THE GENOME AND RECOMBINANT CAPSID PROTEIN OF A NEWLY ISOLATED *Escherichia* PHAGE YD-2008.s

DHARMELA SELLVAM

UNIVERSITI SAINS MALAYSIA

2018

THE GENOME AND RECOMBINANT CAPSID PROTEIN OF A NEWLY ISOLATED *Escherichia* PHAGE YD-2008.s

by

DHARMELA SELLVAM

**Thesis submitted in fulfillment of the requirements
for the degree of
Doctor of Philosophy**

June 2018

ACKNOWLEDGEMENT

Undertaking this PhD study has been a truly challenging experience for me and I am thanking GOD for giving me good health throughout this journey. Foremost, I would like to express my sincere gratitude to my supervisor Associate Professor Dr. Yahya Mat Arip for the continuous support for my PhD study and related research, for his motivation, encouragement, patience, and immense knowledge. I could not have imagined having a better supervisor and mentor to complete my study successfully on time. His lessons always will be a good guidance for my future undertakings.

I would also like to express my appreciation to Ministry of Higher Education under the program of My PhD for the financial assistance throughout my PhD study. Besides, many thanks to Electron Microscope Unit staffs for their patience and guidance in viewing my phage samples. I would like to thank all the staffs from School of Biological Sciences, USM and Institute of Postgraduate Studies for their assistance.

A very special thank you to all my lab-mates from Lab 218 for stimulating discussion and for the fun we had in the last four years. Especially to Ira, Anne, Zatil, Jia Wei, Ikin, Jana and Anjali for always being so supportive to me in numerous stages when I needed the most.

Finally, my heartfelt gratitude goes to my Appa (Mr.Sellvam), Amma (Mrs. Dhanavalli), beloved husband Shangkar (my soul mate) and brothers for their care and always believing in me and encouraging me to follow my dreams. My lovely kids Tharunessh, Shrestaa and Vishvak.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

LIST OF GRAPHS

LIST OF ABBREVIATIONS

A	Adenine
ATCC	American Type Culture Collection
APS	Ammonium sulfate phosphate
bp	base pair
BLAST	Basic Local Alignment Search Tool
C	Cytosine
ddH_2O	Double distilled water
DNA	Deoxyribonucleic acid
DNase 1	Deoxyribonuclease
dATP	Deoxyadenosine triphosphate
dNTP	Deoxyribonucleotide triphosphate
E.coli	*Escherichia coli*
EDTA	Ethylene diaminetetraacetic acid
G	Guanine
HCl	Hydrochloric acid
ICTV	International Community Taxonomy of Viruses
IPTG	Isopropyl β-D-1-thiogalactopyranoside
kbp	Kilobase pair
KCl	Potassium chloride
KDa	KiloDalton
LB	Luria Bertani
$MgCl_2$	Magnesium Chloride
NaCl	Sodium chloride

NaOH	Sodium hydroxide
NCBI	National Centre of Biotechnology Information
NGS	Next Generation Sequencing
nm	Nanometer
OD	Optical density
PCR	Polymerase Chain Reaction
Pfu	Plaque forming unit
RE	Restriction endonuclease
rpm	Revolutions per minute
RBP	Receptor binding protein
RNA	Ribonucleic acid
RNase A	Ribonuclease A
SDS	Sodium dodecyl sulfate
SDS-Page	Sodium dodecyl sulfate Polyacrylamide gel electrophoresis
T	Thymine
Taq	*thermos aquacticus*
TBE	Tris/Borate/EDTA
TEMED	Tetramethylethylenediamine
TEM	Transmission Electron Microscope
Tris-base	Tris (hydroxymethyl)-aminomethane
tRNA	Transfer ribonucleic acid
v/v	Volume/volume
w/v	Weight/volume

GENOM DAN PROTEIN KAPSID REKOMBINAN DARIPADA FAJ YD-2008.S

ABSTRAK

Bilangan faj yang dipencilkan sehingga kini hanyalah merupakan sebahagian kecil dari keseluruhan populasi faj dalam biosfera. Walaubagaimanapun, dengan pencapaian teknologi baharu seperti penjujukan generasi seterusnya (NGS), maklumat mengenai lebih banyak genom faj baharu dapat dikenalpasti dan daftar di dalam pangkalan data Jawatankuasa Antarabangsa Taksonomi Virus (ICTV). Satu faj berekor yang diberi nama *Escherichia* faj YD-2008.s. telah ditemui dan jujukan genom faj lengkap tersebut telah didaftarkan di dalam pangkalan data Pusat Nasional Maklumat Bioteknologi (NCBI) dengan nombor GenBank (KM896878.1). Analisis penjujukan genom menunjukkan faj YD-2008.s dimiliki oleh famili *Siphoviridae* yang mempunyai genom lurus dsDNA bersaiz 44,613bp dengan kandungan 54.6% G+C. Menggunakan perisian anotasi bioinformatik (RAST). Sejumlah enam puluh dua kerangka bacaan (ORFs) telah dikenal pasti untuk genom faj YD-2008.s. Antara ORFs yang dikenalpasti, dua puluh lapan mengkodkan protein berfungsi. Manakala, tiga puluh dua ORFs diklasifikasi sebagai protein hipotesis dan dua lagi ORF mengkodkan protein yang tidak dikenal pasti. Walaupun, majoriti protein putatif yang dikodkan mempunyai identiti /fungsi yang menunjukkan persamaan asid amino yang tinggi dengan faj daripada genus 'HK578likevirus' dan famili *Siphoviridae*, faj YD-2008.s mempunyai keunikan tersendiri. Sebagai pengesahan lanjut mengenai hasil jujukan penuh genom yang diperolehi dari NGS, gen kapsid faj YD-2008.s telah diekspres dan dianalisa. Data ekspresi gen kapsid faj didapati mengesahkan ketulenan maklumat himpunan genom, anotasi dan organisasi genom faj YD-2008.s yang telah diperolehi. NGS data menyediakan maklumat faj genom dan analisis perbandingan dengan faj lain. Data NGS yang tersedia ada memudahkan pengesahan faj genom YD-2008.s berdasarkan maklumat gen kapsidnya. Kejayaan penghasilan kapsid faj rekombinan yang menyamai kapsid faj jenis liar dengan berdasarkan penjujukkan penuh genom dari NGS memberi pengesahan bahawa gen pada genom faj YD-2008.s terletak pada kedudukan, bingkai dan orientasi yang betul. Penemuan bioinformatik seterusnya menempatkan faj YD-2008.s dalam Genus *Hk578virus* dan diiktiraf sebagai spesies baharu seperti yang direkodkan di dalam laporan ICTV yang kesepuluh tahun 2017 dengan *Id taksonomi*:ICTVonline=20164995. Walau bagaimanapun, keanehan sifat perlekatan faj YD-2008.s yang tergolong di dalam kumpulan spesies baharu dari Genus *Hk578virus* menjadi kepentingan dan keistimewaan kajian ini. Perlekatan faj YD-2008.s ke atas sel perumah adalah melalui kapsid dan bukannya melalui struktur ekor. Penemuan ini merupakan pengetahuan yang berlawanan dengan pengetahuan umum di dalam buku teks mengenai faj dari *Siphoviridae family,* di mana perlekatan faj dari family *Siphoviridae* adalah melalui struktur ekor. Berkemungkinan, *Escherichia* faj YD-2008.s boleh membentuk subfamili baharu kepada famili *Siphoviridae.*

