

**A GENOMIC ANALYSIS OF *Paenibacillus  
macerans* ATCC 8244, A GRAM POSITIVE  
NITROGEN FIXING BACTERIUM**

**AHMAD YAMIN BIN ABDUL RAHMAN**

**UNIVERSITI SAINS MALAYSIA**

**2016**

**A GENOMIC ANALYSIS OF *Paenibacillus  
macerans* ATCC 8244, A GRAM POSITIVE  
NITROGEN FIXING BACTERIUM**

by

**AHMAD YAMIN BIN ABDUL RAHMAN**

**Thesis submitted in fulfilment of the requirements  
for the degree of  
Master of Science**

**September 2016**

## ACKNOWLEDGEMENT

First and foremost, Alhamdulillah, all praise be to Allah who with His compassion enabled me to finish my thesis. This thesis not only represents scientific findings of *Paenibacillus macerans* through a genomics and bioinformatics point of view but it is also a proof of that a simple hobby/interest can actually drive scientific research. More importantly, this thesis is my second chance to put my research life back on track. I was at a juncture of turbulence to choose between having a career or to further studies when Prof Nazalan Najimudin offered me the golden opportunity to be under his supervision. Honestly, I am very grateful for his compassion, flexibility, perseverance and encouragement to mold me from an average computer enthusiast to a scientist with a more constructive direction in the field of bioinformatics.

My deepest thanks to Dr Rashidah Abdul Rahim, my ex-supervisor during my bachelor's degree studies who provided me the first step to delve into a bioinformatics-related field. Along the way, I met several key persons who changed my life and they included Prof Nazalan, Prof Razip Samian, and Prof Maqsudul Alam. Without the exposure and the motivational push they gave, I would not even know what the terms “scaffold”, “contig” or “annotation” were all about.

My ultimate thanks goes to my mother, Hafsah binti Yaakub, who has always been encouraging and supportive of my studying further, keeping me sane during my busy days, and constantly reminding me to write my thesis. Not forgetting my family who gave me relentless support emotionally, I must declare my gratitude to all of them. Last but not least, my friends, Kee Shin, Mardani and Haida.

# TABLE OF CONTENTS

<b>ACKNOWLEDGEMENT</b> .....	<b>ii</b>
<b>TABLE OF CONTENTS</b> .....	<b>iv</b>
<b>LIST OF TABLES</b> .....	<b>vii</b>
<b>LIST OF FIGURES</b> .....	<b>viii</b>
<b>LIST OF ABBREVIATIONS</b> .....	<b>x</b>
<b>ABSTRAK</b> .....	<b>xii</b>
<b>ABSTRACT</b> .....	<b>xiii</b>
<b>CHAPTER 1: INTRODUCTION</b> .....	<b>1</b>
1.1 Research Objectives: .....	3
<b>CHAPTER 2: LITERATURE REVIEW</b> .....	<b>4</b>
2.1 <i>Paenibacillus</i> genus.....	4
2.2 <i>Paenibacillus macerans</i> .....	5
2.3 Nitrogen Fixing Microorganisms .....	5
2.4 Importance Of Nitrogen .....	6
2.5 Nitrogen Cycle .....	7
2.6 Nitrogen Fixation Key Enzymes .....	9
2.6.1 MoFe protein .....	9
2.6.2 Fe Protein .....	9
2.7 The Chemical Equation and Stochoimetry of Nitrogen Fixation.....	10
2.8 Regulation Of Nitrogen Fixation.....	10
2.9 <i>Azotobacter vinelandii</i> as a Model of Nitrogen Fixing Organism .....	11
2.9.1 Genes Involved In Nitrogen Fixation of <i>A. vinelandii</i> . .....	12
2.10 Bioinformatics as a Fusion of Genomics and Computing.....	14
2.11 Next Generation Sequencing and Bioinformatics .....	15
2.11.1 Illumina Sequencing .....	15
2.11.2 Roche 454 Pyro Sequencing.....	16
2.11.3 Ion Torrent Sequencing .....	16
2.11.4 Pac Bio SMRT Sequencing .....	17
2.12 Genome Assembly .....	17
2.13 Gene Finding .....	18

2.14	Post Annotation; Analysis, Comparative and Evolutionary Studies.....	19
<b>CHAPTER 3: MATERIALS AND METHODS .....</b>		<b>21</b>
3.1	Schematic Workflow Of This Study .....	21
3.2	Bioinformatics Tools For Assembly, Annotation And Analysis .....	21
3.3	Genomic DNA isolation.....	23
3.4	Raw Sequence Pre-Processing .....	25
3.5	Insert Size Estimation.....	27
3.6	Assembler Benchmarks.....	27
3.7	Assembly Optimisation, Scaffolding And Gap Filling .....	28
3.7.1	Kmer Optimisation .....	28
3.7.2	Scaffolding .....	28
3.7.3	Gap filling.....	29
3.8	Genome Annotation .....	29
3.9	Metabolic Pathway reconstruction .....	30
3.10	Phylogenomics analysis .....	30
3.11	Pan Genome analysis .....	31
3.12	Genomic Island Analysis .....	31
3.13	CRISPR Identification.....	32
<b>CHAPTER 4: RESULTS .....</b>		<b>33</b>
4.1	Quality Control Processing of Raw Data .....	33
4.2	Empirical estimation of insert size .....	33
4.3	Performance of Genome Assembler Software on Assembly Data .....	37
4.4	Assembly Optimisation, Scaffolding And Gap Filling .....	37
4.4.1	Assessment To Get The Best K Value For K-Mer Selection.....	38
4.5	Final assembly.....	40
4.6	General Features of the <i>P. macerans</i> ATCC 8244 Draft Genome.....	41
4.7	SEED Classification.....	41
4.8	Identification of Nitrogen Fixation and Associated Genes .....	44
4.9	Phylogenomic Analysis.....	45
4.10	Pan Genomic Analysis .....	50
4.11	Metabolic Pathway Reconstruction.....	53
4.12	Genomic Island Analysis .....	58
4.13	Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) Detection .....	58

<b>CHAPTER 5: DISCUSSION .....</b>	<b>61</b>
5.1 Nitrogen fixation operon in <i>Paenibacillus macerans</i> ATCC 8244 .....	61
5.2 Bisphenol A (BPA) Degradation.....	63
5.3 Biofuel Potential.....	64
5.4 Genome-based Taxonomic Analysis.....	66
5.5 Evolution of Nitrogenase .....	68
<b>CHAPTER 6: CONCLUSION.....</b>	<b>70</b>
<b>REFERENCES.....</b>	<b>.....</b>
<b>APPENDIX A – Script:Estimate insert size Script.....</b>	<b>.....</b>
<b>APPENDIX B – Gene Annotation Parameters .....</b>	<b>.....</b>
<b>APPENDIX C- Venn Diagram Count Script .....</b>	<b>.....</b>
<b>APPENDIX D – Multi Contigs Into Singular Contig Script.....</b>	<b>.....</b>

