CONSTRUCTION AND COMPREHENSIVE ANALYSIS OF DNA METHYLOME IN *Pandoraea spp.* AT SINGLE BASE RESOLUTION

LIM YAN LUE

FACULTY OF SCIENCE

UNIVERSITY OF MALAYA

KUALA LUMPUR

2018

CONSTRUCTION AND COMPREHENSIVE ANALYSIS OF DNA METHYLOME IN *Pandoraea spp*. AT SINGLE BASE RESOLUTION

LIM YAN LUE

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

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

2018

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: LIM YAN LUE

Registration/Matric No: SHC140131

Name of Degree: DOCTOR OF PHILOSOPHY

Title of Thesis: CONSTRUCTION AND COMPREHENSIVE ANALYSIS OF DNA METHYLOME IN *Pandoraea spp.* AT SINGLE BASE RESOLUTION

Field of Study: GENETICS AND MOLECULAR BIOLOGY

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 form, 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 or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

CONSTRUCTION AND COMPREHENSIVE ANALYSIS OF DNA METHYLOME IN *Pandoraea spp*. AT SINGLE BASE RESOLUTION

ABSTRACT

Prokaryotic DNA methylation is a prevalent epigenetic modification on bacterial DNA bases and is accomplished by DNA methyltransferases which catalyse covalent attachment of methyl groups onto the DNA bases. DNA methylation of prokaryotic genomes, in particular those catalysed by solitary DNA methyltransferases, were demonstrated to contribute to regulation of a plethora of cellular processes including gene expression regulation, DNA replication control, DNA mismatch repair and pathogenicity. However, despite the expansion of prokaryotic epigenetic studies in recent years, characterisation of DNA methylation, especially in the beta-subclass of *Proteobacteria* and environmental isolates largely remain unexploited to date. Therefore, this dissertation aims to characterise the epigenomic landscape of the genus of *Pandoraea*, a newly described taxa in *Betaproteobacteria*. Members of *Pandoraea* are frequently reported as emerging opportunistic pathogens with multi-drug resistance properties. To date, no methylome studies were performed on these bacterium. Single Molecule Real Time (SMRT) sequencing technology was employed to perform complete genome sequencing and detection of epigenetic modifications at base-pair resolution on 10 *Pandoraea* strains, inclusive of 9 type strains acquired from the Leibniz-Institut DSMZ culture collection centre (*Pandoraea apista* DSM 16535^{T}, *Pandoraea faecigallinarum* DSM 23572^{T}, *Pandoraea norimbergensis* DSM 11628^{T}, *Pandoraea oxalativorans* DSM 23570^{T}, *Pandoraea pnomenusa* DSM 16536^{T}, *Pandoraea pulmonicola* DSM 16583^{T}, *Pandoraea sputorum* DSM 21091^{T}, *Pandoraea thiooxydans* DSM 25325^{T}, and *Pandoraea vervacti* DSM 23571^{T}) and 1 in-house landfill isolate, *Pandoraea pnomenusa* strain RB-38. From the whole-genome methylation patterns analysis, 1 palindromic *N*6-methyladenine (^{m6}A) sequence motif, GTWWAC, was identified in the genomes of all 10 *Pandoraea* strains.

Furthermore, from the restriction-modification (R-M) system genes annotation performed, a Type II orphan methyltransferase, was identified to be the corresponding methyltransferase of the conserved motif detected. Subsequent bioinformatics analyses performed indicated that the candidate methyltransferase represents a novel class of orphan methyltransferase within the family of *Burkholderiaceae* and *Oxalabacteraceae*. Comparative genome-wide GTWWAC methylation pattern distribution analysis performed indicated that this methyltransferase, currently designated as *Pandoraea* adenine methylase (PAM), demonstrated analogous function with Dam (DNA adenine methyltransferase) and CcrM (cell-cycle regulated methyltransferase), both are well-characterised orphan methyltransferases prevalent in *Gammaproteobacteria* and *Alphaproteobacteria* respectively. In conclusion, in addition to providing the first comprehensive illustration of the pan-genus methylome profile of *Pandoraea*, the findings of this dissertation also reported the identification of a potential novel class of orphan methyltransferase within the class of *Betaproteobacteria*.

Keywords: *Pandoraea***,** methylome, orphan methyltransferase, GTWWAC

PEMBENTUKAN DAN ANALISIS KOMPREHENSIF METILOM DNA DALAM *Pandoraea spp*. PADA RESOLUSI BES TUNGGAL

ABSTRAK

Pemetilan DNA prokariot adalah modifikasi epigenetik yang prevalen antara genom bakteria dan dimangkinkan oleh enzim metiltransferase yang menyebabkan reaksi pengikatan kumpulan metil secara kovalen kepada bes-bes DNA. Pemetilan DNA genom prokariot, khususnya yang dimangkinkan oleh DNA metiltransferase tunggal, telah terbukti memainkan peranan dalam pengawalaturan pelbagai jenis proses selular, termasuk pengawalaturan ekspresi gen, pengawalan replikasi DNA, pembaikan salah padan DNA, dan kepatogenan. Walaupun pengajian epigenetik prokariot makin berkembang akhir-akhir ini, namun, pencirian aktiviti pemetilan DNA terutamanya pada ahli-ahli sub-kelas beta filum *Proteobacteria* dan juga pada bakteria yang diperoleh dari sumber alam sekitar amatlah kekurangan. Oleh sebab itu, disertasi ini bertujuan untuk mencirikan epigenom ahli-ahli genus *Pandoraea*, suatu taxa yang baharu ditemui dalam kelas *Betaproteobacteria*. Ahli-ahli genus *Pandoraea* adalah patogen oportunis yang mempunyai ciri-ciri rintangan antibiotik. Buat masa ini, masih tiada pengajian metilom pada ahli-ahli genus tersebut. Dalam kajian tersebut, teknologi penjujukan DNA *Single Molecule Real Time* (SMRT) telah digunakan untuk menghasilkan jujukan genom yang sempurna dan data pengubahsuaian epigenetik dari 10 strain *Pandoraea*, merangkumi 9 strain tip yang diperolehi dari pusat himpunan kultur Leibniz-Institut DSMZ (*Pandoraea apista* DSM 16535^{T}, *Pandoraea faecigallinarum* DSM 23572^{T}, *Pandoraea norimbergensis* DSM 11628^{T}, *Pandoraea oxalativorans* DSM 23570^{T}, *Pandoraea pnomenusa* DSM 16536^{T}, *Pandoraea pulmonicola* DSM 16583^{T}, *Pandoraea sputorum* DSM 21091^{T}, *Pandoraea thiooxydans* DSM 25325^{T}, dan *Pandoraea vervacti* DSM 23571^{T}) dan 1 strain *in-house* yang diasingkan dari sampel kambus tanah, *Pandoraea pnomenusa* RB-38. Berdasarkan analisa corak pemetilan genom, 1 motif jujukan *N*6-

metiladenin (^{m6}A) yang berstruktur palindromik, GTWWAC, telah dikenalpasti wujud dalam kesemua genom strain *Pandoraea*. Selain itu, anotasi gen sistem *restriction-modification* (R-M) turut mengenalpasti 1 enzim metiltransferase tunggal Jenis II yang berpotensi sebagai calon metiltransferase bagi motif GTWWAC. Seterusnya, analisa bioinformatik yang dijalankan mencadangkan bahawa calon metiltransferase tersebut mewakili 1 kelas metiltransferase tunggal yang baharu dan boleh dijumpai dalam famili *Burkholderiaceae* dan *Oxalabacteraceae*. Analisis perbandingan corak pemetilan motif GTWWAC turut mencadangkan bahawa enzim metiltransferase tersebut, yang dinamai *Pandoraea adenine methylase* (PAM), mempunyai fungsi yang menyerupai Dam (*DNA adenine methyltransferase*) dan CcrM (*cell-cycle regulated methyltransferase*), kedua-dua metiltransferase tersebut tersebar antara kelas *Gammaproteobacteria* dan *Alphaproteobacteria*. Kesimpulannya, selain membentangkan profil metilom genus *Pandoraea*, kajian ini juga melaporkan penemuan 1 kelas metiltransferase tunggal yang baharu dalam kelas *Betaproteobacteria*.

Kata Kunci: *Pandoraea***,** metilom, metiltransferase tunggal, GTWWAC

ACKNOWLEDGEMENTS

Firstly, my utmost gratitude goes to my supervisor Assoc. Prof Dr. Chan Kok Gan for his continuous encouragement, trust and guidance during the journey of my Ph.D. study. This study would not have come into fruition without his support and mentorship. I am deeply grateful to Sir Richard Roberts who introduced me to the fascinating world of epigenomic landscape of bacteria and whose continual support and enlightening advices have made this thesis possible. A huge thank you goes out to Assoc. Prof. Scott Beatson, Dr. Jayde Gawthorne, Dr. Brian Forde and Ms. Melinda Ashcroft from University of Queensland who have generously shared their advices and expertise whenever I encountered bioinformatics challenges.

Furthermore, I extend my sincere gratitude to Ms. Yin Wai Fong, our laboratory manager whom excellent financial and laboratory management skills have helped ensure smooth conduct of all experiments and research matters pertaining to this study. I would also like to gratefully acknowledge the financial support from High Impact Research (HIR) grants and Graduate Research Assistantship Scheme (GRAS).

My deep appreciation goes to my fellow lab members and scientists whom I have had the fortune to be acquainted or collaborated with during the course of my Ph.D. study. Their excellent work, ideas and advices have inspired and enlightened me in various aspects towards the completion of this work. Last but not least, I would like to say a heartfelt thank you to my family who have given me their invaluable trust, confidence and moral support during this challenging yet rewarding journey.

TABLE OF CONTENTS

CHAPTER 3: MATERIALS AND METHODS

CHAPTER 4: RESULTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

°C	:	Degree Celcius
^{m4}C	:	4-methylcytosines
^{m5}C	:	5-methylcytosine
^{m6}A	:	6-methyladenine
aa	:	Amino acid
bp	:	Base pair
CcrM	:	Cell cycle-regulated methyltransferase
CCS	:	Circular consensus sequencing
CDSs	:	Coding sequences
CLR	:	Continuous long reads
CF	:	Cystic fibrosis
DNA	:	Deoxyribonucleic acid
Dam	:	DNA adenine methyltransferase
Dcm	:	DNA cytosine methytransferase
GR	:	Genic regions
gDNA	:	Genomic DNA
GC	:	Guanine-cytosine
HGAP	:	Hierarchical Genome Assembly Process
HGT	:	Horizontal gene transfer
IGR	:	Intergenic regions
IPD	:	Inter-pulse duration
kb	:	Kilobase
LB	:	Luria Bertani
MTases	:	Methyltransferases
MGE	:	Mobile genetic elements
modQV	:	Modification quality value
MEGA	:	Molecular Evolutionary Genetics Analysis
AHL	:	*N*-Acyl homoserine lactones
nt	:	Nucleotide
n	:	Number of samples
PacBio	:	Pacific Biosciences
PBS	:	Phosphate buffered saline

PFGE	:	Pulsed-field gel electrophoresis
QV	:	Quality Value
QS	:	Quorum sensing
RAST	:	Rapid Annotation using Subsystem Technology
REases	:	Restriction endonucleases
R-M	:	Restriction-modification
RPM	:	Revolution per minute
RNase	:	Ribonuclease
RNA	:	Ribonucleic acid
SAM	:	*S*-adenosyl-methionine
SMRT	:	Single Molecule Real Time
SD	:	Standard deviation
TRD	:	Target recognition domains
REBASE	:	The Restriction Enzyme Database
TBE	:	Tris Borate EDTA
v/v	:	Volume/volume
WGA	:	Whole-genome amplified
ZMW	:	Zero-mode waveguides

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

The genus of *Pandoraea* is a taxa in the beta subclass of *Proteobacteria* described in year 2000 (Coenye et al., 2000). Members of *Pandoraea* are frequently reported as emerging opportunistic pathogens with multi-drug resistance properties and are mainly recovered from patients suffering from cystic fibrosis (Ee et al., 2015; Schneider et al., 2006; Stryjewski et al., 2003). However, despite the mounting documentations, the understanding regarding the epidemiology, clinical significance and pathogenicity of this genus remain partial (Degand et al., 2015). Besides their clinical importance, species of *Pandoraea* were also reported to be recovered from soil environment and have remarkable bioremediation activity (Ee et al., 2014; Gómez-Gil et al., 2007; Han-Jen et al., 2013; Okeke et al., 2002; Pham et al., 2012). To date, no methylome characterisation has been performed for members of the *Pandoraea* genus.

Epigenetics describe the heritable changes in gene expression that occur due to modification of DNA sequence epigenetic signals, wherein DNA methylation represent one of the most common example of epigenetic signalling tool (Casadesús & Low, 2006). These enzymatically mediated base modifications have greatly expanded the structural complexity and information depth of the hereditary genetic information we can derive from the four canonical bases. Three most common prokaryotic base modifications are *N*4-methyl-cytosine (^{m4}C), C5-methyl-cytosine (^{m5}C) and *N*6-methyl-adenine (^{m6}A) (Roberts et al., 2010). Majority of the known bacterial methylome elements comprise restriction-modification (R-M) systems which typically consist of a combination of methyltransferases (MTases) and restriction endonucleases (REases) and can be further classified into four main classes, types I to IV (Murray, 2000; Pingoud et al., 2005; Rao et al., 2013; Roberts et al., 2003; Wilson, 1991; Wilson & Murray, 1991b). The R-M systems are most commonly recognised with their role as the cellular defense system

which protect the host's DNA from invasion of foreign DNA elements although additional cellular portfolios of these systems have also been discovered (Roberts et al., 2010; Vasu & Nagaraja, 2013). Intriguingly, some MTases, particularly the orphan MTases, were discovered to have pivotal regulatory roles in gene expression, phase variation, population evolution, DNA replication control, DNA mismatch repair, and most importantly, pathogenicity (Murphy et al., 2013; Wion & Casadesus, 2006). These orphan MTases are believed to be mobilized and cross transmitted in bacteria *via* horizontal gene transfer events by using mobile genetic elements or could be derived from ancestral R-M systems which lost the cognate REases (Kobayashi, 2001; Murphy et al., 2013). The best examples of these MTases that have been studied in depth were namely Dam (DNA adenine methyltransferase) in *Gammaproteobacteria* and CcrM (cell cycle regulator methyltransferase) in *Alphaproteobacteria* (Low et al., 2001; Marinus & Casadesus, 2009b).

Contrary to DNA base modifications in higher eukaryotes which have been extensively studied, the assessment of their functional importance in prokaryotes have been forthcoming at a slower pace. This is mainly due to the laborious and time-consuming aspect of traditional molecular biology methodologies in prokaryotic methylome studies as well as the difficulties in detection of DNA methylation besides 5-methylcytosine (^{m5}C) (Korlach & Turner, 2012). The advent of single-molecule real time (SMRT) sequencing technology circumvents the methodological limitations and provides a high throughput solution to genome-wide detection and analysis of the ^{m6}A, ^{m4}C and ^{m5}C modifications in bacterial genome (Flusberg et al., 2010). In addition, REBASE (The Restriction Enzyme Database), a comprehensive and fully curated database on the information of R-M systems also provides a valuable resource in simplifying and aiding in the rapid advancement of methylome research. The wealth of data on putative R-M systems in combination with the gold standard reference set of experimentally

characterized R-M components in REBASE enable accurate annotation of putative R-M systems and assignment of the most probable MTase candidates for a specific recognition motif in completely sequenced genome (Roberts et al., 2015). The availability of these predictions provides a robust guideline for downstream experimental analysis required to validate and characterize the functional role of these R-M systems. The integration of data generated from both SMRT sequencing and REBASE enable bioinformatics analysis and characterization of the impact of bacterial methylome in the context of genome welfare.

Although the progress of prokaryotic methylome studies have been accelerated significantly in recent years, our understanding on the roles of the methylome in bacterial genetics and physiology remains fragmentary due to several factors. Firstly, majority of the literature on impact of DNA methylation on pathogenicity focused largely on enteric pathogens and the gamma subdivision of *Proteobacteria*, for instance *Campylobacter jejuni, Bacteroides dorei, Salmonella spp., Escherichia coli, Vibrio cholerae and Yersinia pseudotuberculosis* (Garcia-Del Portillo et al., 1999; Julio et al., 2001; Leonard et al., 2014; Mou et al., 2015; van der Woude & Low, 1994). To date, only a handful of studies were reported on regulatory roles of DNA methylation in relation with global gene expression of respiratory pathogens and other pathogenic species of *Betaproteobacteria* (Chen et al., 2003; Lluch-Senar et al., 2013; Mehling et al., 2007; Sater et al., 2015; Watson et al., 2004). Secondly, there are also few literatures which explore the functional roles of DNA methylation in environmental bacterial strains, especially those with metabolic and bioremediation potentials (Bendall et al., 2013). Therefore, in this study, the first report that illustrates the methylome landscape of the *Pandoraea* genus which encompasses pathogenic strains and environmental strains was performed.

The objectives of this study are:

1. To perform complete genome sequencing of *Pandoraea* spp.
2. To conduct pan-genus genome analyses on the complete genomes of *Pandoraea* spp.
3. To perform methylome profiling and comparative methylome analyses of *Pandoraea* spp.

CHAPTER 2: LITERATURE REVIEWS

2.1 The Genus *Pandoraea*

Pandoraea is a genus described by Coenye et al. (2000), and comprises a group of rod-shaped, gram negative, motile, aerobic, non-fermentative and non-spore-forming organisms. The name of the genus was derived from the name of Pandora's box of Greek mythology, drawing it's meaning as the origin of mankind's diseases. This genus was discovered as a result of an extensive polyphasic taxonomic study on several sputum isolates collected from cystic fibrosis (CF) patient (misidentified as *Burkholderia* cepacia, *Ralstonia paucula* and *Ralstonia pickettii*-like organisms) which resulted in the reclassification of these isolates into this novel genus (Coenye et al., 2000).

To date, following the list of prokaryotic names with standing in nomenclature (http://www.bacterio.net/pandoraea.html), this genus contains a total of 9 validated species, namely: *P. apista*, *P. faecigallinarum*, *P. norimbergensis*, *P. oxalativorans*, *P. pnomenusa*, *P. pulmonicola*, *P. sputorum*, *P. thiooxydans*, and *P. vervacti*, where *P. apista* represents the type species of the genus (Anandham et al., 2010; Coenye et al., 2000; Sahin et al., 2011). Majority of the isolates in this genus were recovered from clinical samples such as respiratory secretions and blood samples, particularly from samples of patients with cystic fibrosis. Various clinical studies have highlighted the pathogenicity, invasive potential, transmissibility, and potential in chronic colonisation of *Pandoraea* spp. (Atkinson et al., 2006; Daneshvar et al., 2001; Stryjewski et al., 2003). These have led to the proposed assignment of possibly emerging pathogen status to the isolates as well as implementation of contact isolation procedures on infected patients (Jørgensen et al., 2003). The potential clinical outcome of *Pandoraea* spp. colonisation as described by some clinical reports are lung function declination, bacteremia and formation of high antibody titers (Atkinson et al., 2006; Jørgensen et al., 2003). However,

due to a lack of data and occurrence of co-infection with other pathogens, the prognosis and impact of colonisation of these isolates remain ambiguous. In addition to its potential in causing serious infection, the multidrug-resistant pattern of the *Pandoraea* spp. clinical isolates also attract concerns from clinicians regarding the emergence of these pathogens among the CF community. These isolates demonstrate resistance to a wide variety of commonly used broad spectrum antimicrobial agents, including majority of the β-lactam agents and aminoglycosides (Daneshvar et al., 2001). Besides its pathogenic role, various *Pandoraea* environmental species were also associated with attractive biotechnological potential such as biodegradation of environmental pollutants as well as lignolytic and oxalotrophic potential (Bandounas et al., 2011; Colbert et al., 2013; Jiang et al., 2009; Jin et al., 2007; Liz et al., 2009; Okeke et al., 2002; Ozaki et al., 2007; Sahin et al., 2011; Siddique et al., 2003).

After almost 16 years since the discovery of this genus, several major knowledge gaps concerning this genus could be identified. Firstly, the pathogenicity mechanism and infection outcome of the pathogenic species of this genus remain incompletely understood. The issue of mixed infection and lack of reliable clinical data remain the main hindrance in studies pertaining to pathogenicity elucidation. Furthermore, many research articles frequently highlight the challenges in accurate species identification of this genus *via* conventional microbiology methods. Availability of high quality completed genomes for this genus could be valuable for filling these knowledge gaps by enabling an *in silico* comprehension of the biology of the *Pandoraea* spp.

2.2 Single Molecule Real Time Sequencing

Single molecule Real Time (SMRT) sequencing technology, is a third generation sequencing technology that have revolutionised genome sequencing, particularly in the aspect of generating completely sequenced genomes, by circumventing the restrictions posed by various second generation sequencing technologies (Schadt et al., 2010). These restrictions are namely short read lengths and amplification biases (Aird et al., 2011; Dohm et al., 2008). The SMRT technology harnesses the inherent properties of DNA polymerases (including speed, fidelity and processitivity) as a sequencing machine and visualises the individual nucleotide incorporation events catalysed in real-time (Eid et al., 2009). SMRT technology encompasses 2 main parts, namely the template preparation and the sequencing process.

2.2.1 SMRTbell Template

SMRTbell™ template is a template format which consists of a double-stranded DNA fragment (termed as insert region) flanked on both ends by single stranded hairpin loops (Travers et al., 2010). The completed template is topologically circular and contains 2 complementary strands information of the target DNA sequence of interest as illustrated in Figure 2.1. The template sequence can subsequently be bound with a primer and a DNA polymerase molecule to generate a sequencing-productive complex.

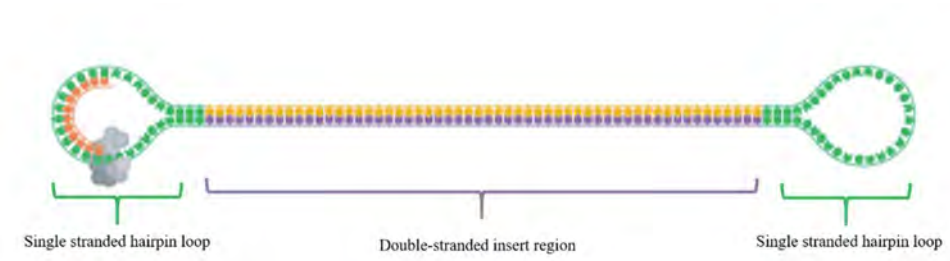


Figure 2.1: Schematic diagram of a SMRTbell™ template. The template format consists of a double-stranded insert region (coloured yellow and purple) capped on both ends by hairpin loops (green). The hairpin loops comprise primer-complementary (coloured orange) single-stranded sequences and are served as a binding point for the DNA polymerase (coloured grey) to initiate DNA synthesis process. The image is reprinted with kind permission from Oxford University Press. Image originally published in Travers et al. (2010), page 3 of 8, Figure 1 (A).

This nature of template format enables the utilisation of SMRT sequencing technology in constructing a high accuracy consensus sequence by using multiple reads generated from repeated observations of both sense and antisense strands of the same insert sequence region. This method of sequencing is referred to as the circular consensus sequencing (CCS) application and is one of the main application of SMRT sequencing (Eid et al., 2009; Larsen & Smith, 2012). The key advantages of this template format are: ability in accommodating a wide range of insert sizes for different sequencing purpose; speed and simplicity of construction; independence from amplification steps and hence bias elimination; and suitability for circular consensus sequencing (Travers et al., 2010).

2.2.2 Sequencing Mechanism

The main principle of SMRT sequencing is capturing real-time kinetic process of DNA polymerisation during the natural replication process of the target DNA molecule (Eid et al., 2009). During the sequencing process, the DNA polymerase, which is bound to the SMRTbell template, incorporates fluorescent-labelled DNA bases as the enzyme reads and determines the DNA sequence of the nascent template strand (Korlach et al., 2010). As the 4 nucleotide bases (A, C, T and G) were labelled with 4 different fluorescent dyes, distinct emission spectrums will be emitted following the incorporation of different

bases and hence enable easy recording of the replication process in the form of "movie" of light pulses and subsequent interpretation of the bases which were synthesized (Figure 2.2) (Eid et al., 2009; Korlach et al., 2010).

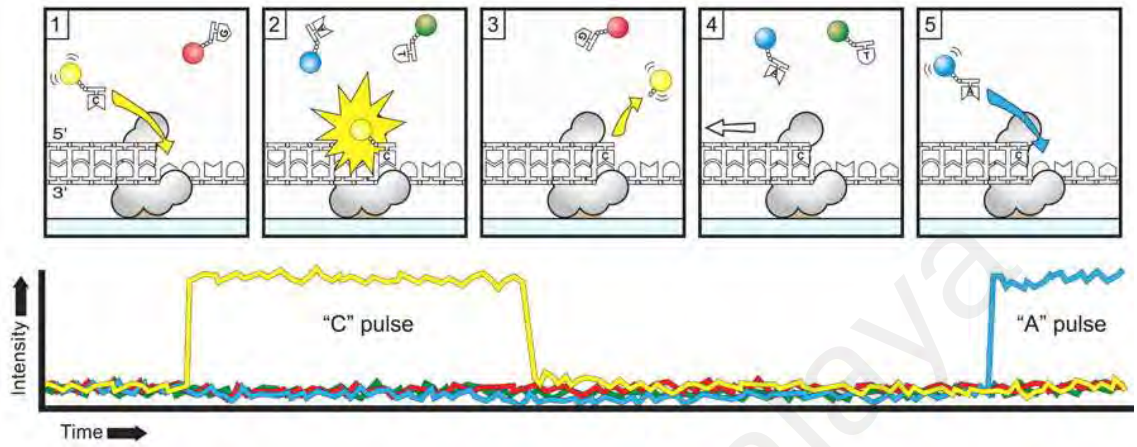


Figure 2.2: Illustration of the principle of SMRT sequencing process. The 4 nucleotide bases (A, C, T, and G) are labelled with blue, yellow, green and red fluorescent dye respectively. Firstly, the DNA polymerase reads the nascent DNA template and binds a fluorescently-labelled nucleotide which is complementary to the template base at its active site. The fluorescence intensity of the colour corresponding to the incorporated base (for example the "C" pulse which is coloured in yellow) will be elevated. Subsequently, the dye-linker-pyrophosphate product will be cleaved from the nucleotide base and diffuse out of the zero-mode waveguides (ZMW) which then terminated the fluorescence pulse. The process then proceeds as the DNA polymerase translocates to the next position and incorporates the next nucleotide and initiates the next fluoresent phase. This figure is from J. Eid et al. (2009), page 134, Figure 1(B), hyperlink: https://d2ufo47lrtsv5s.cloudfront.net/content/sci/323/5910/133/F1.large.jpg. Reprinted with permission from American Association for the Advancement of Science (AAAS).

The key component of efficient real-time monitoring of the sequencing process in SMRT technology is the SMRT cell. SMRT cell is a chip which contains 150,000 sequencing units termed as zero-mode waveguides (ZMWs). ZMWs are photonic nanostructured cylindrical holes made on an opaque metallic cladding film and deposited onto a transparent silica substrate of which the DNA polymerase (bound with the SMRTbell template) are tethered on during the sequencing process (Levene et al., 2003). The design of ZMW, as illustrated in Figure 2.3, creates a highly confined observation volume to the immediate vicinity of the DNA polymerase enzyme. This design provides

the advantage of efficient detection of each fluorophore signal generated by the polymerase activity despite the background noises of free-nucleotide analogs and diffusing fluorophore products (Jonas Korlach & Turner, 2013). Therefore, by using the SMRT cell chip, real time observation of the DNA polymerase base incorporation events can be done at massive parallelism.

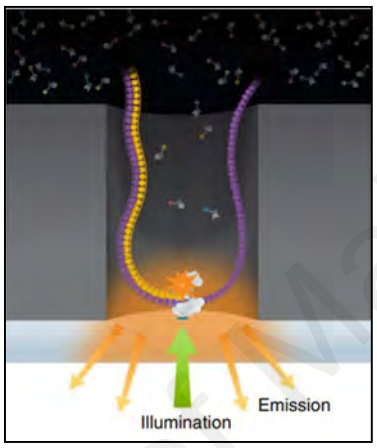


Figure 2.3: Zero-mode waveguide (ZMW) schematic representation. The figure illustrates the example of real-time DNA synthesis which occur in a single ZMW with a DNA polymerase immobilised to the bottom of the well. This figure is reprinted by permission from Springer Customer Service Centre GmbH: Springer Nature, Encyclopedia of Biophysics, (Zero-Mode Waveguides, Jonas Korlach, Stephen W. Turner Korlach J., Turner SW) © European Biophysical Societies' Association (2013).

Furthermore, the main strength of SMRT sequencing is the ability to generate long read lengths (Heiner et al., 2013). Due to flexibility of the SMRTbell construct which could accommodate long insert regions, the main limiting factor of read lengths generation is the lifespan of DNA polymerase during the sequencing process. The latest generation of DNA polymerase (P6) coupled with the proprietary sequencing chemistry (C4) were reported to generate read lengths which are over 10 kb on average and with N50 of more than 20 kb and a maximum read length of over 60 kb (Rhoads & Au, 2015). The long read length generated from SMRT sequencing provides several advantages, including: ease in resolving repetitive elements and complex regions, possibility of high quality *de novo* genome assembly, unambiguous sequence alignments, full length

isoforms characterization, observation of full phased alleles and resolution of structural variants (Liang et al., 2016; Ritz et al., 2014; Roberts et al., 2013; Shin et al., 2013). However, this technology is not without shortfalls. Firstly, the throughput of SMRT sequencing is lower as compared to other second generation sequencing technologies, the typical throughput of the PacBio RS II system is 0.5 to 1 billion bases per SMRT cell (Quail et al., 2012). This is due to the nature of ZMWs which typically only have a 23.3% to 46.7% productivity rate due to loading failure or template molecules overload into a ZMW. This lower throughput could contribute to a higher sequencing cost particularly when sequencing larger genomes. Secondly, the error rates of the sequenced raw reads (termed as continuous long reads (CLR)) are also relatively higher (approximately 11% to 15%) (Quail et al., 2012).

2.3 Application of SMRT Sequencing in Prokaryotic Genome Research

Two applications of SMRT sequencing were utilized in this thesis, namely *de novo* long reads assembly and base modification analysis.

2.3.1 *De novo* Assembly using Hierarchical Genome Assembly Process

Hierarchical Genome Assembly process, abbreviated as HGAP, is a nonhybrid assembly workflow which aim to generate high quality finished genome using the SMRT sequencing long reads exclusively (Chin et al., 2013). This consensus algorithm taps on the features of SMRT reads such long read length, random nature of sequencing errors, and the potential in achieving high consensus accuracy (Ferrarini et al., 2013).

HGAP constitutes 3 well-defined steps as illustrated in Figure 2.4 ("HGAP," 2016). (1) Preassembly. The main aim of preassembly step is to generate highly accurate preassembled reads sequences. This aim is achieved by firstly, generating a seeding sequence data set by selecting the longest portion of reads from the overall filtered subreads (Chin et al., 2013). The amount of seed reads selected is recommended to

represent at least 30 fold of target genome coverage (i.e. if the target bacterial genome is estimated to be 5 MB, the total bases of seed reads should ideally be 150 MB). Subsequently, the remaining shorter single pass reads from the subreads pool are aligned and mapped to the seed reads followed by a directed acyclic graph-based consensus step to error-correct the seed reads. Then, through a quality trimming step (in accordance to Quality Value (QV) score of the consensus sequences which reflect the confidence in the mapped coverage), low-quality and chimeric sequenced reads will be eliminated. From these steps, a highly accurate preassembled reads can be generated. (2) Assembly. The preassembled and corrected reads are subsequently assembled using Celera Assembler, an overlap-layout-consensus approach-based assembler (Denisov et al., 2008). The success of this assembly process in achieving full prokaryotic genome closure is highly reliant on the availability of high coverage (> 50 fold of target genome size) and sufficient length distribution of preassembled reads in order to resolve the regions with high genomic complexity. (3) Consensus Polishing. The consensus polishing step aims to improves accuracy of the assembly by mapping all single pass reads to the assembled contigs and subsequently uses Quiver, a quality-aware consensus algorithm, for variant calling ("Quiver FAQ," 2013). Quiver derives accurate consensus call (achieving > 99.999% accuracies) according to the per-base Quality Values (QV) embedded in the mapped single pass reads ("HGAP," 2016).

Figure 2.4: Illustration of HGAP operating principle. Single pass reads are mapped to the seed reads to construct preassembled reads in the preassembly steps. The preassembled reads were subsequently assembled to produce the final assembly. Quiver consensus-calling step is not shown in this illustration. This figure is reprinted by permission from Springer Customer Service Centre GmbH: Springer Nature, Nature Methods, (Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, Chin et al.) © Nature America, Inc. (2013).

HGAP permits a simplified workflow in generating high quality and contiguous *de novo* prokaryotic genome assemblies with comparable quality to those of Sanger sequencing with the needs of only a single template library preparation. However, the shortfall of this assembly process is the reads coverage requirement for a high quality assembly which could be render the sequencing cost prohibitively expensive for larger genomes.

2.3.2 Base Modification Detection

The nature of SMRT sequencing which monitors DNA polymerase kinetics provides a high-throughput platform for direct detection of native DNA modification at a strand- and base-pair resolution (Feng et al., 2013). The core metric used in DNA modification events detection with SMRT sequencing data is inter-pulse duration (IPD) measurement, which is a precise measurement of the duration between successive fluorescence pulses resulted from nucleotide incorporation events (as described in detail in chapter 2.2.2) catalysed by DNA polymerase. The IPD measurements directly reflect DNA polymerase kinetics which are sensitive to the primary and secondary structure of

the template DNA strand and therefore serve as a metric to detect structural perturbation on the template DNA (Feng et al., 2013; Schadt et al., 2012).

Presence of modified bases on the template DNA strand contribute significantly to increment of IPD due to the impact of the DNA structure alteration onto DNA polymerase kinetics (As illustrated in Figure 2.5). This increment in IPD can be quantified in the form of IPD ratio, which is defined as the normalised ratio between mean IPD of the sequenced DNA template to the null IPD distribution of a homologous site on the unmodified control template sequence ("Methylome Analysis Technical Note," 2017; Flusberg et al., 2010). Current algorithm incorporated within the base modification analysis pipeline of SMRT data permits the utilisation of either a whole-genome amplified (WGA) control sample of which all modifications were erased through the WGA process or an *in silico* control which is a computational model with predicted null average IPD values for each reference position ("White Paper: Base Modifications," 2015; Feng et al., 2013). A position is determined to be modified if the likelihood ratio calculated exceeded a statistical threshold.

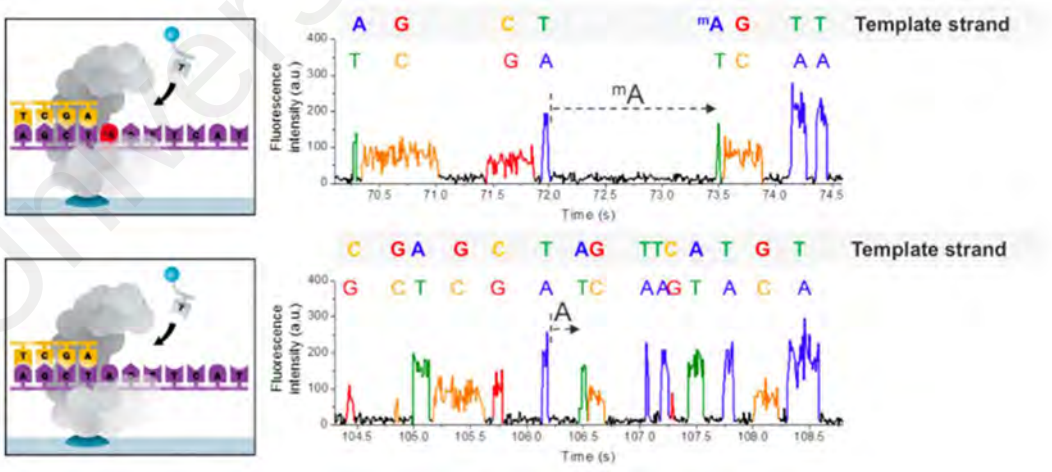


Figure 2.5: Illustration of DNA polymerase kinetics alteration upon encounter of a modified DNA bases during SMRT sequencing. As shown in the top figure, the presence of a methylated adenine base (^{m6}A) resulted in an extended duration between the incorporation of 2 successive nucleostides, and hence a longer interpulse duration (IPD) as compared to the IPD observed between these nucleotides in the unmodified template (bottom figure). This figure is reprinted by permission from Springer Customer Service Centre GmbH: Springer Nature, Nature Methods, (Direct detection of DNA methylation

during single-molecule, real-time sequencing, Flusberg et al.) © Nature America, Inc. (2010).

Each type of DNA base modification generates a reproducible sequence-context dependent IPD alteration pattern which are collective referred to as kinetic signatures (J. Korlach et al., 2010). These kinetic signatures, when incorporated in the modification identification model of SMRT Analysis algorithm, allow an unambiguous detection of different DNA modification types, including: ^{m6}A, ^{m4}C and ^{m5}C (Figure 2.6) (Fang et al., 2012; Flusberg et al., 2010). Detection sensitivity of ^{m5}C can be enhanced *via* Tet1 enzyme treatment which converts 5-methylcytosine bases into 5-carboxylcytosine (Clark et al., 2013; Feng et al., 2013). Moreover, although not integrated in SMRT Analysis algorithm, other modifications such as 8-oxoguanine, O6-methylguanine, 5-hydroxymethyluracil and thymine dimers which represent products of DNA damage could also be detected using SMRT sequence data (Clark et al., 2011).

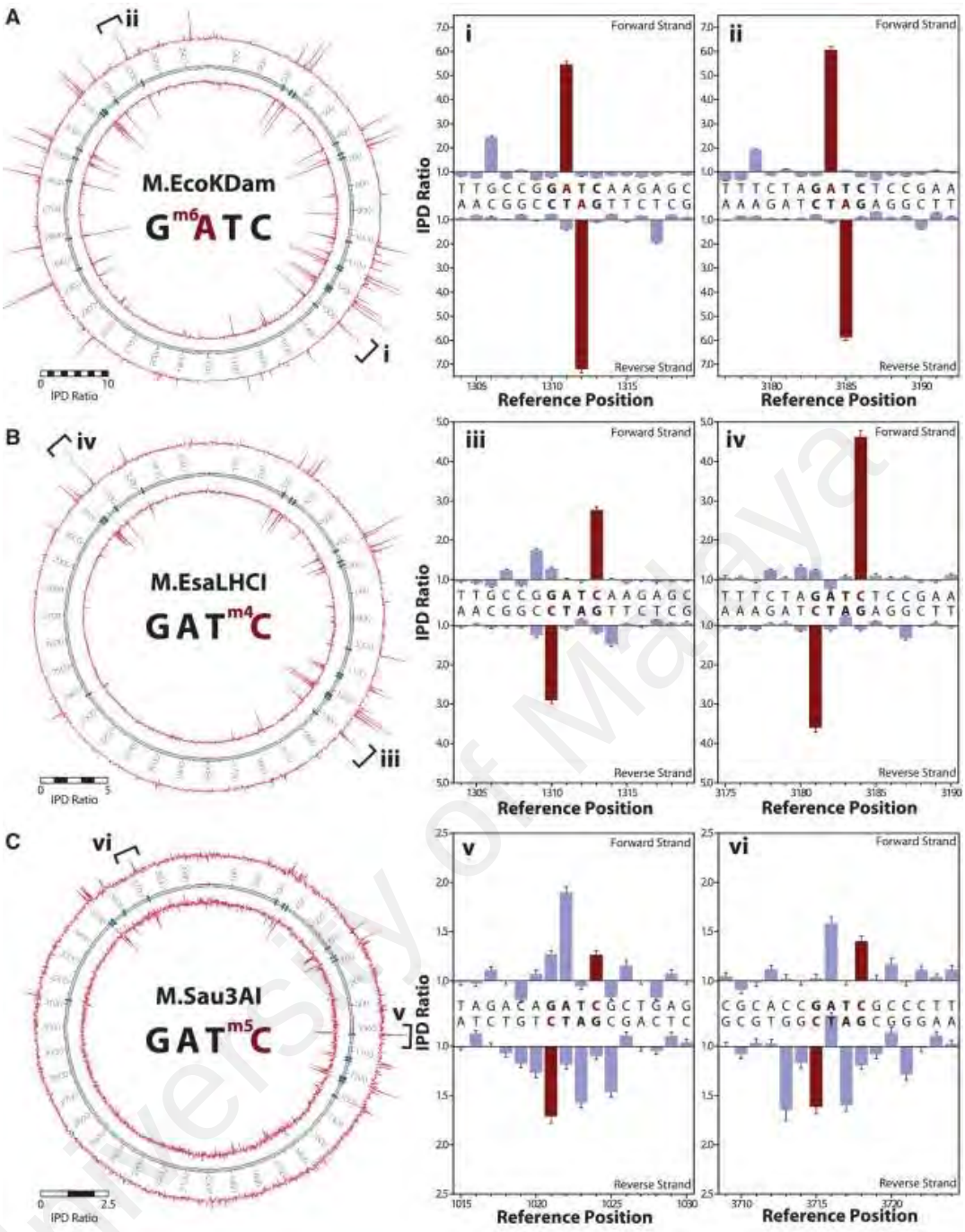


Figure 2.6: Kinetic signatures of (A) ^{m6}A, (B) ^{m4}C, and (C) ^{m5}C as determined from SMRT sequencing. The left panel displays circus plots of IPD ratio of each methylation type on both DNA template strands as detected on the recombinant plasmids which contain the corresponding methyltransferases. The right panel shows a close-up view of IPD ratios on representative template positions containing the target sequence contexts. The bars highlighted in red represent the methylated template positions. All 3 methylation types cause an obvious increment in IPD ratio at the vicinity of the methylated bases. The typical characteristics for each modification type are: ^{m6}A, IPD ratio hike at modified position and 5 bases downstream; ^{m4}C, IPD ratio hike at modified position; ^{m5}C, typical kinetic signal pattern at 2 and 6 bases after the modified methylated position. This figure is reprinted with kind permission from Oxford University Press. Image originally published in Clark et al. (2012), page 7 of 12, Figure 4.

SMRT sequencing and its associated base modification detection application have significant advantages over other base modification techniques, particularly in prokaryotic methylome research, in the aspect of: high throughput, detection of ^{m6}A modification, relative technical ease and the yield of genome-wide modified bases data at base pair and strand-specific resolution (Davis et al., 2013). These benefits have facilitated in expansion of prokaryotic methylome research in the past 3 years. Furthermore, application of SMRT sequencing data in eukaryotic DNA methylation research were also observed to be gaining traction in the past 2 years, including researches on CpG sites and mammalian embryonic stem cells (Grunert et al., 2016; Pfeifer, 2016; Suzuki et al., 2016; Wu et al., 2016).