THE GENOME AND RECOMBINANT CAPSID PROTEIN OF A NEWLY ISOLATED *Escherichia* PHAGE YD-2008.s

ABSTRACT

The numbers of discovered phages are just a minute fraction of its population in the biosphere. However, the advancement of genome sequencing technology as NGS (next generation sequencing) provides opportunity to identify more complete phage genomes data that feasible to register them in ICTV (International Committee on Taxonomy of Viruses). A tailed phage species designated as *Escherichia* phage YD-2008.s was discovered and the complete genome was sequenced and deposited in NCBI (National Centre for Biotechnology Information) database with GenBank accession No. KM896878.1. Genomic sequence analyses revealed phage YD-2008.s genome that belongs to *Siphoviridae* phage family poses a linear dsDNA composed of 44,613 base pairs with 54.6% G+C content. Total number of sixty-two open reading frames (ORFs) were identified on phage YD-2008.s full genome, using bioinformatics annotation software: Rapid Annotation using Subsystem Technology (RAST). Among the ORFs, twenty-eight of them code for functional proteins. Thirty-two were classified as hypothetical proteins and there are two unidentified proteins. Even though majority of the coded putative proteins have high amino acids similarities to phages from the genus *Hk578likevirus* of the *Siphoviridae* family, yet phage YD-2008.s stands with its' own distinctiveness. As further verification of full genome sequencing results by NGS, capsid gene of phage YD-2008.s was expressed and analyzed. Thus, gene expression of the capsid gene further confirmed the authenticity of genome assembly, annotation and organization of phage YD-2008.s.

NGS data provides phage genome information and comparative analysis with other known phages. The available NGS data was useful in verification of the isolated phage YD-2008.s based on the capsid gene. The successful production of recombinant phage capsid using NGS full genome sequencing and that resembling the wild type phage capsid gave the assurance that the genes on the genome are located in correct frames and correct orientation. Further bioinformatics findings place phage YD-2008.s into the genus *Hk578virus* and recognized as a new species as reported in the ICTV 10th report (2017) with *taxon ID*:ICTVonline=20164995. Yet, peculiar attachment behaviour of phage YD-2008.s belonging to new species of genus *Hk578virus* made significant and novelty to this study. Phage YD2008.s attached itself onto the host via the capsid protein rather than the tail structure. This is a violation of the textbook knowledge on *Siphoviridae* phages where the attachment was made through the tail structure. Perhaps, *Escherichia* phage YD-2008.s could form another new subfamily of the *Siphoviridae* family.

CHAPTER 1: INTRODUCTION

The focus in mega biodiversity is often merely on flora and fauna, while the microbial communities are left undisturbed. But in reality, the microbial communities are closely related to flora and fauna that cannot be separated from the existence of cellular life (Greenpeace International Report, 2004). Viruses constitute the major component in the microbial communities. Viruses, specifically bacteriophages are the most abundant life form on the earth and play major roles in the ecological balance of the microbial life forms (Forest, 2003; Pedulla *et al.,* 2003). Based on the International Committee on Taxonomy of Viruses (ICTV) report, viruses could be categorized into four groups: animal, plant, bacteria/archaea, and fungal/protists (Mayo and Pringle, 1998; Murphy *et al.*, 2012). Viruses could be found almost everywhere, in fact wherever cellular life could be found, viruses could be there as well (Ackermann, 2001; Ackermann, 2007). Viruses are dependent biological entities that interact with the genetic material of the host cells (Abeles and Pride, 2014).

Bacteriophage is a bacterial virus which has meaning in its' word itself. Where bacteria act as the host and having viruses as their 'tenants'. Bacteriophages or in short, phages are classified as non-pathogenic viruses to human, animals, and plants (Klumpp *et al.*, 2012). Phages are very specific to bacteria or in other words, bacteria are the natural preys for bacteriophages (Lee and Park, 2012; Klumpp and Loessner, 2013). It was estimated that there were approximately 10^{31} phages with more than 100 million of species exist in the biosphere (Sulakvelidze, 2011). A typical ratio of bacteria to phage is 1:10 (Wommack and Colwell, 2000; McNair *et al.*, 2012). Due to its high abundance and host specificity, phages are excellent agents to control bacteria populations; as well as contributing to the evolution and diversity of the prey species (Chang and Kim, 2011; Rodriguez-Valera *et al.,* 2014).

Even though phages are huge in number but the total number of phages that have been discovered and examined under electron microscope are only in the range of 6,200 (Ackermann and Prangishvili, 2012). Out of that number, over 96% of discovered phages are tailed phages grouped under the order *Caudovirale* and thus formed the biggest group in prokaryote viruses (Ackermann, 1998; Ackermann and Prangishvili, 2012; Yu *et al.,* 2016).Tail phages are divided into three major families: *Siphoviridae* (57.3%), *Myvoviridae* (24.8%) and *Podoviridae* (14.2%) (Ackermann and Prangishvili, 2012; Adriaenssens *et al.,* 2014). In earlier days, phages were identified by culturing and isolation in the laboratory in which only their morphology and physiochemical characteristic could be identified (McNair *et al.,* 2012). As years goes on, isolation of new phages became rapid, simplistic and inexpensive yet the numbers of complete genome sequence of each isolated phages are just a minute fraction of its population in the biosphere. Genome sequencing details are vital to understanding the phage diversity as well as their physiological characteristics (Chang and Kim, 2011; Klumpp *et al.,* 2012). Besides, to understand the diversity as well dynamics of bacterial (host) system, detailed understanding of phages is essential (Cannon *et al.,* 2013). Thus, the development of a high throughput; next generation sequencing (NGS) technology in 2005 has given wide opportunities to virologists to identify more of those unknown phages' identities (Forrester and Hall, 2014; Rodriguez-Valera *et al.,* 2014).