## LIST OF TABLES

		<b>Page</b>
Table 2.1	Nitrogen fixation ( <i>nif</i> ) genes related products and functions	13
Table 3.1	List of freewares used for the overall process of Assembly, Annotation and Analysis of <i>P. macerans</i> ATCC 8244 genome	23
Table 4.1	Summary of raw fastq reads quality of <i>P. macerans</i> ATCC 8244	34
Table 4.2	Summary of statistics before and after the quality control processing on fastq raw reads of <i>P. macerans</i> ATCC 8244	35
Table 4.3	Performance of different Assembler softwares	37
Table 4.4	Kmer values selection and their assembly statistics	38
Table 4.5	Statistics of assembly optimization before and after k-mer 61 Implementation	39
Table 4.6	Final Assembly statistic of <i>P. macerans</i> ATCC 8244 genome	40
Table 4.7	Clusters of Orthologous Groups (COG) function classification of <i>P. macerans</i> ATCC 8244 genome	42
Table 4.8	Nif and nitrogen fixation associated genes product found in of <i>P. macerans</i> ATCC 8244 genome	44
Table 4.9	List of pathways available in <i>Paenibacillus macerans</i> ATCC 8244	54

## LIST OF FIGURES

	<b>Page</b>
Figure 2.1 Nitrogen Cycle	8
Figure 3.1 Schematic workflow overview of overall process of genomic assembly, annotation and analysis of <i>P. macerans</i> ATCC 8244	22
Figure 4.1 Histogram of Inner Distance between Paired Ends reads of <i>P. macerans</i> ATCC 8244.	36
Figure 4.2 SEED classification of <i>Paenibacillus macerans</i>	43
Figure 4.3 Best model for maximum likelihood approximation fit for final tree phylogenomic test	46
Figure 4.4 Phylogenomics tree of selected members of <i>Paenibacillus</i> genus	48
Figure 4.5 Nitrogenase phylogenetic tree	49
Figure 4.6 A Venn diagram showing gene clusters of <i>Paenibacillus</i> sp. JDR1, <i>P. macerans</i> ATCC 8244 <i>P. mucilaginosus</i> 3016 <i>P. polymyxa</i> CR1	51
Figure 4.7 A Venn diagram showing gene clusters of <i>Azorhizobium caulinodans</i> ORS 57 , <i>Azotobacter vinelandii</i> DJ <i>P. macerans</i> and <i>Klebsiella pneumoniae</i> SB3432	52
Figure 4.8 Comparative Bisphenol A Degradation Pathway	56

Figure 4.9	Pathways of nitrogen metabolism in the four selected nitrogen fixing bacteria	57
Figure 4.10	Prediction of Genomic islands in the genome of <i>P.macerans</i> ATCC 8244	59
Figure 4.11	Side by side comparison deduction of phage induced genomic island of <i>P. macerans</i> ATCC 8244	60
Figure 5.1	Formation of 2,3-butanediol from pyruvate	65

## LIST OF ABBREVIATIONS

%	Percentage
°C	degree Celsius
Σ	sigma factor
Λ	lambda clone
Mg	micro gram
μl	micro litre
μM	micro molar
Aa	amino acid
acc. no.	accession number
Atm	standard atmospheric pressure
ADP	adenosine diphosphate
ATP	adenosine triphosphate
bp	base pair
BPA	Bisphenol A
CDS	coding sequences
CRISPR	clustered regularly interspaced short palindromic repeats
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide triphosphates
DTT	Dithiothreitol
G	Gram

IPTG	isopropyl $\beta$ -thiogalactopyranoside
Kb	kilo base pair
kDa	kilo Dalton
M	Molar
Min	Minute
Mm	Millimolar
OD	optical density
ORF	open reading frame
RACE	rapid amplification of cDNA ends
PCR	polymerase chain reaction
RBS	ribosome binding site
RNA	ribonucleic acid
RT	reverse transcription
Rpm	revolution per minute
QC	quality control
Sec	Second

# KAJIAN GENOMIK *Paenibacillus macerans* ATCC 8244, BAKTERIA GRAM POSITIF YANG MENGIKAT NITROGEN

## ABSTRAK

*Paenibacillus macerans* ATCC 8244 adalah bakteri pembentuk spora jenis Gram-positif yang berkemungkinan mempunyai kebolehan metabolik yang paling luas antara ahli genus *Paenibacillus*. *P. macerans* boleh melakukan fermentasi gula heksosa, pentosa, selulosa dan hemiselulosa. Di samping itu, keupayaan untuk mengikat nitrogen mewujudkan potensi agronomi. Kajian ini mengemukakan satu pendekatan genomik menyeluruh untuk pencirian mikroorganisma ini. Analisis filogenomik menunjukkan bahawa genom terdekat yang dikenalpasti adalah *P. macerans* strain ZY18 dengan identiti 98 peratus. Perhimpunan draf genom menghasilkan sejumlah 7.305.296 pasangan bes (bp) yang dibentuk oleh 97 perancah. Nilai N50 bagi perancah adalah 157,355 bp. Saiz purata perancah adalah 75,312 bp dan nisbah GC purata genom draf adalah 53.14%. Ramalan dan anotasi gen yang dibuat pada peringkat perancah menghasilkan 6953 jujukan pengekodan. Ini termasuk 85 gen tRNA dan 9 gen rRNA. Analisis klasifikasi Kelompok Gen Berortolog (COGs) menunjukkan bahawa bilangan tertinggi jujukan pengekodan melibatkan kategori "Pengangkutan dan Metabolisme Karbohidrat " dengan jumlah 845 model gen. Draff genom *P. macerans* ATCC 8244 mendedahkan 15 gen yang terlibat dalam pengikatan nitrogen. Kehadiran gen yang terlibat dalam degradasi Bisphenol A (BPA) menunjukkan potensi dalam mencerna bahan kimia yang berbahaya kepada alam sekitar ini. Di samping itu, ia mempunyai keupayaan untuk menghasilkan 2-3 butanediol menjadikannya sebagai pengeluar biofuel yang berprospektif.