2.4 DNA Methylation in Bacteria

DNA methylation which is the process by which methyl groups are added to the DNA molecules represent one of the most common form of post-replicative DNA modification which orchestrate epigenetic inheritance (Jin et al., 2011). To date, 3 types of methylation types were identified in bacterial genomes, namely *N*6-methyl-adenine (^{m6}A), C5-Methyl-cytosine (^{m5}C), and *N*4-methyl-cytosine (^{m4}C), amongst which only ^{m4}C is found but to be exclusively in bacteria to date (Sanchez-Romero et al., 2015). Formation of these methylation types are catalysed by DNA methyltransferases which recognise specific DNA sequence motifs as substrate and catalyse the transfer of a methyl group from *S*-adenosyl-methionine (SAM) to the target DNA bases (Jeltsch, 2002).

The profound biological consequences of DNA methylation in conveying additional epigenetic information onto prokaryotic genomes are well-established (Casadesus & Low, 2006). However, due to a highly diverse nature of the enzymes which mediate the process and methodological challenges in obtaining base modification data, the nature, extent of distribution and biological significance of prokaryotic DNA methylation remains largely understudied.

2.4.1 Prokaryotic DNA Methyltransferases (MTases)

To date, 3 functional classes of DNA MTases were identified in prokaryotic genomes with two different mechanisms of action (Roberts et al., 2015). While all classes catalyse the transfer of a methyl group from *S*-adenosyl-L-methionine (SAM), the first class utilises the exocyclic amino group of adenine and cytosine bases of duplex DNA as substrate, resulting in ^{m6}A and ^{m4}C whereas the second class transfer onto C5 position on the heterocycle of cytosine to yield 5-methylcytosine (^{m5}C) (Figure 2.7) (Gromova & Khoroshaev, 2003; Wu & Santi, 1987).

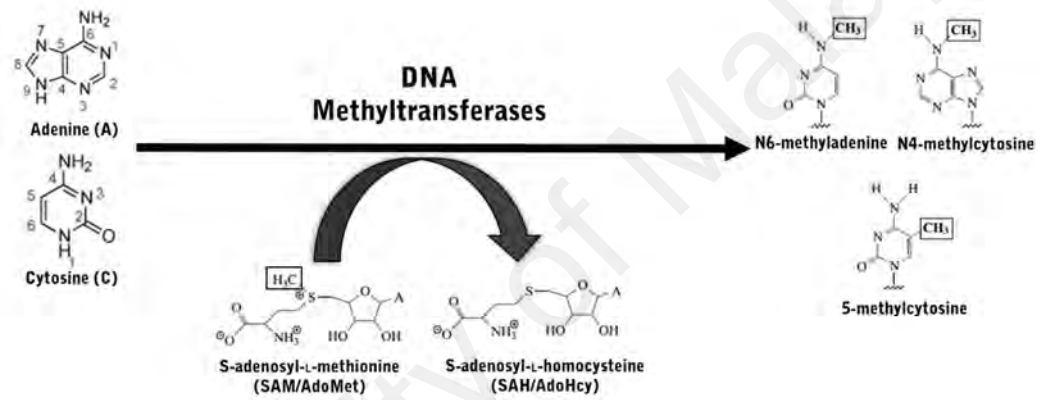


Figure 2.7: Illustration of mechanism of methyl group transfer as catalysed by DNA methyltransferases. The process yield ^{m6}A, ^{m4}C and ^{m5}C modification by *S*-adenosyl-L-methionine (SAM) as cofactor.

2.4.2.1 Classification of DNA MTases

X-ray structures of all 3 types of MTases showed that these MTases constitute 2 main globular proteins which comprise a large and a small domain, separated by a DNA-binding cleft (Cheng & Roberts, 2001; Jeltsch, 2002). The large domain contains conserved regions which encode SAM-binding and catalytic centre of the enzyme (Gromova & Khoroshaev, 2003; Kumar et al., 1994). On the other hand, the small domain of different DNA MTases are known as target recognition domains (TRD) and constitute highly variable regions which are responsible for determination of target DNA sequence

specificities (Jeltsch, 2002). On the other hand, the large domains of all MTases consist of a similar structural core, a β-sheet with 7 strands flanked by 1 or 2 α-chains. The 7th strand is sandwiched between strands 5 and 6 and is positioned in an antiparallel manner to all strands (6↓7↑5↓4↓1↓2↓3↓) (Malone et al., 1995).

The large subdomains contain conserved amino acid motifs which sequential order of arrangement are used to classify DNA MTases. In C5 MTases, a total of 10 conserved amino acid motifs (I – X) which are in consistent motif orders were identified (Cheng et al., 1993b). Motifs I, II, III and X are accountable for SAM binding activity, and are therefore termed as the SAM binding region whereas motifs IV, V, VI, VII and VIII are responsible for catalysis and hence are referred to as the catalytic region (Cheng et al., 1993a; Cheng et al., 1993b; Schluckebier et al., 1995). Motif IX is exclusively identified in the small domain of C5 MTases. On the other hand, N6 and N4 MTases contain 9 conserved motifs, corresponding to the I to VIII and X motifs of the C5 MTases in the aspect of structure and functions. According to the arrangement order of SAM-binding region, catalytic region and TRD region, these N MTases could be categorised into 3 different subgroups (α, β and ɣ) as summarised in Table 2.1 (Malone et al., 1995).

Table 2.1: Classification of DNA methyltransferases according to arrangement order of conserved amino acid motifs.

Family	Group	Motif order
C5		N-I-II-III-IV-V-VI-VII-VIII-TRD-IX-X-C
N6, N4	α	N-X-I-II-TRD-III-IV-V-VI-VII-VIII-C
N6. N4	β	N-IV-V-VI-VII-VIII-TRD-X-I-II-III-C
N6, N4	ɣ	N-X-I-II-III-IV-V-VI-VII-VIII-TRD-C

2.4.2.2 Restriction-Modification Systems

Prokaryotic DNA MTases are commonly linked with a cognate restriction endonuclease (REase) which together these enzymes form a R-M system. Both enzymes in a R-M system exhibit contrasting activity where the MTase catalyse genome-wide

methylation of a specific target recognition sequence and the cognate REase recognises and cleaves the unmethylated target sequences (Bickle & Kruger, 1993).

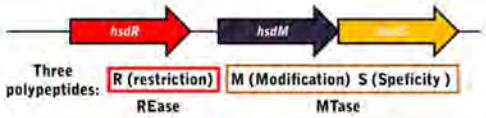
R-M systems are classified into 4 main types (I-IV) according to their enzyme structures, cofactors requirement, target recognition and cleavage patterns (Roberts et al., 2003) (Table 2.2). Type I system constitutes a hetero-oligomeric enzyme complex which comprises 3 protein subunits namely HsdR, HsdM and HsdS (Murray, 2000). *hsdM* and *hsdS* are co-transcribed and the encoded subunits are both necessary for methyltransferase activity whereas *hsdR* is transcribed independently and the product is required for restriction activity. HsdS (also known as specificity (S) subunit) which contains 2 TRDs is responsible for imparting target sequence specificities to both HsdM and HsdR (Obarska-Kosinska et al., 2008). Target sequences recognised by the Type I enzymes are bipartite and asymmetric in structure in which the MTases methylate adenine residues on both strands of the target sequences whereas the associated REases cleaves at approximately 400 – 7000 nt away from the target recognition site (Roberts et al., 2003). In contrast to the interdependency of the Type I systems enzymes, MTase and REase of the Type II system has independent enzymatic activities (Wilson, 1988). Moreover, target recognition sequences of the Type II systems are palindromic in structure and both methylation and DNA cleavage occurs within the symmetrical recognition site (Wilson & Murray, 1991b). Type III system, on the other hand, comprises a multisubunit enzyme complex encoded by 2 genes (*mod* and *res*) (Rao et al., 2013). While the modification (Mod) subunit could catalyse DNA methylation activity independently, restriction activity requires interaction of both subunits (Kauc & Piekarowicz, 1978). The target recognition site of Type III system consists of a short stretch (5 – 6 nt) of uninterrupted and asymmetrical DNA sequences (Rao et al., 2013). Interestingly, full modification of target recognition sites by the Mod subunit is only on 1 strand, leading to the formation of a hemimethylated motif. Due to the nature of this asymmetrical modification, DNA

cleaving requires interaction of the enzyme with 2 copies of unmodified sites in inverse repeat orientation (Meisel et al., 1992). Lastly, Type IV system which represents the latest R-M systems identified, encodes enzymes which cleaves modified (methylated, hydroxymethylated, or glucosyl-hydroxymethylated) DNA bases (Roberts et al., 2003). The sequence specificity of this system remains unknown.

R-M systems are commonly known for their cellular defense role which is akin to the innate immune system of bacteria where the sequence-specific DNA methylation activity of DNA MTases provide a unique "self" epigenetic identity and the REases cleave and degrade DNA sequences which lacks the identity ("non-self") (Arber & Linn, 1969; Wilson & Murray, 1991b). Through this mechanism, the invasion of extraneous DNA elements, particularly phages, which lack the similar epigenetic identity can be effectively curtailed (Bickle, 2004; Tock & Dryden, 2005). The epigenetic identities provided by the R-M systems also contribute towards speciation control and rate of microbial evolution. The restriction of foreign DNAs serve as immigration control of genetic flux between prokaryotes of different lineages which bear different epigenetic identitities and hence contribute to genetic isolation and species identity preservation (Corvaglia et al., 2010; Murray, 2002). Furthermore, following this model, the acquisition and accumulation of epigenetic identity variations due to addition of new R-M genes, could subsequently lead to a distinct variant strain ("biotype") and subsequently to a speciation event (Jeltsch, 2003). Beyond the protective role, the R-M genes also demonstrated attributes of selfish genetic elements, where a dependency is created onto the host genome by the lethal effect imposed by REase-mediated cell death, hence enhancing the survival and relative propagation frequency of these genes (Kobayashi, 2001; Naito et al., 1995). Furthermore, the addiction effect also contribute to a stabilisation effect onto the host genome as the genomic region harbouring R-M genes cannot be easily displaced thereby facilitating in preservation of genomic islands (Vasu

& Nagaraja, 2013). In addition to these roles, R-M systems are also proposed to have functional roles in host resources reallocations, facilitation of recombination machinery, natural genetic engineering, and gene expression regulations (Vasu & Nagaraja, 2013).

Table 2.2: Summary of R-M systems classification according to enzyme structure, recognition sequence pattern, REase cleaving pattern and cofactor requirements.

R-M Types	Type I	Type II	Type III	Type IV
Enzyme structure:				
Recognition sequence :	Asymmetric and bipartite eg. 5'-GCAGNNNNNTCC-3'	4 to 8 nt; Symmetric (continuous/interrupted) eg. 5'-GAATCC-3' / 5'-GANTC-3'	5 to 6 nt; asymmetric (uninterrupted) eg. 5'-CGAAT-3'	Not well studied
Cleaving pattern:	Cleavage occurs at variable distance away from recognition sequence	Cleavage occurs symmetrically within the recognition sequence	Cleavage occurs approximately 25 nt away from the recognition sequence. (requires presence of 2 unmodified recognition sequence in opposite orientation)	REase recognizes and cleaves only methylated DNA

2.4.2.3 Orphan DNA MTases and DNA Adenine Methylation

Beyond R-M systems, the MTases could also be found to occur and function independently without association with a corresponding restriction enzyme, these MTases are termed orphan MTases or solitary MTases (Murphy et al., 2013). Overall, 2 classes of orphan MTases can be identified, namely the *bona fide* orphan MTases and the transiently orphaned MTases, which are solitary MTases generated as a result of R-M systems degradation wherein the REase are selectively degraded as a result of selective pressure (Seshasayee et al., 2012). As compared to the transiently orphaned MTases, *bona fide* orphan MTases have a higher degree of evolutionary stability and lineage coherence and are found to encode core cellular functions (Matveyev et al., 2001). Both types of MTases can serve as a resource for molecular vaccines against lethal effect of R-M systems parasitism (Kobayashi, 2001; Takahashi et al., 2002).

To date, deoxyadenosine methylase (Dam), cell cycle-regulated methyltransferase (CcrM) and DNA cytosine methytransferase (Dcm) represent the three most well-studied orphan MTases. Dam and CcrM catalyse methylation of adenine moieties in the target sequences of 5'-GATC'3' and 5'-GAnTC-3' respectively whereas Dcm methylates the interior cytosine base of 5'-CCWGG-3' sequences (Gonzalez et al., 2014; Marinus & Løbner-Olesen, 2014; Palmer & Marinus, 1994; Urig et al., 2002). Amongst these MTases, Dam and CcrM demonstrate properties of a *bona fide* orphan MTase. In the aspect of lineage coherence and evolutionary stability, Dam homologs are distributed among members of *Gammaproteobacteria* and CcrM within the *Alphaproteobacteria* lineage, both with high level of sequence similarities (Blow et al., 2016; Jeltsch, 2002).

Moreover, these orphan MTases represent an attractive research topic as the DNA adenine methylations catalysed by these MTases were found to have significant biological implications. The implication of DNA adenine methylation onto cellular functions are due to the effect of adenine methylation on DNA protein interactions *via* modification of

DNA's thermodynamical stability and curvature which could cause steric effects onto protein binding (Diekmann, 1987; Engel & von Hippel, 1978; Sternberg, 1985). Hence, the methylation patterns (fully methylated, hemimethylated or unmethylated) of the DNA adenine MTases' recognition sequences have direct impact on expression of associated genes. Various cellular activities which are well-established to be regulated by DNA adenine methylation as catalysed by orphan MTases are summarised in Table 2.3. The implications of DNA adenine methylation in these cellular roles, particularly in the aspect of pathogenicity, have raised the possibility of targeting DNA adenine methylation as a strategy to potentiate antibiotic treatments or as a novel antibacterial treatment approach.

Table 2.3: Summary of well-established cellular roles of Dam and CcrM.

MTase	Cellular roles	Reference
Dam	**DNA mismatch repair (methyl-directed):** Facilitate base mismatch errors which arise from DNA replication by guiding the mismatch repair action of MutH onto newly synthesised strand by using methylation state of GATC motifs for strand discrimination (Unmethylated strand: newly synthesised; Methylated strand: parental strand and used as template).	(Modrich, 1991) (Pukkila et al., 1983) (Hu et al., 2017)
	Regulate chromosome replication: Control timing of replication process by using methylation status of GATC motifs. SeqA binds to hemimethylated GATC motifs in *oriC* and *dnaA* promoter region to sequester DNA replication process and hence ensuring a single occurrence of *oriC* initiation per cell cycle.	(Campbell & Kleckner, 1990) (Bakker & Smith, 1989) (Boye & Løbner-Olesen, 1990)
	Gene expression regulation: Presence and methylation status of GATC motifs in promoter and regulatory sequences can affect binding of RNA polymerases and transcriptional regulators and hence influence gene expression. Gene expression can also be modulated *via* DNA methylation pattern (DMP) of GATC sites as evidenced by Pap (pyelonephritis-associated pilus) phase variation model.	(Marinus & Casadesus, 2009a) (Casadesus & Low, 2006) (Low et al., 2001) (Weyand & Low, 2000)
	Transposition: Synthesis and activities of transposases are coupled with hemimethylation of GATC motifs. This mechanism of control is proposed to curtail potential deleterious effect of excessive transposition. Example of transposons which activity were found to be regulated by Dam methylation are Tn*10*, Tn*5*, TN*903* and insertion element IS*3*.	(Roberts et al., 1985) (Yin et al., 1988)
	Phage genes regulation Methylation status of GATC motifs regulate packaging of phage P1 DNA during lytic stage of the phage cycle and facilitate in phage propagation. Dam-mediated methylation also have transcriptional control over phage protein production and lysogeny maintenance (example: 933W phage which encodes Shiga toxin). In addition, Dam methylation also provide selective advantage in enhancing success of phage invasion.	(Murphy et al., 2013) (Murphy et al., 2008)
	Virulence factor: Association of Dam methylation with virulence was discovered with the observation of virulence attenuation of pathogens including *Escherichia coli*, *Salmonella* spp., *Yersinia pseudotuberculosis* and *Vibrio cholera*. The linkage with virulence is due to the effect of Dam-mediated methylation onto virulence gene expression.	(Marinus & Casadesus, 2009b) (Low et al., 2001)
	Antibiotic stress survival: GATC methylation are indicated to support survival of *E. coli* in antibiotic stress particularly antibiotics from β-lactam and quinolone classes.	(Cohen et al., 2016) (Adam et al., 2008; Balbontín et al., 2006)
CcrM	**Regulate chromosome replication in genome of stalked cell:** The state of methylation of GAnTC motifs orchestrates molecular events of DNA replication and transcription of DNA replication proteins in a manner which is synchronous with the cell cycle progress. The process is largely similar to those coordinated by Dam methylase.	(Stephens et al., 1996) (Reisenauer et al., 1999)
	Gene expression regulation: The methylation status of GAnTC motifs were determined to mediate gene expression regulation in *C. crescentus* genome, including the expression of CcrM.	(Reisenauer & Shapiro, 2002) (Gonzalez & Collier, 2013)
	Virulence CcrM was implicated in supporting pathogenicity of *Brucella abortus*, an intracellular pathogen, by supporting intracellular survival of the pathogen in macrophages.	(Robertson et al., 2000)

CHAPTER 3: MATERIALS AND METHODS

3.1 Reagents

All chemical reagents utilized in this study were obtained from the following sources (in alphabetical order):

1. Amresco® (USA)
2. Lonza (Switzerland)
3. Merck Millipore (Germany)
4. NextGene (Malaysia)
5. Scharlau (Spain)
6. Sigma-Aldrich® (USA)

3.2 Commercial Kits

All commercial kits used in this study were listed as follows (in alphabetical order):

1. Agilent DNA 12000 Kit (Agilent, USA)
2. Agilent High Sensitivity DNA Kit (Agilent, USA)
3. Ampure PB MagBead Kit (Pacific Biosciences, USA)
4. BluePippin™ Cassette Kit PAC20KB (Sage Science, USA)
5. DNA Sequencing Reagent Kit 2.0 (Pacific Biosciences, USA)
6. DNA Sequencing Reagent Kit 4.0 v2 (Pacific Biosciences, USA)
7. DNA/Polymerase Binding Kit P4 (Pacific Biosciences, USA)
8. DNA/Polymerase Binding Kit P6 v2 (Pacific Biosciences, USA)
9. MasterPure™ DNA Purification Kit (Epicentre, USA)
10. PacBio RS II SMRT Cell Oil (Pacific Biosciences, USA)
11. Qubit® dsDNA Broad Range (BR) Assay Kit (Life Technologies, USA)
12. Qubit® dsDNA High Sensitivity (HS) Assay Kit (Life Technologies, USA)
13. SMRTbell Template Prep Kit 1.0 (Pacific Biosciences, USA)

3.3 Equipments

The equipments used in this study were listed as follows (in alphabetical order):

1. Agilent 2100 Bioanalyzer (Agilent, USA)
2. Allegra® X-15R Benchtop centrifuge (Beckman Coulter, USA)
3. BenchMixer™ (Benchmark Scientific Inc., USA)
4. Centrifuge 5424 R (Eppendorf, Germany)
5. DNA LoBind Tubes (Eppendorf, Germany)
6. Force Mini centrifuge (Select Bioproducts, USA)
7. Galileo Biosciences RapidCast Complete Mini-Gel System (Sage Science, USA)
8. Gel documentary image analyser (UV Products, Canada)
9. Glacier NU-9668 Upright Large Capacity -86°C Ultra Low Freezer (NuAire, USA)
10. G-TUBE™ (Covaris, USA)
11. HVE-50 autoclave machine (Hirayama, Japan)
12. Incubator (Memmert GmbH, Germany)
13. Maxymum recovery Pipette tips (Axygen, USA)
14. Mediline Lab Freezer (Liebherr, Switzerland)
15. Microflex MALDI-TOF MS (Bruker Daltonik GmbH, Germany)
16. MilliQ® Integral water purification system (Merck Millipore, Germany)
17. Mini-rotator Bio RS-24 (Biosan, Latvia)
18. MixMate® Vortex Mixer (Eppendorf, Germany)
19. NanoDrop 2000 UV-Vis Spectrophotometer (Thermo Scientific, USA)
20. Pipette tips (Axygen, USA)
21. Pipettes (Eppendorf, Germany)
22. Pippin Pulse Power Supply (Sage Science, USA)
23. PowerPac™ Basic Power Supply (Bio-Rad Laboratories Inc., USA)

24. Qubit 2.0 fluorometer (Life technologies, USA)
25. Single Molecule Real Time (SMRT) RSII sequencer (Pacific Biosciences, USA)
26. Sub-Cell® GT Agarose Gel Electrophoresis System (Bio-Rad Laboratories Inc., USA)
27. T100™ Thermal Cycler (Bio-Rad Laboratories Inc., USA)
28. Thermomixer comfort (Eppendorf, Germany)
29. ThermoStat™ C (Eppendorf, Germany)

3.4 Bacterial Strains

In this study, a total of 9 *Pandoraea* type strains and 3 in-house *Pandoraea* isolates were included. All strains were cultured using Luria Bertani (LB) media at the suitable culture temperature as listed in Table 3.1. The isolation source and culture condition of each strains are as listed in Table 3.1.

Table 3.1: List of *Pandoraea* strains used in this study.

Name	Isolation source	Culture temperature (°C)	Remark
Pandoraea apista DSM-16535^T	Sputum from cystic fibrosis patient (Denmark)	37	Type strain (DSMZ[a])
Pandoraea faecigallinarum DSM 23572^T	Chicken dung (India)	28	Type strain (DSMZ[a])
Pandoraea norimbergensis DSM 11628^T	Oxic water layer above a sulfide-containing lake sediment (Lake Silbersee, Germany)	28	Type strain (DSMZ[a])
Pandoraea oxalativorans DSM 23570^T	Soil litter close to oxalate-producing plants (Turkey)	28	Type strain acquired from DSMZ[a]
Pandoraea pnomenusa DSM-165356^T	Sputum from cystic fibrosis patient (Edinburgh, United Kingdom)	37	Type strain acquired from DSMZ[a]
Pandoraea pulmonicola DSM-16583^T	Sputum from cystic fibrosis patient (Canada)	37	Type strain acquired from DSMZ[a]
Pandoraea sputorum DSM-21091^T	Sputum from cystic fibrosis patient (USA)	37	Type strain acquired from DSMZ[a]
Pandoraea thiooxydans DSM-25325^T	Rhizosphere soils of sesame (Sesamum indicum L.)	30	Type strain acquired from DSMZ[a]
Pandoraea vervacti DSM-23571^T	A soil enrichment culture with 6 g potassium oxalate l 21 as the sole source of carbon and energy	28	Type strain acquired from DSMZ[a]
Pandoraea pnomenusa RB38	Landfill soil (Malaysia)	28	In-house isolates

[a] Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (German Collection of Microorganisms and Cell Cultures GmbH)

3.4.1 Bacterial Strains Maintenance and Storage

For routine cultivation, all strains were cultured using LB medium (Scharlau, Spain) at 28 °C. On the other hand, for storage purpose, all strains were kept in both LB agar slants (in room temperature) and in 20 % (v/v) glycerol stock (in -80 °C). The 20 % glycerol stock comprised a mixture of sterilised 80 % v/v glycerol (200 µL) and bacterial isolates planktonic culture (800 µL).

3.5 Buffer Solution

3.5.1 Phosphate Buffered Saline (PBS)

To prepare 1× phosphate buffered saline (PBS), 10× PBS stock solution was diluted 1:10 (v/v) using milli-Q water and was subsequently sterilized by autoclaving (121 °C, 15 psi, 15 minutes).

3.5.2 Tris Borate EDTA (TBE) Buffer

10× Tris Borate EDTA (TBE) stock solution (pH 8.3) (First Base Laboratory, Malaysia) was diluted 1:10 (v/v) using milli-Q water to the desired concentration of 1× and 0.5× TBE buffer respectively. 1× TBE buffer was used to prepare agarose gels containing 0.8 % (w/v) of agarose (NextGene, Malaysia) whereas 0.5× TBE buffer was used to prepare TBE gel which contains 1 % (w/v) SeaKem® Gold Agarose (Lonza, USA). These buffers were also used as running buffer.

3.6 DNA Ladder Marker

Two types of DNA ladder markers were used in this study, namely GeneRuler™ 1kb DNA ladder (Fermentas International Inc., Canada) and 1 Kb DNA Extension Ladder (Invitrogen, USA).

3.7 Complete Genome Sequencing using Single Molecule Real Time (SMRT) Sequencing Technology

3.7.1 Genomic DNA (gDNA) Isolation

Genomic DNA purification was done using MasterPure^{TM} DNA purification kit per manufacturer's instruction. Each strain was cultured overnight in LB broth and was subjected to cell pellet wash for three times using 1× PBS buffer prior to cell pellet collection *via* centrifugation followed by supernatant removal. The purification process was constituted of two steps, firstly the cell lysis step and followed by DNA precipitation step. For cell lysis, each sample was added with a mixture of Proteinase K (1 µL) and Tissue and Cell Lysis Solution (300 µL) and was mixed thoroughly. Each tube of sample was then incubated at 65 °C along with vortex agitation at every 5 minutes interval. The samples were cooled to 37 °C and were subsequently mixed adequately following addition of RNase A (1 µL, 5 µg/µL). A final incubation at 37 °C was performed for 30 minutes and the samples were chilled for 3 to 5 minutes prior to DNA precipitation process.

For precipitation of gDNA, the lysed samples were firstly added with 175 µL of MPC protein precipitation reagent followed by vigorous mixing. The mixture was subsequently pelleted *via* centrifugation (4 °C, 10 minutes, ≥10,000×*g*) and the resultant supernatant was transferred to a new tube whereas the pelleted debris were discarded. Subsequently, 500 µL of isopropanol was added to each tube of recovered supernatant and the tubes were inverted 30 to 40 times prior to centrifugation (4 °C, 10 minutes). Isopropanol was subsequently discarded and the DNA pellets were rinsed twice with freshly prepared 70 % (v/v) ethanol. Lastly, all residual ethanol was removed by evaporation prior to resuspension in EB buffer (Qiagen, Germany).

3.7.2 gDNA Quantitation and Quality Assessment

Firstly, accurate quantitation of double-stranded DNA concentration was assessed using Qubit Fluorometer (Life Technologies, USA). The Qubit® dsDNA broad range (BR) assay kit was used following manufacturer's protocol. Secondly, NanoDrop® spectrophotometer 2000 (Thermo Scientific, USA) along with the corresponding Nanodrop® software (version 1.5) (Thermo Scientific, USA) were used to assess the purity of the gDNA by determining the A260/280 (ratio of absorbance at wavelength of 260 nm and 280 nm) and A260/230 (ratio of absorbance at wavelength of 260 nm and 230 nm) ratio of each sample. The reading of A260:A280 ratio (acceptable range: approximately 1.8) provided a primary assessment of gDNA purity whereas A260:A230 ratio (acceptable range: 2.0 to 2.2) provided an inference of the quality of gDNA extraction process and a secondary measurement of gDNA purity. Thirdly, 0.8 % (w/v) agarose gel electrophoresis was used to evaluate the integrity of each gDNA sample, approximately 150 ng (according to Qubit measurement) of gDNA was loaded into each lane and GeneRuler™ 1kb DNA ladder was used as marker for the gel run.

Following the guidelines in "PACBIO® GUIDELINES FOR SUCCESSFUL SMRTbell™ LIBRARIES", only samples with intact high molecular DNA band and with both A260/280 and A260/230 ratio that are within range were used to proceed into template preparation.

3.7.3 gDNA Fragmentation

Covaris g-TUBE was used as the shearing device for all gDNA shearing steps, Eppendorf centrifuge model 5424 R was used for g-TUBE centrifugation following manufacturer recommendation. gDNA of all type strains were sheared to an average size of 17 to 20 kb by shearing approximately 10 µg of gDNA (eluted in 100 µL buffer EB, QIAGEN) at several spin cycles of 4800 revolutions per minute (RPM) for 2 minutes. When small volume persisted in the upper chamber of g-TUBE after two spin cycles, an

additional 50 µL of buffer EB was added and mixed adequately *via* pipetting prior to the final spin cycle. On the other hand, gDNA of the in-house strains were sheared into an average size of 8 to 10 kb by shearing 8 µg of gDNA (eluted in 150 µL of buffer EB) at several spin cycles of 6000 RPM for 1 to 2 minutes.

For the gDNA samples of the in-house strains (expected fragment size: 8 to 10 kb), gDNA fragment size assessments were done using Agilent 2100 bioanalyzer with DNA 12000 kit (Agilent Technologies, USA) following manufacturer's protocol. On the other hand, size assessments of the sheared gDNA samples of type strains (expected fragment size of 17 to 20 kb) were done *via* pulsed-field gel electrophoresis (PFGE) utilising the Pippin Pulse electrophoresis instrument (Sage Science, USA) to enable accurate sizing estimation of the large DNA fragments. 1 % TBE gel was used to perform a 14 hours gel run and 0.5× TBE buffer was used as the running buffer. The 1 % (w/v) TBE gel (pH 8.3) was prepared by melting 1.1 g of Lonza SeaKem® agarose powder in 110 mL of 0.5× TBE buffer and 1 kb DNA extension ladder (Invitrogen, USA) was used as the DNA marker. Pre-set protocol which target to resolve DNA size range of 5 to 80 kb (details as listed in Table 3.2) was selected on the Pippin Pulse software (version 1.32) (Sage Science, USA) to control the run voltage, forward and reverse time steps and number of steps per cycle of the pulsed field electrophoresis gel run.

Table 3.2: Details of Pippin Pulse protocol parameters selected for the pulsed-field gel electrophoresis.

Parameter	Function of parameter	Value
V	Electrophoresis voltage	75
A	Forward time at start of run	150
B	Reverse time at start of run	50
C	Increment added to **A** at each step	30
D	Increment added to **B** at each step	10
E	Increment added to **C** at each step	3
F	Increment added to **D** at each step	1
G	Number of steps per cycle	48

3.7.4 SMRTbell™ Template Library Preparation

All type strains were processed into greater than 10 kb SMRTbell library following the "Procedure & Checklist-20 kb Template Preparation using BluePippin™ Size Selection" protocol whereas all the in-house strains were constructed into SMRTbell library with an approximate size of 10 kb following the "Procedure & Checklist - 10 kb Template Preparation and Sequencing" protocol. The sheared gDNA was firstly purified using 0.45× volume ratio of Ampure®® PB Beads followed by several steps leading to the final SMRTbell™ library construction, including: DNA damage repair, ends repair, blunt end-ligation of hairpin adapters, exonuclease digestion, and SMRTbell™ purification. Reagents from the SMRTbell template prep kit 1.0 (Pacific Biosciences, USA) were utilised in the library preparation process.

For all type strains, the purified SMRTbell™ template was subjected to an additional size-selection step using the BluePippin system and the corresponding dye-free gel cassettes (PAC20KB). Cassette definition "0.75 % DF Marker S1 high-pass 6-10 kb vs3" and pre-set protocol "SMRTbell 20kb, 7000 bp High-Pass" were used to perform the size selection run. The eluted size-selected SMRTbell™ was purified with 1.0× volume ratio of Ampure® PB beads prior to eluting in buffer EB.

3.7.5 Annealing, Polymerase Binding and Sequencing Steps

Based on the estimated concentration and the insert size of the SMRTbell™ library as well as the target number of SMRT cells to be used, annealing and binding calculator version 2.1.0.2 and version 2.3.1.1 were used to set up the annealing and binding reactions for the 10 kb and 20 kb SMRTbell™ library respectively. In the primer annealing step, the primer was first diluted and conditioned by going through a 80 °C melting step and was subsequently mixed with the SMRTbell™ template to a final concentration of 0.8333 nM. Further, P4 and P6 DNA polymerases were bound to the primer-annealed 10 kb and 20 kb SMRTbell™ template respectively in the polymerase

binding reaction step with the presence of binding buffer, DTT and nucleotides. SMRT sequencing was carried out on a PacBio RSII sequencer (Pacific Biosciences, USA) using Magbead loading protocol. Stage-start setting and 180-minutes movie collection time was used. The 10 kb SMRTbell™ libraries were sequenced in 4 SMRT cells whereas the 20 kb SMRTbell™ libraries were sequenced in 3 SMRT cells.

3.8 Genome Assembly and Post-Assembly Processing

3.8.1 *De novo* Assembly using HGAP

All genome assemblies were performed using Hierarchical Genome Assembly Process (HGAP) algorithm (version 2 and version 3) included in the SMRT Analysis software suite (version 2.3.0) and was accessed through the SMRT Portal user interface (Chin et al., 2013). The HGAP workflow constitutes 4 main steps, namely the polymerase reads filtering step, preassembly step, *de novo* assembly step and assembly polishing step (Koren & Phillippy, 2015). Between HGAP version 2 (optimised for quality) and version 3 (optimised for speed), the module used in the pre-assembly step and *de novo* assembly step is different, which contribute to the performance difference between these two protocols. The function and SMRT® Pipe module implemented for each components of the workflow are summarised in Table 3.3.

Table 3.3: Summary of the function of each HGAP workflow component and the SMRT® Pipe modules implemented in each component.

Workflow component	Function	SMRT® Pipe modules
Filtering	To filter and trims raw reads (polymerase reads) produced by primary analysis software to generate single pass subreads.	P_filter
Pre-assembly	To perform mapping of all single pass subreads to seed reads to generate highly accurate pre-assembled reads.	HGAP 2: P_PreAssembler HGAP 3: P_PreAssemblerDagcon
De novo assembly	To perform *de novo* assembly of the corrected pre-assembled reads into unitigs.	HGAP 2: P_CeleraAssembler HGAP 3: P_AssembleUnitig
Assembly Polishing	To align all single pass subreads against the assembled contigs and to perform draft assemblies polishing using Quiver	P_Mapping and P_AssemblyPolishing

Briefly, raw reads obtained from sequencing were pre-processed and were further filtered using default parameters to generate subreads. The filtered subreads were subsequently proceeded for *de novo* assembly using the Hierarchical Genome Assembly Process (HGAP). Firstly, all samples will be assembled using the computed subread length cut-off which correspond to 30× of target genome coverage. After the initial assembly, samples which were not assembled into closure were then proceeded with multiple rounds of assembly using HGAP version 3 to determine the optimum seed read length cut-off based on the subread filtering graph. Seed read length which demonstrated improvements in the contiguity of tested assemblies were selected to be used in generating final assembly using HGAP version 2. Contiguity of sequence contigs were examined and visualised using contig adjacency graphs constructed using Contiguity software (version 1.04) (Sullivan et al., 2015). The polished assemblies (FASTA formatted) were exported for further analysis.

3.8.2 Circularisation

The assembled genomes were determined as a complete assembly when self-overlapping ends were identified in the assembled contig. By using Gepard (Krumsiek et al., 2007), a rapid dot plot tool, coordinates of the overlapping regions in each genome were pinpointed and trimmed to generate a blunt-ended circular genome sequence. Additionally, the trimmed genome sequences were imported into SMRT portal to generate final polished consensus using Quiver consensus algorithm (Chin et al., 2013).

3.8.3 Genome Annotation and Bioinformatics Analyses

Various bioinformatics tools were utilised to perform genome annotation and analyses. Firstly, the circularised and rearranged genomes were annotated using SEED-based automated annotation system incorporated in RAST server (Aziz et al., 2008; Tatusova et al., 2013) with default settings and Glimmer2 tool was selected as the gene prediction program (Delcher et al., 1999). From the result of the annotation (**Appendix**

A (i)), location of the *dnaA* gene associated with the *dnaA* operon were identified and all genomes were subsequently rearranged to align at the *dnaA* gene. Furthermore, positional homology multiple genome alignments of these genomes were constructed and visualised using the progressiveMauve program included in the Mauve genome alignment package (version 2.4.0) (Darling et al., 2010).

Prediction of *oriC* of each gen omes were also performed using a combination of three rules: analysis of chromosome asymmetry, DnaA box clusters distribution and *dnaA* gene location (Mackiewicz et al., 2004). Ori-Finder was used to determine DnaA box clusters distribution and to perform chromosome asymmetry analysis (http://tubic.tju.edu.cn/Ori-Finder) (Gao & Zhang, 2008). Additional analysis to detect presence of prophages was performed using Phage Search Tool (PHAST) where only prophages scored as intact were reported and used for further analysis (Zhou et al., 2011). TnpPred tool (Riadi et al., 2012) was also used to identify potential transposases in the genomes.

3.9 Base Modification Analysis

RS_Modification_and_Motif_Analysis.1 pipeline, an accompanying tool in the SMRT analysis software suite version 2.3.0, was used to perform genome-wide detection of modified bases and identification of associated motif. Default settings of the pipeline was used. Each rearranged and circularised genome were imported to be used as *in silico* control, which is the polymerase kinetics computational model utilised to calculate interpulse duration (IPD) ratio (Clark et al., 2012; Flusberg et al., 2010). The IPD ratio is calculated by comparing the average observed IPDs on each site of the native DNA with the IPD of the *in silico* control, a two-sample t-test was used to analyse the comparison of which a quality value (QV) score was calculated following the equation of QV=-log (p-value).

Minimum modification QV (modQV) of 30 (corresponds to p-value of 0.001) were used as a threshold value to identify the sites with base modification event. Furthermore, "modification identification" function of the pipeline was also enabled. The modification identification function enabled comparison of the modification signal identified to an additional computational model which contains the expected characteristic signature of the three known DNA modification types namely, ^{m6}A, ^{m4}C, and ^{m5}C which further improved the detection accuracy of modification signals and enable categorisation of modification type. The associated methyltransferase recognition motifs were identified using the MotifFinder tool integrated in the pipeline by congregating recurring context of modifications.

3.10 R-M Systems Annotation

Prediction of methyltransferases (MTases) and relevant R-M genes from all complete genome sequences were conducted using the SEQWARE computer resources in conjunction with REBASE internal database (Clark et al., 2012; Roberts et al., 2015). R-M components were firstly identified based on sequence matches from BLAST searches, and the annotation was further refined according to known genomic context and order of predictive functional domains within the predicted proteome (Klimasauskas et al., 1989; Pósfai et al., 1989). Furthermore, the candidate MTases were assigned, when possible, to the detected MTase recognition motifs based on matching R-M systems classifications and known characteristics of MTase homologs (Roberts et al., 2015). Each identified R-M components was also named according to proposed nomenclature of R-M components (Roberts et al., 2003). The annotated R-M genes of each genome was deposited in REBASE.

3.11 Cloning and Over-Expression of *M.PpnI* Gene

A methylase (*M.PpnI*) gene was amplified using Q5® High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, USA) with gene-specific primers as listed in Table 3.4. The DNA insert was cloned into PCR-amplified pRRS vector using the Gibson Assembly® Cloning kit (New England Biolabs, Ipswich, USA) and subsequently transformed into cloning vector *E.coli* NEB5α. The ampicillin-resistant transformants were selected by plating on LB-agar containing 100 µg/mL of ampicillin and LB broth with similar concentration of ampicillin was used to maintain the selected transformant. Subsequently, the pRRS vector harbouring *M.PpnI* gene was reisolated from overnight culture of the transformant and the sequence was verified *via* Sanger sequencing. The recombinant construct was then transformed into dam^-/dcm^- chemically competent *E. coli* (New England Biolabs, Ipswich, USA). The genomic DNA of the *E. coli* recombinant strain was subjected to SMRTbell™ template preparation and SMRT sequencing to determine the resulting methylation pattern. Plasmid sequences were confirmed by re-sequencing the PacBio reads against the plasmid reference.

Table 3.4: Gene-specific primers used in cloning of *M.PpnI* gene

Primer	Description	Sequence
pRRS_fwd	Forward sequence of pRRS vector	CCCGGGGAAGATCTAGATCTAGATAG
pRRS_rev	Reverse sequence of pRRS vector with overlapping region of *Ppn* gene	AACTTCCACCTTACCTGCAGGCATGCAAGCTTGGC GTAATCATGG
Ppn_fwd	Forward sequence of M.*PpnI* gene with overlapping region with pRRS vector	GCAGGTAAGGTGGAAGTTATGACCGATCTGACTGA TCGCAAGGCACAG
Ppn_rev	Reverse sequence of *Ppn* gene with overlapping region with pRRS vector	CTAGATCTTCCCCGGGTTACCCTGCGCTTGCAAGGC TTGGAACGTC

3.12 Phylogenetic Analyses of GTWWAC MTases

Phylogenetic analysis were conducted using Molecular Evolutionary Genetics Analysis (MEGA) software, version 6.06 (Tamura et al., 2013). Nucleotide sequences were aligned based on codons using MUSCLE (Edgar, 2004). Maximum likelihood method based on the Kimura 2-parameter model (Rates among sites: Gamma distributed with invariant sites (G+I)) with 1000 bootstrap replicates was used to construct the phylogenetic tree. M.Eco9387Dam, a REBASE gold standard Type II subtype alpha orphan MTase was included as outgroup.

3.13 Genome-Wide GTWWAC Methylation Frequency Distribution Analysis

3.13.1 Analysis of GTWWAC Sequence Motif-Associated Adenine Bases

For each genome, the genome positions of all adenine bases associated with GTWWAC sequence motif (referred as adenine bases from hereon) were obtained from the motifs.gff output file which comprise information of all modified sites, locations of all discovered motifs as well as the combined information of the modification sites and the motifs. SMRT® View genome browser was utilised to visualise and peruse the modification data. Figure 3.1 shows the flow chart of the methodology used to analyse the adenine bases.

Firstly, all adenine bases were categorised as methylated or unmethylated according to per base value of modQV. The unmethylated adenine bases (modQV lower than 30) were manually examined to be classified as the unmethylated adenine bases associated with unmethylated or hemimethylated GTWWAC motifs. Subsequently, the methylated bases associated with the hemimethylated motifs were identified and withdrawn from the dataset of methylated adenine bases.

