Metagenomic Analysis of the Human Mouth Virus Population and Characterisation of Two Lytic Viruses

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester

By

Ahmed N. Al-Jarbou, MSc

Leicester Medical School

Department of Infection, Immunity and Inflammation

University of Leicester

October, 2008.

TABLE OF CONTENT

ACKNOWLEDGEMENTS

Special thanks to Dr Shaun Heaphy for his supervision and for allowing me the opportunity to work in his lab. He provided me with inspiration for a number of new research ideas that have helped me to become involved with excellent scientific works. Additionally, I would like to thank Professor William Grant; Dr Martha Clocki; Dr Richard Haigh; Dr Jinyu and Dr Christopher Bayliss for their help. I was honoured with the opportunity to interact with them, while I was conducting my research.

Much appreciation goes to all lab members, Andrew Wallace, Dr Eulyen Pagaling, Dr Susan Grant for their help and advice along with all the students whom I have shared the lab, seminars, and scientific meetings with. Special thanks also go to all members of the electron microscopy department and the PNACL lab at the University of Leicester.

I would like to express gratitude to my family back home, my wife, and my daughters for being patient and waiting for me to complete my work.

Finally, I would like to sincerely thank the Al-Qassim University in Saudi Arabia, which has allowed me the opportunity to study abroad, firstly for my master' degree in the United States and later to undertake my PhD degree in the UK. The University has given me a great chance to interact with highly educated individuals within institutions of scientific excellence. I would like also to thank my government for its financial support.

Abstract

Viruses are biological agents that infect cellular organisms. Most viruses are bacteriophages, these are the most abundant biological entities on earth. Not much is known about virus diversity in the human mouth, including dental plaque, compared to other environments. A culture-independent based approach was tried using metagenomic analysis to characterize uncultured virus gene fragments in human dental plaque. The isolated viral genomes were amplified using a multiple displacement amplification method. Eighty, eleven and ten clones were sequenced from three volunteers, respectively. TBLASTX analysis showed that 44% of the sequences had significant identities to the GenBank databases. Of these 66% were viral; 12% human; 10% bacterial; 6% mobile and 6% eukarya. These sequences were sorted into six contigs and forty five single sequences. Four contigs and one single sequence were found to have a significant identity to a small region of a putative prophage in the *Corynebcterium diphtheria* genome. The gaps between these were filled by primer walking and PCR to give a continuous contig of 11554 bp.

Two viruses A1 and A2 and their bacterial host were isolated from the human mouth. The 16S rRNA gene sequence of the host had a 99% identity to several *Neisseria* sp. The A1 virus was found to appear spontaneously on soft top agar plates, and might be a lysogenic virus. The A2 virus was a lytic virus. The two viruses have different morphological shapes. A1 has a varied isometric head size that ranges from 32 to 58 nm and no tail; it may belong to the *Tectiviridae* family. It has a linear dsDNA genome with a size between 12 kb and 23kb. A limited amount of the genome of the A1 virus was sequenced. The A2 virus has an icosohedral head with size of 60±3 nm and a sheathed rigid tail about 175 nm long with no detectable base plate or tail fibres. It can be classified into the order *Caudovirales* family *Siphoviridae*. The size of the A2 virus genome is estimated to be 35 to 40 kb. 31703 bp of unique sequence has been determined and sorted into three contigs and 14 single sequences. Further attempts at gap filling using primer walking and PCR were unsuccessful. It has a linear dsDNA genome, with a GC content of 49 mol%. A latent period of 25 min and a burst size of 25±2 particles were determined by a single step growth curve. Bioinformatic approaches were used to identify ORFs in the genome. A2 virion associated proteins were analysed by SDS–PAGE gel electrophoresis, and some proteins sequences were directly related to the translated genomic sequence.

LIST OF TABLES

LIST OF FIGURES

ABBREVIATION

aa	amino acid
ATP	adenosine triphosphate
BHIB	brain heart infusion broth
BLAST	Basic local alignment search tool
bp	base pair
CFU	colony forming unit
cos	Cohesive
D	direction of translation
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide triphosphate
dH_2O	distilled water
dsDNA	double stranded DNA
DTT	Dithiothreitol
EDTA	ethylenediamietetracetic acid
E. coli	*Escherichia coli*
g	Gravity
HBA	horse blood agar
ICTV	The Universal Database of the International on Taxonomy of Viruses
ID	sequence identity
IPTG	isopropyl β-D thiogalactopyranoside
LB	Luria Bertani
Mr	molecular mass
LC-MS/MS	Liquid Chromatography/Mass Spectrometry/Mass Spectrometry
nH_2O	nanopure water
OD	optical density
ORF	open reading frame
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PEG	polyethylene glycol
PFGE	pulse field gel electrophoresis
PFU	plaque forming units
pI	calculated isoelectric point
RBS	ribosomal binding site
RNA	ribonucleic acid
RNase	Ribonuclease
rRNA	ribosomal RNA
SDS	sodium dodecyl sulphate
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
TAE	tris-acetate EDTA
TE	Tris-EDTA
TEM	transmission electron microscopy
Tm	melting temperature
TEMED	N,N,N',N'-tetramethylethylenediamine
X-Gal	5-Bromo-4-Chloro-3-Indolyl-*β*-d Galactopyranoside

Chapter 1: Introduction and Literature Review

1: Introduction

Viruses are biological agents that infect organisms. They cannot replicate themselves outside their hosts for a variety of reasons e.g., because they lack ribosomes and other initiation factors required for protein synthesis, which they must obtain from their hosts (Sarnow *et al.,* 2005). Viruses are a diverse group of organisms differing greatly in their structure and processes of replication. This has been recently been further confirmed using metagenomic analysis, a method which directly extracts genomic DNA from the environment. Most environmental viruses are phages (viruses of bacteria). At least six viral metagenomic DNA libraries have so far been described in the literature: two from near-shore marine water samples (Breitbart *et al.,* 2004), a human faecal sample (Breitbart *et al.,* 2003), an equine faecal sample (Cann *et al.,* 2005), one from a marine sediment sample (Breitbart *et al.,* 2002) and one from Chesapeake Bay virioplankton (Bench *et al.,* 2007). The total global viral population has been estimated to be around 10^{31} viruses (Edwards and Rohwer, 2005). However, knowledge of the viral diversity in the human mouth is limited.

This work used two methods to study viral diversity in the human mouth. Firstly, culture-independent methods were employed to measure viral diversity; secondly, culture-based methods were used to detect unknown lytic viruses.

1.1: A Historical Perspective of Viruses

The discovery of phages may be traced back to 1896, when Ernest Hankin found that the water of the Ganges and Jumna rivers could destroy cholera bacteria (Parfitt, 2005). In 1901 Emmerich and Low stated that autolysed bacterial cultures caused the lysis of other cultures. It is unclear if these observations were due to the action of bacteriophages, bacteriocins or lytic enzyme production (Kutter and Sulakvelidze, 2005). Fredrick Twort at the Brown Institution of London first independently identified phages in the UK in 1915; he believed that viruses could infect bacteria. Later, in 1917, Felix d'Herelle at the Pasteur Institution in Paris named these viruses 'bacteriophages' which means 'eaters of bacteria' (Duckworth, 1976).

1. 2: The Oral Cavity and Microflora

The oral cavity within the human mouth has a wide range of surfaces providing different microenvironments, including the mucosa, tongue and tooth (Figure 1.1). The mucosa includes gingival, buccal and palatal tissues, which are covered by an epithelium, teeming with microorganisms. The gingival microenvironment, with a low-level of O_2 is inhabited by a large number of Gram-negative anaerobes. The tongue is covered with papillae, providing a sheltered habitat for microbial colonization. The teeth are considered the hardest tissue in the body, and are also colonised by complex bacterial communities (Rogers, 2008).

Briefly, cleaned teeth are coated with salivary proteins called the pellicle, which enables microbial attachment to the tooth surface. The organisms produce a large accumulated mass called dental plaque (Rogers, 2008). The term 'dental plaque' refers to biofilm formed on teeth; however, this term is now used in reference to biofilms on all the oral surfaces (Lamfon *et al.*, 2003). Biofilms consist of complex microbial communities. The most common oral microbial communities' residents are known to be bacteria, virus, fungi and protozoa (Macarthur and Jacques, 2003). Fungi are also common in the mouth, and it estimated that 50% of the population carry harmless fungi in the form of *Candida* species; however, they can also cause opportunistic infections (Cannon *et al.*, 2001).

Protozoa exist in the human mouth, which feed on bacteria and food debris (Wantland *et al.*, 1958). Pathogenic protozoa also exist, for example, the *Entamoeba gingivalis* found in patient with destructive periodontal disease, attacks and destroys both erythrocytes and leukocytes (Lyons *et al.*, 1983). Eukaryotic viruses are present in the human mouth, for example the herpes simplex virus present cold sores (Miller *et al.*, 2005). Examples of viruses found in the mouth during systemic viral infections including the Human Immunodeficiency Virus (HIV), herpes and rabies virus, also, paramyxovirus may present in the mouth.

Figure 1.1: An image of a healthy human mouth.

Figure 1.2: False-coloured scanning electron micrograph of a cavity

1.3: Ecology of Mouth Microbes and Viruses

Microbes from different environments are known to have important roles in global nutrient cycles, e.g. the carbon cycle (Breitbart *et al., 2005*). In clinical environments such as the human mouth, the roles of microbes are likely to be symbiotic, commensal, parasitic or pathogenic. Many of the bacteria in the mouth are anaerobes, particularly those inhabiting the gap between the gums and the teeth. Bacteria in the mouth can cause diseases which require treatment, such as inflammation of the gums, known as gingivitis, which can lead to a serious condition known as periodontal disease, where bone damage will result in the loss of teeth (Figure1. 2) (Spratt, 2008). In addition, the mouth is an important reservoir of bacteria that can cause infections in other sensitive areas of the body. Bacteria can migrate in the blood to the brain or the heart valves. This can also lead to septicaemia (Logan *et al.,* 2006). Individuals who have diets high in sucrose may be at risk of dental caries caused by certain bacteria such as *Streptococcus mutans* (Spratt, 2008).

In the human oral cavity, some bacteria ferment carbohydrates to produce lactic acid and produce antibacterial materials. These bacteria are known to be gram-positive; some of these are *Lactobacillus lactis, Lactococcus lactis, Enterococcus faecalis, Lactobacillus thermophilus, Streptococcus thermophilus* and *Enterococcus durans.* This inhibits the growth of some other harmful bacteria, but not Streptococcus *mutans* (Hegde *et al.,* 2005).

Bacteriophages have been found to have significant influence on bacterial abundance and gene transfer in various microbial environments; however, very little is known about their impact on oral ecology (Bachrach *et al.,* 2003). In addition, the extent of

the diversity of viruses in the human mouth is unknown. The culture-based and culture-independent studies reported here measured the diversity of these communities to give insight into the structure of microbial communities from different environments.

1.4: General Diversity of Bacteria

The diversity of organisms in a sample is a given number that results from measuring the species *richness* and *evenness*. The richness of species is defined as the total number of different species, while evenness refers to the distribution of the richness among the species. The diversity of organisms in an environment depends on the conditions of the environment, the number of species and their distribution in the community. The higher the number of species and the more even their distribution in the sample, the greater is the diversity of the community (Magurran, 2004). Both culture-based and culture-independent methods have been used to assess the microbial communities in different environments, then the diversity has been measured using various statistical approaches to determine a biodiversity index.

Culture-based methods using plates have been in use since the early days of microbiology. This approach allows researchers to determine some features of the microbial physiology and the ecology of an organism. However, culturing all bacteria with standard techniques is difficult (Amann *et al.,* 1995). It has been estimated that 99% of environmental bacteria (Torsvik *et al.,* 1996; Amann *et al.,* 1995) and 50-60% of the flora from the human oral cavity (Krose *et al.,* 1999; Paster *et al.,* 2001; Kumar *et al.,* 2005) could not be cultured using standard techniques. The fact that laboratory media and environmental conditions are different may explain why most microorganisms fail to grow (Keller *et al.,* 2004). The estimation of uncultured bacteria in various environments has therefore become necessary, using culture-independent methods. Firstly, microscopy showed that different cell morphologies were not accounted for by the growth measured using plate culture methods (Roszak *et al.,* 1987; Staley and Konopka, 1985); this was called “the great plate count anomaly” (Staley and Konopka, 1985). The difference between microscopic observation and results from culture methods has driven the development of non-culture methods, including the direct analysis of nucleic acids (Stephen, 2007).

The structure of microbial communities has become clearer since the introduction of nucleic acid analysis. One powerful tool has been the use of 16S ribosomal RNA (rRNA) gene sequence analysis. This gene is found in all bacteria and can be employed to identify and determine the evolutionary relationships among them (Rogers 2008). In this method polymerase chain reaction (PCR) is used to amplify the 16S rRNA gene from extracted environmental DNA without cultivation of the organisms (Giovannoni *et al.,* 1990). The amplified sequence can then be determined and compared to known examples. This showed that most microbes present in a wide variety of different environments were not found in the cultured group (Rappe *et al.*, 1997). The growing database that has been generated by this method shows that the diversity of the microbial world is much larger than had been estimated before the advent of the molecular techniques (Pace 1997; Hugenholtz *et al.,* 1998).

Many environmental and clinical microbial communities have been characterized using culture-independent methods. For example, it is predicted that there are more than 10^{10} bacteria cells in one gram of soil (Torsvik *et al.,* 1996), about 10^{8} bacteria cells in every millilitre of saliva ((Logan *et al.,* 2006) and up to 10^{11} bacteria cells in one gram of human faeces (Suau *et al.,* 1999).

Other culture-independent methods are used to detect, identify and characterise bacteria. This has also improved the understanding of microbial communities. Some of these methods are amplified ribosomal DNA restriction analysis (ARDRA), ribosomal intergenic spacer analysis (RISA), terminal restriction fragment length polymorphism (t-RFLP), random amplified polymorphic DNA (RADP), denaturing gradient gel electrophoresis (DGGE), temperature gradient gel electrophoresis (TGGE) and fluorescent in-situ hybridization (FISH). Most of these techniques require PCR amplification of the target DNA for bacterial analysis. The products of the amplification can then be further characterised using the above techniques, depending on their sequence polymorphism or based on their separation sizes, using gel electrophoresis (Rogers, 2008). However, with the advent of cheap sequencing technologies, the method of choice for determining bacterial diversity is probably 16S rRNA sequencing and increasing direct sequence analysis of isolated nucleic acid.

Since 1987 more than 85 novel bacterial phyla have been discovered using both methods (Achtman and Wagner, 2008). Figure 1.3 shows the number of bacterial and archaeal phyla identified using culture-based and culture-independent methods, which has been greatly increased by the development of the latter.

Figure 1.3: Numbers of phyla among bacteria and archaea since 1987.

1.5: Diversity of Viruses

1.5.1: Culture-based Methods of Measuring Viral Diversity

The culture-based method is applied to both prokaryotic and eukaryotic viruses (Breitbart *et al.,* 2005). In the case of prokaryotic viruses, the plaque assay method is used, resulting in a clear area in the soft top agar due to the lysis from the host. Soft top agar is a semifluid gel giving viruses more flexibility to diffuse and move to attack other bacteria growing nearby, which results in the formation of a clear area called a plaque (Breitbart *et al.,* 2005). The results obtained using this method have proved that environmental viruses are more diverse than their hosts are: one marine bacterium can be infected with at least one or two types of virus, some of which are specific to only one host (Pull *et al.,* 1995; Sullivan *et al.,* 2003). Another study found that *E. coli* can be infected with more than 50 phage types (Rohwer, 2003). These studies strongly proved that phage types are more diverse than their host, probably by a ratio of >10 phages per microbe. Therefore 100 million types of phage may exist, based on the estimation of 10 million free-living and eukaryotic-associated microbial species in the world (Rohwer, 2003).

However, only a few studies have isolated bacteriophages from the oral cavity (Hitch *et al.,* 2004). In one of these studies, when human saliva from 31 donors was screened for the presence of bacteriophages using a wide range of gram positive bacteria, only a bacteriophage specific for *Enterococcus faecalis* was found. It should be noted that the saliva donors had not received any antibiotics for three months before the

collection of the samples (Bachrach *et al.*, 2003). An extensive search for lytic bacteriophages was also conducted on oral cavity samples from 23 volunteers, but the only lytic phage isolated was that for *Pr. mirabilis*, which is not recognized as an inhabitant of the oral cavity (Hitch *et al.*, 2004).

Isolated phages are currently classified by the morphology of phage particles, by nucleic acid type and by resistance to chemical solvents (Ackermann, 2001). Their host range, restriction mapping, hybridisation analysis and genome sequencing further characterize cultured phages. Presently 503 phages have been reported as completely sequenced (www.ncbi.nlm.nih.gov/genomes/static/phg.html). The analysis of cultured phages has extended the information on virus diversity. Single bacteria can be infected with many different phages and whole genome sequences of interested phages can be easily obtained from cultured phages.

1.5.2: Culture-independent Methods of Measuring Viral Diversity

1.5.2.1: Morphology of Viruses Using Electron Microscopy

Since 1959, more than 5500 phages have been examined using electron microscopy (Ackermann, 2007). Electron microscopy has helped to characterise viral communities based on their morphological features, e.g. their capsid diameter and tail length. Viruses have been visualised by electron microscopy from different environments and the results showed high morphological diversity among the viral particles, some of which were from very high temperature water (>80°C) (Prangishvili and Garrett, 2004) and others from sediments (Haring *et al.*, 2005). Electron microscopy analysis has demonstrated that the morphological diversity of phages obtained by culturing methods is very different from that of phages obtained from natural environmental communities (Ackermann, 2001). For example, more filamentous viruses with elongated capsid were found in soil compared to the known viruses that were obtained by cultured phage isolate studies (Williamson *et al.*, 2005). Cultured phage isolates were also found to be larger on average than those found in the environment, indicating that culturing methods produce different sized viruses compared to environmental viruses (Borsheim, 1993).