Currently, National Centre for Biotechnology Information (NCBI) phage genome database contains 1864 complete sequences of phage genomes (http://www.ncbi.nlm.nih.gov/genome/?term=PHAGE) [accessed on, Aug 2016], compared to the year 2008 where only ~500 phage genomes were deposited in NCBI phage genome database (Hatfull, 2008) Thus, more phage complete genome sequences have been deposited in NCBI GenBank with NGS technology (Rodriguez-Valera *et al.*, 2014). The first master list of virus species was released in the year 2005 by ICTV. Since then, the numbers of viruses as well as phage species listed in ICTV taxonomy database showed a solid increase. According to the ICTV list, only 36 phage species of *Caudovirale* were registered in the year 2005 but in the year 2017, a total number of 954 phage species of *Caudovirale* have been registered (King *et al.*, 2011; ICTV 10th report). This indicates that, advancement in NGS technology enhanced phage genome identification as well as phage diversity.

An isolated phage belonging to *Siphoviridae* family was designated as *Escherichia* phage YD-2008.s (ICTV 10th report). Previous work has described the isolation details, physiochemical properties and partial genome study of *Escherichia* phage YD-2008.s (Sellvam, 2011). Thus, the objectives of present work were:

I) To conduct full-length genome sequencing of *Escherichia phage* YD-2008.s

II) To carry out bioinformatics analysis and full genome annotation

III) To construct recombinant clone carrying recombinant capsid gene

IV) To study attachment behavior of this phage to its host cell

V) To propose new model of phage-host interaction

CHAPTER 2: LITERATURE REVIEW

2.1 History of viruses

The study of virology was established in the 19th century. Recognition to the discovery of this small, nanometre in size "living thing" went to Sir Louis Pasteur. This French bacteriologist declared that rabies was caused by a microbe, which is smaller than bacteria. Once thought to be a poison, he later termed this infectious as virus in 1884 (Raven *et al.,* 2005; Abedon, 2011b; Liu *et al.,* 2012). Following that, the development of porcelain bacterial filter in same year of 1884 by Sir Charles Chamberland gave huge advantages in the discovery of new viruses. The particles that retained within the porcelain filter were bacteria but viruses passed through the filter (Park and Chess, 2008; Rob and Johan, 2012).

At the beginning of 20th century, it was proven that filterable biological entities were different from bacteria and they had the abilities to cause diseases in plants and human as well. Later on, it was revealed that these smallest microbes were even able to attack and lyse bacterial communities. Hence, further research works verified that viruses were complexes of nucleic acids and proteins that were able to replicate only in living cells which acted as their hosts (Lucas and David, 2002; Liu *et al.,* 2012). Viruses are the most abundant life form on earth that are found almost everywhere. Regardless to their origins, viruses have been the great biological success since all living organisms are their potential hosts (Ackermann, 2001). Thus, viruses are dependent biological entities which are able to interact with the genetic materials of host cells (Bandea, 2009; Abeles and Pride, 2014). Their hosts range from prokaryotes to eukaryotes. Viruses are principally species-specific with respect to their hosts and do not cross the species boundaries (Esteban *et al.*, 2008; Liu *et al.*, 2012; Abeles and Pride, 2014).

Generally, size of viruses ranges from 10 to 400 nanometres (nm). Thus, they are also viewed as a type of natural nanobiomaterial (Liu *et al.*, 2012). Due to their size, they are capable of entering their hosts via varieties of routes such as direct contact, ingestion and inhalation and subsequently infecting the cells (Cannon *et al.*, 2013; Abeles and Pride, 2014). This nanometre sized parasites require host cells for replication, since they are only able to reproduce by residing in a host cell. As they gain access to their host cells, they would employ their host cells to manufacture substances needed for their own replication (Harrison, 2007; Cannon *et al.*, 2013). Although viruses are dominant entities in the biosphere, without their host cells, they are unable to produce their own proteins and generate energy or even replicate. Viruses are only active within host cells but externally they are in a dormant state. Therefore, they exist in a world between the living and non-living (Clark and March, 2006; Harrison, 2007).

Since their discovery, these intracellular parasites were defined based on their viral particles properties. These particles are highly specialized structures which are used by viruses for their transmission to the new host cells (Bandea, 2009; Rodriguez-Valera *et al.*, 2014). The viral particles could be made up of single or double-stranded DNA or RNA, and these genetic materials are obligatory for their replication process. These viral hereditary materials are protected within the protein capsid/ shell (Matsuzaki *et al.*, 2005; Harrison, 2015). The general view of virus structure is presented in **Figure 2.1**.

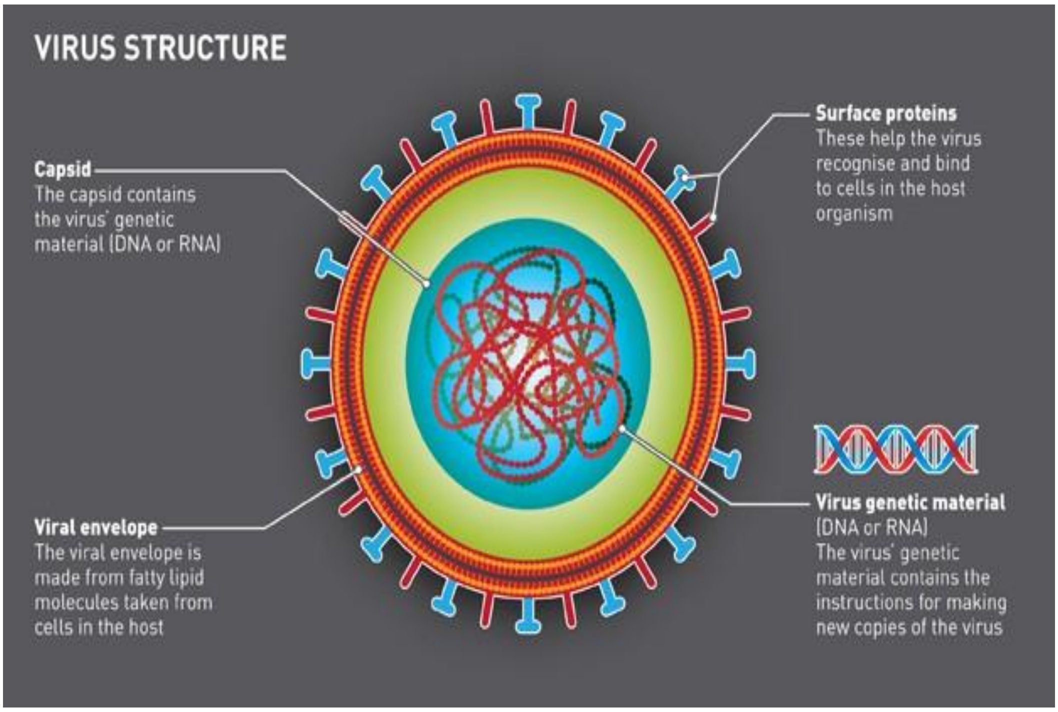

VIRUS STRUCTURE
Capsid
The capsid contains the virus' genetic material (DNA or RNA)
Surface proteins
These help the virus recognise and bind to cells in the host organism
Viral envelope
The viral envelope is made from fatty lipid molecules taken from cells in the host
Virus genetic material
(DNA or RNA)
The virus' genetic material contains the instructions for making new copies of the virus