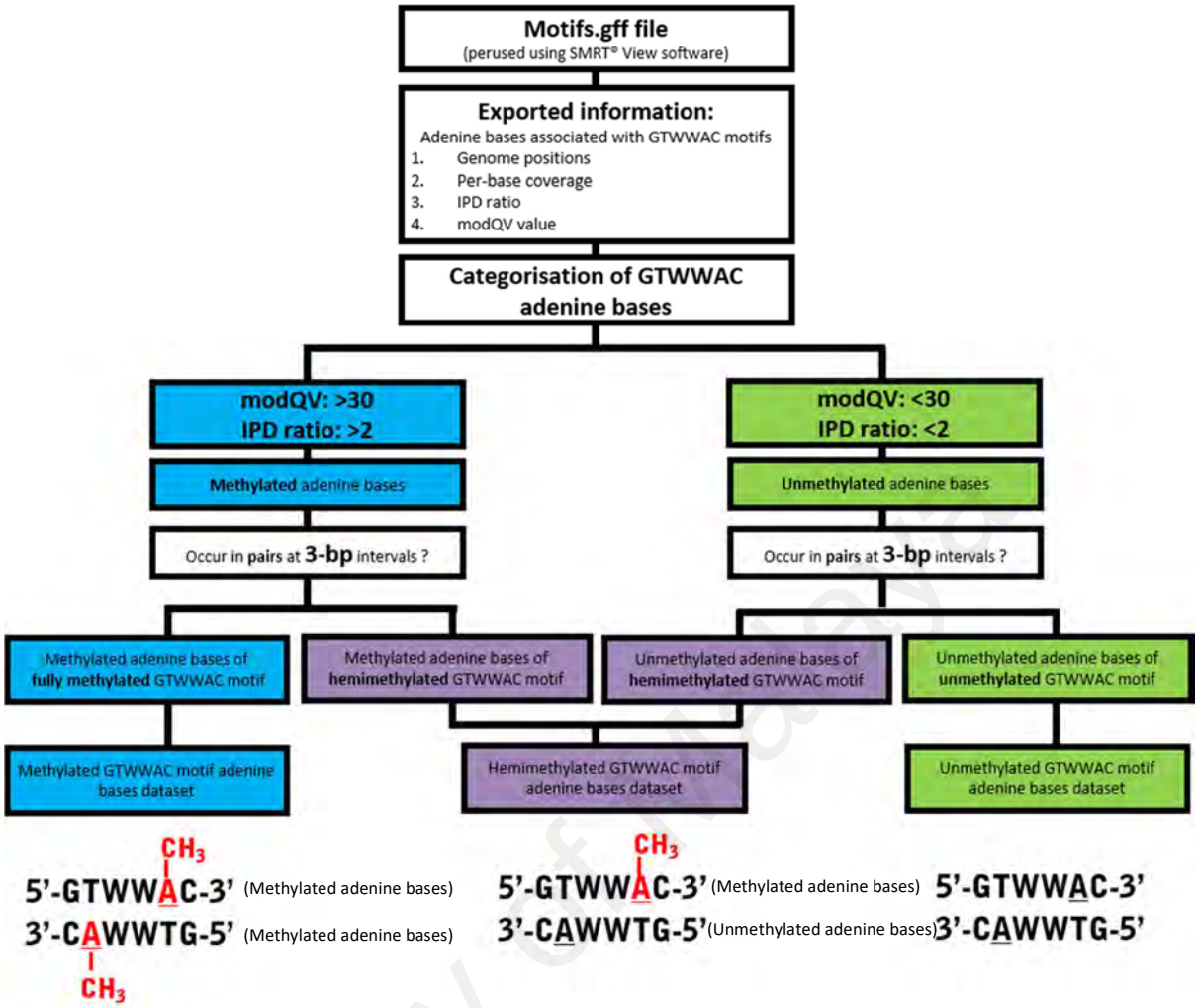


Figure 3.1: Flow chart of systematic methodology used to analyse the adenine bases associated with the GTWWAC motif. Examples which illustrate the methylation state (fully methylated, hemimethylated and unmethylated) of the motif are shown at the bottom of the chart.

3.13.2 Intragenic and Intergenic Distribution Analysis of Fully Methylated, Hemimethylated and Unmethylated GTWWAC Motifs

BEDTools *intersect* utility (bedtools version 2.26.0) was used to determine distribution of GTWWAC motif (fully methylated, hemimethylated and unmethylated) in intergenic regions (IGR) and genic regions (GR) of the chromosome and plasmid of all genomes (Quinlan, 2014). For each genome, the *intersect* utility compare genome positions of the adenine bases (associated with fully-methylated, hemi-methylated and unmethylated GTWWAC motifs respectively) with the annotated genome and overlaps of the adenine bases with the genomic regions were reported (Quinlan & Hall, 2010).

Furthermore, the hemimethylated and unmethylated GTWWAC motifs present within IGR regions were further analysed using BEDtools *closest* utility. The *closest* utility enable identification of left and right flanking genes of these intergenic motifs. Figure 3.2 shows a flow chart which summarised the workflow of these analyses.

3.13.2.1 *Intersect* Utility Worflow

Firstly, genome positions of GTWWAC motif adenine bases were formatted into a three-column tab delimited Bed file (file A) (columns: chromosome or plasmid name; start coordinate; stop coordinate) whereas genome annotations in general feature format (GFF3) file (downloaded from RAST annotation server) were formatted into a nine-column tab delimited Bed file (file B) (columns: chromosome or plasmid name; start; stop; FIG; type of genomic feature; note 1; strand; note 2; description of genome feature). Each genome position in file A was compared to file B using the BEDtools *intersect* utility.

Secondly, the adenine bases which overlapped the coding sequences (present within intragenic region) were detected and reported using the command printed below. The –wb option was used to report the original entry in file B which the genome positions in file A overlapped. This analysis generates a tab delimited output file (intragenic analysis output file) which contains information on the genome positions of the adenine

bases in file A as well as the genomic features (coding sequences) in file B which those adenine bases overlapped.

$ bedtools intersect -a File A -b File B -wb > intragenic analysis output file

Thirdly, the adenine bases which intersected the intergenic regions were detected and reported using the command as printed below. The –v option was used to report exclusively genome positions in file A which have no overlaps with the genomic features in file B. This analysis produces a three-column tab delimited output file (file C: intergenic analysis output file) which contains the genomic positions of all intergenic GTWWAC motif.

$ bedtools intersect -a File A -b File B -v > intergenic analysis output file

3.13.2.2 *Closest* Utility Workflow

Firstly, the output file generated from intergenic region intersection analysis (for both unmethylated and hemimethylated GTWWAC motifs (file C) and file B were used as input files.

Secondly, the command as printed below was used to determine the right flanking genes of the intergenic motifs. The option of –D (variable –b) reports genomic features in file B that are closest to intergenic motifs in file C along with information on relevant distance between the genomic features and the motif genome positions. Subsequently, the option of –iu on the other hand, restrict the output file to contain only information on the right flanking genes of the intergenic motifs.

$ bedtools closest -a File C -b File B -D b –iu > right flanking genes list (output file)

Thirdly, the command as printed below were used to determine the left flanking genes of the intergenic motifs. The option of –id ignores the features which are downstream of the genome positions in File A and reports only information on the left flanking genes of the intergenic motifs.

$ bedtools closest -a File C -b File B -D b –id > left flanking genes list (output file)

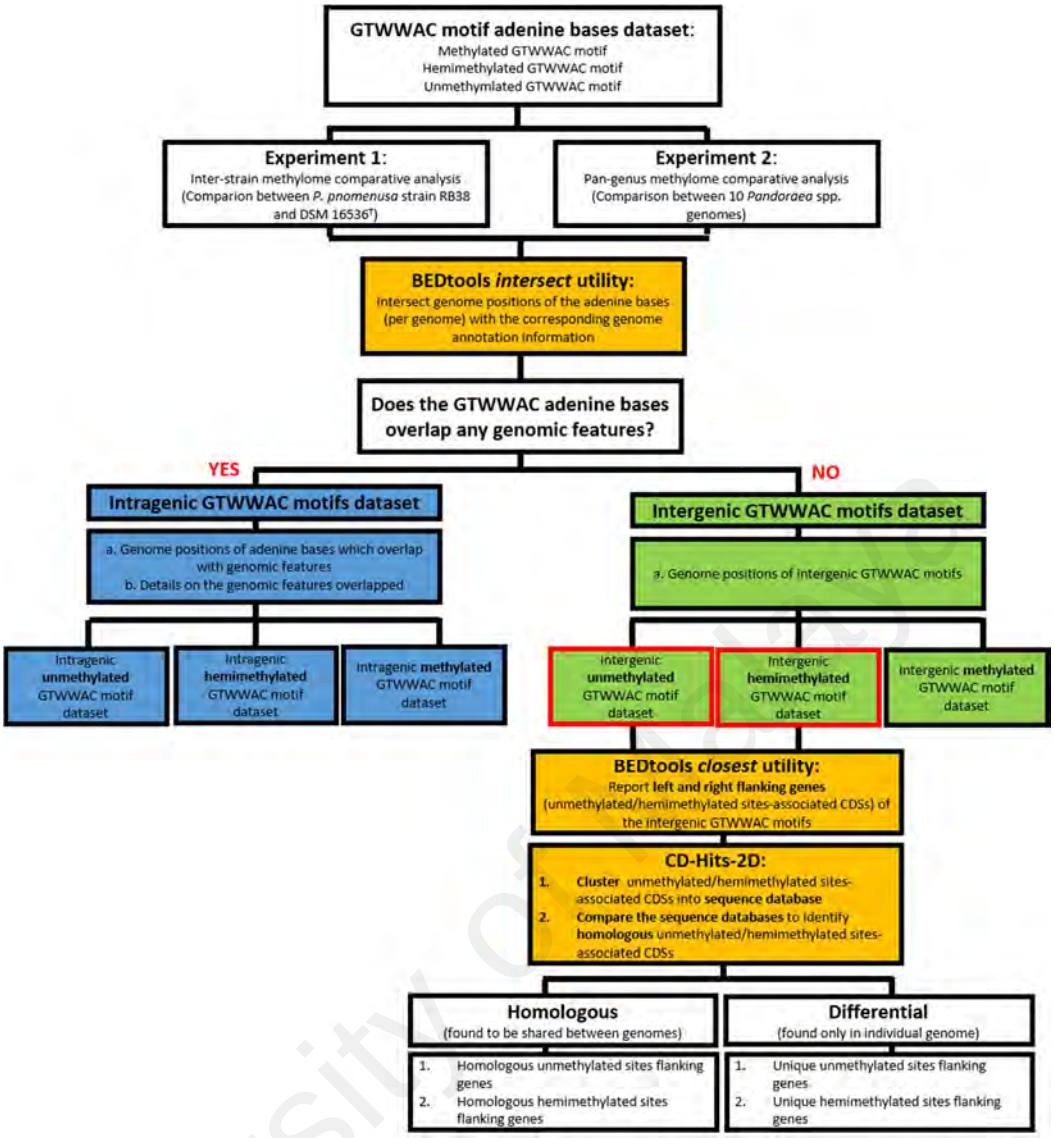


Figure 3.2: Flow chart of analysis workflow on GTWWAC motif intergenic and intragenic region distribution analysis.

3.13.3 GTWWAC Methylome Bins Analysis

The curated methylated adenine bases datasets of each genome were used to generate genome-wide GTWWAC methylation frequency distribution plots in which density of methylated GTWWAC motifs per 1 kb interval were calculated. Subsequently, the calculated motif density distribution datasets (per 1 kb bin width) of all genomes were pooled and the mean and standard deviation (*SD*) of the pooled dataset were estimated. These values were used to determine two threshold values for the purpose of methylome genome bins categorisation in each genome. Based on the calculation, the methylome

genome bins of each genome were categorised into hypermethylated bins (methylation frequency: mean + 6 *SD*) and methylation hotspot bin (methylation frequency: mean + 3 *SD*).

An AWK script was used to annotate coding sequences (CDSs) present within the hypermethylated bins and methylation hotspot bins of all genomes. The AWK script used for this analysis is provided in **Appendix B**.

3.13.3.1 Identification of Homologous Methylome Bins-Associated CDSs

Predicted peptide sequences of the methylome bins-associated CDSs were clustered into sequence databases and were subsequently compared to identify homologous sequences using the CD-HIT-2D algorithm (included in CD-HIT program version 4.6.5) (Huang et al., 2010). Two sets of experiments were done using this algorithm, which are firstly the inter-strain comparison of methylome bins-associated CDSs between *P. pnomenusa* RB38 and *P. pnomenusa* DSM 16536^T (Experiment 1) and secondly, the pan-genus comparison of methylome bins-associated CDSs of each genome with a pooled methylome bins-associated CDSs sequence database of all *Pandoraea* genomes (Experiment 2). The command and the explanation of each option used are as shown below:

$ cd-hit-2d –i db1 –i2 db2 –o output file name –c 0.7 –n 5 –d 200 –M 16000 –T 8 –s2 0.9

CD-HIT-2D (options)	**Value**	**Explanation**
-i	db1	Protein dataset 1: the query sequences dataset in fasta format
-i2	db2	Protein dataset 1: the reference sequences dataset in fasta format
-o	output file name	Output files: two files will be produced from this command, namely one text file which lists the similar sequences found in both input datasets and a FASTA-formatted file which contain all sequences in db2 which has no hit with sequences in db 1.
-c	0.7	Sequence identity threshold used in sequence comparison between the two datasets, sequences with more than 70 % sequence identity will be accepted included in the homologous gene list.
-n	5	Word length, value of 5 which matches the sequence similarity threshold (following recommendation on cd-hits program user guide) was used.
-d	200	Length of description on the homologous gene list for each sequence.
-M	16000	Memory limit
-T	8	Number of threads
-s2	0.9	Length difference cutoff

Similar approach of comparison was also used to identify conserved sequences among flanking genes of unmethylated and hemimethylated motifs (Figure 3.2). The summarised workflow of methylome bins analysis is presented in Figure 3.3.

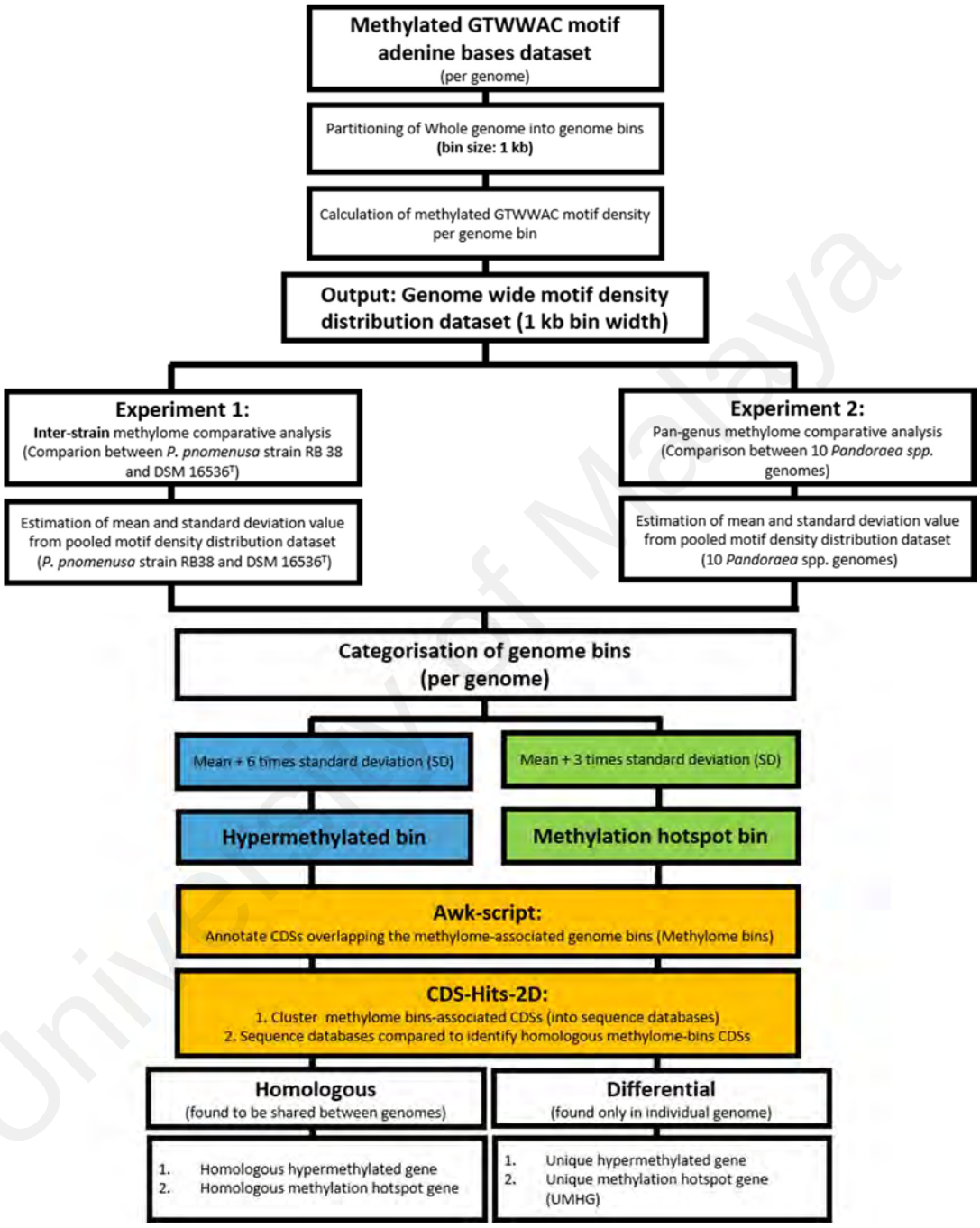


Figure 3.3: Flow chart of methylome bins analysis workflow

CHAPTER 4: RESULTS

4.1 Complete Genome Sequencing

Ten genomes of *Pandoraea* spp. which comprised all named species of the genus were sequenced and assembled to completion in this study.

4.1.1 Genome Sequencing of *Pandoraea* spp.

Genomes of 10 *Pandoraea* spp. genomes were sequenced using single molecule real time technology sequencing technology where all type strains were sequenced in the form of 20 kb library and the in-house strain namely strain RB-38, was sequenced in the form of 10 kb library. The outputs from the sequencing processes were assessed based on two parameters: the pre- and post- filtered polymerase read metrics and the subreads metrics. The polymerase reads metrics reflected the quality of the contiguous sequences (raw reads) captured from each zero-mode waveguide (ZMW) before and after the primary filtering whereas the subreads metrics represented condition of the sequences generated following removal of adapter sequences from filtered polymerase reads.

From the result (as summarised in Table 4.1), in the aspect of polymerase reads metrics, the pre-filter polymerase reads quality were low among all sequenced libraries, where the average read quality was 0.25 (n = 10, SD = 0.12) with the highest being 0.507 (sequenced 10 kb library of strain RB38). However, following the primary filtering step, after the filtering of an average of 16.3% of low quality polymerase bases, an averagely 4 folds improvement of the polymerase reads quality can be observed among all of the sequenced 20 kb libraries whereas only 1.6 folds of read quality improvement was observed following removal of 26.8 % of low quality polymerase reads of the 10 kb library. The differences between the pre- and post- filtered polymerase reads metrics are due to the raw-read trimming process in the primary analysis where the low quality region within the raw reads were removed. This trimming process resulted in the improvement

on the overall polymerase read quality. The discrepancies between the rate of improvement between the 10 kb and 20 kb libraries suggest that a small portion of sequences generated from the 20 kb libraries have very low per base read quality.

On the other hand, it was evident that the N50 read length of the 20 kb libraries were significantly longer (approximately 4 folds) than the 10 kb library, this length difference can be attributed to two factors, namely the difference in shearing parameters and the addition of a size-selection step in the 20 kb template libraries preparation process. The size-selection step in particular, contributed significantly to the sequenced read length distribution difference, as the elimination of short reads through the size-selection step minimised the loading bias of shorter fragments into ZMWs and hence increased the average length of the template library fragments loaded into the ZMWs. The shearing parameter used in the 20 kb library preparation process which utilised lower centrifugation speed also ensured that the DNA insert fragments are longer.

Table 4.1: Genome sequencing statistics of all *Pandoraea* spp.

	P. apista DSM 16535^T	*P. faecigallinarum* DSM 23572^T	*P. norimbergensis* DSM 11628^T	*P. oxalativorans* DSM 23570^T	*P. pnomenusa* DSM 16536^T	*P. pulmonicola* DSM 16583^T	*P. sputorum* DSM 21091^T	*P. thiooxydans* DSM 25325^T	*P. vervacti* DSM 23571^T	*P. pnomenusa* RB38 (in-house)
SMRT cell number	3	3	3	5	4	4	4	3	5	4
Raw reads filtering										
Pre-filter polymerase read bases	2029173877	1972139198	1285826909	891802355	1677947984	1093320789	1206964935	1060273892	2286074343	1743877188
Pre-filter polymerase reads	450876	450876	450876	751460	601168	601168	601168	450876	751460	601168
Pre-filter polymerase read N50	13828	19735	18154	11698	25374	7623	25628	14308	13007	6412
Pre-filter Polymerase read length	4500	4374	2851	1186	2791	1818	2007	2351	3042	2900
Pre-filter Polymerase read quality	0.37	0.284	0.198	0.143	0.184	0.276	0.121	0.179	0.276	0.507
Post-filter polymerase read bases	1637976426	1761249874	1138864861	805121977	1505797571	990755032	1088397722	790382121	1593526699	1276773164
Post-filter polymerase reads	153084	134691	93051	110948	114584	189065	75785	69571	164156	285138
Post-filter polymerase read N50	14909	20296	18877	12354	26200	7666	26457	15001	14950	6554
Post-filter Polymerase read length	10699	13076	12239	7256	13141	5240	14361	11360	9707	4477
Post-filter Polymerase read quality	0.847	0.858	0.858	0.849	0.857	0.83	0.859	0.838	0.836	0.844
Subread filtering										
Mean subread length	7391	7701	7789	3842	3225	4,683	4337	8932	4050	3521
Total number of bases (subreads)	1628098818	1755641873	1135409740	767987372	1458616765	989,595,418	1074659346	788951388	1558034929	1264734283
Subreads N50	9060	9524	9943	5238	4306	6,869	5812	11872	6029	4655
Number of reads (subreads)	220270	227962	145769	199845	452,255	211,277	247737	88327	384670	359123

4.1.2 *De Novo* Assembly

Genomes of all *Pandoraea* genomes were *de novo* assembled into closure using HGAP workflow. With the exception of *P. apista* and *P. sputorum*, the preasssembly step of all genomes were performed using seed read length cut-off value which generated seeding sequence data sets that represented approximately 30 folds target genome coverage. The seed read lengths of *P. apista* and *P. sputorum* were adjusted during assembly optimization to provide roughly 210 and 76 folds target genome coverage respectively.

From the genome assembly statistics which the output statistic from each assembly step (pre-assembly, *de novo* assembly, and circularization) were reported (Table 4.2), several observations were made. Firstly, in the preassembly step, only an average of 71 % (n = 10, SD = 12.44) of total bases were retained from the seeding sequences to form pre-assembled reads (represented by pre-assembled yield metric). This reduction in the number of bases observed are due to the mapping and trimming parameters used in the preassembly stage which resulted in removal of a portion of bases from end trimming and filtering of spurious reads.

Secondly, coverages and consensus concordances of the assembled contigs of all genomes increased following the circularization process. The consensus accuracy of assembled genomes are generally slightly lower and the size larger than the original genome size. This is due to the nature of Celera assembler, utilised as the assembler in the HGAP workflow, that assumes a genome to be linear and hence result in assembly of a finished contig which is linear and contains overlapping ends (S. Koren et al., 2012). The overlapping ends caused ambiguous mapping of the subreads which in turns result in lower reads mapping quality. However, as observed in the post-circularisation read metrics, significant increment in coverage per contig and consensus accuracy scores (to at least 99.9999%) for all contigs occured. This is due to unambiguous mapping of the subreads onto the blunt-ended reference sequence.

Table 4.2: *De novo* assembly statistics of all *Pandoraea* spp. genomes.

	P. apista DSM 16535^T	*P. faecigallinarum* DSM 23572^T	*P. norimbergensis* DSM 11628^T	*P. oxalativorans* DSM 23570^T	*P. pnomenusa* DSM 16536^T	*P. pulmonicola* DSM 16583^T	*P. sputorum* DSM 21091^T	*P. thiooxydans* DSM 25325^T	*P. vervacti* DSM 23571^T	*P. pnomenusa* RB38 (in-house)
Pre-assembly										
Polymerase Read Bases	1628098818	1755641873	1135409740	767987372	1458616765	679,770,185	1074659346	788951388	1558034929	1264734283
Length Cutoff	8000	17362	16536	8671	9306	9,934	7000	16892	13159	8810
Seed Bases	1047927506	150019304	150017142	150040821	165025102	150,032,748	381464221	150017142	150015118	150057066
Pre-Assembled bases	448036437	120860412	126871586	125964551	128550359	98,304,289	281268190	103500823	106652165	94591469
Pre-Assembled Yield	0.428	0.806	0.846	0.84	0.779	0.655	0.737	0.69	0.711	0.63
Pre-Assembled Reads	61159	8009	8360	15444	16072	16,627	39722	7493	10959	15882
Pre-Assembled Reads Length	7325	15090	15176	8156	7998	5,912	7080	13813	9731	5955
Pre-Assembled N50	9601	18335	18047	9233	9626	8,024	7995	17118	13305	8372
Polished assembly										
Polished Contigs	2	3	1	6	1	1	1	1	2	1
Contig Lengths (separated by semicolon)	Contig 1: 5524266; Contig 2: 78779	Contig 1: 5260522; Contig 2: 402221; Contig 3: 124370	6167937	Contig 1: 5639015; Contig 2: 639933; Contig 3: 135958; Contig 4: 85768; Contig 5: 46274;	5398127	5877157	5760347	4484974	Contig 1: 5655022: Contig 2: 105133	5390532
Coverage per contig	Contig 1: 253.07; Contig 2: 233.3	Contig 1: 274.72; Contig 2: 178.48; Contig 3: 80.6	165.09	Contig 1: 104.19; Contig 2:91.96; Contig 3: 70.12; Contig 4: 98.92; Contig 5: 96.65;	244.3	122.7	169.99	144.07	Contig 1: 185.03; Contig 2: 191.99	187.96
Mean coverage	252.79	263.86	165.09	103.88	244.3	122.7	169.99	144.07	185.16	187.96
Consensus concordance per contig		Contig 1: 99.99%; Contig 2: 99.98%; Contig 3: 99.97%	99.99%	Contig 1: 99.99%; Contig 2: 99.96%; Contig 3: 99.98%; Contig 4: 99.99%; Contig 5: 99.99%;	100%	100%	100%	100%	Contig 1: 99.98%; Contig 2: 98.72%	99.97%
Post-circularization										
Contig Lengths (separated by semicolon)	Contig 1: 5493973; Contig 2: 61145	Contig 1: 5241671; Contig 2: 386625; Contig 3: 104370	6167370	Contig 1: 5630311; Contig 2: 633357; Contig 3: 126976; Contig 4: 75108; Contig 5: 34979;	5389285	5867621	5742997	4464186	Contig 1: 5636000; Contig 2: 100282	5378916
Consensus concordance per contig	Contig 1: 100 %; Contig 2:100 %;	Contig 1: 100%; Contig 2: 100% Contig 3: 100%	100%	Contig 1: 100%; Contig 2: 100%; Contig 3: 100%; Contig 4: 100%; Contig 5: 100%	100%	100%	100%	100%	Contig 1: 100%; Contig 2: 99.99%	100.00%
Coverage per contig	Contig 1: 254.29; Contig 2: 290.23	Contig 1: 282.02 Contig 2: 190.51; Contig 3: 93.83	169.57	Contig 1: 107.58; Contig 2: 96.34; Contig 3: 76.45; Contig 4: 115.21; Contig 5: 125.71;	244.62	123.57	188.58	165.52	Contig 1: 215.95; Contig 2: 249.08	207.64

4.1.3 Genome Features

4.1.3.1 Pan-genus Genomic Feature Overview

The genomic features of all *Pandoraea* genomes were analysed to allow pan-genus genome features overview and comparison (Table 4.3). The genomic GC content of all species within the genus are consistent and within the range of 62.65 to 64.87 %. On the aspect of chromosome size, the size ranged from 4.46 Mb to 6.17 Mb, with the smallest being the chromosome of *P. thiooxydans* and the largest being the chromosome of *P. norimbergensis*. Furthermore, among all 10 sequenced *Pandoraea* strains, only 4 strains contained plasmids, all circular in structure. Both *P. vervacti* and *P. apista* contained one plasmid whereas *P. faecigallinarum* and *P. oxalativorans* contained 2 and 4 plasmids respectively. The sizes of the plasmids varied greatly, the genome of *P. oxalativorans* harboured both the smallest and the largest plasmid among all genomes analysed.

On the other hand, with the exception of *P. thiooxydans* which had 2 copies of rrn operon (comprised of genes encoding the 16S rRNA, 23S rRNA and 5S rRNA), all *Pandoraea* strains harboured 4 copies of rrn operon. As the number of rrn operon reflects the responding capacity of a microbe to surrounding resources, the low copy number of rrn operons in *P. thiooxydans*, besides possibly correlated with its small genome size, is also likely to suggest the adaptation of this organism towards a niche with low nutrient availability by having a lower protein synthesis capacity (Klappenbach et al., 2000; Lee et al., 2009). On the other hand, for the rest of *Pandoraea* strains, the number of rrn operons were close to average copy number of bacteria with rapid response to resource availability, indicating their capacity to respond rapidly to fluctuating growth condition which could render a fitness advantage (Klappenbach et al., 2000). The number of tRNAs, on the other hand, positively correlated with the number of rrn operons in the genome, where the number of tRNAs was lowest in *P. thiooxydans* at 49 whereas among the rest

of genomes the number of tRNAs were in the range of 64 to 73. Moreover, the total number of coding sequences (CDSs) were also positively correlated with the total size of each genome, where *P. oxalativorans*, with its total genome size of 6,515,938 bp, was predicted to harbour 6166 CDSs, whereas *P. thiooxydans* contained only 4319 CDSs.

Table 4.3: Genome features of all *Pandoraea* spp. genomes.

Strains	*P. apista* DSM 16535^T	*P. faecigallinarum* DSM 23572^T	*P. norimbergensis* DSM 11628^T	*P. oxalativorans* DSM 23570^T	*P. pnomenusa* DSM 16536^T	*P. pulmonicola* DSM 16583^T	*P. sputorum* DSM 21091^T	*P. thiooxydans* DSM 25325^T	*P. vervacti* DSM 23571^T	*P. pnomenusa* RB38 (in-house)
Chromosome size (bp)	5,493,973	5241671	6167370	5,630,311	5389285	5867621	5742997	4464186	5,636,000	5378916
Plasmid size (bp)	61,145	386,625; 104,370		633,357; 126,976; 75,108; 34,979;					100,282	
GC%	62.65	63.45	63.06	63.06	64.87	64.3	62.77	63.19	63.52	64.76
RNAs										
# rRNA	12	12	12	12	12	12	12	6	12	12
# tRNA	64	66	65	65	64	73	64	49	64	65
CDS										
Total number	5213	5456	5557	6166	5018	5389	5225	4319	5235	4982
# Hypothetical proteins	1440	1392	1516	1913	1088	1423	1375	897	1393	1039
# Functional assignment	3773	4064	4041	4253	3930	3966	3850	3422	3842	3943
# With EC assignment	1280	1249	1324	1272	1266	1267	1292	1173	1325	1283
# With GO assignments	1159	1132	1198	1153	1147	1148	1176	1048	1198	1165
# With Pathway assignments	1014	995	1056	1016	1012	1014	1035	929	1050	1026

Genome similarity assessments of all 10 genomes were performed using BLAST-based average nucleotide identity (ANI) analysis. ANI is one of the most robust genome-relatedness estimation metrics and is based on pair-wise comparison calculations of the shared sequences between genomes. This metric was found to demonstrate congruence with core genome phylogeny analysis and represent a powerful method in high resolution prokaryotic species definition (Chan et al., 2012; Zhang et al., 2014). From the orthologous ANI values between all genomes as shown in Figure 4.1, it can be observed that with the exception of *P. pnomenusa* DSM 16536^T and *P. pnomenusa* RB38, which represent two strains of the same species, the ANI between each genome of different *Pandoraea* species were observed to be lower than 93 %. This is in line with the 95 % species boundaries numerical cut-off proposed by Goris et al. (2007). Furthermore, among species, based on the ANI values, it can also be discerned that *P. oxalativorans* and *P. sputorum* demonstrated high degree of genomic relatedness whereas *P. thiooxydans* showed the lowest and *P.norimbergensis* the second lowest degree of relatedness with all other species within the genus. This level of distance in the aspect of genome relatedness is not surprising as both these genomes were reported to demonstrate a relatively distant phylogenetic positions with the other *Pandoraea* spp. (Coenye et al., 2000). The topology of the phylogenetic trees constructed using 16S rDNA sequences and OrthoANI values showed close agreement.

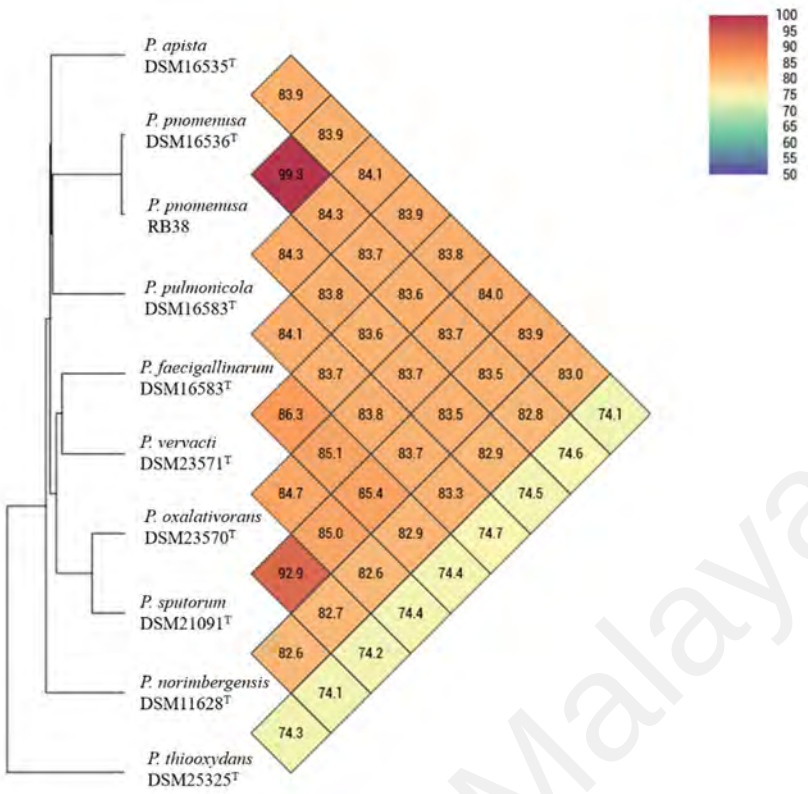


Figure 4.1: OrthoANI results calculated with the genomes of *Pandoraea* spp. The OrthoANI value between two genomes are stated at their diagonal point on the chart.

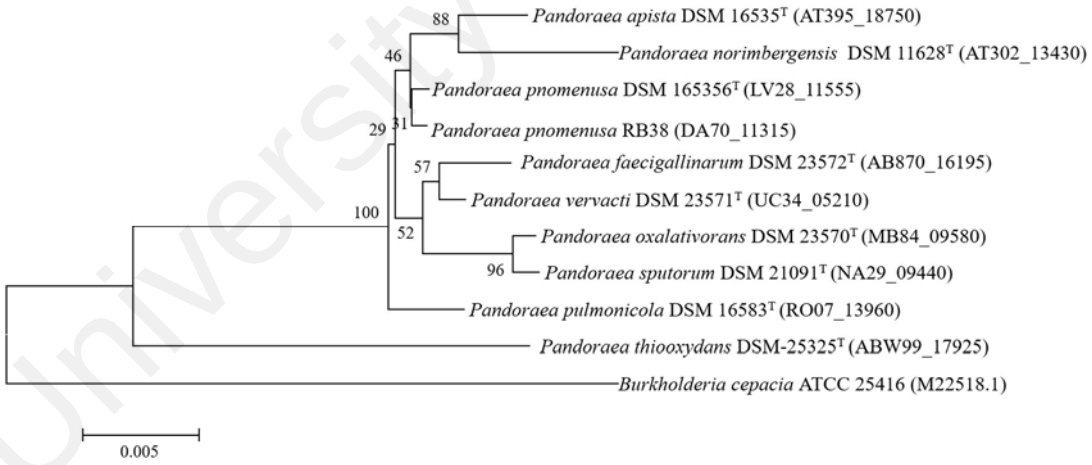


Figure 4.2: 16S rDNA phylogenetic analysis of 10 *Pandoraea* genomes. The tree was inferred by the neighbour joining method (Saitou & Nei, 1987). The optimal tree with the sum of branch length = 0.08908537 is shown. The percentage of replicate trees in which the associated taxon clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). Bar, 0.005 substitution per nucleotide position.

Furthermore, whole genome alignments of 10 rearranged *Pandoraea* spp. chromosomes were performed using Mauve as shown in Figure 4.3. The alignment

revealed that among all species, a large number of locally collinear blocks (LCBs) were conserved. LCBs represent homologous genome segments which are internally free from genome rearrangements (Darling et al., 2004). Overall, between the genomes of *P. apista*, *P. pnomenusa*, *P. pulmonicola*, *P. faecigallinarum*, *P. vervacti*, and *P. oxalativorans*, the genomes showed largely similar gene arrangement patterns, showing agreement to the degree of genome-relatedness as shown by ANI analysis (Figure 4.2). However, between *P. sputorum* and *P. oxalativorans*, which ANI similarity value suggested highly similar gene content, an occurrence of inversion in the middle segment of the genome of *P. sputorum* can be observed. On the other hand, the genome of *P. norimbergensis* showed a relatively similar genome arrangement with *P. sputorum*, however, presence of a large amount of breakpoint regions can be observed. These regions indicated that the *P. norimbergensis* genome contained a large sum of species-specific gene contens. Furthermore, also in concurrence with the ANI value, the genome of *P. thiooxydans* showed a substantial gene content difference relative to the rest of the *Pandoraea* genomes, as evidenced by the small amount of LCBs present within the genome.

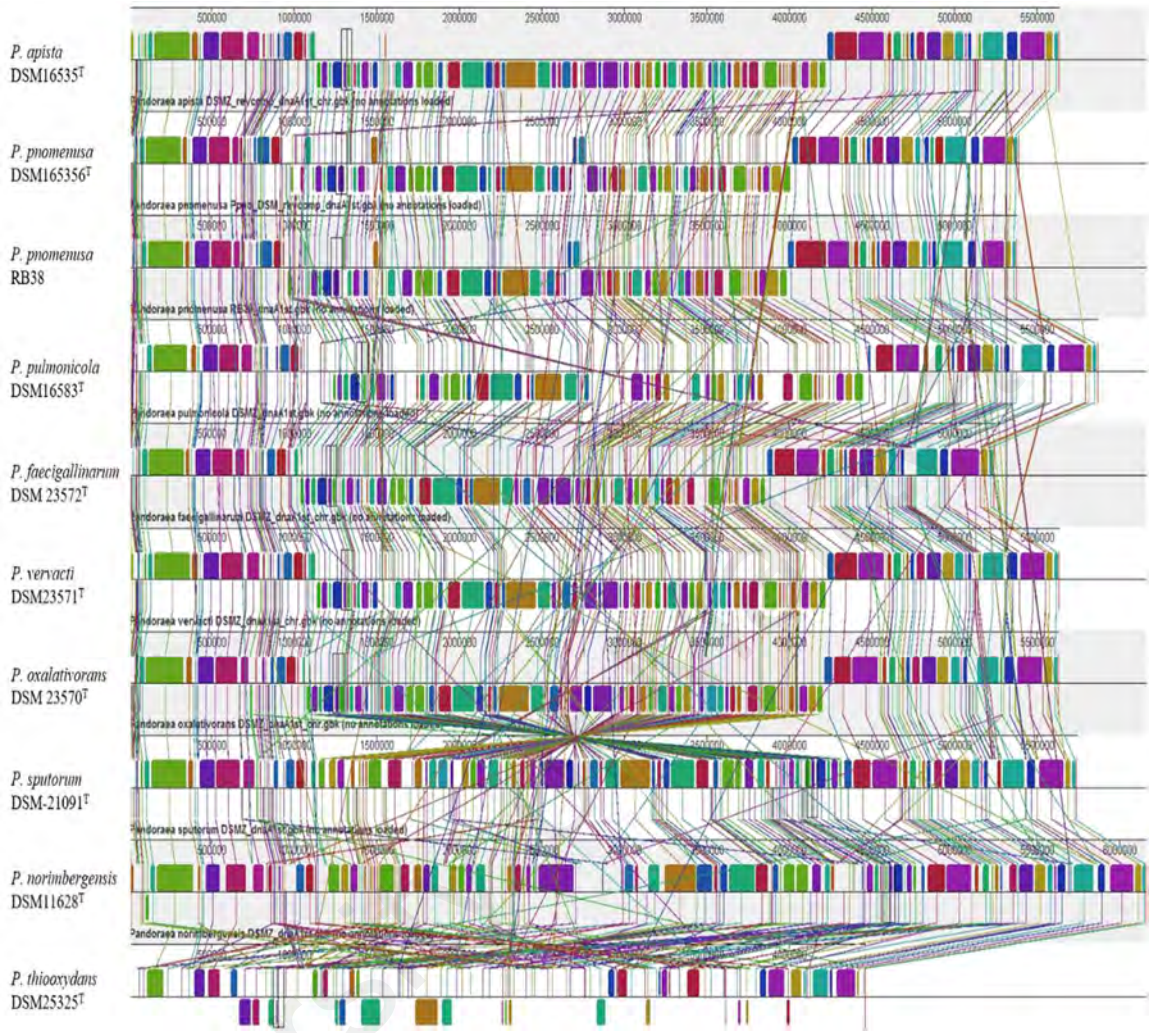


Figure 4.3: Multiple genome alignment of 10 *Pandoraea* genomes (rearranged to align at the DnaA gene). The alignment was performed using Mauve software (Darling et al., 2004). The boxes with identical colours between genomes represent local colinear blocks (LCBs), indicating the homologous genome segments identified.

4.1.3.2 Functional Classification of Annotated Genes

RAST subsystem classification was used to assess the overall functional classification statistics of annotated genes (Figure 4.4). Among all subsystem categories, the "amino acids and derivatives" category was shown to contain the highest number of genes whereas the "dormancy and sporulation" category contained the least. The "photosynthesis' category was the only subsystem category which was not found to be annotated in the genomes of all *Pandoraea* species.

Overall, majority of the subsystem categories were shown to contain approximately similar number of genes in all species. The slight variation in gene numbers per species positively corresponded to genome sizes where on average, *P. norimbergensis* harboured the highest number of genes whereas *P. thiooxydans* harboured the least in most subsystem categories. However, 5 categories were found to be an exception to these distribution patterns, namely the categories of "prophages, transposable elements, plasmids", "Iron acquisition and metabolism", "Motility and chemotaxis", "secondary metabolism" and "sulfur metabolism".