1.5.2.2: Conserved Gene Studies of Viral Diversity

Eukaryota, bacteria and archaea have marker genes that can be used to identify and characterise them. However, no single gene is sufficient to classify and characterise diverse viral communities. There are some conserved genes among certain related viral taxonomic groups which share similarities (Breitbart *et al., 2005*). A short sequence (592 bp) of the structure g20 gene of the cyanophages was used extensively to test the diversity and genetic richness of the cyanophage communities and other phages in the environment (Dorigo *et al.,* 2004; Short and Suttle, 2005; Wilhelm *et al.,* 2006). These studies suggested that the diversity of the cyanophages is high and that there are genetic relationships between viruses within different environments. Another recent study has designed general primers to target the major capsid protein (MCP) gene of the large dsDNA algal viruses. One aim of this study was to use the MCP gene to determine the genetic diversity of algal viruses in culture and their phylogenetic relationships with marine viral assemblages. A further aim was to determine the diversity of this gene among the viral communities in different environments. The result of amplifying this conserved gene was that nine new additional MCP genes of algal viruses were detected, suggesting that this gene can be useful as a genetic marker in the construction of preliminary phylogenetic trees among the algal viruses (Larsen *et al.,* 2008).

1.5.2.3: Metagenomic Studies of Viral Diversity

Conserved gene studies of viral diversity have not yet given enough information to obtain a complete analysis of the virus groups (Breitbart *et al.,* 2005). Since no single conserved gene has been found for viruses and due to the difficulties of culturing the viral host, metagenomic analysis, which can counter these problems, is used. Metagenomics is a term applied to the direct extraction of all the genomes from an environment, including the sequences and complete analysis of the extracted genomes (Edwards and Rohwer, 2005). The metagenomic study of viruses is a new technique, which began in 2002 with two publications concerning uncultured marine viral communities (Edwards and Rohwer, 2005). At the time of writing (January 2008), seven viral metagenomic libraries have been described in the literature; six of these contain only sequences from double-stranded DNA (dsDNA) viruses (Edwards and Rohwer, 2005; Bench *et al.,* 2007), and at least one study of RNA virus diversity in the human gut (Zhang *et al.,* 2006). The analysis of these libraries showed that about

75% of the sequences were unknown and did not match any gene in the non-redundant GenBank database. Therefore, culture-independent studies of viral diversity have proved that the majority of viral diversity is still unknown (Breitbart *et al.*, 2005). As indicated above, metagenomic analysis has been used to explore virus diversity in many environments; however, there are some difficulties that may be faced by the metagenomic method of obtaining novel virus sequences.

1.5.2.3.1: Methods for Metagenomics Analysis of Viral Genomes

There are some problems associated with viral genomes, which makes cloning difficult. These are the abundance of free DNA and host DNA in samples, low content of viral DNA and the presence of lethal genes such as holins, lysozymes and modified viral DNA (Edwards and Rohwer, 2005). These problems have been at least partially solved by filtration, digestion of the free DNA (Edwards and Rohwer, 2005) and increasing the content of viral genomes by using PCR-based methods (Abulencia *et al.*, 2006). Increasing the viral genomes can be achieved by many techniques, such as sequence-independent single primer amplification (SISPA), linker amplified shotgun library (LASL) (Figure 1.4), arbitrary primed PCR (AP-PCR), random PCR amplification and multiple displacement amplification (MDA) (Figure 2.2). All these methods have greatly increased the discovery of new viruses that have not yet been characterized (Delwart, 2007).

1.5.2.3.2: Purification of Viruses

The majority of phages and eukaryotic viruses are recovered using purification methods; however, some viruses may be lost using this approach because of their size, density and sensitivity to chemicals. Virus particles must be separated from microorganisms and free DNA before the extraction of the viral genomes. Different sizes of filter are applied to clear samples from particles that are larger than virus particles; but however large it is, the filter may stop viruses. The sample is usually filtered through a filter with a larger pore size, such as 0.45 nm or more, then a smaller size, such as 0.2 nm, is used. Free nucleotides (DNA and RNA) can be removed from the sample by treatment with DNase and RNase. Solid samples can be re-suspended in neutral solution before the filtration step. In the case of large volumes or large amounts of solid sample, the proportion of viral particles can be increased by using concentration methods such as polyethylene glycol (PEG) precipitation. The

concentrated viral particles are loaded into a caesium chloride gradient, then centrifuged in an ultracentrifuge at 55,000 rpm. Fractions from the sample are collected and dialysed, then virus genomes are extracted (Sambrook *et al.*, 1989).

If large samples cannot be obtained, an alternative is to use a PCR-based method (Abulencia *et al.*, 2006) such as PEG precipitation, then extraction of the precipitated viral particles, followed by additional amplification methods.

1.5.2.3.3: Amplification and Sequencing of Viral Genomes

Viral metagenomic libraries from different environments can be created using the LASL approach. In this procedure, the total extracted viral genomes are physically and randomly sheared and end-repaired, then dsDNA linkers are ligated to the ends and the fragments are amplified using polymerase. The resulting fragments are then ligated into a vector. The vector with its insert is transformed into competent cells (Breitbart *et al.*, 2002). In the present project an alternative method was employed to make a virus library from the human mouth by using an isothermal MDA kit. The extracted viral genomes were directly amplified using the MDA method (Figure 2.2), the amplified genomes were physically sheared, fragments were end-repaired and inserted into a vector, then the vector and the insert were transformed into competent cells. Figure 1.4 shows the differences between the LASL and MDA methods of creating virus libraries for metagenomic studies.

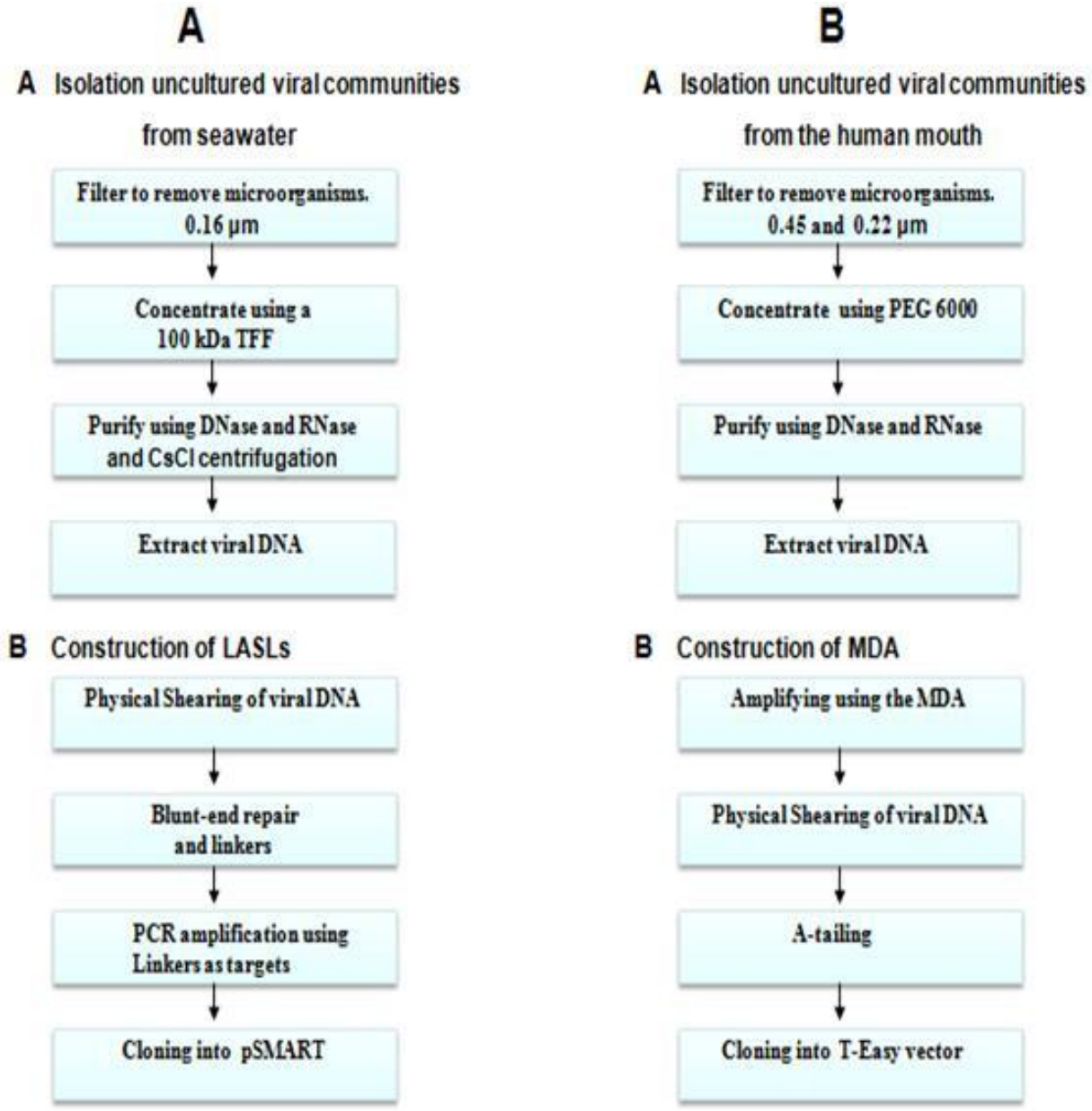


Figure 1.4: Comparison of two methods (LASLs and MDA).
The two methods have been used to access the viral communities. Chart A shows the method of construction of a shotgun library (Edwards and Rohwer 2005) which has been used in several environments. Chart B illustrates the construction of an MDA library, used in this project to amplify viral genomes that were directly extracted from the human mouth.

1.5.2.3.4: Multiple Displacement Amplification

MDA is a technique to amplify whole genomes (Ambrose and Clewley, 2006). It has been used to overcome the limitation of the small amount of DNA from different sample sources (Abulencia *et al.,* 2006). In addition, it has recently been used to amplify a complex DNA library. The results were compared to a DNA library made by using the classic protocol and the differences between the two methods were minimal (Fullwood *et al*., 2008). In addition, several billion-fold of genomic DNA was amplified from a single bacterium using MDA, and a length of 662 bp of the 16S rRNA gene was sequenced from the amplified bacteria (Raghunathan *et al.,* 2005). 20-30 µg of product was also amplified from as few as 1-10 copies of human genomic DNA using MDA. MDA could be employed in the discovery of new species, population and polymorphism analysis, diagnostics and rapid detection of pathogens (Raghunathan *et al.,* 2005).

The MDA method has some disadvantages, one of which is that during its manufacture, background DNA products can be produced, even in the absence of templates. Amplification bias and chimera formation have also been detected using this method (Dean *et al.,* 2001; Lasken *et al.,* 2007), and these problems will be discussed briefly in chapter 3.

The MDA method, which amplifies single or double linear DNA templates, has been also used to amplify circular viral DNA (Tanaka *et al.,* 2001). The MDA method uses the bacteriophage Phi29 DNA polymerase, which has the ability to cause strand displacement and random start points using random primers (Aviel-Ronen *et al.*, 2006). These processes occur without thermal cycling during incubation at 30°C for 16-18 h. The products generated can be over 10 kb in size (Dean *et al.*, 2002). The quantity of product generated by this method is estimated to be 1 µg DNA from 1 ng DNA (protocol). The polymerase has an error rate of 1 in 10^6 to 10^7 nucleotides, when compared with Taq polymerase, which has an error of 3 in 10^4 nucleotides. One of the main advantages of using the MDA method is that in one step a high content of genomic DNA is generated with large fragments (Ambrose and Clewley, 2006).

1.6: The Current State of Bacterial and Archaeal Diversity in the Human Mouth

As described earlier, the human mouth is a complex environment, which has various sites offering different ecological niches. In order to measure bacterial diversity in the human mouth, specimens should be collected from different sites of the mouth, different ages, sexes and individuals having both a healthy and diseased oral status. These very important issues must be taken into account in order to best estimate microbial diversity.

Prior to the advent of independent molecular techniques, culture-based studies did not offer a full picture regarding the diversity of bacterial communities (Roger, 2008). Culture-based methods were only used to detect bacteria that caused or were associated with disease. However, after discovering the independent molecular method the knowledge of the bacteria has been extended to cover and explore the uncultured microbial community structure in various environments and clinical samples. Of these, the 16s rRNA gene sequence have been widely used as a marker to characterise and estimate bacterial diversity in the human mouth (Roger, 2008).

In 1994, The 16s rRNA gene was amplified from a subgingival plaque sample in order to estimate the genetic diversity of cultivable and uncultivable spirochaetes. This was the first clone library analysis of bacteria in the human mouth. The result fell into 23 clusters differing by about 1-2%, displaying unbelievable diversity from a one-patient sample (Choi *et al.,* 1994). A study to characterise and analysis the bacterial diversity of the middle and front of carious dental lesions obtained from five sample using two different methods, cultural-based and the molecular methods (16S rRNA). Table 1.1 represent the names and the frequency numbers of isolation by these methods. In total 95 taxa were detected based on the 16S rRNA gene sequence identity from 496 isolates and 1,577 clones. Of the 95 taxa, only 44 taxa were identified by the molecular method (Wade *et al.,* 2004). Subsequently, many studies have aimed to characterise and determine the diversity amongst oral bacterial communities. Most study samples were collected from specific sites of the mouth, such as the subgingival site (Muyzer *et al.*, 1993; Krose et al, 1999; Paster *et al.*, 2001; Sakamoto *et al.*, 2002; Kumar *et al.,* 2003; Kumar *et al.*, 2005; Kumar *et al.*,

2006). The diversity of bacteria in human subgingival plaque based on these studies (analysis of 2522 16S rRNA gene clones) was estimated to be 347 species falling into 9 bacterial phyla. Thus, the estimated number of species present in the oral cavity is between 415 species and 500 species (if the number of the other oral surfaces such as cheek, tongue and teeth were added) (Paster *et al.*, 2001).

Another extensive study intended to extend the knowledge of bacterial diversity in the human mouth used the same method. Samples from different sites of the mouth were collected from five volunteers, and 2589 clones were amplified and sequenced. The analysis of the sequences demonstrated that more than 700 different bacteria species or phylotypes inhibit the oral cavity of the human mouth. These two studies revealed that over 50% of the bacteria within the human mouth have not been cultured (Aas *et al*., 2005), which indicates that based on the accounted number of cultived oral bacteria (509 taxa) (More and More, 1994) the human mouth may be inhabited by at least 1000 different bacteria (Roger, 2008)

Professor William Wade from King's College London Dental Institute stated that; *"The healthy human mouth is home to a tremendous variety of microbes including viruses, fungi, protozoa and bacteria. The bacteria are the most numerous: there are 100 million in every millilitre of saliva and more than 600 different species in the mouth. Around half of these have yet to be named and we are trying to describe and name the new species"* (Society for General Microbiology, 2008).

All the studies above show that diversity amongst the bacterial communities is vast and more than half is found to be uncultured. The high percentage of cultivable bacteria is due to significant effort which has extended the cultivate oral bacteria (Paster, 2001), which has resulted in the discovery of new species. Recently, new species inhabiting the human mouth found to be associated with various oral diseases and infections in other parts of the body (Downes *et al*., 2008) were discovered and called *Prevotella histicola.* This indicates that the human mouth may still have species that have not yet been discovered, therefore, further efforts are needed in order to give more exact estimations of bacterial diversity in the human mouth.

The presence or detection of the domain *Archaea* from the human mouth is limited, only members such as *Methanobrevibacter oralis*-like species have been isolated from subgingival plaque samples. It is found that archaea are more abundant with periodontal disease, and only detected in patients with severe disease (Leep *et al.*, 2003; Vickerman *et al.*, 2007).

As indicated, the diversity of virus in the human mouth has not yet been fully characterised, and only a few reports describe the isolation of phages from the human mouth. The estimation of viral diversity could be predicted from the estimation of the bacterial diversity, as it has been proven that viruses are more diverse than microbial prey, on average by a ratio of >10 phages per microbe (Rohwer, 2003) and several types. As indicted above there are more than 10^8 bacteria in every ml of saliva within the human mouth, thus, the number of viruses in every ml could be as high as 10^9 consisting of several thousand different types. These would be expected to have a significant effect on the oral bacterial flora.

Table 1.1: Identificantion of bacterial strains isolated by two differents methods

1.7: The Order *Caudovirales* Bacteriophages

The word *Caudovirales* is Latin in origin; 'cauda' means 'a tail' (Kutter and Sulakvelidze, 2005). *Caudovirales* is an order of tailed viruses to which 96% of known bacteriophages belong (Ackermann, 2003).

1.7.1: Taxonomy

The phage classification system started in the 1920s and 1930s, and was based on different host specificities of different phage types. In the 1940s and 1950s, phages were classified using electron microscopy based on their morphotypical features, such as size, length and capsid shapes (Figure 1.5) (Nelson, 2004). During the 1960s new methods were developed which helped to isolate and characterise the nucleic acid of viruses as single stranded DNA (ssDNA), double-stranded DNA (dsDNA), single-stranded RNA (ssRNA) or double-stranded RNA (dsRNA). This classification has improved the taxonomy of phages (Nelson 2004; Ackermann *et al.*, 1978). The order *Caudovirales* consists of three families, *Myoviridae*, *Siphoviridae* and *Podoviridae*, which are all the tailed phages. Other families have different features which have still not been grouped into orders (Table 1.1) (Ackermann, 2007).