Figure 2.1: General structure of an enveloped virus. This virus structure diagram briefly explains the general components in the virus particle [Adapted from Elizabeth, 2014].

2.2 Viral life cycle

There are seven steps in the viral life cycle (**Figure 2.2),** which are: i) attachment (adsorption) of the virion to the host cell surface and fusion of virions' envelopes with the cell membrane, ii) penetration of either the viral genome only or whole nucleocapsid into the host cell, iii) uncoating of the viral capsid, iv) viral genome replication, v) assembly or packaging of viral particles, vi) maturation of the completely assembled viral particles and vii) exit or release of the virions from the cell (Ackermann and DuBow, 1987; Summers, 2001). The final releasing process varies between the non-enveloped and enveloped viruses. For most non-enveloped viruses, the exit is simple, in which the cells break open and release the viruses (Birge, 2013). For the enveloped viruses, the release from the host cell is by budding. They acquire host lipid membrane as they bud out through the cell membrane (Mackenzie and Westaway, 2001).

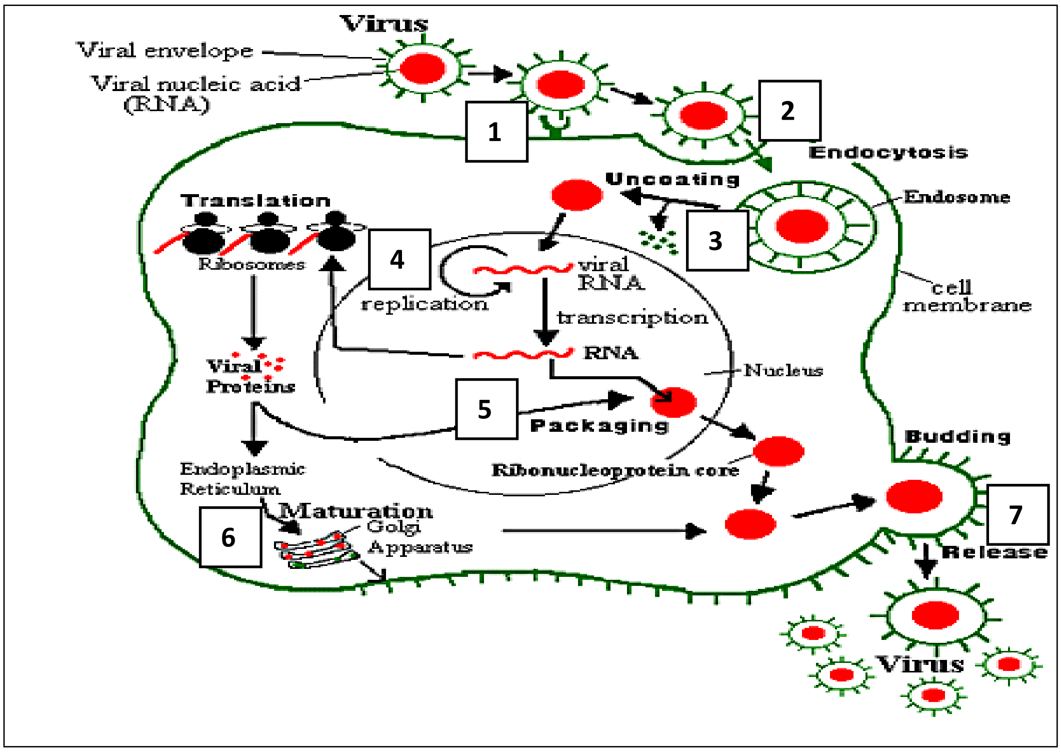


Figure 2.2: The general steps of a viral life cycle. The steps are (1) attachment, (2) penetration, (3) uncoating, (4) replication, (5) assembly/packaging, (6) maturation and (7) release [Adapted from TODAR, 2009].

2.3 Introduction to bacteriophage

Bacteriophages, also known as phages, mean 'devour' or 'to eat' in Greek (Matsuzaki *et al.*, 2005). In the 20th century (1915-1917), Sir Frederick Twort and Sir Felix d’Herelle were the first to introduce the bacteriophage (Twort, 1915; d’Herelle, 1917). Sir d’Herelle noted that when virus suspension was spread on top of bacteria lawn on an agar, clear circular zones (plaque) were formed. Those plaques indicate as bacteriophages (d’Herelle, 1917; Cormier and Janes, 2014). Phages are bacterial viruses that specifically infect and lyse bacterial cells (Matsuzaki *et al.,* 2005; Abedon, 2011b). This group of viruses that could infect the domain of Bacteria contrast with the viruses infecting domains of Archae and Eukarya (Pina *et al.,* 2011; Pietilä *et al.,* 2014). Phages are very specific to bacteria whereby, bacteria are natural preys of bacteriophages. Thus, phages are non-pathogenic viruses to human, animals and plants (Lee and Park, 2012; Klumpp and Loessner, 2013).

Bacteria and their viruses (phages) have coexisted and coevolved for approximately three to five billion years ago (Abeles and Pride, 2014). By nature, bacteria and phages have co-evolutionary relationship; an arm race (Cannon *et al.,* 2013). The survival of phages over billions of years show their ability to overcome the bacterial resistance mechanism by constantly evolving in parallel with their hosts (Klumpp *et al.*, 2012). It is estimated that there are ~10^{30} bacterial cells on the planet and as typical ratios phages tend to be present in ratios of about ten to one bacteria (Stern and Sorek, 2011; McNair *et al.,* 2012). Current estimates suggest that ~ 100 million-phage species exist globally. Thus, bacteriophages are ubiquitous and that make them highly diverse among themselves, as well as other microorganisms (Wommack and Colwell, 2000; Ackermann, 2011; McNair *et al.,* 2012).