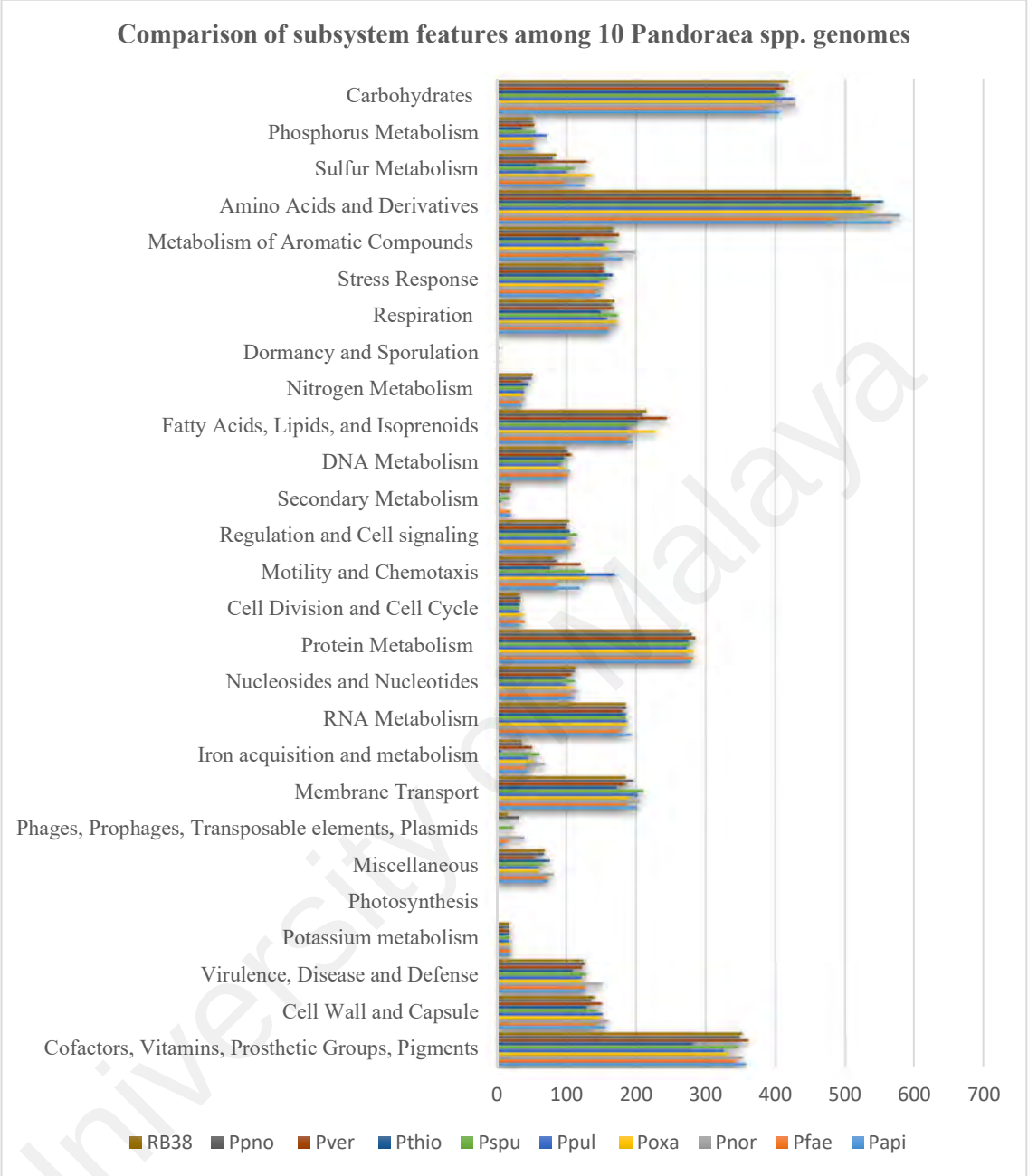


Figure 4.4: Functional comparisons among the *Pandoraea* spp. genomes. The comparison was performed based on the RAST subsystems classification. The y-axis represent the 27 RAST subsystem categories whereas the x-axis denotes the number of genes present in each subsystem categories. Denotations of abbreviation: Papi: *Pandoraea apista* DSM-16535^T; Pfae: *Pandoraea faecigallinarum* DSM 23572^T; Pnor: *Pandoraea norimbergensis* DSM 11628^T; Poxa: *Pandoraea oxalativorans* DSM 23570^T; Ppul: *Pandoraea pulmonicola* DSM-16583^T; Pspu: *Pandoraea sputorum* DSM-21091^T; Pthio: *Pandoraea thiooxydans* DSM-25325^T; Pver: *Pandoraea vervacti* DSM-23571^T; Ppno: *Pandoraea pnomenusa* DSM-165356^T; RB38: *Pandoraea pnomenusa* RB38.

4.1.3.3 *OriC* Region Prediction Analysis in The *Pandoraea* spp. Chromosomes

In all rearranged chromosomes (with *dnaA* gene as the first CDS), a distinct DNA asymmetry in the aspect of G-C parameter was observed with the lowest skew located on average 15 kb upstream of the *dnaA* gene and the highest skew approximately in the middle of the chromosome (Table 4.4). Each genome was also identified to contain several clusters of DnaA boxes and majority of these clusters were found to be located upstream of the *dnaA* gene (Table 4.5). These results provided an estimated range of the putative *oriC* region location. Location of DnaA boxes clusters were further analysed to pinpoint the clusters which represent the putative *oriC* region.

Based on the search of DnaA boxes which sequences differed by one or two mismatches from *E. coli* perfect DnaA box motif (TTATCCACA), only 7 genomes were found to have a putative *oriC* region which demonstrated perfect agreement of the DnaA boxes cluster location with the position of the *dnaA* gene. These genomes were namely *P. apista*, *P. faecigallinarum*, *P. norimbergensis, P. pulmonicola*, *P. sputorum*, *P. pnomenusa*, and *P. pnomenusa* RB38 (Table 4.5). The *oriC* region of *P. oxalativorans* was identified using similarity searches of the genome based on the identified *oriC* region of the other genomes. The predicted *oriC* regions in these genomes were on average 797 bp in size, with a mean AT-content of 39.7 %, and were located in an intergenic region within the *rnpA-rmpH–oric-dnaA-dnaN-gyrB* genes cluster (within the intergenic region between the annotated *dnaA* and *rnpA* gene). On the other hand, the *oriC* regions of *P. thiooxydans* and *P. vervacti* were unable to be irrevocably identified as the results of DnaA boxes clusters locations and the gene clusters associated with replication initiation process showed imperfect agreement. Therefore, for the purpose of subsequent analyses, the location of the intergenic region located between the *dnaA* and *rnpA* gene were assumed to be the putative *oriC* region for these two genomes.

In all genomes, in addition to the putative *oriC* region, additional DnaA boxes clusters were also identified. Overrepresentation of DnaA boxes clusters is commonly observed in majority of bacterial chromosomes. These additional DnaA boxes are proposed to be involved in replication initiation or serve a regulatory role in chromosome replication by partaking in titration of free DnaA protein (Mackiewicz et al., 2004). Furthermore, in agreement to the finding of Mackiewicz et al. (2004), one DnaA box clusters could be found approximately 19.6 kb upstream of the putative *oriC* regions of all *Pandoraea* genomes.

Table 4.4: Summary of the maximum and minimum point of RY, MK, AT and GC disparity curves of each *Pandoraea* genomes. The genome coordinates (nt) of the *dnaA* genes along with the associated GenBank locus tag of each genome is included.

Genomes	*P.apista* DSM 16535^T	*P. faecigallinarum* DSM 23572^T	*P. norimbergensis* DSM 11628^T	*P. oxalativorans* DSM 23570^T	*P. pnomenusa* DSM 16536^T	*P. pulmonicola* DSM 16583^T	*P. sputorum* DSM 21091^T	*P. thiooxydans* DSM 25325^T	*P. vervacti* DSM 23571^T	*P. pnomenusa* RB38 (in-house)
Coordinates of *dnaA* gene (nt) (accession)	1 .. 1560 (AT395_RS24795)	1 .. 1542 (AB870_RS00005)	1 .. 1569 (AT302_RS00005)	1 .. 1545 (MB84_04650)	1 .. 1533 nt (LV28_RS48200)	1 .. 1551 (RO07_RS00005)	1 .. 1545 (NA29_RS00005)	1 .. 1473 (ABW99_RS20720)	1 ..1566 (UC34_RS00005)	1 .. 1533 (DA70_RS00005)
The extremes of GC disparity	5482949 nt (minimum), 2639235 nt (maximum)	5229376 nt (minimum), 2467006 nt (maximum)	6152217 nt (minimum), 2934586 nt (maximum)	5614383 nt (minimum), 2650142 nt (maximum)	5376762 nt (minimum), 2652795 nt (maximum)	5848823 nt (minimum), 2825329 nt (maximum)	5728110 nt (minimum), 2751911 nt (maximum)	4442176 nt (minimum), 2155783 nt (maximum)	5618438 nt (minimum), 2754488 nt (maximum)	5366381 nt (minimum), 2620881 nt (maximum)
The extremes of AT disparity	2455067 nt (minimum), 29906 nt (maximum)	2701407 nt (minimum), 22724 nt (maximum)	3507849 nt (minimum), 6130 nt (maximum)	2966496 nt (minimum), 15295 nt (maximum)	2740973 nt (minimum), 10237 nt (maximum)	3020314 nt (minimum), 33269 nt (maximum)	2579005 nt (minimum), 31578 nt (maximum)	2318444 nt (minimum), 4347801 nt (maximum)	2984076 nt (minimum), 53947 nt (maximum)	2712326 nt (minimum), 5700 nt (maximum)
The extremes of RY disparity	5473449 nt (minimum), 2625769 nt (maximum)	5220207 nt (minimum), 2466928 nt (maximum)	6145384 nt (minimum), 2935070 nt (maximum)	5607874 nt (minimum), 2642010 nt (maximum)	5369032 nt (minimum), 2645196 nt (maximum)	5847201 nt (minimum), 2864389 nt (maximum)	5721545 nt (minimum), 2754771 nt (maximum)	4449446 nt (minimum), 2099584 nt (maximum)	5611823 nt (minimum), 2729769 nt (maximum)	5358677 nt (minimum), 2616915 nt (maximum)
The extremes of MK disparity	2635723 nt (minimum), 5484824 nt (maximum)	2562195 nt (minimum), 5231296 nt (maximum)	2933159 nt (minimum), 6156806 nt (maximum)	2674766 nt (minimum), 5619155 nt (maximum)	2695623 nt (minimum), 1030 nt (maximum)	2820556 nt (minimum), 5857362 nt (maximum)	2749082 nt (minimum), 5732923 nt (maximum)	2252026 nt (minimum), 4461991 nt (maximum)	2851750 nt (minimum), 5623029 nt (maximum)	2658708 nt (minimum), 5369778 nt (maximum)

Table 4.5: Distribution of putative DnaA boxes (differs by one or two mismatches from *E. coli* perfect DnaA box motif TTATCCACA) in genomes of *Pandoraea* spp. The location of DnaA box clusters which represent the putative *oriC* region (demonstrated perfect agreement of the DnaA boxes cluster location with the position of the *dnaA* gene) are highlighted in grey.

Genomes	***P.apista*** **DSM 16535^T**	***P. faecigallinarum*** **DSM 23572^T**	***P. norimbergensis*** **DSM 11628^T**	***P. oxalativorans*** **DSM 23570^T**	***P. pnomenusa*** **DSM 16536^T**	***P. pulmonicola*** **DSM 16583^T**	***P. sputorum*** **DSM 21091^T**	***P. thiooxydans*** **DSM 25325^T**	***P. vervacti*** **DSM 23571^T**	***P. pnomenusa*** **RB38 (in-house)**
DnaA boxes clusters (C)	C1: 5418143..5419295 nt; 5	C1: 52675..53177 nt; 4	C1: 1570..1868 nt; 3	C1: 6891..7761 nt; 3	C1: 5316016..5317040 nt; 5	C1: 92444..93576 nt; 4	C1: 66505..67003 nt; 6	C1: 5305919..5306688 nt; 5	C1: 4450805 to 4451317; 3	C1: 1567..1815 nt; 3
(Location; # of DnaA boxes)	C2: 5472878..5474085 nt; 4	C2: 5159435..5160060 nt; 5	C2: 12546..12880 nt; 6	C2: 5538425..5538806 nt; 5	C2: 5368456..5369869 nt; 4	C2: 5791860..5794270 nt; 4	C2: 5666796..5669658 nt; 4	C2: 5358101..5359514 nt; 4	C2: 4414715 to 4415118; 3	C2: 89278..89740 nt; 3
	C3: 5493178..5493973 nt; 4	C3: 5219990..5221059 nt; 4	C3: 39989..40433 nt; 4	C3: 5604543..5604683 nt; 3	C3: 5388488..5389285 nt; 4	C3: 5845547..5846960 nt; 4	C3: 5721331..5722379 nt; 4	C3: 5378119..5378916 nt; 4		C3: 90050..90640 nt; 4
		C4: 5240876 to 5241671 nt; 4	C4: 6091341..6091878 nt; 5	C4: 5608195..5608697 nt; 3		C4: 5866825..5867621 nt; 4	C4: 5742199..5742997 nt; 4			C4: 96598..96818 nt; 3
			C5: 6142012..6142125 nt; 3	C5: 5629512 to 5630311; 3						C5: 5551556..5553895 nt; 5
			C6: 6145644..6146228 nt; 4							C6: 5611851..5612676 nt; 4
			C7: 6166578 to 6167370; 4							

4.1.3.4 Prophage Diversity in The Genomes of *Pandoraea* spp.

By using the available complete genome data of the *Pandoraea* genus, analysis of prophage-like elements present in the genome were analysed with the aim to (i) address the information gap on the prophages distribution in the genus of *Pandoraea* and to (ii) understand the role of prophages in lateral gene transfer of DNA methyltransferases (MTases) in the *Pandoraea* spp. genomes.

All *Pandoraea* spp. genomes were analysed for presence of prophages using both PHAST and PHASTER pipeline (Arndt et al., 2016; Zhou et al., 2011). In all genomes, a total of 82 prophages were identified (Figure 4.5 and Table 4.6). The genome of *P. oxalativorans* harboured the largest number of prophages and *P. sputorum* the least. Furthermore, no prophages were identified in the pPA35 and pPO70-4 plasmids which were the two smallest plasmids among all genomes (smaller than 62 kb) suggesting that genome size could be a factor which contribute to the abundance of prophages distribution. In all genomes, with the exception of *P. pulmonicola* and *P. thiooxydans*, both potentially intact and functional phage as well as defective prophages (annotated as incomplete or questionable prophages) were identified (Figure 4.5). As observed in Figure 4.5, the number of defective prophages identified are either higher or at equal amount as intact prophages. Defective prophages are prophages which comprised incomplete phage structural genes and are most likely dormant as a result of mutational decay or genetic degradation over multiple host generations. The relatively high abundance of these prophage types are common in bacterial genomes (Casjens, 2003).

In the phyla of *Betaproteobacteria*, the patterns of phage infection are largely unknown, therefore, the identity of prophages detected in the *Pandoraea* genomes were analysed. In general, no significant pattern can be observed with the prophage type distribution pattern. Only 7 prophage types (closest hits to the prophages detected) were identified more than once in these genomes: Stx2-converting phage 1717 (NC_011357),

Staphlococcus phage SPbeta-like (NC029119), *Ralstonia* phage phiRSA1 (NC_009382), Shigella phage SfIV (NC_022749), Enterobacteria phage BP-4795 (NC_004813), Bacillus phage G (NC_023719) and *Paenibacillus* phage Xenia (NC_028837). However, these results are not completely representative of the prophage types of the *Pandoraea* genomes as the hits were confined to the available records within the prophage database integrated in the PHAST and PHASTER prophage analysis pipeline and therefore novel prophages were not annotated.

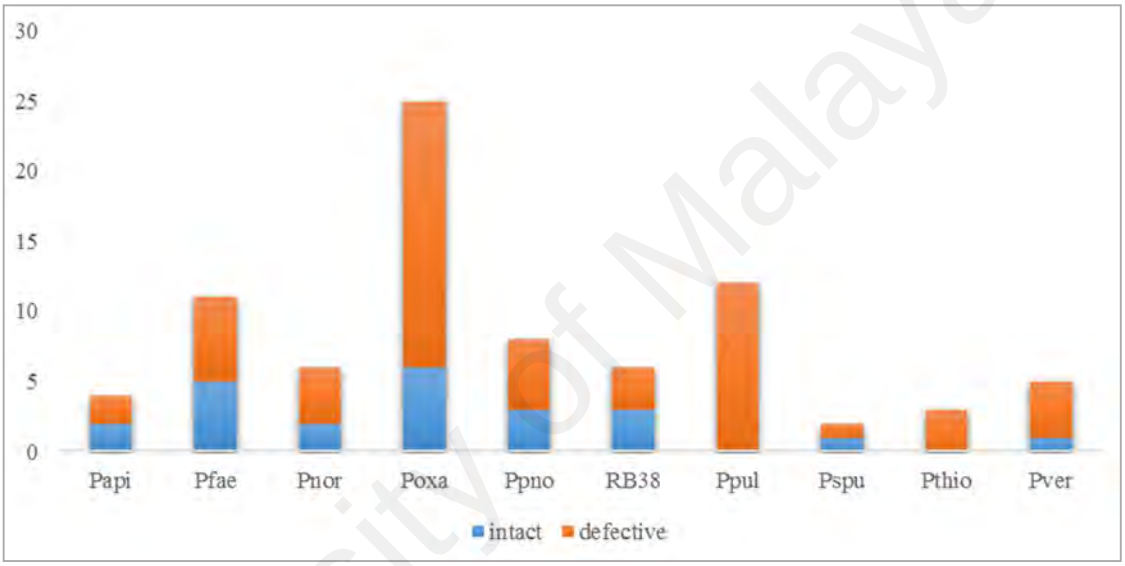


Figure 4.5: Bar chart which depicts the ratio of intact and defective prophages annotated in each *Pandoraea* spp. genome. Denotations of abbreviation: Papi: *Pandoraea apista* DSM-16535^T; Pfae: *Pandoraea faecigallinarum* DSM 23572^T; Pnor: *Pandoraea norimbergensis* DSM 11628^T; Poxa: *Pandoraea oxalativorans* DSM 23570^T; Ppul: *Pandoraea pulmonicola* DSM-16583^T; Pspu: *Pandoraea sputorum* DSM-21091^T; Pthio: *Pandoraea thiooxydans* DSM-25325^T; Pver: *Pandoraea vervacti* DSM-23571^T; Ppno: *Pandoraea pnomenusa* DSM-165356^T; RB38: *Pandoraea pnomenusa* RB38.

Table 4.6: Prophages annotated in the 10 *Pandoraea* species genomes using PHAST and PHASTER.

Genome	Region	Region Length	Completeness	Score	Region Position	Possible Phage	GC %
***P. apista* DSM 16535 ^{T}**							
Chromosome	1	22 kb	incomplete	30	2235842-2257916	PHAGE Shigel SfII NC 021857	58.59%
	2	27.8 kb	intact	150	3450136-3477952	PHAGE Shigel SfIV NC 022749	59.03%
	3	23.8 kb	intact	150	3454125-3477952	PHAGE Entero SfV NC 003444	58.88%
	4	14.1 kb	incomplete	60	3494834-3509012	PHAGE Sinorh PBC5 NC 003324	61.87%
***P. faecigallinarum* DSM 23572 ^{T}**							
Chromosome	1	10.7 kb	questionable	90	1010973-1021718	PHAGE Pseudo PaBG NC 022096	57.07%
	2	25 kb	intact	150	3253158-3278248	PHAGE Shigel SfIV NC 022749	58.86%
	3	46.8 kb	intact	110	3428344-3475228	PHAGE Ralsto RSB1 NC 011201	60.34%
Plasmid pPF72-1	1	26.9 kb	questionable	80	51001-77953	PHAGE Staphy SPbeta like NC 029119	60.49%
	2	24.4 kb	questionable	80	54836-79271	PHAGE Burkho phiE12 2 NC 009236	
	3	13.4 kb	questionable	70	156075-169493	PHAGE Shigel SfIV NC 022749	60.85%
	4	40.1 kb	intact	150	165479-205654	PHAGE Staphy SPbeta like NC 029119	59.77%
	5	19.8 kb	intact	150	178547-198429	PHAGE Stx2 c 1717 NC 011357	59.12%
	6	6 kb	questionable	70	347141-353237	PHAGE Singap grouper iridovirus NC 006549	59.83%
Plasmid pPF72-2	1	56.7 kb	intact	150	31500-88217	PHAGE Staphy SPbeta like NC 029119	59.07%
	2	10.8 kb	questionable	70	40411-51215	PHAGE Staphy phiPV83 NC 002486	57.53%
***P. norimbergensis* DSM 11628 ^{T}**							
Chromosome	1	41.1 kb	intact	93	1234661-1275787	PHAGE Burkho BcepMu NC 005882	60.31%
	2	48.1 kb	intact	93	3084614-3132806	PHAGE Burkho phiE255 NC 009237	60.03%
	3	17.9 kb	incomplete	30	3449692-3467615	PHAGE Vibrio VP58.5 NC 027981	64.65%
	4	14.4 kb	incomplete	20	3496411-3510835	PHAGE Burkho DC1 NC 018452	65.71%
	5	37.7 kb	incomplete	40	3496411-3534129	PHAGE Burkho BcepIL02 NC 012743	65.55%
	6	23.9 kb	incomplete	20	3513199-3537142	PHAGE Rhizob RR1 A NC 021560	65.21%

Table 4.6, continued

Genome	Region	Region Length	Completeness	Score	Region Position	Possible Phage	GC %
***P. oxalativorans* DSM 23570^T**							
Chromosome	1	8.2 kb	incomplete	30	539052-547293	PHAGE Thermu OH2 NC 021784	62.97%
	2	8 kb	questionable	70	1036844-1044923	PHAGE Stx2 c 1717 NC 011357	57.18%
	3	6.3 kb	incomplete	40	1045233-1051611	PHAGE Paenib Xenia NC 028837	58.24%
	4	12.9 kb	questionable	70	1079205-1092191	PHAGE Stx2 converting 1717 NC 011357	61.79%
	5	6.1 kb	incomplete	10	1839140-1845245	PHAGE Bacill G NC 023719	62.05%
	6	7.2 kb	questionable	90	3406352-3413602	PHAGE Stx2 c 1717 NC 011357	60.12%
	7	27.8 kb	intact	150	3643936-3671791	PHAGE Entero BP 4795 NC 004813	58.40%
	8	9.8 kb	questionable	90	4226501-4236398	PHAGE Burkho phiE125 NC 003309	59.41%
	9	7.9 kb	incomplete	60	4969764-4977719	PHAGE Paenib Xenia NC 028837	59.36%
	10	7.4 kb	questionable	80	5281398-5288858	PHAGE Stx2 c 1717 NC 011357	60.60%
	11	9.7 kb	questionable	70	5567946-5577703	PHAGE Entero fiAA91 ss NC 022750	58.61%
Plasmid pPO70-1	1	11.7 kb	questionable	90	294381-306178	PHAGE Stx2 c 1717 NC 011357	60.53%
	2	20.1 kb	intact	150	294381-314512	PHAGE Stx2 converting 1717 NC 011357	60.15%
	3	7.5 kb	questionable	80	316422-324003	PHAGE Mannhe vB MhS 1152AP2 NC 028956	60.38%
	4	6.5 kb	questionable	70	346563-353068	PHAGE Shigel SfIV NC 022749	60.82%
	5	7.8 kb	questionable	80	356128-363934	PHAGE Ralsto RSA1 NC 009382	58.65%
	6	14.3 kb	questionable	90	365383-379690	PHAGE Stx2 c 1717 NC 011357	60.22%
	7	12.7 kb	intact	110	400150-412923	PHAGE Stx2 converting 1717 NC 011357	61.96%
Plasmid pPO70-2	1	19.7 kb	incomplete	60	99304-119070	PHAGE Caulob karma NC 019410	59.68%
	2	19.7 kb	questionable	80	26752-46500	PHAGE Staphy SPbeta like NC029119	62.70%
	3	41.6 kb	intact	120	33218-74910	PHAGE Staphy SPbeta like NC029119	60.45%
	4	5.6 kb	incomplete	60	114322-119985	PHAGE Paenib Xenia NC 028837	58.85%
Plasmid pPO70-3	1	29.3 kb	questionable	80	537-29897	PHAGE Mycoba 32HC NC 023602	60.61%
	2	25.9 kb	intact	100	13137-39064	PHAGE Ralsto RSA1 NC 009382	60.05%
	3	21.9 kb	intact	120	52656-74617	PHAGE Staphy SPbeta like NC029119	57.30%
***P. pnomenusa* RB38**							
Chromosome	1	25.7 kb	intact	150	1023855-1049597	PHAGE Entero BP 4795 NC 004813	57.98%
	2	8.1 kb	incomplete	60	1078077-1086233	PHAGE Papiin herpesvirus 2 NC 007653	63.18%
	3	58.2 kb	intact	150	1429243-1487480	PHAGE Salmon 64795 sal3 NC 031918	59.98%
	4	28.2 kb	intact	150	1444389-1472632	PHAGE Burkho Bcep176 NC 007497	59.39%
	5	12.5 kb	incomplete	50	4074268-4086853	PHAGE Entero BP 4795 NC 004813	60.38%
	6	5.9 kb	incomplete	50	4077533-4083513	PHAGE Bacill G NC 023719(2)	60.94%

Table 4.6, continued.

Genome	Region	Region Length	Completeness	Score	Region Position	Possible Phage	GC %
***P. pnomenusa* DSM 16536^T**							
Chromosome	1	34.6 kb	intact	120	445468-480076	PHAGE Ralsto RSA1 NC 009382	62.43%
	2	10.3 kb	questionable	70	3168200-3178529	PHAGE Ralsto RSL1 NC 010811	62.00%
	3	17.8 kb	questionable	90	3168200-3186052	PHAGE Ralsto RSA1 NC 009382	61.88%
	4	15.4 kb	incomplete	10	3186428-3201865	PHAGE Escher HK639 NC 016158	62.13%
	5	13.1 kb	incomplete	20	3188693-3201865	PHAGE Haemop HP2 NC 003315	61.68%
	6	20.7 kb	incomplete	30	3195301-3216087	PHAGE Pseudo phi297 NC 016762	62.13%
	7	28.2 kb	intact	150	4351988-4380276	PHAGE Entero BP4795 NC 004813	61.35%
	8	35 kb	intact	150	4418923-4454003	PHAGE Shigel SfIV NC022749	61.75%
***P. pulmonicola* DSM 16583^T**							
	1	10.6 kb	incomplete	10	189537-200198	PHAGE Entero CAjan NC028776	64.96%
	2	7.1 kb	incomplete	20	1344156-1351277	PHAGE Helico 2 NC 004156	65.57%
	3	8.7 kb	incomplete	10	1434540-1443277	PHAGE Bacill MG B1 NC 021336	64.39%
	4	8.5 kb	incomplete	10	1482513-1491025	PHAGE Synech S SSM7NC 015287	65.15%
	5	5.8 kb	incomplete	10	2065756-2071599	PHAGE Bacill G NC023719	65.06%
	6	9.7 kb	incomplete	10	2134733-2144439	PHAGE Megavi chiliensis NC016072	65.21%
	7	12.6 kb	incomplete	10	2325722-2338349	PHAGE Acanth moumouvirus NC020104	64.51%
	8	9.2 kb	incomplete	10	3274822-3284097	PHAGE Plankt PaV LD NC016564	64.03%
	9	6.8 kb	incomplete	10	3478650-3485484	PHAGE Burkho BcepC6B NC005887	55.23%
	10	8.5 kb	incomplete	10	4540415-4548922	PHAGE Acinet Acj9 NC014663	63.66%
	11	7.7 kb	incomplete	10	4566945-4574739	PHAGE Sphing PAU NC019521	64.43%
	12	14.5 kb	incomplete	20	4738833-4753396	PHAGE Bacill G NC023719	65.83%
***P. sputorum* DSM 16583^T**							
	1	15.3 kb	incomplete	40	2151971-2167331	PHAGE Salmon SEN34 NC 028699	60.08%
	2	38.8 kb	intact	150	5251438-5290302	PHAGE Ralsto phiRSA1 NC 009382	60.13%
***P. thiooxydans* DSM 25325^T**							
	1	35.8 kb	incomplete	40	504168-539985	PHAGE Pandor inopinatum NC026440	64.01%
	1	6.7 kb	incomplete	10	3687923-3694705	PHAGE Entero phi92 NC 023693	55.68%
	2	9.4 kb	incomplete	50	4437965-4447371	PHAGE Stx2 c 1717 NC 011357	57.02%
***P. vervacti* DSM 23571^T**							
Chromosome	1	13.9 kb	intact	100	3334731-3348653	PHAGE Stx2 converting 1717 NC 011357	59.76%
	2	13.5 kb	incomplete	30	1074543-1088064	PHAGE Entero JSE NC 012740	57.32%
Plasmid pPV15	1	18 kb	incomplete	40	70851-88875	PHAGE Entero P1 NC 005856	60.84%
	2	4.8 kb	incomplete	20	2216127-2220937	PHAGE Salmon ST160 NC 014900	59.45%
	3	6.3 kb	questionable	80	3336760-3343139	PHAGE Stx2 c 1717 NC 011357	60.22%

4.1.4 Genome Data Deposition

The genome sequence data of 10 *Pandoraea* species genomes sequenced in this study were deposited in NCBI GenBank and the accession numbers are as summarised in Table 4.7.

Table 4.7: GenBank accession numbers of all *Pandoraea* spp. genomes sequenced in this study.

Organism	Contig (chromosome/plasmid)	GenBank accession number
Pandoraea apista DSM 16535^T	Chromosome	CP013481.2
	plasmid pPA35	CP013482.1
Pandoraea faecigallinarum DSM 23572^T	Chromosome	CP011807.3
	plasmid pPF72-1	CP011808.2
	plasmid pPF72-2	CP011809.2
Pandoraea norimbergensis DSM 11628^T	Chromosome	CP013480.3
Pandoraea pnomenusa DSM 16536^T	Chromosome	CP009553.3
Pandoraea pnomenusa RB38	Chromosome	CP007506.3
Pandoraea pulmonicola DSM 16583^T	Chromosome	CP010310.2
Pandoraea sputorum DSM 21091^T	Chromosome	CP010431.2
Pandoraea thiooxydans DSM 25325^T	Chromosome	CP011568.3
Pandoraea vervacti DSM 23571^T	Chromosome	CP010897.2
	plasmid pPV15	CP010898.2
Pandoraea oxalativorans DSM 23570^T	Chromosome	CP011253.3
	Plasmid pPO70-1	CP011518.2
	Plasmid pPO70-2	CP011519.2
	Plasmid pPO70-3	CP011520.2
	Plasmid pPO70-4	CP011521.2

4.2 Pan-genus Methylome Profile

The methylome profiles of the 10 *Pandoraea* spp. genomes were analysed in this study.

4.2.1 Pan-genus Motif Analysis

The RS_modification_and_motif analysis pipeline generated 4 output files namely: Modifications.csv file which contained statistical analysis data of polymerase kinetics at every genome bases; Modifications.gff file which included sequence contexts of putative modification sites (positions with significantly high IPD ratio, defined by modQV higher than 20 which correspond to p-value of 0.01 or higher); Motif_summary.csv file which contained genome-wide summary of MTases binding specificities (also termed as MTases recognition motif or sequence motifs) detected in the

genome; and Motifs.gff file which contained detailed information of all modified sites including the genomic positions and contexts, types of modification, modification QV score (modQV), and coverage.

A total of 28 motifs were identified in all analysed *Pandoraea* species, corresponding to 19 ^{m6}A sequence motifs, 1 ^{m5}C sequence motif and, and 8 sequence motifs which the modification type could not be identified (labeled as unknown) (Table 4.8). The observation where ^{m6}A methylated motifs represented the predominant motif type was similar to the occurrence frequencies of ^{m6}A motifs observed in the REBASE database of known MTase specificities (Roberts et al., 2015). Furthermore, the abundance of ^{m6}A motifs detection can also be attributed to its ease of detection through SMRT sequencing in which the DNA polymerase is sensitive to structural perturbations caused by ^{m6}A modifications (Clark et al., 2012). On the other hand, a total of 10 sequence motifs (Table 4.8: highlighted in grey) demonstrated characteristics such as low mean modification QV value (< 60), expanded motif structure, low methylation event detection (in the range of 16.23 % to 39.62 %) and undetermined modification types. These are most likely motifs associated with ^{m5}C modifications which have a relatively subtle and dispersed polymerase kinetic signature and are difficult to be detected at high statistical confidence, therefore rendering accurate motifs identification a challenge (Clark et al., 2012). The presence of these motifs suggested that more ^{m5}C methylated motifs could be present in the *Pandoraea* genomes. However, as the determination of the valid recognition motif is not within the scope of this dissertation, these false positive motifs were excluded from this study.

A total of 18 motifs were categorised as genuine motifs, comprised 17 ^{m6}A motif and 1 ^{m5}C motif. With the exception of the only ^{m5}C motif, CGATCG, all motifs had high mean modification QV (higher than 69) and the percentage of methylation events were in the range of 74.35 % to 100 %. In contrast, for CGATCG motif, the mean modification

QV was 48.18 and the methylation event detection percentage was 34.78 %. On the aspect of R-M types of these motifs, 16 of the motifs were categorised as Type II R-M systems whereas there were only 2 Type I motifs. A total of 4 motifs were detected to be novel, in which they were not documented and reported before in literature or within the REBASE database, namely CAYNNNNNNNCTCC (partner motif: GGAGNNNNNNNRTG), AGGNNNNNCTGA (partner motif: TCAGNNNNNCCT), CGATCG, and ATGAGC. The observation where these sequence motifs were novel motifs is consistent with the findings of Blow et al. (2016) where there is a high discovery rate of novel enzyme specificities among Type I and Type III R-M systems. The tendency of formation of novel specificities of Type I system are as a result of the high frequency of occurrence of genetic recombination on the DNA recognition architecture, the target recognition domain of the S subunit, of this system which result in the new enzyme specificities (Wilson & Murray, 1991a). With the exception of the CTGCAG, all motifs were not reported in literature (Molnarova et al., 1999).

Across all 10 *Pandoraea* genomes, majority of the motifs were unique to individual strains. Only GTWWAC motif was consistently observed to be present across all 10 *Pandoraea* genomes. CTGCAG motif on the other hand, was found to be present in two type species, namely *P. oxalativorans* and *P. vervacti*, with almost similar genome-wide methylation frequency.

Table 4.8: Motif summary of 10 *Pandoraea* spp. genomes analysed in this study. The motif rows which were of undetermined modification type and potentially represent motifs of ^{m5}C modifications were coloured grey.

Strain	Motifs	Modified position	R-M types	Type	% Motifs Detected	# of Motifs Detected	# of Motifs in Genome	Mean Modification QV	Mean Motif Coverage	Partner Motif
P. apista DSM 16535^T	GTWW**A**C	5	II beta	^{m6}A	96.16%	1102	1146	168.82	119.94	GTWWAC
P. faecigallinarum DSM 23572^T	GTWW**A**C	5	II beta	^{m6}A	98.17%	1394	1420	170.14	123.29	GTWWAC
	CGATCG	1	II	^{m5}C	34.78%	960	2760	48.18	128.94	
	CRGTGTCGA	3	II	unknown	28.97%	104	359	42.81	129.28	
	CGGTGTGNGND	3	II	unknown	28.47%	84	295	42.68	129.49	
P. norimbergensis DSM 11628^T	GTWW**A**C	5	II beta	^{m6}A	93.57%	932	996	112.89	79.11	GTWWAC
	C**A**YNNNNNNNCTCC	2	I gamma	^{m6}A	99.69%	970	973	126.78	81.57	GGAGNNNNNNNRTG
	GG**A**GNNNNNNNRTG	3	I gamma	^{m6}A	99.59%	969	973	123.06	80.71	CAYNNNNNNNCTC
P. oxalativorans DSM 23570^T	GTWW**A**C	5	II beta	^{m6}A	95.31%	1424	1494	69.73	48.94	GTWWAC
	CTGCAG	1	II	^{m6}A	99.46%	2224	2236	85.16	50.11	CTGCAG
	GCCGGCYR	1	II	unknown	26.32%	428	1626	43.68	50.7	
P. pnomenusa DSM 16536^T	GTWWAC	5	II beta	^{m6}A	95.2%	1169	1228	149.8	114.08	GTWWAC
	GAAMGTGGV	1	II	unknown	32.48%	114	351	41.31	124.39	
	AKGCCGCA	1	II	^{m6}A	25.22%	172	682	59.83	125.89	
	GWGVDTKG	1	II	unknown	24.04%	1280	5324	44.51	123.37	
	VAKRYASYW	2	II	^{m6}A	20.85%	567	2720	59.31	122.54	
	GTBTNVGG	1	II	unknown	16.23%	752	4633	44.56	120.75	
P. pulmonicola DSM 16583^T	GTWWAC	5	II beta	^{m6}A	95.77%	701	732	86.42	57.16	
	CGATCG	3	II	^{m6}A	74.35%	1713	2304	60.94	60.17	
P. sputorum DSM 21091^T	GCGATCGC	4	II beta	^{m6}A	95.3%	568	596	100.38	91.83	GCGATCGC
	GTWWAC	5	II beta	^{m6}A	94.63%	1111	1174	114.3	87.49	GTWWAC
P. thiooxydans DSM 25325^T	ATGAGC	4	III	^{m6}A	100.0%	2275	2275	123.83	79.07	
	GTWWAC	5	II beta	^{m6}A	97.98%	1113	1136	108.81	79.6	GTWWAC
P. vervacti DSM 23571^T	AGGNNNNNCTGA	1	I gamma	^{m6}A	100.0%	611	611	145.15	100.84	TCAGNNNNCCT
	TCAGNNNNNCCT	3	I gamma	^{m6}A	99.84%	610	611	159.98	104.15	AGGNNNNNCTGA
	CTGCAG	5	II	^{m6}A	100.0%	1536	1536	159.08	105.81	CTGCAG
	GTWWAC	5	II beta	^{m6}A	96.32%	1152	1196	121.74	100.18	GTWWAC
	CCAGGAAR	2	II	unknown	39.62%	124	313	47.89	90.83	
	CCWGGVNYD	2	II	unknown	22.61%	1208	5343	46.18	103	
P. pnomenusa RB38	GTWWAC	5	II beta	^{m6}A	95.59%	1191	1246	145.72	101.52	GTWWAC

4.2.2 R-M Genes Annotation and Assignment in The *Pandoraea* Genus

Annotations of all *Pandoraea* genomes identified a total of 36 putative DNA MTase-encoding genes (Table 4.9). In congruence with the positive correlation between number of R-M genes and genome size observed in a study conducted by Vasu and Nagaraja (2013), *P. norimbergensis* which had the largest chromosomal contig was annotated with the highest number of R-M systems (a total of 7 MTases). On the other hand, *P. thiooxydans* which had the smallest chromosomal contig were annotated with 3 R-M systems. However, it is also interesting to note that *P. pulmonicola* which had genome size of 5.87 Mb were detected to contain only 2 R-M systems.

From the R-M systems assignment analysis, majority of the detected Type II sequence motifs were able to be unambiguously assigned with a candidate corresponding MTase based on both R-M system types pairing as well as sequence similarity to known DNA MTases in the REBASE database (Table 4.10). These MTases (and their associated R-M genes) were named following the nomenclature guidelines to reflect the unequivocal assignment (Roberts et al., 2003). In all genomes, a Type II MTase was unambiguously assigned to be the corresponding MTase for the GTWWAC motif based on the high sequence similarity (>60%) of its predicted proteome to MTases which were documented to recognise GTWWAC sequence motif. These MTases were determined to be orphan MTases as no corresponding REase were detected in the close vicinity of these MTases. On the other hand, as all Type I sequence motifs detected in the *Pandoraea* genomes were novel motifs, the MTase assignment were made *via* R-M systems pairing. Furthermore, as only one set of Type I R-M genes were annotated in all genomes with Type I motifs, the assignment of the MTase were unambiguous. Similar R-M systems pairing method was also used to assign the candidate corresponding MTase for the Type III motif (ATGAGC) detected in *P. thiooxydans* genome and for two Type II ^{m6}A sequence motifs,

namely CGATCG and GCGATCGC, detected in the genome of *P. pulmonicola* and *P. sputorum* respectively.

Furthermore, 6 *Pandoraea* genomes were annotated to contain putative ^{m5}C MTases, of which 3 of these genomes were found to contain motifs with expanded structures, matching the prediction where those motifs potentially represented ^{m5}C motifs. On the other hand, in the genome of *P. vervacti*, despite the detection of 2 expanded motifs which potentially represented CCWGG motifs, no candidate ^{m5}C MTase gene were annotated. This most likely indicated the occurrence of a potentially missing plasmid from the assembly.

Interestingly, among some of the *Pandoraea* genomes, several genomes were identified to contain MTases which were homologous to known MTases documented to recognise and methylate recognition sequences such as CTGCAG (M.Pfa23572ORF23735P), CAGCTG (M.Pno11628ORF23340P), CCGCGG (M.Ppn16536ORF12180P), GCGATCGC (M.Ppn16536ORF12340P), TGGCCA (M.Ppn16536ORF24555P) and GTCGAC (M.PpnRB38ORF11825P). However, these predicted motifs were not observed in the respective genomes, suggesting that the enzymes were inactive. Several factors which could possibly contribute to the observed inactivity are namely mutation, xenogeneic silencing, or inactivity of the MTase during time of analysis.

Table 4.9: Putative R-M systems annotated in the analysed *Pandoraea* spp. genomes. The rows coloured in grey indicate R-M genes assigned to the recognition motifs detected in the motif analysis.

