitis*,	(Desai & Banat, 1997), (Käppeli & Finnerty, 1979)

2.1.4 Biosurfactants applications

2.1.5 Microbial Enhanced Oil Recovery (MEOR) and Bioremediation

MEOR is a substitute tertiary for oil recovery technology where microbial metabolites and activities are used to enhance the recovery of residual oil from consumed and marginal under oil reservoirs. This technology takes advantage of the ability of indigenous or injected microorganisms to synthesize useful products by fermenting low cost raw materials. MEOR offer major advantages over conventional CEOR, due to the low consumption of energy as the thermal processes, not depending on the price of crude oil as compared to various chemical processes. Moreover, microbial products have low toxicity level and biodegradable (Sen, 2008; Youssef, Elshahed, & McInerney, 2009). One more important method in MEOR is the biodegradation of heavy oil portions by microorganisms. In this process, heavy oil portions are converted into lighter ones, reducing the viscosity of crude oil and improving its mobility through the reservoir, which increases oil recovery (Gudina et al., 2013).

The application of biosurfactants in the remediation of organic compounds, such as hydrocarbons, is to increase their bioavailability or mobilising and removing the contaminants by pseudosolubilisation and emulsification in a washing treatment, while the remediation of inorganic compounds such as heavy metals, on the other hand, is targeted at chelating and removal of ions during the washing step facilitated by the chemical interactions between the amphiphiles and the metal ions (Banat et al., 2010. Bioremediation usually consists of the application of nitrogenous and phosphorous fertilizers, adjusting the pH and water content, if necessary, supplying air and often adding suitable bacteria. Amphiphiles are able to alter the physico-chemical conditions at the interfaces affecting the distribution of the chemicals among the phases. Microbial surfactants can promote the growth of bacteria on hydrocarbons by increasing the

surface area between oil and water and through emulsification and increasing pseudosolubility of hydrocarbons through partitioning into micelles (Banat et al., 2010).

2.1.5.1 Biosurfactants in food industry

In the food industry, the most useful property is the ability to form stable emulsions, which improves the texture and creaminess of dairy products. Biosurfactants are also used to retard staling, solubilise flavour oils and improve organoleptic properties in bakery and ice cream formulations and as fat stabilisers during cooking of fats. Although the addition of rhamnolipids has been suggested to improve dough characteristics of bakery products, the use as food ingredients of compounds derived from an opportunistic pathogen such as *P. aeruginosa* is not practically feasible. Instead, it has been suggested to use biosurfactants obtained from yeasts or *Lactobacilli*, which are generally recognised as safe and are already involved in several food-processing technologies (Nitschke & Costa, 2007).

Surfactins are used to maintain the texture, stability, and volume and also to help in the emulsification of fat in order to control the aggregation of fat globules. Recently, some lipopeptides isolated from bacterial group, *Enterobacteriaceae*, have been introduced into the food industry with their high emulsifying properties at enhanced viscosity and in acidic pH. Often various food preservatives are used by food manufacturers during processing to avoid rapid food spoilage. Among biopreservatives, several antimicrobial compounds have been accepted. These compounds effectively control food poisoning microbes, additionally Lipopeptides biosurfactants plays essential role in food safety where it used to inhabit the growth of pathogens (Table 2.2) (Meena & Kanwar, 2015).

Table 2.2: Lipopeptides in food industry (Meena & Kanwar, 2015)

Plant disease	Phytopathogen	Lipopeptide producing microorganism	Lipopeptide inhibiting the phytopathogen
Damping-off bean	*Pythium ultimum*	*Bacillus subtilis* M4	Iturin/Fengycin
Gray mold disease of apple	*Botrytis cinerea*	*Bacillus subtilis* M4	Fengycin
Arabidopsis root infection	*Pseudomonas syringae*	*Bacillus subtilis* 6051	Surfactin
Powdery mildew of cucurbits	*Podosphaera fusca*	*Bacillus subtilis*	Iturin/Fengycin
Fusarium head blight (FHB) in wheat, barley and ear rot in corn	*Gibberella zeae (anamorph of Fusarium graminearum)*	*Bacillus subtilis* JA; JA026	Fengycin
Sugar beet seed infection	*Rhizoctonia solani*	*Pseudomonas fluorescens* strain 96.578	Tensin
Root and foliar diseases of soybeans	*Xanthomonas axonopodis PV. Glycines*	*Bacillus amyloliquefaciens* KPS46	Surfactin
Sclerotinia stem rot disease	*Sclerotinia sclerotiorum*	*Bacillus amyloliquefaciens*	Surfactin/Fengycin
Rice blast	*Magnaporthe grisea*	*Chromobacterium sp* C61	Chromobactomycin

2.1.5.2 Biomedical applications

The high demand for new antimicrobial agents following increased resistance shown by pathogenic microorganisms against existing antimicrobial drugs has drawn attention to biosurfactants as antibacterial agents. Some biosurfactants have been reported to be suitable alternatives to synthetic medicines and antimicrobial agents and may therefore be used as effective and safe therapeutic agents (Table 2.3) (Meena & Kanwar, 2015). Among several categories of biosurfactants, lipopeptides are particularly interesting because of their high surface activities and antibiotic potential against an array of phytopathogens. Surfactin, produced by *B. subtilis*, is the best-known lipopeptide (Arima et al., 1968). Surfactins can act as antiviral agents, antibiotics, antitumor agents, immunomodulators or specific toxins inhibitors (Meena & Kanwar, 2015). Other antimicrobial lipopeptides include fengycin, iturin, bacillomycins and mycosubtilins produced by *B. subtilis* (Vater et al., 2002), Lichenysin, pumilacidin and polymyxin B (Landman et al., 2008), are other antimicrobial lipopeptides produced by *B. licheniformis*, *Bacillus pumilus* and *Bacillus polymyxa*, respectively. The production of antimicrobial lipopeptides by *Bacillus* probiotic products is one of the main mechanisms by which they inhibit the growth of pathogenic microorganisms in the gastrointestinal tract (Hong & Cutting, 2005).

Table 2.3: Lipopeptides in medical applications (Meena & Kanwar, 2015)

Microorganisms	Biosurfactant type	Activity/application
Bacillus subtilis MZ-7 and *B. amyloliquefaciens* ES-2	Surfactin	Antimicrobial and antifungal activities; inhibition of fibrin clot formation; hemolysis and formation of ion channels in lipid membranes; antitumor activity against Ehrlich's ascites carcinoma cells and antiviral activity against HIV-1; high concentration of Surfactin affects the aggregation of amyloid β-peptide into fibrils, a key pathological process associated with Alzheimer's disease; antifungal, antiviral, antitumor, insecticidal, and antimycoplasma activities.
Bacillus subtilis, *B. amyloliquefaciens* B128 and *B. amyloliquefaciens* PPCB004	Iturin	Antimicrobial activity and antifungal activities against profound mycosis. Effect on the morphology and membrane structure of yeast cells. Increase in the electrical conductance of bimolecular lipid membranes and acting as nontoxic and nonpyrogenic immunological adjuvant.
Bacillus subtilis	Iturin and Surfactin	Both bioagents show broad hypocholesterolemic activities and can act as antibiotics, antiviral, and antitumor agents; immuno-modulators; specific toxins; and enzyme inhibitors.

2.2 DNA sequencing technologies

The origin of sequencing method originally known as the Sanger chemistry which uses specific ladled nucleotides to detect DNA read through a DNA template throughout DNA synthesis. This method requires a specific primer to initiate the read at a particular position on the DNA template and record the distinguished labels for each nucleotide in the sequence. The Sanger method had reached the ability to read between 1,000 to 1,200 base pair (bp), though, this method is unable to read beyond two kilo base pair (Zhang et al., 2011). In order to sequence longer region within a DNA sequence, the Shotgun Sequencing approach was introduced during the Human Genome Project (Lander et al., 2001; Venter et al., 2001). Shotgun sequencing method based sequencing DNA sequence enzymatically or mechanically broken down into smaller fragments and cloned into sequencing vectors in which cloned DNA fragments can be sequenced individually. The complete sequence of a long DNA fragment can be eventually generated by these methods by alignment and reassembly of sequence fragments based on partial sequence overlaps. Shotgun sequencing was a momentous advantage from Human Genome Project, in addition it made sequencing the entire human genome possible (Zhang et al., 2011). The core concept of massive parallel sequencing which is used in NGS technology is adapted from shotgun sequencing approach (Venter et al., 2003). An important advantage of sequence data is its quality, robustness and low noise. It should be noted that a successful NGS project requires expertise both at the wet lab as well as the bioinformatics side in order to warrant high quality data and data interpretation (Buermans & Den Dunnen, 2014).

Three widely used platforms for massively parallel DNA sequencing read production were Roche/454 FLX (30), the Illumina/ Solexa Genome Analyzer, and the Applied Biosystems SOLiDTM System. Very recently, another two massively parallel systems were being used. The Helicos Heliscope^{TM} and Pacific Biosciences SMRT

instruments. The Helicos system only recently became commercially available (Mardis, 2008). Table 2.4 shows the comparison among various NGS sequencing plateforms .

2.2.1 NGS technology limitations

In the last few years, a series of high throughput sequencing platforms have been commercially introduced based on different sequencing chemistries and detection techniques. Three platforms for massively parallel DNA sequencing read production are in reasonably widespread use at present: the Roche/454 FLX, the Illumina/ Solexa Genome Analyzer and the Applied Biosystems SOLiDTM System. Recently, another two massively parallel systems were announced: the Helicos Heliscope^{TM} and Pacific Biosciences SMRT (Mardis, 2008).

Since NGS introduction in 2005, high throughput NGS technologies have faced several challenges. The first has been the improvement in sequencing output, in terms of read length and accuracy. The second challenge has been the total output of the sequencing experiment in relation to the cost and the labour expenses. The third challenge is related to the amplification step prior to sequencing. This final challenge includes different sources of PCR bias, formation of chimeric sequences and secondary structure-related issues (Shokralla et al., 2012). The main disadvantage of Illumina systems is the relative short-read length because of optical signal decay and dephasing. This limits the application of these technologies in situations where no reference sequence is available to align, assign and annotate the short sequences generated (Zhou et al., 2010).

Table 2.4: Comparison of NGS platforms (Mardis, 2017)

Platform	Read length	Applications
454/Roche	400 bp (single end)	Bacterial and viral genomes, multiplex-PCR products, validation of point mutations, targeted somatic-mutation detection.
Illumina	150–300 bp (paired end)	Complex genomes (human, mouse and plants) and genome-wide NGS applications, RNA-seq, hybrid capture or multiplex-PCR products, somatic-mutation detection, forensics, noninvasive prenatal testing.
ABI SOLiD	75 bp (single end) or 50 bp (paired end)	Complex genomes (human, mouse, plants) and genome-wide NGS applications, RNA-seq, hybrid capture or multiplex-PCR products, somatic-mutation detection.
Pacific Biosciences	Up to 40 kb (single end or circular consensus)	Complex genomes (human, mouse and plants), microbiology and infectious-disease genomes, transcript-fusion detection, methylation detection.
Ion Torrent	200–400 bp (single end)	Multiplex-PCR products, microbiology and infectious diseases, somatic-mutation detection, validation of point mutations.
Oxford Nanopore	Variable: depends on library preparation (1D or 2D reads)	Pathogen surveillance, targeted mutation detection, metagenomics, bacterial and viral genomes.
Qiagen GeneReader	107 bp (single end)	Targeted mutation detection, liquid biopsy in cancer.

2.2.2 Illumina sequencing platform

The Illumina Genome Analyzer currently is the most widely used sequencing platform. The Illumina system uses a sequencing by synthesis approach (Mardis, 2008), in which all four nucleotides are added simultaneously into oligo-primed cluster fragments in flow cell channels along with DNA polymerase. Bridge amplification extends cluster strands with all four fluorescently labelled nucleotides for sequencing Figure 2.1. Illumina Genome Analyzer is widely recognized as the most adaptable and easiest to use sequencing platform (Zhang et al., 2011). Higher data quality and proper read lengths have made it the system of choice for many genome sequencing projects. To date, the majority of published NGS scientific papers have described methods using the short sequence data produced with the Illumina Genome Analyzer. At present, the new Illumina HiSeq 2000 is capable of producing single reads of 2 X 100 basepairs (pair-end reads), and generates about 200 giga basepair (Gbp) of short sequences per run. The raw base accuracy is greater than 99.5% (Zhang et al., 2011).

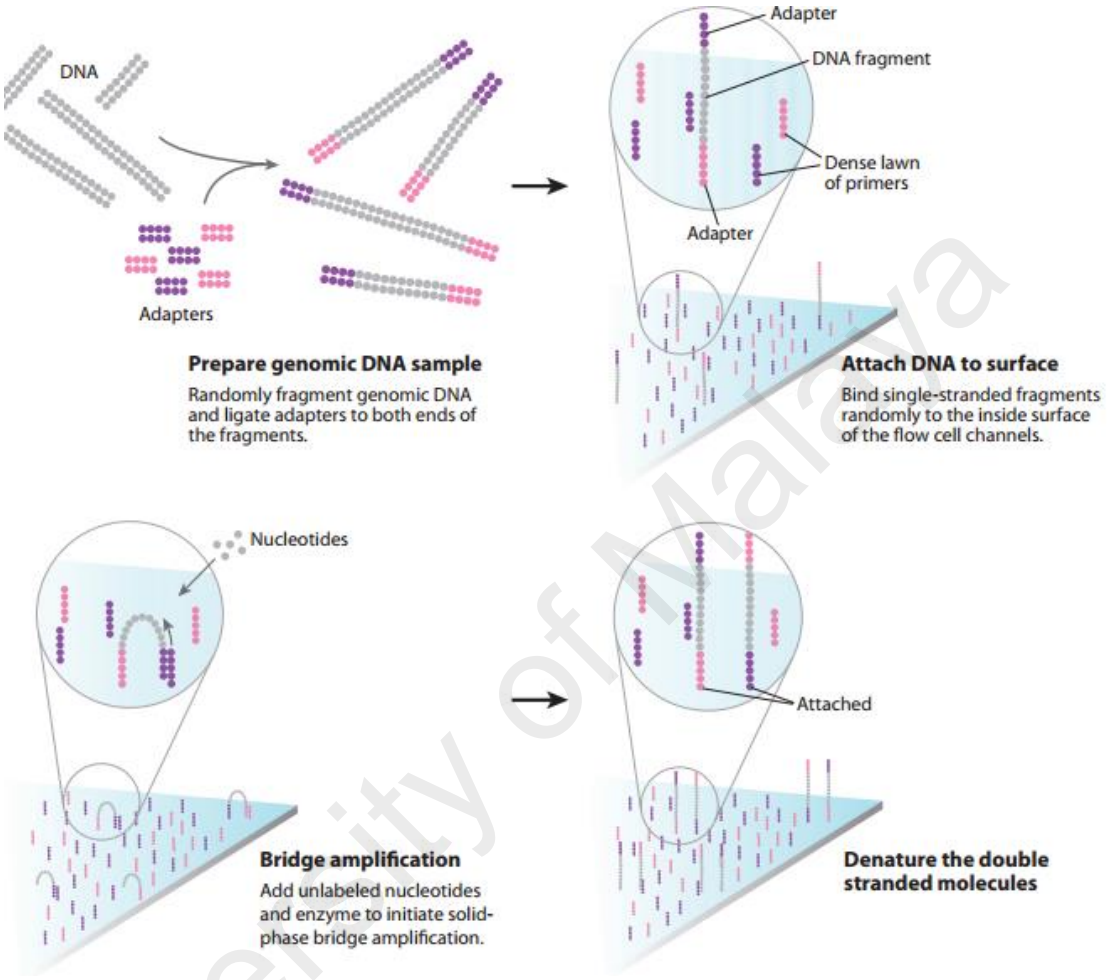

Figure 2.1: The Illumina sequencing-by-synthesis approach. Cluster strands created by bridge amplification are primed and all four fluorescently labeled, 3 -OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation step, the unused nucleotides and DNA polymerase molecules are washed away, a scan buffer is added to the flow cell, and the optics system scans each lane of the flow cell by imaging units called tiles. Once imaging is completed, chemicals that effect cleavage of the fluorescent labels and the 3 -OH blocking groups are added to the flow cell, which prepares the cluster strands for another round of fluorescent nucleotide incorporation (Mardis, 2008).

2.2.3 NGS applications in genomics

2.2.3.1 Whole Genome Sequencing (WGS)

Genomics is a relatively young field arguably, started in 1976 through a published RNA genome of bacteriophage MS2 (Fiers et al., 1976). The first time this term was used in 1986, and it was defined by Thomas Roderick as "encompassed the structure and function of genes, and comparative genomics elucidated the hereditary relationships and evolution within and between different species" (Kuska, 1998). Since the introduction of NGS, the meaning of "genomics" is narrowed more towards mapping the structure and organization of genomes and differentiating between *de novo* sequences, resequenced genomes, exonic or targeted sequences, and metagenomic sequences (Hocquette et al., 2009; Kulski, 2015) Genome Wide Association Studies (GWAS) were the most common applied approach to genomics. Though numerous primary GWAS studies reported potentially promising results, the majority of GWAS studies were disappointing because of insufficient sample size, limitation of arrays for certain genetic variations, and/or heterogeneity in phenotype (Daly, 2010). These obstacles overcome by new genomics NGS technology (Zhang et al., 2011). The most comprehensive approached applied in genomic studies is Whole Genome Sequencing (WGS). The rapid sequencing cost reduction and the capability of WGS to generate large amount of data makes it effective tool for genomic research. WGS generally associated with sequencing human genome, however due to the flexibility of the technology, makes it also valuable for sequencing any species such as disease related microbial genomes and plant genomes. WGS has massive impact in bacterial and virus genomes (Ladner et al., 2014). It has become progressively easier, faster, and cheaper because of technological improvements and the availability of hundreds of sequenced genomes that can be used as references for annotation (Kulski, 2015).

2.2.3.2 Targeted sequencing

In targeted sequencing, a specific subset of genes or regions within a genome are isolated and sequenced. This application allows researchers to analyse data on particular area of interest as it provides sequencing at high level of coverage. The usual WGS study archives at coverage level of 30x or 50x, while targeted sequencing study allow the targeted region of study at 500x to 1000x or higher coverage. With this high level of coverage, researchers are able to identify rare variants which are too expensive to identify using WGS. The mostly used and known targeted sequencing application is exome sequencing. In exome sequencing, the protein-coding regions in a genome are selectively captured and sequenced. Using this application, a wide range of variant can be identified in many studies such as population genetics, cancer studies and genetic disease studies. Exome sequencing has been extensively used for clinical studies in the recent years, and is giving rise to promising novel diagnostic tools that have the potential to transform medical healthcare in the near future (Dijk et al., 2014).

2.2.3.3 ChIP-Seq sequencing

The Combination of Chromatin Immunoprecipitation (ChIP) assay followed by sequencing (ChIP-Seq) is an affective technique to identify genome-wide profiling of DNA-binding sites, histone modifications or nucleosomes and other proteins. ChIP-seq application enables genome-wide mapping of protein binding and epigenetic marks which is essential for understanding transcriptional regulation. An accurate map of binding sites for transcription factors, primary transcriptional machinery and other DNA-binding proteins is vital for interpreting the gene regulatory networks that underlie various biological processes (Park, 2009). The ChIP-seq approach is dependent on the cross-linking of proteins to specific DNA elements, followed by antibody enrichment of the protein–DNA complexes, and high-throughput sequencing of the recovered DNA fragments. In principle, ChIP-seq using an antibody specific for an

enhancer-binding protein could provide a conservation-independent approach for the identification of candidate enhancer sequences (Visel et al., 2009). The application of NGS to ChIP has revealed insights into gene regulation events that play a major role in various diseases and biological pathways, such as development and cancer progression. ChIP-Seq enables thorough examination of the interactions between proteins and nucleic acids on a genome-wide scale.

2.2.4 NGS data analysis using bioinformatics

Bioinformatics plays a major role in NGS technology in overcoming the rising challenges of storage, analysis, and interpretation of NGS data (Land et al., 2015). When using the NGS platforms, there are at least three steps of nucleotide sequence analysis that need to be considered. The first step is generation of sequence reads using the software integrated within the sequencing instruments that convert the raw signals into base calling with short reads of nucleotide sequences and associated with quality scores. The second is the alignment and assembly of raw reads, contigs, scaffolds and variant detection. The third is annotation, data integration, and visualization of the assembled sequence (Kulski, 2015). Figure 2.2 illustrate the stages of NGS data analysis using bioinformatics.

The raw sequencing signals generated by the manufacturer's sequencing instrument or system are converted into nucleotide bases of short read data which known as base calling with base quality scoring using the system's FASTQ format or the native raw data file formats such as Illumina, SFF, HDF5, CG, or SOLID. Storage of raw signal and sequencing data as short read archives in the FASTQ format or native raw data file formats is a problem in regard to computing resources for many research sequencing laboratories and commercial service providers. Thus, the conversion of

FASTQ files to the more compact Sequence Alignment Map (SAM) format and its compressed Binary Alignment Map (BAM) format is recommended because it is easier to read and process for later bioinformatics analysis (Kulski, 2015). The right way of storing the original raw sequences is essential for the NGS data analysis, because it is the main source of the initial sequencing errors which are either left or filtered out of the final assembled sequence. The quality assessment is needed to remove sequence errors and reads with low Phred score. Sequences such as adapters, primers, vectors, and tails that were introduced experimentally during the preparation of the sequencing libraries also should be remove (Horner et al., 2010; Kulski, 2015).