# A GENOMIC ANALYSIS OF *Paenibacillus macerans* ATCC 8244, A GRAM POSITIVE NITROGEN FIXING BACTERIUM

## ABSTRACT

*Paenibacillus macerans* ATCC 8244 is a Gram-positive, spore-forming bacterium with possibly the widest metabolic capabilities among the *Paenibacillus* genus. It is able to perform fermentation of hexoses, deoxyhexoses, pentoses, cellulose, and hemicelluloses. In addition, its ability to fix nitrogen creates an agronomic potential. This study presents a whole genomic approach to give a holistic view of the features of this microorganism. A phylogenomic analysis revealed that the nearest genome identified was *Paenibacillus macerans* strain ZY18 with 98 percent identity. The draft genome assembly yielded a total of 7,305,296 base pairs (bp) formed by 97 scaffolds. The N50 value of the scaffold was 157,355 bp. The average scaffold size was 75,312 bp and the average GC ratio of the draft genome was 53.14 %. Gene prediction and annotation performed at the scaffold level yielded 6953 coding sequences. These included genes for 85 tRNAs and 9 rRNAs. A Clusters of Orthologous Groups (COGs) classification analysis revealed that the highest number of coding sequences involved the “Carbohydrate Transport and Metabolism” category with a total of 845 gene models. The draft genome of *P. macerans* ATCC 8244 revealed 15 genes that are involved in nitrogen fixation. The presence of genes involved in Bisphenol A (BPA) degradation showed a potential in degrading this environmentally hazardous chemical. In addition, it has the capability to produce 2-3 butanediol making it as a prospective producer of biofuel.

## CHAPTER 1: INTRODUCTION

Following the first isolation of *Bacillus subtilis* in 1872, a lot of bacteria were classified under the genus of *Bacillus* since they have common characteristics such as rod-shaped, endospore forming, and possessing either aerobic or facultative anaerobic lifestyle. Members of the genus *Paenibacillus* were originally classified as *Bacillus* as well due to their similar characteristics. The word 'paene' means "nearly or almost" in Latin and hence "Paenibacillus" means "almost bacillus" (De Vos 2009). Members of the genus *Paenibacillus* are usually curved or straight rod in shape, ranging from 0.5-1.0 x 2-6  $\mu\text{m}$  in size and with G+C content ranging around 39 to 59 percent. They are Gram-positive and have been found in many environments. They can be also belong to any of the different classes of bacteria such as psychrophile, mesophile, thermophile, alkaliphile, neutrophile, aerobic or anaerobic.

The variety of adaptations of many members of this genus allows them to play unique and important roles in diverse environments. For example, *Paenibacillus larvae* became a pathogen of the cultured honey bee *Apis mellifera*, especially the first and second instar larvae (Genersch 2010). Others such as *Paenibacillus sp. Aloe-11* reportedly invade human intestine (Li et al. 2012). However, members of this genus are also famous for their association with plant communities (McSpadden Gardener 2004). They can act as plant growth enhancers with features that make them biotechnologically beneficial. Furthermore their ability to adapt in the human intestinal environment could potentially make them as beneficial probiotic components (Hoyles 2012).

An interesting species under this genus is *Paenibacillus macerans*. It can be considered as one with the broadest metabolic capabilities. *P. macerans* N234A has been known to anaerobically ferment glycerol (Gupta et al. 2009). It is also able to ferment hexoses, deoxyhexoses, pentoses, cellulose, and hemicelluloses. Since it has been reported to be associated with several pseudo bacteremia cases, *P. macerans* has been described as having features of opportunistic pathogen (Noskin et al. 2001). More notably, *P. macerans* has the ability to fix nitrogen that is very beneficial towards crops productivity. Biological nitrogen fixers accounted for supplying nearly 60 % of world's new ammonia source annually (Schlesinger, 1991). It is vital to harness research understanding on biological nitrogen fixation to maximize its potential, especially on the Gram positive bacteria. Although there is tremendous research effort done on the Gram-negative nitrogen fixers such as *Klebsiella*, *Bradyrhizobium* and *Azotobacter* species, there is less information on Gram-positive members such as *P. macerans*.

The field of life sciences has been tremendously transformed with the advent of genome sequencing. The ability to sequence in a high throughput fashion has not only advanced our fundamental understanding of how genes and genomes are assembled, it also has yielded extremely in depth knowledge of the structure of evolutionary trees, increased our understanding of genetics and development, and led to the growth of new biotechnologies. The genomic information obtained enabled us to understand the strategies of the microbes to survive under diverse environmental conditions.

A genomic study on *P. macerans* would uncover the biological mechanisms and functions of the interacting genes responsible for its special physiological features and adaptation capabilities. Currently there is no *P. macerans* complete genome available in the public database using next generation sequencing approach. This study aims to study the *P. macerans* genome using whole genome approach.

### **1.1 Research Objectives:**

The investigation was carried out to fulfill the following objectives:

1. To analyze the structure of a *P. macerans* draft genome
2. To characterize the genes related to nitrogen fixation and unique metabolic pathways belongs to *P. macerans*
3. To perform a functional and comparative analysis of *P. macerans* genome with other species

## CHAPTER 2: LITERATURE REVIEW

### 2.1 *Paenibacillus* genus

Members of the bacterial genus *Paenibacillus* are usually curved or straight rod in shape, ranging from 0.5-1.0 x 2-6  $\mu\text{m}$  in size with a G+C content ranging around 39 to 59 percent (De Vos et al. 2010). They are Gram-positive and have been found in many environments. The varieties of niches that many members of this genus reside in reflect their adaptability. Various psychrophilic, mesophilic and thermophilic representatives have been successfully isolated from a range of different niches reflecting how diverse they are in terms of growth temperatures. Isolates which are alkaliphilic or acidophilic play unique roles in their diverse environments. *Paenibacillus* members capable of growing under aerobic or anaerobic have also been successfully isolated.

Some species are pathogenic to animals and their pathogenicity towards certain pests means that they can serve as effective biological pesticides. For example, *P. larvae* is pathogenic to bees and have been proposed to be used as a biological control of wild honey bees (Genersch 2010). Another species, *Paenibacillus* sp HGF -5, even invade the human intestine and this ability to adapt in the intestinal environment could be taken advantage of by moulding it into a beneficial probiotic component in human (Hoyle et al. 2012). There are also members of this genus well known for their association with the plant communities (McSpadden Gardener 2004; Khan et al. 2008). Their ability to produce plant growth enhancers makes them biotechnologically beneficial.