Strain	Gene Name	Gene	R-M Systems (Type)	Subtype	GenBank Locus Tag	Predicted Recognition Sequence[a]	Meth type[b]	Located in MGE?[c]
P. apista DSM 16535^T	*M.Pap16535I*	M	II	beta	AT395_RS10700	GTWW^{m6}**A**C	^{m6}A	No
	M.Pap16535ORF7915P	M	II	-	AT395_RS16935	-	^{m5}C	No
P. faecigallinarum DSM 23572^T	*M.Pfa23572I*	M	II	beta	AB870_RS08545	GTWW^{m6}**A**C	^{m6}A	No
	M.Pfa23572II	M	II		AB870_RS21945	^{m5}**C**GATCG	^{m5}C	No
	Pfa23572IIP	R	II	P	AB870_RS21950	-	-	No
	V.Pfa23572IIP	V	II		AB870_RS21940	-	-	No
	M.Pfa23572ORF2855P	M	II	-	AB870_RS14400	-	^{m5}C	No
	Pfa23572ORFBP	RM	II	G	AB870_RS06290	-	^{m6}A	No
	M.Pfa23572ORF23735P	M	II	gamma	AB870_23735	CTGCAG	-	plasmid pPF72-1
	Pfa23572ORF23735P	R	II	P	AB870_23740	CTGCAG	-	plasmid pPF72-1
P. norimbergensis DSM 11628^T	*M.Pno11628II*	M	II	beta	AT302_RS16260	GTWW^{m6}**A**C	^{m6}A	No
	M.Pno11628I	M	I	gamma	AT302_RS00035	C^{m6}**A**YNNNNNNNCTCC/ GG^{m6}**A**GNNNNNNNRTG	^{m6}A	No
	S.Pno11628I	S	I	-	AT302_RS00030	C^{m6}**A**YNNNNNNNCTCC/ GG^{m6}**A**GNNNNNNNRTG	^{m6}A	No
	M.Pno11628ORF13620P	M	II	alpha	AT302_RS05535	-	-	Intact phage
	M.Pno11628ORF21685P	M	II	alpha	AT302_RS13585	-	-	Intact phage
	M.Pno11628ORF23340P	M	II	beta	AT302_RS15235	CAGCTG	-	Incomplete phage
	M.Pno11628ORF27105P	M	II	-	AT302_RS15410	-	^{m5}C	Incomplete phage
P. oxalativorans DSM 23570^T	*M.Pox23570II*	M	II	beta	MB84_13795	GTWW^{m6}**A**C	^{m6}A	No
	M.Pox23570I	M	II	-	MB84_29725	CTGC^{m6}**A**G	^{m6}A	pPO70-2
	Pox23570IP	R	II	P	MB84_29720	CTGC^{m6}**A**G	^{m6}A	pPO70-2
	M.Pox23570ORF28765P	M	II	-	MB84_28765	-	^{m5}C	pPO70-1
	Pox23570ORF28765P	R	II	-	MB84_28775	-		pPO70-1
P. pnomenusa DSM 16536^T	*M.PpnI*	M	II	beta	LV28_RS38940	GTWW^{m6}**A**C	^{m6}A	No
	M.PpnORF12340P	M	II	beta	LV28_RS26195	GCGATCGC	-	Intact Phage
	M.PpnORF12180P	M	II	-	LV28_RS26340	CCGCGG	^{m5}C	No
	V.PpnORF12180P	V	II	-	LV28_RS26335	CCGCGG	-	No
	M.PpnORF24555P	M	II	alpha	LV28_RS38375	TGGCCA	-	Incomplete phage
	M.PpnORF24550P	M	II	-	LV28_RS38380	-	-	Incomplete phage
	M.PpnORF19115P	M	II	alpha	LV28_RS43660	-	-	Intact phage
P. pnomenusa RB38	*M.PpnRB38I*	M	II	beta	DA70_RS09320	GTWW^{m6}**A**C	^{m6}A	No
	M.PpnRB38ORF11825P	M	II			GTCGAC	^{m5}C	No
	PpnRB38ORF11825P	R	II	P	DA70_RS04380	GTCGAC	-	No
	C.PpnRB38ORF11825P	C	II			GTCGAC	-	No
	V.PpnRB38ORF11825P	V	II		DA70_RS04400	GTCGAC	-	No
	M.PpnRB38ORF9795P	M	II		DA70_RS06350	-	^{m5}C	Intact phage
	M.PpnRB38ORF8250P	M	II	alpha	DA70_RS07810	-	-	No
P. pulmonicola DSM 16583^T	*M.Ppu16583I*	M	II	beta	RO07_RS09895	GTWW^{m6}**A**C	^{m6}A	No
	M.Ppu16583ORFAP	M	II	alpha	RO07_RS10540	CG^{m6}**A**TCG	^{m6}A	No

Table 4.9, continued.

Strain	Gene Name	Gene	R-M System (Types)	Subtype	GenBank Locus Tag	Predicted Recognition Sequence[a]	Meth type[b]	Located in MGE? [c]
P. sputorum DSM 21091^T	M.Psp21091I	M	II	beta	NA29_RS14775	GTWW^{m6}**A**C	^{m6}A	No
	M.Psp21091II	M	II	beta	NA29_RS23275	GCG^{m6}**A**TCGC	^{m6}A	Intact phage
	M.Psp21091ORF7300P	M	II	alpha	NA29_RS03035	-		No
P. thiooxydans DSM 25325^T	M.Pth25325I	M	II	beta	ABW99_RS13875	GTWW^{m6}**A**C	^{m6}A	No
	M.Pth25325II	M	III	beta	ABW99_RS04315	ATG^{m6}**A**GC	^{m6}A	No
	Pth25325IIP	R	III		ABW99_RS04310	-		No
P. vervacti DSM 23571^T	M.PveNS15III	M	II	beta	UC34_RS09400	GTWW^{m6}**A**C	^{m6}A	No
	M.PveNS15I	M	I	gamma	UC34_RS00055	TC^{m6}**A**GNNNNNCCT	^{m6}A	No
	PveNS15IP					TC^{m6}**A**GNNNNNCCT	^{m6}A	No
	S.PveNS15I				UC34_RS00065	TC^{m6}**A**GNNNNNCCT	^{m6}A	No
	M.PveNS15II	M	II	beta	UC34_RS24965	CTGC^{m6}**A**G	^{m6}A	pPV15 and intersect with incomplete phage
	PveNS15IIP	R			UC34_RS24970	CTGC^{m6}**A**G	^{m6}A	pPV15 and intersect with incomplete phage

[a] Modified bases were highlighted in bold and labelled with the modification type in superscript, the corresponding modified base on the complementary strand were underlined. The recognition sequences in the uncoloured rows represented the motifs documented to be methylated by the homologues of the corresponding MTases (REBASE database).
[b] Meth type: Types of modification of the recognition specificities encoded by the putative MTase.
[c] Results obtained from integration of genome coordinates of the R-M genes with plasmid and prophage annotation data performed in chapter 4.1.

Table 4.10: Reasoning details on MTases assignments to respective recognition specificities.

Strain	Methyltransferases	Predicted Recognition Sequence	Details on assignment made
P. apista DSM 16535^T	M.Pap16535I	GTWWAC	M.Pap16535I showed high sequence similarity (92% - 100%) to various active Type II subtype beta MTases which recognised the GTWWAC sequence motif.
P. faecigallinarum DSM23572^T	M.Pfa23572I	GTWWAC	M.Pfa23572I showed high sequence similarity (93% -98%) to various active Type II subtype beta MTases which recognised the GTWWAC sequence motif.
	M.Pfa23572II	CGATCG	M.Pfa23572II showed high sequence similarity (63-86%) to various Type II ^{m5}C DNA MTases which recognised CGATCG, including 2 gold standard ^{m5}c MTases (M.XorKI and M.XorII).
P. norimbergensis DSM 11628^T	M.Pno11628II	GTWWAC	M.Pno1628II showed high sequence similarity (88-92%) to various active Type II subtype beta MTases which recognised the GTWWAC sequence motif.
	M.Pno11628I	CAYNNNNNNCTCC / GGAGNNNNNNRTG	M.Pno11628I was assigned based on matching R-M types and was unambiguously assigned as it was the only Type I R-M system detected within the genome.

Table 4.10, continued**.**

Strain	Methyltransferases	Predicted Recognition Sequence	Details on assignment made
P. oxalativorans DSM 23570^T	M.Pox23570II	GTWWAC	M.Pox23570II was assigned based on high sequence similarity (97 %-100 %) to active MTases known to recognise GTWWAC motif.
	M.Pox23570I	CTGCAG	M.Pox23570I was assigned based on high sequence similarity (74 %-86 %) to various MTases predicted to recognise the CTGCAG motifs.
	M.Pox23570ORFBP	GCCGGCYR	M.Pox23570ORFBP was assigned based on its identity as the sole ^{m5}C methyltransferase present within the genome and showed sequence similarity to MTases predicted to recognise GCCGGC motif (the most probable motif which generated the GCCGGCYR overcalled motif).
P. pnomenusa DSM 16536^T	M.Ppn16536I	GTWWAC	M.Ppn15636I showed high sequence similarity (93-100 %) to various active Type II subtype beta MTases which recognised the GTWWAC sequence motif.
P. pulmonicola DSM 16583^T	M.Ppu16583I	GTWWAC	M.Ppu16583I showed high sequence similarity (92-97 %) to various active Type II subtype beta MTases which recognised the GTWWAC sequence motif.
	M.Ppu16583ORFAP	CGATCG	M.Ppu16583ORFAP was assigned based on matching R-M types.
P. sputorum DSM 21091^T	M.Psp21091II	GCGATCGC	M.Psp21091II was assigned due to its presence as the only remaining Type II beta MTase available in the genome.
	M.Psp21091I	GTWWAC	M.Psp21091I showed high sequence similarity (93-99 %) to various active Type II subtype beta MTase which recognised the GTWWAC sequence motif.
P. thiooxydans DSM 25325^T	M.Pth25325ORF580P	ATGAGC	M.Pth25325ORF580P was assigned to be the best candidate for ATGAGC based on matching R-M types. ATGAGC motif matched the asymmetric pattern of Type III motif and M.Pth25325ORF580P is the only Type III R-M system present in the genome.
	M.Pth25325I	GTWWAC	M.Pth25325I shows high sequence similarity (84%) to various active Type II subtype beta MTases which recognised the GTWWAC sequence motif.
P. vervacti DSM 23571^T	M.PveNS15I	AGGNNNNNCTGA/ TCAGNNNNNCCT	M.PveNS15I was assigned based on matching R-M type and was unambiguously assigned as it was the sole Type I R-M system detected within the genome.
	M.PveNS15II	CTGCAG	M.PveNS15II was assigned based on its high sequence similarity (76-92 %) to various active Type II N6-adenine DNA MTases which recognised the CTGCAG motifs
	M.PveNS15III	GTWWAC	M.PveNS15III showed high sequence similarity (93-99 %) to various active Type II subtype beta MTases which recognised the GTWWAC sequence motif.
P. pnomenusa RB38	M.PpnRB38I	GTWWAC	M.PpnRB38I showed high sequence similarity (93-100 %) to various active Type II subtype beta MTases which recognised the GTWWAC sequence motif.

4.2.2.1 Association of R-M Systems with Mobile Genetic Elements (MGEs)

Various studies reported that the wide distribution and rapid evolution of R-M systems in the prokarytic genomes were associated with horizontal gene transfer events particularly by the means of mobile genetic elements (MGEs) such as plasmids and prophages (Kobayashi, 2001; Oliveira et al., 2014). In this study, the association of the annotated R-M genes with plasmids and prophages were investigated.

Firstly, with the exception of R-M systems of *P. apista*, *P. pulmonicola*, and *P. thiooxydans*, all other genomes were shown to have R-M genes which were associated with MGEs (Table 4.9). Interestingly, these three genomes were also among the genomes which contained the least amount of R-M systems. Overall, a total of 4 R-M systems were associated with plasmid sequences whereas a total of 10 R-M systems were associated with prophages. Among these, only the genes harboured in plasmids were complete R-M systems (comprising both MTases and restriction endonuclease (REase)) whereas most genes associated with prophages were solitary MTases. Furthermore, only Type II R-M systems were associated with these MGEs.

The plasmids of *P. faecigallinarum*, *P. oxalativorans* and *P.vervacti* were found to harbour an identical complete R-M system which recognition sequence were predicted to be C**T**GC^{m6}**A**G. These MTases were determined to belonged to the Type II systems and of the ɣ subclass. This categorisation was according to firstly, the matching of recognition specificities consensus sequence (TNN<u>A</u> where adenine represent the methylated base) and the corresponding conserved amino acid sequence motif I and II (motif I: G-G-G and motif II: NPPY) as proposed by Malone et al. (1995). BLASTx search of the predicted protein sequences of these MTases against the non-redundant (nr) protein sequences database demonstrated the presence of their homologs in a divergent genera of bacteria suggesting an association with high rate of horizontal gene transfer occurrences, possibly mediated by plasmid transfer (Figure 4.6). Intriguingly, despite the

presence of this complete R-M system within the genome of *P. faecigallinarum*, the motif of CTGCAG were not detected to be methylated. Further observation of truncation and presence of a relatively large amount of potentially inactivating frameshift mutations in the *P. faecigallinarum* CTGCAG R-M genes indicated that this system was most likely inactivated due to accumulation of debilitating mutation.

Description	Max score	Total score	Query cover	E value	Ident	Accession
SAM-dependent methyltransferase [Pandoraea oxalativorans]	962	962	99%	0.0	100%	WP_052654676.1
SAM-dependent methyltransferase [Pandoraea vervacti]	955	955	99%	0.0	99%	WP_044458869.1
SAM-dependent methyltransferase [Pandoraea faecigallinarum]	861	861	96%	0.0	92%	WP_047909215.1
SAM-dependent methyltransferase [Acetobacter sp. DsW_47]	818	818	99%	0.0	83%	WP_088652119.1
SAM-dependent methyltransferase [Sulfuriferula sp. AH1]	802	802	98%	0.0	83%	WP_087448437.1
SAM-dependent methyltransferase [Dickeya sp. S1]	798	798	99%	0.0	81%	WP_049954844.1
SAM-dependent methyltransferase [Enterobacter cloacae complex sp. SMART_783]	797	797	99%	0.0	81%	WP_058657688.1
SAM-dependent methyltransferase [Gluconobacter oxydans]	796	796	99%	0.0	82%	WP_052454480.1
MULTISPECIES: SAM-dependent methyltransferase [Proteobacteria]	795	795	99%	0.0	80%	WP_043496682.1
PREDICTED: modification methylase BsuBI-like [Diachasma alloeum]	793	793	98%	0.0	82%	XP_015125780.1
SAM-dependent methyltransferase [Salmonella enterica]	790	790	99%	0.0	80%	WP_048348777.1
SAM-dependent methyltransferase [Citrobacter koseri]	789	789	99%	0.0	80%	WP_049008614.1
SAM-dependent methyltransferase [Klebsiella pneumoniae]	785	785	99%	0.0	80%	WP_080883549.1
MULTISPECIES: SAM-dependent methyltransferase [Xanthomonas]	785	785	99%	0.0	79%	WP_033483225.1
MULTISPECIES: SAM-dependent methyltransferase [Enterobacteriaceae]	784	784	99%	0.0	80%	WP_062939558.1
SAM-dependent methyltransferase [Edwardsiella tarda]	783	783	99%	0.0	80%	WP_078000494.1
restriction endonuclease subunit M [Escherichia coli]	783	783	99%	0.0	79%	WP_001410808.1
SAM-dependent methyltransferase [Pantoea sp. A4]	783	783	99%	0.0	79%	WP_017347759.1
MULTISPECIES: SAM-dependent methyltransferase [Enterobacterales]	782	782	99%	0.0	79%	WP_060441676.1
SAM-dependent methyltransferase [Klebsiella pneumoniae]	782	782	99%	0.0	79%	WP_085898539.1
restriction endonuclease subunit M [Escherichia coli]	780	780	99%	0.0	79%	WP_001386842.1
SAM-dependent methyltransferase [Thiobacillus sp. 65-29]	780	780	99%	0.0	79%	OJZ19037.1
SAM-dependent methyltransferase [Aeromonas australiensis]	778	778	99%	0.0	79%	WP_040097874.1
SAM-dependent methyltransferase [Edwardsiella hoshinae]	776	776	99%	0.0	81%	WP_024523835.1
SAM-dependent methyltransferase [Pseudomonas syringae]	774	774	99%	0.0	78%	WP_003380659.1
SAM-dependent methyltransferase [Pseudomonas syringae group genomosp. 3]	771	771	99%	0.0	78%	WP_054088665.1
modification methylase BsuBI [Xanthomonas citri pv. mangiferaeindicae LMG 941]	768	768	97%	0.0	79%	CCG39364.1
SAM-dependent methyltransferase [Pseudomonas sp. HPB0071]	759	759	99%	0.0	76%	WP_010799773.1
SAM-dependent methyltransferase [Snodgrassella alvi]	758	758	99%	0.0	77%	WP_084665159.1
SAM-dependent methyltransferase [Aeromonas jandaei]	757	757	99%	0.0	76%	WP_041207768.1
SAM-dependent methyltransferase [Comamonas aquatica]	751	751	98%	0.0	77%	WP_042420086.1
SAM-dependent methyltransferase [Escherichia coli]	746	746	92%	0.0	82%	WP_023277465.1

Figure 4.6: Result of BLASTx search of CTGCAG MTase sequence against the non-redundant (nr) protein sequences database.

4.3 Identification of A Novel Class of Orphan Methyltransferase, GTWWAC Methyltransferase

4.3.1 GTWWAC Motif Analysis

A Type II palindromic ^{m6}A sequence motif, GTWW**A**C (from the direction of 5' to 3', the methylated adenine base is indicated as **A**, and the thymine pairing of the methylated adenine base on the complementary strand was represented as T) was identified in all 10 analysed *Pandoraea* spp. genomes. The methylation state of this sequence motif in all genomes were consistently incomplete where the average percentage of methylation events were 95.87 % (SD: 0.01), suggesting that orphan MTases were most likely responsible for the methylation of these motifs. Based on its structure which matches the consensus structure of (G/C)$N_{(0-3)}$M$N_{(0-2)}$(G/C), this recognition motif was grouped into the β subgroup of type II ^{m6}A MTase target motif (Malone et al., 1995). A slight variation in the total number of motifs detected within each genome was observed, where genome of *P. oxalativorans* was found to have the most number of motifs (1494), most likely as a result of the higher number of extrachromosomal contigs present in the genome, whereas the genome of *P. norimbergensis* had the least number of motifs (996) (Table 4.8).

4.3.1.2 Search of GTWWAC Motif within REBASE Database

The conservation of GTWWAC motif across *Pandoraea* genus indicated a possibility of ancient antecedent of its corresponding MTase in this genus and a representation of a conserved genetic element with essential gene function. In order to understand the prevalence of this gene, a REBASE recognition sequence search of GTWWAC motif was performed.

The output result from the search showed that within the REBASE database, a large number of MTases were documented or predicted to recognize the GTWWAC motif.

The identity of the organisms which encode these MTases were examined and was found to belong to six genera, namely *Burkholderia*, *Oxalabacteraceae*, *Cupriavidus*, *Janthinobacterium*, *Massilia*, and *Pandoraea*. These genera belonged in the order of *Burkholderiales* which prompted the hypothesis that the GTWWAC motif could potentially be a widespread MTase recognition motif within the order of *Burkholderiales.*

In order to test the idea, a comprehensive search of the deposited *Burkholderiales* PacBio methylation data available in the REBASE database was subsequently performed to identify genomes with active GTWWAC MTases. At the time of this study, a total of 19 *Burkholderiales* PacBio sequenced genomes were available (among 642 PacBio sequenced genomes deposited) which comprised genomes of organisms from the *Alcaligenaceae*, *Burkholderiaceae*, *Comamonadaceae*, and *Oxalobacteraceae* families. Among these genomes, only 12 were detected to contain methylated GTWWAC motifs where 11 of them belonged to the family of *Burkholderiacea* and 1 belonged to the family of *Oxalabacteraceae*. On the other hand, *Burkholderiales* genomes which did not contain methylated GTWWAC motifs were: 5 genomes from the families of *Alcaligenaceae* and *Comamonadaceae* and 2 genomes from the *Burkholderiaceae* family (Table 4.11). From these results, the presence of GTWWAC motifs appeared to be localised in the genomes of *Burkholderiaceae* and *Oxalobacteraceae* families. Although more methylation data will be needed to test the generality of this finding, the identifcation of this sequence motif and its prevalence among the organisms of these two families called for an exploration of the potential biological significance of its corresponding MTase.

Table 4.11: Distribution of GTWWAC motif within *Burkholderiales* genomes deposited in REBASE database.

Family	Genus	Rebase PacBio data	Accession	GTWWAC[a]	Methylation %	GTWWAC MTase homolog [b]
Alcaligenaceae	*Achromobacter*	*A. denitrificans USDA-ARS-USMARC-56712*	CP013923	no	-	no
		A. xylosoxidans subsp. xylosoxidans	LN831029	no	-	no
	Advenella	Nil	-		-	-
	Alcaligenes	Nil	-		-	-
	Bordetella	*B. pertussis* 137	CP010323	no	-	no
	Brackiella	Nil	-	-	-	-
	Castellaniella	Nil	-	-	-	-
	Derxia	Nil	-	-	-	-
	Kerstersia	Nil	-	-	-	-
	Oligella	Nil	-	-	-	-
	Pelistega	Nil	-	-	-	-
	Pigmentiphaga	Nil	-	-	-	-
	Pusillimonas	Nil	-	-	-	-
	Taylorella	Nil	-	-	-	-
	Tetrathiobacter	Nil	-	-	-	-
Burkholderiaceae	*Burkholderia*	*B. cenocepacia* DDS 22E-1	CP007783; CP007784	yes	98.8	yes (M.Bce22E1II)
		B. cenocepacia Jx2315	AM747720; AM747721; AM747723	yes	94.7	yes (M.BceJIV)
		B. cepacia ATCC 25416	CP012981; CP012982	yes	98.8	yes(M.Bce25416II)
		B. cepacia DDS 7H-2	CP007786; CP007787	no	98.8	yes (M.Bce7H2ORF4094P)
		B. cepacia LMG 16656	JTDP01000001; JTDP01000002; JTDP01000004; JTDP01000005	yes	94.8	yes (M.Bce16656II)
		B. mallei 2000031063	CP008731; CP008732	yes	99.4	yes (M.BmaBMKII)
		B. mallei China5	JPNX02000001	yes	98.1	yes (M.BmaBMZII)
		B. oklahomensis C6786	CP009555; CP009556	no	-	yes (Locus Tag:BG90_4691)
		B. pseudomallei 7894	CP009535; CP009536; NEBC8394	yes	98.2	yes (M.Bps7894II)
		B. pseudomallei Bp1651	CP012041; CP012042	yes	98.3	yes (M.Bps1651II)
		B.pseudomallei Pasteur 52237	CP009898; CP009899; JPNT01000009	yes	98.6	yes (M.BpsBEMORF3513P)
	Chitinimonas	Nil	-		-	-
	Cupriavidus	Nil	-		-	-
	Lautropia	Nil	-		-	-
	Limnobacter	Nil	-		-	-
	Pandoraea	*P. pnomenusa* DSM 16536	-	yes	95	M.Ppn16536I
		P. pnomenusa RB38	-	yes	95.5	M.PpnRB38I
	Paucimonas	Nil	-	-	-	-
	Polynucleobacter	Nil	-	-	-	-
	Thermothrix	Nil	-	-	-	-

Table 4.11, continued.

Family	Genus	Rebase PacBio data	Accession	GTWWAC[a]	Methylation %	GTWWAC MTase homolog [b]
Comamonadaceae	*Acidovorax*	*Acidovorax sp.* JHL-3	JAFU01000001; JAFU01000003; JAFU01000004	no		No
	Aquabacterium	Nil	-	-	-	-
	Brachymonas	Nil	-	-	-	-
	Comamonas	Nil	-	-	-	-
	Curvibacter	Nil	-	-	-	-
	Delftia	Nil	-	-	-	-
	Hydrogenophaga	Nil	-	-	-	-
	Ideonella	Nil	-	-	-	-
	Leptothrix	Nil	-	-	-	-
	Limnohabitans	Nil	-	-	-	-
	Malikia	Nil	-	-	-	-
	Pelomonas	Nil	-	-	-	-
	Polaromonas	*Polaromonas sp.* EUR3 1.2.1	JIBH01000001	no		No
	Rhodoferax	Nil	-	-	-	-
	Roseateles	Nil	-	-	-	-
	Sphaerotilus	Nil	-	-	-	-
	Tepidimonas	Nil	-	-	-	-
	Thiomonas	No methylation data available	-	-	-	-
	Variovorax	Nil	-	-	-	-
Oxalobacteraceae	*Collimonas*	Nil	-	-	-	-
	Duganella	Nil	-	-	-	-
	Glaciimonas	Nil	-	-	-	-
	Herbaspirillum	Nil	-	-	-	-
	Herminiimonas	Nil	-	-	-	-
	Janthinobacterium	Nil	-	-	-	-
	Massilia	Nil	-	-	-	-
	Naxibacter	Nil	-	-	-	-
	Oxalicibacterium	Nil	-	-	-	-
	Oxalobacter	*O. bacterium* AB14 (unclassified)	yes	yes	not available	M.Oba14I
	Telluria	Nil	-	-	-	-
	Undibacterium	Nil	-	-	-	-
Ralstoniaceae	*Ralstonia*	Nil	-	-	-	-
Sutterellaceae	*Parasutterella*	Nil	-	-	-	-
	Sutterella	Nil	-	-	-	-

[a] Presence of methylated GTWWAC motif in the deposited PacBio methylation data of the relevant organism.

[b] Presence of the GTWWAC MTase homolog within the genome, gene name as deposited in REBASE is stated when available. GenBank locus tag will be provided when the MTase was not annotated in REBASE but could be identified in the genome.

4.3.2 GTWWAC MTases Analyses

Recent studies suggested that orphan MTases have high degree of evolutionary conservation at the level of genus and family (Blow et al., 2016; Seshasayee et al., 2012). Since the GTWWAC motif analysis result indicated that this motif was found exclusively within the *Pandoraea* genus and within family of *Burkholderiaceae* and *Oxalobacteraceae*, it was hypothesised that the GTWWAC MTases could potentially be a new class of orphan MTases within these families. Therefore, several analyses were performed to characterise the GTWWAC MTases and to further discerned their evolutionary relationship.

Firstly, all MTases which were documented or predicted in REBASE to recognise GTWWAC motif were examined. A total of 101 Type II MTase sequences (inclusive of 5 REBASE gold standard MTases and 10 MTases annotated in the genomes of the *Pandoraea* spp. sequenced in this study) were included in the analysis. All these MTases were determined to be a potential orphan MTase as no cognate REase were identified in proximity of these genes. Conserved domain analysis indicated the predicted proteome of all 101 MTases contained a C-terminal MTase specific domain (cl17173: AdoMet_MTases) and a N^6 – N^4 MTase Pfam domain (pfam01555), the characteristic domains for both N^4 cytosine-specific and N^6 adenine-specific DNA methylases. Furthermore, sequence alignment analysis performed using the predicted amino acid sequences of all MTases also revealed presence of several stretches of highly conserved amino acid sequence blocks as shown in Figure 4.7. Motif IV, V, VI, VII, X and II were identified and the arrangement order of these motifs, particularly with the preceding position of motif IV (DPPY) relative to motif I (F-G-G), indicated that these MTases belong to the β subgroup of ^{m6}A MTases, matching the predicted R-M types of the GTWWAC motif (Malone et al., 1995; Wilson & Murray, 1991b).

Figure 4.7: Graphical representation of the consensus amino acid sequence alignments of 101 candidate GTWWAC MTases. Motifs (IV, V, VI, VII, X, I and II) were labelled according to their positions. The arrangement of catalytic region (represented by motif IV-VII) preceding the S-adenosyl methionine (AdoMet) -binding region (represented by motif X, I and II) is a signature of β subgroup ^{m6}A MTases. The size of each symbol corresponds to the frequency of the amino acid residue at specific position.

Subsequently, pairwise amino acid sequence similarity analysis demonstrated that all candidate MTases shared an overall high primary sequence similarity (gene sequence similarity: 70-100 %; translated amino acid sequence similarity values: 62-100 %) as shown in Figure 4.8, indicating significant homology. Based on the neighbour-joining clustering analysis, these MTases were separated into 3 main clusters: the cluster of

Pandoraea genus, *Burkholderia* genus and of the *Oxalabacteraceae* family. MTases of the *Pandoraea* genus demonstrated high inter-species sequence identity (>96 %), with the exception of the MTase of *P. thiooxydans*, in congruence with the genome relatedness assessment performed (Subchapter 4.1.3.1). On the other hand, for GTWWAC MTases of the *Burkholderia* genus, clustering of the MTases were in accordance to species and demonstrated agreement with the 16S rRNA phylogenetic positions of *Burkholderia* spp. (Coenye et al., 2001).

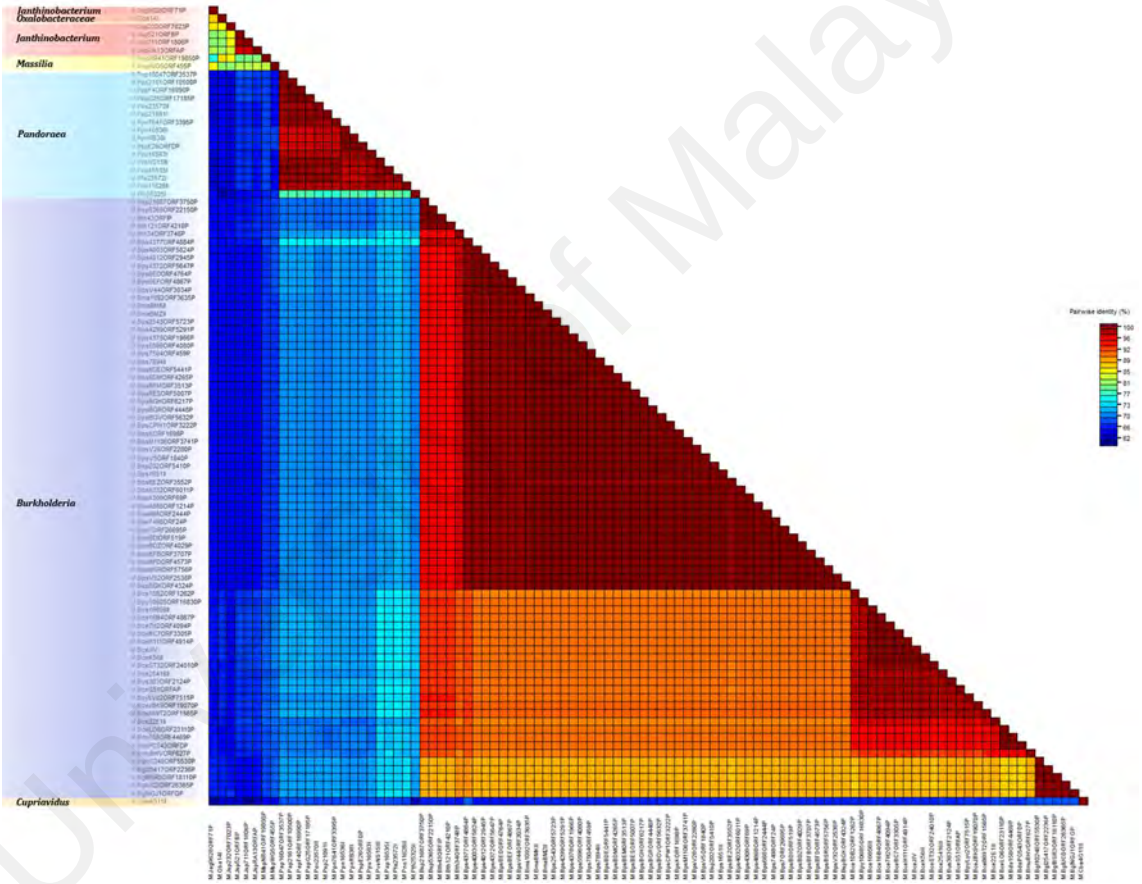
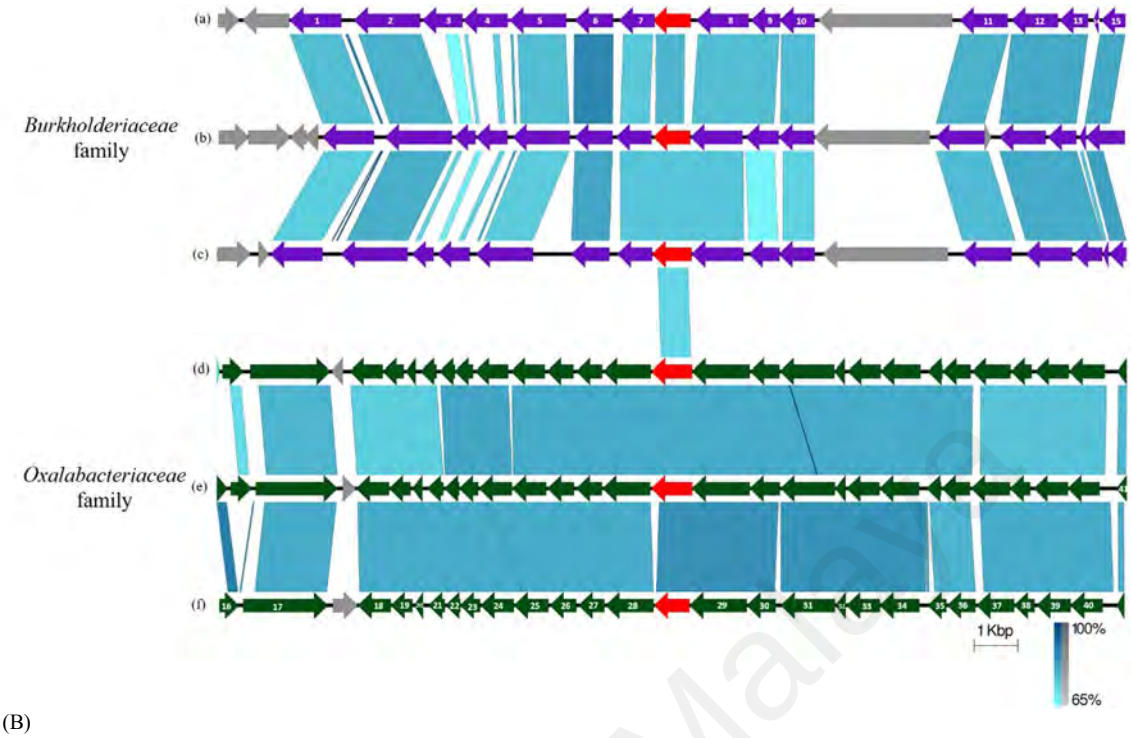


Figure 4.8: Colour-coded pairwise identity matrix for 101 candidate GTWWAC MTases. NW algorithms implemented in MUSCLE alignment tool was used to compute the pairwise identities of each MTase. The range of percentage identity score of all sequences are between 62 to 100. The corresponding genus of each MTase were displayed at the side bar. All MTases were shown to display high sequence similarities, particularly among organisms of the same species.

Each MTase was further examined from the aspect of gene size and neighbouring gene context. The average gene size of the GTWWAC MTase was 863 nt (SD: 34.4) and the size of the MTases showed high intra-species similarity (**Appendix C**). In addition, analysis of neighbouring genomic context (10kb upstream and downstream neighbouring gene blocks) demonstrated that the GTWWAC MTases within genomes of each family were located in a highly syntenic region where a high level of gene neighbourhood profile conservation was observed as shown in Figure 4.9. The degree of conservation of neighbouring genes context was observed to be higher in the family of *Oxalabacteraceae* as compared to those of *Burkholderiaceae*. The neighbouring genes included operons predicted for amino acid biosynthesis, protein transports, ABC transporters, cell wall maintenance and biogenesis as well as single genes with functional role in amino acid biosynthesis, nucleic acid biosynthesis, defense, metabolite biosynthesis, pseudouridine biosynthesis, cell wall biogenesis, transcription, replication, recombination and repair. With the exception of the GTWWAC MTase, no significant inter-family neighbouring genes nucleotide sequence similarities were observed. However, when comparing the predicted proteome of the neighbouring genes (Figure 4.9 (B)), a low degree (>20 %, e-value: 0.0001) of sequence similarities can be observed. Interestingly, in both families, the GTWWAC MTases were located in the middle of an amino acid biosynthetic operon, namely the *trp* (tryptophan) operon in the *Burkholderiaceae* genomes and the *his* (histidine) operon in the *Oxalabacteraceae* genomes.

(A)



(B)

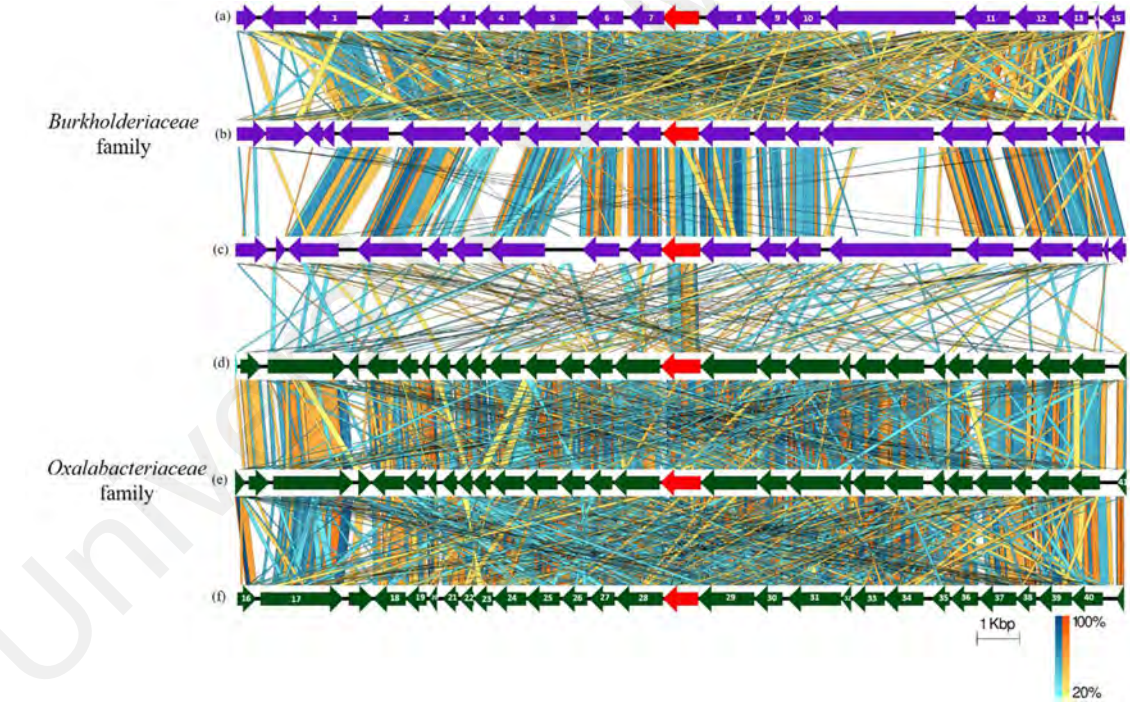


Figure 4.9: (A) BLASTN (B) TBLASTX comparison of GTWWAC MTases. The comparison is made along with the adjacent 10kb block neighbouring genes of the *Burkholderiaceae* and *Oxalabacteraceae* family: (a) *Pandoraea* spp. (b) *Burkholderia* spp. (c) *Cupriavidus* spp. (d) *Janthinobacterium* spp. (e) *Massilia* spp. (f) *Oxalobacteraceae* spp. Vertical blocks between sequences indicate regions of shared similarity (blue for matches in the same direction or orange for inverted matches). The CDSs which represent GTWWAC MTase are coloured in red. (C) Labels of each CDS (table form), CDSs which constitutes an operon structure are highlighted in grey.

(C)

Number	Gene	Gene Name
1	*metZ*	O-succinylhomoserine sulfhydrylase
2	*purF*	amidophosphoribosyltransferase
3	*cvpA*	colicin V synthesis protein/bacteriocin production protein
4	-	SPOR domain-containing protein
5	*folC*	bifunctional folylpolyglutamate synthase/dihydrofolate synthase
6	*accA*	acetyl-CoA carboxylase subunit beta
7	*trpA*	tryptophan synthase subunit alpha
8	*trpB*	tryptophan synthase subunit beta
9	*trpC*	N-(5'-phosphoribosyl)anthranilate isomerase
10	*truA*	tRNA pseudouridine(38-40) synthase
11	*asd*	aspartate-semialdehyde dehydrogenase
12	*leuB*	3-isopropylmalate dehydrogenase
13	*leuD*	3-isopropylmalate dehydratase small subunit
14	-	entericidin
15	*leuC*	3-isopropylmalate dehydratase large subunit
16	-	helicase SNF2
17	-	hypothetical protein
18	-	hypothetical protein
19	*tatC*	twin arginine-targeting protein translocase TatC
20	*tatB*	Sec-independent protein translocase TatB
21	*tatA*	Sec-independent protein translocase TatA
22	*hinT*	histidine triad nucleotide-binding protein
23	*hisE*	phosphoribosyl-ATP pyrophosphatase
24	*hisI*	phosphoribosyl-AMP cyclohydrolase
25	*hisF*	imidazole glycerol phosphate synthase subunit HisF
26	*hisA*	1-(5-phosphoribosyl)-5-((5-phosphoribosylamino)methylideneamino)imidazole-4-carboxamide isomerase
27	*hisH*	imidazole glycerol phosphate synthase subunit HisH
28	*hisB*	imidazoleglycerol-phosphate dehydratase
29	*hisC*	histidinol-phosphate aminotransferase
30	*hisD*	histidinol dehydrogenase
31	*hisG*	ATP phosphoribosyltransferase
32	*murA*	UDP-N-acetylglucosamine 1-carboxyvinyltransferase
33		BolA family transcriptional regulator
34	*yadH*	ABC-2 type transport system permease protein
35	*ccmA*	ABC-2 type transport system ATP-binding protein/ABC transporter ATP-binding protein
36	*mlaB*	phospholipid transport system transporter-binding protein
37	*mlaC*	phospholipid transport system substrate-binding protein
38	*mlaA*	phospholipid-binding lipoprotein MlaA
39	*mlaD*	outer membrane lipid asymmetry maintenance protein MlaD
40	*mlaE*	phospholipid/cholesterol/gamma-HCH transport system permease protein
41	*mlaF*	phospholipid/cholesterol/gamma-HCH transport system ATP-binding protein
42	*gltD*	glutamate synthase subunit beta

Figure 4.9, continued.

Distribution of GTWWAC homologs beyond the organisms annotated by the REBASE database was determined by searching against the NCBI non redundant (nr) database. Firstly, in the genus of *Burkholderiaceae* and *Oxalabacteraceae* family which have reliable genome data available, the homologs of GTWWAC MTases were found to be present. Interestingly, beyond these two families, several GTWWAC MTase homologs were also detected to be present as shown in Table 4.12. The homologs detected in the genome of *Mumia flava* and the species of *Betaproteobacteria* shared largely similar gene context as those of the *Burkholderiaceae* genomes whereas the GTWWAC MTase

homolog found in *Rugamonas rubra*, a *Gammaproteobacteria*, shared gene context similarities with the *Oxalabacteraceae* genomes.

Lastly, assessment on the relatedness between the candidate GTWWAC MTases was performed by using phylogenetic analysis (Figure 4.10). From the analysis, the clustering pattern of the MTases demonstrated phylogenetic congruence where the MTases were grouped into 5 major groups: *Burkholderia* genus group, *Pandoraea* genus group, *Betaproteobacteria* bacterium group, *Oxalabacteraceae* family group, and unclassified GTWWAC MTase homologs group. For MTases within the *Burkholderia* genus and *Pandoraea* genus, majority of GTWWAC MTases from the same species clustered together in the same branch whereas all GTWWAC MTases of the *Oxalabacteriaceae* family formed a single clade. Interestingly, the GTWWAC MTase homolog found in the *M. flava* genome, a novel Actinobacterial strain, was clustered among the *Burkholderia cepacia complex* group, suggesting a potential horizontal gene acquisition. Similarly horizontal gene transfer event indication was also observed in the grouping of the GTWWAC MTase homolog of *R. rubra* in the *Oxalabacteraceae* family clade. As both of these genomes contained only a single genome record at the genus level in GenBank, the degree of distribution of GTWWAC MTase homologs among these genomes were unable to be investigated. However, in both genomes, a putative transposase was found to be located upstream of the GTWWAC MTase homolog at approximately similar distance (16.9 kb and 18.3 kb), potentially representing the horinzontal gene transfer mechanism. Even more interesting, a similarly spaced putative transposase was also found in the genomes of *Oxalobacteraceae bacterium* AB14 and *B. cepacia* ATCC 25416, the closest GTWWAC MTase homolog phylogenetic neighbour of *R. rubra* and *M. flava*. On the other hand, in the other genomes, no transposase was observed in the vicinity of gene clusters bearing GTWWAC MTases, which is suggestive that these MTases were not of recent acquisitions or were acquired *via* the means of

vertical transmission. Furthermore, the GTWWAC MTase homologs of the *Betaproteobacteria bacterium* genomes can be observed to form a cluster with *Cupriavidus basilensis*, which is indicative that these genomes represented part of the *Burkholderiales* family with close phylogenetic relationship with the *Cupriavidus* genus.