Gene ontology is one of the bioinformatics analysis initiative that provides various information about genome, it defined terms representing gene product properties and pathways covering biological domains such as cellular components, molecular function, and biological processes with their various subcategories, and it provides functional annotation tools to find functions for large gene lists. It sits somewhere between the third stage in NGS data analysis, which known as annotation, and the fourth stage of is known as structural analysis. The first major Gene Ontology (GO) project was founded in 1998 to address a need for standard filtered descriptions of gene products across different databases. GO is a collaborative effort that started between three model organism databases, FlyBase (*Drosophila*), the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD) but now incorporates many databases for plant, animal, and microbial genomes. The GO Contributors page lists all member organizations. Some other ontology providers among many are the Open Biological and Biomedical Ontologies (OBBO), Reactome, DAVID, and the KEGG Pathway database (Kulski, 2015).

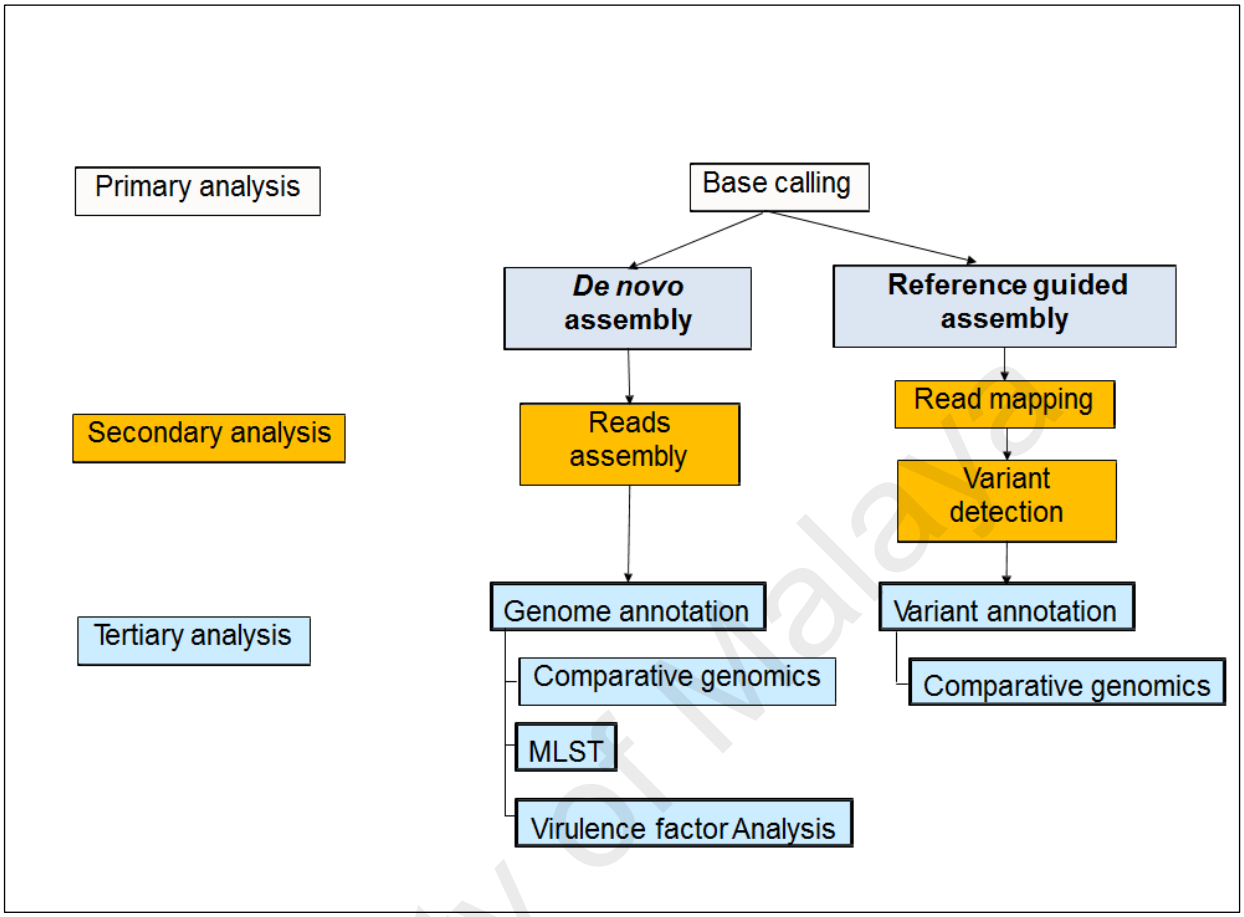

Figure 2.2: NGS data analysis using bioinformatics.

2.3 Bacterial identification using traditional and modern methods

Using 16S rDNA sequencing, has facilitates the discovery of novel genera and species (Woo et al., 2008). 16S rRNA gene sequencing techniques provide genus and species identification for isolates that do not fit any recognized biochemical profiles, for strains generating only a low likelihood or acceptable identification according to commercial systems, or for taxa that are rarely associated with human infectious diseases. The cumulative results from a limited number of studies to date suggest that 16S rRNA gene sequencing provides genus identification in most cases (90%) but less so with regard to species (65 to 83%). Although 16S rRNA gene sequencing is highly useful in regards to bacterial classification, it has low phylogenetic power at the species level and poor discriminatory power for some genera, and DNA relatedness studies are necessary to provide absolute resolution to these taxonomic problems. The genus *Bacillus* is a good example of this (Janda & Abbott, 2007). In general, the 16S rRNA gene is universal in bacteria, and so relationships can be measured among all bacteria (Clarridge, 2004).

Traditionally, identification of bacteria was performed using phenotypic tests, including Gram smear and biochemical tests, taking into account culture requirements and growth characteristics. However, these methods of bacterial identification have major limitations. First, organisms with biochemical characteristics that do not fit into the patterns of any known genus and species are occasionally encountered. Second, they cannot be used for uncultivable organisms. Third, identification of some particular groups of bacteria, such as anaerobes and mycobacteria, would require additional equipment and expertise, which are not available in most microbiology laboratories.

CHAPTER 3: MATERIALS AND METHODS

3.1 Experimental and data analysis design

This research contains integrated approaches of biochemical, bioinformatics and genomics analysis. Figure 3.1 demonstrates the overall approached used in this study.

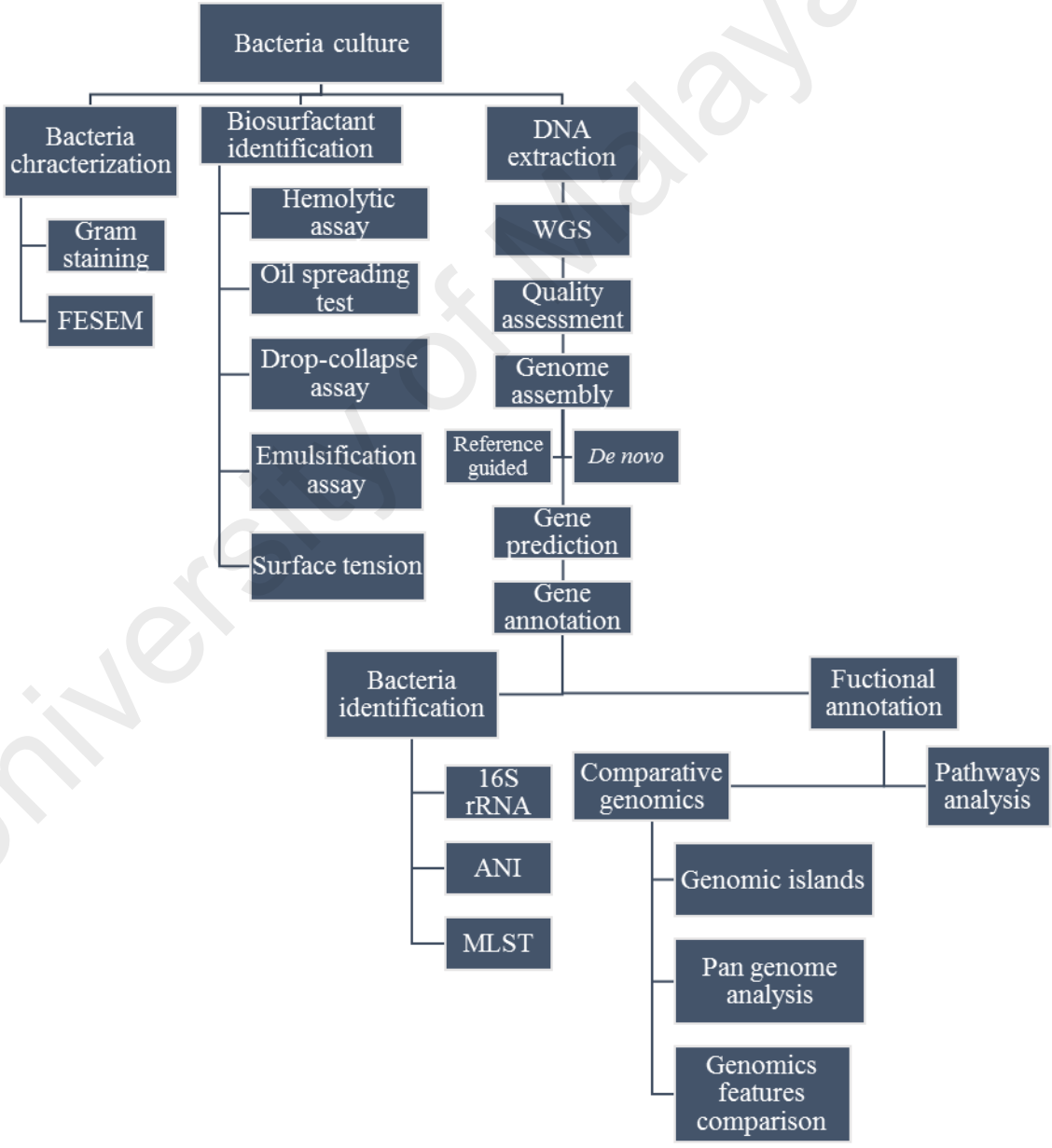

Figure 3.1: Overall approached used in this study.

3.2 Sample source and preparation

Bacillus subtilis UMX-103 (originally known as TS 8003) was isolated from a hydrocarbon contaminated site in Terengganu Malaysia (Abdelhafiz et al., 2017). The bacterium was cultured in Tryptone Soya Agar (TSA) (Merck KGaA, Germany) and incubated overnight at 30ºC. Optimal colony was selected from the cultured bacteria and inoculated in 50 ml of Tryptone Soya Broth (TSB) (Merck KGaA, Germany) using 200 ml conical flask. The broth was incubated overnight at 30º C in orbital shaker at 121 rpm.

3.3 Bacterial identification

UMX-103 was initially identified using Gram staining. Field Emission Scanning Electron Microscope **(**FESEM) was used to determine the bacteria morphology. The bacteria cells were observed under 10 000x and 20 000x magnifications.

3.3.1 Gram staining

Gram staining was conducted using Gram staining kit (Gainland Chemical Co, UK). The prepared smear slides were firstly flooded with crystal violet oxalate solution for one minute. Then the dye was washed using distilled water for five seconds. Followed by flooding the slide with iodine solution and kept for one minute. The iodine solution was washed from the slide using distilled water. Next, the slide was flooded with decolorizer for five seconds. Then the decolorizer was rinsed using distilled water. The slide was flooded with safranin and allowed to remain for one minute. Then the slide was washed and dried. The slides were observed under light microscope for identification of the Gram stain of the bacteria.

3.3.2 Field Emission Scanning Electron Microscope (FESEM)

Bacteria cells were observed using FESEM (Quanta 450 FEG, USA) to determine the morphology of the bacteria. The bacteria cells were observed under magnification of 10 000x and 20 000x.

3.3.3 16S rRNA gene

16S ribosomal DNA was obtained from the whole genome sequence and aligned with other 16S rRNA genes from different *Bacillus* strains. The 16S genes were extracted from each reference strains using RNAmmer (Lagesen et al., 2007). Molecular Evolutionary Genetics Analysis (MEGA) version 7 (Kumar et al., 2016) was used to align and construct the distanced phylogenetic tree. The 16S genes were aligned with ClustalW (Li & Kuo-Bin, 2003) then the phylogenetic tree and distance were constructed using Neigbour-joing method. 16S rRNA gene is widely used in identification of bacterial isolates and discovery of novel species (Janda & Abbott, 2007; Woo et al., 2008).

3.3.4 Average Nucleotide Identity (ANI)

The Average Nucleotide Identity of UMX-103 was determined using GGDC 2.1 server (Meier et al., 2013). ANI uses DNA-DNA Hybridization (DDH) which is a method widely used to measure overall similarity between two genomes (Kim et al., 2014).

3.3.5 Multilocus Sequence Typing (MLST)

MLST is a method that is widely used to identify bacteria. This method is based on PCR amplification and sequencing of essential seven housekeeping genes within species which are spread around the bacteria genome. The seven housekeeping genes of *Bacillus subtilis* UMX-103 and the other genome references used in this research were predicted using (MultiLocus Sequence Typing MLST) server v 1.8 (Larsen et al., 2012).

3.4 Screening and detection of biosurfactant production

The ability of biosurfactant production by UMX-103 was tested using 5 different methods: hemolytic assay (Yonebayashi et al., 2000), oil spreading test (Youssef et al., 2004), drop-collapse assay (Shoeb et al., 2015), emulsification assay (Cai et al., 2015) and surface tension measurements (Pereira et al., 2013).

3.4.1 Hemolytic activity test

The isolated strain was streaked onto blood agar plates and incubated at 30^{o} C for 24 hours. The plate was visually inspected for clear zone formed around the colonies, which is indicative of biosurfactant production. Hemolysis activity on blood agar plates has been widely used as a method to screen surfactant producing bacteria (Banat, 1993; Morán et al., 2002; Mulligan et al., 1984; Yonebayashi et al., 2000).

3.4.2 Oil spreading and drop-collapse assays

The oil spreading test is used to determine the clear zone diameter which is a result of dropping a biosurfactent or surfactant solution on an oil-water interface. A volume of 15µl of 10W-40 Shell® was added on the top of 40ml of distilled water in a petri dish (150 mm in diameter) to form a thick layer of oil on the surface. Then 15µl of culture supernatant were added to the central of the oil layer (Morikawa et al., 1993;

Morikawa et al., 2000; Youssef et al., 2004). Water was used as negative control (Shoeb et al., 2015). Diameter of the clear zone formed on the top surface of oil was observed and measured after 30 seconds (Morikawa et al., 2000). Triton® X-100 (Merck KGaA, USA) as the positive control.

Qualitative test of the biosurfactant produced by UMX-103 was conducted on polystyrene lid of a 96-microwell (12.7×8.5) plate. A volume of 2µl of 10-40 Shell® was added to each well and the lid was equilibrated for 1 hour at the room temperature. Then 5µl of the culture supernatant was placed on the oil surface. Water was used as negative control (Bodour et al., 2003; Shoeb et al., 2015). The shape of the drop on the oil surface was observed after 1 min.

3.4.3 Emulsification assay

Emulsification assay was performed to check the ability of biosurfactant produced by UMX-103 to emulsify the hydrocarbon. Initially, 5ml of 50mM Tris buffer (8.0 pH) was added to 30-ml screw-caped test tube. Then, 5ml of 10-40W Shell® was added to the above solution and vortex-shaken for 2 min and let to stand for 24 h. The absorbance of aqueous phase was measured by spectrophotometer (Spectroquant® Pharo100, USA) at wavelength of 400 nm. Distilled water was used as negative control, while Triton-X as the positive control (Shoeb et al., 2015). The emulsification activity was calculated (Cai et al., 2015) as stated below:

EA= Sample Emulsification Abs/Optimum Emulsification Abs x 100%.

EA= Emulsification Absorbance

Abs = Absorbance

3.4.4 Surface tension measurement

Culture sample was centrifuged at 3000 rpm for 25 min to harvest the bacteria cells. The surface tension of the culture supernatant was determined by the Du Nouy ring method using interfacial tensiometer (Force Tensiometer, Sigma700, Biolin Scientific) at room temperature. The measurements of the surface tension were repeated three times and an average value was obtained (Cai et al., 2015; Pereira et al., 2013; Vaz et al., 2012).

3.5 Whole genome sequencing and data analysis

3.5.1 Whole genome sequencing using Illumina HiSeq 2000 platform

The whole genome sequence of *Bacillus subtilis* UMX-103 was obtained from Illumina HiSeq 2000 sequencing platform (Illumina, USA). The DNA was extracted using phenol-chloroform method and the quality and quantity of the DNA was measured using QIAxpert (QIAGEN, Germany). The sample was run on 1.2% (w/v) agarose gel to determine the integrity of genomic DNA. Fragmentation of the DNA was performed using Covaris S220 (Covaris Inc, USA). Ligation to NEBNext adapters conducted using NEBNext Ultra, while the PCR-enrichment used DNA Library Prep Kit (NEB, USA). The final library was quantified using KAPA kit (KAPA Biosystem, USA). Library size was confirmed using Agilent Bioanalyzer High Sensitive DNA Chip (Agilent, USA). The prepared library was sequenced using an Illumina flow cell, consisting of 2x100 cycles.

3.5.2 *De novo* assembly by Velvet and mapping the reads to reference genome by BWA

Quality control assessment was performed using Trimmomatic 0.35 (Bolger et al., 2014). The generated dataset was assembled using Velvet 1.2.10 (Zerbino & Birney,

2008) which is a *de novo* assembly software that use de Bruijn graph algorithm. SSPACE-Standard v3.0 (Boetzer et al., 2011) was used for scaffolding the generated contigs from Velvet assembler. GapFiller v1.10 (Boetzer & Pirovano, 2012) was used to close the gaps and replace unknown nucleotide with known nucleotides. The scaffolds were sorted along with the reference genome (*Bacillus subtilis* strain 168; accession number NC_000964.3) using Mauve 2.3.1 (Darling et al., 2004). Burrows-Wheeler Aligner (BWA) (Li, 2013) was used for mapping the reads to the reference genome.

3.5.3 Gene prediction and annotation

Gene prediction for protein-coding genes was conducted using Prodigal (Hyatt et al., 2010). The tRNA and rRNA screenings were performed using tRNAscan-SE v1.3.1 (Lowe & Eddy, 1997) and RNAmmer v1.2 (Lagesen et al., 2007), respectively. Gene annotation was conducted using Prokka v1.11 (Seemann, 2014). The functional annotation was performed using EggNOG-mapper 4.5.1 database (Huerta-Cepas et al., 2017). The annotated genes were submitted to IslandViewer3 (Dhillon et al., 2015) to identify the genomic islands in the genome. IslandViewer3 contains three methods which are integrated to identify genomic islands; SIGI-HMM, IslandPath-DIMOB and Integrated method (Langille & Brinkman, 2009). SIGI-HMM is a sequence composition GI prediction method that uses Hidden Markov Model (HMM) and measure codon usage to identify possible GIs (Waack et al., 2006), where IslandPath-DIMOB is a method designed to identify prokaryotic genomic islands by detecting abnormal sequence composition or the presence of genes that functionally related to gene horizontal transfer (Hsiao et al., 2003). The third method is the combination of SIGI-HMM and IslandPath-DIMOB. All of the three methods shows high accuracy (Langille, Hsiao et al., 2008). Pangenome analysis and comparison between all the selected reference genomes were conducted using Roary version 3.6.1 (Larsen et al., 2012) and

BPGA 1.3 (Chaudhari et al., 2016). The comparative genomic analysis of UMX-103 against the most close related genomes which are publicly available (Table 3.1). This analysis can highlight the unique features that are present in this strain.

CHAPTER 4: BACTERIAL IDENTIFICATION AND BIOSURFACTANTS SCREENING

4.1 Introduction

Bacillus species are key workhorses for many biotechnologies, industry and applications, as their products are in the list of GRAS (Generally Regarded As Safe), the list which is provided by United States Food and Drug Administration (USFDA). They are able to produce a various products including biosurfactants, extracellular enzymes, biopesticides and biopolymers (Joshi et al., 2012). Surfactants are classified in two major classes, which are biosurfactants and synthetic surfactants. Biosurfactants usually produced through biological processes which are extracted extracellularly by many microorganisms such as bacteria and fungi (Gudina et al., 2013), while chemical surfactants are produced by chemical reaction using petroleum feedstocks (Vaz et al., 2012). Biosurfactants are potential alternative neutral surfactants to the chemical synthetic surfactants due to their amphiphilic structure and properties. They have surface active features where it can lower surface tension and interfacial tension which can be used in oil and gas industry especial in Enhanced Oil Recovery application. Also it can be employed in petrochemicals and pharmaceutical industries (Gudina et al., 2013). Thus most of the used methods for screening biosurfactant production in strains are based on surface activity of culture supernatant (Walter et al., 2010).

The objective of this chapter is to identify and characterize the functional features of UMX-103 to produce biosurfactants as outlined under objective1 (see page 3).

4.2 Bacteria and sample preparation

Figure 4.1 shows the growth of UMX-103 in nutrient agar TSA after incubating the culture overnight, where the pigmentation of colony is grey and the consistency is opaque. Also, the formation of the colonies is irregular.