## 2.2 *Paenibacillus macerans*

*Paenibacillus macerans*, previously called *Bacillus macerans* and *Bacillus acetoethylicum*, is a Gram-positive, spore-forming bacterium belonging to the genus *Paenibacillus* (Ash et al. 1993). *P. macerans* can be considered as a member species with the broadest metabolic capability within this genera. It was reported that *P. macerans* N234A able to ferment hexoses, deoxyhexoses, pentoses, cellulose, and hemicelluloses. It has also been known anaerobically ferment glycerol (Gupta et al. 2009). In addition, *P. macerans* also has features of an opportunistic pathogen with reported associations to several pseudo-bacteraemia cases (Noskin et al. 2001). *Paenibacillus macerans* ATCC 8244 was first isolated from potatoes by Schardinger in 1905 and was mentioned in 1942 by (Tilden and Hudson 1942).

## 2.3 Nitrogen Fixing Microorganisms

Biological nitrogen fixers are microorganisms that are capable of converting the gaseous form of nitrogen into ammonia. These microorganisms are either bacteria or archaea and they can be in a symbiotic relationship or free living. Biological Nitrogen Fixation (BNF) accounted for 60 % supply of the world's new source of ammonia annually (Schlesinger 1991). The role of nitrogen fixation is clearly important in the field of agriculture. Availability of fixed nitrogen correlates proportionally with crops productivity. It is therefore vital to harness a deeper understanding on BNF to maximize its potential.

Most of the studies on biological nitrogen fixation were performed on the Gram-negative bacteria such as *Klebsiella*, *Bradyrhizobium* and *Azotobacter*. For the Gram-positive group, studies had been done on *Clostridium pasteurinum* but its

difficulty to grow and lack of genetic tractability prevented rapid advancement of its research (Chen and Johnson 1993). Thus a suitable model is needed to dissect the nitrogen fixation system in the Gram-positive group of bacteria. *P. macerans* offers a possible model due to its ease of growth in aerobic condition and the availability of genetic amenability such as transposons and plasmids.

## **2.4 Importance Of Nitrogen**

Nitrogen constitutes 78 percent of the earth's atmosphere. It is an essential component for composition of proteins that are required for all living things including bacteria, plants and humans (Stewart and Gallon 1980; Clark et al. 1981). Nitrogen can exist both in organic and inorganic forms. The organic form can be ammonium ( $\text{NH}_4^+$ ), nitrite ( $\text{NO}_2^-$ ), nitrate ( $\text{NO}_3^-$ ), nitrous oxide ( $\text{N}_2\text{O}$ ) and nitric oxide ( $\text{NO}$ ). The inorganic nitrogen form would be in the inert nitrogen gas ( $\text{N}_2$ ) itself. Nitrogen serve as a component for the building units that form nucleic acids (such as DNA and RNA) as well as for the amino acids that form proteins. Unfortunately, the actual nitrogen gas cannot be utilized directly by most of organisms; it needs to undergo a series of processes that convert it to an organic form. It needs to go through a fixation process to become utilizable. Circulation of nitrogen from the atmosphere to organic compounds by biological nitrogen fixation and then back to the atmosphere (denitrification) is called the "nitrogen cycle". Nitrogen cycle affects the major rate of growth and decay of many ecosystems through primary production and decomposition (Fowler et al. 2013; Isobe and Ohte 2014).

## 2.5 Nitrogen Cycle

The cycle supply chain starts with the nitrogen fixation process in which nitrogen is converted to ammonia (Mancinelli 1996). This fixation process is very energy intensive and can happen in multiple ways: (i) by lightning strikes, (ii) through industrial synthesis by humans (for fertilizers), or (iii) by biological nitrogen fixers. Biological nitrogen fixers are microorganisms that are symbiotic (examples: *Frankia sp.*, *Rhizobium sp.*, *Bradyrhizobium sp.*) or free living (such as *Azotobacter sp.*). The ammonia produced is then assimilated by majority of plants and others through the glutamate synthase cycle (Mifflin and Lea 1975). When the organisms die, diverse bacteria and fungi can convert the decay products of the body into ammonium. Bacteria such as *Nitrosomonas sp.* and *Nitrobacter sp.* are able to assimilate and convert the inorganic ammonium into nitrite and nitrate in a 2-step process called nitrification. The nitrate produced is a source of energy for groups of bacteria such as *Pseudomonas sp.* in anaerobic respiration. They use nitrate as an electron acceptor in place of oxygen and this will convert back nitrogen into its inert gas form through a process called denitrification. This completes the nitrogen cycle. Figure 2.1 shows overview process of nitrogen cycle. It has been noted that human influence on the nitrogen cycle through over-usage of fertilizers has caused imbalance leading to environmental problems such as acidification and eutrophication. These effects on the environment can be reduced if more effort is given to optimize biological nitrogen fixation instead of relying of industrially produced chemical fertilizers.

Key processes in nitrogen cycle involves:

- Fixation -the process of  $N_2$  being converted into  $NH_3/NH_4^+ / NO_3^-$ ,
- Nitrification- the process of  $NH_3$  being oxidized to  $NO_3^-$  and  $NO_2^-$
- Ammonification- the process of organic nitrogen are converted to  $NH_3$
- Denitrification, the process of reduction of  $NO_3^-$  to  $N_2$

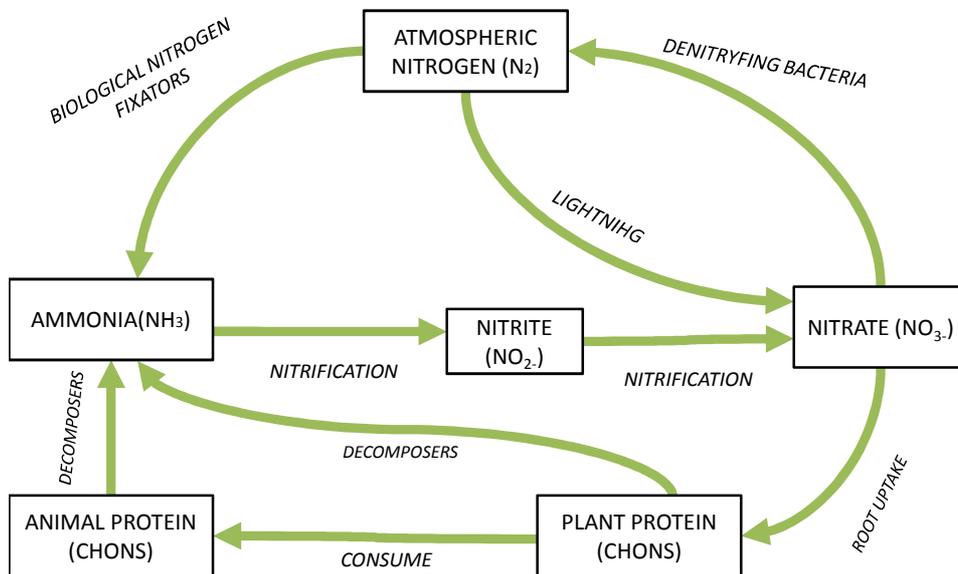


Figure 2.1 Nitrogen cycle

## **2.6 Nitrogen Fixation Key Enzymes**

Atmospheric nitrogen or dinitrogen (N<sub>2</sub>) is converted to biological form through key enzyme nitrogenase (EC 1.18.6.1). This enzyme is essentially a metalloprotein complex consisting of MoFe and Fe proteins. The Fe protein acts as an electron donor to the catalytic site on the MoFe protein (Rees et al. 2005).