Table 4.12: The organisms (not within the *Burkholderiaceae* and *Oxalabacteriaceae* families) found to harbour GTWWAC homologs in their genomes.

Organism	accession	Sequence identity to M.PpnRB38I	Query cover	Neighbouring gene context*
Mumia flava	KHL12633.1	76%	80%	yes. *Burkholderiaceae*
Betaproteobacteria bacterium	OGA17090.1	73%	78%	yes. *Burkholderiaceae*
Betaproteobacteria bacterium	OFZ90379.1	73%	78%	yes. *Burkholderiaceae*
Betaproteobacteria bacterium	OGA15275.1	72%	78%	yes. *Burkholderiaceae*
Betaproteobacteria bacterium	OGA45437.1	72%	78%	yes. *Burkholderiaceae*
Omnitrophica bacterium	OGW94241.1	63%	80%	No
Candidatus Nomurabacteria bacterium	KKR78860.1	62%	79%	No
Bellilinea caldifistulae	WP_061918248.1	65%	77%	No
Candidatus Kuenenbacteria bacterium	OGG91171.1	63%	79%	No
Candidatus Giovannonibacteria bacterium	OGF78721.1	64%	78%	No
Parcubacteria group bacterium	OIP79067.1	61%	78%	No
Rugamonas rubra	SFM36086.1	69%	76%	yes. *Oxalabacteraceae*
Parcubacteria group bacterium	KKT40487.1	64%	78%	No
Candidatus Kuenenbacteria bacterium	OGG98827.1	63%	80%	No
Candidatus Kuenenbacteria bacterium	OGG95971.1	67%	73%	No
Parcubacteria group bacterium	KKS49708.1	61%	77%	No
Candidatus Synechococcus spongiarum	WP_074456991.1	60%	77%	No

* The neighbouring gene context of the detected GTWWAC MTase homolog which shows similarity to the GTWWAC MTases of the *Burkholderiaceae* or the *Oxalabacteraceae* family.

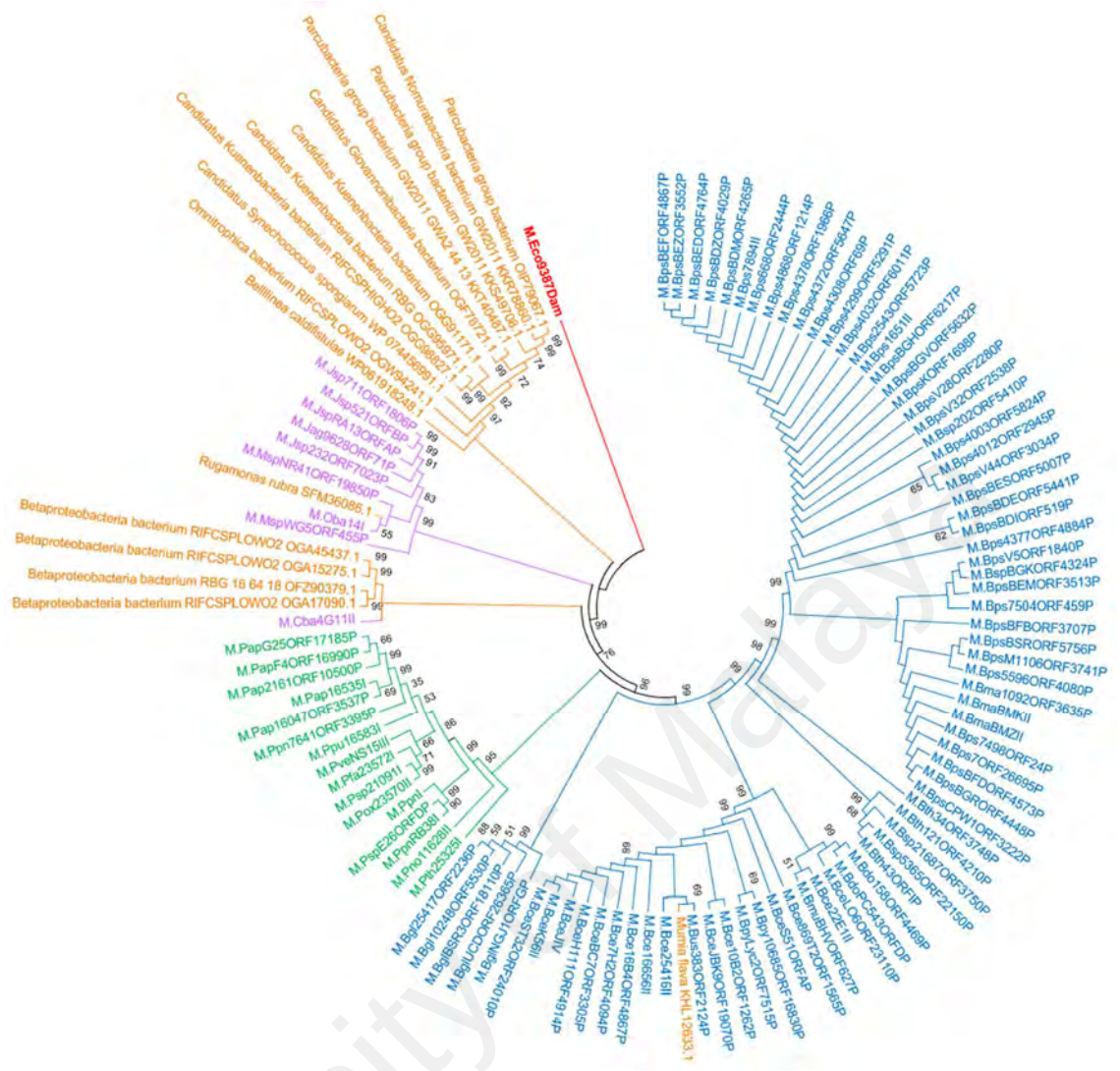
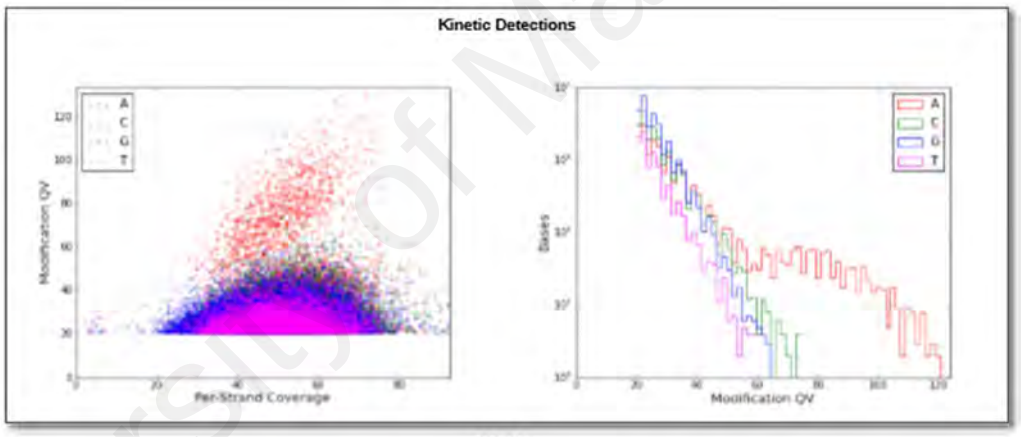


Figure 4.10: Phylogenetic analyses of all candidate GTWWAC MTases. The tree was inferred by the maximum likelihood method using Kimura 2-parameter (K2+G+1) model. The analysis included 118 MTases nucleotide sequences. Each MTase belonging to the *Burkholderiaceae* and *Oxalabacteraceae* families were represented by their enzyme name as listed in REBASE database. The GTWWAC MTase homologs identified beyond these two families were labelled with the organism name and GenBank accession number according to Table 4.12. Major clusters were indicated in different colours: *Burkholderia* genus group (blue), *Pandoraea* genus group (green), *Oxalabacteraceae* family group (purple), and GTWWAC MTase homologs group (orange). Only bootstrap value higher than 50 were displayed.

4.3.3 Verification of Activity and Specificity of GTWWAC MTase

In order to confirm the recognition motif specificity of the putative GTWWAC MTases, *M.PpnI* gene was cloned and expressed into a Dam and Dcm deficient *E. coli* strain. As expected, SMRT sequencing of the genomic DNA from this recombinant strain revealed that the GTWWAC target motifs of the *E. coli* genome were modified with methylation at the adenine base, matching the result detected in the *Pandoraea* species genomes, hence confirming the activity and specificities of GTWWAC MTases (Figure 4.11). AACNNNNNNGTGC (partner motif: GCACNNNNNNGTT) represent the positive control MTase, EcoKI type I R-M system.

(a)



(b)

Motifs

Motif Summary

Motifs	Modified Position	Type	% Motifs Detected	# Of Motifs Detected	# Of Motifs In Genome	Mean Modification QV	Mean Motif Coverage	Partner Motif
AACNNNNNNGTGC	2	m6A	95.23%	599	629	75.39	49.90	GCACNNNNNNGTT
GCACNNNNNNGTT	3	m6A	94.91%	597	629	77.26	50.64	AACNNNNNNGTGC
GTWWAC	5	m6A	33.47%	3219	9618	38.05	53.41	GTWWAC

Figure 4.11: Methylome determination of *E. coli* recombinant construct harbouring *M.PpnI* gene. (a) Modifications report (Modification QV vs Coverage scatterplot and Modification QV Histogram) which demonstrated a relatively distinct modification of the adenine bases (b) Motifs report which contains genome-wide summary of the MTase recognition motif detected in the genome.

4.4 Comparative Analysis of *Pandoraea* spp. Genome-wide GTWWAC Motif Distribution

Variation in the DNA methylation patterns of orphan MTases are known to influence gene expression regulation and DNA replication control (Casadesús & Low, 2006; Low et al., 2001). Therefore, in this analysis, genome wide distribution of GTWWAC motif and the interaction of these motif locations with annotated CDSs were analysed in all *Pandoraea* spp. genomes. The analysis was made with a hypothesis that CDSs which were associated with specific methylome sites, namely those with significantly higher than average methylation frequency or reproducibly unmethylated motif sites, could provide an indication of the functional role of the GTWWAC methylation in the *Pandoraea* spp. The raw data used in this analysis are available in **Appendix A (ii)**.

4.4.1 Genome-wide GTWWAC Motif Distribution of The *Pandoraea* spp.

Genome-wide GTWW^{m6}AC motif methylation density distribution plots were generated to visualise the methylome distribution in all genomes (Figure 4.12). Inspection of these plots revealed a generally even distribution of the methylation frequency across the genome with several distinct regions of hyper-frequency. To further characterise these regions, genome bins were grouped into two categories, namely the hypermethylated bins and methylation hotspot bins. The pan-genus mean frequency value was 0.2 (±standard error of mean 0.0028, n = 55016) whereas the standard deviation (*SD*) value was 0.66. Therefore, each 1kb genome bin with an occurrence frequency of methylated motifs higher than 2 was considered as a methylation hotspot bin whereas each bin with a motif occurrence frequency greater than 4 was considered as hypermethylated. Each hypermethylated and methylation hotspot bin was annotated to determine the coding sequences (CDSs) present within.

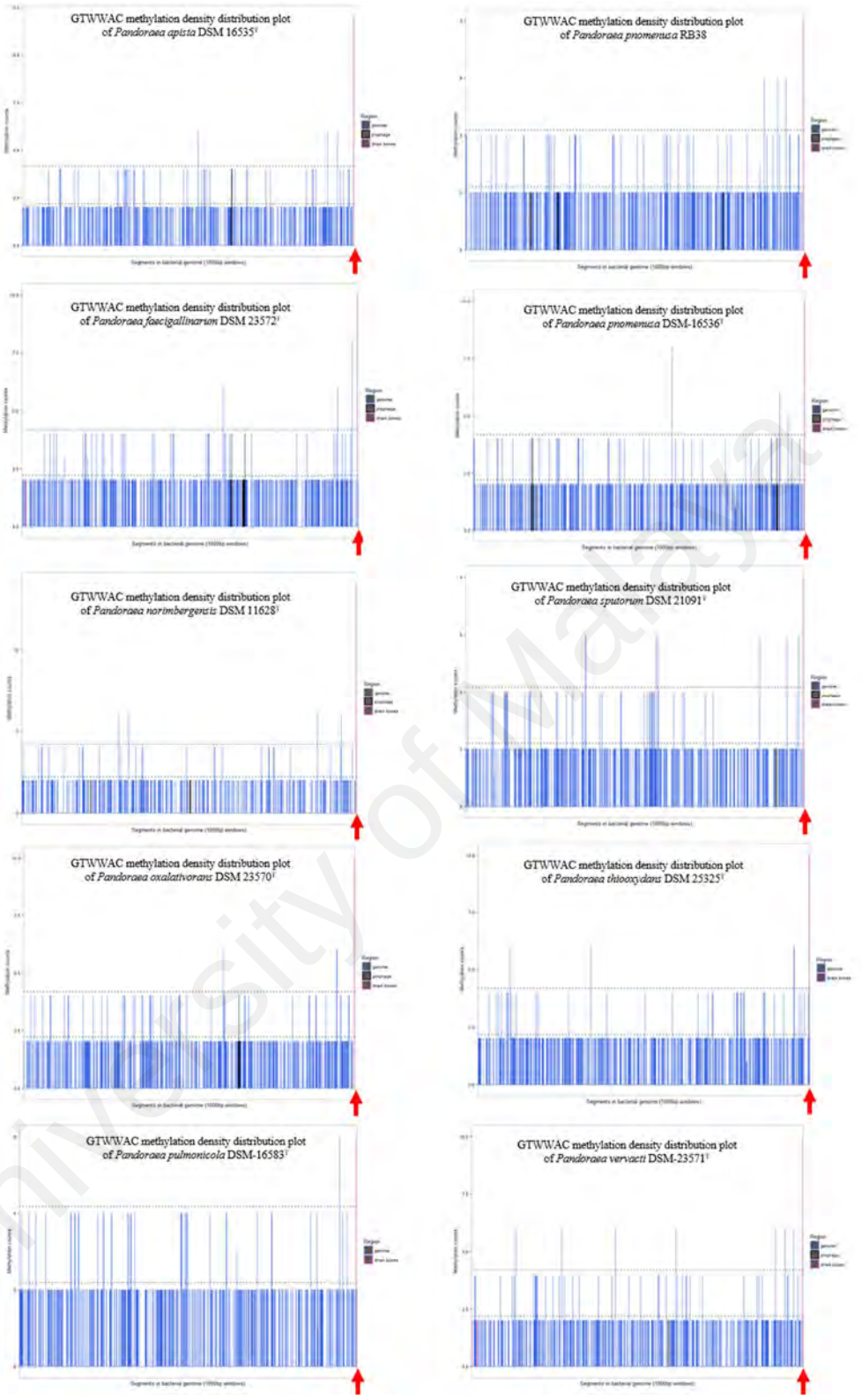


Figure 4.12: Genome wide GTWWAC motif methylation frequency distribution plot of *Pandoraea* spp. Y-axis represent the frequency of motif methylation detected per 1kb genome window whereas X-axis represent the genome positions (bp). The methylome bins are coloured according to the overlapped region: genomic region (blue), putative prophages (black), dnaA boxes (pink). The red arrows denote the locations of putative origin of replication (*oriC*) sites. The methylation frequency threshold value which categorise the genome bins into methylation hotspot bins and hypermethylated bins are represented by 2 dotted lines.

4.4.1.1 Intragenic and Intergenic Distribution of GTWWAC Motif

The intra-genic (GR) and inter-genic (IGR) distribution of the GTWWAC motif of three different methylation status (methylated, hemimethylated and unmethylated) were analysed in all *Pandoraea* genomes and the results were summarised in Table 4.13. Firstly, in all chromosomal regions, the methylated motifs were highly enriched within the GR regions where 58.52 % to 69.66 % of these motifs intersected with CDSs. Similar GR methylated motif enrichment pattern was also observed with the GATC motifs (Hénaut et al., 1996). Unmethylated and hemimethylated GTWWAC motifs within the chromosomal regions on the other hand, were largely found within the IGR region (Unmethylated motifs in IGRs: 75 % to 100%; Hemimethylated motifs in IGRs: 52.94 % to 100 %). Similar IGR location bias of unmethylated sites were also reported previously (Zhu et al., 2015).

For the extrachromosomal contigs, with the exception of pPF72-1, pPO70-1 and pPO70-3, majority of the fully methylated GTWWAC motifs present were preferentially located within intragenic regions. Distribution of hemimethylated and unmethylated motifs within extrachromosomal regions were sporadic. Firstly, only pPV15 contained both hemimethylated (100% in IGR) and unmethylated (100% in GR) GTWWAC motifs. Secondly, only pPA35 harboured unmethylated motifs (100% in IGR) and only extrachromosomal contigs of *P. oxalativorans* (pPO70-1, -2 and -3) harboured hemimethylated motifs.

Table 4.13: Comparison of GTWWAC motif distribution. GTWWAC motifs (methylation status: fully methylated, hemimethylated and unmethylated) distribution in intra-genic (GR) and inter-genic (IGR) regions of all *Pandoraea* spp. genomes are compared and listed as follows.

Genome	Methylated GTWWAC			Hemimethylated GTWWAC			Unmethylated GTWWAC		
Chromosome	Total	in GR (%)	in IGR (%)	Total	in GR (%)	in IGR (%)	Total	in GR (%)	in IGR (%)
Pandoraea apista DSM-16535^T	1068	625 (58.52)	443 (41.48)	10	2 (20)	8 (80)	42	2 (4.76)	40 (95.24)
Pandoraea faecigallinarum DSM 23572^T	1154	666 (57.71)	488 (42.29)	6	0	6 (100)	24	0	24 (100)
Pandoraea norimbergensis DSM11628^T	920	477 (51.85)	443 (48.15)	20	0	20 (100)	56	2 (3.57)	54 (96.43)
Pandoraea oxalativorans DSM 23570^T	1154	664 (57.54)	490 (42.46)	28	13 (46.43)	15 (53.57)	34	0	34 (100)
Pandoraea pnomenusa DSM-165356^T	1160	747 (64.4)	413 (35.6)	22	4 (18.18)	18 (81.82)	46	5 (10.87)	41 (89.13)
Pandoraea pnomenusa RB38	1152	718 (62.33)	434 (37.67)	40	12 (30)	28 (70)	54	3 (5.56)	51 (94.44)
Pandoraea pulmonicola DSM-16583^T	1084	698 (64.39)	389 (35.89)	34	16 (47.06)	18 (52.94)	38	2 (5.26)	36 (94.74)
Pandoraea sputorum DSM-21091^T	1086	620 (57.09)	466 (42.91)	38	14 (36.84)	24 (63.16)	50	0	50 (100)
Pandoraea thiooxydans DSM-25325^T	1104	769 (69.66)	339 (30.71)	16	0	16 (100)	16	4 (25)	12 (75)
Pandoraea vervacti DSM-23571^T	1098	601 (54.74)	497 (45.26)	18	2 (11.11)	16 (88.89)	40	0	40 (100)
Plasmid									
Plasmid pPA35	24	16 (66.67)	8 (33.33)	nil	nil	nil	2	0	2 (100)
Plasmid pPF72-1	182	60 (32.97)	122 (67.03)	nil	nil	nil	nil	nil	nil
Plasmid pPF72-2	54	28 (51.85)	26 (48.15)	nil	nil	nil	nil	nil	nil
Plasmid pPO70-1	162	73 (45.06)	90 (55.56)	6	2 (33.33)	4 (66.67)	nil	nil	nil
Plasmid pPO70-2	32	20 (62.5)	12 (37.5)	6	4 (66.67)	2 (33.33)	nil	nil	nil
Plasmid pPO70-3	34	8 (23.53)	26 (76.47)	2	2 (100)	0	nil	nil	nil
Plasmid pPO70-4	16	9 (56.25)	7 (43.75)	nil	nil	nil	nil	nil	nil
Plasmid pPV15	36	20 (55.56)	16 (44.44)	2	0	2 (100)	2	2 (100)	0

4.4.1.2 Analysis of Hypermethylated Bins

All *Pandoraea* spp. genomes contained several hypermethylated genome bins ranging from the lowest number in genome of *P. pulmonicola* (2 hypermethylated bins) and the highest in genome of *P. vervacti* (8 hypermethylated bins). The average methylation frequency of all hypermethylated bins were 6.9.

In each genome, one hypermethylated bin with the highest methylation frequency was observed. Interestingly, all motifs located of these hypermethylated bins were located within an intergenic region and overlapping a cluster of DnaA boxes. When analysed further, these hypermethylated bins were also found to be located within a syntenic segment across all genomes (Figure 4.13) and were on average 20 kb upstream of the putative *oriC* regions, suggesting a functional significance in this arrangement. Similar hypermethylation pattern was reported previously in *M. pneumonia* with a suggested linkage to a role of DNA methylation in DNA replication (Lluch-Senar et al., 2013). However, in contrast to the observation of adenine motifs enrichment within the *oriC* region in *Campylobacter jejuni* clinical isolate, *Shewanella oneidensis* and *E. coli*, only 2 GTWWAC motifs (fully methylated) were found within the putative oriC sites of these

genomes (Bendall et al., 2013; Mou et al., 2015). Furthermore, SeqA protein, a DNA-binding protein which is crucial in regulating timing of *E. coli* chromosomal replication initiation *via* its interaction with hemimethylated DNA sequences in *oriC* (von Freiesleben et al., 1994; Slater et al., 1995), was not detected within the genomes of *Pandoraea* spp.. These results are suggestive that methylation of GTWWAC motif potentially has DNA replication regulatory role similar to Dam and CcrM MTases, most likely through a different mechanism.

(A)

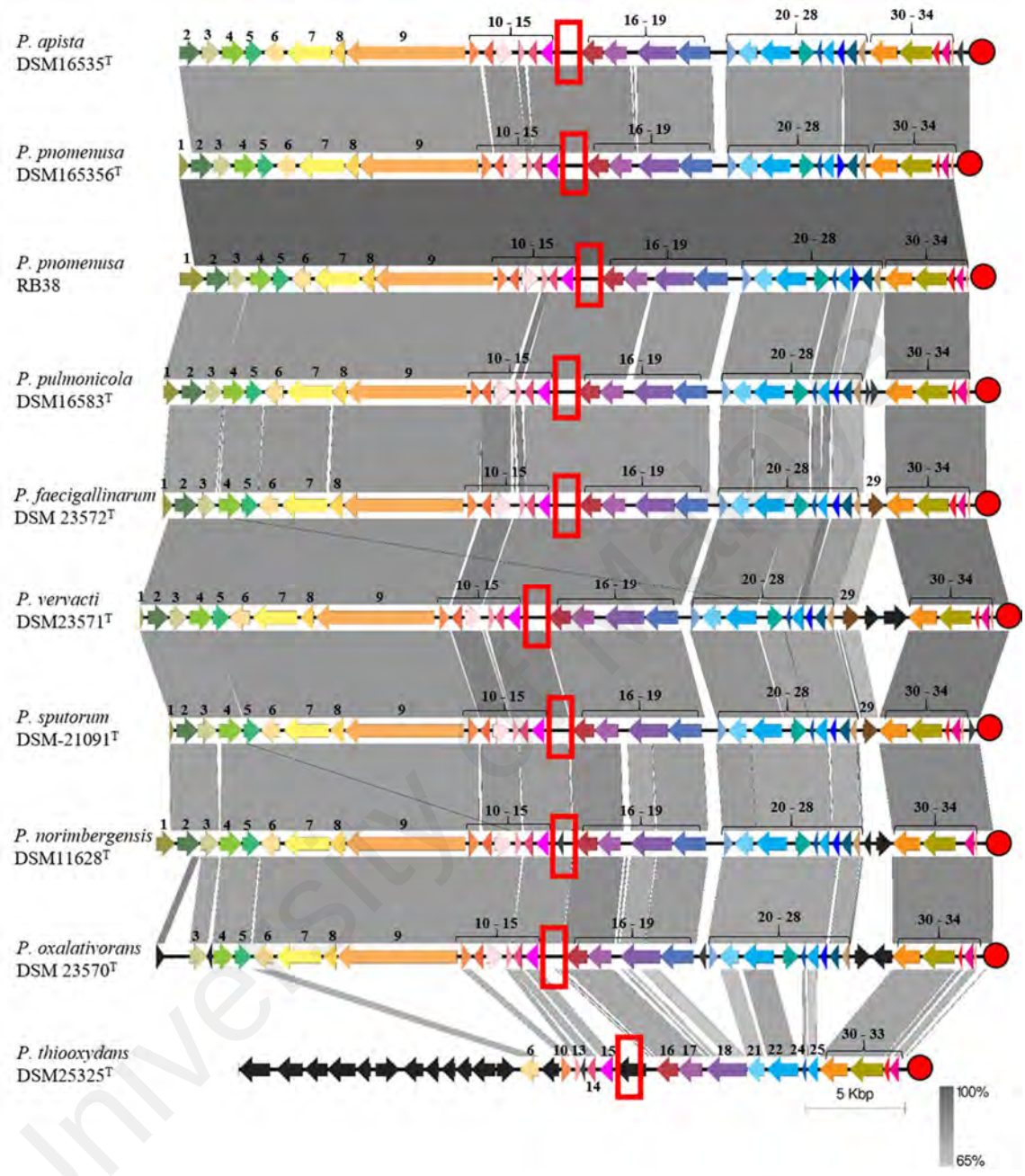


Figure 4.13: (A) BLASTN comparison of 20 kb block flanking genes of the homologous hypermethylated bin (highest methylation frequency) in all *Pandoraea* spp. genomes. Vertical blocks (coloured grey, intensity of colour reflect degree of similarity) between sequences indicate regions of shared similarity. The location of the intergenic regions where the GTWWAC motifs of the hypermethylated bins are distributed are labeled in red rectangular frames whereas the location of the putative *oriC* regions are labeled with red circles. (B) Labels of each CDS.

(B)

Number	Gene Name
1	Membrane protein
2	Hemin-degrading factor
3	Hemin ABC transporter substrate-binding protein
4	Hemin ABC transporter permease
5	Hemin importer ATP-binding subunit
6	Ornithine cyclodeaminase
7	Penicillin-binding protein 1C
8	Lysoplasmalogenase
9	Alpha-2-macroglobulin
10	Fusaric acid resistance protein FusBC
11	AsnC family transcriptional regulator
12	Phenylalanine 4-monooxygenase
13	4a-hydroxytetrahydrobiopterin dehydratase
14	SET domain-containing protein-lysine N-methyltransferase
15	Transcriptional regulator (MarR family)
16	Glutamate--cysteine ligase
17	Sodium:proton exchanger
18	Hypothetical protein
19	Chloride channel protein
20	MarR family transcriptional regulator
21	Membrane protein
22	Multidrug MFS transporter
23	MHYT domain-containing hypothetical protein
24	Hypothetical protein
25	2'-5' RNA ligase
26	Hypothetical protein
27	Glycine zipper family protein
28	Hypothetical protein
29	hypothetical protein
30	tRNA modification GTPase
31	Membrane protein insertase YidC
32	Membrane protein insertion efficiency factor
33	Ribonuclease P protein component
34	50S ribosomal protein L34

Figure 4.13, continued**.**

Furthermore, all hypermethylated bins were analysed to determine the CDSs associated with these genome bins in more than one *Pandoraea* genome (Table 4.14). A total of 9 genes were found and only indolepyruvate ferredoxin oxidoreductase (Ior) was identified in the hypermethylated bins of all *Pandoraea* genomes. Ior is an essential gene which is involved in the biosynthesis of 2-oxoacid (Tersteegen et al., 1997). On the other hand, only the hypermethylated bins of *Pandoraea vervacti* were found to contain large number of transposases which belonged to the IS5 (subfamily IS427) and IS66 families (**Appendix D**). The hypermethylated motifs were largely located in the upstream region of these transposases, suggesting a potential of GTWWAC motif involvement in the activity of these transposases.

Table 4.14: Annotated CDSs which overlap the hypermethylated genome bins of *Pandoraea* spp.

	Pandoraea spp.									
Genes associated (overlapped/flanking) hypermethylated genome bins	Papi	Pfae	Pnor	Poxa	Ppul	Pspu	Pthio	Pver	Ppno	RB38
Indolepyruvate ferredoxin oxidoreductase (alpha and beta subunits)	+	+	+	+	+	+	+	+	+	+
Hypothetical protein (213 bp-restricted to *Pandoraea* genomes)	+	+	-	+	+	+	+	+	+	+
Hypothetical protein	+	+	+	-	+	+	-	+	-	+
Glutamate Aspartate periplasmic binding protein precursor GltI (TC 3.A.1.3.4)	-	+	+	+	-	+	-	+	-	-
Potassium-transporting ATPase A chain	-	-	+	+	-	+	-	+	-	-
Hypothetical protein	-	-	+	+	-	+	-	-	-	-
Outer membrane protein Imp (required for envelope biogenesis	-	-	+	-	-	+	-	-	-	-
ABC transporter ATP binding protein	-	-	-	-	-	-	-	-	+	+
Glycine oxidase ThiO (EC 1.4.3.19)	-	-	-	-	-	-	-	-	+	+

Denotations of abbreviation: Papi: *Pandoraea apista* DSM-16535^T; Pfae: *Pandoraea faecigallinarum* DSM 23572^T; Pnor: *Pandoraea norimbergensis* DSM 11628^T; Poxa: *Pandoraea oxalativorans* DSM 23570^T; Ppul: *Pandoraea pulmonicola* DSM-16583^T; Pspu: *Pandoraea sputorum* DSM-21091^T; Pthio: *Pandoraea thiooxydans* DSM-25325^T; Pver: *Pandoraea vervacti* DSM-23571^T; Ppno: *Pandoraea pnomenusa* DSM-165356^T; RB38: *Pandoraea pnomenusa* RB38. "+": gene identified within hypermethylated genome bins.

4.4.1.3 Pan-genus Methylation Hotspot Analysis

The number of methylation hotspot bins in the *Pandoraea* spp. genome ranged from 41 (*P. faecigallinarum* and *P. pnomenusa* RB38) to 24 (*P. norimbergensis*) with the average number of hotspot bins at 35 (SD = 5.12, n = 10). From the analysis, a total of 71 hotspot genes were found to be shared between different *Pandoraea* genomes (Table 4.15). The top 4 genes associated with pan-genus hotspots (shared by more than 5 genomes) were genes with essential roles in bacterial physiology owing to their role in translation, DNA recombination and repair, prokaryotic metabolism pathway and amino acid metabolism, namely: *S*-adenosylmethionine:tRNA ribosyltransferase-isomerase (QueA), DNA helicase (RecG), acetoacetyl-CoA synthase, and branched-chain amino acid aminotransferace (IlvE).

Various hotspot genes were associated with virulence and antibiotic resistance properties in various pathogens. The first example was UDP-4-amino-4-deoxy-L-arabinose--oxoglutarate aminotransferase (*arnB*) and polymyxin resistance protein (*arnC*), two genes that are linked to the *arn* operon and encode proteins involved in modification of the L-Ara4N lipid A pathway (Olaitan et al., 2014). These genes are responsible for the intrinsic resistance of various microorganisms to polymyxin which are

currently the last resort antibiotics, including the close phylogenetic neighbour of *P. pnomenusa* namely *Burkholderia cepacia* complex, (Loutet & Valvano, 2011). Secondly, aminoglycoside phosphotransferase (AphA) which contributes to drug resistance *via* catalysis of ATP-dependent phosphorylation of aminoglycosides (Llano-Sotelo et al., 2002). Lastly, LPS assembly outer membrane protein (LptD) which is involved in antibiotic resistance *via* manipulation of outer membrane permeability and YaeQ protein, which enhances transcription of bacterial virulence factors (Sampson et al., 1989; Wong et al., 1998). These proteins are largely distributed in the hotspot bins of *P. pnomenusa*, *P. vervacti*, *P. sputorum*, and *P. oxalativorans*.

Table 4.15: Pan-genus hotspot bins genes.

	Pandoraea spp.									
Flanking genes	Papi	Pfae	Pnor	Poxa	Ppul	Pspu	Pthio	Pver	Ppno	RB38
S-adenosylmethionine:tRNA ribosyltransferase-isomerase (EC 5.-.-.-)	+	+	+	+	+	+	+	+	+	+
ATP-dependent DNA helicase RecG (EC 3.6.1.-)	NA	+	+	+	+	+	+	+	+	+
Acetoacetyl-CoA synthetase (EC 6.2.1.16) / Long-chain-fatty-acid--CoA ligase (EC 6.2.1.3)	+	NA	+	+	NA	NA	+	+	+	+
Branched-chain amino acid aminotransferase (EC 2.6.1.42)	+	NA	+	+	NA	NA	NA	+	+	+
Predicted regulator PutR for Proline utilization GntR family	+	+	NA	+	+	NA	NA	NA	+	+
COG3178: Predicted phosphotransferase related to Ser/Thr protein kinases	NA	NA	+	+	NA	+	NA	NA	+	+
Enoyl-CoA hydratase (EC 4.2.1.17)	NA	+	NA	+	NA	+	NA	NA	+	+
LPS assembly protein LptD	NA	NA	+	+	NA	+	NA	NA	+	+
Polymyxin resistance protein ArnC glycosyl transferase (EC 2.4.-.-)	NA	NA	NA	+	NA	+	NA	+	+	+
UDP-4-amino-4-deoxy-L-arabinose--oxoglutarate aminotransferase (EC 2.6.1.-)	NA	NA	NA	+	NA	+	NA	+	+	+
Hypothetical protein	NA	+	NA	+	NA	NA	NA	NA	+	+
FIG00453175: hypothetical protein	+	+	+	+	NA	+	NA	NA	NA	NA
Transcriptional regulator associated with Tricarboxylic transport	+	+	NA	+	NA	+	NA	NA	NA	+
Cyanophycin synthase (EC 6.3.2.29)(EC 6.3.2.30)	+	+	NA	NA	NA	+	NA	+	NA	NA
ABC transporter, transmembrane region	NA	+	NA	NA	NA	+	+	+	NA	NA
Phenylacetic acid degradation protein PaaD, thioesterase	NA	+	+	+	NA	+	NA	NA	NA	NA
Hypothetical protein	NA	+	NA	+	NA	+	NA	NA	NA	+
Outer membrane protein A precursor	NA	NA	+	NA	+	NA	NA	+	+	NA
Transcriptional regulator, LacI family	NA	+	NA	NA	+	NA	NA	+	+	NA
Isovaleryl-CoA dehydrogenase (EC 1.3.8.4)	+	+	NA	+	NA	NA	NA	NA	NA	NA
Glutamate Aspartate periplasmic binding protein precursor GltI (TC 3.A.1.3.4)	+	NA	NA	NA	NA	NA	NA	NA	+	+
3-polyprenyl-4-hydroxybenzoate carboxy-lyase (EC 4.1.1.-)	NA	+	NA	NA	+	NA	NA	+	NA	NA
Transcriptional regulator, LacI family	NA	+	NA	NA	+	NA	NA	NA	+	NA
Two-component system response regulator OmpR	NA	+	NA	+	NA	+	NA	NA	NA	NA
Probable glutathione S-transferase (EC 2.5.1.18), YfcF homolog	NA	+	NA	NA	NA	+	NA	NA	NA	+
DNA gyrase subunit A (EC 5.99.1.3)	NA	NA	+	NA	NA	NA	NA	+	+	NA
Predicted transcriptional regulator LiuR of leucine degradation pathway, MerR family	NA	NA	+	+	NA	+	NA	NA	NA	NA
ATP synthase epsilon chain (EC 3.6.3.14)	NA	NA	+	NA	NA	NA	NA	+	NA	+
Outer membrane protein (porin)	NA	NA	NA	+	NA	+	NA	NA	NA	+
4-hydroxybenzoate transporter	NA	NA	NA	+	NA	+	NA	NA	+	NA
Soluble lytic murein transglycosylase and related regulatory proteins	NA	NA	NA	NA	NA	+	NA	+	+	NA
2-hydroxy-3-oxopropionate reductase (EC 1.1.1.60)	NA	NA	NA	NA	NA	NA	NA	+	+	+
Hypothetical protein	NA	NA	NA	NA	+	NA	NA	+	+	NA
YaeQ protein	NA	NA	NA	NA	NA	+	NA	NA	+	+
4-hydroxybenzoyl-CoA thioesterase family active site	NA	NA	NA	NA	NA	NA	NA	+	+	+
Probable glutathione S-transferase (EC 2.5.1.18) YfcF homolog	NA	+	NA	NA	NA	+	NA	NA	NA	+
2-dehydropantoate 2-reductase (EC 1.1.1.169)	NA	+ (2)	NA	NA	NA	NA	NA	+	NA	NA
tRNA dihydrouridine synthase A	+	+	NA	NA	NA	NA	NA	NA	NA	NA
Prolipoprotein diacylglyceryl transferase (EC 2.4.99.-)	+	NA	NA	+	NA	NA	NA	NA	NA	NA
RhtB family transporter	+	NA	NA	+	NA	NA	NA	NA	NA	NA
OsmC/Ohr family protein	+	NA	+	NA	NA	NA	NA	NA	NA	NA
Seryl-tRNA synthetase (EC 6.1.1.11)	NA	+	NA	NA	NA	NA	NA	+	NA	NA
Fatty acid desaturase family protein	NA	+	NA	+	NA	NA	NA	NA	NA	NA
TRAP transporter solute receptor, unknown substrate 4	NA	+	NA	NA	NA	NA	NA	NA	+	NA
Bll3346 protein	NA	+	NA	NA	NA	NA	NA	+	NA	NA
2-dehydropantoate 2-reductase (EC 1.1.1.169)	NA	+	NA	NA	NA	NA	NA	+	NA	NA
N-formylglutamate deformylase (EC 3.5.1.68)	NA	+	NA	NA	NA	NA	NA	NA	NA	+
Lactate-responsive regulator LldR in Enterobacteria, GntR family	NA	+	NA	NA	NA	NA	NA	NA	NA	+
Methyl-accepting chemotaxis protein I (serine chemoreceptor protein)	NA	NA	+	NA	NA	+	NA	NA	NA	NA
LSU ribosomal protein L13p (L13Ae)	NA	NA	+	NA	+	NA	NA	NA	NA	NA
Deoxyribodipyrimidine photolyase (EC 4.1.99.3)	NA	NA	NA	+	NA	+	NA	NA	NA	NA
Aldehyde dehydrogenase (EC 1.2.1.3)	NA	NA	NA	+	NA	NA	NA	NA	+	NA
UDP-N-acetylglucosamine 4,6-dehydratase (EC 4.2.1.-)	NA	NA	NA	+	NA	NA	NA	NA	+	NA
Transcriptional regulator, GntR family	NA	NA	NA	+	+	NA	NA	NA	NA	NA
Potassium-transporting ATPase A chain (EC 3.6.3.12) (TC 3.A.3.7.1)	NA	NA	NA	NA	+	NA	NA	NA	+	NA
D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95)	NA	NA	NA	NA	NA	+	NA	NA	+	NA
FIG00348850: hypothetical protein	NA	+	NA	NA	NA	NA	NA	NA	NA	+
Transcriptional regulator GntR family	NA	NA	NA	+	+	NA	NA	NA	NA	NA
Predicted nucleotidyltransferase	NA	NA	NA	NA	+	NA	NA	+	NA	NA
Cytochrome c-type biogenesis protein ResA	NA	NA	NA	NA	+	NA	NA	+	NA	NA
2-hydroxy-3-oxopropionate reductase (EC 1.1.1.60)	NA	NA	NA	NA	NA	NA	NA	+	NA	+
Chromosomal replication initiator protein DnaA	NA	NA	NA	NA	NA	NA	NA	NA	+	+
FOG: GGDEF domain	NA	NA	NA	NA	NA	NA	NA	NA	+	+
DNA polymerase III beta subunit (EC 2.7.7.7)	NA	NA	NA	NA	NA	NA	NA	NA	+	+
Nicotinate-nucleotide adenylyltransferase (EC 2.7.7.18)	NA	NA	NA	NA	NA	NA	NA	NA	+	+
Electron transfer flavoprotein, beta subunit	NA	NA	NA	NA	NA	NA	NA	NA	+	+
ortholog of Bordetella pertussis (BX470248) BP2750	NA	NA	NA	NA	NA	NA	NA	NA	+	+
Ribosomal silencing factor RsfA (former Iojap)	NA	NA	NA	NA	NA	NA	NA	NA	+	+
Universal stress protein family	NA	NA	NA	NA	NA	NA	NA	NA	+	+
Hypothetical protein	NA	NA	NA	NA	NA	NA	NA	NA	+	+
Lysophospholipase (EC 3.1.1.5) Monoglyceride lipase (EC 3.1.1.23) putative	NA	NA	NA	NA	NA	NA	NA	NA	+	+
Electron transfer flavoprotein beta subunit	NA	NA	NA	NA	NA	NA	NA	NA	+	+

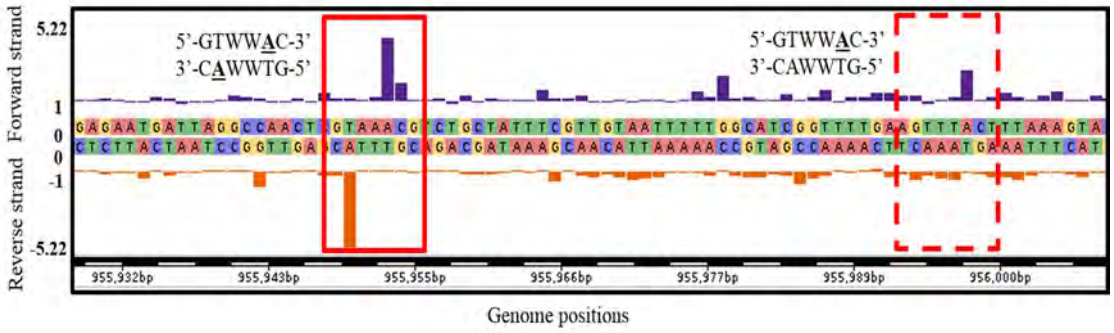
Denotations of abbreviation: Papi: *Pandoraea apista* DSM-16535^T; Pfae: *Pandoraea faecigallinarum* DSM 23572^T; Pnor: *Pandoraea norimbergensis* DSM 11628^T; Poxa: *Pandoraea oxalativorans* DSM 23570^T; Ppul: *Pandoraea pulmonicola* DSM-16583^T; Pspu: *Pandoraea sputorum* DSM-21091^T; Pthio: *Pandoraea thiooxydans* DSM-25325^T; Pver: *Pandoraea vervacti* DSM-23571^T; Ppno: *Pandoraea pnomenusa* DSM-165356^T; RB38: *Pandoraea pnomenusa* RB38. NA: not applicable; "+": gene identified within hotspot bins.