Table 1.2: Overview of phage taxonomy

Figure 1.5: The morphological structure of bacteriophages

1.7.2.1: Head (Capsid)

Using electron microscopy, the capsid of the tailed phages appears smooth, thin walled, not enveloped (Lwoff *et al*., 1962) and its diameter varies from 34 to 160 nm. The capsid is composed of protein subunits called capsomeres and is seldom visible (Bradley, 1967). The capsids of tailed phages are found to have either isometric or prolate icosahedral shapes (Xiang *et al.*, 2006). The capsid and tail are connected by a small disc called a connector, which is located inside the capsid at the site of the tail attachment. The connector plays important roles in DNA encapsidation and head assembly (Ackermann, 1998). Within the capsid, DNA is present as tightly packed coils which have no bound proteins.

1.7.2.2: Tail

The tail consists of proteins forming a tubular shape that is connected to the capsid. Its length varies according to family from 10 to 800 nm. In the case of the family *Myoviridae*, tails are found to be long, rigid and contractile, while those of the family *Siphoviridae* are long, flexible and non-contractile; and those of the family *Podoviridae* are found to be short. The tail shafts in these families have six-fold symmetry. In addition, tailed phages can have various numbers of base plates, tail spikes and tail fibres (Ackermann, 1998). These have different roles in infection of bacteria and will be described later (attachment and penetration).

Figure 1.6: Bacteriophage structure

1.7.3: Genomic Structure

The tailed phages of the order *Caudovirales* have linear dsDNA genomes. The size of the DNA in these phages ranges from 17 kb to 700 kb, with an average of about 50 kb. The genetic map is known to be complex and genes that have related functions are found to cluster together (Ackermann, 2003). The head and tail genes are generally separated from each other, so that for example the former come before the latter (Casjens, 2003). It was found that there are about 290 genes in phage T4 and there may be more genes in larger phages (Ackermann 2003; Ackermann, 1998). "*The genome has end redundant sequences. The double stranded DNA may have single-stranded gaps, and have covalently bound terminal proteins that may be linked at both ends. The end of the linear molecule can be blunt, or have complementary protruding 5'– or 3'– ends (cohesive or sticky ends, which can base pair to circularize the molecule). Nucleotide sequences at the 3'–terminus are complementary to similar regions on the 5' end"* (http://www.ncbi.nlm.nih.gov/ICTVdb/ICTVdB/02.htm).

1.7.4: Lifecycle

The complete lifecycle of prokaryotic viruses has many steps that are common to all viruses (Weinbauer, 2004). These include attachment of virus receptors to the surface of the host, penetration of the host cell wall and replication of the viral nucleic acids inside their host.

1.7.4.1: Attachment and Penetration

In order to initiate infection, the tailed phages have to attach to specific receptors on the cell surface of the host. Simply, without a recognition signal between phage and host, attachment cannot be initiated; this is called specificity. Some phages infect more than one host, while others infect just one specific host. When the tail is bound to a specific receptor on the cell surface, the base plate is brought closer to the cell surface. In the case of T4-like phages, the baseplate will provide the energy for infection and will change from a hexagonal to a star shape. When three or more of the long tail fibres fix the baseplate to the cell surface, six short tail fibres will be bound to the lipopolysaccharide (LPS) inner core (Crawford and Goldberg 1980; Riede *et al*., 1985; Montag *et al*., 1987). As the conformation of the base plate is changed, this will cause the tail sheath to contract and push the inner tail tube into the cell

membrane (Kanamaru *et al*., 2004). The peptidoglycan layer of the host is digested by the enzyme lysozyme that is present in the baseplate (Leiman *et al*., 2004).

1.7.4.2: Replication

When the viral genome is inserted into the bacterial host cell, phages undergo either lytic, lysogenic, pseudolysogenic or chronic infections, as shown in Figure 1.7 (Weinbauer, 2004). In the case of the lytic cycle, phages replicate directly without integrating their genome into the host's genome, which results in the destruction of the host. An example of a lytic phage is T4 bacteriophage, which lyses the host a short time after infection, releasing new phage particles (Hadas *et al*., 1997)

In contrast, in the lysogenic cycle, phages either integrate their genomes into the host's genome, or replicate along with its offspring until the lytic cycle is induced. Phages that follow the lysogenic cycle are known as temperate phages; one example is the bacteriophage lambda (λ). The third type of phage life cycle is chronic infection, where after infection phages bud or exit from the host cells without lysis occurring. The best example of this type is the M13 bacteriophage (Weinbauer, 2004).

The final type life cycle is pseudolysogeny, otherwise known as the phage carrier state; the term 'carrier state' is applied to bacteria with a plasmid-like prophage (Bergh, 1989). In the carrier state, after infection, phages can undergo lytic and pseudolysogenic cycles. In the latter case, after infection, phage genomes are not integrated into the host genome, nor do the phage genomes segregate and replicate equally into all progeny cells (Kutter and Sulakvelidze, 2005).

Figure 1.7: Different types of viral life cycle

1.7.5: Lysis

When the viral genome is inserted into host cells, two steps can be taken to release new virions from them. The first is immediate recognition by the host RNA polymerase, which will lead to the transcription of the early genes, which are needed for the replication process of the viral genome and include genes encoding DNA polymerase, ligase, helicase and primase (Mikhailov and Rohrmann, 2002). These genes will protect the viral genome by blocking some of the host proteases and restriction enzymes or destroying some of the host proteins (Kutter and Sulakvelidze, 2005). The second step is the transcription of the late genes, which include those encoding for the assembly of the head and tail. Finally, all tailed phages use two factors for lysis, which are holin and lysin. The former is a protein that creates or forms a hole in the inner membrane of the host cells, providing access for the latter to lyse the peptidoglycan layer and to rupture the cell wall of the host (Kutter and Sulakvelidze, 2005). These very complex steps allow new progeny phages to be released within an hour.

1.8: Single-Step Growth Curve

An understanding of the interaction between viruses and their hosts can be obtained by drawing a single-step growth curve. Simply, the viruses and their host are mixed at low multiplicity of infection in an appropriate growth medium (Ellis and Delbruck, 1939). Samples are then taken at various time points and plated to determine the latent period and burst size. The number of plaques remains constant during the latent period, after which it increases sharply. The burst size is determined by the ratio between the number of plaques obtained before and after lysis (Figure 4.10). The length of the latent period varies according to the phage species, the incubating temperature and the condition of the medium (Kutter and Sulakvelidze, 2005). In addition, the burst size that is calculated from a single cell is known to have different estimations, from a few to 500 phages (Weinbauer, 2004).

1.9: Phage Therapy

Many pathogenic bacteria are resistant to existing antibiotics, which is a serious problem to which a solution needs to be found. Before the development of ampicillin, phages were used as treatment tools. In 1917 Felix d'Herelle was interested in using bacteriophages to treat dysentery after publishing his first paper on them. In the

summer of 1919, he started the treatment at a hospital in Paris, under the clinical supervision of Professor Victor-Henri Hutinel. The first known treatment with phages in humans was in August 1919, where a 12-year-old boy who had severe dysentery producing about 10 to 12 bloody stools per day. After microbiological analysis of the stool samples, d'Herelle administered 2 ml of an anti-dysentery phage preparation. The condition of the boy improved quickly and he had recovered by the next day. In September 1919, three brothers were reported as having bacterial dysentery after their sister had died of the same symptoms. They started to recover within 24 hours of the anti-dysentery phage preparation treatment. The doses of phage preparation ingested by these early volunteers were 100-fold higher than the therapeutic dose, but none of them showed any side effects one day after treatment (Kutter and Sulakvelidze, 2005)

More recently, three men who had ulcerated wounds caused by radiation poisoning became infected with *Staphylococcus aureus*. Treatment of the wounds did not succeed after one month because the bacterium was resistant to antibiotics. A successful treatment occurred with a preparation of biodegradable polymer impregnated with ciprofloxacin, with a mix of bacteriophages. This eliminated the infection and healing of the wounds occurred after seven days of exposure to the polymer (Jikia *et al*., 2005). In addition, bacteriophages have been proved to pass the peripheral blood and migrate to the site of infection (Dabrowska *et al.,* 2005).

In the oral cavity, it is found that infection with *Enterococci,* specifically *E. faecalis,* is restricted to the root canal system of the teeth and is present as a major bacterium when endodontic treatment has failed (Peciuliene *et al.,* 2000; Hancock *et al*., 2001). This bacterium has the ability to survive in extreme environments and is resistant to medication and other irrigants used during endodontic treatment (Siqueira *et al.,* 2000). A study isolating lytic phages from this pathogenic bacterium found them in 22% of human donor saliva. This suggests that phages may play an important role in controlling the outbreak of these and other bacteria, thus protecting the tooth root system (Bachrach *et al.,* 2003).

1.10: THESIS AIMS AND OBJECTIVES

This thesis consists of two parts:

First part is: To measure the viral diversity in the human mouth based on the aims and objectives below:

- Metagenomic analysis of the viral diversity in the human mouth
- Using sensitive nucleic acid amplification method, multiple displacement amplification method (MDA), to overcome the difficulties of cloning and sequencing viral genes.
- Sequencing unknown viral genes fragments

The second part is: Isolate lytic viruses from the human mouth based on the aims and objectives below:

- Detect lytic phages on bacterial lawns
- Determine virus morphological structure by transmission electron microscopy
- Perform a typical single-step growth curve for the isolated lytic virus
- Sequence the genome of lytic virus and annotate and characterise the genes
- Protein analysis of the isolated virus particles using Mass spectral technique

Chapter 2: Materials and Methods

Chapter 2

2.1: Materials

The GenomiPhi DNA Amplification Kit was supplied by Amersham Biosciences, the 1 kb DNA ladders by Invitrogen, the antibiotic ampicillin by Sigma, the QIAEX® II Gel Extraction Kit and the QIAquick PCR Purification Kit by Qiagen and the Wizard® *Plus* SV Minipreps kit by Promega. Another gel extraction kit, called Sephaglas™ BandPrep, was supplied by Amersham Biosciences. Oligonucleotides were synthesised by VH Bio. A nebulizer unit was supplied by Invitrogen, protein molecular weight markers by Fermentas Life Sciences, the restriction enzymes, DNase I, Exonuclease III, T4 DNA ligase and Phusion High-Fidelity DNA Polymerase by New England Biolabs (NEB), *Taq* polymerase by ABgene, Herculase® II Fusion DNA Polymerase by Stratagene, Brilliant Blue G-colloidal by BIO-RAD, the lambda DNA purification kit Vira prep Lambda by Cambio Ltd and a low range PFG marker by NEB. *E. coli* JM109 competent cells and the pGEM T-Easy TA cloning kit were supplied by Promega.

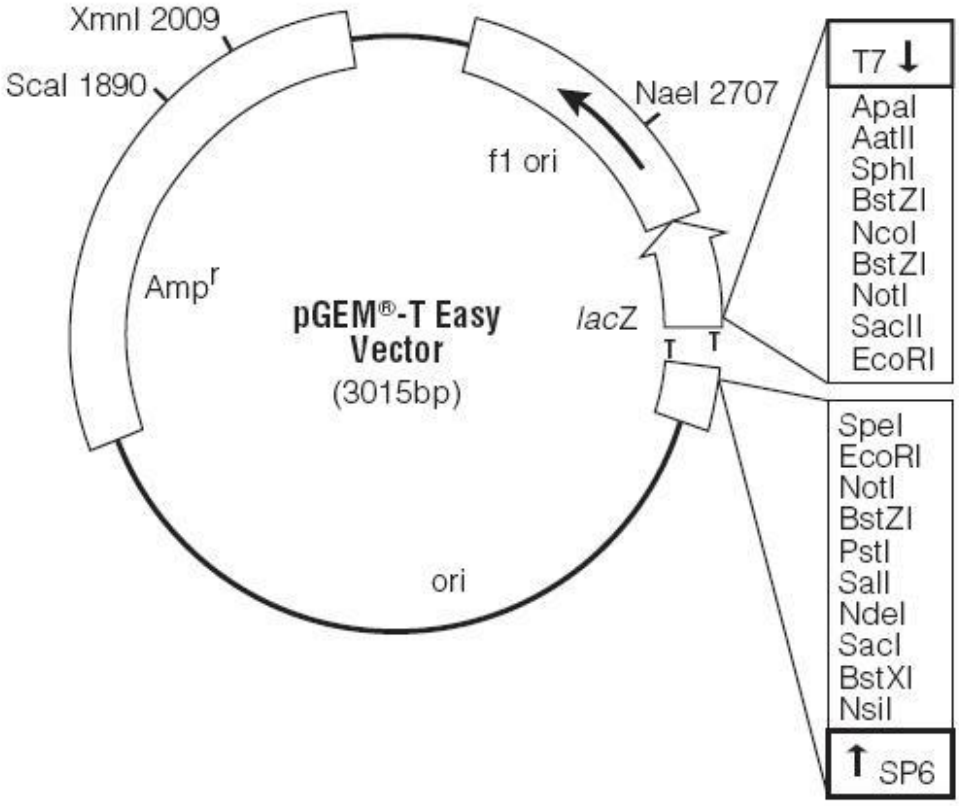


Figure 2.1: pGEM-T Easy Vector circle map and sequence reference points

(Source: Promega).

Table 2.1: PCR primers used to amplify the 16S rRNA gene from the isolated virus host

16S RNA	Sequence (5' to 3')	Reference
Forward	AGA GTT TGA TCC TGG CTC AG	(Weisburg *et al.*, 1991)
Reverse	ACG GHT ACC TTG TTA CGA CTT	

Table 2.2: Vector sequencing primers

Vector Primers	Forward sequence (5' to 3')	Reverse sequence (5' to 3')	Reference
M13F	GTT TTC CCA GTC ACG AC		Promega
M13R		CAG GAA ACA GCT ATG AC	
T7F	TAATACGACTCACTATAGGG		AGOWA
T7R		GCTAGTTATTGCTCAGCGG	

See appendix for the list of virus genome sequences and contig PCR primers.

2.2: Sterilisation

All media, glassware, equipment and toothpicks were sterilised by autoclaving at 121°C for 15 min. Deionised distilled water (dH_2O), produced by reverse osmosis using Elga water purification equipment, was used to prepare all the media and solutions, which were then sterilised by autoclaving at 121°C for 15 min. Nanopure water (nH_2O) from Sigma was used for sensitive work such as polymerase chain reaction (PCR).

2.3: Media

Luria Bertani Broth (LB)

	Per litre
NaCl	10 g
Tryptone	10 g
Bacto-yeast Extract	5 g

The pH was adjusted to 7.0 and the broth autoclaved at 121°C for 15 min.

Luria Bertani Agar (LA)

LB broth with addition of 0.5% w/v agar.

Brain Heart Infusion Broth (BHIB)

Distilled water was added to 37 g brain heart infusion broth to make up to 1 litre. The pH was adjusted to 7.4 and the broth autoclaved at 121°C for 15 min.

Blood agar

Following the manufacturer's instructions, 39 g of Columbia Agar Base was added to 1 L of distilled water. The mixture was boiled until completely dissolved, then sterilised by autoclaving at 121°C for 15 min. The autoclaved medium was cooled to 50°C in a water bath, then 5% horse blood was added and shaken well to mix, and the mixture poured out into plates. These were inverted after the blood agar had solidified and left overnight at room temperature, then stored at 10°C for up to one month.

Soft top agar

0.35% agarose was added to LB broth with the addition of 10 mM $MgSO_4$. This was autoclaved at 121°C for 15 min.

SOC broth

The first four ingredients listed in the table below were combined and dissolved in dH_2O to 95 ml. This was autoclaved at 121°C for 15 min and cooled to room temperature. Then 1 ml of 2M $MgSO_4$ and 1 ml of 2 M glucose were added, and this was made up to a volume of 100 ml with dH_2O. The SOC was then filter-sterilised and stored at -20°C until used.

SOC broth	**per 100 ml**
Tryptone	2 g
Yeast extracts	0.5 g
1 M NaCl	1 ml
1 M KCl	0.25 ml
2 M Mg	1 ml
2 M glucose	1 ml
dH_2O to make up to 100 ml	

2.4: Solutions

Solutions were made up according to the ingredients listed in the tables below

Tris-acetate (TAE) buffer (50x)	**per litre**
Tris	242 g
Glacial acetic acid	57.1 ml
0.5M EDTA (pH 8.0)	100 ml
Distilled water to make up to 1 L	

Phosphate buffer saline (PBS)	**per litre**
NaCl	8 g
Na_2HPO_4	1.44 g
KH_2PO_4	0.24 g
KCl	0.2 g

The PBS was made up to 900 ml with dH_2O and adjusted to Ph 8 before the volume was made up to 1 L with dH_2O. It was then autoclaved at 121°C for 15 min.

SM Buffer	**per litre**
NaCl	5.8 g
$MgSO_4$	2 g
1M tris-HCl (pH 8)	5 ml
Gelatin	2 % (v/v)

The SM buffer was also made up to 1 L with dH_2O and autoclaved at 121°C for 15 min.

Sodium dodecyl sulphate (SDS) (10 %)

dH_2O was added to 100 g SDS to make up to one litre.