### **2.6.1 MoFe protein**

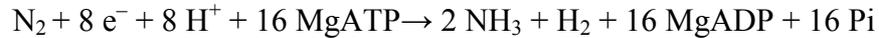
Structurally, the size of the MoFe protein is about 200-250 kDa and it is a tetramer formed by four subunits: 2 alpha and 2 beta subunits. The gene *nifD* encodes the alpha subunit and *nifK* encodes the beta subunit. The MoFe protein contains the substrate-binding site with six prosthetic groups of two types of clusters: 4 8Fe-7S and 2 Mo-7Fe-9S clusters. The former is also known as the P-cluster while the latter is commonly known as the FeMo cofactor. It contains the substrate binding site for a reduction process (Seefeldt et al. 2009). Both clusters act as a bridge to conduct electron transport between the Fe-protein and MoFe protein.

### **2.6.2 Fe Protein**

Fe-protein is about 55 to 65 kDa in size and is a dimer formed by only the alpha subunits. It contains a 4Fe-4S metal cluster that connects the two alpha subunits via covalent bond. The 4Fe-4S cluster in the active site modulates between the reduced and oxidised states of the protein during the electron transfer process to the FeMo-protein (Seefeldt et al. 2009).

## 2.7 The Chemical Equation and Stoichiometry of Nitrogen Fixation

Under normal condition, the equation is:



The process starts by reduction of the Fe protein by ferredoxin or flavodoxin or other electron donors (Dixon and Kahn 2004). Then, a single electron is transferred from the Fe protein to the MoFe protein and this transfer is dependent on a hydrolysis of an ATP molecule. Upon reaching the MoFe protein, the electron is then transferred internally from the P cluster to the FeMo-co located at the active site.

## 2.8 Regulation Of Nitrogen Fixation

The nitrogen fixation activity is controlled according to the environmental ammonium content as well as the level of oxygen. Bacteria repress transcription when there is a high level of nitrogen or oxygen. There are two regulatory proteins involved: NifL and NifA. In *Azotobacter vinelandii*, NifL is a redox- and nitrogen-responsive regulatory flavoprotein and functions as a regulatory protein that controls the transcription of nitrogen fixation genes by regulating the activity of the transcriptional activator NifA. This involves a direct protein-protein interaction. When the levels of oxygen is high, the NifL protein, in the presence of an oxidised Flavin adenine dinucleotide (FAD), inhibits the NifA protein causing inhibition of *nif* operons. When the oxygen is low, NifL disassociate from NifA enabling it to transcribe the *nif* genes.

Similarly, a high level of ammonia would trigger repression of the nitrogen fixation genes. Ammonia modifies the structure of glutamine synthase via covalent adenylation. It also acts as a corepressor that changes the regional DNA binding site region known as *nifR* in the *nifRLA* operon. This results in the prevention of transcription of the operon (Halbleib and Ludden 2000).

## **2.9 *Azotobacter vinelandii* as a Model of Nitrogen Fixing Organism**

One of the best studied model to understand the mechanisms involving nitrogen fixation is the Gram-negative bacterium *Azotobacter vinelandii* (Schmitz *et al.* 2002). *A. vinelandii* is a free living diazotroph that is able to fix nitrogen under aerobic condition. It has a single circular chromosome of approximately 5.37 Mbp in size with a total number of 5051 genes (Setubal *et al.* 2009). Ambient oxygen concentration is required (around 20 percent) for *A. vinelandii* to catalyze the nitrogen fixation process (Curatti *et al.* 2005). *A. vinelandii* is special because it possesses 3 different types of enzymes to perform its nitrogen fixation: molybdenum nitrogenase, iron nitrogenase, and vanadium nitrogenase (Pau *et al.* 1989; Hales 1990). These are sometimes simply abbreviated to the Mo-, V- and Fe- dependant enzymes.

### 2.9.1 Genes Involved In Nitrogen Fixation of *A. vinelandii*.

Inside *A. vinelandii*, the genes for the three oxygen sensitive nitrogenases (Mo-, V- and Fe-dependant enzymes) are located in three separate regions (Setubal et al. 2009). The molybdenum nitrogenase is made up of the dinitrogenase reductase coded by the gene *nifH* and the second component, the MoFe protein, is coded by *nifDK*. The iron and molybedum (Mo-Fe) cofactors are situated at the active site of the MoFe-protein (Seefeldt et al. 2009). The NifH protein supplies electron to the MoFe-protein and the latter acts as the site for nitrogen reduction. Nitrogen is reduced at the active site to produce the final product, ammonium, along with the release of hydrogen. This reduction process is accompanied by the hydrolysis of ATP. Alternative nitrogenases such as the V- and Fe-nitrogenases comprised of homologous VnfHDK and AnfHDK proteins (Hales 1990). These alternative nitrogenases require additional structural components of the dinitrogenase that include the VnfG and AnfG as subunits. A lot of accessory proteins participate in the synthesis and assembly of the transition metal cofactors and these are encoded within the *nif* gene clusters. Table 2.1 describes nitrogen fixation (*nif*) genes related products and functions.