2.4 Structural identification of phages

Bacteriophages could be divided into few groups based on their morphology. They could be filamentous, icosahedra with or without tail, rod-shaped, polyhedral, pleomorphic and some are lipid containing enveloped phages (Ackermann, 2007; Orlova, 2012). Phage head /capsids are made of copies of one or more different proteins. Moreover, function of the capsid is to protect the genetic hereditary (nucleic acids) that inside it (Lucas and David, 2002; Matsuzaki *et al.*, 2005). So far, ~ 6,200 bacteriophages have been examined under electron microscopes (EM). But approximately, 96% of those phages examined under EM are tailed bacteriophages (Ackermann, 2011; Ackermann and Prangishvili, 2012)

2.5 Bacterial virus replication system

Even though bacteriophages are under the virus category, their replication pathway slightly differs from plant and animal viruses (Ackermann and DuBow, 1987). In animal viruses there are seven steps, whereas in prokaryotic viruses (bacteriophages and archae, there are only six steps (Summers, 2001). Most of bacteriophages do not have an uncoating step, since only their nucleic acid will penetrate into the host cytoplasm and not the whole nucleocapsid as in eukaryotic viruses (Matsuzaki *et al.*, 2005; Rodriguez-Valera *et al.*, 2014). Phages from *Microviridae*, *Plasmaviridae* and *Cystoviridae* families are exceptional, whereby their mechanisms resemble those utilized by eukaryotic viruses (Krupovič *et al.*, 2010; Jakutyte, 2011). Aside from heredity and structural classification, phages also could be classified based on their life cycles.

Accordingly, bacteriophages will undergo either the lysogenic (temperate phages) or lytic (virulent phages) replication pathway (Clark and March, 2006; Birge, 2013). Temperate phages are less harmful to their host cells as they do not lyse their host cells totally. These phages will live in semi-stable lifestyle as prophage. Eventually the phage genome multiply cooperatively with the host bacteria without destroying it (Birge, 2013). Whereby, the bacterial strains which are able to integrate with the phage genomes are termed as lysogens (Matsuzaki *et al.,* 2005; Kropinski, 2006).

Mostly, lytic cycle ends in death of the host cells. Thus, virulent phages are very harmful to their host cells. As the viral particles entered into the host cytoplasm, they would defeat the host metabolic activities and dominate the host cell. They would initiate the replication process and produce the new progeny of virulent viral particles (Kropinski, 2006; Li *et al.,* 2010; Gan *et al.,* 2013). The life cycle pathways of bacteriophage are shown in **Figure 2.3.**

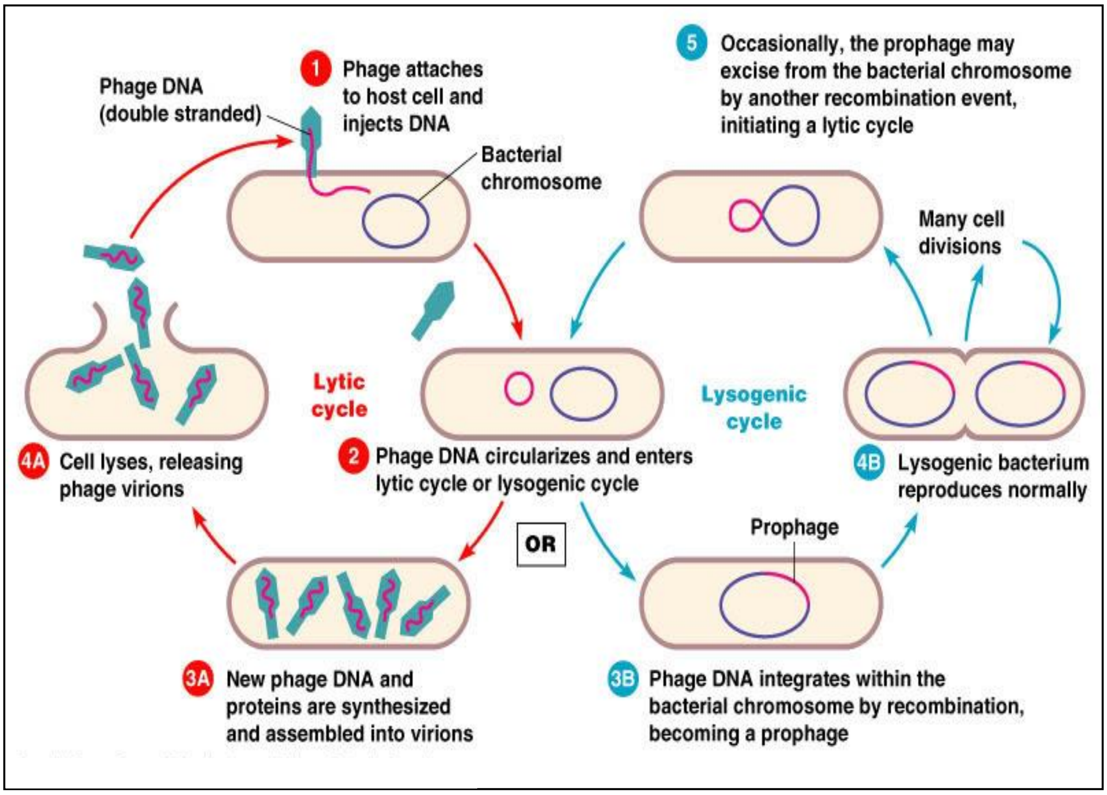

Phage DNA (double stranded)
1 Phage attaches to host cell and injects DNA
Bacterial chromosome
5 Occasionally, the prophage may excise from the bacterial chromosome by another recombination event, initiating a lytic cycle
Many cell divisions
Lytic cycle
Lysogenic cycle
4A Cell lyses, releasing phage virions
2 Phage DNA circularizes and enters lytic cycle or lysogenic cycle
4B Lysogenic bacterium reproduces normally
OR
Prophage
3A New phage DNA and proteins are synthesized and assembled into virions
3B Phage DNA integrates within the bacterial chromosome by recombination, becoming a prophage

Figure 2.3: Lytic and lysogenic pathways in bacteriophage life cycle. This diagram illustrates the different pathways that are involved in bacteriophage life cycle. Bacteriophages have either lytic cycle or lysogenic cycle depending upon the phages [Adapted from Krasner, 2010].