2.5: Bacterial Methods

2.5.1: Culturing Organisms from the Human Mouth

Toothpicks and dental floss (Johnson and Johnson, REACH) were used to collect materials attached to the dental plaque and between the teeth in the human mouth. These were first collected from three volunteers for bacterial host isolation. Each sample was mixed with 1 ml of PBS buffer to dissolve dental plaque materials. 10 µl from each mixed sample was added to separate tubes of 10 ml of LB and BHIB. These tubes were incubated aerobically over night at 37°C with shaking at 150 rpm in an orbital shaker. Serial dilutions from overnight growth were streaked out onto blood, LB and BHI agar plates and incubated at 37°C for 24 and 48 h. Colonies were selected according to their colour, morphological features and size. Isolated colonies were picked and plated again to ensure purity of the isolated bacteria.

Next, 20 µl from each mixed sample was streaked out directly into blood and LB agar plates to grow bacteria that might not grow well in the broth media. These were incubated for 24 and 48 h, then colonies were selected and purified as described above.

2.5.2: Quantification of Host Cells

Bacterial cell concentrations were determined using the Miles-Misra assay (Miles *et al*., 1938). Serial dilutions of the host growth in PBS buffer were applied and 50 µl of each dilution was spread out onto blood agar plates, which were inverted and incubated overnight at 37°C. Colonies on the plates were counted and the average was taken. Colony forming units (CFU) per ml were calculated according to the following equation:

CFU/ml = average colony count x 20 x dilution

Plates were prepared for cloning by taking LB as indicated with the addition of 100 mg/ml ampicillin, 20 mg/ml 5-Bromo-4-Chloro-3-Indolyl-*β*-d Glactopyranoside (X-Gal) and 23.8 mg/ml isopropyl β-D thiogalactopyranoside (IPTG).

2.6: Virus Methods

2.6.1: Plaque Assay

The plaque assay is a technique that is used to study viruses for many reasons. It is a clear area in the soft agar that results from the lysis of the host. This happens when the bacterial host grows in the soft agar to form a confluent lawn, then viruses propagate and replicate in the host cells and kill them, which results in a clear area called a plaque. The number of infected particles in a plate can be determined by counting the number of plaques.

2.6.2: Detecting Lytic Phages

LB agar was used as a bottom agar, while LB and BHIB containing 0.35% agarose were used as the soft top agar, which was prepared and kept molten in a water bath at 50°C. For infection, 100 µl of filtered sample was added to 300 µl of the host cell culture that had been grown overnight. The virus particles were allowed to adsorb onto the host cells for 15 min at room temperature, then the infected cells were added to 3 ml of the molten soft top agar in universal tubes and mixed well before being poured onto the bottom agar. This was left to set for a few minutes, then the plates were inverted and incubated at 37°C. After 24 to 48 h, they were checked for the appearance of plaques.

2.6.3: Increasing the Titre of Virus Particles

A single plaque was selected from a lawn and used to infect subsequent cultures to increase the yield of the same virus particles as described. 1 ml of SM buffer was added to plates confluent with virus plaques. The soft top agar was scraped from 15 plates and collected into a 250 ml Sorvall tube. 40 ml of SM buffer was added to the collection tube and mixed well, then incubated overnight at 10°C to allow the virus particles to diffuse from the soft top agar into the buffer. The tube was centrifuged in a Beckman centrifuge at 250 x g for 25 min, then the supernatant was transferred to a

fresh tube. The supernatant was filtered with Millipore filters of sizes 0.45 μm and then 0.22 μm to ensure the removal of agar and cell debris.

2.6.4: Determining Titre of Phages

The soft top agar was used to determine the titre of the phage as plaque forming units (PFU/ml), as described by Adams (1959). Serial dilutions were prepared from phage stock in SM buffer. 0.1 ml of each dilution was added to 0.2 ml of actively growing culture (10^8 CFU/ml) and was added to 3 ml of top soft agar. This was mixed and immediately poured on top of the first layer of LB agar plate. Plates were kept for 20 min at room temperature to allow the agar to solidify. They were then inverted and incubated overnight at 37°C. Plaques were counted as PFU/ml to determine the titre of the phage stock.

PFU/ml = Total number of plaques on plate / d x V

where d is the dilution factor and V is the volume of viruses that is added to the plate.

2.6.5: One Step Growth Curve

The latent period and burst size were determined by the one step growth curve method, as described by Ellis and Delbruck (1939). The cells were infected with phage at low multiplicity of infection (MOI), see below calculation, to ensure that each cell was infected with only one phage. The MOI is defined as the ratio of the number of virus particles to that of bacterial cells. After infection the culture media was diluted at 100-fold to avoid multiple cycles of growth and lysis inhibition.

Basically, a phage stock was added to an overnight cell culture containing 10 mM of $MgSO_4$ and incubated on ice for 10 min. The infected cells were centrifuged at 10,000 x g for 5 min at 4ºC. The supernatant was removed and infected cells were washed twice with PBS buffer, and then pelleted. They were next resuspended in fresh LB broth containing 10 mM $MgSO_4$ to aid adsorption and subsequently incubated in a water bath at 37ºC. Samples were taken every 5 min up to 95 min, and were immediately titred by the plaque assay method. The one step growth experiment was repeated three times to observe if there was any variation in the results.

Calculation of the MOI was used:

Total number of phage was used:

0.01 ml x (12×10^6) / ml = 12×10^4 PFU/ml

Total number of cells was used:

0.5 x (145×10^7) / ml = 725×10^6 CFU/ml

Therefore the MOI was 6042 cells for each phage

2.6.6: Host Range Experiments

Several *Neisseria* sp were obtained as indicated and grown overnight in 5 ml of BHIB at 37ºC in an orbital shaker at 150 rpm. 300 µl of growing cells at the stationary face were infected with the different dilutions of the two isolated lytic viruses. These were incubated for 20 min at 37ºC and mixed with 3 ml of soft top agar of BHI. This mixture was poured onto a BHI agar plate. The plaques were checked as indicated.

2.7: Storage of Isolates

Isolated bacteria were kept on beads at -80°C to be used later, because most of the isolated bacteria could not survive for a long time when they were kept at 4°C or -20 °C. 30% (v/v) of sterile glycerol was added to bacteria grown overnight and stored at -80°C until needed.

2.8: Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (SDS-PAGE)

SDS-PAGE is a method used in protein analysis; SDS is an anionic detergent that has the ability to denature and cover proteins with a negative charge. Two gel layers were used: resolving gel as the bottom layer and stacking gel as the top layer. A 12% acrylamide mix of resolving gel was prepared by mixing the following components in a falcon tube: 32% of dH_2O, 40% (v/v) of 30% acrylamide/bis-acrylamide (37.5:1) solution, 26% (v/v) of 1.5 M Tris-HCl (pH 8.8), 0.1% (w/v) SDS, 0.1% (w/v) ammonium persulphate and 0.1% (v/v) TEMED. The ammonium persulphate and TEMED were added last, then the mixture was poured immediately into the gap between the glass plates and left to polymerise. To avoid air bubbles on the top of the resolving gel and to form a smooth horizontal surface, isopropanol was poured immediately onto the resolving gel. When the gel had set, the isopropanol was poured

off and the top of the resolving gel was washed three times with dH_2O. The remaining dH_2O was dried from the top of the gel, which was left for 30 min at room temperature.

A small space on the top of the resolving gel was left for the 5% stacking gel, which consisted of 68% (v/v) dH_2O, 17% (v/v) of 30% acrylamide mix, 13% (v/v) 1.0 M Tris-HCl, pH 6.8, 0.1% (w/v) SDS, 0.1% (w/v) ammonium persulphate and 0.1% (v/v) TEMED. As indicated, the ammonium persulphate and TEMED were added last. This mixture was quickly poured onto the resolving gel after the insertion of a suitable comb. Once the stacking gel was set, the comb was removed to form wells. The prepared SDS-PAGE gel was placed in the tank and covered with running buffer, consisting of 3.03 g/L Tris base, 14.4 g/L glycine and 1.0 g/L SDS, which were dissolved in dH_2O. The protein samples were loaded and run at 100 mV until they passed the staking gel, then the voltage was reduced to 30 mV when the samples separated in the resolving gel and the samples were left overnight.

2.8.1: SDS-PAGE Staining and Destaining

After electrophoresis, the proteins were fixed for 1 h in a solution of 7% glacial acetic acid in 40% (v/v) methanol. Following the manufacturer's protocol, Brilliant Blue G-Colloidal (Sigma) was used as the protein staining suspension, into which the gel was placed for 1-2 h. The gel was destained with 10% acetic acid in 25% (v/v) methanol for 60 s with shaking, rinsed with 25% methanol (which was discarded), then destained with 25% methanol overnight.

2.8.2: Determining Sizes of Protein Bands

The Marker 12™ Unstained Standard, a protein standard marker, was run in one side of the protein sample to determine the sizes of the bands. This protein standard consists of the following fragments (kDa): 116, 66.2, 40, 35, 25, 18.4 and 14.4. A Kodak EDAS 290 camera was also used with the software provided, which automatically constructed a molecular weight distance calibration curve when the position of the standard was entered, allowing the position of the protein bands of unknown size to be calculated.

2.8.3: Protein Sequence Using LC-MS/MS

Bands of interest were excised from the gel and an in-gel trypsin digest performed (Speicher *et al*., 2000). The bands were destained using 200 mM ammonium bicarbonate/20% acetonitrile, followed by reduction (10 mM dithiothreitol, Melford Laboratories Ltd., Suffolk, UK), alkylation (100 mM iodoacetamide, Sigma, Dorset, UK) and enzymatic digestion (sequencing grade modified porcine trypsin, Promega, Southampton, UK) using an automated digest robot (Multiprobe II Plus EX, Perkin Elmer, UK).

LC-MS/MS was carried out upon each sample using a 4000 Q-Trap mass spectrometer (Applied Biosystems, Warrington, UK). Peptides resulting from in-gel digestion were loaded at high flow rate onto a reverse-phase trapping column (0.3 mm i.d. x 1 mm), containing 5 µm C18 300 Å Acclaim PepMap media (Dionex, UK) and eluted through a reverse-phase capillary column (75µm i.d. x 150 mm) containing Jupiter Proteo 4 µm 90 Å media (Phenomenex, UK) that was self-packed using a high pressure packing device (Proxeon Biosystems, Odense, Denmark). The output from the column was sprayed directly into the nanospray ion source of the 4000 Q-Trap mass spectrometer.

Fragmentation spectra generated by LC-MS/MS were searched using the MASCOT search tool (Matrix Science Ltd., London, UK) against the nucleotide contig sequences supply using appropriate parameters. The criteria for protein identification were based on the manufacturer's definitions (Matrix Science Ltd). Candidate peptides with probability-based Mowse scores exceeding threshold ($p < 0.05$), thus indicating a significant or extensive homology, were referred to as 'hits'. In addition, as the contig sequence was essentially just a very long single sequence, it was necessary to manually apply a peptide ion score cut-off of 40 in order to remove multiple low-scoring peptide matches from the data set.

2.9: Nucleic Acid Isolation

2.9.1: Direct Extraction of Viral Nucleic Acids for Metagenomic Studies

Samples were collected from healthy volunteers. Materials attached to the dental plaques were collected from the mouth by toothpick and dental floss (Johnson and Johnson, REACH). Each sample collected was dissolved in 1 ml of PBS in an Eppendorf tube and mixed well to release the sample from the toothpick and the dental floss. The sample was then vortexed well to release the viral particles from attached materials. The sample was passed through a 0.45 μm filter and then a 0.2 μm filter to free the viral particles from other contaminants like food, blood cells and other cell debris. The viral particles were precipitated by the addition of 1 M NaCl and 10% (w/v) PEG 6000, then incubated for 2 h on ice and pelleted by centrifugation at 16000 x g for 10 min. The supernatant was removed and the pellets were resuspended in 0.5 ml of 10 mM Tris pH 7.5, 10 mM $MgCl_2$ and 100 mM NaCl. The free nucleic acids were digested by adding 10 U of DNase and 10 μg/ml RNase A, then incubated at 37°C for 30 min. Phenol:chloroform was used to extract the viral nucleic acids, as described below.

2.9.1.1: Isothermal Amplification of Extracted Viral Nucleic Acids

The concentration of viral nucleic acids directly extracted from the samples was insufficient to be visualised on agarose gel. Therefore, the viral nucleic acid was amplified using a Genomiphi DNA Amplification kit, which amplifies both single and double stranded linear DNA templates. The concentration generated can be microgram quantities from a nanogram template of starting material after an overnight incubation at 30°C. This kit has three components: sample buffer, reaction buffer and enzyme mix (Phi29 DNA polymerase with random primer). According to the manufacturer's instructions, during amplification at 30°C, strands displace each other when being synthesised, while other random primers anneal to the new strands that are synthesized (Figure 2.2). This process is repeated, resulting in a high concentration of amplified DNA.

1 μl of template was added to 9 μl of sample buffer and heated to 95°C for 3 min to denature the DNA template. This was cooled and added to a cooled mixture of 9 μl of reaction buffer and 1 μl of enzyme mix, then incubated at 30°C overnight (about 16-

18 h). The temperature was then increased to 65°C for 10 min to stop enzyme activity. The mixture was cooled by placing it on ice for a few minutes, then stored at -20°C for later use. Two controls were run: a positive one containing Lambda DNA and a negative one with no template material, only nH_2O, which still amplifies products because of the sensitivity of this technique. According to the manufacturer's instruction, these products are not expected to be used in downstream processes. The amplified viral genomes were visualised by agarose gel electrophoresis, as shown in the figures 3.1 and 3.8.

Figure 2.2: Schematic diagram shows the steps of the amplification process using the MDA method.

2.9.1.2: Shearing the Amplified Viral Nucleic Acids

The amplified viral genomes were broken into small fragments that could be cloned and sequenced by means of a nebuliser, which had two ports: one was a wide blocked outlet port and the other was a narrow inlet port connected to a source of nitrogen under pressure. This method gives completely randomly sheared DNA products that have either blunt ends or short 5' or 3' overhangs, with terminal 5' phosphate and 3' hydroxyl groups.

25 µl of 2 µg of the isothermally amplified viral DNA was added to 725 µl of TE buffer pH 8 containing 10% glycerol. This mixture was pipetted into the bottom of the nebuliser, which was placed on ice to keep the DNA cold. The DNA was sheared for 60 s at 9-10 psi, then the sample was transferred to a sterile microcentrifuge tube and ethanol precipitated as previously described. The pellet was resuspended in 30 µl of TE buffer pH 8.0 and 3 µl of the sheared DNA was visualised on a 1% agarose gel (see Figure 3.2).

2.9.2: Purification and Concentration Techniques

Three methods were used to concentrate the viral particles from soft top agar plates before extraction.

2.9.2.1: PEG Precipitation Method

Polyethylene glycol 6000 (PEG) is a high molecular weight polymer of ethylene oxide used in many applications, one of which is to concentrate virus particles. The filtered viral particles (section 2.6.3) were precipitated by the addition of 1M NaCl and 10% (w/v) PEG 6000, then incubated for 30 min on ice. The particles were pelleted by centrifugation at 12000 x g in a Beckman centrifuge for 30 min, the supernatant was removed and the pellet was resuspended in 1 ml of 10 mM Tris pH 7.5, 10 mM $MgCl_2$ and 100 mM NaCl, then transferred to a fresh 1.5 ml Eppendorf tube. Free nucleic acids were digested by adding 10 U of DNase and 10 µg/ml RNase A and incubating for 30 min at 37°C.

2.9.2.2: ViraPrep™ Lambda kit

The second method of purifying the virus particles was to use a ViraPrep™ Lambda kit. According to the manufacturer's instructions (Biotech Support Group), this is designed to purify viral DNA based upon the unique virus binding reagent, Viraffinity™ Matrix, whose performance characteristics are listed in Table 2.3. The polymer matrix can capture virus particles from plate or liquid. The matrix with virus was pelleted and washed of contaminants using the buffer supplied, then the particles were lysed to release the DNA while the coat proteins and exonuclease remained bound. The viral DNA was concentrated by ethanol precipitation. Among the many advantages of this method are that it is non-hazardous, it produces a high yield, it is simple to use and the entire protocol can be completed in a short time.

Table 2.3: Performance characteristics of ViraPrep™ Lambda kit

Culture Conditions	Titer	DNA Yield	% Bound
150mm plate lysate, solubilised in 10ml ViraPrep™ Lambda HS1 buffer, and clarified	10^9 pfu/ml (approximate)	10 - 20 µg	>95
10ml liquid lysate, plus addition of LL1 buffer,and clarified	10^9 pfu/ml (approximate)	10 - 20 µg	>95

Items required	ViraPrep™ Lambda	Storage
LL1, Buffer (for liquid lysates)	Supplied	4 °C
HS1, Solubilization Buffer (for plate lysates)	Supplied	4 °C
RNase Cocktail	Supplied	-20 °C
HL2, Lysis Buffer	Supplied	4 °C
AA3, Ammonium Acetate	Supplied	4 °C
V1062, Viraffinity™	Supplied	4 °C
Ethanol	Not Supplied	--
Growth Media	Not Supplied	--
Final Resuspension Buffer	Not Supplied	--

10 ml of HS1 buffer was added to a plate with confluent lysis and incubated for 30 min at room temperature. The soft top agar was scraped into a 50 ml centrifuge tube and centrifuged in a Beckman centrifuge at 1250 x g for 10 min. The supernatant was transferred to a fresh 15 ml tube, 2 ml of vortexed Viraffinity™ was added, then the tube was inverted 10 times to mix and incubated at room temperature for 5 min. The tube was then centrifuged at 300 x g for 10 minutes to pellet the phage-matrix complex, the supernatant was removed and the pellet was resuspended and washed with 10 ml of buffer HS1. After a second centrifuge and wash with HS1 buffer, 200 µl of RNase was added and the tube incubated for 15 minutes at 37°C. The tube was again centrifuged at 300 x g for 10 min and the supernatant was removed. The pellet was resuspended with 2 ml of HL2 lysis buffer and heated at 65°C for 10 min to release the DNA. The tube was centrifuged at 16000 x g for 10 min and the supernatant containing the viral DNA was transferred to a fresh tube. 200 µl of AA3 buffer was added and mixed well, then 5.5 ml of 95% ethanol was added and mixed well; this was left to stand for 10 min at room temperature. To pellet the viral genome, the tube was again centrifuged at 16000 x g for 10 min. The supernatant was removed and the pellet was air dried at room temperature. The pellet was resuspended with 50 µl TE buffer pH 8.0.