Table 2.1 : Genes involved in nitrogen fixation adapted from (Lee et al. 2000)

<b>Gene (<i>nif</i>)</b>	<b>Function</b>
Q	Incorporation of molybdenum into FeMo-cofactor
B	FeMo-cofactor precursor synthesis
A	Positive regulation of <i>nif</i> (transcriptional activator)
L	Negative regulation of <i>nif</i> (negative regulator)
F	Flavodoxin (electron carrier)
M	Nitrogenase reductase maturation process
W	Dinitrogenase stability; oxygen protection of FeMo-protein
V	Homocitrate-synthase involved in FeMo-cofactor synthesis
S	Involved in mobilization of S for Fe-S cluster synthesis and repair.
U	FeMo-protein processing
X	FeMo cofactor synthesis. Negative regulation
E	Synthesis and insertion of FeMo-cofactor into dinitrogenase protein
N	Required for FeMo-cofactor synthesis
T	Nitrogenase maturation
K	Dinitrogenase $\beta$ subunit
D	Dinitrogenase $\alpha$ subunit
H	Dinitrogenase reductase (electron donor)
J	Pyruvate-flavodoxin oxidoreductase (electron transport)

## **2.10 Bioinformatics as a Fusion of Genomics and Computing**

The nitrogenous bases "Adenine (A), Cytosine (C), Guanine (G), Thymine (T)" are basically the four significant compounds that hold the key towards the meaning of life. Understanding the sequential order of the bases mobilized researchers to investigate the genetic information hidden within a biological system. Looking at the bigger scale, study of genomics includes understanding how the genetic and non-genetic makeup interacts and form the biological system of an organism.

A genome can be defined as the genetic material of an organism and it is made up from nucleic acids. It consists of coding and non-coding parts that together hold the mechanisms on how the organism lives and maintains its existence. This includes gene function, expression, regulation as well as its evolution. Recent advancement in the next generation sequencing technologies has increased the capability to mine biological information exponentially from various aspects. The speed of sequencing and the amount of data produced allow more discoveries and biological understanding to be made. The progress made at the level of data resolutions (or sensitivity) itself was phenomenal. However, connecting the dots that involve analysis and interpretation often face obstacles, especially in handling huge amount of biological information made up of genomics sequences. With advances in computing, most of the challenges faced can be answered through bioinformatics, a hybrid intersection of the disciplines of Biology and Computation. Bioinformatics is a way to use computing capabilities to systematically gather, manage, and make connections of biological information that aids more insights into understanding life and the scientific meaning behind it (Kelly 1989).

## **2.11 Next Generation Sequencing and Bioinformatics**

Sequencing technology refers to the process of nucleotide identification in a chain of nucleic acid polymers which can be either be referred to as DNA or RNA. The development of DNA sequencing has opened a new window of understanding for genomic sciences especially on genome organization and compositions among different organisms. This has further enhanced by development of the “Next Generation Sequencing” (NGS) platforms which make data generation grow exponentially. This phenomenon triggered the expansion of bioinformatics as a discipline to cope with large and complex datasets (Metzker 2005).

### **2.11.1 Illumina Sequencing**

Illumina sequencing technology relies on fluorescent-labeled nucleotides that enable identification of single bases as they are introduced into the DNA strands (Fedurco et al. 2006; Bentley et al. 2008). The process starts with the DNA being sheared and hybridised in the flow cell. Inside the flow cell, fragments are amplified in an isothermal reaction creating multiple copies of the same read or fragment. The flow cell then is then flooded with fluorescent-labeled nucleotides specific for each type of base. To ensure one base is added at a time, terminators are included in the reaction mixture. Coordinated image positioning location on slide was captured based on fluorescent signal (which happen only when bases has been added). The fluorescent signal are then removed so that other signal that comes from the next base to take place. This emphasis was given to prevent signal contamination/mix-up from happening. These coordinated fluorescent images are then analyzed by server to construct sequence. The read length of the latest version is about 150 base-long (Bentley et al. 2008).

### **2.11.2 Roche 454 Pyro Sequencing,**

The principle of the Roche 454 sequencing technology involves ligating nebulized fragments of DNA with adaptors (Ronaghi et al. 1996; Ronaghi et al. 1998). These ligated DNA molecules are then captured into beads enveloped in oil-water emulsions. The fragments on the beads were amplified using PCR. The beads containing the amplified DNA bound to them are then placed on a Picotiter plate. An enzymatic combination of DNA polymerase, luciferase and ATP sulfurylase are also packed into each well of the plate. The plate are placed into the sequencing machine called GS FLX where light peaks are produced in a “pyrogram format” when nucleotides are added. These light peaks are generated as a result of a reaction involving luciferase and ATP. This technology can give read length of up to 1 kilobases (Droege and Hill 2008).

### **2.11.3 Ion Torrent Sequencing**

The principle of the Ion Torrent technology is different from Illumina and Roche 454 in a sense that it does not use optical signal to identify the bases. Instead it relies on pH values. The primary principle relies on the fact that addition of a dNTP molecule to a DNA fragment will involve the release of H<sup>+</sup> ion (Rothberg et al. 2011). The process starts by flooding the slide with a single type of dNTP together with buffers and polymerase and one NTP at a time. Each H<sup>+</sup> ion released in each well is measured as a decrease of pH value. The values of the pH changes allow the machine to determine the profile of the base added. The sequences of changes are progressively read to develop the sequencing reads. The average read of this method is about 200 bases.

#### **2.11.4 Pac Bio SMRT Sequencing**

Pacific Biosciences (PacBio) developed a single molecule real time (SMRT) sequencing method that uses parallel sequencing of single DNA molecules by synthesis (Eid et al. 2009). It uses zero –mode waveguide (ZMW) which guides light energy into a volume that is compact in all dimension compared to the wavelength of the light (Levene et al. 2003). A single molecule of DNA polymerase with a single molecule of DNA template was affixed at bottom of the ZMW. The ZMW that was equipped with high-resolution camera records the activity of the fluorescence signal of incorporation of nucleotide in a movie-sequential format. This method is capable of generating reads maximum length over 40000bp with average of 14000bp.

#### **2.12 Genome Assembly**

The shotgun sequencing route was first introduced on the DNA sequencing of phiX174 bacteriophage genome and it still remains an important technique for genome sequence assembly (Sanger et al. 1977a; Sanger et al. 1977b). Shotgun sequencing method involves obtaining random sequence reads from a genome (which may ranging from a small plasmid to a complete chromosome of an organism) and merging them together into contigs (a linear representation of nucleotides) using overlaps as the basis of contigs construction. Paired end reads are then used to link together distinct sequence contigs into super contigs using forward and reverse end reads. Contigs that are joined together are called supercontigs, and a combination of supercontigs becomes scaffold. Assemblies are measured by the size and accuracy of their contigs and scaffolds. Assessment of assemblies usually consider contig accuracy and size. N50 refers to contig size that sum of sorted length of 50 percent of total assembly is reached.