2.6 Bacterial and bacteriophage co-evolution

Bacteria and bacterial viruses (bacteriophages) have coexisted and coevolved for approximately three to five billion years ago, but only recently in 20th century their relation was revealed (Bandea, 2009, Klumpp *et al.,* 2012). The existence of bacteriophage is denoted by the existence of their host bacteria cells. Phages are capable of controlling bacterial population and maintaining colony diversity (Abedon, 2011a; Stern and Sorek, 2011). Many phages only parasitize within certain species and even within a subset of that species (Abedon, 2011a). Phage provides great evolutionary pressure on bacteria by stimulus adaptations, mutations, as well as changing the existing gene pool. As an example; in an *in vitro* study, *Pseudomonas fluorescens* and its phage showed a 10-100 fold increase in mutation rate over 200 bacterial generations compared to bacteria grown in the absence of phage (Brockhurst *et al.*, 2007).

Phages have been a great vehicle in horizontal gene transfer in living organisms especially among various bacterial species even at level of genera (Pedulla *et al.,* 2003; Van Dessel *et al.,* 2005). In most cases, the horizontal gene transfer occur mainly in the temperate phages that undergo lysogenic life pathway as lysogens (Kropinski, 2006; Clark and March, 2006). Although phages use horizontal gene transfer as a tool to continuously develop counter strategies to evade the defences developed by their host bacteria, concurrently, this tool provides equivalent benefits to their host as well (Lee and Park, 2012; Abeles and Pride, 2014). Genes are introduced to bacteria through prophage may offer protection to the host from infection of other phages from similar family (Matsuzaki *et al.*, 2005). Besides this, by introducing virulence genes to the their host, phages offer indirect advantage to enhance survival rate to these bacterial communities. Thus, these assure temperate phages have been powerful symbionts and agents in enhancing genetic diversity in microbial populations and adaptation among bacteria (Abedon, 2011a; Stern and Sorek, 2011; Abeles and Pride, 2014).

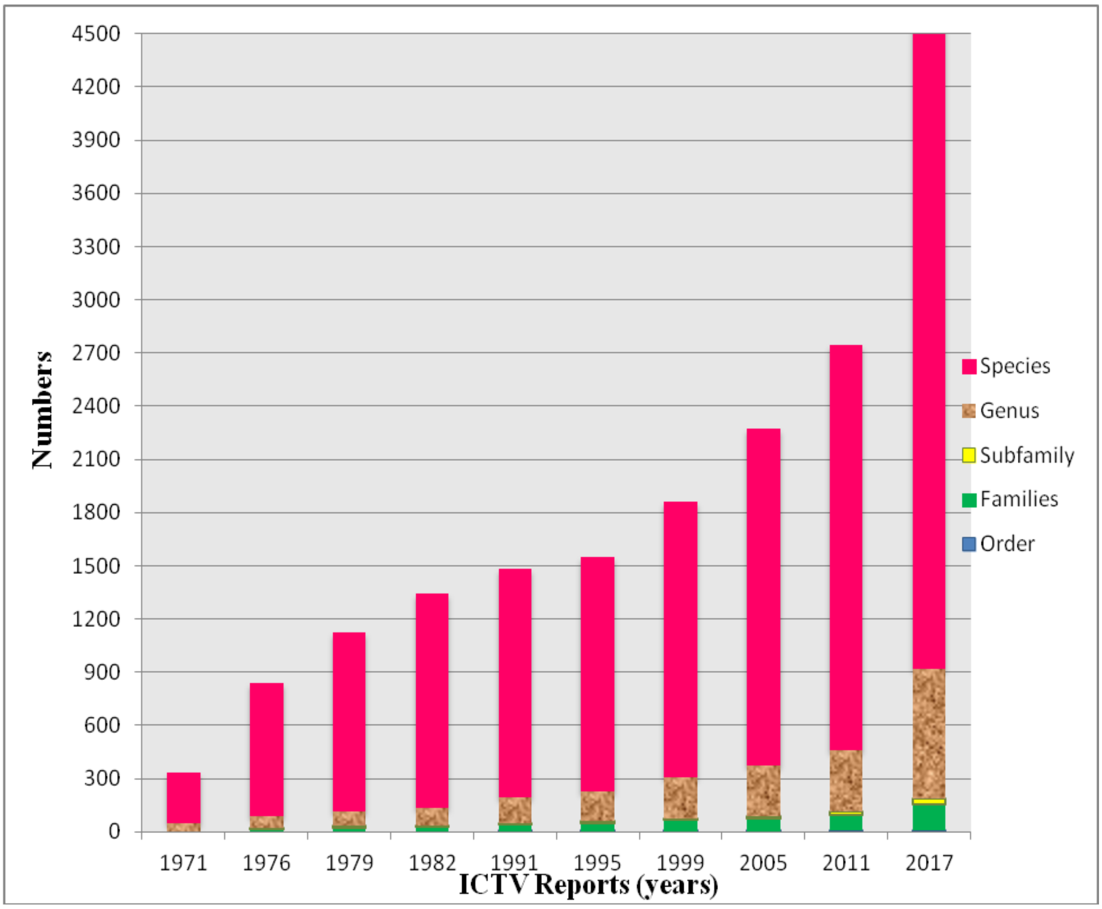
2.7 Virus taxonomy

International Committee on Taxonomy of Viruses (ICTV) is a committee which authorizes and organizes the taxonomic classification of viruses. This committee has been given the full governing authority in terms of viruses' classification according to the universal taxonomic scheme (Mayo and Pringle, 1998; Ackermann, 2011). ICTV has subcommittees for invertebrate, vertebrate, plant, bacteria, protozoa and fungal viruses (Mayo and Pringle, 1998).

The viruses in ICTV hierarchy followed as; order, family, sub-family, genus and species. The first ICTV report was announced in 1971 (Wildy, 1971) , with only two families, 43 genera and 290 species from all group of viruses. In this first report, only six phage genera were identified among the 43 genera. The first six genera were filamentous phage, lambda phage, lipid phage PM2, ribophage group, phix group and *T-even* phages (Mayo and Pringle, 1998; Murphy *et al.*, 2012).

Based on the ICTV reports, the total number of viruses from order to species arrangements showed increment. Till to date viruses had been classified into eight orders compared to the ICTV first report in 1971 with zero order. These registered seven orders are *Bunyavirales* (enveloped and spherical), *Caudovirales* (tailed), *Herpesvirales* (icosahedral and enveloped), *Ligamenvirales* (filamentous with helical nucleocapsid), *Mononegavirales* (helical and enveloped), *Nidovirales* (enveloped), *Picornavirales* (non-enveloped and icosahedral) *and Tymovirales* (isometric and flexuous filamentous).