4.4.1.4 Analysis of Unmethylated Sites

Unmethylated sites, for the purpose of this analysis, were defined as adenine bases of the GTWWAC motifs that were kept in an unmethylated state and resulted in a hemimethylated or unmethylated motifs as depicted in Figure 4.14. Hemimethylated motifs could be found on a newly synthesized daughter DNA molecules as a result of semi-conservative replication of a fully methylated template DNA molecule whereas the most likely cause of an unmethylated motif is due to interacting protein occupancy which prevents the methylating action of MTase on the motif (Low et al., 2001). The presence of unmethylated or hemimethylated DNA MTase recognition motifs within intergenic region often indicates a role of DNA methylation-related transcriptional regulation of downstream genes (Casadesús & Low, 2006).

The number of unmethylated sites in all *Pandoraea* genomes ranged from 94 (*P. pnomenusa* RB38) to 30 (*P. faecigallinarum*) with an average number of 63 (SD = 21.23, n = 10), demonstrating the high degree of variation in numbers of unmethylated sites between different species. Furthermore, in all genomes, the number of unmethylated motifs were higher than the hemimethylated motifs, with the exception of *P. thiooxydans* which had equal number of both.

(A)



(B)

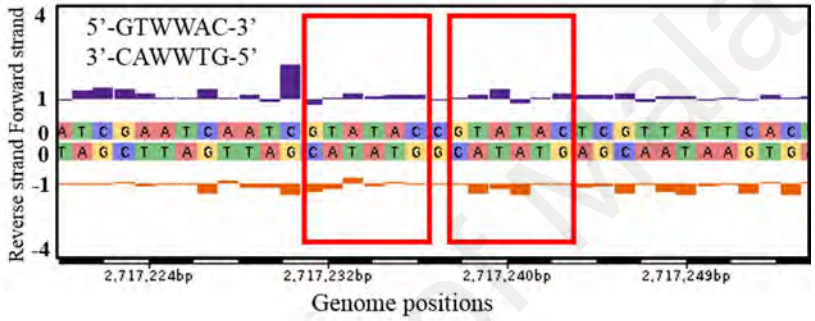
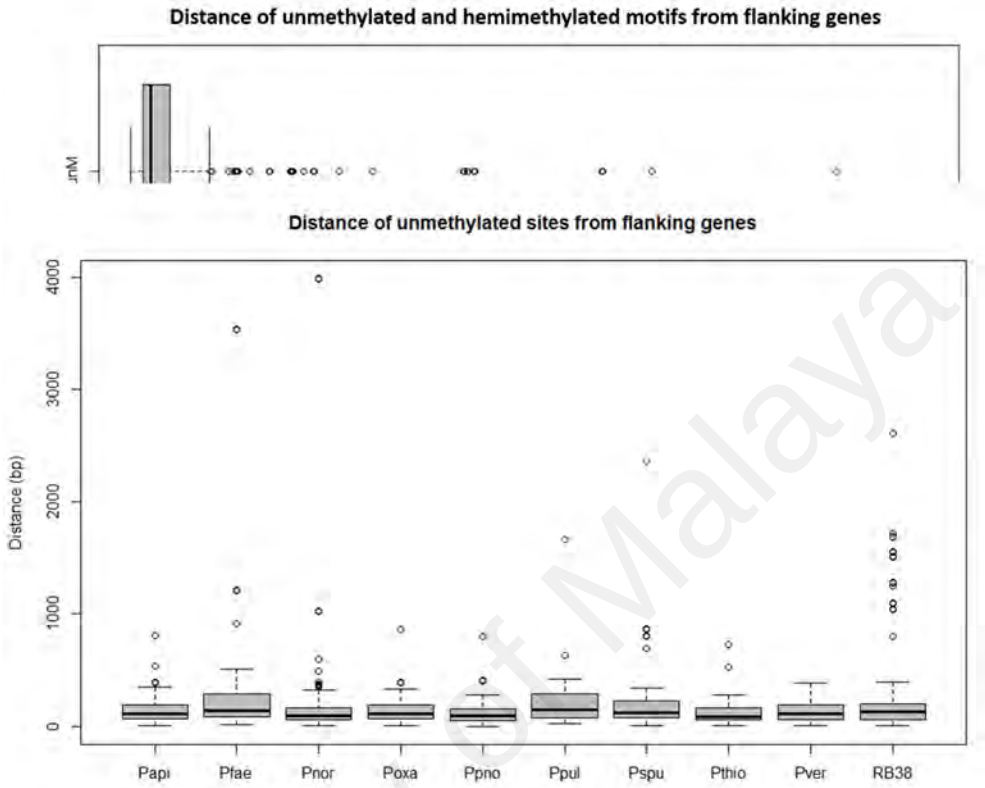


Figure 4.14: Example of IPD ratio plots of GTWWAC motifs. (A) Fully methylated and hemimethylated GTWWAC motifs. The left box shows a fully methylated palindromic GTWWAC/CAWWTG motif adjacent to a hemimethylated GTWWAC motif (right box with dotted lines) (B) Unmethylated GTWWAC motifs.

The distance between the unmethylated sites and the adjacent flanking genes in all genomes were subsequently analysed. Firstly, both unmethylated and hemimethylated motifs were observed to have a largely similar range of distance length with their flanking genes (Figure 4.15). When the distance were analysed with individual *Pandoraea* genomes as the grouping variable, the unmethylated sites were observed to be located at a largely consistent distance from the adjacent flanking genes where the median value ranged from 81.5 to 143, a distance which potentially represent an enrichment of these sites at gene regulatory regions. *P. faecigallinarum*, *P. pulmonicola* and *P. sputorum* had a relatively high variation in the aspect of distance length distributions where the distance

data above the median value demonstrated a higher variation, possibly indicating that this is not a consistent occurrence.

(A)



Denotation of abbreviation: unM: unmethylated motifs; hemiM: hemimethylated motifs.

(B)

Denotations of abbreviation: Papi: *Pandoraea apista* DSM-16535^T; Pfae: *Pandoraea faecigallinarum* DSM 23572^T; Pnor: *Pandoraea norimbergensis* DSM 11628^T; Poxa: *Pandoraea oxalativorans* DSM 23570^T; Ppul: *Pandoraea pulmonicola* DSM-16583^T; Pspu: *Pandoraea sputorum* DSM-21091^T; Pthio: *Pandoraea thiooxydans* DSM-25325^T; Pver: *Pandoraea vervacti* DSM-23571^T; Ppno: *Pandoraea pnomenusa* DSM-165356^T; RB38: *Pandoraea pnomenusa* RB38.

Figure 4.15: Boxplots showing distance of unmethylated sites from flanking genes with the grouping variable of (A) motif types and (B) *Pandoraea* species.

A recent study demonstrated that analysis of unmethylated sites, particularly those that present in clusters and are conserved across multiple genomes, represent a useful approach for *in silico* prediction of potential methylation regulatory sites (Blow et al., 2016). Therefore, firstly, the homologous unmethylated motifs which were present in adjoining locations were analysed and the flanking genes were identified (Table 4.16). No unmethylated sites clusters were observed in the genomes of *P. oxalativorans*, and *P.*

thiooxydans. A total of 11 genes were found to be shared by not more than 4 *Pandoraea* genomes. Interestingly, the pattern of motifs located within these homologous sites were similar across the genomes, whether they are clusters of completely unmethylated motifs or adjoining unmethylated and hemimethylated motifs. Transcriptional regulators constituted 45 % (13/29) of the genes associated with these unmethylated sites clusters. Interestingly, among the unmethylated sites clusters-associated genes, ferredoxin, cysteine desulferase, and polyketide cyclase matched the gene types (functional association with iron-sulfur cluster proteins and polyketide synthesis) which were identified to be conserved among the unmethylated sites of *Mycobacterium tuberculosis* complex (MTBC) strains (Zhu et al., 2015). In addition, ferredoxin was also identified within major hypomethylated areas (genome bins with four or less methylation sites) in *Campylobacter jejuni* strains (Mou et al., 2015). Furthermore, two genes associated with phenylacetic acid degradation were also identified among the flanking genes.

Table 4.16: Flanking genes associated with unmethylated sites clusters.

	Pandoraea spp.									
Flanking genes	Papi	Pfae	Pnor	Poxa	Ppul	Pspu	Pthio	Pver	Ppno	RB38
* Ferredoxin	NA	NA	-/-(2)	NA	-/-(2)	NA	NA	NA	-/-(2)	-/-(2)
* Ornithine cyclodeaminase	NA	NA	NA	NA	-/-(2)	NA	NA	NA	-/-(2)	-/-(2)
* Cysteine desulfurase , IscS subfamily	-/-(2)	NA	NA	NA	NA	NA	NA	NA	-/-(2)	-/-(2)
* Nitrilotriacetate monooxygenase component B	-/-(2)	NA	NA	NA	NA	NA	NA	NA	-/-(2)	-/-(2)
Transcriptional regulator, RpiR family	NA	NA	NA	NA	NA	NA	NA	NA	-/-(2)	-/-(2)
* Enoyl-CoA hydratase	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA
* NAD-dependent formate dehydrogenase gamma subunit	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA
GntR family transcriptional regulator	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA
* GntR family transcriptional regulator	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA
XRE family transcriptional regulator	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA	NA	-/-(1), +/-(1)	NA
* Phenylacetic acid degradation protein PaaI	NA	NA	NA	NA	-/- (2)	NA	NA	NA	-/-(1), +/-(1)	NA
* XRE family transcriptional regulator	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA	NA	NA	NA
* GNAT family N-acetyltransferase	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA	NA	NA	NA
* GntR family transcriptional regulator	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA	NA	NA	NA
* Glyoxalase/bleomycin resistance/dioxygenase family protein	NA	-/-(1), +/-(1)	NA	NA	NA	NA	NA	NA	NA	NA
Polyketide cyclase	NA	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA
Transcriptional regulator, LacI family	NA	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA
* Pyrroloquinoline quinone biosynthesis protein PqqC (putative)	NA	NA	NA	NA	NA	-/-(2)	NA	NA	NA	NA
Transcriptional regulator, RpiR family	NA	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA
* DNA-binding transcriptional regulator, LacI/PurR family	NA	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA
* MurR/RpiR family transcriptional regulator	NA	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA
* TetR family transcriptional regulator	NA	NA	NA	NA	NA	NA	NA	-/-(2)	NA	NA
Bll6423 protein	NA	NA	NA	NA	NA	NA	NA	-/-(2)	NA	NA
Phenylacetic acid degradation protein PaaD, thioesterase	NA	NA	NA	NA	NA	NA	NA	-/-(2)	NA	NA
* Hypothetical protein	NA	NA	NA	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA
* Hypothetical protein	NA	NA	NA	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA
Hypothetical protein	NA	NA	NA	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA
* MerR family transcriptional regulator	NA	NA	NA	NA	NA	NA	NA	NA	-/-(1), +/-(1)	NA
* Benzoate anaerobic degradation transcriptional regulator BadR, MarR family	NA	NA	NA	NA	NA	NA	NA	NA	-/-(1), +/-(1)	NA

Denotations of abbreviation: Papi: *Pandoraea apista* DSM-16535^T; Pfae: *Pandoraea faecigallinarum* DSM 23572^T; Pnor: *Pandoraea norimbergensis* DSM 11628^T; Poxa: *Pandoraea oxalativorans* DSM 23570^T; Ppul: *Pandoraea pulmonicola* DSM-16583^T; Pspu: *Pandoraea sputorum* DSM-21091^T; Pthio: *Pandoraea thiooxydans* DSM-25325^T; Pver: *Pandoraea vervacti* DSM-23571^T; Ppno: *Pandoraea pnomenusa* DSM-165356^T; RB38: *Pandoraea pnomenusa* RB38. "-/-" denotes unmethylated motifs wheras "+/-" denotes hemimethylated motifs. The number of the respective motif type present within the cluster is indicated in parenthesis. "*" sign associated with gene names indicates genes which are identified as flanking genes of singly unmethylated sites.

Subsequently, CDSs associated with singly homologous unmethylated sites were also studied (Table 4.17). A total of 81 genes were identified to flank the singly unmethylated sites of more than one *Pandoraea* genome of which 36 of them were reproducibly identified in more than 5 *Pandoraea* genomes. 66% (19/29) of the unmethylated sites clusters genes were also identified to flank singly occurred unmethylated sites in other genomes, indicating that those regions are conserved to be unmethylated. Overall, several significant functional categories of genes can be observed: membrane transporters (associated with antibiotic resistance activity), cell wall components, proteins with roles in DNA replication and transcription initiation process, osmoregulation and various microbial metabolism activities. Among these genes, besides the transcriptional regulators (constitutes 30% (24/81) of all genes) which were previously identified to contain substantial enrichment of conserved unmethylated sites upstream in other organisms, a TonB-dependent receptor which was identified to be located in a conserved downstream location from clusters of unmethylated Dam sites was also identified to be downstream of GTWWAC unmethylated sites (Blow et al., 2016).

Table 4.17: Flanking genes associated with singly unmethylated sites.

	Pandoraea spp.								
Flanking genes	Papi	Pfae	Pnor	Poxa	Ppul	Pspu	Pthio	Pver	Ppno	RB38
* NAD-dependent formate dehydrogenase gamma subunit	-/-	-/-	-/-(1), +/-(1)	-/-	+/-	-/-	NA	-/-(1), +/-(1)	-/-	-/-
* Enoyl-CoA hydratase	NA	-/-	-/-(1), +/-(1)	-/-	-/-	-/-	NA	-/-(1), +/-(1)	-/-	-/-
* GntR family transcriptional regulator	NA	-/-	-/-(1), +/-(1)	-/-	-/-	-/-	NA	-/-(1), +/-(1)	-/-	-/-
* Glyoxalase/bleomycin resistance/dioxygenase family protein	-/-	-/-(1), +/-(1)	-/-	-/-	NA	-/-	NA	-/-	-/-	-/-
Outer membrane protein (porin)	-/-	-/-	-/-	-/-	NA	-/-	NA	-/-	-/-	-/-
Ureidoglycolate lyase/ 5-oxopent-3-ene-1,2,5-tricarboxylate decarboxylase	-/-	-/-	-/-	-/-	NA	-/-	NA	-/-	-/-	-/-
D-3-phosphoglycerate dehydrogenase	-/-	-/-	-/-	-/-	NA	-/-	NA	-/-	-/-	-/-
Acetaldehyde dehydrogenase (acetylating)	-/-	-/-	-/-	-/-	NA	-/-	NA	-/-	-/-	-/-
GntR family transcriptional regulator	-/-	-/-	-/-	-/-	-/-	-/-	NA	NA	-/-	-/-
GntR family transcriptional regulator	-/-	+/-	-/-	-/-	NA	-/-	NA	-/-	-/-	-/-
Predicted arabinose efflux permease, MFS family	-/-	+/-	-/-		NA	-/-	NA	-/-	-/-	-/-
ABC transporter substrate-binding protein	-/-	-/-	NA	-/-	NA	-/-	NA	-/-	-/-	-/-
Hypothetical protein	-/-	+/-	-/-		NA	-/-	NA	-/-	+/-	+/-
GntR family transcriptional regulator	+/-	-/-	NA	-/-	NA	-/-	NA	-/-	-/-	+/-
Glucan biosynthesis protein G	NA	NA	+/-	+/-	+/-	+/-	NA	+/-	+/-	+/-
Hypothetical protein	-/-	-/-	-/-	-/-	NA	-/-	NA	-/-	-/-	NA
Glutamyl-tRNA(Gln) amidotransferase subunit A; gatA	-/-	NA	-/-	NA	+/-	+/-	NA	NA	-/-	-/-
Sodium:sulfate symporter	+/-	NA	NA	-/-	NA	-/-	NA	-/-	-/-	+/-
Anaerobic benzoate catabolism transcriptional regulator	+/-	NA	+/-	+/-	NA	NA	NA	+/-	+/-	-/-
Enoyl-CoA hydratase (2)	+/-	NA	+/-	+/-	NA	NA	NA	+/-	+/-	-/-
Lipid A palmitoyltransferase PagP precursor (pagP)	-/-	-/-	-/-	NA	NA	-/-	NA	NA	-/-	-/-
Transcriptional regulator, PadR family	-/-		-/-	NA	NA	-/-	NA	-/-	-/-	-/-
GntR family transcriptional regulator	-/-	-/-	-/-	-/-	NA	-/-	-/-	NA	NA	NA
Taurine catabolism dioxygenase TauD, TfdA family	NA	-/-	-/-	-/-	-/-	-/-	NA	-/-	NA	NA
Benzoate MFS transporter BenK	NA	NA	+/-	NA	+/-	NA	NA	+/-	+/-	+/-
* Phenylacetic acid degradation protein PaaI	NA	NA	+/-	NA	-/-(2)	NA	NA	+/-	-/-(1), +/-(1)	+/-
* XRE family transcriptional regulator	NA	NA	-/-(1), +/-(1)	+/-	NA	NA	NA	-/-	-/-	-/-
* GNAT family N-acetyltransferase	NA	NA	-/-(1), +/-(1)	+/-	NA	NA	NA	-/-	-/-	-/-
Putative sialic acid transporter, MFS superfamily	-/-	NA	-/-	NA	NA	-/-	NA	NA	-/-	-/-
Alpha/beta hydrolase family protein	-/-	NA	-/-	NA	-/-	NA	NA	NA	-/-	-/-
* Pyrroloquinoline quinone biosynthesis protein PqqC (putative)	-/-	NA	-/-	NA	NA	-/-(2)	NA	NA	-/-	-/-
TetR family transcriptional regulator	-/-	NA	-/-	NA	-/-	NA	NA	NA	-/-	-/-
* Ferredoxin	-/-	NA	-/-(2)	NA	-/-(2)	NA	NA	NA	-/-(2)	-/-(2)
* GntR family transcriptional regulator	NA	-/-	-/-(1), +/-(1)	-/-		-/-	NA	-/-	NA	NA
Two component smolarity sensor histidine kinase EnvZ	-/-	+/-	NA	+/-	-/-	-/-	NA	NA	NA	NA
Alkyl hydroperoxide reductase subunit AhpC (peroxiredoxin)	-/-	+/-	NA	+/-	-/-	-/-	NA	NA	NA	NA
* MerR family transcriptional regulator	NA	NA	+/-	NA	NA	NA	NA	+/-	-/-(1), +/-(1)	+/-
* Cysteine desulfurase (EC 2.8.1.7), IscS subfamily	-/-(2)	NA	NA	NA	-/-	NA	NA	NA	-/-(2), +/-(1)	-/-(2)
* Nitrilotriacetate monooxygenase component B	-/-(2)	NA	NA	NA	-/-	NA	NA	NA	-/-(2), +/-(1)	-/-(2)
* DNA-binding transcriptional regulator, LacI/PurR family	NA	NA	-/-	-/-	NA	-/-(1), +/-(1)	NA	NA	NA	+/-
GntR family transcriptional regulator	+/-	NA	-/-	+/-	NA	NA	NA	NA	NA	+/-
* Hypothetical protein	NA	NA	-/-	NA	-/-	+/-	NA	-/-(1), +/-(1)	NA	NA
RNA polymerase subunit sigma-70	NA	NA	NA	-/-	-/-	+/-	NA	-/-	NA	NA
* Benzoate anaerobic degradation transcriptional regulator BadR, MarR family	-/-	NA	NA	NA	NA	NA	NA	NA	-/-(1), +/-(1)	-/-
3-ketoacyl-(acyl-carrier-protein) reductase	-/-	NA	NA	NA	NA	NA	NA	NA	-/-	-/-
Primosomal assembly protein PriA	NA	NA	NA	NA	NA	-/-	NA	NA	+/-	-/-
Sulfite:cytochrome c oxidoreductase subunit A	NA	NA	NA	NA	NA	-/-	NA	NA	+/-	-/-
* Ornithine cyclodeaminase (EC 4.3.1.12)	NA	NA	NA	NA	-/-(2)	NA	NA	NA	-/-(2)	-/-(2)
Hypothetical protein	+/-	NA	NA	+/-	NA	NA	NA	NA	NA	+/-
IS110 family transposase	NA	NA	NA	NA	-/-	NA	NA	-/-	NA	+/-
* Hypothetical protein	NA	NA	-/-	NA	NA	+/-	NA	-/-(1), +/-(1)	NA	NA
TonB-dependent siderophore receptor	NA	NA	-/-	-/-	NA	NA	NA	-/-	NA	NA
Cobalamin biosynthesis protein CobS	NA	NA	+/-	+/-	NA	+/-	NA	NA	NA	NA
IclR family transcriptional regulator	NA	NA	+/-	+/-	NA	+/-	NA	NA	NA	NA
FAD-binding monooxygenase	NA	NA	NA	-/-	NA	-/-	NA	-/-	NA	NA
D-amino acid dehydrogenase small subunit (EC 1.4.99.1)/Glycine/D-amino acid oxidase (deaminating)	NA	NA	NA	NA	NA	NA	NA	NA	-/-	-/-
Transcriptional regulator, XRE family with cupin sensor	NA	NA	NA	NA	NA	NA	NA	NA	-/-	-/-
MFS transporter (arabinose ABC transporter permease?)	+/-	NA	NA	NA	NA	NA	NA	NA	NA	+/-
Cytochrome O ubiquinol oxidase subunit II (EC 1.10.3.-)	+/-	NA	NA	NA	NA	NA	NA	NA	NA	+/-
HPr kinase/phosphorylase (EC 2.7.1.-)	NA	NA	NA	NA	NA	+/-	NA	NA	+/-	NA
Phosphate starvation-inducible protein PsiF	NA	NA	NA	NA	NA	+/-	NA	NA	+/-	NA
* MurR/RpiR family transcriptional regulator	-/-	NA	NA	NA	NA	-/-(1), +/-(1)	NA	NA	NA	NA
CDP-6-deoxy-delta-3,4-glucoseen reductase	-/-	NA	-/-	NA	NA		NA	NA	NA	NA
LacI family transcriptional regulator	NA	-/-	NA	-/-	NA		NA	NA	NA	NA
Outer membrane protein (porin)	NA	NA	-/-	NA	NA	-/-	NA	NA	NA	NA
C4-dicarboxylate ABC transporter	NA	NA	+/-	NA	NA		NA	+/-	NA	NA
Hypothetical protein	NA	NA	-/-	NA	NA	-/-	NA	NA	NA	NA
Transcriptional regulator, GntR family	NA	NA	-/-	NA	NA	-/-	NA	NA	NA	NA
AsnC family transcriptional regulator	NA	NA	+/-	NA	NA		NA	+/-	NA	NA
Transcriptional regulator, XRE family with cupin sensor	NA	NA	-/-	NA	NA	-/-	NA	NA	NA	NA
Hypothetical protein	NA	NA	-/-	NA	NA	-/-	NA	NA	NA	NA
MarR family transcriptional regulator	NA	NA	-/-	NA	NA	-/-	NA	NA	NA	NA
ArsR/Lrp family transcriptional regulator	NA	NA	-/-	-/-	NA		NA	NA	NA	NA
RidA/YER057c/UK114 family protein (hypothetical protein)	NA	NA	-/-	NA	NA	-/-	NA	NA	NA	NA
Excinuclease ABC subunit A, dimeric form; UvrA	NA	NA	NA	-/-	NA	+/-	NA	NA	NA	NA
N5,N10-methylene tetrahydromethanopterin reductase	NA	-/-	NA	-/-	NA		NA	NA	NA	NA
4'-phosphopantetheinyl transferase (EC 2.7.8.-)	NA	NA	NA	-/-	NA	+/-	NA	NA	NA	NA
* TetR family transcriptional regulator	NA	NA	NA	-/-	NA	-/-(2)	NA	NA	NA	NA
Glutamate--cysteine ligase	NA	NA	NA	NA	NA	+/-	NA	NA	NA	NA
Hypothetical protein	NA	NA	NA	-/-	NA	+/-	NA	NA	NA	NA
2,4-dienoyl-CoA reductase [NADPH](EC 1.3.1.34)	NA	NA	NA	NA	NA		-/-	-/-	NA	NA

Denotations of abbreviation: Papi: *Pandoraea apista* DSM-16535; Pfae: *Pandoraea faecigallinarum* DSM 23572^{T}; Pnor: *Pandoraea norimbergensis* DSM 11628^{T}; Poxa: *Pandoraea oxalativorans* DSM 23570^{T}; Ppul: *Pandoraea pulmonicola* DSM-16583^{T}; Pspu: *Pandoraea sputorum* DSM-21091^{T}; Pthio: *Pandoraea thiooxydans* DSM-25325^{T}; Pver: *Pandoraea vervacti* DSM-23571^{T}; Ppno: *Pandoraea pnomenusa* DSM-165356^{T}; RB38: *Pandoraea pnomenusa* RB38. "-/-" denotes unmethylated motifs wheras "+/-" denotes hemimethylated motifs. The number of the respective motif type present within the cluster is indicated in parenthesis. "*" sign associated with gene names indicates genes which are identified as flanking genes of unmethylated sites clusters.

CHAPTER 5: DISCUSSION

5.1 Complete Genome Sequencing of *Pandoraea* Species

In this study, the genomes of the type strains of all accepted species in the *Pandoraea* genus and one in-house *P. pnomenusa* landfill isolate were sequenced and assembled to completion with manual curation to ensure full representation of the genome. The main aims of finishing these genomes were to expand the comprehension of this genus *via* the means of genomic data and to generate *in silico* reference data for methylome analysis. In addition, the annotated genomic features data were also analysed in conjunction with the methylation motif distribution data to glean into the correlation between DNA methylation and potential gene regulatory activities.

From the pan-genomic analysis performed, the following information were discerned: average GC content of all species; range of chromosome sizes; distribution and location of extrachromosomal elements (plasmids and prophages); enumeration and functional categorisation of coding sequences; estimation of genomic relatedness among all species; location of putative *oriC* regions and distribution of DnaA boxes clusters. Among this information, the data that are of particular interest for this thesis are the genome-relatedness estimation, location and annotation of extrachromosomal elements and the elements associated with bacterial DNA replication as they could unravel the functional role of adenine methylation.

According to the available GenBank record, as of March 2017, there are 26 deposited *Pandoraea* genome records of which 9 are draft genomes and another 17 are complete genomes (Table 5.1). Among the 17 complete genome in GenBank records, 10 are contributed by this study. These genome data shed light on several interesting metabolic, biotechnological and pathogenicity of the *Pandoraea* species *via* a top-down approach. Firstly, the pathogenic potential and mechanism of *P. pnomenusa* DSM 16536^T

were evaluated using its virulence and antimicrobial resistance genes profile (Lim et al., 2016). Furthermore, the underlying metabolic pathway of the oxalatrophic potential of *P. vervacti*, *P. pnomenusa* RB38, *P. oxalativorans* and the thiosulfate oxidation activity of *P. thiooxydans* were determined *via* genome mining analysis (Chan et al., 2016; Ee et al., 2015; Lim et al., 2015; Yong et al., 2016). The potential of these strains to be used as biocontrol and bioremediation agents are also revealed with the identification of genes with aromatic compounds catabolism and heavy metal and toxic resistance properties. In addition, the complete quorum sensing system of *P. pnomenusa* which represent a new evolutionary branch of AHL-based QS system were also identified and characterised with the aid of these genome data (Lim et al., 2015).

Overall, this information acquired also shed light on the genomic features of the *Pandoraea* spp. that were previously unknown. Further, the availability of these complete genomes in the taxonomical level of genera and deposited in public databases could facilitate future exploration of the metabolic network, biotechnological potential, pathogenicity, molecular drug targets prediction, and most importantly, provide a useful resource in facilitating taxonomic assignment and comprehensive comparative genomic studies of this genus (Kim et al., 2014; Thompson et al., 2013).

Table 5.1: Available GenBank complete genome records of the *Pandoraea* spp.

Organism name	Assembly	Assembly level	Source
P. pnomenusa DSM 16536^T	ASM76761v3	Complete genome	This study
P. pnomenusa RB38	ASM60406v3	Complete Genome	This study
P. sputorum DSM 21091^T	ASM81484v2	Complete Genome	This study
P. pulmonicola DSM 16583^T	ASM81510v2	Complete Genome	This study
P. vervacti DSM 23571^T	ASM93460v2	Complete Genome	This study
P. oxalativorans DSM 23570^T	ASM97278v3	Complete Genome	This study
P. thiooxydans DSM 25325^T	ASM101777v3	Complete Genome	This study
P. faecigallinarum DSM 23572^T	ASM102910v3	Complete Genome	This study
P. norimbergensis DSM 11628^T	ASM146554v3	Complete Genome	This study
P. apista DSM 16535^T	ASM146559v2	Complete Genome	This study
P. pnomenusa RB-44	ASM50458v2	Complete Genome	University of malaya
P. pnomenusa 3kgm	ASM59049v2	Complete Genome	University of malaya
P. apista TF81F4	ASM82696v3	Complete Genome	UCSF
P. apista TF80G25	ASM101078v1	Complete Genome	UCSF
P. apista AU2161	ASM102726v1	Complete Genome	UCSF
P. pnomenusa MCB032	ASM163627v1	Complete Genome	Wuhan Institute of Virology
P. thiooxydans ATSB16	ASM193167v1	Complete Genome	Chungbuk National University

5.2 Methylome Diversity in The Genus *Pandoraea*

Diversity of prokaryotic MTases were increasingly studied in the last decade with the rising availability of DNA base modification data. However, majority of the studies on methylome distribution and functional assessments focused largely on Alpha- and *Gammaproteobacteria* branch with emphasis on the Dam and CcrM solitary MTases. Expanding the study of methylome across the bacterial taxon is crucial to expand the understanding of the impact of DNA methylation on shaping the prokaryotic genome composition. As *Pandoraea* genus comprise a recent branch in the clade of *Betaproteobacteria*, this study could help close gaps in the current understanding of prokaryotic methylomes. In this study, the complete methylome profile of the *Pandoraea* genus, comprising information of DNA methylation target motifs and R-M systems, were studied by using DNA modification detection function incorporated in SMRT sequencing technology. The identification of the methylation motifs and their associated genome-wide methylation pattern present crucial information in comprehension of the enzyme specificities and the potential functional significance of the corresponding R-M genes (Blow et al., 2016).

Overall, the results obtained in this study are in agreement with previously reported findings. Firstly, matching the reported pervasiveness of DNA methylation in most prokaryotic genomes, similar pattern of ubiquity were observed in the genomes of the *Pandoraea* spp. with genome size appearing to be the main influencing factor of the R-M systems distribution patterns (Oliveira et al., 2014). Secondly, the relatively high abundance of solitary MTases are in line with previously reported assessment where degradation of complete R-M systems and horizontal gene transfer events are postulated to be the main contributing factors (Blow et al., 2016; Seshasayee et al., 2012). These MTases are more likely to be retained in the genome over a long evolutionary time due to a neutral selection or the presence of selectable benefits for the host genome

(Seshasayee et al., 2012). Furthermore, matching the results obtained in previous studies, rapid turnover of R-M genes were also observed in the *Pandoraea* spp. genomes, particularly in the mutational decay of M.Pfa23572ORF23735P and its cognate REase gene (Oliveira et al., 2014; Oliveira et al., 2016). The loss of selective advantage of this R-M system could be as a result of the presence of another R-M system (M.Pfa23572II) in the genome.

Furthermore, comparative analysis of the MTases demonstrated a high rate of inter- and intra- species diversity of the MTases, which indicated possible occurences of lateral gene acquisition events. In order to determine the source of the diversity observed, interaction between the R-M genes and mobile genetic elements were analysed. Although a portion of MTases were indeed associated with MGEs, majority of these MTases were observed to be inactive in this study. The most possible explanation is that these MTases conferred an advantage to the MGEs during the course of host infection which favoured their high rate of transfer (Oliveira et al., 2014; Vasu & Nagaraja, 2013). However, upon integration into the host genome, the selective advantages of these MTases were lost as a result of competing MTases present in the genome or due to high metabolic cost for maintenance of the gene hence resulted in a relaxed selection and in eventual gene loss. Furthermore, in agreement with the observation made by Oliveira et al. (2014), majority of the MTases which are active and contributed to the methylation patterns observed are not associated with prophages or plasmids. This suggest that either the R-M genes were acquired by these genomes *via* other MGE forms not examined in this study for instance CRISPR-cas and transposases, or certain R-M systems exhibit behaviour as an independent mobile unit.

In summary, the result obtained in this study provide new insights to the methylome profile in the genus of *Pandoraea* spp. and have expand the knowledge on

current R-M systems repertoire by identifying new R-M systems with novel sequence specificities.

5.3 Discovery of A Class of Novel Orphan Methyltransferase with Analogous Genomic Properties to CcrM and Dam

Orphan MTases, which are defined as MTases that exist independently without a cognate REase partner, are one of the major topic in the field of prokaryotic methylome research (Blow et al., 2016; Murphy et al., 2013). These MTases are of interest due to their roles in various important cellular processes, including genome replication, population evolution and gene expression regulation (Boye & Løbner-Olesen, 1990; Palmer & Marinus, 1994; Schlagman et al., 1986). Furthermore, an increasing pathogenicity associated roles of orphan MTases are being uncovered in recent years, including antibiotic stress survival and regulation of virulence genes (Cohen et al., 2016; Marinus & Casadesus, 2009b).

Various studies have demonstrated the abundance and wide distribution of orphan MTases in prokaryotic genome, owing their ubiquitous distribution to horizontal gene transfer events and the selective advantages conferred by these genes to the host (Blow et al., 2016; Matveyev et al., 2001; Seshasayee et al., 2012). However, despite the abundance of orphan MTases observed and reported, evolutionarily conserved orphan MTases, defined as MTases with orthologs present in >50% of species across a respective taxonomic class (genus, family or class) and are associated with crucial cellular functions, are relatively rare. To date, only three widely-conserved orphan MTases were identified, namely Dcm MTase family (conserved in *E. coli*), Dam MTase family (conserved in *Gammaproteobacteria*) and CcrM MTase family (conserved in *Alphaproteobacteria*) (Militello et al., 2012; Wright et al., 1997). These orphan MTases share several similar genomic properties such as evolutionary stability, sequence conservation and lineage

coherence. To the best of my knowledge, no conserved orphan MTase family was identified in the class of *Betaproteobacteria*.

In this study, a novel class of orphan MTases, which encode methylation of the GTWWAC motif and were highly conserved in the *Burkholderiaceae* and *Oxalabacteraceae* families was identified. The homologous relationship of the analysed MTases were demonstrated with the high sequence similarity determined *via* pairwise sequence similarity analysis. Furthermore, a high degree of evolutionary conservation of these MTases were also evidenced by the conservation of amino acid residues and functional domains which indicate the application of a rigorous negative selection pressure against amino acid changes onto this MTases. The consistent methylation pattern of the GTWWAC motifs in the genomes of *Pandoraea* spp. also indicate conservation in the functional aspect. Further evolutionary conservation of these MTases were observed from the phylogenetic analysis where the pattern of clustering of the genes indicate a genealogical relationship. This conservation pattern of GTWWAC MTase homologs over evolutionary time can be considered as an indication of essentiality or functional significance among these genomes (Luo et al., 2015).

In addition, among the respective *Burkholderiaceae* and *Oxalabacteraceae* genomes, an intra-family gene neighbourhood synteny was observed. GTWWAC MTases of the *Burkholderiaceae* and *Oxalabacteraceae* families were localised within the tryptophan and histidine biosynthesis operons respectively and flanked by a largely conserved neighbouring gene block (>10 kb). Both operons are ancient essential aromatic amino acid biosynthetic operon which are known for their properties as a whole-pathway operon and were reported to have interconnected metabolic relationship (Xie et al., 2003). It is also notable that Dam methylase were also found to be located in a multicistronic operon, and associated with the *aroK* (shikimate kinase I) and *aroB* (3-dehydroquinate synthase) genes which are involved in the biosynthesis of aromatic amino acids

(Lyngstadaas et al., 1995). Through the analysis of other genomes (beyond the *Burkholderiaceae* and *Oxalabacteraceae* families) which were detected to harbour GTWWAC MTases homolog, the evolutionary and functional relevance of the associated neighbouring genes can be gleaned. Firstly, among the genomes (*R. rubra* and *M. flavia*) which the GTWWAC MTase homologs were postulated to be acquired *via* the means of transposase-mediated horizontal gene transfer, the MTases was observed to be transferred along with the a conserved segment of neighbouring genes. This manner of gene string conservation is indicative of a functional advantage in the aspect of gene expression or its regulation (Lathe et al., 2000; Rogozin et al., 2002). This manner of transfer could also potentially explain the difference in gene neighbourhood context observed between the families.

From this analysis, GTWWAC MTases potentially constitute a novel class of orphan MTases with a largely taxonomically restricted distribution pattern within the class of *Betaproteobacteria*. The genomic properties exhibited by these MTases were analogous to those observed in Dam and CcrM in the aspect of evolutionary conservation and lineage coherence. Several questions remained unanswered following the results obtained in this analysis: (1) From the observation where GTWWAC MTase homologs represent extraneous insertion into ancestral biosynthetic operon indicated that these genes are of a recent evolutionary state, what could have led to the emergence of this class of orphan MTases? (2) What evolutionary events could have contributed to the family-specific GTWWAC MTase associated neighbouring gene context? (3) What is the functional advantage or association between the GTWWAC MTases and the conserved neighbouring genes? (4) What is the functional role of these orphan MTases within the host genome? (5) Could this class of orphan MTases represent the source of a new conserved class of orphan MTase which could be distributed widely across the class of *Betaproteobacteria* with evolutionary time?

5.4 Methylome Distribution Analysis of GTWWAC motif in *Pandoraea* genomes

In this study, genome-wide comparative GTWWAC methylome distribution analysis in the *Pandoraea* spp. genome was performed to obtain *in silico* prediction of potential functional role of GTWWAC MTases.

Prokaryotic DNA MTases, particularly the solitary DNA MTases were discovered to have important functional roles in timing of DNA replication, chromosome partitioning, DNA repair, control of transposition, and conjugal transfer of plasmids and gene regulation (Casadesus & Low, 2006; Murphy et al., 2013; Wion & Casadesus, 2006). One of the proposed mechanisms of influence is the structural effects contributed by the methylation status of the adenine bases in the target recognition motifs which could lead to modulation of DNA-protein interaction, for example the interaction of transcription factor with their target DNA binding sites (Sanchez-Romero et al., 2015). The comparable genomic properties of GTWWAC MTases with Dam and CcrM led to the hypothesis that GTWWAC MTases could potentially have equivalent biological functions. Therefore, genome-wide distribution of GTWWAC motif with different methylation status were analysed to explore this hypothesis. Furthermore, analysis of these result in a pan-genus scale could significantly enhanced the statistical significance of the functional role prediction.

From the pan-genus comparative methylome distribution analyses, several distinctive features were identified. Firstly, the identification of a highly conserved hypermethylated region which overlapped a DnaA box clusters (with a potential role in regulation of chromosomal replication and replication initiation control) in all analysed *Pandoraea* genomes is suggestive that GTWWAC MTases have a potential regulatory role in DNA replication. Both Dam and CcrM-mediated methylation activity were also discovered to play integral role in the regulation of DNA replication, particularly in the aspect of DNA replication initiation control, albeit with different mechanisms suggesting

a possibility that DNA replication regulation is a conserved function in these classes of orphan MTases (Boye & Lobner-Olesen, 1990; Reisenauer et al., 1999; Wright et al., 1997). Secondly, implication of GTWWAC MTases in transcriptional control of gene expressions are also indicated with the association of a large portion of motif, particularly the unmethylated sites, with transcriptional regulators of various families. Thirdly, the identification of various essential bacterial outer membrane proteins and cell wall components associated with methylation hotspots and unmethylated sites, are also indicative that GTWWAC MTases, like Dam, could have a regulatory role in cell membrane biogenesis and membrane stability maintenance (Pucciarelli et al., 2002). Additionally, matching the finding of various studies which implicated the affiliation of Dam methylome with antibiotic resistance activity (Adam et al., 2008; Cohen et al., 2016), the methylome sites of GTWWAC motifs were also found to be associated with various genes which could confer antibiotic resistance *via* means of enzymatic detoxification, cell envelope modification or active efflux mechanism (Chen et al., 2009; Llano-Sotelo et al., 2002; Ma et al., 1996; Sampson et al., 1989; Wright et al., 1998; Zhang et al., 2008). This result is particularly interesting, as multiple literature highlighted the remarkable multidrug resistance properties of *Pandoraea* strains and targeting adenine methylation could serve as an attractive approach to potentiate antibiotics treatment (Daneshvar et al., 2001; Stryjewski et al., 2003).

In summary, the pan-genus methylome distribution analysis demonstrated that GTWWAC MTases potentially have analogous function to Dam and CcrM MTases in the aspect of gene expression control, DNA replication regulation, cell membrane integrity maintenance and structural support for antibiotic resistance. These data could be valuable in aiding in the experimental design of further downstream exploratory and confirmatory studies pertaining to the functional role of GTWWAC MTases.

CHAPTER 6: CONCLUSION

In this thesis, the first pan-genus complete genomes analysis and comprehensive methylome characterisation of the *Pandoraea* genus were presented. The results and data obtained from this study have bridged several information gaps in current prokaryotic genomic and methylome research. Firstly, the advantage of SMRT sequencing in substantially expanding the comprehension of prokaryotic biology could be evidenced in this study *via* facile construction of high quality complete genomes along with their methylome information in base-pair resolution. The pan-genus genomic analyses performed on all representative strains of the named *Pandoraea* species have revealed the fundamental genomic features of this genus. Furthermore, the availability of these data could provide a roadmap for future studies in answering various research questions pertaining to this recently established genus particularly in the aspect of its antibiotic resistance properties, underlying pathogenicity, treatment strategies and their potential biotechnology applications. Secondly, while the current understanding of prokaryotic methylome landscape remains largely limited, particularly in the beta subdivision of *Proteobacteria*, the methylome profile analysis conducted in this study have advanced the current knowledge on the diversity of prokaryotic R-M systems with the identification and characterisation of novel MTases. Additionally, a potential novel class of solitary MTase were identified *via* the identification of GTWWAC motif which has a conserved pattern of distribution among the *Pandoraea* species. This class of MTases were found to be widely distributed in the *Burkholderiaceae* and *Oxalabacteraceae* families' genomes and demonstrated analogous genomic properties with the Dam and CcrM MTases. Further analysis of the genome-wide motif distribution have also suggested that these MTases may have a comparable functional role with Dam and CcrM MTases in the aspect of gene expression control, DNA replication regulation and cell membrane integrity maintenance.

I acknowledge that more work is needed to provide a complete appreciation of the findings in this study such as analysis of knockout mutants to experimentally verify the *in silico* functional role prediction of GTWWAC methylation. The correlation of motif types and their associated distribution data with the genuine functional role should also be studied rigorously in order to determine the value and feasibility of *in silico* functional prediction *via* motif distribution analysis. This would contribute significantly in guiding the future direction of prokaryotic methylome research. However, as our current understanding of the connection between DNA methylation and its regulatory role remains ambiguous, the findings in this study hold value in providing new insights to fill the chasm of knowledge in the functional significance of DNA methylation in prokaryotes.

REFERENCES

Adam, M., Murali, B., Glenn, N. O., & Potter, S. S. (2008). Epigenetic inheritance based evolution of antibiotic resistance in bacteria. *BMC Evolutionary Biology, 8*, 52.

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., . . . Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology, 12*(2), R18.