Basically there are two standard methods of genome assembly, namely, the Overlapping Consensus Layout (OCL) method and the De Bruijn method (Miller et al. 2010). The OCL method usually works well with long reads. It relies on analyzing the overlap of reads through graphing process. The ‘contiging’ or assembly process require the potential reads to be joined by ‘make do’ where the reads are joined exactly one time through overlap. Examples of current softwares that use the OCL method are Newbler, Phast, and Celera Assembler (Miller et al. 2010).

The De bruijn method has the same principle of using overlapping information (Compeau et al. 2011). However, the main difference between this and the OCL method is in the process of contig forming. The algorithm permits paths of the graphing process to accept potential reads that can be joined by different edges of overlaps by one or more times (Miller et al. 2010). This characteristic is most beneficial towards short reads that have less specificity in the sequence. Examples of software using this method are IDBA-UD, Velvet and ABySS (Simpson et al. 2009).

### **2.13 Gene Finding**

Upon genome assembly, the next process is to find biologically meaningful regions associated with protein coding genes, RNA genes and regulatory regions. The search for these often utilizes tools known as “gene finders”. In prokaryotes, the gene finding process is associated with finding the longest transcribe-able coding region that is uninterrupted by a stop codon. Since coding regions will be translated, they are characterized by the fact that three successive bases in the correct frame

define a codon which will be translated into a specific amino acid in the final protein. The gene finding process will avoid the intergenic region where a coding DNA sequence (CDS) cannot be found.

Generally there are two types of gene finders and they are categorized as intrinsic and extrinsic (Korf 2004). The intrinsic approach refers to an *ab initio* method in which gene finders are trained just based on the assembled DNA sequences. Statistical probabilities are tabulated based on specific signals (for example translation initiation site recognition) that indicate the presence of a gene nearby. Examples of software that use the intrinsic gene finding method are Prodigal, Glimmer and Genemark. The extrinsic gene finding method refers to a process that looks for evidences based on known similarity from other evidences. Usually, other characterised genes from establish databases such as Uniprot, NCBI Refseq, are used to find similar regions that have protein-coding sequences. Examples of software that use the external gene finding approach are IPRSCAN and FASTA (Alves and Buck 2007).

#### **2.14 Post Annotation; Analysis, Comparative and Evolutionary Studies**

The availability of NGS in the genomics era has impacted the value of science tremendously. The ability to sequence in a high throughput fashion has not only advanced our fundamental understanding of how genes and genomes are assembled, it also has yielded in depth knowledge of the structure of evolutionary trees, increased our understanding of genetics and development, and led to the growth of new biotechnologies. The genomic information obtained enabled us to understand the strategies of the microbes to survive under diverse environmental

conditions. Upon gathering genome assembly and annotation, post genome analysis are crucial in determining novel features of the organism, relationship, susceptibility, exploitability, uniqueness, pattern . These features can be tracked by comparing features between one and another. The principle revolves around the idea of analyzing conserved motif/sequence/features/genes could account for biological function by comparing the entity to one another. These are the essence of comparative genomics (Bejerano et al. 2004). The study of these conserved motif/sequence/features/genes ancestry lineage would give light to evolutionary information (Alfoldi and Lindblad-Toh 2013).

## CHAPTER 3: MATERIALS AND METHODS

### 3.1 Schematic Workflow Of This Study

The overall process of genomic assembly, annotation and analysis of *Paenibacillus macerans* ATCC 8244 is as portrayed in Figure 3.1 .

### 3.2 Bioinformatics Tools For Assembly, Annotation And Analysis

Various softwares were employed for the whole genome assembly, annotation and analysis. All of the softwares are available as freewares. The specific function of each software is described in all the subsections below which delineate the all bioinformatic processes involved. Table 3.1 lists out the softwares involved as well as their web addresses for downloading purposes.

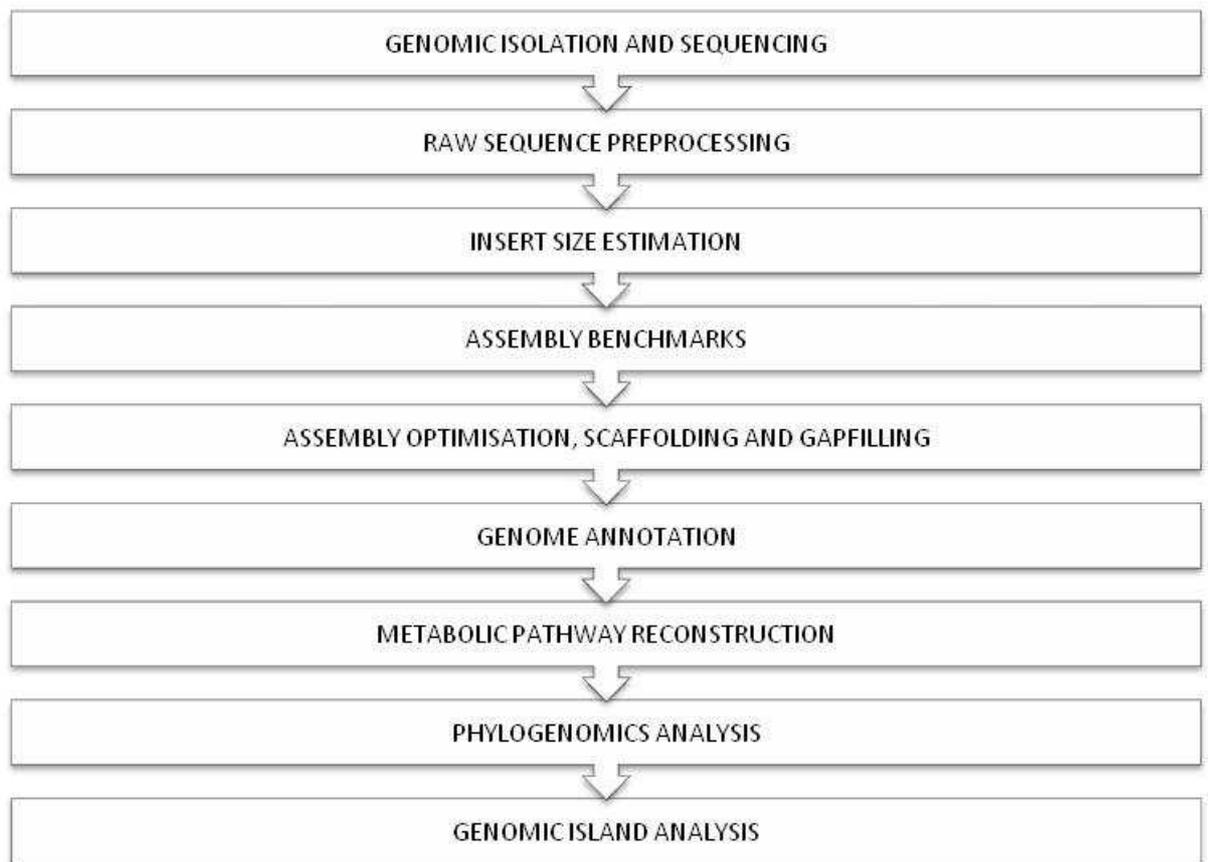


Figure 3.1 Schematic workflow overview of overall process of genomic assembly, annotation and analysis of *P. macerans* ATCC 8244.