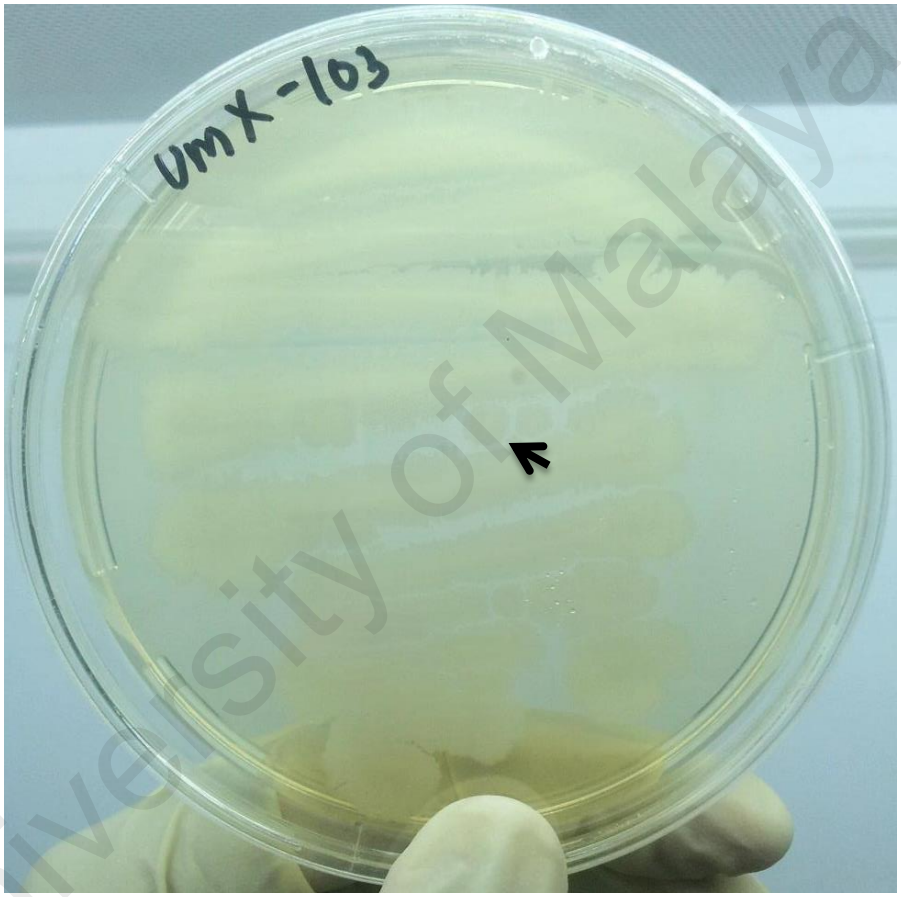

Figure 4.1: UMX-103 culture in TSA plate.

4.3 Bacterial identification

4.3.1 Gram staining and FESEM

The Gram staining results showed that UMX-103 is Gram positive as all the cells colour are purple (Figure 4.2)**.** UMX-103 was visualized under the FESEM, showing the morphology of the colony which is rod and the size of 1.954 µm in length and 540.9 nm in width (Figure 4.3 and 4.4).

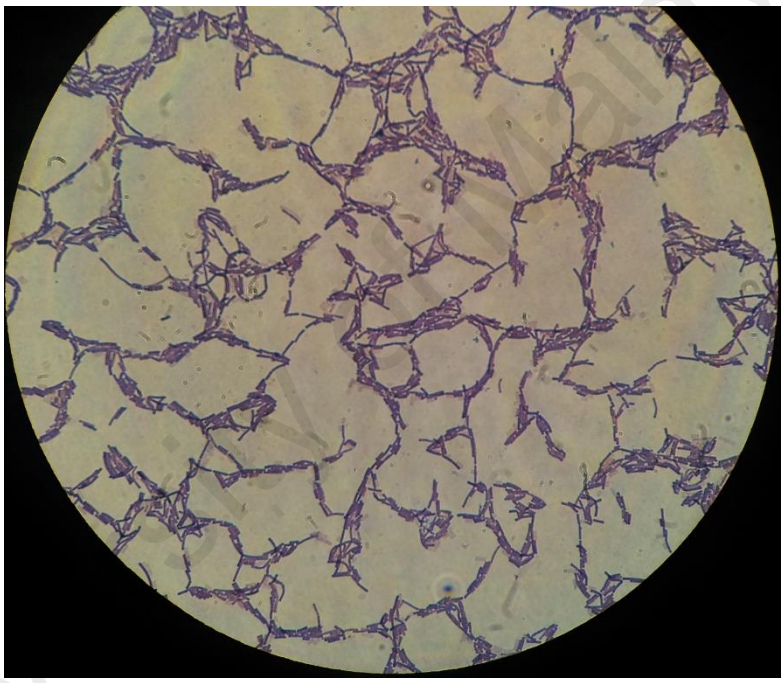

Figure 4.2: Gram staining result of *Bacillus subtilis* UMX-103.

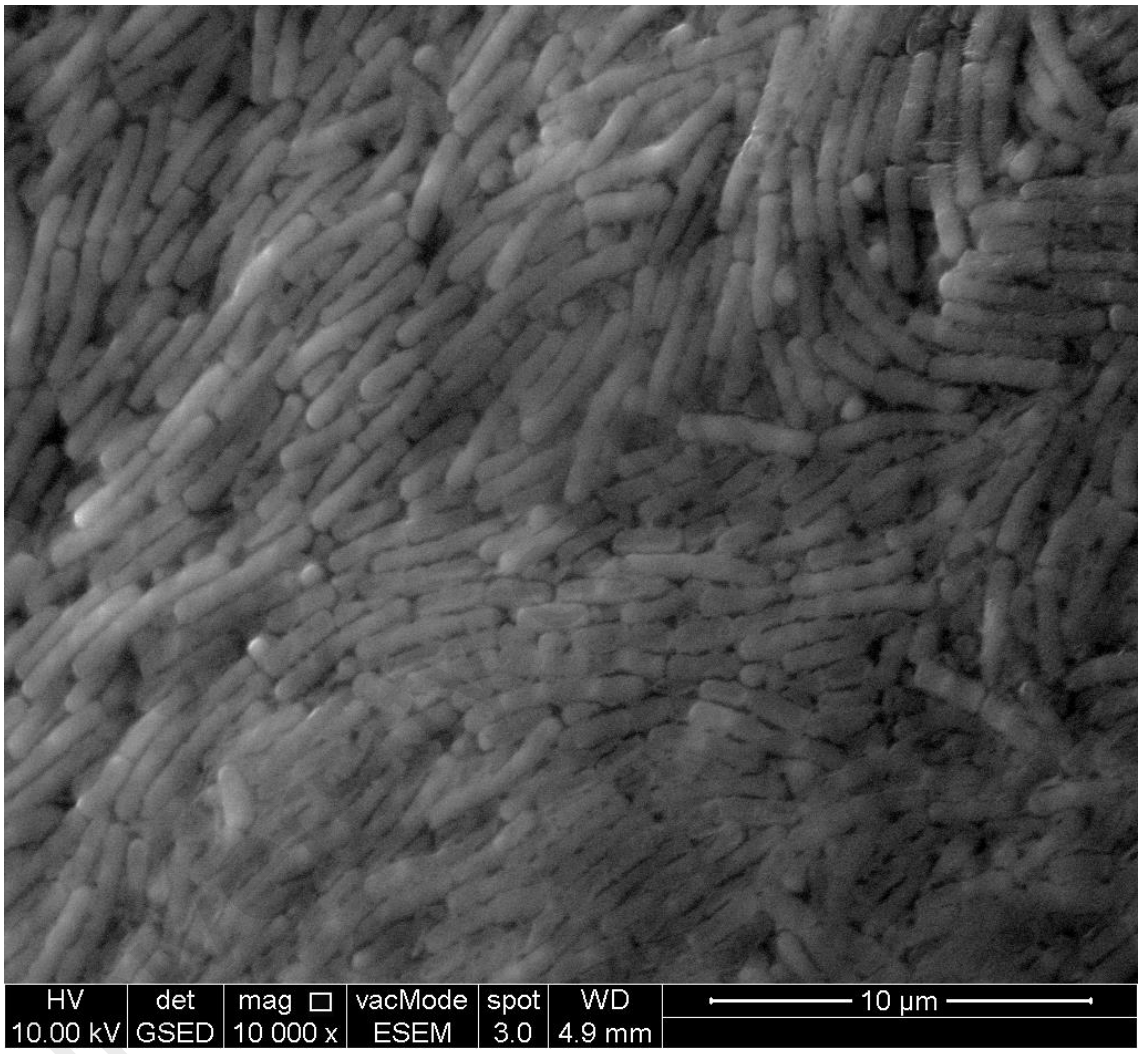

Figure 4.3: The bacterial morphology of UMX-103 under 10 000x magnification. The figure shows UMX-103 cells after overnight incubation at (10 000x magnification).

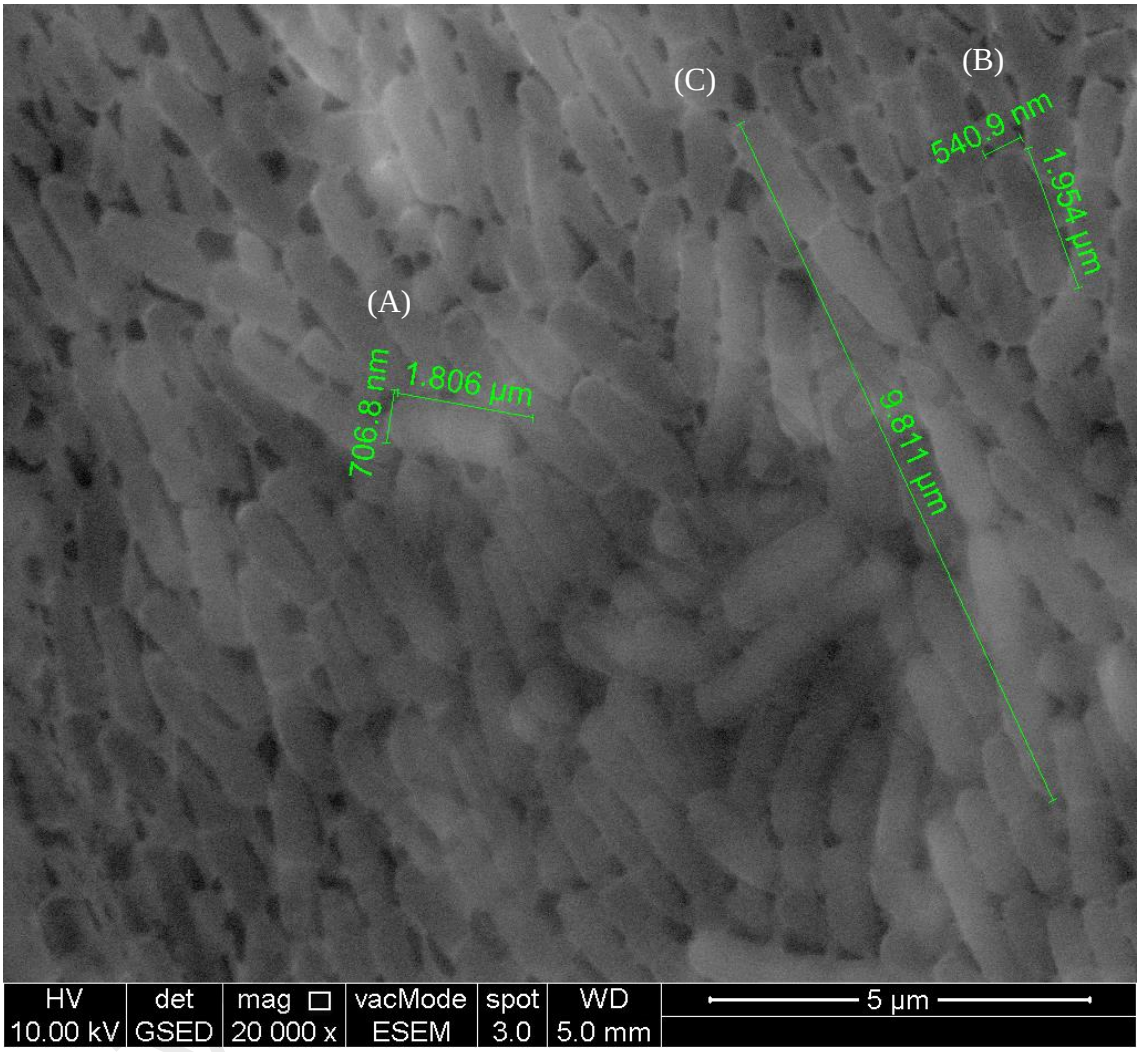

Figure 4.4: The bacterial morphology of UMX-103 under 10 000x magnification. (A) Cell size of 1.806 µm in length and 706.8 nm in width. (B) Cell size of 1.954 µm in length and 540.9 nm in width. (C) Rod shape cells and in chain formation with chain length 9.811 µm at (20 000x magnification).

4.4 Biosurfactant activity

The ability of UMX-103 biosurfactant production was tested using five different methods; i) Hemolytic assay, ii) Oil spreading test, iii) Drop-collapse assay, iv) Emulsification assay and v) Surface tension measurements. The results are shown in Table 4.1.

4.4.1 Hemolytic activity

Hemolysis assay on blood agar plates has been widely used as a method to screen surfactants producing bacteria (Banat, 1993; Morán et al., 2002; Mulligan et al., 1984; Yonebayashi et al., 2000). Thus method was also used in surfactin screening (Morán et al., 2002). UMX-103 was streaked onto blood agar plates and incubated at 30^{o} C for 24 hours. The plate was visually inspected for clear zone formation around the colonies (Figure 4.5), which is an indicative of biosurfactant production. UMX-103 demonstrated beta lysis as it produced a clear zone around the colony which determines the biosurfactant production by the strain.

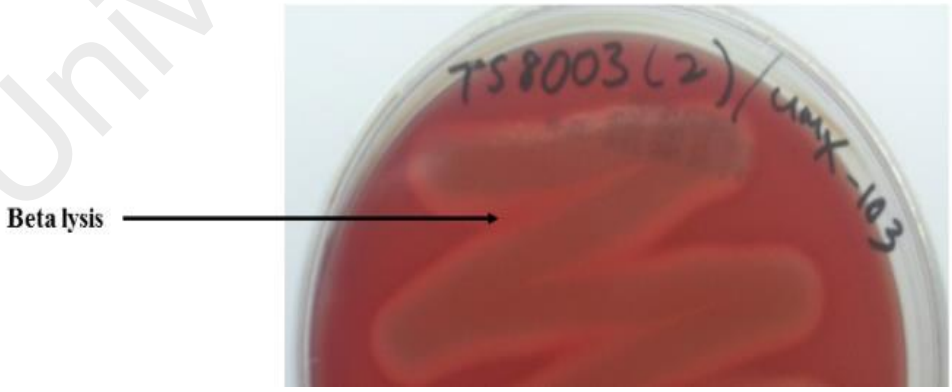

Figure 4.5. Hemolytic activity of UMX-103. Shows beta lysis the formation of clear zone around the colonies.

Table 4.1: Biosurfactant producing capability tests conducted on UMX103; hemolysis assay, oil-spreading, drop-collapse, emulsification assay and surface tension measurement

Bacterial sample/ Controls	**Test type**				
	Hemolytic activity	**Oil-spreading**	**Drop-collapse**	**Emulsification Activity**	**Surface tension (mN/m)**
UMX-103	+++	+++	+++	++	26.4 ± 0.02
Triton-X	X	++++	++++	++++	34.3 ± 0.003
Hexane	X	++++	++++	X	18.1 ± 0.06
TSB	X	X	X	+	52.0 ± 0.31
Deionized water	X	X	X	X	70.3 ± 0.91
Distilled water	X	-	-	+	X

Symbol means: (-) = no result; (+) = week result; (++) = average result; (+++) = good result; (++++) = high result, (x) = not applicable. (TSB) = Tryptone Soya Broth.

4.4.2 Oil spreading assay and drop-collapse test

Oil spreading assay is based on the formation of a clear zone and a displacement area, in the presence of biosurfactant in the culture supernatant. The diameter of this clear zone on the oil surface correlates to the amount of biosurfactant produced. Supernatant of UMX-103 culture formed a clear zone and oil displacement region about 2 cm for as indication of biosurfactant production.

The drop-collapse test depends on the destabilization of liquid droplets by the biosurfactants produced by the bacterial isolate. The stability of drops is dependent on biosurfactant concentration and correlates with the surface and interfacial tension (Sari et al., 2014). Distilled water was used as a negative control and there is no droplet collapsing was observed. Biosurfactant produced by UMX-103 was tested positive, where the droplet was collapsed.

4.4.3 Emulsification test and surface tension activity

The emulsification test was used to evaluate the emulsification ability of UMX-103. A positive activity of the strain was observed where it emulsifies the oil surface (Figure 4.6). In this study, Triton-X was used as positive control due to it emulsification ability and it have been widely used as positive control (Shoeb et al., 2015).

The measurement of surface tension using Du Nouy ring method is based on measurement of the force required to detach the ring from the culture supernatant surface. The detachment force is directly proportional to the interfacial tension. Test results showed that UMX-103 has a higher ability to reduce the surface tension which is up to 26.4 ± 0.02 mN/m compared to Triton X (34.3 ± 0.003 mN/m). Hexane showed the lowest value of surface tension measurement (18.1 ± 0.06 mN/m). A summary of surface tension measurement is presented in Table 4.2.

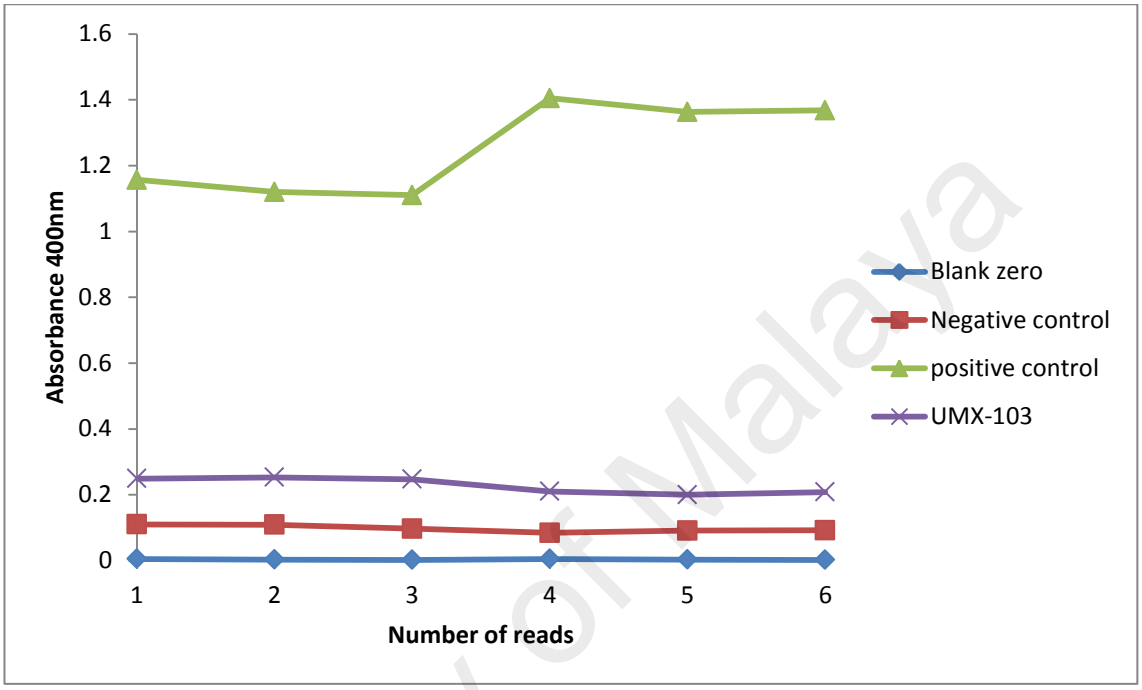

Figure 4.6: Emulsification assay result. Distilled water is used as negative control (red colour line) while positive control is Triton X (green colour line). UMX-103 (purple colour line) and blank reads (blue colour line).

Table 4.2: Surface tension measurements

Samples	Classification	Time [s]	ST [mN/m]	F/L [mN/m]	Force [mN]	Position [mm]	Speed [mm/min]	T [C]
Deionized water	Aqueous	93.4 ± 62.7	68.5 ± 0.1	73.1 ± 0.05	8.8 ± 0.006	3.4 ± 0.02	5.0 ± 0.0	27.2 ± 0.04
UMX-103	Biosurfactant	79.3 ± 53.5	26.4 ± 0.02	30.2 ± 0.02	3.6 ± 0.002	1.7 ± 0.009	5.0 ± 0.0	27.2 ± 0.03
Triton-X	Emulsifier	83.5 ± 56.3	34.3 ± 0.003	39.1 ± 0.003	4.7 ± 0.0004	2.4 ± 0.01	5.0 ± 0.0	27.6 ± 0.03
Hexane	Dispersant	83.5 ± 56.2	18.1 ± 0.06	20.6 ± 0.06	2.5 ± 0.007	2.2 ± 0.04	5.0 ± 0.0	27.9 ± 0.02
TSB	Culture media	86.8 ± 58.4	52.0 ± 0.31	56.6 ± 0.31	6.8 ± 0.038	2.7 ± 0.015	5.0 ± 0.0	26.8 ± 0.03

(s) = per second; (ST) = Surface Tension; (F/L) = force per liter; (T[C]) =Temperature in Celsius

4.5 Discussion

Bacillus species are Gram positive rod shape bacteria which often presented in pairs or chains with rounded shape ends with a single endospore and divided into Gram positive and Gram variable (England, 2015). The Gram staining result along with the FESEM result showed that UMX-103 belongs to the genus *Bacillus.*

An effective screening approach is the major element of success in discovering novel biosurfactant producers. The screening techniques applied in the current study are amongst the most used methods in biosurfactant production determination, in which all assays are mostly based on physical effects of biosurfactants produced (Cai et al., 2015; Kosaric & Sukan, 2014; Sari et al., 2014; Shaligram et al., 2016; Walter et al., 2010).

Bernheimer & Avigad (1970) reported that surfactin produced by *Bacillus subtilis* lyse the red blood cells. There is an association between hemolysis activity and surfactant production, since then hemolytic assay is recommended as a primary technique to screen biosurfactant producers (Youssef et al., 2004). Therefore, this assay was employed in this research. Biosurfactants are well known to have haemolytic, antibacterial, and antiviral activity, owning a precise mechanism that has impact on the membrane permeability and eventually leading to cell disruption (Heerklotz & Seelig, 2007).