2.9.2.3: Caesium Chloride Gradient Method

Caesium chloride (CsCl) is used in a purification method that was developed in 1957 and is widely used today for many purposes, one of which is the purification and concentration of virus particles. Ultracentrifugation is used to separate the virus particles from other components on the basis of their buoyant density. Usually the sample is sedimented based on a step density gradient using a high concentration of CsCl. Components in the sample move to the appropriate position during the ultracentrifuge procedure, in a process known as equilibrium sedimentation.

1 g/ml of CsCl was added directly to the filtered virus particles (see section 2.6.3). This was mixed well until the CsCl was dissolved. 8 ml of this mixture was transferred to four ultracentrifuge tubes, which were centrifuged at 55000 rpm for 2 h at 4ºC. After centrifugation the tubes were checked for the appearance of a layer. A 21-gauge needle was used to puncture the plastic centrifuge tube at the level of the layer. A small piece of autoclaved dialysis tubing membrane was clamped at one end and the purified phage sample was gently pipetted into the dialysis tubing, then the upper end was clamped. 1 L of SM buffer was used as a dialysis buffer to remove the CsCl from the purified phage. The SM buffer was stirred during dialysis and changed twice.

2.9.3: Plasmid Extraction

Plasmids were extracted from overnight broth culture using a Wizard® *Plus* SV Minipreps kit supplied by Promega. A single white colony was added to 10 ml of LB broth containing 100 mg/ml ampicillin. This was incubated overnight at 37°C with shaking at 150 rpm in an orbital shaker. Following the manufacturer's protocol, 1.5 ml of overnight broth culture was centrifuged at 16000 x g for 5 min. The pellet was resuspended by pipetting with 250 µl of cell resuspension solution. 250 µl of cell lysis solution was added to lyse the cells and the tube was inverted four times to mix. 10 µl of alkaline protease solution was added and the tube was incubated for 5 min at room temperature to inactivate endonuclease and other proteins that can affect the quality of the isolated plasmid.

Composition of Buffers and Solutions

Cell Resuspension Solution (CRA)	50 mM Tris-HCl (pH 7.5, 10 mM EDTA, 100 µg/ml RNase A
Cell Lysis Solution (CLA)	0.2M NaOH, 1% SDS
Alkaline Protease Solution	----------
Neutralization Solution (NSB)	4.09 M guanidine hydrochloride, 0.759 M potassium acetate 2.12 M glacial acetic acid, final pH approximately 4.2.
Column Wash Solution (CWA)	162.8 mM potassium acetate, 22.6 mM Tris-HCl (pH 7.5) 0.109 mM EDTA (pH 8.0), 95% ethanol added
Nuclease-Free Water	----------

350 µl of neutralisation solution was added and the tube was immediately inverted four times before being centrifuged at 16000 x g for 10 min at room temperature. A spin column was placed into a 2 ml collection tube and the cleared lysate was transferred to the spin column. This was centrifuged at 16000 x g for 1 min, the flow-through was discarded and the spin column was reinserted into the same collection tube. 750 µl of column wash solution was added to the spin column and centrifuged at 16000 x g for 1 min. This step was repeated with 250 µl of column wash solution and it was centrifuged at 16000 x g for 2 min. The spin column was transferred to a new sterile 1.5 ml Eppendorf tube. To elute the plasmid from the spin column, 100 µl of nuclease-free water was added to the centre of the spin column and centrifuged at 16000 x g for 1 min. The eluted plasmid was stored at -20°C for later use.

2.10: Manipulation of Nucleic Acid

2.10.1: Agarose Gel Electrophoresis

Agarose gel electrophoresis was used to separate and purify DNA fragments based on the molecular size. As DNA has a negative charge at neutral pH, DNA fragments migrate towards the anode when an electric field is applied (Sambrook and Russel 2001). The electrophoresis was carried out on Tris-acetate (TAE) buffer, consisting of 24.2% (w/v) Tris-base, 5.71% (v/v) acetic acid and 10% (v/v) 0.5 M EDTA, pH 8, 84.29% dH_20. Based on the size of the DNA fragments, different percentages of agarose were dissolved in TAE buffer, which was then heated until molten and poured into a plastic tray containing a suitable comb to develop the wells. This was left to set at room temperature, then the comb was removed and the gel with its tray were

transferred to the platform inside the electrophoresis tank. Before loading the DNA samples into the gel wells, they were mixed with loading buffer (0.25 % (w/v) bromophenol blue, 0.25 % (w/v) xylene cyanol FF and 15 % (w/v) Ficoll in dH_2O). These mixes were loaded into the gel wells.

2.10.2: Pulse Field Gel Electrophoresis (PFGE)

2% low melting point agarose (Seaplaque® CTG agarose) in 0.5 × TBE (20 mg in 1.0 ml) was prepared. This was difficult to dissolve, so the eppendorf tube was held in boiling water to keep the agarose warm. 40 µl of sample was added into an eppendorf tube, followed by the addition of 40 µl of agarose, which was mixed gently by pipetting to avoid the presence of bubbles. The mixed sample was transferred into a plug mould and left on the bench for 2 hours. 1 ml of lysis buffer was added to the eppendorf tube, followed by the agarose plug, pushing the agarose out from the back. This was incubated overnight in a water bath at 55°C. On the second day, 200 ml pulsed field certified agarose (1%) in 0.5 × TBE was prepared and was then melted by microwave for 3 min. The melted agarose was poured into a plastic tray with a suitable comb; this was left at room temperature for 2 hours to set. The agarose plugs were washed three times using 1 × TE buffer. Dried plugs and PFGE ladder were put into the wells. Gels were run for 18 hours at 14°C using a CHEF DR-III system (Bio-Rad) at 6.0 V/cm with initial and final pulse times of 5 s and 13 s respectively. Following electrophoresis, the gel was stained for 30 min in 0.5 × TBE containing Gel Red (Biotium). The gel image was captured using a G:BOX gel documentation system (Syngene).

2.10.3: Phenol Chloroform Extraction

An equal volume of phenol:chloroform:isoamyl alcohol (25:24:1) was used to extract the viral nucleic acid from the viral capsid and denatured proteins associated with DNA. An equal volume of phenol chloroform was added to the sample. The mixture was inverted for about 30 s, then centrifuged at 16000 x g for 1 min. It formed two layers; the upper aqueous layer containing the viral DNA was carefully removed to a fresh tube. The viral DNA then was precipitated by ethanol precipitation, as described below.

2.10.4: Ethanol Precipitation

The extracted viral DNA was precipitated by the addition of 0.1 times the sample volume of 3 M sodium acetate (pH 5.2) and twice the sample volume of 95% ethanol. The tube was mixed and chilled on ice for 15 min, or incubated at -20°C overnight, then the sample was centrifuged at 16000 x g for 10 min. The supernatant was removed and the pellet was washed with 70% ethanol. The DNA pellet was air dried at room temperature for 15 min, then dissolved in 30 µl TE buffer. This was stored at -20°C until needed.

2.10.5: Extraction of the Gel Slice

Viral DNA fragments 1-5 kb long were extracted from agarose gels using a Sephaglas™ BandPrep Kit, whose components were: Sephaglas BP (a DNA binding reagent), gel solubilizer (containing sodium iodide, to dissolve the gel slice and help to promote binding of the DNA), washing buffer (to remove gel contamination such as proteins, nucleotides and linkers from matrix-bound DNA) and elution buffer (to elute the DNA). Following the manufacturer's protocol, 1 µl of gel solubilizer was used for each mg of agarose. The tube was incubated at 60°C with vortexing for 5 to 10 min or until the agarose slice was dissolved. The Sephaglas was added for scale-up to 5 µl for each estimated µg of template DNA in the gel slice. This was incubated at room temperature for five minutes with gentle vortexing every minute to resuspend the Sephaglas. The tube was centrifuged at 12000 x g for 30 s to pellet the Sephaglas and the DNA. The supernatant was removed and the tube was centrifuged again to remove any residual liquid. The pellet was washed with buffer for scale up to 16 times the volume of the added Sephaglas. This was centrifuged at 12000 x g for 30 s to pellet the Sephaglas and the supernatant was removed. The washing step was repeated three times. The residual ethanol was removed from the matrix by air drying and a low ionic buffer was used to elute the DNA from the dried matrix.

2.10.6: Determining Sizes of DNA Fragments

The 1 kb DNA ladder and the Lambda DNA Hind III Digest marker were used to determine the sizes of the DNA fragments. The 1 kb ladder was used to determine the sizes of fragments between 100 and 12000 bp, while the Lambda DNA Hind III was used for larger fragments, ranging from 1503 to 23000 bp.

2.10.7: Measuring DNA Concentration

Two methods were used to quantify the DNA. One was using the 1 kb DNA ladder band 1636 bp, which gave 200 ng if 2 µg of the 1 kb DNA ladder was loaded. Simply, unknown DNA concentration was visualized by agarose gel electrophoresis and compared to the 1636 bp band to estimate the unknown DNA concentration.

Ultraviolet spectrophotometry was used to measure the quantity of DNA. Although UV can be used to estimate the DNA or RNA in a sample, it cannot distinguish between them. The DNA sample was placed in a cuvette and quantified by the absorbance at wavelengths of 260 nm and 280 nm. First, the cuvette was washed with nH_2O and a blank control solution of nH_2O was put in it to set the zero reading. Then the blank control was removed using a pipette and the DNA sample was placed in the cuvette and read at 260 nm and 280 nm. The following formula was used to calculate the DNA concentration:

[DNA] (µg) = (A 260 – A 280) x 50 (dilution factor)

2.11: Polymerase Chain Reaction Amplification

In this project PCR was used to fill the gaps between the contigs of the viral genome. Several forward and reverse primers were designed at the ends of the contigs. Low concentrations of DNA template were used to amplify the unknown fragments of viral DNA. The components of the PCR used were reaction buffer, $MgCl_2$, dNTPs mix, forward primer, reverse primer and Taq polymerase.

For each reaction carried out, the following were mixed:

- 34.75 µl of nH_2O
- 5 µl reaction buffer
- 6 µl of 25 mM $MgCl_2$
- 1 µl dNTPs mix
- 1 µl forward primer
- 1 µl reverse primer
- 1 µl DNA template
- 0.25 µl Taq polymerase
- 50 µl final volume

A negative control was always carried out with each reaction, using 1 µl of nH_2O in place of the DNA template. Initially, the strands of template DNA were denatured at 95°C for 5 min, then cycles were carried out as follows:

- Step 1: 95°C held for 5 min
- Step 2: 95°C for 1 min
- Step 3: (annealing) 55°C for 45 seconds
- Step 4: (extension) 72°C for 1 min
- Step 5: repeat steps 2 to 4 for more 29 cycles
- Step 6: (extension) 72°C for 10 min
- Step 7: incubate at 15°C for 96 hours.

The annealing temperature might be varied according to the primer length and G+C content; primers with a higher G+C content require a higher annealing temperature. The extension at step 4 may also be varied according to the length of the amplified DNA.

2.11.1: Gradient PCR

The gradient PCR is used to detect the optimal annealing temperature, in order to avoid amplifying multi, or non-specific bands that may have appeared after the PCR reaction. To obtain the specificity of a PCR reaction, 10°C below and above the calculated temperature of the primer melting point (T_m) were used. Twelve different annealing temperatures could be run simultaneously using the universal block of the Peltier thermal cycler machine. Thus, the gradient process was set dependant up on the T_m of the primers used. For example, if the T_m of the used primers were 55°C, the gradient was set between 45 to 65°C.

2.11.2: PCR of 16S ribosomal RNA (rRNA) Genes

To identify the origin of bacteria isolated and used in this study, 16S rRNA genes were amplified by PCR as indicated. First the bacterial DNA was extracted, then the 16S rRNA gene was amplified using the specific forward and reverse primers in a total volume of 50 µl, as indicated.

2.11.3: PCR Product Purification

A QIAquick PCR Kit, supplied by QIAGEN, was used to clean up the amplified DNA fragments generated by PCR. It is designed to purify single or double stranded DNA. According to the manufacturer's protocol, using this kit will clean all the enzymatic reactions, primers, nucleotides and salts that were used in the PCR reaction.

A QIAquick spin column was placed in a 2 ml collection tube. Five volumes of buffer PB were added to the PCR sample and mixed well. To bind the DNA, the mixture was transferred to the QIAquick column and centrifuged at 16000 x g for 30-60 s, the flow-through was discarded and the QIAquick column was replaced in the collection tube. 750 µl of PE buffer was added to the QIAquick column to wash the bound DNA and then centrifuged at 16000 x g for 30-60 s. The flow-through was discarded and the QIAquick column was replaced in the same tube. This was again centrifuged at 16000 x g for 1 min to remove residual ethanol. To elute the DNA, the QIAquick column was placed in a clean 1.5 ml centrifuged tube and 50 µl of elution buffer was added in the centre of the QIAquick membrane. This was centrifuged at 16000 x g for 1 min and the eluted DNA was stored at -20°C for later use.

2.12: Cloning

2.12.1: A-Tailing for the Sheared Viral DNA

The TA cloning vector needs the DNA fragments to be A-tailed because the cloning vector has T-overhangs at the cloning site. A-tailing will fill all gaps at the ends of fragments and add an A-overhang. The following were mixed in a 0.5 ml Eppendorf tube:

- 25 µl DNA fragments
- 5 µl 10x Taq polymerase reaction buffer
- 5 µl 25 mM $MgCl_2$
- 1 µl deoxynucleotide triphosphate mix (dNTPs)
- 1 µl Taq polymerase
- nH_2O to a final volume of 50 µl

This was incubated at 72°C for 30 min. The A-tailed sample was cleaned of Taq polymerase, salts and free dNTPs using the PCR purification kit (Qiagen) and stored at -20°C.

2.12.2: Ligation of A-tailed DNA Fragments

Following the manufacturer's protocol, the A-tailed DNA fragments were ligated to the pGEM-T Easy vector.

- 5 µl ligation buffer
- 1 µl pGEM-T Easy vector (50 ng)
- 2 µl of the sheared viral fragments
- 1 µl T4 DNA Ligase (3 Weiss units/µl)
- 1 µl dH_2O
- 10 µl final volume

This preparation was mixed by pipetting and incubated for 1 h at room temperature or overnight at 4°C if the maximum number of transformants were required.

2.12.3: Transformation by Heat Shock

Following the manufacturer's protocol, 50 µl of frozen *E. coli* JM109 competent cells were thawed on ice for 5 min and added to 2 µl of the ligation preparation. The tube was gently flicked to mix the cells and placed on ice for 20 min. The tube was heat shocked by incubating at 42°C for exactly 45 to 50 s, then quickly returned to ice for 2 min. Next, 950 µl of SOC was added to the tube, which was incubated at 37°C in an orbital shaker for 1.5 h with shaking at 150 rpm. 100 µl of the grown cells were plated out onto a medium containing ampicillin, IPTG and X-Gal, then the plates were incubated overnight at 37°C. The cells containing the inserts were able to grow because the ampicillin resistance gene was present on the vector. The plates were stored at 4°C to facilitate blue/white screening.

2.13: DNA Sequencing

DNA fragments were sequenced by Lark Technologies (Essex, UK) or AGOWA (Berlin, Germany) or PNACL, University of Leicester.

2.14: Transmission Electron Microscopy

Preparations of virus particles were visualised by transmission electron microscopy at the Electron Microscopy facility of the University of Leicester. Approximately 3 µl of virus sample was placed on fresh glow-discharged Pioloform-coated grids and fixed in glutaraldehyde vapour for 2 min. Excess solution was removed from the grid using filter paper. The grid surface was washed with dH_2O, then left to dry at room temperature for about 3-5 min and negatively stained with 1% (v/v) uranyl acetate and viewed on a JEOL 1220 microscope (Pagaling, *et al*., 2007).

2.15: Computer Analysis

2.15.1: Viewing Gels

A White/Ultraviolet Transilluminator (Ultra Violet Products) was used to visualise polyacrylamide and agarose gels. Images were captured and transferred using a Kodak EDAS 290 camera.

2.15.2: Viewing DNA Sequence Data

Chromas version 2 (Technelysium Pty Ltd) was used to visualise and edit chromatograms of DNA sequences.

2.16: Assembling DNA Sequences

The genomic viral DNA sequences were assembled using the Lasergene SeqMan version 7.0 program (DNAStar).

2.17: Homology and Annotation of the Viral Genomes

The nucleic acid sequences were first analysed using the online Basic Local Alignment Search Tool (BLAST), a search program that can be accessed at http://www.ncbi.nlm.nih.gov/. This program has the ability to find similarities between a protein or DNA query sequence and any of the available sequences in the GenBank databases. BLAST uses a heuristic algorithm that seeks out local, as opposed to global alignments; therefore, it is able to detect relationships amongst

sequences that only share isolated regions of similarity (Altschul *et al.*, 1990). BLASTN was first used to check if there was any known match to gene homology in the GenBank, via comparing each nucleotide query sequence against a nucleotide sequence database. Latter, TBLASTX was used to uncover the sequence identity by comparing the six-frame translations of each query sequence, against the six-frame translations of the database.