Table 3.1: List of freewares used for the overall process of Assembly, Annotation and Analysis of *P. macerans* ATCC 8244 genome

Software	Version	URL
NGS QC Toolkit	2.33	<a href="http://www.nipgr.res.in/ngsqctoolkit.html">http://www.nipgr.res.in/ngsqctoolkit.html</a>
VELVET	1.1	<a href="http://www.molrev.org/software/genomics/velvet">http://www.molrev.org/software/genomics/velvet</a>
BOWTIE	1.01	<a href="http://bowtie-bio.sourceforge.net/index.shtml">http://bowtie-bio.sourceforge.net/index.shtml</a>
MIRA	3	<a href="https://sourceforge.net/projects/mira-assembler/">https://sourceforge.net/projects/mira-assembler/</a>
SOAP DENOVO	r218	<a href="https://github.com/aquaskyline/SOAPdenovo2">https://github.com/aquaskyline/SOAPdenovo2</a>
ABYSS	1.3	<a href="http://www.bcgsc.ca/platform/bioinfo/software/abyss">http://www.bcgsc.ca/platform/bioinfo/software/abyss</a>
SSPACE	2	<a href="https://omictools.com/sspace-tool">https://omictools.com/sspace-tool</a>
GAPFILLER	1.4	<a href="https://sourceforge.net/projects/gapfiller/">https://sourceforge.net/projects/gapfiller/</a>
PRODIGAL	2.6	<a href="http://prodigal.ornl.gov/">http://prodigal.ornl.gov/</a>
ARAGORN	1.2.36	<a href="http://mbio-serv2.mbioekol.lu.se/ARAGORN/">http://mbio-serv2.mbioekol.lu.se/ARAGORN/</a>
RNAMMER	1.23	<a href="http://www.cbs.dtu.dk/services/RNAmmer/">http://www.cbs.dtu.dk/services/RNAmmer/</a>
BLAST+	2.227	<a href="ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/">ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/</a>
HMMER	3	<a href="http://www.ebi.ac.uk/Tools/hmmer/">http://www.ebi.ac.uk/Tools/hmmer/</a>
ASGARD	1.5	<a href="https://sourceforge.net/projects/asgard-bio/">https://sourceforge.net/projects/asgard-bio/</a>
MUSCLE	3.3	<a href="http://www.drive5.com/muscle/">http://www.drive5.com/muscle/</a>
TRIMAI	1.2	<a href="http://trimal.cgenomics.org/">http://trimal.cgenomics.org/</a>
PHYML	3	<a href="http://www.atgc-montpellier.fr/phyml/binaries.php">http://www.atgc-montpellier.fr/phyml/binaries.php</a>
PROTTEST	2	<a href="https://github.com/ddarriba/prottest">https://github.com/ddarriba/prottest</a>
ORTHOMCL	2	<a href="http://orthomcl.org/common/downloads/">http://orthomcl.org/common/downloads/</a>
MCL	1.4	<a href="http://micans.org/mcl/">http://micans.org/mcl/</a>
ISLANDVIEWER	2	<a href="http://www.pathogenomics.sfu.ca/islandviewer/">http://www.pathogenomics.sfu.ca/islandviewer/</a>
CRISPRFINDER	2014	<a href="http://crispr.i2bc.paris-saclay.fr/Server/">http://crispr.i2bc.paris-saclay.fr/Server/</a>

### 3.3 Genomic DNA isolation

The overall process of genomic DNA isolation was done as follows (kindly executed by Mr Sim Kee Shin, Lab 414, School of Biological Sciences, Universiti Sains Malaysia). A single colony of *P. macerans* ATCC 8244 was cultured into LB medium (5 ml) and incubated at 37 °C with 180 rpm agitation for about 16 hours. This *P. macerans* culture was used as an inoculum into a larger volume of LB medium (200 ml) for DNA isolation. The culture was checked for contamination by streaking it on an LB agar medium and observed for colonial consistency. The

isolation protocols were performed by using QIAGEN Genomic tips – 100G according to the manufacturer’s protocol. The 200 ml overnight culture was chilled on ice before centrifuging at 5,000 x G for 5 minutes at 4 °C. The pellet was resuspended in 3.5 ml of Buffer B1 with 10 µl RNase A (100 mg/ml, Thermo Scientific, USA), 100 µl of lysozyme (100 mg/ml, Sigma Aldrich, USA), and 100 µl of Proteinase K (100 mg/ml, Sigma Aldrich, USA). The mixture was then incubated at 37 °C in a water bath for at least 1 hour until the solution turned sticky or clear. After this incubation, 1.2 ml of Buffer B2 was added and the mixture was further incubated at 55 °C for at least 1 hour or until the solution became totally clear. The lysate was centrifuged for 10 minutes at 5,000 x G at 4 °C to precipitate any non-soluble particle. Then, the QIAGEN Genomic-tip 100G was equilibrated with 4 ml of Buffer QBT and allowed to empty by gravitational flow. The column was loaded with the cell lysate and the latter was allowed to flow out by gravitational flow. The column was washed with 7.5 ml of Buffer QC and this process was repeated. The genomic DNA was then eluted from the column by flowing through 5 ml of Buffer QF. The genomic DNA was precipitated by adding 3.5 ml of isopropanol and immediately spooled by using a sterile glass rod. The DNA was washed with 75 % ethanol twice by spooling gently. The genomic DNA was rehydrated by dipping the end of the glass rod bearing the DNA into an Eppendorf tube containing 1 ml of nuclease free water.

The genomic DNA (gDNA) was subjected to agarose gel electrophoresis for DNA integrity assessment. The absorbance value of OD<sub>230nm</sub>, OD<sub>260nm</sub>, and OD<sub>280nm</sub> were used to assess the quality of DNA by determining the ratio of OD<sub>260nm</sub>/OD<sub>280nm</sub> and OD<sub>260nm</sub>/OD<sub>230nm</sub>, respectively.