The drop-collapse test depends on the destabilization of liquid droplets by the biosurfactants produced by the bacterial isolate tested. Consequently, drops of the culture free-cell supernatant on the microplate surface will result in either stable or spreading or even collapsing droplets depending on the presence of biosurfactants. If the supernatant does not contain biosurfactants, the polar water molecules are repelled from hydrophobic surface and the droplets spread or even collapse because the force or interfacial tension between the liquid drop and the hydrophobic surface is reduced. The

stability of drops is dependent on biosurfactant concentration and correlates with surface and interfacial tension (Sari et al., 2014).

The oil spreading assay, on the other hand, is based on the formation of a clear zone and a displacement area when biosurfactant activity is present in the culture free-cell supernatant displacing the oil in the assay. The diameter of this clearing zone on the oil surface correlates the biosurfactant activity.

The emulsification assay was performed to determine the emulsification ability of UMX-103. The result showed that UMX-103 has the ability to emulsify oil surface. In general biosurfactants are emulsifiers which have correlation with the value of Hydrophilic-Lipophilic Balance (HLB) (Cai et al., 2015). The stability of micelles vary amongst high HLB and low HLB, where the high HLB stabilize oil in water emulsions and low HLB stabilize water in oil emulsion (Pacwa-Płociniczak et al., 2011).

The biosurfactants assays conducted in this study determined that UMX-103 is a biosurfactant producer, interestingly the strain has the ability to lower the surface tension up to 26 mN/m which is the best compared to other *Bacillus* species. Other *Bacillus* species and *Bacillus subtilis* strains were reported to exhibit lower surface tension from 72 mN/m to a range between 39 mN/m and 27 mN/m (Cai et al., 2015; Dadrasnia & Ismail, 2015; Joshi et al., 2012; Shoeb et al., 2015; Vaz et al., 2012). (Figure 4.7) shows comparison of the surface tension reported from various *Bacillus* species.

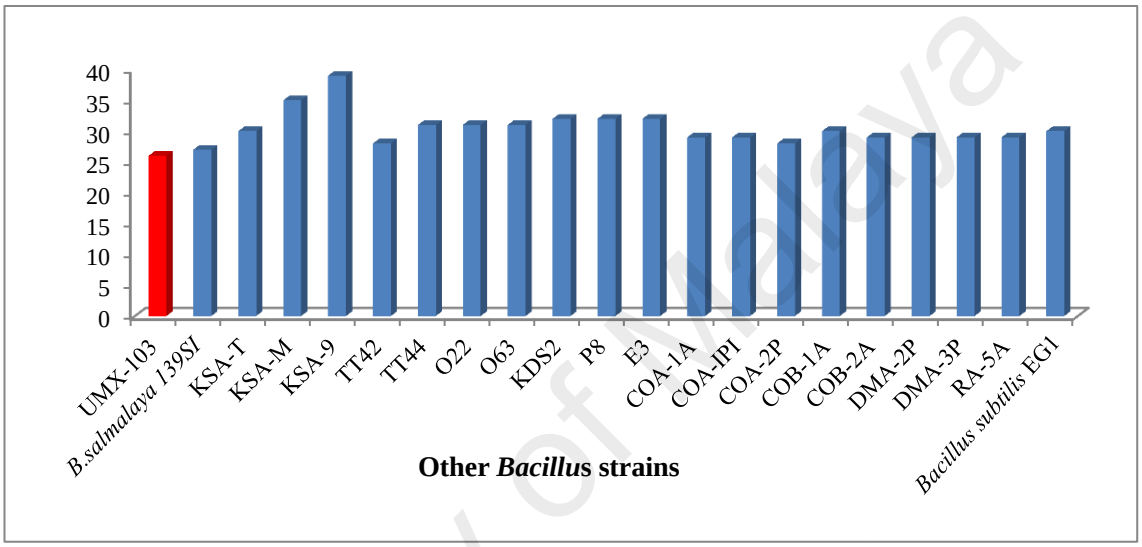

Figure 4.7: Surface tension comparison with other *Bacillus* species. UMX-103 has the lowest surface tension value which is 26 mN/m followed by *B. salmalaya* 139SI (value of 27 mN/m).

In summary, the Gram staining along with the FESEM results revealed that UMX-103 is a Gram positive rod shape bacteria belongs to the genus *Bacillus*. The determination of biosurfactant production by UMX-103 was identified using the most known methods up to date in screening biosurfactants producers. Haemolytic activity determined the ability of the bacteria to lysis the red blood cells in the blood agar plate which also reflected the antibacterial properties of UMX-103. Additionally, the oil spreading test and drop-collapse assay confirmed the ability of biosurfactnt production by the UMX-103 due to the formation of the clear zoon on the oil surface and destabilize the droplet on the oil surface causing the droplet to collapse. Biosurfactants are emulsifiers have the ability to emulsify oil surface. The emulsification assay results showed that UMX-103 has emulsification properties. Generally, biosurfactants have the ability to lower surface tension from 72 mN/m to 27 mN/m. Interestingly, UMX-103 has the ability to reduce surface tension up to 26 mN/m which is the best compared to the other *Bacillus* producers reported.

All of the sampling information regarding the isolate UMX-103 was deposited in DDBJ/EMBL/GenBank BioSample and BioProject databases under accession number SAMD00051050 and PRJDB4745, respectively. The bacterium was labeled as UMX-103.

CHAPTER 5: WHOLE GENOME SEQUENCE AND DATA ANALYSIS

5.1 Introduction

In recent years, biosurfactants have gained extensive attention due to their widespread applications in pharmaceutical, food and many other industries. These biologically and industrially valuable biomolecules cover variety of surface active compounds such as lipopeptides, terpenoids and bacteriocins. Surfactants have been extensively studied for years, both with functional and structural characteristics. Apart from environmental and industrial uses, these surface active compounds have wide biological applications (Ongena & Jacques, 2008).

Recent significant evolution in Whole Genome Sequencing (WGS) and computing technologies which can handle huge volumes of data using super computers with high speed and capacity RAMs have enabled the assembly and determination of a bacterial genome. Microbial WGS sequencing has a great potential not only for medical applications and public health microbiology but also for understanding production mechanisms of useful materials by microorganisms (Kamada et al., 2014). Using these technologies, several attempts have been made to determine various bacterial genomes such as *Bacillus sp*. AM13 (Shaligram et al., 2016) and *Bacillus subtilis* natto (Kamada et al., 2014; Nishito et al., 2010).

The objective of this chapter is to analyze the whole genome sequence of UMX-103 as outlined under objective 2 (see page 3).

5.2 Genomic data pre-processing

A total of 565,068,437 paired-end reads with a length of 101 bp were generated using Illumina HiSeq 2000 platform, with an average insertion size of 534 bp. Low quality bases and reads were filtered to obtain an optimal quality score of 30 or higher at each base (Figure 5.1). There was bubble in the flow cell used during the sequencing which caused error in base calling at position 54 in each read, therefore the quality score is below 30 at this position. As the sequencing coverage is high and the availability of reference genomes it aid in overcoming this error. The preprocessed reads were exported as Fastq files for further bioinformatics data analysis.

5.3 Genome mapping and assembly of UMX-103

The reads were mapped to *Bacillus subtilis* strain 168 and other reference genomes which are used in this study (Table 5.1). The mapping results are shown in Table 5.8, The mapping result to the *Bacillus subtilis* strain 168 showed 93.44 % of genomic similarity to UMX-103.

Velvet assembler which was used for *de novo* assembly generated a total of 69 contigs with average length of 61,362 bp, with maximum and minimum length of 869,096 bp and 137 bp, respectively. All contigs generated by Velvet were used to generate the scaffolds using SSPACE-Standard software (Boetzer et al., 2011). Scaffolding process generated a total of 39 scaffolds, with average scaffold size of 108,565 bp with maximum and minimum scaffold sizes of 1,059,836 bp and 144 bp, respectively. Then GapFiller software (Boetzer & Pirovano, 2012) was used to close the gaps in the generated scaffolds. A total of 34 gaps from 41 gaps were closed. In addition, the result after gap closing shows total of 39 scaffold with average size of 1,085,80 bp (Table 5.2). The scaffolds were sorted according to the reference genome

using MAUVE (Rissman et al., 2009) (Figure 5.2). The assembled genome was deposited in DDBJ/EMBL/GenBank under the accession number BDCV01000000.

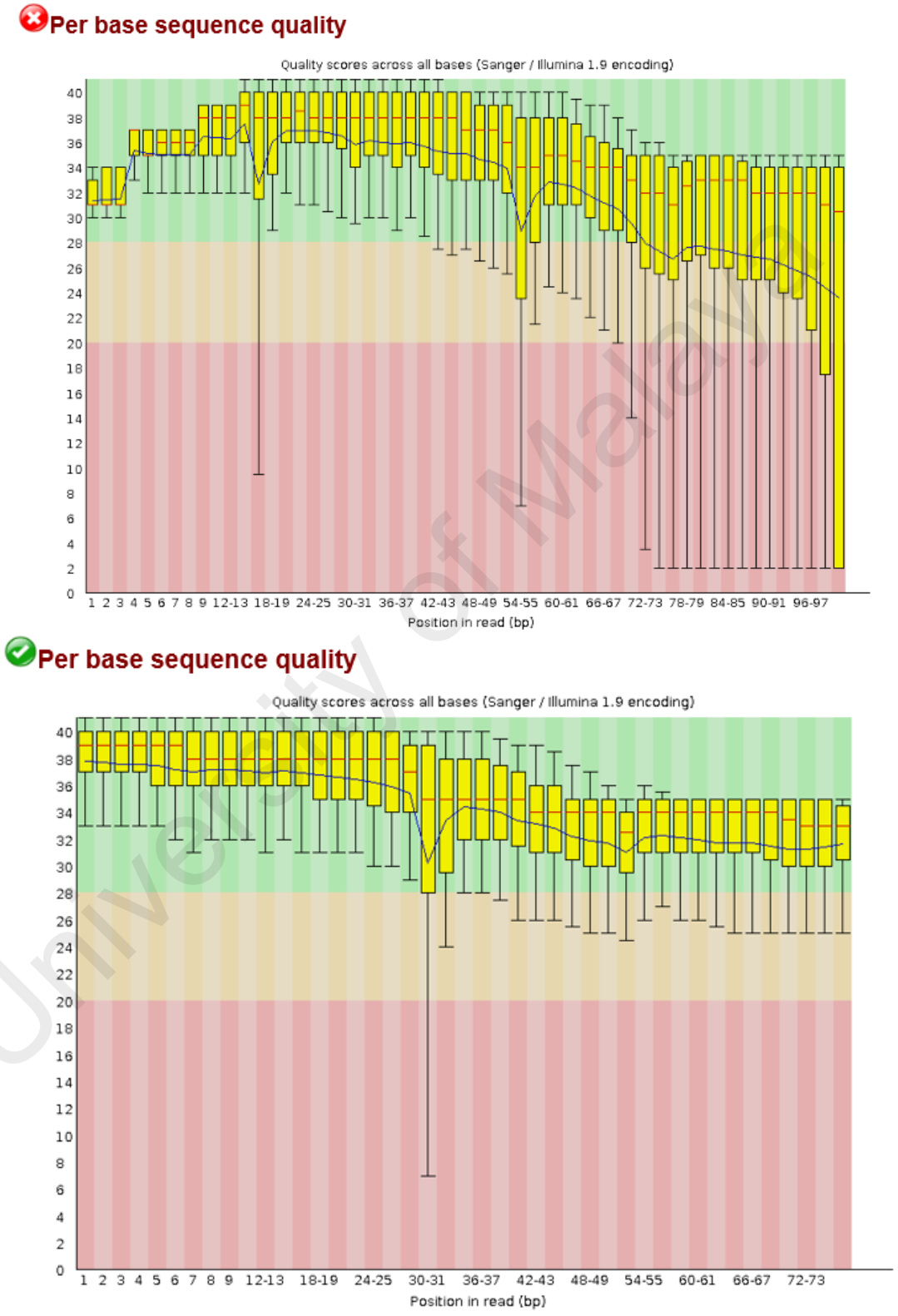

Figure 5.1: Quality control of the generated data before and after trimming process. Upper panel: before trimming; lower panel after trimming

Table 5.1: Genomes used in this study. Species name and accession numbers of genomes selected in this study

Genomes	BioProject	Assembly	Accession No
Bacillus subtilis 168	PRJNA76	GCA_000009045.1	NC_000964.3
Bacillus subtilis LM 4-2	PRJNA277611	GCA_000978495.1	NZ_CP011101.1
Bacillus subtilis BEST7003	PRJDB111	GCA_000523045.1	NZ_AP012496.1
Bacillus subtilis KCTC1028	PRJNA81651	GCA_000971925.1	NZ_CP011115.1
Bacillus subtilis RO-NN-1	PRJNA68559	GCA_000227485.1	NC_017195.1

Table 5.2: Summary of *de novo* assembly of UMX-103

Software	Number of contigs / scaffolds	Average size (bp)	Maximum size (bp)	N50 (bp)	Number of Ns
Velvet	69	61362	869096	320133	1571
SSPACE-Standard	39	108565	1059836	810791	2358
GapFiller	39	108580	1059595	810618	7

2351 out of 2358 unknown nucleotides were replaced with known nucleotide. (**Ns**) = unknown nucleotides; (**bp**) = base pair.

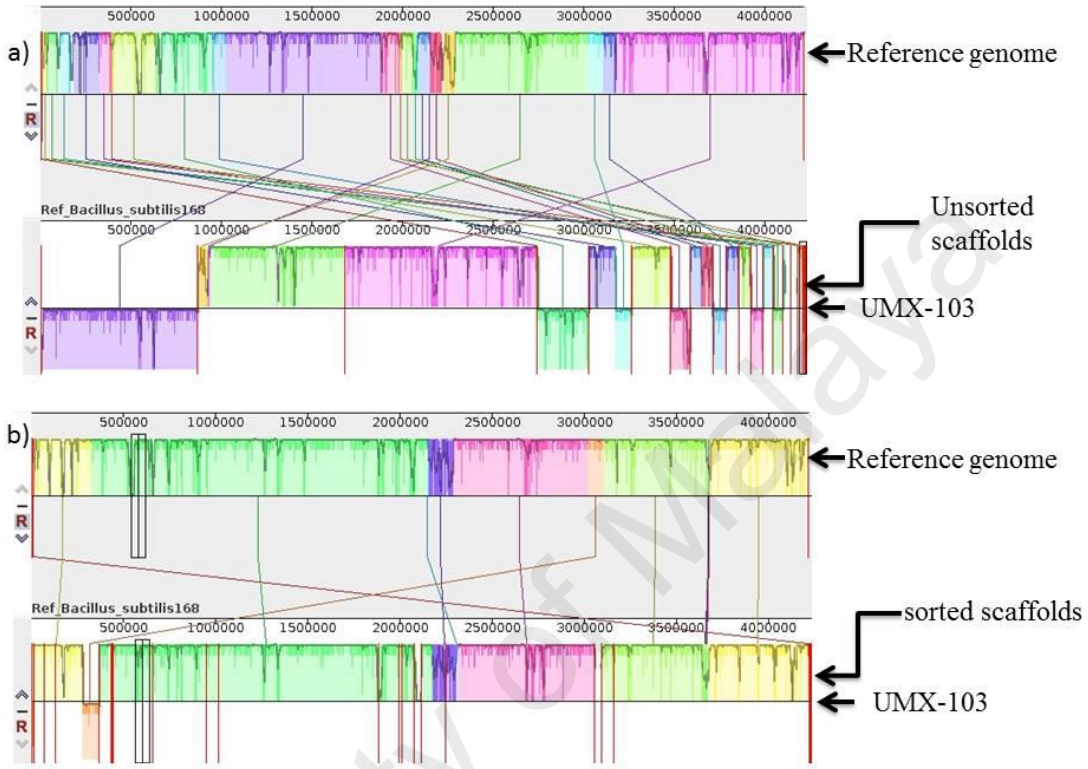

Figure 5.2: Scaffold sorting. Panel a) shows the scaffold before sorting them. Panel b) shows the scaffolds after sorting.

5.4 Gene predication and annotation

Bacillus subtilis UMX-103 contains a single circular chromosome of the size 4,234,627 bp with an average G+C content of 43.41% (Table 5.3). The assembled genome consists of 39 scaffolds with an average scaffold size of 108,580 bp using a combination of several gene-prediction software and manual inspection, a total of 4,301 protein-coding genes and 98 RNA genes were identified in this strain (Figure 5.3)**.**

Table 5.3: Key features of *Bacillus subtilis* UMX-103

Feature	**Genome**
DNA, total number of bases	4,234,627 (bp)
GC content	43.41%
Total number of genes	4399
Protein coding genes	4301
RNA genes	98
rRNA genes	4
5S rRNA	2
16S rRNA	1
23S rRNA	1
tRNA	94

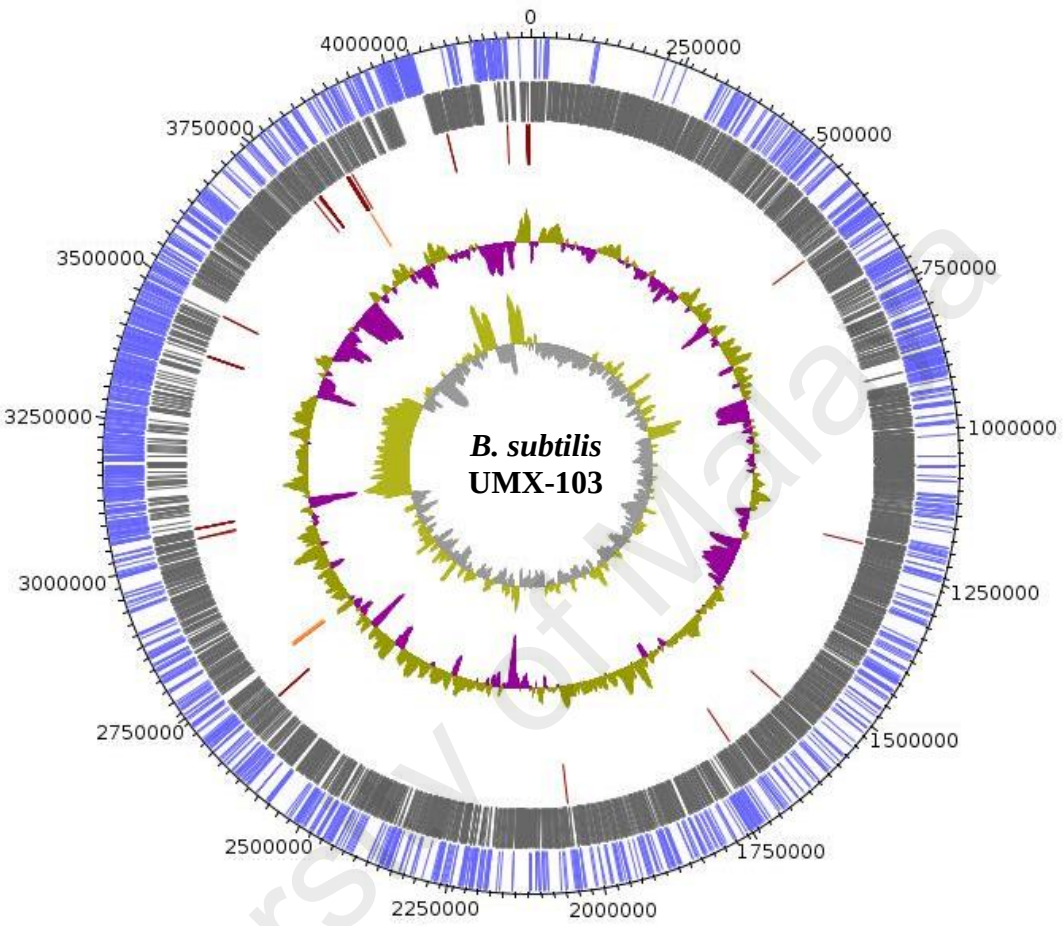

Figure 5.3: *Bacillus subtilis* UMX-103 genome features. The two outmost concentric circles denote the predicted protein-coding genes represented as forward strand (external blue circle) and the reverse strand (internal grey). The third concentric circle (purple) represents tRNAs while the fourth concentric circle (light brown) shows rRNAs genes. The fifth concentric (green and purple) represents the GC content. The green colour shows GC content more than the average while the purple colour shows the GC content below average. Purple and green in the last inner concentric represent GC skew.