The total number of species have increased up to 4,404 in the year 2017 (ICTV 10th report) compared to the first report in 1971 with only 290 species. The latest release in year 2017 showed that there are eight orders, 145 families, 35 subfamily, 730 genera/genus and 4,404 species. Besides this, among the 145 families, 84 virus families are still not assigned to any order. This includes all groups of viruses that is registered in ICTV master list update on March 2017, (ICTV Master Species List for 2016). Therefore, without doubt there are many more viruses waiting to be discovered and classified. The taxonomy of viruses according to ICTV reports shown in **Graph 2.1**. The detailed of each ICTV reports on taxonomy attached in **Appendix G**.



Graph 2.1: ICTV reports on virus taxonomy and their respective numbers. This shows number of virus order, families, subfamily, genus and species from the first ICTV report (1971) until the latest 10th ICTV report (2017).

2.7.1 Bacteriophage taxonomy

Globally there are estimated ten million of free-living and eukaryote associated microbial species on earth. If each of these microbe is a host for at least ten different phages, then phage species richness is huge, with a predicted total of 100 million phage species (Forest, 2003; Ackermann, 2011). However, majority of phages and their sequences are still undiscovered in ICTV. According to d'Herelle, there was one phage species with many races (d'Herelle, 1917). In 1967, the first phage survey indicated that there were 111 negatively stained phages. Among the 111 phages, 99 were tailed, nine cubic and three filamentous phages were observed under EM (Eisenstark, 1967; Ackermann and Prangishvili, 2012). Combination of electron microscopic morphology and different forms of nucleic acids are the basic classification scheme that preset by Bradley that is used until today (Ackermann, 2007; Adriaenssens *et al.*, 2014).

Currently, there are two orders (*Caudovirales* and *Ligamenvirales)* with 13 established families for bacteriophages in ICTV database (King *et al.*, 2011; Adriaenssens *et al.*, 2014). However, a huge number of phages are categorized under order of *Caudovirales* with three main families ; *Siphoviridae* (phages with long non-contractile tailed), *Myoviridae* (phages with long contractile tailed) and *Podoviridae* (phages with short contractile tailed) (Ackermann, 1998; Leiman and Shneider, 2012). Besides, phages from family of *Lipothrixviridae* and *Rudiviridae* are assigned to order of *Ligamenvirale* (Ackermann, 2001; Adriaenssens *et al.*, 2014). Hence, the remaining 8 families of phages; *Microviridae, Corticoviridae, Tectiviridae, Leviviridae, Cystoviridae, Iniviridae, Plasmaviridae and Fuselloviridae* are still not assigned to any order (Orlova, 2012; Murphy *et al.*, 2012). The thirteen (13) families of phages are listed in **Figure 2.4**.

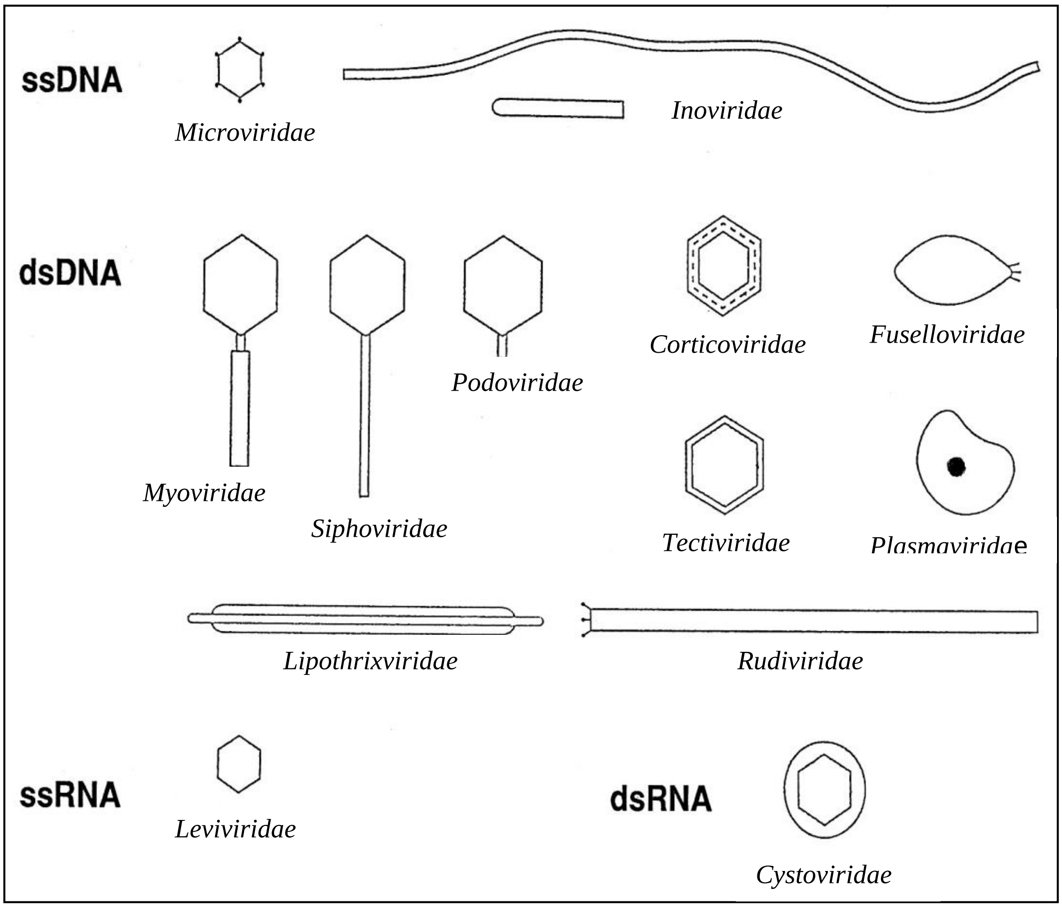


Figure 2.4: Classification of bacteriophages. Thirteen families of bacteriophages grouped according to their nucleic acid details and their structural features as presented in the diagram [Adapted from Ackermann, 2011].

These 13 phage families have their own criteria which groups them under different families. Thus, families that have similar criteria such in morphology, are grouped under one order (Matsuzaki *et al.*, 2005; Ackermann and Prangishvili, 2012; Orlova, 2012). The details of each known phage family under ICTV are illustrated in Table 2.1. There are also important and well studied phages that categorised as unassigned to any order, for example; non tailed and filamentous phage M13. This phage classified under family of *Inoviridae* that composed of ssDNA (Aksyuk and Rossmann, 2011). This non-tailed filamentous phage been an important in early development of gene sequencing as well phage display technology. The robustness of filamentous bacteriophage M13, made it used widely in phage display technology (Forrester and Hall, 2014; Rodriguez-Valera *et al.*, 2014).