Open Reading Frame Finders (ORFs) at http://www.ncbi.nlm.nih.gov/projects/gorf/ and GeneMark at http://opal.biology.gatech.edu/GeneMark/index.html were used to predict and analyse the genes present in the sequences. The predicted ORFs were compared to the GenBank databases using the BLASTP search program, which compares an amino acid query sequence against a protein sequence database. Statistics for each of the ORFs were calculated using the ProtParam program (http://www.expasy.ch/tools/protparam.html). GC content in the virus countings were calculated using the online base composition tools at http://atmolbiol-tools.ca/Jie_Zheng.

2.18: Statistical Analyses

2.18.1: Phylogenetic Tree Evolutionary Relationships

The Molecular Evolutionary Genetics Analysis (MEGA) version 3.1program (Kumar et al., 2004) was download freely from http://www.megasoftware.net/ . It was used to align and compare the 16S rRNA generated from the isolated bacterial host with other related species.

2.18.2: Richness Estimation

The abundance of viral communities was calculated using Chao1, which is used to estimate the number of classes in a population (Chao, 1984).

$$S_{Chao1} = S_{obs} + \frac{n_1\ n_1 - 1}{2\ n_2 + 1}$$ **When n_1>0 and n_2≥0 and when n_1=0 and n_2=0**

$$S_{Chao1} = S_{obs} + \frac{n_1^2}{2n_2}$$ **When n_1=0 and n_2≥0**

In this equation, S_{Chao1} is the total number of different clones in a population, S_{obs} is the number of observed clones, n_1 is the number of clones observed once and n_2 is the number of clones observed more than once.

2.18.3: Measuring Biodiversity

The Shannon-Weaver Index is calculated using DOTUR program for every distance level that was used, by the equation:

$$H_{Shannon} = -\sum_{i=1}^{S_{obs}} \frac{S_i}{N} \ln \frac{S_i}{N}$$

Species Evenness

The species evenness can be calculated from values for $H_{Shannon}$ by the equation:

$$E = \frac{H_{Shannon}}{\log N}$$

Chapter 3

Cloning and Sequencing of Uncharacterized Virus Gene Fragments Isolated from Human Dental Plaque

Chapter 3

3: Results and Discussion

3.1: Introduction

The aim of this project was to produce a metagenomic virus library from dental plaque in the human mouth. This seems desirable in order to measure the virus diversity present in an important niche of the human mouth. This is the first such study to attempt to estimate viral diversity in the human mouth. Such knowledge has the potential both to reveal novel virus sequences and to provide more information about the virus community in the human mouth.

In principle, when measuring virus diversity using metagenomic methods, samples should either contain a high population of viruses or be of a large volume, which is then concentrated to increase the virus population. Dental plaque samples are obviously difficult to collect in large volumes when compared to samples previously used for virus metagenomic analysis i.e. water, sediment and faeces. For example, in previous studies the volume of samples collected to analyse viral diversity using metagenomic methods were 200 litres of water (Breitbart *et al.*, 2002), 1 kg of sediment (Breitbart *et al.*, 2004), approximately 500 g human faeces (Breitbart *et al.*, 2003) and 500 g of horse faeces (Cann *et al.*, 2005), all of which were enough to create good viral metagenomic libraries and provide significant information about the virus diversity in these different environments.

Collecting dental plaque samples in large quantities from the human mouth thus presented a major problem in this project. Small amounts of sample yielded a low content of extracted viral genomes, which could not be visualized on agarose gels when stained with ethidium bromide or SYBR Green I, and which were not suitable for the linker amplified shotgun library (LASL) (section 1.5.2.3.3). One solution to these difficulties in accessing or exploring the viral diversity in some clinical samples or other environments is to use recently-developed nucleic acid amplification (Abulencia *et al.*, 2006). The recently commercialized technique involving the multiple displacement amplification kit (MDA) was used to amplify the viral genes

extracted from the human mouth (section 2.9.1.1). This allows the technically facile amplification of ng amounts of nucleic acid to μg amounts. It has been widely used to amplify 'precious' DNA samples for community distribution and was initially reported as doing so in an unbiased and reliable fashion (Abulencia *et al.*, 2006). For these reasons we chose to attempt to amplify virus nucleic acid sequences isolated from dental plaque by MDA and to take this amplified material as the starting point for community sequence analysis.

The MDA method was used to generate a metagenomic virus library from the human mouth, beginning in the year 2005, when we were not aware of any disadvantages to the MDA method. Obviously the template to be amplified should be as pure as reasonably possible, undesired amplification of contaminant DNA has been observed (Dean *et al.*, 2002; Hosono *et al.*, 2003). MDA uses the Phi29 proof reading polymerase enzyme. One feature of the use of this enzyme is that it generates amplified DNA products even in the absence of input DNA template; this is called a DNA background (Abulencia *et al.*, 2006; Blanco *et al.*, 1989). According to the manufacturer's protocol, the DNA background generated in negative controls is artefacts such as primer-derived multimers and cannot be cloned (Lasken *et al.*, 2007). However, in a previous study from our laboratory (Pagaling, 2007), samples from the negative control were able to be cloned and sequenced. These sequences had no significant matches to the GenBank databases. Another possibility is that in negative control samples, trace nucleic acid contamination associated with the purification of the enzyme from its bacterial expression system is amplified; such contamination is a well known feature of enzyme reagent manufacture. Template DNA, especially if added in high concentrations relative to any 'endogenous contamination', would be expected to preferentially amplify in this system and reduce or eliminate the 'background' DNA. Although we are not aware of systematic experiments supporting this supposition, its widespread use and manufacturers claims would seem to justify such a supposition.

Perhaps more seriously both amplification bias and chimeric rearrangements have also been observed using the MDA method (Dean *et al.*, 2001; Dean *et al.*, 2002; Lasken and Egholm 2003; Lasken and Stockwell 2007). Some modifications of the MDA procedure to ameliorate these problems have been described. One study used a

combination of MDA and rolling circle amplification methods to amplify a circular 7 kb DNA template by reducing the standard volume of the MDA reaction from 50 µl to 600 nl, which improved the specificity of the amplification (Hutchison *et al.,* 2005). However, the effect of the lower volume on amplification bias was not determined. Another study which aimed to amplify a single cell genome using MDA showed that reducing the volume from microlitres to nanolitres reduced the bias, while the specific amplification was increased (Marcy *et al.,* 2007).

Notwithstanding the attempts to address them, these disadvantages created doubts about the origin of some sequences that had no significant matches to the databases. Sequences which were not generated from the specific input template DNA are a concern that affects the statistical analysis of the virus diversity in the human mouth. Nevertheless, using the MDA method was the only choice to explore the virus genomes in such a DNA sample obtained from the dental plaque of the human mouth at that time.

3.2: Collection of Samples

Materials attached to the dental plaque and between the teeth were collected from the mouths of three volunteers with no history of taking antibiotics in the three months prior to sampling. Toothpicks and dental floss (Johnson and Johnson, REACH) were used to collect the samples. The volunteers were asked not to brush their teeth the night and morning immediately before collection of the samples and not to have breakfast on the day of collection. Thus, dental plaque mass would be increased without undue contamination with recent food debris. Each sample, with toothpicks and dental floss, was added to 1 ml of PBS buffer and rigorously agitated to disperse the dental plaque material into solution. Samples were obviously contaminated with saliva and blood, indicative of mild gingivitis.

A virus genomic library was constructed for each volunteer from these samples. Briefly, viral nucleic acids were extracted from each sample and were then amplified, sheared, cloned and sequenced. Eighty, ten and eleven clones were sequenced from the first, second and third volunteers respectively. These results and the problems associated with using the MDA method to amplify genomic DNA are described below.

3.3: Extraction and Amplification of the Viral Genomes

Each plaque sample in 1 ml of PBS was filtered twice: first through a 0.45 nm filter and then through a 0.2 nm filter to separate viral particles from bacteria and other debris. A final concentration step, precipitation with PEG 6000, as described in section 2.9.1, was applied to concentrate the viral particles. The concentrated virus samples in 0.5 ml of 10 mM Tris pH 7.5, 10 mM $MgCl_2$ and 100 mM NaCl were then treated with DNase and RNase to degrade extracellular, non-viral nucleic acids (section 2.9.2.1). Virus genomes were then extracted by adding an equal volume of phenol: chloroform (section 2.10.3). Viral nucleic acids were then precipitated with ethanol and the pellet was dissolved in 30 µl of TE buffer. 10 µl of the sample was run on 0.8% agarose gel and stained with ethidium bromide or SYBR Green I; no bands were detected.

This was obviously an unpromising start for library construction. Accordingly, an isothermal amplification of the extracted viral nucleic acid was used to increase the amount of genomic DNA (section 2.9.1.1). High molecular weight amplified viral genomes were obtained, as shown in Figure 3.1. Band 2 is the control reaction using lambda DNA (1 µl of 10 ng/µl) added as a starting template. By comparison with DNA markers (section 2.10.6), about 4 µg of product was generated, which corresponds with the amount expected to be generated within 16-18 h at 30°C, according to the manufacturer's protocol. Band 3 shows that the quantity of DNA was about 2 µg. Band 4 is the negative control reaction; although no input template was used, this still generated a background smear (section 2.9.1.1).

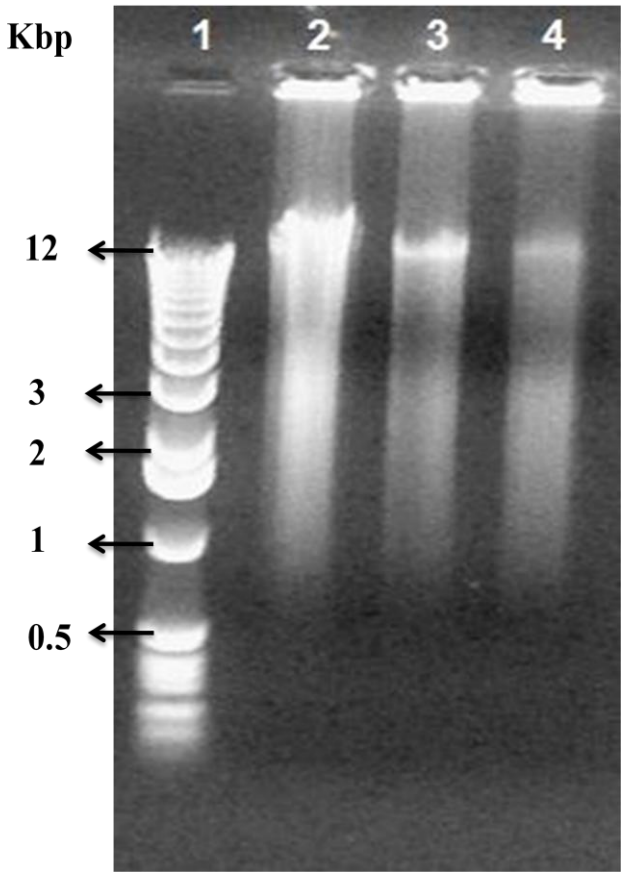


Figure 3.1: 0.8% agarose gel showing the amplified viral genomes from the first volunteer
Lane 1: 1 kb marker
Lane 2: control reaction (using lambda DNA (1 µl of 10 ng/µl))
Lane 3: amplified viral genomes that show increase of level of viral genoic DNA.
Lane 4: no input template DNA; still showed a background.

3.4: Fragmenting and Sequencing the Amplified Viral Genomes

The amplified nucleic acid, presumptive viral genomes (band 3, Figure 3.1) were physically sheared by passing them through a nebulizer (section 2.9.1.2). This resulted in shearing of the nucleic acid sample to small fragments, which were electrophoresed on a 1% agarose gel; see Figure 3.2. Fragments between 1 and 5 kb were extracted from the gel slice using the Sephaglas™ BandPrep Kit. The purified fragments were A-tailed and then cloned into the pGEM-T Easy vector (section 2.12). Minipreps were made to purify the recombinant plasmids. Then endonuclease *Eco*RI was used to check the size of the insert (Figure 3.3). Clone sizes varied from 2-3kb to 0.5kb. Inserts having the same size on the gel were not sent to be sequenced, in order to avoid duplicates. Fragments of interest were sent for sequencing by Lark in the UK and AGOWA in Germany

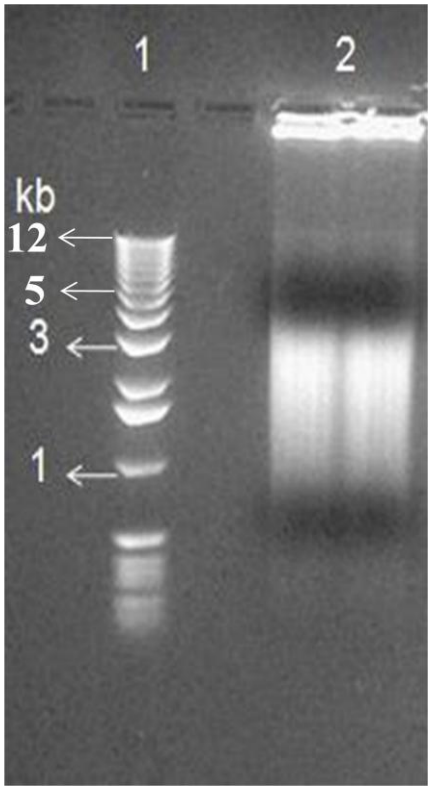


Figure 3.2: Nebulizer sheared amplified viral genomes
Lane 1: 1 kb marker

Lane 2: The amplified viral genomes was sheared for 60 s at 9-10 psi, and sample was run on a 1% agarose gel, band shows the fragments estimated sizes.

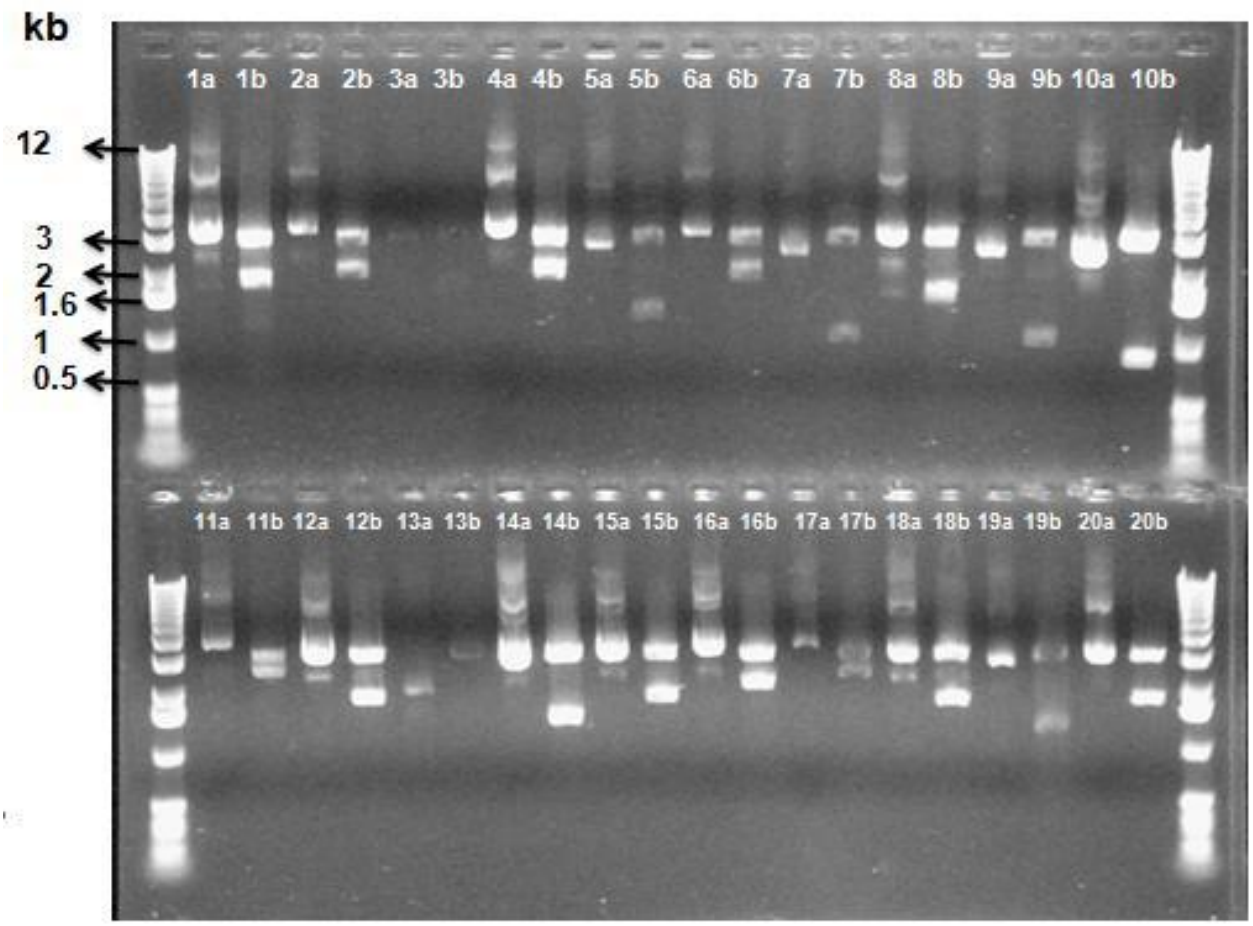


Figure 3.3: Plasmid with its inserts digested with endonuclease *Eco*RI
Restriction enzyme digestion of sheared DNA optain from folountter 1. The 1 kb marker was run on the left and right sides of the gel. Bands marked "a" are the uncut plasmid, bands marked "b" are those where the insert was cut from its plasmid using the *Eco*RI enzyme. Fragments generated different in size suggesting sequence diversity.