5.5 Genome similarity and Phylogenetic analysis

Average Nucleotide Identity (ANI) of UMX-103 was determined by comparing the whole genome with the selected references (Table 5.4). The highest ANI was detected with KCTC 1028 and 168 strains which is 89%. The 16S rRNA from UMX-103 was used to carry out the phylogenetic analysis where it was aligned with other 16S rRNA genes of *Bacillus* strains which including; *B. subtilis* LM 4-2, *B. subtilis* BEST7003, *B. subtilis* KCTC 1028, B. *subtilis* 168, *B. subtilis* RO-NN-1, *B. amyloliquefaciens* DSM7, *B. licheniformis*, *B. pumilus* GR-8 and *Paenibacillus macerans* (Figure 5.4). Additionally the seven housekeeping genes used to determine the species of UMX-103 were detected using Multilocus Sequence Typing (MLST) server 1.8 (Larsen et al., 2012) (Table 5.5). All of the housekeeping genes in UMX-103 are highly identical with the housekeeping genes of *Bacillus subtilis.*

Table 5.4: Average nucleotide identity of UMX-103

Query genome	Reference genome	DDH	Distance	Prob. DDH >= 70%	G+C difference
UMX-103	*B. amyloliquefaciens* DSM7	20.5	0.2139	0	2.67
UMX-103	*B. subtilis* 168	89	0.0132	95.45	0.11
UMX-103	*B. subtilis* LM4-2	89.1	0.0131	95.5	0.42
UMX-103	*B.subtilis* BEST7003	89	0.0132	95.44	0.48
UMX-103	*B. subtilis* KCTC1028	89	0.0132	95.46	0.11
UMX-103	*B. subtilis* RO-NN-1	82.7	0.0202	92.43	0.46
UMX-103	*B. licheniformis*	18.8	0.2337	0	2.47
UMX-103	*Paenibacillus macerans*	27.9	0.1544	0.04	9.16
UMX-103	*B. pumilus GR-8*	17.9	0.2449	0	1.98

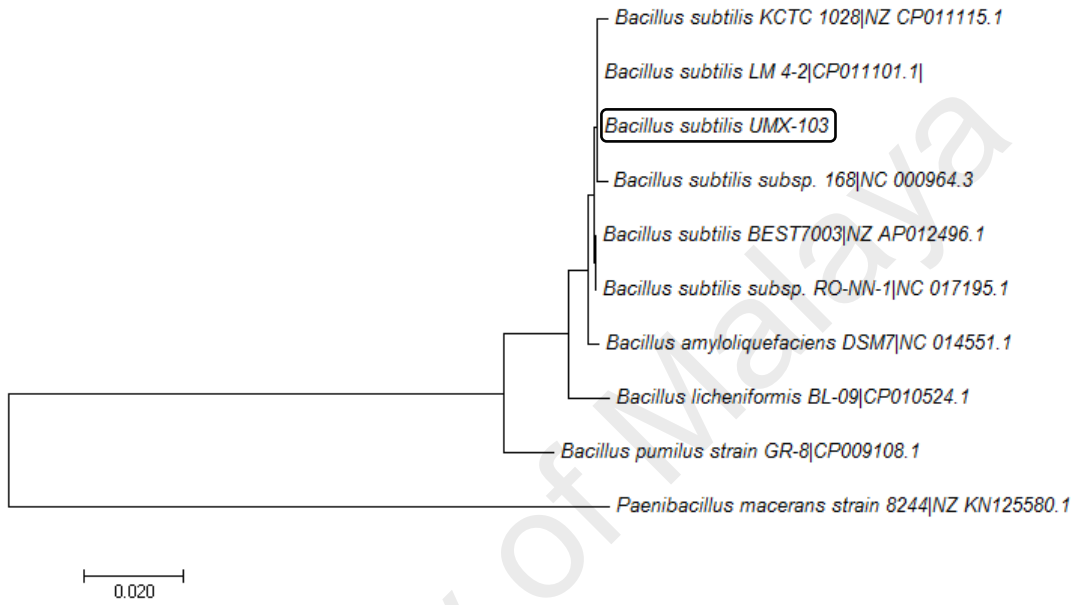

Figure 5.4: Phylogenetic analysis based on 16S rRNA gene. Phylogenetic reconstruction was performed based on the sequence of 16S rRNA gene using MEGA7 (Kumar et al., 2016). The 16S rRNA genes sequence of *Bacillus pumilus* GR-8 and *Paenibacillus macerans* was used as outgroup.

Table 5.5: MLST of the 7 housekeeping genes in *Bacillus subtilis*

Gene	***Bacillus subtilis*** **UMX-103**	***Bacillus subtilis*** **LM 4-2**	***Bacillus subtilis*** **BEST7003**	***Bacillus subtilis*** **KCTC1028**	***Bacillus subtilis*** **168**	***Bacillus subtilis*** **RO-NN-1**
glpf	100%	100%	100%	100%	100%	100%
ilvd	100%	99.79%	100%	100%	100%	100%
pta	100%	100%	100%	100%	100%	100%
purh	100%	100%	100%	100%	100%	100%
pyca	100%	100%	100%	100%	100%	100%
rpod	100%	100%	100%	100%	100%	100%
tpia	100%	100%	100%	100%	100%	100%

5.6 Functional annotation

The annotated genes which performed using Prokka software were used for functional annotation analysis. The functional annotation was conducted using EggNOG-mapper, the summary of functional categories of annotated genes is shown on Table 5.6. There are total of 3712 protein-coding genes in UMX-103 annotated based on their function. A total of 618 genes involve in information storage and processing, 672 genes involve in cellular processing and singling, 1332 genes involve in bacteria metabolism and 1090 genes are poorly characterized. Among the 618 genes involved in information storage and processing, there are 166 genes implicate in translation and biogenesis (Appendix A); 269 transcriptional genes (Appendix B); 155 genes involve in DNA replication, recombination and repair (Appendix C).

The result reveled existence of biosynthetic cluster of genes which are known for coding surfactin. This gene cluster belongs to Non-ribosomal Peptide Synthetase (NRPS) family, particularly to the microbial surfactants group. These genes usually present in secondary metabolites biosynthesis, transport and catabolism (Doroghazi et al., 2014). This cluster of genes is further elaborated in Chapter 6.

5.7 Genomic islands

Genomic islands analysis is widely used to compare bacteria strains and identify essential genes in bacterial genome (Dobrindt et al., 2004; Langille et al., 2010). Basically, genomic islands associate with Horizontal Gene Transfer (HGT) which also known as mobile genetic elements. There are 15 genomic islands in UMX-103 that was predicted by IslandViewer 3 (Dhillon et al., 2015) and the localization of the predicted genomic islands is shown in Figure 5.5. The 15 predicted genomic islands consist of 331 genes (Appendix D). These genomic islands are possibly having a significant role in adapting and surviving the bacteria to different abiotic stress and antimicrobial

resistance, which may occur after the bacteria was exposed to different environment including the hydrocarbon contaminated soil. It is possible that *Bacillus subtilis* UMX-103 has witnessed a number of Horizontal Gene Transfer events. Moreover, the identification of genomic islands revealed identical genes available in other *Bacillus* species. Features of the genomic islands are given in Table 5.7.

Table 5.6: Functional annotation of the predicted genes of *Bacillus subtilis* UMX-103

INFORMATION STORAGE AND PROCESSING	
Translation, ribosomal structure and biogenesis	166
Transcription	296
Replication, recombination and repair	155
Chromatin structure and dynamics	1
CELLULAR PROCESSES AND SIGNALING	
Cell cycle control, cell division, chromosome partitioning	31
Defense mechanisms	65
Signal transduction mechanisms	130
Cell wall/membrane/envelope biogenesis	297
Cell motility	43
Intracellular trafficking, secretion, and vesicular transport	34
Posttranslational modification, protein turnover, chaperones	106
METABOLISM	
Energy production and conversion	184
Carbohydrate transport and metabolism	281
Amino acid transport and metabolism	296
Nucleotide transport and metabolism	93
Coenzyme transport and metabolism	114
Lipid transport and metabolism	99
Inorganic ion transport and metabolism	200
Secondary metabolites biosynthesis, transport and catabolism	65
POORLY CHARACTERIZED	
Function unknown	1090

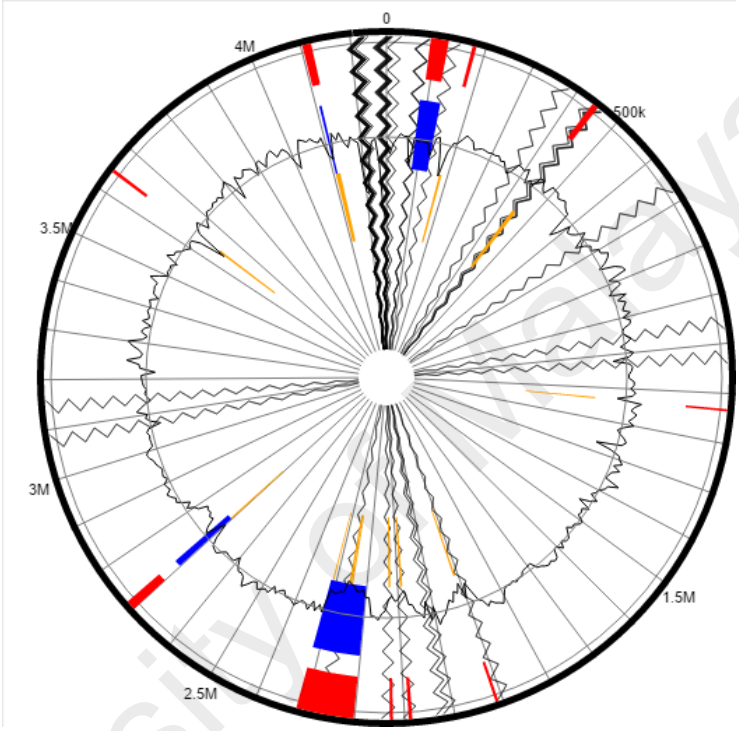

Figure 5.5: Genomic Islands of *Bacillus subtilis* UMX-103. Red colour defines predicted genomic islands using integrated method. The blue colour shows genomic islands predicted by IslandPath-DIMOB while yellow colour shows genomic islands predicted by SIGI-HMM method. The broken lines represent scaffolds borders.

Table 5.7: Genomic islands feature of *B. subtilis* UMX-103

Genomic islands	Start (bp)	End (bp)
G1	88253	122890
G2	171240	179050
G3	426292	431318
G4	1129385	1133873
G5	1895277	1917461
G6	1907520	1914315
G7	2088719	2094260
G8	2240947	2356838
G9	2284268	2292538
G10	2347793	2351919
G11	2739783	2758593
G12	2739806	2745071
G13	3680804	3687700
G14	4142627	4157049
G15	4145918	4160821

5.8 Genomic comparison of UMX-103 with close related bacteria

5.8.1 Comparative genomics

Comparative genomics of *B. subtilis* UMX-103 with six other related reference genomes (Table 5.8) showed that *B. subtilis* UMX-103 is most closely related to *Bacillus subtilis* KCTC 1028 and *Bacillus subtilis* 168 with the genome sequence similarity of 93.99% and 93.44%, respectively. Analysis showed that *B. subtilis* UMX-103 has the largest genome size compared with the other bacteria strains studied. The genome contains the highest number of genes and the lowest GC contents which is 43.41%. The ANI of *Bacillus subtilis* UMX-103 with the reference genomes range between 82.2% and 89.10%. This ANI percentage and the high sequence similarity based on mapping the genome to the reference genomes supports our findings that UMX-103 belongs to *Bacillus subtilis* species. All reference genomes statistics were retrieved from https://img.jgi.doe.gov.

The genomic islands comparison (Table 5.9) showed that *B. subtilis* UMX-103 has the same number of genomic islands with *Bacillus subtilis* LM 4-2; however, the total number of genes in the genomic islands of *B. subtilis* UMX-103 is 331 genes, where only 108 genes were found in *Bacillus subtilis* LM 4-2. The genomic islands of the respective genomes are presented in Figure 5.6.

5.8.2 Pangenome analysis

Pangenome analysis revealed total of 735 essential genes in UMX-103 (Appendix E). Pangenome analysis resulted in the identification of 3434 core genes which present in the entire *Bacillus* strains studied (Figures 5.7 and 5.8). Pangenome composed of the essential genes in species. It also used as a method in identification unknown bacteria (Lasken & McLean, 2014).

Table 5.8: Genomic comparisons with closely related bacteria strains

RefSeq		NZ_CP011101.1	NZ_AP012496.1	NZ_CP011115.1	NC_000964.3	NC_017195.1
Genome Features	*Bacillus subtilis* UMX-103	*Bacillus subtilis* LM 4-2	*Bacillus subtilis* BEST7003	*Bacillus subtilis* KCTC1028	*Bacillus subtilis* 168	*Bacillus subtilis* RO-NN-1
DNA, (total number of bases)	4234627	4069266	4043042	4215633	4215606	4011949
GC content %	43.41	43.83	43.89	43.51	43.51	43.87
Total number of genes	4399	4143	4133	4369	4354	4257
Protein coding genes	4301	3994	4011	4215	4176	4141
RNA genes	98	149	122	154	178	116
rRNA genes	4	30	30	30	30	30
5S rRNA	2	10	10	10	10	10
16S rRNA	1	10	10	10	10	10
23S rRNA	1	10	10	10	10	10
tRNA	94	86	92	86	86	86
Mapping %	-	92.89	91.81	93.99	93.44	90.58
Average Nucleotide Identity %	-	89.10	89	89	89	82.8

Table 5.9: Genomic islands comparison of *Bacillus subtilis* UMX-103 with close related genomes

RefSeq		**NZ_CP011101.1**	**NZ_AP012496.1**	**NZ_CP011115.1**	**NC_000964.3**	**NC_017195.1**
Genome	*Bacillus subtilis* UMX-103	*Bacillus subtilis* LM 4-2	*Bacillus subtilis* BEST7003	*Bacillus subtilis* KCTC1028	*Bacillus subtilis* 168	*Bacillus subtilis* RO-NN-1
Number of genomic islands	15	15	12	16	22	17
Number of genes in genomic islands	331	108	75	125	440	269

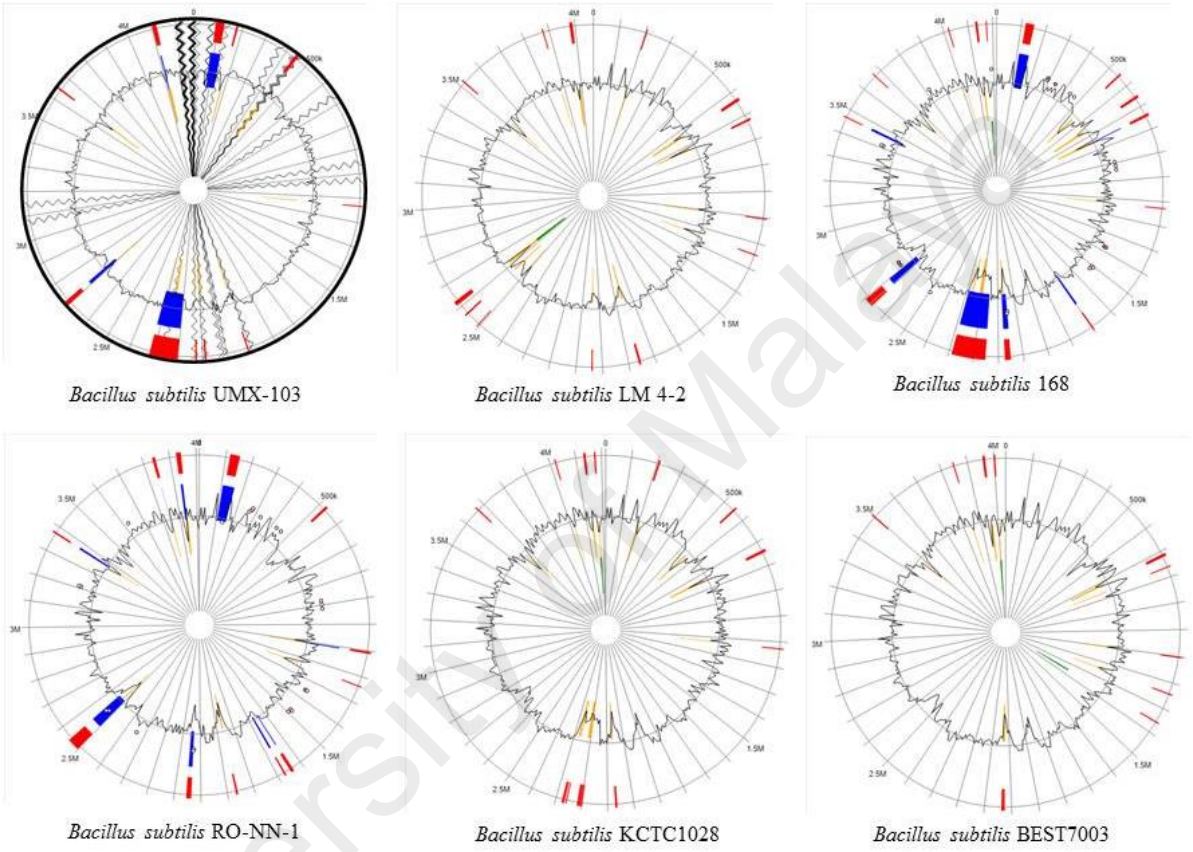

Figure 5.6: Genomic islands of UMX-103 and other reference genomes used in this study. Red colour defines predicted genomic islands using integrated method. The blue colour shows genomic islands predicted by IslandPath-DIMOB while yellow colour shows genomic islands predicted by SIGI-HMM method.

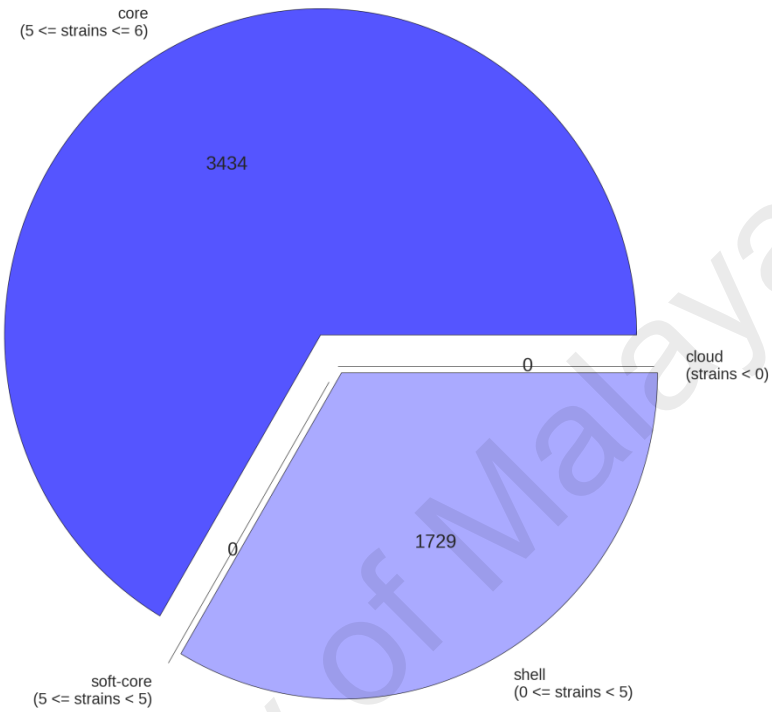

Figure 5.7: Core genes within UMX-103 and reference genomes. The 3434 core genes are present in the 6 genomes which include; *B. subtilis* UMX-103, *B. subtilis* LM 4-2, *B. subtilis* BEST7003, *B. subtilis* KCTC1028, *B. subtilis* 168 and *B. subtilis* RO-NN-1.

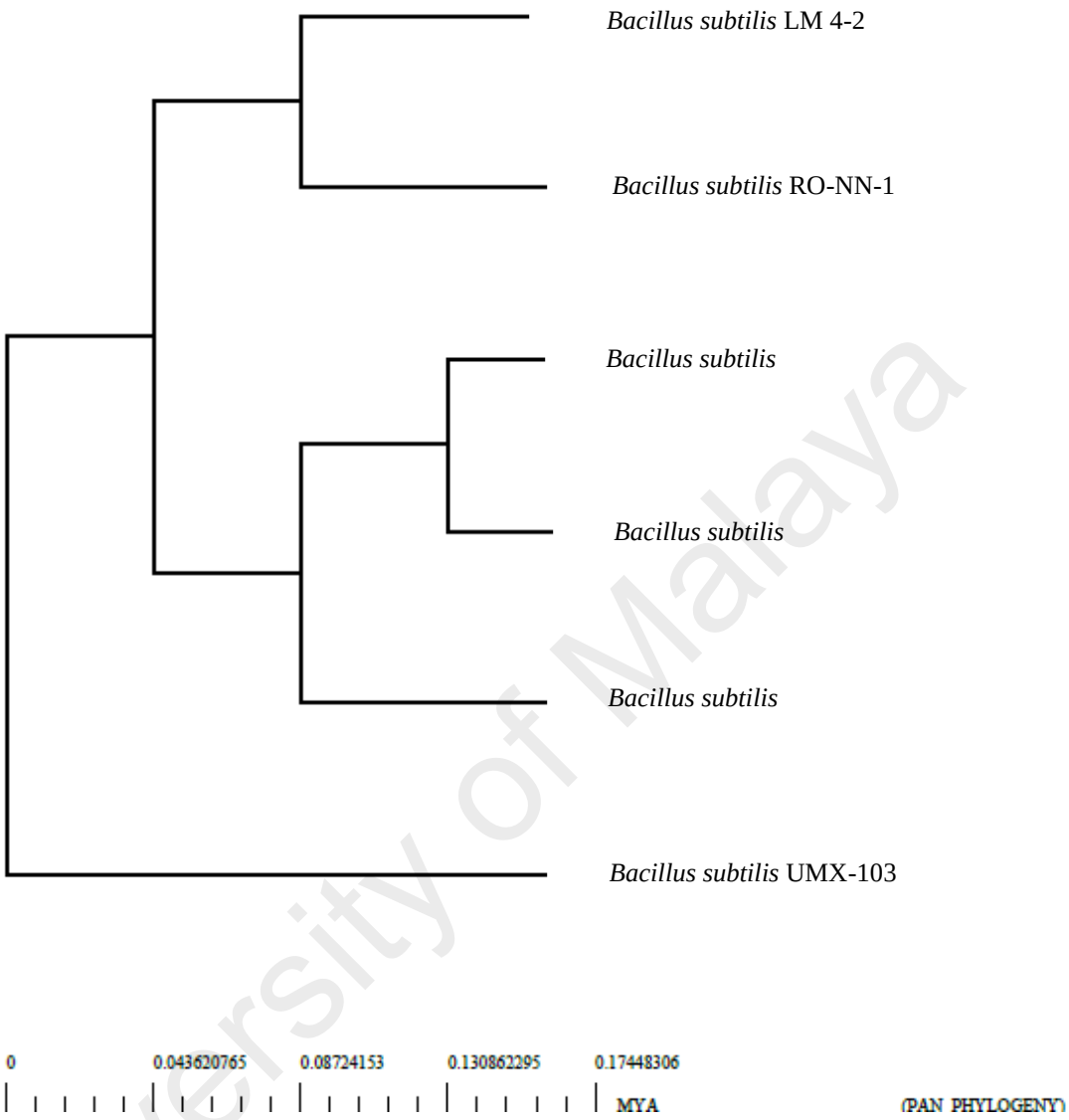

Figure 5.8: Phylogenetic tree of UMX-103. The tree was constructed using all of the genes shared among all 6 strains (3434 genes).