Table 2.1: Bacteriophage taxonomy and classification according to ICTV. The nucleic acid and morphology characteristics of each bacteriophage family with respect to their order are listed below. (Ackermann, 2011; King *et al.*, 2011)

Order	Family	Nucleic acid	Morphology characteristics
Caudovirales (constitute ~96% of identified phages)	*Siphoviridae* (61% leading the phage families)	dsDNA	Long non-contractile tails
	Myoviridae (25%)	dsDNA	Contractile tail with a sheath and a central tube
	Podoviridae (14%)	dsDNA	Short tails
Ligamenvirales	*Lipothrixviridae*	dsDNA	Consists of lipo protein envelope and rod-like shape
	Rudiviridae	dsDNA	Straight rigid rods with envelope
Unassigned to any order	*Plasmaviridae*	dsDNA	Filamentous and with lipo protein envelope (known as nucleoprotein granule)
	Fuselloviriidae	dsDNA	Lemon shaped capsid with short spike at one end
	Tectiviridae	dsDNA	Icosahedral capsid that envelope with lipo protein vesicle
	Corticoviridae	dsDNA	Icosahedral capsid enclose lipid bilayer and with spikes
	Inoviridae	ssDNA	Long rigid or flexible filaments with coat structure protein
	Microviridae	ssDNA	Icosahedral capsid without envelope
	Leviviridae	ssRNA	Small icosahedral capsid
	Cystoviridae	Segmented dsRNA	Spherical with envelope double structure capsid

Beginning from the year 2005 up until 2017, there are nine master species lists in ICTV database. In the first master species list, the total numbers of viruses were 1,898 and in the latest release in 2017 which indicates the total number of viruses are 4,404 species (http://ictvonline.org/virusTaxonomy.asp) [Accessed on March 2017]. Therefore, there is an increase of 2,506 species of viruses within 12 years period. Thus in average, there are ~ 20 new virus species isolated and registered under ICTV database per month.

Up until the year 2005, only one order with 37 phage species was reported *Caudovirales*, and in year 2012, a new order *Ligamenvirales* was added with twelve phages species. In 2017 master species list, there were 954 phage species listed in *Caudovirale* (ICTV Master Species List for 2016). The detailed counting of the ICTV master species list from year 2005 until 2017 is tabulated in Table 2.2.

Table 2.2: ICTV Master Species Lists from year 2005 to 2017. This table includes the total number of viruses as well as bacteriophages from the 1st till latest master species list.

Master species list year	Viruses	Bacteriophages	
		Caudovirales	*Ligamenvirales*
2005	1,898	36	-
2008	2,084	36	-
2009	2,284	78	-
2011	2,480	135	-
2012	2,617	161	12
2013	2,827	161	12
2014	3,186	456	12
2015	3,704	655	12
2017	4,404	954	12

2.7.1 (a) *Siphoviridae*

Phages that belong to *Siphoviridae* family are those composed of an icosahedral capsid/head with a long non-contractile tail and mostly are double-stranded DNA (Ackermann, 1998; Orlova, 2012). Among the 665 phage species of *Caudovirale*, approximately 373 species (~57%) belongs to *Siphoviridae* family (Ackermann and Prangishvili, 2012; Mahony and Van Sinderen, 2014). Until end of year 2013, there were 10 genus listed under *Siphoviridae* family. They were *C2likevirus, L5likevirus, Lambdalikevirus, N15likevirus, Phic3unalikevirus, Psimunalikevirus, Spbetalikevirus, T5likevirus, Tunalikevirus* and *Yualikevirus* (Grose and Casjens, 2014; Adriaenssens *et al.*, 2014). However, at the end of year 2014, 39 new genera that consist of 216 phage species were added to *Sipho* family. In this 39 new genera, the phages are grouped by their DNA and protein composition in addition with morphological and physiological criteria (Adriaenssens *et al.*, 2014).

Phage genomes deposited in NCBI Genbank is additional evidence to assure that discovery of new and novel *sipho* phages are vast growing. By mid of 2007, there were only ~ 500 phage genomes deposited in NCBI Genbank (Hatfull, 2008; Savalia *et al.*, 2008). By mid of 2014, ~ 1,200 phage genomes have been registered in NCBI (Adriaenssens *et al.*, 2014). Up to mid of 2016, the numbers of phages genome increased to 1864 species (http://www.ncbi.nlm.nih.gov/genome/?term=phage) [Accessed on Aug 2016]. In this count, more than 50% of the total phage genomes are denoted in *sipho* family, which means, that there are 939 *sipho* phages are deposited in NCBI database (http://www.ncbi.nlm.nih.gov/genome/?term=Siphoviridae)[Accessed on Aug 2016], compared to May 2014 whereby only 300 phage genomes of *sipho* family were deposited (Grose and Casjens, 2014; Adriaenssens *et al.*, 2014).

2.9 Sequencing of phage genome

The first viral genome sequencing took place in 1976 by Sir Walter Fiers, which was the RNA-genome of bacteriophage MS2. The following year, the first DNA-genome of bacteriophage phiX174 was sequenced by Fred Sanger using the chain termination method, also known as Sanger sequencing (Koonin and Galperin, 2003; Hatfull, 2008). Hence, the Sanger sequencing method gave more opportunities to scientists to study entire molecular biology system of an organism instead of focusing on particular gene and protein (Forrester and Hall, 2014). However, more recently Sanger sequencing has been supplanted by the next generation sequencing (NGS) method. This current wave of high throughput method was established in 2005 that facilitates the sequencing of more new genomes in shorter period of time (Klumpp *et al.*, 2012; Rodriguez-Valera *et al.*, 2014).

Isolation and propagation of novel phages are facilitated by phage abundances in the biosphere. However, sequencing of the bacteriophages genomes will improve our knowledge of phage and bacterial evolution as well (Li *et al.*, 2010; Bibby, 2014). The emergence of high-throughput sequencing technology has elucidated how genomic has influenced the field of virology. Furthermore, the understanding on phage molecular biology system has gained interest among virologist (Chang and Kim, 2011; Forrester and Hall, 2014). In microbial communities, there is a rule of thumb whereby, there are 10 phages per cell, but unfortunately the known number of phage genomes are less compared to bacteria (Hatfull, 2008; Rodriguez-Valera *et al.*, 2014). Less than 2,000 phage genomes are available at the NCBI genomic database (http://www.ncbi.nlm.nih.gov/genome/?term=phage) [Accessed on Aug 2016], in