3.5: Sequence Analysis

Clones from the first volunteer were sent to be sequenced only from forward primers P7 and M13. Figure 2.1 shows where the insert was cloned and where these primers were primed on the plasmid. Eighty sequences were obtained and contiguous overlapped sequences were identified after assembling all the sequences using the Lasergene SeqMan version 7.0 program (DNAStar). Assembly of the 80 sequences at 98% identity and a minimum overlap length of 20 bp was used to identify contigs. Resulted in 6 contigs and 45 single sequences appeared, as shown in Table 3.1.

The only contigs to be used for the population analysis were those created from two different sequences, not from one clone that occured twice (Breitbart *et al.*, 2002). Note that the second and third volunteers' sequences were not used for the estimation of the viral diversity or for the population analysis reported in this chapter, because the number of sequences in each case was too small. Only the identities of these sequences were searched against the GenBank databases (section 3.5 and 3.6).

Table 3.1: The outcome of assembling the 80 sequences

Name	Contig Length	Total Sequence Length	Number of Sequences	Name	Contig Length	Total Sequence Length	Number of Sequences
Contig 1	3137	8248	11	seq 27	287	287	1
Contig 2	2420	7049	9	seq 28	287	287	1
Contig 3	2053	3024	4	seq 29	217	217	1
Contig 4	1161	1411	3	seq 30	865	865	1
Contig 5	1059	2076	3	seq 31	400	400	1
Contig 6	273	442	2	seq 32	298	298	1
seq 7	888	1740	2	seq 33	354	354	1
seq 8	744	1468	2	seq 34	400	400	1
seq 9	449	886	2	seq 35	855	855	1
seq 10	565	565	1	seq 36	873	873	1
seq 11	515	515	1	seq 37	890	890	1
seq 12	864	864	1	seq 38	289	289	1
seq 13	881	881	1	seq 39	761	761	1
seq 14	345	345	1	seq 40	857	857	1
seq 15	789	789	1	seq 41	892	892	1
seq 16	875	875	1	seq 42	152	152	1
seq 17	780	780	1	seq 43	904	904	1
seq 18	820	820	1	seq 44	853	853	1
seq 19	724	724	1	seq 45	833	833	1
seq 20	714	714	1	seq 46	698	698	1
seq 21	285	285	1	seq 47	213	213	1
seq 22	640	640	1	seq 48	628	628	1
seq 23	437	437	1	seq 49	400	400	1
seq 24	486	486	1	seq 50	639	693	1
seq 25	489	489	1	seq 51	189	189	1
seq 26	852	852	1	Total	37804	51964	80

Seq: single sequence

TBLASTX was used to try to identify the origin of these sequences, which were compared to sequences in the GenBank non-redundant (nr) database. Sequences

compared to the GenBank were considered significant if they had an E-value of less than 0.001 (Breitbart *et al.*, 2002). Sequences that resulted in significant matches to those in GenBank were then divided into groups based on the sequence annotation. The groups obtained were: viruses, bacteria, mobile elements or eukarya. A sequence was classified as a virus if it appeared within the top five significant database matches (Breitbart *et al.*, 2002). This significant match to a virus was then classified into a family group based on the International Committee on Taxonomy of Viruses (ICTV) classification.

3.5.1: Estimates of Viral Community Diversity

Metagenomic analysis overcomes the difficulties of estimating the viral diversity in any given sample. Sequences must first be assembled to check the production of overlapping sequences; if many contigs are identified, this means that the viral diversity is not as high as when fewer contigs are identified (Breitbart *et al.*, 2004). The results shown in Table 3.1 immediately suggest that my library exhibits low clone diversity a conclusion supported by the statistical analysis, described below.

3.5.2: Population Analysis of the Sequences

The *Shannon-Wiener Index*, also known as the *Shannon-Weaver Index*, is used to estimate the diversity of species in terms of richness (the number of genotypes) and evenness (the relative abundance of each genotype) (Magurran, 2004). The average length of the 80 sequences was 649 nucleotides, ranging from 152 to 926 nucleotides. The total length of the 80 sequences was 52294 base pairs. Eight sequences were found to appear twice; thus the total number of unique sequences was 72 only. Based on the Shannon Index (Shannon, 1997), the value of species diversity and evenness was found to be 1.9, by applying the following formula:

1/80 (72 times), 8/80 (once)

$$E = \frac{H_{Shannon}}{\log N}$$

N is the number of individual sequences

Shannon index = - 72 x (1/80 log (1/80)) - 1 x (8/80 log (2/80)) = 1.9

This value is much lower than the reported values for virus communities, which ranged from 5.6 in equine faeces (Cann *et al.*, 2005) and 6.4 in human faeces (Breitbart *et al.*, 2003) to 9 in marine sediments (Breitbart *et al.*, 2002).

3.5.2.1: Richness Estimation

The abundance of viral communities was calculated using Chao 1, which is used to estimate the number of classes in a population (Chao, 1984).

SChao1 = Sobs+ (n_1^2 /2n2)

In this equation: S_{Chao1} is the total number of different clones in a population, S_{obs} is the number of the observed clones, n_1 is the number of clones observed once and n_2 is the number of clones observed more than once.

$= 80 + (72^2 /2 \times 8) = 404$

Based on this equation, this library has 404 distinct clones with an average size of 649 bp. As the average virus genome size is 50 kb, typical of the tailed bacteriophages that constitute the majority of virus clones in this library (see below), then 50000/649 or 77 clones comprise a complete virus genome and the equation (1) library contains (404/77) or 5 different virus genomes, , assuming all the clones are derived from virus sequences.

3.6: Sequences Identity in Sample from the First Volunteer

BLASTN analysis was carried out on the 72 sequences and only three matches to the databases were detected. Two of them matched to the human genome that came from a read of 400 and 237 nucleotides respectively (accession numbers AC083867 and AC107070), and showed 97% identity over the 310 nucleotides, with an E value of 2e-144. The second was shorter, over 110 nucleotides, and showed 89% identity, with an E-value of 1e-30. The third match came from a read of 688 and showed a match over only 67 nucleotides, with 92% identity to a partial 16S rRNA gene of *Streptomyces* sp SHX-102 (accession number AM889493.1). The matches were not over the whole clone length; this may be due the formation of chimera which occurred during the multiple displacement amplification.

3.6.1: Sequences with no Similarity to the Databases

After analysis with TBLASTX, 40 of the 72 clones (55.5%) were found to have no significant matches to the GenBank database with E-values less than 0.001 (Figure 3.4A). These unknown sequences could be unknown virus sequences generated by the MDA method or other unknown genomic contaminates e.g. bacterial genome. If the unknown sequences belonged to virus genomes, the results show that the diversity of viral communities in the human mouth is a little lower than the 59% of unknown sequences observed from a genomic library of human faeces (Breitbart *et al.*, 2003).

It is impossible to detect the origin of single virus sequence when no similarities are detected with the GenBank databases using TBLASTX, but it is easy to establish the origin of an unknown sequence when it overlaps with other known sequences and forms a contig. Recently, independent amplification methods have built up a large fraction of sequences (5%-30%) from animal and environmental samples including viruses that have no significant similarities to the current GenBank database sequences (Delwart *et al.*, 2007).

3.6.2: Sequences with Significant Similarities to the Databases

Among the 72 sequences, 32 (44.5%) showed identity with known sequences using TBLASTX. Thus, the proportion of known sequences obtained in this library was slightly higher than those of the five DNA viral metagenomic libraries that use the random shotgun cloning method, which are between 21% and 41% (Breitbart *et al.*, 2002; Breitbart *et al.*, 2003; Breitbart *et al.*, 2004; Cann *et al.*, 2005). Of the 32 known sequences, twenty-one matched to viruses, two matched to mobile elements, four matched to human DNA, three matched to bacteria and two matched to eukarya; see Figure 3.4B and Table 3.2 lists the identities and other characteristics of these except viruses, which they will be discussed below the table.

Table 3.2: Significant matches to human, bacteria, mobile and eukara sequences were detected using TBLASTX analysis

Organism	Seq name	Total seq length bp	LMN bp	E-value	Identity	Accession number	Similarity using TBLASTX
Human	Seq 38	289	93	3e-10	77%	AC107070	Homo sapiens BAC clone RP11-293H4
	Seq 31	400	264	2e-41	91%	AC083867	Homo sapiens chromosome 7 clone RP11-183I20
	Seq 47	213	99	6e-09	66%	AC107070	Homo sapiens BAC clone RP11-293H4
	Contig 2	417	87	1e-12	82%	AC107070	Homo sapiens BAC clone RP11-293H4
Bacteria	Seq 32	298	288	2e-73	95%	AE017283	DeaD/DeaH box helicase, Propionibacterium acnes KPA171202
	Seq 40	857	729	2e-109	65%	AM398681	Citrate (Si)-synthase *Flavobacterium psychrophilum* JIP02
	Contig 3	731	255	5e-19	31%	BX248359	Hypothetical exported protein *Corynebacterium diphtheriae* gravis NCTC13129
Mobile	Seq 18	820	81	2e-04	48%	AL939131	Secreted protein of *Streptomyces coelicolor* A3
	Seq 30	865	213	7e-05	33%	AE014275	Hypothetical protein *Streptococcus agalactiae* 2603V/R
Eukara	Contig 1	750	138	2e-04	36%	AC132444	Mus musculus BAC clone RP23-160D6
	Seq 44	853	171	0.001	28%	X56010	Sorghum vulgare hydroxyproline-rich glycoprotein gene

LMN: length of matched nucleotides over the total sequence

Among the virus groups, there were seven significant hits to *Siphoviridae*, one to *Myoviridae* and one to *Podoviridae*, which are the major tailed bacteriophage families. Two sequences were similar to the *Bicaudaviridae* family, five were similar to prophages and the remaining five sequences were unclassified viruses; see Figure 3.4C. The number of *Siphoviridae* (7 sequences) and prophages (5 sequences) were thus found to be the highest, possibly because *Siphoviridae* is the most common family of cultivated and uncultivated temperate phages. It was found that the siphophages and prophages were strongly represented and more abundant in the marine sediment (Breitbart *et al.*, 2004) and faecal libraries (Breitbart *et al.,* 2003) than in the seawater libraries (Breitbart *et al.,* 2002). No matches were observed for the T7-like podophages or Y-like siphophages, while in the marine viral communities these were found to be the most abundant groups (Breitbart *et al.*, 2004).

Based on BLASTP with predicted ORF finder and TBLASTX, 61% of the known viral sequences were found to match genes of known function, as indicted in Table 3.3. These functions were phage portal protein, terminase, minor tail protein, tail accessory factor gp26, predicted ATPase, binding protein gp32, proteases gp76, gp28 and gp50, two-tailed virus, endolysin, RTX toxins and related $Ca2^{+}$-binding protein, and repeat sequences of human herpes virus 6B. It was found that 39% of known sequences were similar to proteins of unknown function or presented as hypothetical proteins. Most of the hypothetical proteins may be classified as prophage sequences, based on their location in the bacterial genome.

Table 3.3: Categories of significant matches to uncultured virus proteins in the database

Protein classification	Number of matches
Unknown	14
Phage portal protein	6
Terminase	3
Minor tail protein	2
Tail accessory factors gp26	2
Predicted ATPase	2
Binding protein gp32	1
Protease /gp76	1
gp28	1
gp50 two-tailed virus	1
Endolysin	1
RTX toxins and related Ca2+-binding protein	1
Repeated sequence of human herpes virus 6	1

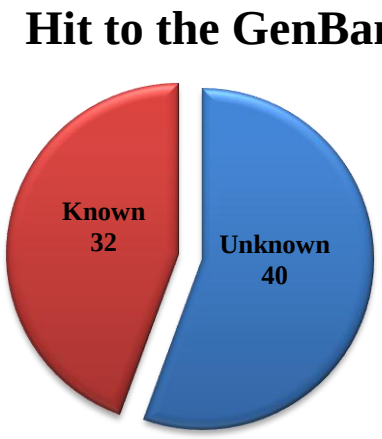


Figure 3.4A: Number of known and unknown sequences using TBLASTX

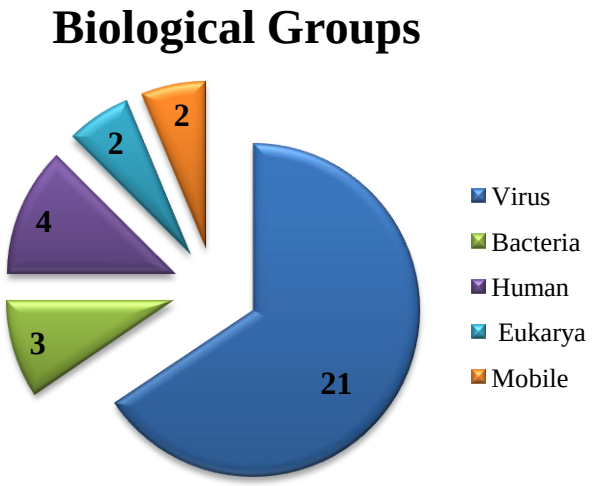


Figure 3.4B: The biological groups of known sequences

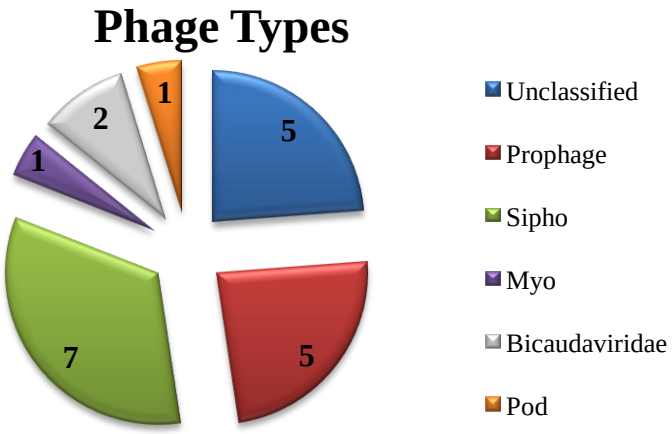


Figure 3.4C: Phage types detected

Contigs Analysis

Using TBLASTX and BLASTP analysis of the predicted ORF finder results, I noticed that four contigs, 1, 2, 3 and 4, and one single sequence, seq 12 (Table 3.1), had a significant similarity to a putative prophage genome in the *Corynebcterium diphtheria* genome. These contigs matched a small region in the bacterial genome of about 12,000 bp. Four differently sized gaps exist between these contigs, as shown in Figure 3.4A. Primers were designed at the end of each contig. Several primer walking and PCR reactions were used to join the contigs using the MDA amplified extracted viral nucleic acid sample genomes as a template DNA, and all the gaps were filled successfully (Figure 3.5A). However, the PCR product of gap D gave several bands, so a gradient PCR was set up to observe more specific bands (Figure 3.7). All the contigs were assembled after the gaps had been filled using Lasergene SeqMan (DNAStar) and CAP3 (http://pbil.univ-lyon1.fr/pbil.html) to ensure that the gaps were filled correctly; see appendix for the total nucleotides (11554 bp) of the joined contigs.

It was found that the lengths of the PCR products for filling the gaps were not exactly the same length as the predicted gaps using TBLASTX and BLASTP analysis: the PCR products of gaps A, B, C and D were about 400, 1840, 50 and 690 bp respectively. GeneMark and ORF Finder were used to predict the genes of the joined contigs. TBLASTX was also used to find the function and the origin of these genes (Table 3.2).

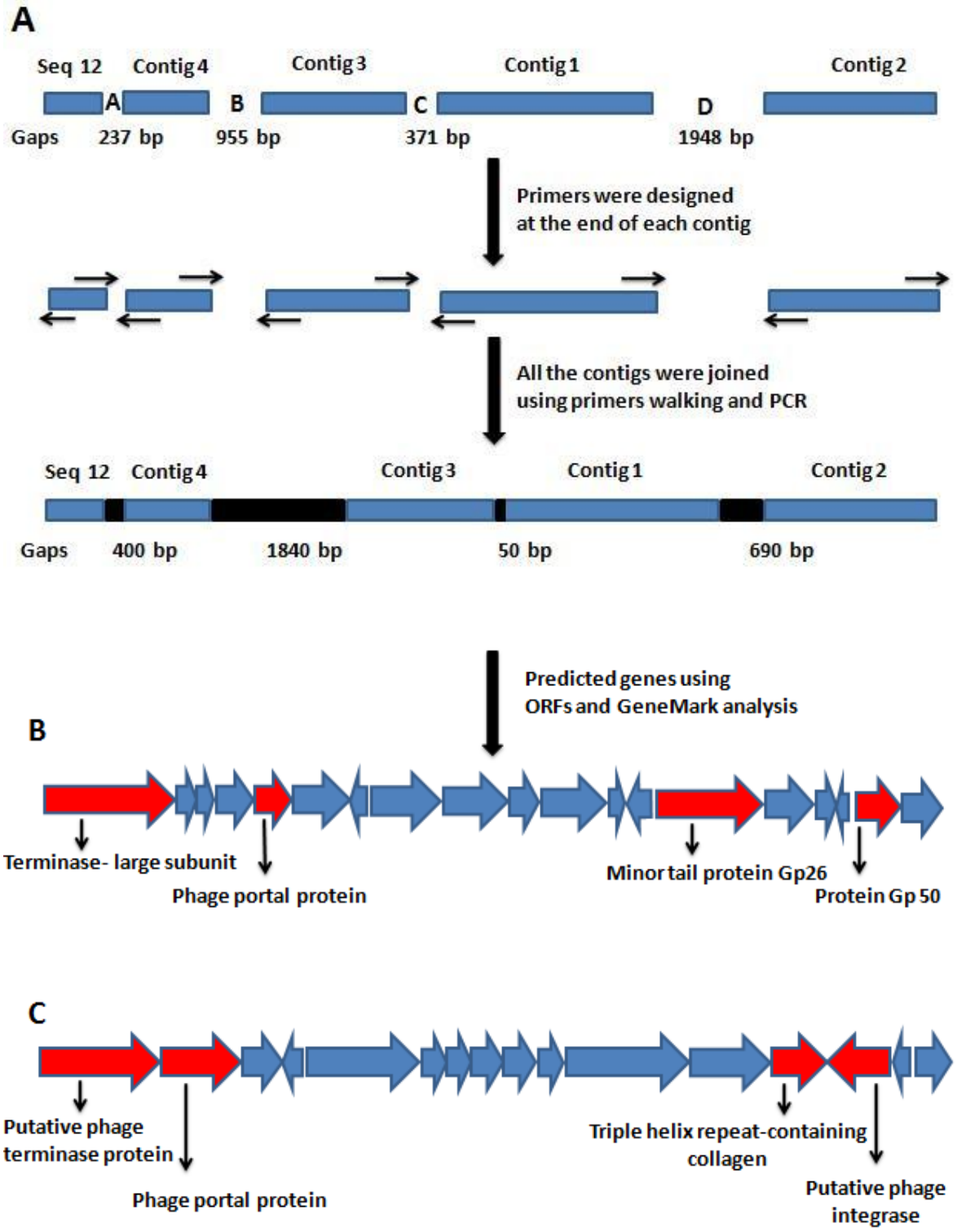


Figure 3.5: The predicted contigs and gap filling of the overlapping sequences

Part A shows the order of the contigs that were based on similarities to the partial phage in the *Corynebacterium diphtheriae* genome using BLASTP and TBLASTX analysis. All the contigs were joined by primer walking and PCR sets. Part B shows the predicted genes of the joined contigs using ORFs and GeneMark; the red arrows represent genes of known function and the blue arrows those of unknown function. Part C shows the order of genes in the partial phage in the *Corynebacterium diphtheriae* genome.