Pangenome analysis also revealed the Cluster Orthologues Groups (COGs). The COG distribution pattern showed involvement of more core genes in carbohydrate transport and metabolism and translation, ribosomal structure and biogenesis, while accessory and unique genes appear to be enriched in Transcription, Replication, recombination and repair and Cell wall /membrane/envelope biogenesis related functions (Figure 5.9).

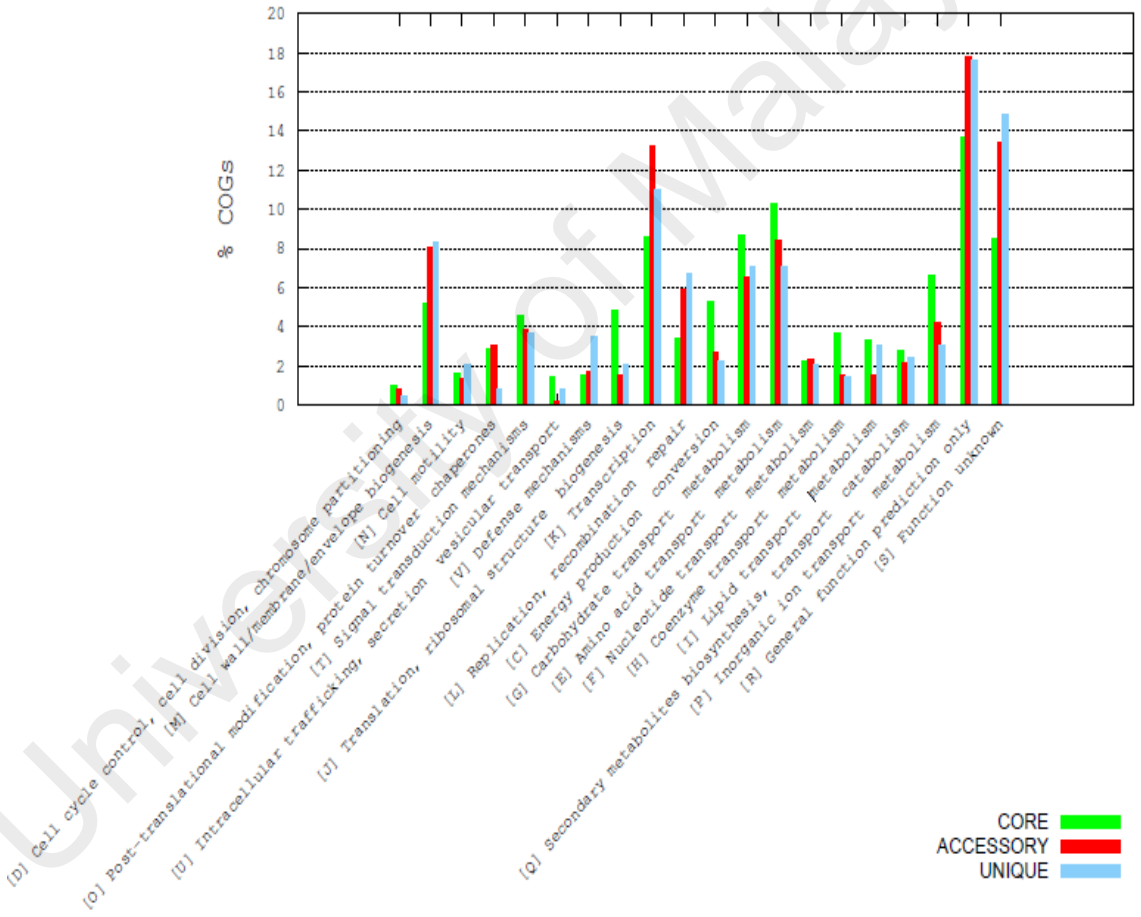

Figure 5.9: COG distribution. Shows most of the core genes categorized in amino acid transport metabolism, carbohydrate transport metabolism and transcription. The unique genes mostly fall in transcription and envelope biogenesis category.

5.9 Discussion

The rapid development of NGS technology and the advances in bioinformatics have had a major impact on understanding genomics and functional genomics of microbial genomes (Land et al., 2015). It is possible for research groups to generate draft genome sequences for any organism of interest. In addition, it assists in reducing errors and improves *de novo* assemblies (Escalona et al., 2016; Kulski, 2015). Although, Illumina sequencing technology has >1% error rate (Escalona et al., 2016) the quality assessment for the raw data generated from this study were conducted to ensure high quality *de novo* assembly. In this study, a combination of *de novo* and reference-guided assembly were performed to determinate accurate assembly of UMX-103 genome. This hybrid approach in genome assembly has been applied in assembling various microbial genomes (Nishito et al., 2010; Shaligram et al., 2016; Yan et al., 2016). *Bacillus subtilis* strain 168 was used as reference genome, because it is the most studied *Bacillus* strain and it has been widely used in genetic research (Barbe et al., 2009; Nishito et al., 2010; Srivatsan et al., 2008). The reference assembly results in similar genome size with the reference genome (4,215,606 bp) this is due to the mapping of generated reads to the reference genome which shows high genome similarity. Additionally, the mapping results showed a number of the reads which did not mapped to any region in the reference genome. Subsequently, *de novo* assembly based on overlapping using de Bruijn graph algorithm (Zerbino & Birney, 2008) resulted in 4,234,627 bp which is a larger genome size compared to the reference genome. Therefore, comparative genomics analysis was conducted to highlight the unique features in UMX-103 genome.

For many years, 16S rRNA gene is used as a primary tool to study bacterial taxonomic assignment and phylogenetic trees. It has been widely used as the most common housekeeping genetic marker in bacterial genome due to several reasons. It is present mostly in all bacteria and often present as a multigene family or operons. The

functions of 16S rRNA gene do not change overtime suggesting that the gene sequence is highly conserve among genus and species. The size of 16S rRNA gene is about (1500 bp) which is sufficient to provide information on bacteria family, genus and species (Janda & Abbott, 2007; Land et al., 2015). Using only 16S rRNA genes approach to identify bacteria is not recommended, because some *Bacillus* species share 16S rRNA genes >99.5% similarity. Therefore, it is recommended to use ANI along with this approach (Janda & Abbott, 2007). The ANI approach is widely used in bacteria determination (Kim et al., 2014). Although, 16S rRNA approach showed the UMX-103 belongs to species *Bacillus subtilis.* MLST approach was used to improve the taxonomic resolution of these groups. The integrative approach, which includes use of 16S rRNA gene, seven MLST genes and ANI aided in determining the taxonomic classification of UMX-103. The phylogenetic analysis conducted along with the ANI approach successfully determined the taxonomic classification as a new *Bacillus subtilis.* The genome wide comparison conducted emphasized the level of similarity and complexity in genomes of these closely related species within *B. subtilis* group.

Genome comparison with close related bacterial genomes showed that strain KCTC1028 and *B. subtilis* 168 are very close to UMX-103, however these genomes contain fewer number of genes compared to UMX-103. This distinguishes the genetic composition of UMX-103. Genomic islands analysis highlighted several mobile elements in UMX-103, interestingly, *srfAB* gene presented in the predicted GIs. This may be associated to the biosurfactants properties of the strain.

The functional annotation revealed 1332 genes involved in metabolisms which reflect the metabolism rate in UMX-103 is high (Figure 5.10). Amongst the 296 genes annotated based on COG which belongs to the transcription category (Appendix B). There are 11 transcriptional regulators genes of *LysR* type (COG0583), this genes are

the most enormous type of transcriptional regulators in the kingdom of prokaryotic. They play essential role in regulation genes involved in catabolism of aromatic compounds, quorum sensing and cell motility (Binnewies et al., 2006). Additionally, a total of 8 genes were found to be RNA polymerase sigma factor subunits (COG1595). Also there are 4 *Arac* DNA-binding regulators genes (COG2207) and 2 Arginine utilization regulatory genes, *RocR* type (COG3829) were found in UMX-103 genome. One regulatory gene represses a number of genes involved in the response to DNA damage including *recA* and *lexA*. In the presence of single-stranded DNA *RecA* interacts with *LexA* causing an autocatalytic cleavage which disrupts the DNA binding part of *LexA* leading to derepression of the SOS respond regulon and eventually DNA repair (COG1974). Six genes were found to be repressors of the *marRAB* operon (COG1846) which is involved in the activation of both antibiotic resistance and oxidative stress genes.

Functional annotation revealed genes cluster involved in biosurfactant production. These genes are present in secondary metabolism category. These genes known to be NRPS surfactin biosynthetic gene cluster. It composes of *srfA* operon that synthesis surfactin in *Bacillus subtilis* (Arima et al., 1968; Cosmina et al., 1993; Marahier, Nakano, & Zuber, 1993; Nakano et al., 1991). In addition, it revealed genes involve in surfactin regulation which is further discussed in Chapter 6.

Arima et al., (1968) reported the discovery of surfactin which was produced by *Bacillus subtilis.* The most well-known microbial surfactants lipopeptides are surfactin, polymyxin B, and daptomycin that produced by *Streptomyces roseosporus* (Baltz et al., 2005). Extensive researches were conducted to study and characterize surfactin (Davis et al., 2001; Dexter & Middelberg, 2008; Nakano et al., 1992; Noah et al., 2002; Sen & Swaminathan, 2004).

Structurely surfactin composed of seven amino acids (Figure 5.11). Amongst the cyclic lipopeptide biosurfactants, surfactin which produce by *Bacillus subtilis* ATCC-21332 is the most effective and used in industries. As it has the ability to lower surface tension from 72 to 27.9 mN/m (Desai & Banat, 1997; Meena & Kanwar, 2015).

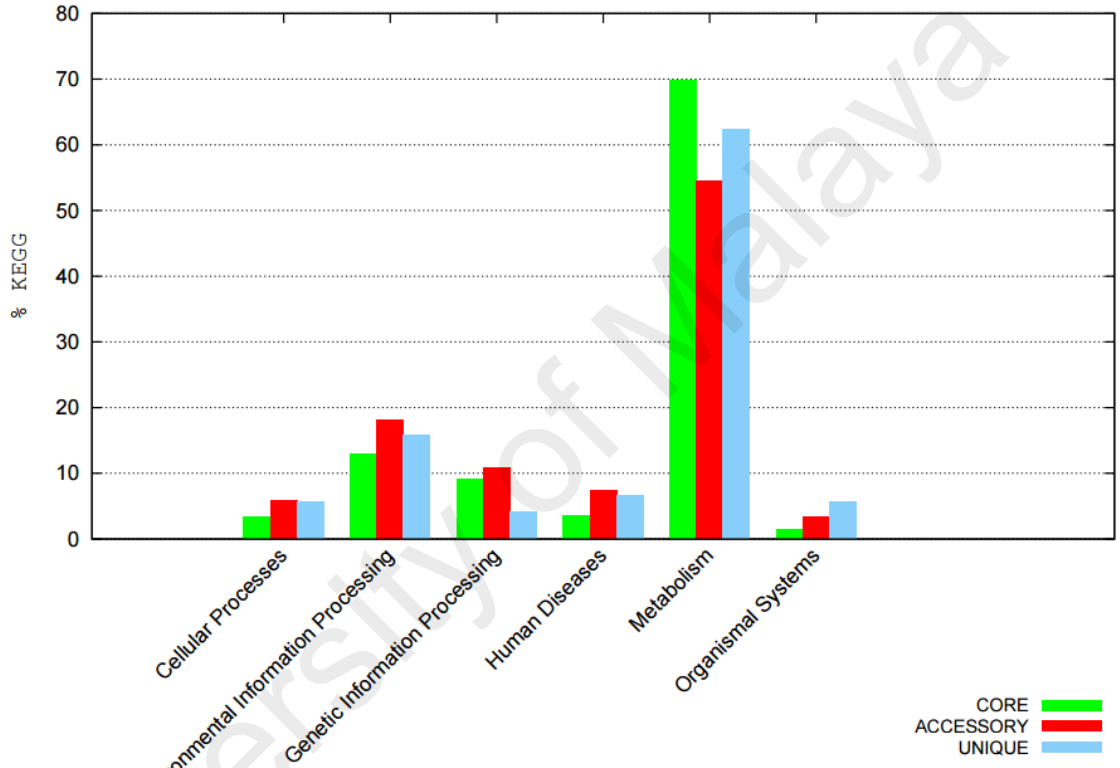

Figure 5.10. KEGG distribution. Most of the genes are involve in metabolism.

Lipopeptide antibiotics

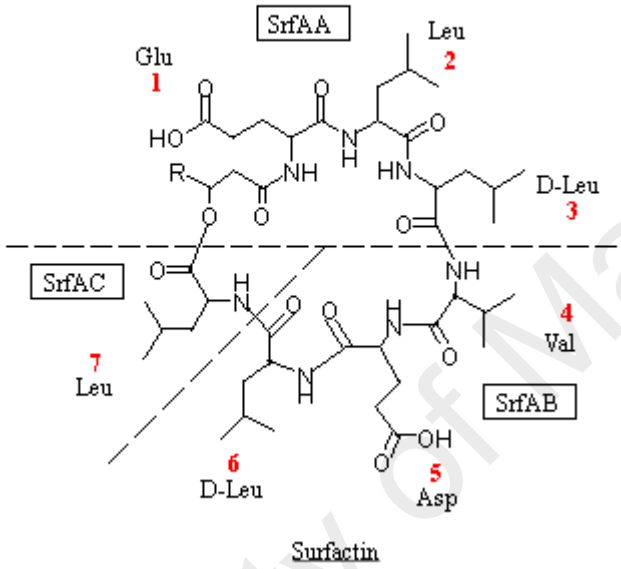

Figure 5.11: Surfactin structure from KEGG database. The lipopeptide comprise of the seven amino acids.

In summary, the whole genome sequence of UMX-103 was obtained from Illumina HiSeq 2000 platform. The genome was assembled using combination of both *de novo* and reference-guided assembly approaches. The assembly resulted in 39 scaffolds with genome size of 4,234,62 bp and 4,301 total number of genes. UMX-103 species was identified using 16S, ANI and MLST analysis. The analysis revealed that UMX-103 belongs to the *Bacillus subtilis* species. The comparative genomics was conducted to highlight the unique features in UMX-103 compared to the closest related genomes. UMX-103 has the largest genome size. Genomic islands revealed a total of 331 genes in UMX-103 which related to HGT in the genome. Functional annotation revealed gene cluster that is involve in the biosynthesis of surfactin which is a type of biosurfactant produced by *Bacillus subtilis*. Pangenome analysis revealed the essential genes in UMX-103 as well as the COGs.

CHAPTER 6: BIOSURFACTANT GENES AND PATHWAYS

6.1 Introduction

In recent years, natural products which produced by many bacterial have gained attention due to their widespread applications in many industries such as pharmaceutical and food as well as many environmental applications. These bioproducts contain a variety of active compounds such as lipopeptides, bacteriocins, etc (Anuradha, 2010). These complexed biomolecules require complicated mechanism for their biosynthesis. Nonribosomal peptide synthetases (NRPS) are one of the complex groups of proteins which are amongst the various known mechanisms necessary to produce bacterial natural products (Wang et al., 2014).

NGS is the current and popular method for the generation of genomic data, producing massive amounts of information rapidly and at a low cost. These techniques allow us to sequence DNA and RNA very quickly, facilitating the acquisition of massive genomic, transcriptomic, DNA–protein interaction and epigenomic data sets; they are also radically changing the way we look at genomes and microbial mechanisms. In addition introduction of NGS technologies has had dramatically improved various approaches like deciphering novel metabolic pathways, genome-based phylogeny, and to carry out comparative genomics to understand the genome wide variations in closely related organisms (Horner et al., 2010; Land et al., 2015).

The objective of this chapter is to identify pathways responsible for biosurfactant production in UMX-103 as outlined under objective 3 (see page 3).

6.2 Biosurfactant genes and pathways

Twenty-five genes were identified in UMX-103 which are involved in biosurfactant production. The list of the identified genes is presented in Table 6.1. These genes involved in the biosynthesis and regulation of surfactin, which is a type of biosurfactant produced by *Bacillus subtilis* with high industrial value (Płaza et al., 2015). The genes that involved in the biosynthesis of biosurfactant are including; 4-phosphopantetheinyl transferase (*sfp*), Glucose-1-Phosphate thymidyly transferase (*rmlA*), dTDP-glucose 4,6-dehydratase (*rmlB*), dTDP-4-dehydrorhamnose 3,5-epimerase (*rmlC*), dTDP-4-dehydrorhamnose reductase (*rmlD*) (Das et al., 2015), non-ribosomal peptide synthetase (*dhbF*) (May et al., 2001).

In this study, two operons *srfA* (Nakano et al., 1991) and *pps* (Coutte et al., 2010) are involved in coding the non-ribosomal peptide synthetase (NRPS) subunits that catalyse the incorporation of the seven amino acid form surfactin (Coutte et al., 2010; Peypoux et al., 1999). The *srfA* operon contains four genes; *srfAA, srfAB, srfAC* and *srfAD*, while the *pps* operon contains five genes; *ppsA, ppsB, ppsC, ppsD* and *ppsE*. The *srfA* operon encode surfactin synthetase subunits (Płaza et al., 2015). Surfactin is made of seven amino acids which are (Glu-Leu-(D)Leu-Val-Asp-(D)Leu-Leu) (Cosmina et al., 1993) (Figure 6.1). The gene *srfAA* encode the peptide synthesis subunit which involved in the makeup of amino acids Glu, Leu and D-Leu. Whereas, *srfAB* encode the subunit that involved in catalyzing of L-Val, L-Asp and D-leu. The third gene in the operon *srfAC* functions in the foundation of Leu amino acid (Figure 6.2). The surfactin synthase thioesterase subunit is produced by *srfAD* (Marahier et al., 1993).The activating enzyme *sfp* plays essential role in surfactin biosynthesis, as it transforms the inactive protein that changes surfactin synthetase into an active form (Nakano et al., 1992; Płaza et al., 2015).

Table 6.1: Genes involve in the biosynthesis and regulation of biosurfactants in UMX-103

Genes involve in Surfactin biosynthesis						
Gene	**Begin**	**End**	**Annotated function**	**Locus tag**	**COG**	**KEGG PATHWAYS**
srfAA	7599	9128	Surfactin synthase subunit 1	BDCV01000001_04376	COG1020	Map01054
srfAB	59	2002	Surfactin synthase subunit 2	BDCV01000001_03279	COG1020	Map01054
srfAC	2039	5866	Surfactin synthase subunit 3	BDCV01000001_03280	COG1020	Map01054
srfAD	5895	6623	Surfactin synthase thioesterase subunit	BDCV01000001_03281	COG3208	
sfp	10934	11608	4-phosphopantetheinyl transferase	BDCV01000001_03287	COG2091	Map00770
ppsB	1893	9575	Plipastatin synthase subunit B	BDCV01000001_04373	COG1020	Map01054
ppsC	23	4144	Plipastatin synthase subunit C	BDCV01000001_03513	COG1020	Map01054
ppsD_1	4170	10268	Plipastatin synthase subunit D	BDCV01000001_03514	COG1020	Map01054
ppsD_2	10331	14968	Plipastatin synthase subunit D	BDCV01000001_03515	COG1020	Map01054
ppsE	14997	18815	Plipastatin synthase subunit E	BDCV01000001_03516	COG1020	Map01054
dhbF	101986	109122	Dimodular nonribosomal peptide synthase	BDCV01000001_01829	COG1020	Map01053
rmlA	719060	719800	Glucose-1-phosphate thymidylyltransferase	BDCV01000001_02440	COG1209	Map00521
rmlB	718113	719060	dTDP-glucose 4,6-dehydratase	BDCV01000001_02439	COG1088	Map00521
rmlC	716800	717255	dTDP-4-dehydrorhamnose 3,5-epimerase	BDCV01000001_02437	COG1088	Map00521
rmlD	717248	718099	dTDP-4-dehydrorhamnose reductase	BDCV01000001_02438	COG1091	Map00521
Genes involve in regulatory of surfactin						
comA	74243	74887	Transcriptional regulatory protein ComA	BDCV01000001_01802	COG2197	Map02020
comP	74968	77280	Sensor histidine kinase ComP	BDCV01000001_01803	COG4585	Map02020
spo0A_1	299321	300124	Stage 0 sporulation protein A	BDCV01000001_01171	COG0784	Map02020
spo0A_2	88986	89930	Stage 0 sporulation protein A	BDCV01000001_03141	COG0784	Map02020
abrB	45379	45663	Transition state regulatory protein AbrB	BDCV01000001_04229	COG2002	
resD	198670	199392	Transcriptional regulatory protein SrrA	BDCV01000001_01057	COG0745	Map02020
liaR_1	215890	216525	Transcriptional regulatory protein LiaR	BDCV01000001_01942	COG2197	Map02020
liaR_2	828495	829151	Transcriptional regulatory protein LiaR	BDCV01000001_02549	COG2197	Map02020
liaR_3	10590	11237	Transcriptional regulatory protein LiaR	BDCV01000001_03818	COG2197	Map02020
sigA	380481	381596	RNA polymerase sigma factor SigA	BDCV01000001_01271	COG0568	Map05111
rpoN	332153	333463	RNA polymerase sigma-54 factor	BDCV01000001_02055	COG1508	Map02020
csrA	456099	456323	Carbon storage regulator	BDCV01000001_02175	COG1551	Map02020
dnaK	406380	408215	Chaperone protein DnaK	BDCV01000001_01299	COG0443	Map03018
lytR_1	751964	752689	Sensory transduction protein LytR	BDCV01000001_01671	COG3279	Map02020
lytR_2	485622	486542	Transcriptional regulator LytR	BDCV01000001_02204	COG1316	
lytR_3	117850	118551	Sensory transduction protein LytR	BDCV01000001_03170	COG3279	

In addition, genes implicated in regulation of surfactin; *comA* and *comP* which comprise a signal transduction system that involved in the competence development pathway and is required for the transcription of *srfA* (Marahier et al., 1993; m. Nakano et al., 1992) were also identified. The remaining of the genes are involved in DNA-binding response, sporulation, phosphate regulon transcription, carbon storage and sensory transduction protein (Table 6.1).