Anandham, R., Indiragandhi, P., Kwon, S. W., Sa, T. M., Jeon, C. O., Kim, Y. K., & Jee, H. J. (2010). *Pandoraea thiooxydans* sp. nov., a facultatively chemolithotrophic, thiosulfate-oxidizing bacterium isolated from rhizosphere soils of sesame (*Sesamum indicum* L.). *International Journal of Systematic and Evolutionary Microbiology, 60*(1), 21-26.

Arber, W., & Linn, S. (1969). DNA modification and restriction. *Annual Review of Biochemistry, 38*(1), 467-500.

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research, 44*(W1), W16-21.

Atkinson, R. M., LiPuma, J. J., Rosenbluth, D. B., & Dunne, W. M. (2006). Chronic Colonization with *Pandoraea apista* in Cystic Fibrosis Patients Determined by Repetitive-Element-Sequence PCR. *Journal of Clinical Microbiology, 44*(3), 833-836.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., . . . Kubal, M. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics, 9*(1), 75.

Bakker, A., & Smith, D. W. (1989). Methylation of GATC sites is required for precise timing between rounds of DNA replication in *Escherichia coli*. *Journal of Bacteriology, 171*(10), 5738-5742.

Balbontín, R., Rowley, G., Pucciarelli, M. G., López-Garrido, J., Wormstone, Y., Lucchini, S., . . . Casadesús, J. (2006). DNA Adenine Methylation Regulates Virulence Gene Expression in *Salmonella enterica* Serovar Typhimurium. *Journal of Bacteriology, 188*(23), 8160-8168.

Bandounas, L., Wierckx, N. J., de Winde, J. H., & Ruijssenaars, H. J. (2011). Isolation and characterization of novel bacterial strains exhibiting ligninolytic potential. *BMC Biotechnology, 11*(1), 94.

Bendall, M. L., Luong, K., Wetmore, K. M., Blow, M., Korlach, J., Deutschbauer, A., & Malmstrom, R. R. (2013). Exploring the roles of DNA methylation in the metal-reducing bacterium *Shewanella oneidensis* MR-1. *Journal of Bacteriology*, *195*(21), 4966-4974.

Bickle, T. A. (2004). Restricting restriction. *Molecular Microbiology, 51*(1), 3-5.

Bickle, T. A., & Kruger, D. H. (1993). Biology of DNA restriction. *Microbiological Reviews, 57*(2), 434-450.

Blow, M. J., Clark, T. A., Daum, C. G., Deutschbauer, A. M., Fomenkov, A., Fries, R., . . . Roberts, R. J. (2016). The Epigenomic Landscape of Prokaryotes. *PLOS Genetics, 12*(2), e1005854.

Boye, E., & Lobner-Olesen, A. (1990). The role of dam methyltransferase in the control of DNA replication in *E. coli*. *Cell, 62*(5), 981-989.

Campbell, J. L., & Kleckner, N. (1990). *E. coli oriC* and the *dnaA* gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork. *Cell, 62*(5), 967-979.

Casadesus, J., & Low, D. (2006). Epigenetic gene regulation in the bacterial world. *Microbiology and Molecular Biology Reviews, 70*(3), 830-856.

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Molecular Microbiology, 49*(2), 277-300.

Chan, J. Z.-M., Halachev, M. R., Loman, N. J., Constantinidou, C., & Pallen, M. J. (2012). Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. *BMC Microbiology, 12*(1), 302.

Chan, K.-G., Yong, D., Ee, R., Lim, Y.-L., Yu, C.-Y., Tee, K.-K., . . . Ang, G.-Y. (2016). Complete genome sequence of *Pandoraea oxalativorans* DSM 23570^{T}, an oxalate metabolizing soil bacterium. *Journal of Biotechnology, 219*, 124-125.

Chen, L., Paulsen, D. B., Scruggs, D. W., Banes, M. M., Reeks, B. Y., & Lawrence, M. L. (2003). Alteration of DNA adenine methylase (Dam) activity in *Pasteurella multocida* causes increased spontaneous mutation frequency and attenuation in mice. *Microbiology, 149*(Pt 8), 2283-2290.

Chen, L. X., He, S., Li, C., & Ryu, J. (2009). Sublethal kanamycin induced cross resistance to functionally and structurally unrelated antibiotics. *Journal of Experimental and Immunology, 13*, 53-57.

Cheng, X., Kumar, S., Klimasauskas, S., & Roberts, R. J. (1993a). Crystal structure of the HhaI DNA methyltransferase. *Cold Spring Harbor Symposia on Quantitative Biology, 58*, 331-338.

Cheng, X., Kumar, S., Posfai, J., Pflugrath, J. W., & Roberts, R. J. (1993b). Crystal structure of the HhaI DNA methyltransferase complexed with S-adenosyl-L-methionine. *Cell, 74*(2), 299-307.

Cheng, X., & Roberts, R. J. (2001). AdoMet-dependent methylation, DNA methyltransferases and base flipping. *Nucleic Acids Research, 29*(18), 3784-3795.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., . . . Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods, 10*(6), 563-569.

Clark, T. A., Lu, X., Luong, K., Dai, Q., Boitano, M., Turner, S. W., . . . Korlach, J. (2013). Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biology, 11*(1), 4.

Clark, T. A., Murray, I. A., Morgan, R. D., Kislyuk, A. O., Spittle, K. E., Boitano, M., . . . Korlach, J. (2012). Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Research, 40*(4), e29.

Clark, T. A., Spittle, K. E., Turner, S. W., & Korlach, J. (2011). Direct detection and sequencing of damaged DNA bases. *Genome Integrity, 2*, 10.

Coenye, T., Falsen, E., Hoste, B., Ohlén, M., Goris, J., & Govan, J. R. (2000). Description of *Pandoraea* gen. nov. with *Pandoraea apista* sp. nov., *Pandoraea pulmonicola* sp. nov., *Pandoraea pnomenusa* sp. nov., *Pandoraea sputorum* sp. nov. and *Pandoraea norimbergensis* comb. nov. *International Journal of Systematic and Evolutionary Microbiology, 50*, 887-899.

Coenye, T., Vandamme, P., Govan, J. R., & LiPuma, J. J. (2001). Taxonomy and identification of the *Burkholderia cepacia* complex. *Journal of Clinical Microbiology, 39*(10), 3427-3436.

Cohen, N. R., Ross, C. A., Jain, S., Shapiro, R. S., Gutierrez, A., Belenky, P., . . . Collins, J. J. (2016). A role for the bacterial GATC methylome in antibiotic stress survival. *Nature Genetics, 48*(5), 581-586.

Colbert, C. L., Agar, N. Y. R., Kumar, P., Chakko, M. N., Sinha, S. C., Powlowski, J. B., . . . Bolin, J. T. (2013). Structural Characterization of *Pandoraea pnomenusa* B-356 Biphenyl Dioxygenase Reveals Features of Potent Polychlorinated Biphenyl-Degrading Enzymes. *PLoS ONE, 8*(1), e52550.

Corvaglia, A. R., Francois, P., Hernandez, D., Perron, K., Linder, P., & Schrenzel, J. (2010). A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical *Staphylococcus aureus* strains. *Proceedings of the National Academy of Sciences of the United States of America, 107*(26), 11954-11958.

Daneshvar, M. I., Hollis, D. G., Steigerwalt, A. G., Whitney, A. M., Spangler, L., Douglas, M. P., . . . Weyant, R. S. (2001). Assignment of CDC weak oxidizer group 2 (WO-2) to the genus *Pandoraea* and characterization of three new *Pandoraea* genomospecies. *Journal of Clinical Microbiology, 39*(5), 1819-1826.

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.

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE, 5*(6), e11147.

Davis, B. M., Chao, M. C., & Waldor, M. K. (2013). Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Current Opinion in Microbiology, 16*(2), 192-198.

Degand, N., Lotte, R., Decondé Le Butor, C., Segonds, C., Thouverez, M., Ferroni, A., . . . Carrère, J. (2015). Epidemic spread of *Pandoraea pulmonicola* in a cystic fibrosis center. *BMC Infectious Diseases, 15*(1), 1-7.

Delcher, A. L., Harmon, D., Kasif, S., White, O., & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Research, 27*(23), 4636-4641.

Denisov, G., Walenz, B., Halpern, A. L., Miller, J., Axelrod, N., Levy, S., & Sutton, G. (2008). Consensus generation and variant detection by Celera Assembler. *Bioinformatics, 24*(8), 1035-1040.

Detecting DNA Base Modifications Using Single Molecule, Real-Time Sequencing. (2015). Retrieved 15 May 2017, from http://www.pacb.com/wp-content/uploads/2015/09/WP_Detecting_DNA_Base_Modifications_Using_SMRT_Sequencing.pdf

Detecting DNA Base Modifications: SMRT Analysis of Microbial Methylomes. (2017). Retrieved 12 May 2017, from https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Methylome-Analysis-Technical-Note

Diekmann, S. (1987). DNA methylation can enhance or induce DNA curvature. *The EMBO Journal, 6*(13), 4213-4217.

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research, 36*(16), e105-e105.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5), 1792-1797.

Ee, R., Ambrose, M., Lazenby, J., Williams, P., Chan, K. G., & Roddam, L. (2015). Genome Sequences of Two *Pandoraea pnomenusa* Isolates Recovered 11 Months Apart from a Cystic Fibrosis Patient. *Genome Announcements, 3*(1), e01389-14.

Ee, R., Lim, Y. L., Kin, L. X., Yin, W. F., & Chan, K. G. (2014). Quorum sensing activity in Pandoraea pnomenusa RB38. *Sensors,, 14*(6), 10177-10186.

Ee, R., Yong, D., Lim, Y. L., Yin, W.-F., & Chan, K.-G. (2015). Complete genome sequence of oxalate-degrading bacterium *Pandoraea vervacti* DSM 23571^{T}. *Journal of Biotechnology, 204*, 5-6.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., . . . Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science, 323*(5910), 133.

Engel, J. D., & von Hippel, P. H. (1978). Effects of methylation on the stability of nucleic acid conformations. Studies at the polymer level. *The Journal of Biological Chemistry, 253*(3), 927-934.

Fang, G., Munera, D., Friedman, D. I., Mandlik, A., Chao, M. C., Banerjee, O., . . . Schadt, E. E. (2012). Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nature Biotechnology, 30*(12), 1232-1239.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution, 39*(4), 783-791.

Feng, Z., Fang, G., Korlach, J., Clark, T., Luong, K., Zhang, X., . . . Schadt, E. (2013). Detecting DNA Modifications from SMRT Sequencing Data by Modeling Sequence Context Dependence of Polymerase Kinetic. *PLOS Computational Biology, 9*(3), e1002935.

Ferrarini, M., Moretto, M., Ward, J. A., Surbanovski, N., Stevanovic, V., Giongo, L., . . . Sargent, D. J. (2013). An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genomics, 14*, 670.

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., . . . Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods, 7*(6), 461-465.

Gao, F., & Zhang, C.-T. (2008). Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. *BMC Bioinformatics, 9*(1), 79.

Garcia-Del Portillo, F., Pucciarelli, M. G., & Casadesus, J. (1999). DNA adenine methylase mutants of *Salmonella typhimurium* show defects in protein secretion, cell invasion, and M cell cytotoxicity. *Proceedings of the National Academy of Sciences of the United States of America, 96*(20), 11578-11583. Gómez-Gil, L., Kumar, P., Barriault, D., Bolin, J. T., Sylvestre, M., & Eltis, L. D. (2007). Characterization of Biphenyl Dioxygenase of *Pandoraea pnomenusa* B-356 As a Potent Polychlorinated Biphenyl-Degrading Enzyme. *Journal of Bacteriology, 189*(15), 5705-5715.

Gonzalez, D., & Collier, J. (2013). DNA methylation by CcrM activates the transcription of two genes required for the division of *Caulobacter crescentus*. *Molecular Microbiology, 88*(1), 203-218.

Gonzalez, D., Kozdon, J. B., McAdams, H. H., Shapiro, L., & Collier, J. (2014). The functions of DNA methylation by CcrM in *Caulobacter crescentus*: a global approach. *Nucleic Acids Research, 42*(6), 3720-3735.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology, 57*(Pt 1), 81-91.

Gromova, E. S., & Khoroshaev, A. V. (2003). Prokaryotic DNA methyltransferases: the structure and the mechanism of interaction with DNA. *Molecular Biology, 37*(2), 300-314.

Grunert, M., Dorn, C., Cui, H., Dunkel, I., Schulz, K., Schoenhals, S., . . . Sperling, S. R. (2016). Comparative DNA methylation and gene expression analysis identifies novel genes for structural congenital heart diseases. *Cardiovascular Research, 112*(1), 464-477.

Han-Jen, R. E., Wai-Fong, Y., & Kok-Gan, C. (2013). *Pandoraea* sp. RB-44, a novel quorum sensing soil bacterium. *Sensors, 13*(10), 14121-14132.

Heiner, C., Wang, S., Ashby, M., Guo, Y., Underwood, J., & Baybayan, P. (2013). Greater than 10 kb Read Lengths Routine when Sequencing with Pacific Biosciences' XL Release. *Journal of Biomolecular Techniques, 24*(Suppl), S43.

Hénaut, A., Rouxel, T., Gleizes, A., Moszer, I., & Danchin, A. (1996). Uneven Distribution of GATC Motifs in the *Escherichia coli* Chromosome, its Plasmids and its Phages. *Journal of Molecular Biology, 257*(3), 574-585.

HGAP. (2016, 2 September). Retrieved 11 May 2017, from https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP

Hu, C., Zhao, Y., Sun, H., & Yang, Y. (2017). Synergism of Dam, MutH, and MutS in methylation-directed mismatch repair in *Escherichia coli*. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 795*, 31-33.

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics, 26*(5), 680-682.

Jeltsch, A. (2002). Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. *Chembiochem, 3*(4), 274-293.

Jeltsch, A. (2003). Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems? *Gene, 317*(1-2), 13-16.

Jiang, X.-W., Liu, H., Xu, Y., Wang, S.-J., Leak, D. J., & Zhou, N.-Y. (2009). Genetic and biochemical analyses of chlorobenzene degradation gene clusters in *Pandoraea* sp. strain MCB032. *Archives of Microbiology, 191*(6), 485-492.

Jin, B., Li, Y., & Robertson, K. D. (2011). DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy? *Genes & Cancer, 2*(6), 607-617.

Jin, Z. X., Wang, C., Dong, W., & Li, X. (2007). Isolation and some properties of newly isolated oxalate-degrading *Pandoraea* sp. OXJ-11 from soil. *Journal of Applied Microbiology, 103*(4), 1066-1073.

Jørgensen, I. M., Johansen, H. K., Frederiksen, B., Pressler, T., Hansen, A., Vandamme, P., . . . Koch, C. (2003). Epidemic spread of *Pandoraea apista*, a new pathogen causing severe lung disease in cystic fibrosis patients. *Pediatric pulmonology, 36*(5), 439-446.

Julio, S. M., Heithoff, D. M., Provenzano, D., Klose, K. E., Sinsheimer, R. L., Low, D. A., & Mahan, M. J. (2001). DNA adenine methylase is essential for viability and plays a role in the pathogenesis of *Yersinia pseudotuberculosis* and *Vibrio cholerae*. *Infection and Immunity, 69*(12), 7610-7615.

Kauc, L., & Piekarowicz, A. (1978). Purification and Properties of a New Restriction Endonuclease from *Haemophilus influenzae* Rf. *European Journal of Biochemistry, 92*(2), 417-426.

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*(Pt 2), 346-351.

Klappenbach, J. A., Dunbar, J. M., & Schmidt, T. M. (2000). rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria. *Applied and Environmental Microbiology, 66*(4), 1328-1333.

Klimasauskas, S., Timinskas, A., Menkevicius, S., Butkienè, D., Butkus, V., & Janulaitis, A. (1989). Sequence motifs characteristic of DNA[cytosine-N4]methyltransferases: similarity to adenine and cytosine-C5 DNA-methylases. *Nucleic Acids Research, 17*(23), 9823-9832.

Kobayashi, I. (2001). Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Research, 29*(18), 3742-3756.

Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology, 23*, 110-120.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., . . . Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology, 30*, 693-700.

Korlach, J., Bjornson, K. P., Chaudhuri, B. P., Cicero, R. L., Flusberg, B. A., Gray, J. J., . . . Turner, S. W. (2010). Real-time DNA sequencing from single polymerase molecules. *Methods in Enzymology, 472*, 431-455.

Korlach, J., & Turner, S. W. (2012). Going beyond five bases in DNA sequencing. *Current Opinion in Structural Biology, 22*(3), 251-261.

Korlach, J., & Turner, S. W. (2013). Zero-Mode Waveguides. In G. C. K. Roberts (Ed.), *Encyclopedia of Biophysics* (pp. 2793-2795). Berlin, Heidelberg: Springer Berlin Heidelberg.

Krumsiek, J., Arnold, R., & Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics, 23*(8), 1026-1028.

Kumar, S., Cheng, X., Klimasauskas, S., Mi, S., Posfai, J., Roberts, R. J., & Wilson, G. G. (1994). The DNA (cytosine-5) methyltransferases. *Nucleic Acids Research, 22*(1), 1-10.

Larsen, P., & Smith, T. (2012). Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire. *BMC Immunology, 13*, 52.

Lathe, W. C., 3rd, Snel, B., & Bork, P. (2000). Gene context conservation of a higher order than operons. *Trends in Biochemical Sciences, 25*(10), 474-479.

Lee, Z. M.-P., Bussema, C., & Schmidt, T. M. (2009). rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Research, 37*(Database issue), D489-D493.

Leonard, M. T., Davis-Richardson, A. G., Ardissone, A. N., Kemppainen, K., Drew, J. C., Ilonen, J., . . . Triplett, E. W. (2014). The methylome of the gut microbiome: disparate Dam methylation patterns in intestinal *Bacteroides dorei*. *Frontiers in Microbiology, 5*, 361.

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., & Webb, W. W. (2003). Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science, 299*(5607), 682.

Liang, M., Raley, C., Zheng, X., Kutty, G., Gogineni, E., Sherman, B. T., . . . Huang, D. W. (2016). Distinguishing highly similar gene isoforms with a clustering-based bioinformatics analysis of PacBio single-molecule long reads. *BioData Mining, 9*(1), 13.

Lim, Y.-L., Ee, R., How, K.-Y., Lee, S.-K., Yong, D., Tee, K. K., . . . Chan, K.-G. (2015). Complete genome sequencing of *Pandoraea pnomenusa* RB38 and Molecular Characterization of Its N-acyl homoserine lactone synthase gene *ppnI*. *PeerJ, 3*, e1225.

Lim, Y.-L., Ee, R., Yong, D., Tee, K.-K., Yin, W.-F., & Chan, K.-G. (2015). Complete genome of *Pandoraea pnomenusa* RB-38, an oxalotrophic bacterium isolated from municipal solid waste landfill site. *Journal of Biotechnology, 214*, 83-84.

Lim, Y.-L., Ee, R., Yong, D., Yu, C.-Y., Ang, G.-Y., Tee, K.-K., . . . Chan, K.-G. (2016). Complete Genome Sequence Analysis of *Pandoraea pnomenusa* Type Strain DSM 16536^{T} Isolated from a Cystic Fibrosis Patient. *Frontiers in Microbiology, 7*, 109.

Liz, J. A. Z.-E., Jan-Roblero, J., de la Serna, J. Z.-D., de León, A. V.-P., & Hernández-Rodríguez, C. (2009). Degradation of polychlorinated biphenyl (PCB) by a consortium obtained from a contaminated soil composed of *Brevibacterium*, *Pandoraea* and *Ochrobactrum*. *World Journal of Microbiology and Biotechnology, 25*(1), 165-170.

Llano-Sotelo, B., Azucena, E. F., Jr., Kotra, L. P., Mobashery, S., & Chow, C. S. (2002). Aminoglycosides modified by resistance enzymes display diminished binding to the bacterial ribosomal aminoacyl-tRNA site. *Chemistry & Biology, 9*(4), 455-463.

Lluch-Senar, M., Luong, K., Lloréns-Rico, V., Delgado, J., Fang, G., Spittle, K., . . . Serrano, L. (2013). Comprehensive Methylome Characterization of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* at Single-Base Resolution. *PLOS Genetics, 9*(1), e1003191.

Loutet, S. A., & Valvano, M. A. (2011). Extreme antimicrobial Peptide and polymyxin B resistance in the genus *Burkholderia*. *Frontiers in Microbiology, 2*, 159.

Low, D. A., Weyand, N. J., & Mahan, M. J. (2001). Roles of DNA Adenine Methylation in Regulating Bacterial Gene Expression and Virulence. *Infection and Immunity, 69*(12), 7197-7204.

Luo, H., Gao, F., & Lin, Y. (2015). Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. *Scientific Reports, 5*, 13210.

Lyngstadaas, A., Løbner-Olesen, A., & Boye, E. (1995). Characterization of three genes in the dam-containing operon of *Escherichia coli*. *Molecular and General Genetics, 247*(5), 546-554.

Ma, D., Alberti, M., Lynch, C., Nikaido, H., & Hearst, J. E. (1996). The local repressor AcrR plays a modulating role in the regulation of *acrAB* genes of *Escherichia coli* by global stress signals. *Molecular Microbiology, 19*(1), 101-112.

Mackiewicz, P., Zakrzewska-Czerwinska, J., Zawilak, A., Dudek, M. R., & Cebrat, S. (2004). Where does bacterial replication start? Rules for predicting the *oriC* region. *Nucleic Acids Research, 32*(13), 3781-3791.

Malone, T., Blumenthal, R. M., & Cheng, X. (1995). Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes. *Journal of Molecular Biology, 253*(4), 618-632.

Marinus, M. G., & Casadesus, J. (2009). Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more. *FEMS Microbiology Reviews, 33*(3), 488-503.

Marinus, M. G., & Løbner-Olesen, A. (2014). DNA Methylation. *EcoSal Plus, 6*(1). doi: 10.1128/ecosalplus.ESP-0003-2013

Matveyev, A. V., Young, K. T., Meng, A., & Elhai, J. (2001). DNA methyltransferases of the cyanobacterium *Anabaena* PCC 7120. *Nucleic Acids Research, 29*(7), 1491-1506.

Mehling, J. S., Lavender, H., & Clegg, S. (2007). A Dam methylation mutant of Klebsiella pneumoniae is partially attenuated. *FEMS Microbiology Letters, 268*(2), 187-193.

Meisel, A., Bickle, T. A., Kruger, D. H., & Schroeder, C. (1992). Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. *Nature, 355*(6359), 467-469.

Militello, K. T., Simon, R. D., Qureshi, M., Maines, R., Van Horne, M. L., Hennick, S. M., . . . Pounder, S. (2012). Conservation of Dcm-mediated Cytosine DNA Methylation in *Escherichia coli*. *FEMS Microbiology Letters, 328*(1), 78-85.

Modrich, P. (1991). Mechanisms and biological effects of mismatch repair. *Annual Review of Genetics, 25*, 229-253.

Molnarova, V., Pristas, P., & Javorsky, P. (1999). Prevalence of CTGCAG Recognizing Restriction and Modification Systems in Ruminal Selenomonades. *Anaerobe, 5*(1), 37-41.

Mou, K. T., Muppirala, U., Severin, A., Clark, T., Boitano, M., & Plummer, P. J. (2015). A comparative analysis of methylome profiles of *Campylobacter jejuni* sheep abortion isolate and gastroenteric strains using PacBio data. *Frontiers in Microbiology, 5*.

Murphy, J., Mahony, J., Ainsworth, S., Nauta, A., & van Sinderen, D. (2013). Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. *Applied and Environmental Microbiology, 79*(24), 7547-7555.

Murphy, K. C., Ritchie, J. M., Waldor, M. K., Løbner-Olesen, A., & Marinus, M. G. (2008). Dam Methyltransferase Is Required for Stable Lysogeny of the Shiga Toxin (Stx2)-Encoding Bacteriophage 933W of Enterohemorrhagic *Escherichia coli* O157:H7. *Journal of Bacteriology, 190*(1), 438-441.

Murray, N. E. (2000). Type I Restriction Systems: Sophisticated Molecular Machines (a Legacy of Bertani and Weigle). *Microbiology and Molecular Biology Reviews, 64*(2), 412-434.

Murray, N. E. (2002). 2001 Fred Griffith review lecture. Immigration control of DNA in bacteria: self versus non-self. *Microbiology, 148*(Pt 1), 3-20.

Naito, T., Kusano, K., & Kobayashi, I. (1995). Selfish behavior of restriction-modification systems. *Science, 267*(5199), 897-899.

Obarska-Kosinska, A., Taylor, J. E., Callow, P., Orlowski, J., Bujnicki, J. M., & Kneale, G. G. (2008). HsdR Subunit of the Type I Restriction-Modification Enzyme EcoR124I: Biophysical Characterisation and Structural Modelling. *Journal of Molecular Biology, 376*(2), 438-452.

Okeke, B. C., Siddique, T., Arbestain, M. C., & Frankenberger, W. T. (2002). Biodegradation of gamma-hexachlorocyclohexane (lindane) and alpha-hexachlorocyclohexane in water and a soil slurry by a *Pandoraea* species. *Journal of Agricultural and Food Chemistry, 50*(9), 2548-2555.

Olaitan, A. O., Morand, S., & Rolain, J. M. (2014). Mechanisms of polymyxin resistance: acquired and intrinsic resistance in bacteria. *Frontiers in Microbiology, 5*, 643.

Oliveira, P. H., Touchon, M., & Rocha, E. P. (2014). The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Research, 42*(16), 10618-10631.

Oliveira, P. H., Touchon, M., & Rocha, E. P. C. (2016). Regulation of genetic flux between bacteria by restriction–modification systems. *Proceedings of the National Academy of Sciences, 113*(20), 5658-5663.

Ozaki, S., Kishimoto, N., & Fujita, T. (2007). Change in the Predominant Bacteria in a Microbial Consortium Cultured on Media Containing Aromatic and Saturated Hydrocarbons as the Sole Carbon Source. *Microbes and Environments, 22*(2), 128-135.

Palmer, B. R., & Marinus, M. G. (1994). The dam and dcm strains of *Escherichia coli*--a review. *Gene, 143*(1), 1-12.

Pfeifer, G. P. (2016). Epigenetics: An elusive DNA base in mammals. *Nature, 532*(7599), 319-320.

Pham, T. T., Tu, Y., & Sylvestre, M. (2012). Remarkable ability of *Pandoraea pnomenusa* B356 biphenyl dioxygenase to metabolize simple flavonoids. *Applied and Environmental Microbiology, 78*(10), 3560-3570.

Pingoud, A., Fuxreiter, M., Pingoud, V., & Wende, W. (2005). Type II restriction endonucleases: structure and mechanism. *Cellular and Molecular Life Sciences, 62*(6), 685-707.

Pósfai, J., Bhagwat, A. S., Pósfai, G., & Roberts, R. J. (1989). Predictive motifs derived from cytosine methyltransferases. *Nucleic Acids Research, 17*(7), 2421-2435.

Pucciarelli, M. G., Prieto, A. I., Casadesus, J., & Garcia-del Portillo, F. (2002). Envelope instability in DNA adenine methylase mutants of *Salmonella enterica*. *Microbiology, 148*(Pt 4), 1171-1182.

Pukkila, P. J., Peterson, J., Herman, G., Modrich, P., & Meselson, M. (1983). Effects of high levels of DNA adenine methylation on methyl-directed mismatch repair in *Escherichia coli*. *Genetics, 104*(4), 571-582.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., . . . Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics, 13*(1), 341.

Quinlan, A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current Protocols in Bioinformatics*, *47*(1), 11.12.1-11.12.34.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842.

Quiver FAQ. (2013). Retrieved 28 April 2017, from http://bioinfo-master.ird.fr:8080/smrtanalysis/doc/bioinformatics-tools/GenomicConsensus/doc/QuiverFAQ.html#quiver-faq

Rao, D. N., Dryden, D. T. F., & Bheemanaik, S. (2013). Type III restriction-modification enzymes: a historical perspective. *Nucleic Acids Research*, *42*(1), 45-55.

Reisenauer, A., Kahng, L. S., McCollum, S., & Shapiro, L. (1999). Bacterial DNA methylation: a cell cycle regulator? *Journal of Bacteriology, 181*(17), 5135-5139.

Reisenauer, A., & Shapiro, L. (2002). DNA methylation affects the cell cycle transcription of the CtrA global regulator in *Caulobacter*. *The EMBO Journal, 21*(18), 4969-4977.

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics, 13*(5), 278-289.

Riadi, G., Medina-Moenne, C., & Holmes, D. S. (2012). TnpPred: A web service for the robust prediction of prokaryotic transposases. *Comparative and Functional Genomics, 2012*, 678761.

Ritz, A., Bashir, A., Sindi, S., Hsu, D., Hajirasouliha, I., & Raphael, B. J. (2014). Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics, 30*(24), 3458-3466.

Roberts, D., Hoopes, B. C., McClure, W. R., & Kleckner, N. (1985). IS10 transposition is regulated by DNA adenine methylation. *Cell, 43*(1), 117-130.

Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickle, T. A., Bitinaite, J., . . . Dybvig, K. (2003). A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Research, 31*(7), 1805-1812.

Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biology, 14*(6), 405.

Roberts, R. J., Vincze, T., Posfai, J., & Macelis, D. (2003). REBASE: restriction enzymes and methyltransferases. *Nucleic Acids Research, 31*(1), 418-420.

Roberts, R. J., Vincze, T., Posfai, J., & Macelis, D. (2010). REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research, 38*(Database issue), D234-236.

Roberts, R. J., Vincze, T., Posfai, J., & Macelis, D. (2015). REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research, 43*(D1), D298-D299.

Robertson, G. T., Reisenauer, A., Wright, R., Jensen, R. B., Jensen, A., Shapiro, L., & Roop, R. M. (2000). The *Brucella abortus* CcrM DNA Methyltransferase Is Essential for Viability, and Its Overexpression Attenuates Intracellular Replication in Murine Macrophages. *Journal of Bacteriology, 182*(12), 3482-3489.

Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., . . . Koonin, E. V. (2002). Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Research, 30*(10), 2212-2223.

Sahin, N., Tani, A., Kotan, R., Sedláček, I., Kimbara, K., & Tamer, A. U. (2011). *Pandoraea oxalativorans* sp. nov., *Pandoraea faecigallinarum* sp. nov. and *Pandoraea vervacti* sp. nov., isolated from oxalate-enriched culture. *International Journal of Systematic and Evolutionary Microbiology, 61*(9), 2247-2253.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*(4), 406-425.

Sampson, B. A., Misra, R., & Benson, S. A. (1989). Identification and characterization of a new gene of Escherichia coli K-12 involved in outer membrane permeability. *Genetics, 122*(3), 491-501.

Sanchez-Romero, M. A., Cota, I., & Casadesus, J. (2015). DNA methylation in bacteria: from the methyl group to the methylome. *Current Opinion in Microbiology, 25*, 9-16.

Sater, M. R. A., Lamelas, A., Wang, G., Clark, T. A., Röltgen, K., Mane, S., . . . Schmid, C. D. (2015). DNA Methylation Assessed by SMRT Sequencing Is Linked to Mutations in *Neisseria meningitidis* Isolates. *PLoS ONE, 10*(12), e0144612.

Schadt, E., Banerjee, O., Fang, G., Feng, Z., Wong, W., Zhang, X., . . . Kasarskis, A. (2012). Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Research, 23*(1), 129-141.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics, 19*(R2), R227-240.

Schlagman, S. L., Hattman, S., & Marinus, M. G. (1986). Direct role of the *Escherichia coli* Dam DNA methyltransferase in methylation-directed mismatch repair. *Journal of Bacteriology, 165*(3), 896-900.

Schluckebier, G., O'Gara, M., Saenger, W., & Cheng, X. (1995). Universal catalytic domain structure of AdoMet-dependent methyltransferases. *Journal of Molecular Biology, 247*(1), 16-20.

Schneider, I., Queenan, A. M., & Bauernfeind, A. (2006). Novel carbapenem-hydrolyzing oxacillinase OXA-62 from *Pandoraea pnomenusa*. *Antimicrobial Agents and Chemotherapy, 50*(4), 1330-1335.

Seshasayee, A. S., Singh, P., & Krishna, S. (2012). Context-dependent conservation of DNA methyltransferases in bacteria. *Nucleic Acids Research, 40*(15), 7066-7073.

Shin, S. C., Ahn, D. H., Kim, S. J., Lee, H., Oh, T.-J., Lee, J. E., & Park, H. (2013). Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. *PLoS ONE, 8*(7), e68824.

Siddique, T., Okeke, B. C., Arshad, M., & Frankenberger, W. T., Jr. (2003). Enrichment and isolation of endosulfan-degrading microorganisms. *Journal of Environmental Quality, 32*(1), 47-54.

Slater, S., Wold, S., Lu, M., Boye, E., Skarstad, K., & Kleckner, N. (1995). *E. coli* SeqA protein binds *oriC* in two different methyl-modulated reactions appropriate to its roles in DNA replication initiation and origin sequestration. *Cell, 82*(6), 927-936.

Stephens, C., Reisenauer, A., Wright, R., & Shapiro, L. (1996). A cell cycle-regulated bacterial DNA methyltransferase is essential for viability. *Proceedings of the National Academy of Sciences of the United States of America, 93*(3), 1210-1214.

Sternberg, N. (1985). Evidence that adenine methylation influences DNA-protein interactions in *Escherichia coli*. *Journal of Bacteriology, 164*(1), 490-493.

Stryjewski, M. E., LiPuma, J. J., Messier, J. R. H., Reller, L. B., & Alexander, B. D. (2003). Sepsis, Multiple Organ Failure, and Death Due to *Pandoraea pnomenusa* Infection after Lung Transplantation. *Journal of Clinical Microbiology, 41*(5), 2255-2257.

Sullivan, M. J., Ben Zakour, N. L., Forde, B. M., Stanton-Cook, M., & Beatson, S. A. (2015). Contiguity: Contig adjacency graph construction and visualisation. *PeerJ PrePrints, 3*, e1273.

Suzuki, Y., Korlach, J., Turner, S. W., Tsukahara, T., Taniguchi, J., Qu, W., . . . Morishita, S. (2016). AgIn: measuring the landscape of CpG methylation of individual repetitive elements. *Bioinformatics, 32*(19), 2911-2919.

Takahashi, N., Naito, Y., Handa, N., & Kobayashi, I. (2002). A DNA Methyltransferase Can Protect the Genome from Postdisturbance Attack by a Restriction-Modification Gene Complex. *Journal of Bacteriology, 184*(22), 6100-6108.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution, 30*(12), 2725-2729.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Ciufo, S., & Li, W. (2013). Prokaryotic genome annotation pipeline. In The NCBI Handbook (2nd ed.) (Prokaryotes). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK143764/toc/?report=reader

Tersteegen, A., Linder, D., Thauer, R. K., & Hedderich, R. (1997). Structures and functions of four anabolic 2-oxoacid oxidoreductases in *Methanobacterium thermoautotrophicum*. *European Journal of Biochemistry, 244*(3), 862-868.

Thompson, C. C., Chimetto, L., Edwards, R. A., Swings, J., Stackebrandt, E., & Thompson, F. L. (2013). Microbial genomic taxonomy. *BMC Genomics, 14*(1), 913.

Tock, M. R., & Dryden, D. T. (2005). The biology of restriction and anti-restriction. *Current Opinion in Microbiology, 8*(4), 466-472.

Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S., & Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research, 38*(15), e159.

Urig, S., Gowher, H., Hermann, A., Beck, C., Fatemi, M., Humeny, A., & Jeltsch, A. (2002). The *Escherichia coli* dam DNA methyltransferase modifies DNA in a highly processive reaction. *Journal of Molecular Biology, 319*(5), 1085-1096.

Van der Woude, M. W., & Low, D. A. (1994). Leucine-responsive regulatory protein and deoxyadenosine methylase control the phase variation and expression of the *sfa* and *daa* pili operons in *Escherichia coli*. *Molecular Microbiology, 11*(4), 605-618.

Vasu, K., & Nagaraja, V. (2013). Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiology and Molecular Biology Reviews, 77*(1), 53-72.

Von Freiesleben, U., Rasmussen, K. V., & Schaechter, M. (1994). SeqA limits DnaA activity in replication from *oriC* in *Escherichia coli*. *Molecular Microbiology, 14*(4), 763-772.

Watson, M. E., Jr., Jarisch, J., & Smith, A. L. (2004). Inactivation of deoxyadenosine methyltransferase (dam) attenuates *Haemophilus influenzae* virulence. *Molecular Microbiology, 53*(2), 651-664.

Weyand, N. J., & Low, D. A. (2000). Regulation of Pap phase variation. Lrp is sufficient for the establishment of the phase off *pap* DNA methylation pattern and repression of *pap* transcription *in vitro*. *Journal of Biological Chemistry, 275*(5), 3192-3200.

Wilson, G. G. (1988). Type II restriction--modification systems. *Trends in Genetics, 4*(11), 314-318.

Wilson, G. G. (1991). Organization of restriction-modification systems. *Nucleic Acids Research, 19*(10), 2539-2566.

Wilson, G. G., & Murray, N. E. (1991). Restriction and modification systems. *Annual Review of Genetics, 25*, 585-627.

Wion, D., & Casadesus, J. (2006). N6-methyl-adenine: an epigenetic signal for DNA-protein interactions. *Nature Reviews Microbiology, 4*(3), 183-192.

Wong, K. R., Hughes, C., & Koronakis, V. (1998). A gene, *yaeQ*, that suppresses reduced operon expression caused by mutations in the transcription elongation gene *rfaH* in *Escherichia coli* and *Salmonella typhimurium*. *Molecular and General Genetics, 257*(6), 693-696.

Wright, G. D., Berghuis, A. M., & Mobashery, S. (1998). Aminoglycoside antibiotics. Structures, functions, and resistance. *Advances in Experimental Medicine and Biology, 456*, 27-69.

Wright, R., Stephens, C., & Shapiro, L. (1997). The CcrM DNA methyltransferase is widespread in the alpha subdivision of proteobacteria, and its essential functions are conserved in *Rhizobium meliloti* and *Caulobacter crescentus*. *Journal of Bacteriology, 179*(18), 5869-5877.

Wu, J. C., & Santi, D. V. (1987). Kinetic and catalytic mechanism of HhaI methyltransferase. *The Journal of Biological Chemistry, 262*(10), 4778-4786.

Wu, T. P., Wang, T., Seetin, M. G., Lai, Y., Zhu, S., Lin, K., . . . Xiao, A. Z. (2016). DNA methylation on N^6-adenine in mammalian embryonic stem cells. *Nature, 532*(7599), 329-333.

Xie, G., Keyhani, N. O., Bonner, C. A., & Jensen, R. A. (2003). Ancient origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiology and Molecular Biology Reviews, 67*(3), 303-342.

Yin, J. C., Krebs, M. P., & Reznikoff, W. S. (1988). Effect of *dam* methylation on Tn5 transposition. *Journal of Molecular Biology, 199*(1), 35-45.

Yong, D., Ee, R., Lim, Y.-L., Yu, C.-Y., Ang, G.-Y., How, K.-Y., . . . Chan, K.-G. (2016). Complete genome sequence of *Pandoraea thiooxydans* DSM 25325^T, a thiosulfate-oxidizing bacterium. *Journal of Biotechnology, 217*, 51-52.

Zhang, D. F., Jiang, B., Xiang, Z. M., & Wang, S. Y. (2008). Functional characterisation of altered outer membrane proteins for tetracycline resistance in *Escherichia coli*. *International Journal of Antimicrobial Agents, 32*(4), 315-319.

Zhang, W., Du, P., Zheng, H., Yu, W., Wan, L., & Chen, C. (2014). Whole-genome sequence comparison as a method for improving bacterial species definition. *The Journal of General and Applied Microbiology, 60*(2), 75-78.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., & Wishart, D. S. (2011). PHAST: A Fast Phage Search Tool. *Nucleic Acids Research*, *39*(Web Server issue), W347-W352.

Zhu, L., Zhong, J., Jia, X., Liu, G., Kang, Y., Dong, M., . . . Chen, F. (2015). Precision methylome characterization of *Mycobacterium tuberculosis* complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. *Nucleic Acids Research, 44*(2), 730-743.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

A LIST OF PUBLICATIONS

Lim, Y.-L., Ee, R., How, K.-Y., Lee, S.-K., Yong, D., Tee, K. K., . . . Chan, K.-G. (2015). Complete genome sequencing of *Pandoraea pnomenusa* RB38 and Molecular Characterization of Its *N*-acyl homoserine lactone synthase gene *ppnI*. *PeerJ, 3*, e1225.

Lim, Y.-L., Ee, R., Yong, D., Tee, K.-K., Yin, W.-F., & Chan, K.-G. (2015). Complete genome of *Pandoraea pnomenusa* RB-38, an oxalotrophic bacterium isolated from municipal solid waste landfill site. *Journal of Biotechnology, 214*, 83-84.

Lim, Y.-L., Ee, R., Yong, D., Yu, C.-Y., Ang, G.-Y., Tee, K.-K., . . . Chan, K.-G. (2016). Complete Genome Sequence Analysis of *Pandoraea pnomenusa* Type Strain DSM 16536^{T} Isolated from a Cystic Fibrosis Patient. *Frontiers in Microbiology, 7*.

Chan, K.-G., Yong, D., Ee, R., **Lim, Y.-L.**, Yu, C.-Y., Tee, K.-K., . . . Ang, G.-Y. (2016). Complete genome sequence of *Pandoraea oxalativorans* DSM 23570T, an oxalate metabolizing soil bacterium. *Journal of Biotechnology, 219*, 124-125.

Ee, R., **Lim, Y.-L.**, Yin, W.-F., & Chan, K.-G. (2014). *De novo* assembly of the quorum-sensing *Pandoraea* sp. strain RB-44 complete genome sequence using PacBio single-molecule real-time sequencing technology. *Genome Announcements, 2*(2), e00245-00214.

Ee, R., Yong, D., **Lim, Y.-L.**, Yin, W.-F., & Chan, K.-G. (2015). Complete genome sequence of oxalate-degrading bacterium *Pandoraea vervacti* DSM 23571^{T}. *Journal of Biotechnology, 204*, 5-6.

Yong, D., Ee, R., **Lim, Y.-L.**, Yu, C.-Y., Ang, G.-Y., How, K.-Y., . . . Chan, K.-G. (2016). Complete genome sequence of *Pandoraea thiooxydans* DSM 25325^{T}, a thiosulfate-oxidizing bacterium. *Journal of Biotechnology, 217*, 51-52.

B LIST OF PRESENTATIONS

Lim, Yan Lue, & Chan, K. G. (2014, July). "*Aeromonas caviae*: Can they talk?" Oral presentation in Monash Science Symposium 2014 which took place at Monash University, Selangor, Malaysia.

Lim, Yan Lue, & Chan, K. G. (2014, July). "The language of *Serratia fonticola*-A multilingual bacteria" oral presentation in UTAR National Postgraduate Fundamental and Applied Science Seminar 2014 (UTAR NPFASS 2014) which took place at UTAR Perak Campus, Kampar, Malaysia. (Awarded with Best Oral Presenter Award)