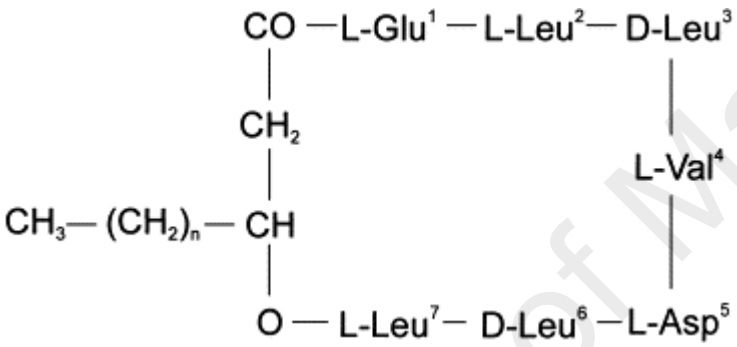

Figure 6.1: Amino acids structure in surfactin. Comprise of 7 amino acids (Carrillo et al., 2003)

Figure 6.2: Surfactin genes organisation. The figure demonstrate *srfAA* operon and the amino acids synthesis by each gene (Soberón-Chávez, 2010).

6.3 Discussion

The results presented in this chapter showed the presence of biosurfactant genes in UMX-103, thus confirming that UMX-103 has the ability to produce surfactin which is lipopeptide biosurfactant. These results are in agreement with the earlier screening assays which were conducted in this study (Chapter 4). NRPS are diverse family of natural products with a broad range of biological activities and pharmacological properties. They include toxins, siderophores, pigments, antibiotics, cytostatics, and immunosuppressants. NRPS products have remarkably diverse structures and can be linear or cyclic or have branched structures. They can be further reengineered to produce complex products with exotic chemical structures and biological activities (Wang et al., 2014).

In *Bacillus* species, starvation leads to the activation of a number of processes that affect the ability to survive during periods of nutritional stress. Activities that are induced include the development of genetic competence, sporulation, the synthesis of degradative enzymes, motility, and antibiotic production. The genes that function in these processes are activated during the transition from exponential to stationary phase and are controlled by mechanisms that operate primarily at the level of transcription initiation. One class of genes functions in the synthesis of special metabolites such as the cyclic lipopeptide surfactin. These genes include the *srfA* operon of *Bacillus subtilis* which encodes the enzymes of the surfactin synthetase complex (Marahier et al., 1993).

The transcription of *srfA* depends on the positive control of *comA, comP and spo0A.* the main role of *comA* and *comP* in surfactin production and competence development is to activate *srfA* transcription. The gene *comA* encodes a response regulatory protein which contains helix-turn-helix motif characteristics of DNA-binding proteins (Marahier et al., 1993). While *comP* encodes a putative membrane protein with

sequence similarity to the histidine protein kinase class of two-component regulatory proteins (Marahier et al., 1993). These two genes involve in two-component signal transduction systems (KEGG: Map02020) which enable bacteria to sense, respond, and adapt to changes in their environment or in their intracellular state. Each two-component system consists of a sensor protein-histidine kinase (HK) and a response regulator (RR). In the prototypical two-component pathway, the sensor HK phosphorylates its own conserved histidine residue in response to a signal(s) in the environment. Subsequently, the phosphoryl group of HK is transferred onto a specific Aspartic acids (Asp) residue on the RR. The activated RR can then effect changes in cellular physiology, often by regulating gene expression. Two-component pathways thus often enable cells to sense and respond to stimuli by inducing changes in transcription (Figure 6.3).

The activation of *srfA* requires two response regulators which can be through direct interaction with *srfA* promoter or by indirect interaction way. In the case of *spo0A* (KEGG: Map02020), it is possible that it is required to activate the expression of a gene encoding another sensor kinase that catalyses *comA*-phosphate formation (Figure 6.4), or *spo0A* could interact in some way with *comA* protein forming a novel heterodimer (Marahier et al., 1993). The initial gene expression for development of spores in *B. subtilis* is regulated by *spo0A* transcriptional factor. This protein requires phosphorylation to be active and the level of its phosphorylation determines whether a cell will divide or sporulate (Burbulys et al., 1991). Phosphorylated *spo0A* activates the sporulation gene transcription and represses the transcription of the *abrB* gene, which encodes the "transition state" regulator *abrB* protein. The gene *spo0A*-mediated repression of *abrB* leads to depletion of the *abrB* protein from the cell and activation of genes under this negative control protein (Strauch et al., 1990).

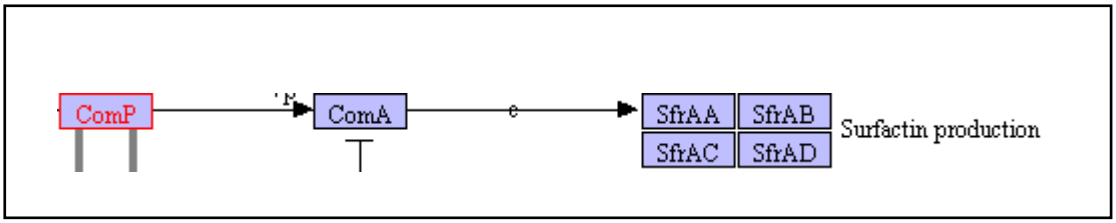

Figure 6.3: Pathway map (Map02020) from KEGG database of *comP* and *comA* in surfactin production. Where its shows *comP* function as sensor histidine kinase and *comA* functions as transcriptional regulatory protein for *srfA* genes.

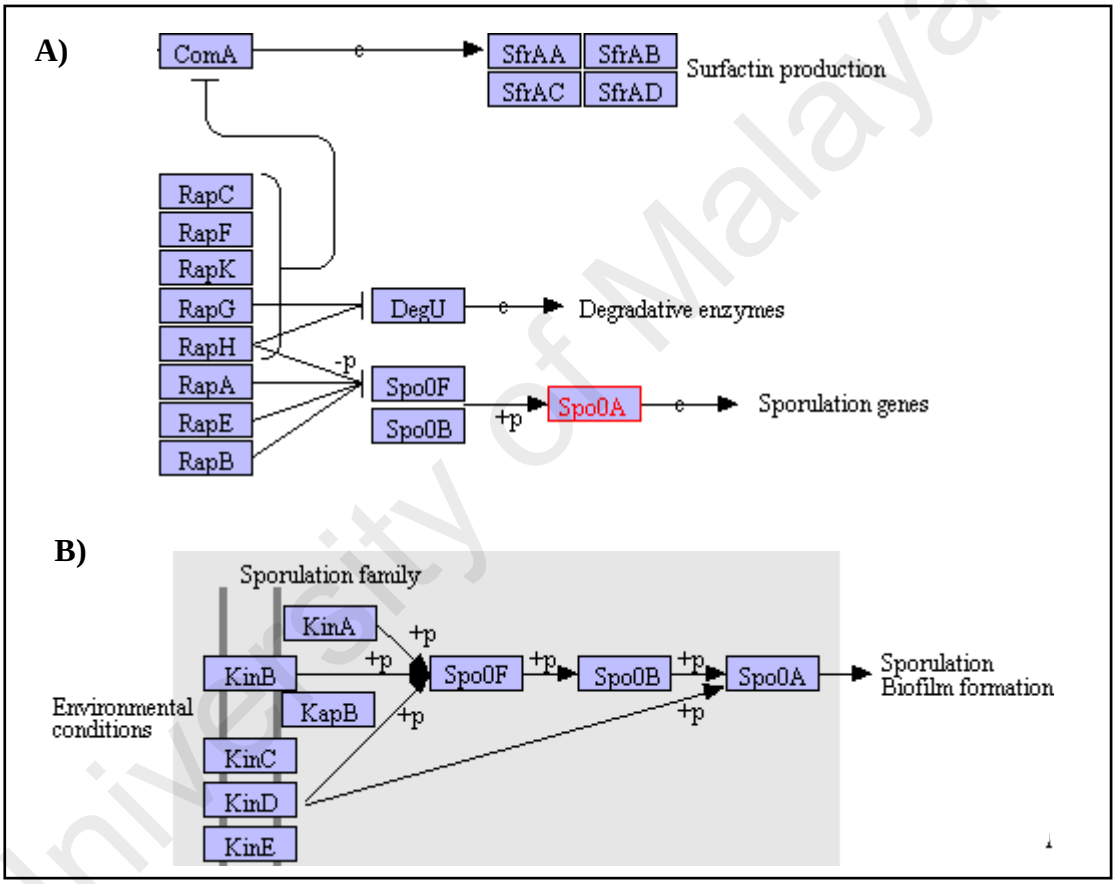

Figure 6.4: Pathway map (Map02020) from KEGG database of sporulation gene *spo0A*. (A) Shows *spo0A* is required to activate the expression of a gene encoding another sensor kinase that catalyses *comA*-phosphate formation which is *degU (Marahier et al., 1993).* (B) Shows the activation of *spo0A* under environmental conditions.

The identified genes which involved in biosurfactant production were compared with closely related genomes of environmental isolates of *Bacillus* strains (Table 2). Two operons were found to be present only in UMX-103 and *B. subtilis* 168, while they are absent in the other strains. These operons are *srfA* and *pps*, the *srfA* operon contains four genes (*srfAA, srfAB, srfAC, srfAD*). The operon *pps* contains four genes (*ppsB,ppsC,ppsD,ppsE*). These operons involve in biosurfactant synthesis. Among these genes there are few genes that only present in UMX-103. These genes are *rmlA, rmlB, rmlC and rmlD*. The genes *comA, comP, rpoN, abrB and ResD* are presented in both UMX-103 and *B. subtilis* 168. Three genes are presented in all the *Bacillus* strains which are *sigA, DnaK and LytR* (Table 6.2).

In summary, surfactin biosynthesis is triggered by environment condition where there are few number of genes regulating the production of surfactin. The operon *srfA* which contains four genes is responsible for surfacin production. These genes are regulated by *comP* and *comA*. The gene *comP* functions as histidine kinase sensor and it is regulating the gene *comA* which initiate the transcription of the genes inside the operon *srfA*. The gene that transforms the surfactin synthetase protein to active form is *sfp*, this shows the essential role of this gene in surfactin production. As the production of surfactin is related to environment condition the gene *spo0A* is important player in sporulation and regulation of *comA*.

Table 6.2: Compression of UMX-103 biosurfactants genes with closely related *Bacillus* strains.

RefSeq		NZ_CP011101.1	NZ_AP012496.1	NZ_CP011115.1	NC_000964.3	NC_017195.1
Gene name	*Bacillus subtilis* UMX-103	*Bacillus subtilis* LM 4-2	*Bacillus subtilis* BEST7003	*Bacillus subtilis* KCTC1028	*Bacillus subtilis* 168	*Bacillus subtilis* RO-NN-1
srfAA	•	X	X	X	•	X
srfAB	•	X	X	X	•	X
srfAB	•	X	X	X	•	X
srfAD	•	X	X	X	•	X
Spf	•	X	X	X	•	X
ppsB	•	X	X	X	•	X
ppsC	•	X	X	X	•	X
ppsD	•	X	X	X	•	X
ppsE	•	X	X	X	•	X
dhbF	•	X	X	X	•	X
rmlA	•	X	X	X	•	X
rmlB	•	X	X	X	X	X
rmlC	•	X	X	X	X	X
rmlD	•	X	X	X	X	X
comA	•	X	X	X	•	X
comP	•	X	X	X	•	X
ResD	•	X	X	X	•	X
LiaR	•	X	X	X	•	X
spo0A	•	X	•	•	•	•
rpoN	•	X	X	X	•	X
crsA	•	X	X	X	X	X
sigA	•	•	•	•	•	•
abrB	•	X	X	X	•	X
DnaK	•	•	•	•	•	•
LytR	•	•	•	•	•	•

(•) = Biosurfactants genes present in the bacterial genome; (X) Biosurfactants genes absent in the bacterial genome

CHAPTER 7: GENERAL DISCUSSION AND CONCLUSION

7.1 General discussion

Biosurfactants producers create a diverse varity group of extracellular surfactants and are known to occur in a range of chemical structures, such as glycolipids, lipopeptides and lipoproteins, phospholipids and fatty acids, polymeric and particulate structures. These bioactive molecule producers have many features which makes them potential alternative to the chemical surfactants. The production of biosurfactants has gained attention in the past years due to their commercial potentials. The majority of biosurfactants surface tension activity are impervious to many ecological factors such as pH and temperature (Vijayakumar & Saravanan, 2015). The biosurfactant lichenysin produced from *Bacillus licheniformis* strain JF-S was reported that it could resists temperature up to 50^0C and pH among 4.5 and 9.0 (McInerney et al., 1990). The availability of raw materials and their specificity make them amongst the most preferred surfactants.

Biosurfactnats are known for their high biodegradability and foaming properties, low toxicity level and their stability under high temperature and pH levels (Shoeb et al., 2015). Distinguished chemical synthetic surfactants which are generally obtained from petroleum feedstock, alternatively these surfactants can be produced by microbial fermentation procedures using waste materials and low cost agro based substrates (Al-Bahry et al., 2013). Various types of biosurfactants have different properties depending on the biosurfactant producers. The properties of different biosurfactants producers have been extensively studied (Abdel-Mawgoud et al., 2008; Cai et al., 2015; Joshi et al., 2012; Shoeb et al., 2015).

Among the various bacterial genera, the genus *Bacillus* has already been known to have a biosynthetic potential to produce a range of antimicrobial compounds. Interestingly, they are also reported to have a genomic basis for the biosynthesis of nonribosomal peptide derivatives due to the presence of nonribosomal peptide-synthetase (NRPS). So newly identified *Bacillus* species with genes already known for antimicrobial compound biosynthesis can be expected to have the same or novel product coded by the same gene cluster in a strain-specific manner. Exploring the chemical basis of this can have many biocontrol applications. Bacteria belonging to the species *Bacillus* have been reported to produce secondary metabolites of the class lipopeptides which have broad bioactivity. The biosynthesis of these metabolites is mediated by nonribosomal peptide synthase

The advance NGS technologies are based on sequencing entire DNA of a given genome randomly. Basically, this is conduct by fragmenting the entire genome into DNA fragments, then ligating those fragments of DNA to designated adapters for random read during DNA synthesis, this mehod is also known as sequencing by synthesis. Thus, NGS technology is called as massively parallel sequencing. The NGS technology generates reads, this reads refer to the actual number bases sequenced. NGS generates shorter reads compared to Sanger sequencing, as it provides reads length between 50 to 500 basepire, therefore, the sequencing results known as short reads. However, evolving NGS technologies such as single-molecule sequencing able to generate longer reads than Sanger methods. Due to the short length of reads generate by current NGS technology, determining of the sequencing coverage is essential for high quality sequencing. Coverage is defined as the number of short reads that overlap each other within a specific genomic region. Appropriate coverage is critical for accurate assembly of the genomic sequence (Zhang et al., 2011).

After NGS reads have generated, the reads are mapped to a known reference sequence or assembled using *de novo* approach. The decision to use either approach depends on the intended biological analysis as well as cost, effort and time considerations (Metzker, 2010). Also it depends on the complexity of the project for example small genomes such as bacteria and virus are less complex and faster to perform sequence mapping or alignment compared to large genomes such as human and mammals where they are more complex and require more time to compute mapping to the reference genome (Kulski, 2015). In the case if there is no reference genome available to map with the generated reads by NGS platforms usually *de novo* assembly approach is used (Metzker, 2010). *De novo* assemblies have been reported for bacterial genomes (Kamada et al., 2014; Shaligram et al., 2016; Yan et al., 2016). The accuracy of *de novo* assembly can be confirmed or improved by integrating it with comparative alignment mapping to reference genomic sequences. Sequencing assemblers may employ different graph construction algorithms and pre-processing and post-processing filter computations to flag, correct, or eliminate sequencing errors with no single computation solution (Kulski, 2015). Another way to improve the quality of sequencing and assembly is to apply a hybrid approach by using two or more different sequencing platforms such as combing short reads from Illumina sequencer platform with long reads form PacBio sequencer platform (Kamada et al., 2014).

In this study, the new strain UMX-103 belongs to *Bacillus subtilis* species. It is a Gram positive bacteria, rod shape and with a length of 1.954 µm and a diameter of 540.9 nm. All the five different biosurfactant producing tests showed that UMX-103 has the capability to produce biosurfactant. In addition, the genome of the strain UMX-103 was successfully assembled using a combination of both *de novo* and reference-guided assembly methods. The genome was assembled into 39 scaffolds with a size of 4,234,627 bp.Interestingly, 25 genes were identified which are involved in biosurfactant

production, where 14 genes involved in biosynthesis and 11 genes associated with the gene regulation. Genomic analysis revealed that UMX-103 has the genes which promote biosurfactant production. Future work will be conducted to characterize the unknown function genes as well as biosurfactant genes using various omics approaches.

7.2 Conclusion

This thesis report a genomic analysis of a newly identified *B. subtilis* strain isolated from hydrocarbon contaminated site in Malaysia. This strain was identified to produce biosurfactant. The determination of biosurfactant production by this strain was conducted using the latest known biochemical approaches. The whole genome sequencing analysis revealed the novelty of the new strain and the capability of the strain to produce biosurfactant. Also it revealed the functional features of the strain.

REFERENCES

Abdelhafiz, Y. A., Manaharan, T., BinMohamad, S., & Merican, A. F. (2017). Draft Genome Sequence of a Biosurfactant-Producing *Bacillus subtilis* UMX-103 Isolated from Hydrocarbon-Contaminated Soil in Terengganu, Malaysia. *Current Microbiology, 74*(7), 803-805.

Abdel-Mawgoud, A. M., Aboulwafa, M. M., & Hassouna, N. A. H. (2008). Optimization of surfactin production by *Bacillus subtilis* isolate BS5. *Applied Biochemistry and Biotechnology, 150*(3), 305-325.

Al-Bahry, S., Al-Wahaibi, Y., Elshafie, A., Al-Bemani, A., Joshi, S., Al-Makhmari, H., & Al-Sulaimani, H. (2013). Biosurfactant production by *Bacillus subtilis* B20 using date molasses and its possible application in enhanced oil recovery. *International Biodeterioration & Biodegradation, 81*, 141-146.

Alvarez, V. M., Jurelevicius, D., Marques, J. M., de Souza, P. M., de Araújo, L. V., Barros, T. G., . . . Seldin, L. (2015). *Bacillus amyloliquefaciens* TSBSO 3.8, a biosurfactant-producing strain with biotechnological potential for microbial enhanced oil recovery. *Colloid Surface B, 136*, 14-21.

Amaral, P., Da Silva, J., Lehocky, M., Barros-Timmons, A., Coelho, M., Marrucho, I., & Coutinho, J. (2006). Production and characterization of a bioemulsifier from *Yarrowia lipolytica*. *Process Biochemistry, 41*(8), 1894-1898.

Anuradha S, N. (2010). Structural and molecular characteristics of lichenysin and its relationship with surface activity. *Biosurfactants*, 304-315.

Arima, K., Kakinuma, A., & Tamura, G. (1968). Surfactin, a crystalline peptidelipid surfactant produced by *Bacillus subtilis*: Isolation, characterization and its inhibition of fibrin clot formation. *Biochemical and Biophysical Research Communications, 31*(3), 488-494.

Baltz, R. H., Miao, V., & Wrigley, S. K. (2005). Natural products to drugs: daptomycin and related lipopeptide antibiotics. *Natural Product Reports, 22*(6), 717-741.

Banat, I. M. (1993). The isolation of a thermophilic biosurfactant producing *Bacillus sp*. *Biotechnology Letters, 15*(6), 591-594.

Banat, I. M., Franzetti, A., Gandolfi, I., Bestetti, G., Martinotti, M. G., Fracchia, L., . . . Marchant, R. (2010). Microbial biosurfactants production, applications and future potential. *Applied Microbiology and Biotechnology, 87*(2), 427-444.

Banat, I. M., Makkar, R. S., & Cameotra, S. (2000). Potential commercial applications of microbial surfactants. *Applied Microbiology and Biotechnology, 53*(5), 495-508.

Barbe, V., Cruveiller, S., Kunst, F., Lenoble, P., Meurice, G., Sekowska, A., . . . Médigue, C. (2009). From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later. *Microbiology, 155*(6), 1758-1775.

Bernheimer, A. W., & Avigad, L. S. (1970). Nature and properties of a cytolytic agent produced by *Bacillus subtilis*. *Microbiology*, *61*(3), 361-369.

Binnewies, T. T., Motro, Y., Hallin, P. F., Lund, O., Dunn, D., La, T., ... & Ussery, D. W. (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Functional & Integrative Genomics*, *6*(3), 165-185.

Bodour, A. A., Drees, K. P., & Maier, R. M. (2003). Distribution of biosurfactant-producing bacteria in undisturbed and contaminated arid southwestern soils. *Applied and Environmental Microbiology*, *69*(6), 3280-3287.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, *27*(4), 578-579.

Boetzer, M., & Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome Biology*, *13*(6), R56.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Buermans, H., & Den Dunnen, J. (2014). Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1842*(10), 1932-1941.

Burbulys, D., Trach, K. A., & Hoch, J. A. (1991). Initiation of sporulation in *B. subtilis* is controlled by a multicomponent phosphorelay. *Cell*, *64*(3), 545-552.

Cai, Q., Zhang, B., Chen, B., Song, X., Zhu, Z., & Cao, T. (2015). Screening of biosurfactant-producing bacteria from offshore oil and gas platforms in North Atlantic Canada. *Environmental Monitoring and Assessment*, *187*(5), 284.

Carrillo, C., Teruel, J. A., Aranda, F. J., & Ortiz, A. (2003). Molecular mechanism of membrane permeabilization by the peptide antibiotic surfactin. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1611*(1), 91-97.

Chaudhari, N. M., Gupta, V. K., & Dutta, C. (2016). BPGA-an ultra-fast pan-genome analysis pipeline. *Scientific Reports*, *6*, 24373.

Choudhary, D. K., & Johri, B. N. (2009). Interactions of *Bacillus spp*. and plants–with special reference to induced systemic resistance (ISR). *Microbiological Research*, *164*(5), 493-513.

Cirigliano, M. C., & Carman, G. M. (1985). Purification and characterization of liposan, a bioemulsifier from *Candida lipolytica*. *Applied and Environmental Microbiology*, *50*(4), 846-850.

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews*, *17*(4), 840-862.

Cosmina, P., Rodriguez, F., Ferra, F., Grandi, G., Perego, M., Venema, G., & Sinderen, D. (1993). Sequence and analysis of the genetic locus responsible for surfactin synthesis in *Bacillus subtilis*. *Molecular Microbiology, 8*(5), 821-831.

Coutte, F., Leclère, V., Béchet, M., Guez, J. S., Lecouturier, D., Chollet Imbert, M., . . . Jacques, P. (2010). Effect of pps disruption and constitutive expression of srfA on surfactin productivity, spreading and antagonistic properties of *Bacillus subtilis* 168 derivatives. *Journal of Applied Microbiology, 109*(2), 480-491.

Dadrasnia, A., & Ismail, S. (2015). Biosurfactant production by *Bacillus salmalaya* for lubricating oil solubilization and biodegradation. *International Journal of Environmental Research and Public Health, 12*(8), 9848-9863.

Daly, A. K. (2010). Genome-wide association studies in pharmacogenomics. *Nature Reviews Genetics, 11*(4), 241-246.

Darling, A. C., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research, 14*(7), 1394-1403.

Das, D., Baruah, R., Roy, A. S., Singh, A. K., Boruah, H. P. D., Kalita, J., & Bora, T. C. (2015). Complete genome sequence analysis of *Pseudomonas aeruginosa* N002 reveals its genetic adaptation for crude oil degradation. *Genomics, 105*(3), 182-190.

Davis, D., Lynch, H., & Varley, J. (2001). The application of foaming for the recovery of surfactin from *B. subtilis* ATCC 21332 cultures. *Enzyme and Microbial Technology, 28*(4), 346-354.

Desai, J. D., & Banat, I. M. (1997). Microbial production of surfactants and their commercial potential. *Microbiology and Molecular Biology Reviews, 61*(1), 47-64.

Dexter, A. F., & Middelberg, A. P. (2008). Peptides as functional surfactants. *Industrial & Engineering Chemistry Research, 47*(17), 6391-6398.

Dhillon, B. K., Laird, M. R., Shay, J. A., Winsor, G. L., Lo, R., Nizam, F., ... & Brinkman, F. S. (2015). IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Research, 43*(W1), W104-W108.

Dobrindt, U., Hochhut, B., Hentschel, U., & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology, 2*(5), 414-424.

Doroghazi, J. R., Albright, J. C., Goering, A. W., Ju, K. S., Haines, R. R., Tchalukov, K. A., . . . Metcalf, W. W. (2014). A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nature Chemical Biology, 10*(11), 963-968.

England, P. H. (2015). UK Standards for Microbiology Investigations ID 9: Identification of *Bacillus* species. *Public Health England*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/407156/ID_9i3.pdf

Escalona, M., Rocha, S., & Posada, D. (2016). A comparison of tools for the simulation of genomic next-generation sequencing data. *Nature Reviews Genetics, 17*(8), 459-469.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., . . . Van den Berghe, A. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature, 260*(5551), 500-507.

Gautam, K., & Tyagi, V. (2006). Microbial surfactants: a review. *Journal of Oleo Science, 55*(4), 155-166.

Gudiña, E. J., Fernandes, E. C., Rodrigues, A. I., Teixeira, J. A., & Rodrigues, L. R. (2015). Biosurfactant production by *Bacillus subtilis* using corn steep liquor as culture medium. *Frontiers in Microbiology, 6*, 59.

Gudina, E. J., Pereira, J. F., Costa, R., Coutinho, J. A., Teixeira, J. A., & Rodrigues, L. R. (2013). Biosurfactant-producing and oil-degrading *Bacillus subtilis* strains enhance oil recovery in laboratory sand-pack columns. *Journal of Hazardous Materials, 261*, 106-113.

Gurjar, M., Khire, J., & Khan, M. (1995). Bioemulsifier production by *Bacillus stearothermophilus* VR8 isolate. *Letters in Applied Microbiology, 21*(2), 83-86.

Heerklotz, H., & Seelig, J. (2007). Leakage and lysis of lipid membranes induced by the lipopeptide surfactin. *European Biophysics Journal, 36*(4-5), 305-314.

Hocquette, J., Cassar-Malek, I., Scalbert, A., & Guillou, F. (2009). Contribution of genomics to the understanding of physiological functions. *Journal of Physiology and Pharmacology, 60*(3), 5-16.

Hong, H. A., & Cutting, S. M. (2005). The use of bacterial spore formers as probiotics. *FEMS Microbiology Reviews, 29*(4), 813-835.

Horner, D. S., Pavesi, G., Castrignanò, T., De Meo, P. D. O., Liuni, S., Sammeth, M., . . . Pesole, G. (2010). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics, 11*(2), 181-197.

Hsiao, W., Wan, I., Jones, S. J., & Brinkman, F. S. (2003). IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics, 19*(3), 418-420.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through

Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution, 34*(8), 2115-2122.

Hyatt, D., Chen, G.L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics, 11*(1), 119.

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology, 45*(9), 2761-2764.

Johansson, I., & Svensson, M. (2001). Surfactants based on fatty acids and other natural hydrophobes. *Current Opinion in Colloid & Interface Science, 6*(2), 178-188.

Joshi, S. J., Suthar, H., Yadav, A. K., Hingurao, K., & Nerurkar, A. (2012). Occurrence of biosurfactant producing *Bacillus spp*. in diverse habitats. *ISRN Biotechnology, 2013*.

Kamada, M., Hase, S., Sato, K., Toyoda, A., Fujiyama, A., & Sakakibara, Y. (2014). Whole genome complete resequencing of *Bacillus subtilis* natto by combining long reads with high-quality short reads. *PloS One, 9*(10), e109999.

Käppeli, O., & Finnerty, W. (1979). Partition of alkane by an extracellular vesicle derived from hexadecane-grown *Acinetobacter*. *Journal of Bacteriology, 140*(2), 707-712.

Kim, M., Oh, H.-S., Park, S.-C., & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International journal of Systematic and Evolutionary Microbiology, 64*(2), 346-351.

Kosaric, N., & Sukan, F. V. (2014). *Biosurfactants: Production and Utilization—Processes, Technologies, and Economics*: CRC Press.

Jerzy K. Kulski (2016). Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications, Next Generation Sequencing - Advances, Applications and Challenges, Dr. Jerzy Kulski (Ed.), InTech, DOI: 10.5772/61964. Available from: https://www.intechopen.com/books/next-generation-sequencing-advances-applications-and-challenges/next-generation-sequencing-an-overview-of-the-history-tools-and-omic-applications

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular biology and evolution*, *33*(7), 1870-1874.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A., Alloni, G., Azevedo, V., . . . Borchert, S. (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature, 390*(6657), 249-256.

Kuska, B. (1998). Beer, Bethesda, and biology: how "genomics" came into being. *Journal of the National Cancer Institute, 90*(2), 93-93.

Kuyukina, M. S., Ivshina, I. B., Baeva, T. A., Kochina, O. A., Gein, S. V., & Chereshnev, V. A. (2015). Trehalolipid biosurfactants from nonpathogenic Rhodococcus actinobacteria with diverse immunomodulatory activities. *New Biotechnology, 32*(6), 559-568.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . FitzHugh, W. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921.

Ladner, J. T., Beitzel, B., Chain, P. S., Davenport, M. G., Donaldson, E., Frieman, M., . . . Sabeti, P. C. (2014). Standards for sequencing viral genomes in the era of high-throughput sequencing. *MBio, 5*(3), e01360-01314.

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research, 35*(9), 3100-3108.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., . . . Wassenaar, T. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics, 15*(2), 141-161.

Landman, D., Georgescu, C., Martin, D. A., & Quale, J. (2008). Polymyxins revisited. *Clinical Microbiology Reviews, 21*(3), 449-465.

Langille, M. G., & Brinkman, F. S. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics, 25*(5), 664-665.

Langille, M. G., Hsiao, W. W., & Brinkman, F. S. (2008). Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics*, *9*(1), 329.

Langille, M. G., Hsiao, W. W., & Brinkman, F. S. (2010). Detecting genomic islands using bioinformatics approaches. *Nature Reviews Microbiology, 8*(5), 373-382.

Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., … Lund, O. (2012). Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria. *Journal of Clinical Microbiology*, *50*(4), 1355–1361.

Lasken, R. S., & McLean, J. S. (2014). Recent advances in genomic DNA sequencing of microbial species from single cells. *Nature Reviews Genetics*, *15*(9), 577-584.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Preprint arXiv:1303.3997*.

Li, K. B. (2003). ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics, 19*(12), 1585-1586.

Loman, N. J., & Pallen, M. J. (2015). Twenty years of bacterial genome sequencing. *Nature Reviews. Microbiology*, *13*(12), 787

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research, 25*(5), 0955-0964.

Marahier, M., Nakano, M., & Zuber, P. (1993). Regulation of peptide antibiotic production in *Bacillus*. *Molecular Microbiology, 7*(5), 631-636.

Mardis, E. R. (2017). DNA sequencing technologies: 2006-2016. *Nature Protocols, 12*(2), 213-218.

Mardis, E. R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics, 9*(1), 387-402.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., . . . Chen, Z. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*(7057), 376-380.

May, J. J., Wendrich, T. M., & Marahiel, M. A. (2001). The dhb operon of *Bacillus subtilis* Encodes the biosynthetic template for the catecholic siderophore 2, 3-dihydroxybenzoate-glycine-threonine trimeric ester *Bacillibactin. Journal of Biological Chemistry, 276*(10), 7209-7217.

McInerney, M. J., Javaheri, M., & Nagle Jr, D. P. (1990). Properties of the biosurfactant produced by *Bacillus licheniformis* strain JF-2. *Journal of Industrial Microbiology, 5*(2-3), 95-101.

Meena, K. R., & Kanwar, S. S. (2015). Lipopeptides as the antifungal and antibacterial agents: applications in food safety and therapeutics. *BioMed Research International, 2015*, 473050.

Meier, J. P., Auch, A. F., Klenk, H.-P., & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics, 14*(1), 60.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics, 11*(1), 31-46.

Morán, A. C., Martinez, M. A., & Siñeriz, F. (2002). Quantification of surfactin in culture supernatants by hemolytic activity. *Biotechnology Letters, 24*(3), 177-180.

Morikawa, M., Daido, H., Takao, T., Murata, S., Shimonishi, Y., & Imanaka, T. (1993). A new lipopeptide biosurfactant produced by *Arthrobacter sp*. strain MIS38. *Journal of Bacteriology, 175*(20), 6459-6466.

Morikawa, M., Hirata, Y., & Imanaka, T. (2000). A study on the structure–function relationship of lipopeptide biosurfactants. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 1488*(3), 211-218.

Mulligan, C. N. (2009). Recent advances in the environmental applications of biosurfactants. *Current Opinion In Colloid and Interface Science, 14*(5), 372-378.

Mulligan, C. N., Cooper, D. G., & Neufeld, R. J. (1984). Selection of microbes producing biosurfactants in media without hydrocarbons. *Journal of Fermentation Technology, 62*(4), 311-314.

Nakano, M., Magnuson, R., Myers, A., Curry, J., Grossman, A., & Zuber, P. (1991). srfA is an operon required for surfactin production, competence development, and efficient sporulation in *Bacillus subtilis*. *Journal of Bacteriology, 173*(5), 1770-1778.

Nakano, M. M., Corbell, N., Besson, J., & Zuber, P. (1992). Isolation and characterization of sfp: a gene that functions in the production of the lipopeptide biosurfactant, surfactin, in *Bacillus subtilis. Molecular and General Genetics, 232*(2), 313-321.

Nishito, Y., Osana, Y., Hachiya, T., Popendorf, K., Toyoda, A., Fujiyama, A., . . . Sakakibara, Y. (2010). Whole genome assembly of a natto production strain *Bacillus subtilis* natto from very short read data. *BMC Genomics, 11*(1), 1.

Nitschke, M., & Costa, S. (2007). Biosurfactants in food industry. *Trends in Food Science & Technology, 18*(5), 252-259.

Noah, K. S., Fox, S. L., Bruhn, D. F., Thompson, D. N., & Bala, G. A. (2002). Development of continuous surfactin production from potato process effluent by *Bacillus subtilis* in an airliftreactor. In *Biotechnology for Fuels and Chemicals* (pp. 803-813). Humana Press.

Ongena, M., & Jacques, P. (2008). *Bacillus* lipopeptides: versatile weapons for plant disease biocontrol. *Trends in Microbiology, 16*(3), 115-125.

Pacwa-Płociniczak, M., Płaza, G. A., Piotrowska-Seget, Z., & Cameotra, S. S. (2011). Environmental applications of biosurfactants: recent advances. *International Journal of Molecular Sciences, 12*(1), 633-654.

Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics, 10*(10), 669-680.

Pereira, J. F., Gudiña, E. J., Costa, R., Vitorino, R., Teixeira, J. A., Coutinho, J. A., & Rodrigues, L. R. (2013). Optimization and characterization of biosurfactant production by *Bacillus subtilis* isolates towards microbial enhanced oil recovery applications. *Fuel, 111*, 259-268.

Peypoux, F., Bonmatin, J., & Wallach, J. (1999). Recent trends in the biochemistry of surfactin. *Applied Microbiology and Biotechnology, 51*(5), 553-563.

Płaza, G., Chojniak, J., Rudnicka, K., Paraszkiewicz, K., & Bernat, P. (2015). Detection of biosurfactants in *Bacillus* species: genes and products identification. *Journal of Applied Microbiology, 119*(4), 1023-1034.

Rissman, A. I., Mau, B., Biehl, B. S., Darling, A. E., Glasner, J. D., & Perna, N. T. (2009). Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics, 25*(16), 2071-2073.

Rosenberg, E., Zuckerberg, A., Rubinovitz, C., & Gutnick, D. (1979). Emulsifier of *Arthrobacter* RAG-1: isolation and emulsifying properties. *Applied and Environmental Microbiology, 37*(3), 402-408.

Sari, M., Kusharyoto, W., & Artika, I. M. (2014). Screening for Biosurfactant-producing Yeast: Confirmation of Biosurfactant Production. *Biotechnology, 13*(3), 106.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics, 30*(14), 2068-2069.

Sen, R. (2008). Biotechnology in petroleum recovery: the microbial EOR. *Progress in Energy and Combustion Science, 34*(6), 714-724.

Sen, R., & Swaminathan, T. (2004). Response surface modeling and optimization to elucidate and analyze the effects of inoculum age and size on surfactin production. *Biochemical Engineering Journal, 21*(2), 141-148.

Shaligram, S., Kumbhare, S. V., Dhotre, D. P., Muddeshwar, M. G., Kapley, A., Joseph, N., . . . Pawar, S. P. (2016). Genomic and functional features of the biosurfactant producing *Bacillus sp.* AM13. *Functional & Integrative Genomics, 16*(5), 557-566.

Shoeb, E., Ahmed, N., Akhter, J., Badar, U., Siddiqui, K., ANSARI, F., . . . Shaikh, Q. U. A. (2015). Screening and characterization of biosurfactant-producing bacteria isolated from the Arabian Sea coast of Karachi. *Turkish Journal of Biology, 39*(2), 210-216.

Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next generation sequencing technologies for environmental DNA research. *Molecular Ecology, 21*(8), 1794-1805.

Soberón-Chávez, G. (Ed.). (2010). *Biosurfactants: from genes to applications* (Vol. 20). Springer Science & Business Media.

Soberón-Chávez, G., Lépine, F., & Déziel, E. (2005). Production of rhamnolipids by *Pseudomonas aeruginosa. Applied Microbiology and Biotechnology, 68*(6), 718-725.

Srivatsan, A., Han, Y., Peng, J., Tehranchi, A. K., Gibbs, R., Wang, J. D., & Chen, R. (2008). High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS Genetics, 4*(8), e1000139.

Stein, T. (2005). *Bacillus subtilis* antibiotics: structures, syntheses and specific functions. *Molecular Microbiology, 56*(4), 845-857.

Strauch, M., Webb, V., Spiegelman, G., & Hoch, J. A. (1990). The SpoOA protein of *Bacillus subtilis* is a repressor of the abrB gene. *Proceedings of the National Academy of Sciences, 87*(5), 1801-1805.

Syldatk, C., & Wagner, F. (1987). Production of biosurfactants. *Biosurfactants and Biotechnology, 25*, 89-120.

Transparency Market Research. (2011). Biosurfactant market global scenario, raw material and consumption trends, industry analysis, size, share and forecasts, 2011-2018. Retrieved from http://www.transparencymarketresearch.com/biosurfactants-market.html

Van Bogaert, I. N., Saerens, K., De Muynck, C., Develter, D., Soetaert, W., & Vandamme, E. J. (2007). Microbial production and application of sophorolipids. *Applied Microbiology and Biotechnology, 76*(1), 23-34.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next generation sequencing technology. *Trends in Genetics, 30*(9), 418-426.

Vater, J., Kablitz, B., Wilde, C., Franke, P., Mehta, N., & Cameotra, S. S. (2002). Matrix-assisted laser desorption ionization-time of flight mass spectrometry of lipopeptide biosurfactants in whole cells and culture filtrates of *Bacillus subtilis* C-1 isolated from petroleum sludge. *Applied and Environmental Microbiology, 68*(12), 6210-6219.

Vaz, D. A., Gudiña, E. J., Alameda, E. J., Teixeira, J. A., & Rodrigues, L. R. (2012). Performance of a biosurfactant produced by a *Bacillus subtilis* strain isolated from crude oil samples as compared to commercial chemical surfactants. *Colloids and Surfaces B: Biointerfaces, 89*, 167-174.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Holt, R. A. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351.

Venter, J. C., Levy, S., Stockwell, T., Remington, K., & Halpern, A. (2003). Massive parallelism, randomness and genomic advances. *Nature Genetics, 33*, 219-227.

Vijayakumar, S., & Saravanan, V. (2015). Biosurfactants-Types, Sources and Applications. *Research Journal of Microbiology, 10*(5), 181.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., . . . Chen, F. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature, 457*(7231), 854-858.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., . . . Merkl, R. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics, 7*(1), 142.

Walter, V., Syldatk, C., & Hausmann, R. (2010). Screening concepts for the isolation of biosurfactant producing microorganisms. *Biosurfactants, 672*, 1-13.

Wang, H., Fewer, D. P., Holm, L., Rouhiainen, L., & Sivonen, K. (2014). Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proceedings of the National Academy of Sciences, 111*(25), 9259-9264.

Wayman, M., Jenkins, A., & Kormady, A. (1984). Biotechnology for oil and fat industry. *Journal of American Oil Chemists Society, 61*, 129-131.

Woo, P., Lau, S., Teng, J., Tse, H., & Yuen, K. Y. (2008). Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinical Microbiology and Infection, 14*(10), 908-934.

Yan, Y., Zhang, L., Yu, M., Wang, J., Tang, H., Yang, Z., & Wan, P. (2016). The genome of *Bacillus aryabhattai* T61 reveals its adaptation to Tibetan Plateau environment. *Genes & Genomics, 38*(3), 293-301.

Yonebayashi, H., Yoshida, S., Ono, K., & Enomoto, H. (2000). Screening of microorganisms for microbial enhanced oil recovery processes. *Sekiyu Gakkai Shi, 43*(1), 59-69.

Youssef, N., Elshahed, M. S., & McInerney, M. J. (2009). Microbial processes in oil fields: culprits, problems, and opportunities. *Advances in Applied Microbiology, 66*, 141-251.

Youssef, N. H., Duncan, K. E., Nagle, D. P., Savage, K. N., Knapp, R. M., & McInerney, M. J. (2004). Comparison of methods to detect biosurfactant production by diverse microorganisms. *Journal of Microbiological Methods, 56*(3), 339-347.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research, 18*(5), 821-829.

Zhang, J., Chiodini, R., Badr, A., & Zhang, G. (2011). The impact of next generation sequencing on genomics. *Journal of Genetics And Genomics, 38*(3), 95-109.

Zhou, X., Ren, L., Li, Y., Zhang, M., Yu, Y., & Yu, J. (2010). The next generation sequencing technology: a technology review and future perspective. *Science China Life Sciences, 53*(1), 44-57.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

1. Abdelhafiz, Y.A., Manaharan, T., Mohamad, S.B., and Merican, A.F. (2017) Whole genome sequencing and functional features of UMX-103: a new *Bacillus* strain with biosurfactant producing capability. *Genes & Genomics* 39: 877-886.

2. Abdelhafiz, Y. A., Manaharan, T., BinMohamad, S., & Merican, A. F. (2017). Draft Genome Sequence of a biosurfactant-producing *Bacillus subtilis* UMX-103 isolated from hydrocarbon-contaminated soil in Terengganu, Malaysia. *Current Microbiology*, *74*(7), 803-805.