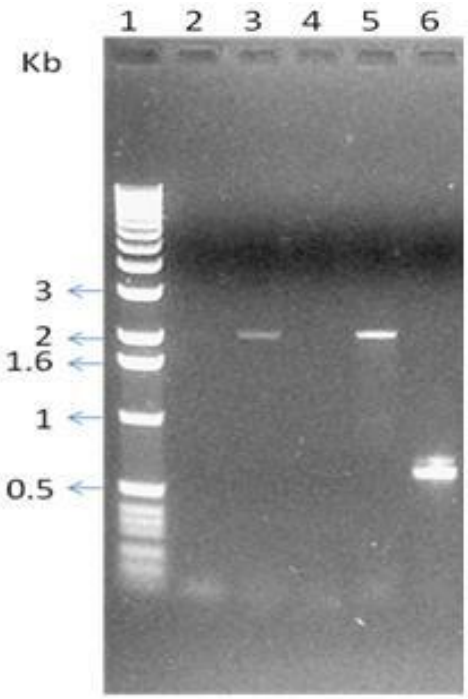


Figure 3.6: Gap filling using PCR reaction
PCR primer walking to fill sequence gaps in Figure 3.5.
Lane 1: 1 kb marker; lanes 2 and 4: no product was generated; lanes 3 and 5: same product of 2 kb was generated for gap B; lane 6: product for gap A.

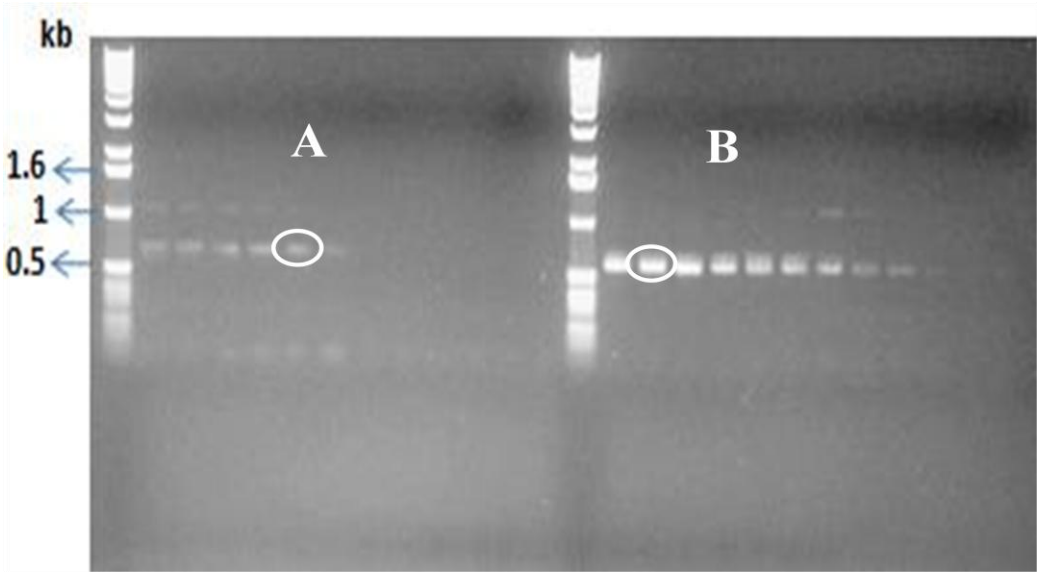


Figure 3.7: Gradient PCR reaction to avoid non-specific generated bands
A gradient PCR (section 2.11.1) was set up to avoid multiple non-specific bands. Band A and B represent the filling of gap D. The circled bands are sharp ones which were chosen to be sequenced. Bands at A had higher molecular weights than bands at B, due to the use of different primers sites.

Table 3.2: Sequence analysis of the partial phage of *Corynebacterium diphtheriae* using ORFs and TBLASTX analysis

ORF	Start	Stop	D	aa	E-value	Identities (%)	Function (domain)	Significant Matches (accession number)
1	70	1704	+	520	7e-34	31	Terminase – large subunit (COG4626)	C. *diphtheria* NCTC 13129 (NP_940162)
2	1733	1984	+	83	--------	--------	No match	
3	1988	2194	+	68	--------	-------	No match	
4	2229	2705	+	158	2e-14	35	Unknown	C. *diphtheria* NCTC 13129 (NP_940163)
5	2878	3321	+	147	2e-14	34	Phage portal protein	C. *diphtheria* NCTC 13129 (NP_940163)
6	3324	4061	+	245	11e-36	46	Unknown (pfam01510)	Lactococcus phage asccphi28 (YP_001687532)
7	4093	4287	-	64	--------	--------	No match	
8	4326	5201	+	291	4e-26	30	Unknown	C. *diphtheria* NCTC 13129 (NP_940166)
9	5290	6156	+	288	1e-06	27	Unknown	C. *diphtheria* NCTC 13129 (NP_940166)
10	6157	6525	+	122	-------	-------	No match	
11	6535	7362	+	275	22-24	46	Unknown	C. *diphtheria* NCT C 13129 (NP_940169)
12	7359	7586	+	75	-------	-------	No match	
13	7652	7990	-	112	-------	-------	No match	
14	7881	9227	+	448	1e	29%	Minor tail subunit	Mycobacterium phage D29 (NP_046842)
15	9286	9927	+	213	3e-08	30%	Unknown	*Clostridium hiranonis* DSM 13275 (ZP_03292433)
16	10116	10364	+	82	-------	-------	No match	
17	10430	10603	-	57	-------	-------	No match	
18	10597	11166	+	189	3e-04	50%	gp50	Acidianus two-tailed virus (YP_319881)
19	11034	11553	3	173	-------	-------	No match	

D = Direction of translation
aa = Number of amino acids

3.7: Sequence Identities in Samples from the Second Volunteer

As indicated, only 10 sequences were obtained. BLASTN sequencing was done to check whether there were any matches to the GenBank databases. Two sequences, B3 and B4, (Table 3.4) overlapped to form a contig (the overlapped length was 478 bp with an identity of 98%). This contig had a significant identity (E= 9e-139 with an identity of 77% over 838 nucleotides) to a genomic sequence from *Ralstonia solanacearum* (accession number AL646053). Sequence B6 also had a short match (E= 7e-47 with an identity of 100% over 105 nucleotides) to a partial 16S rRNA gene of *Pseudomonas acephalitica* (accession number AM407893). BLASTX was also used to detect the identities of these sequences. Two sequences (B2 and B6) and one contig (B3 and B4) of the second volunteer had significant matches to the GenBank database (Table3.4).

Table 3.4: Sequences analysis of the second volunteer using TBLASTX

Seq number	Seq length	Matched length bp	E-Value	Identity (%)	Nearest match to the GenBank identified by TBLASTX
B1	812				No match detected
B2	765	213	6e-19	38%	Corynebacterium phage P1201
B3 +B4	1269	807	0.0	76%	Anaerobic ribonucleoside triphosphate reductase, *Acidovorax sp*. JS42
B5	762				No match detected
B6	537	105	2e-14	100%	partial 16S rRNA gene of *Pseudomonas acephalitica*
B7	810				No match detected
B8	783				No match detected
B9	538				No match detected
B10	811				No match detected

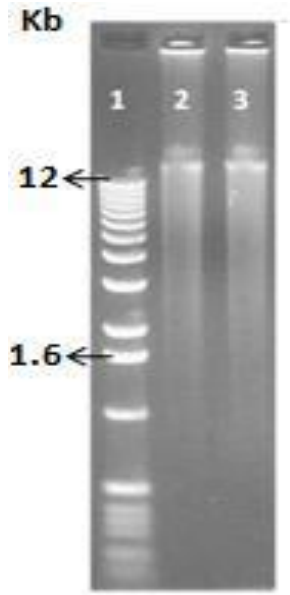


Figure 3.8: Amplified viral genomes of the second volunteer using the MDA method
Lane 1: 1 kb marker
Lane 2: control reaction
Lane 3: amplified viral genomes

3.8: Sequence Identities of the Third Volunteer

Only 11 clones were sequenced for the third volunteer. No matches were detected against the GenBank database using BLASTN analysis, which indicates that these sequences are highly novel and not contaminated with other known genomic DNA rather than virus genomes. BLASTX analysis was done for the 11 sequences of the third volunteer, showing significant matches to 4 different phages, a plasmid and the genomic DNA of *Streptomyces avermitilis* (Table 3.5). All the sequences of the three volunteer were aligned against each other, however, no similarity match was found.

Table 3.5: Sequence analysis of the third volunteer using TBLASTX

Sequence number	Sequence length pb	Matched length bp	E-Value	Identity (%)	Nearest match to the GenBank identified by TBLASTX
C1	409				No match detected
C2	483				No match detected
C3	629	162	1e-05	44%	Oligomycin biosynthetic gene of *Streptomyces avermitilis*
C4	617				No match detected
C5	520				No match detected
C6	602	213	1e-17	43%	Genomic DNA, Rhodococcus phage YF1
C7	537	303	2e-24	39%	Genomic DNA, Lactococcus phage 1706
C8	491	375	3e-30	43%	Plasmid pAYS DNA, *Nitrosomonas* sp.
C9	610				No match detected
C10	427	225	7e-10	41%	Phage-related terminase, *Leifsonia xyli* subsp. xyli str. CTCB07
C11	456	174	8e-19	67%	Genomic DNA, Mycobacterium phage Kostya

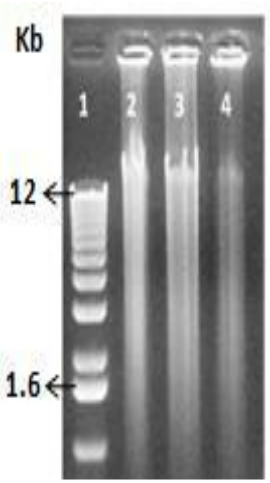


Figure 3.9: Amplified viral genomes of the third volunteer using the MDA method
Lane 1: 1 kb marker
Lane 2: control
Lane 3: amplified viral genomes
Lane 4: no input template DNA; still showed background

3.9: Conclusion

The new culture-independent methods have solved many problems of accessing the viral communities in a variety of samples. The six earlier metagenomic viral libraries were constructed using the LASL method, which required large samples. The average percentage of unknown sequences of these libraries against the GenBank databases is about 67% (75-59) when they were initially analyzed. However, the reported metagenomic dental plaque viral library discussed in this part of the thesis was derived from samples of limited size (human dental plaque) which were then amplified using the MDA method. The unknown sequences with reference to the GeneBank databases were found to constitute more than half (55%). This result shows that the percentage of unknown viral sequences was still high, even though the GenBank non-redundant database had doubled in size (Delwart 2007; Edwards and Rohwer 2005).

All the viral sequence libraries received were annotated and were found to contain novel sequences, except for sequences that had no similarity to the GenBank database. However, these sequences could be novel sequences if they were not artefact sequences that had been generated by the MDA method. Thus, use of the MDA method created doubt about the origin of the unknown sequences, and the effect of artefact sequences could be reduced, as indicated above.

Most viruses were found to be related to phages in the dental plaque viral library as well as to other published metagenomic viral libraries. The *Siphoviridae* and prophage members were strongly represented, forming 57% of the total viral matches in the dental plaque library. These were also identified strongly as ~80% in the faecal (Breitbart *et al.*, 2003) and marine sediment (Breitbart *et al.*, 2004) libraries. However, they were found to constitute fewer than 50% of the total viral matches in the seawater libraries. On the other hand, the members of the *Myoviridae* and *Podoviridae* families were found to be as high as 83% of the total viral matches of the Chesapeake Bay Virioplankton library (Bench *et al.*, 2007).

It was hoped that this library would allow us to estimate the viral diversity in the human mouth and characterise the viral genes present. The viral diversity in this library appears to be very low: according to the Shannon Index calculation, the value of species diversity

and evenness is 1.9, which is lower than the reported virus libraries. Even if we assume all the clones in this library are virus related it appears that the library only contains the equivalent of 5 or so bacteriophages. One might reasonably assume that the ~1000 bacterial species occupying the oral cavity have ~1-10, 000 associated bacteriophage types. Why this obvious discrepancy is unclear. Perhaps the most reasonable suggestion is that the source of DNA used in this work was limiting. It could be scaled up considerably by using pooled volunteer samples. Future work could then compare results obtained with MDA and other nucleic acid amplification methods to results obtained using unamplified DNA samples. It would be very surprising if such work did not identify a much richer virus flora associated with the oral cavity.

As well as identifying a low population complexity in our analysis we also identified many clones related to a putative phage from *Corynebacterium diphtheriae*. We might conclude that this phage is real and was actively lytic and at relatively high titre (compared to other viruses) in the sample we determined. This itself may account for the low complexity of our library 80 unique sequences would have led to very different numbers and the conclusion of high sample richness. It would appear that most viruses in the mouth are phage related and most are unknown, but much more work is required to confirm all these hypotheses.

Chapter 4

Characterization of Two Lytic Bacteriophages Isolated from Human Dental Plaque

Chapter 4

4: Results and Discussion

4.1: Detecting Lytic Phages in the Human Mouth

Lytic phages probably play an important role in the ecology of the human mouth because they control and interact with the population of bacteria. They could also be used to treat bacterial diseases (Skurnik and Strauch 2006), as some of the multidrug-resistant pathogenic bacteria have increased their resistance to a variety of available antibiotics (Fortuna *et al*., 2008; Jado *et al*., 2003; Wang *et al*., 2006). Few attempts have been made to isolate lytic phages from the human mouth (Hitch *et al*., 2004), and detecting more of them would increase our knowledge of oral lytic phages and their hosts, and would further increase our understanding of diseases and viral structures.

Detecting a lytic phage using culture-based methods, usually plaque assays, requires patience and careful examination of the small plaques formed by lysis of the bacterial host. It can be difficult to even see let alone distinguish between different plaques, and even if the plaques have the same morphological appearance different viruses may cause them. The examination of every plaque detected, however, is impractical.

More than one hundred different bacterial colonies were isolated from the mouth in this study, using morphological features such as colour, shape and size on blood, LB and brain heart infusion agar plates. The colonies were re-plated to ensure purity. All bacteria were assessed for their ability to form a lawn in soft-top agar. The majority of bacterial isolates formed lawns. Different types of soft-top, LB and brain heart infusion, which contained 0.35% w/v agar, were used. Soft tops were supplemented with horse blood to various percentages (2%, 3%, 5% v/v) for bacteria that grew better in the presence of blood.

Dental plaque containing saliva from volunteers was mixed, filtered and used to infect overnight cultures of the isolated bacteria (section 2.6.2); these infected cultures were plated in soft tops, resulting in many instances in the appearance of plaques. A single plaque was then used to re-infect the same host culture; however, this procedure produced plaques only at the first and second times of propagation, after which plaques did not appear. The reason why infection upon serial passage was so often lost is unknown but

fairly common. It would be interesting to find the cause of the change in the virus host interaction leading to the inability to form plaques. Possible reasons are lost receptors, host resistance and growth conditions affecting the host virus interactions.

One putative plaque-forming virus was an exception and continued to re-plaque on soft-top agar through multiple rounds of propagation. Its host, called Oral Isolated Bacterium (OIB), turned out to have two different plaque morphologies, caused by two viruses. This strain gave confluent lawns with no plaques when grown with low passage from -80°C cultures. Upon repeated passaging plaques spontaneously formed in this strain due to the activation of a prophage ; this was called A1 virus. The other plaque was due to a lytic phage and was called A2 virus. The lytic phage was isolated from saliva following the infection experiment.

4.2: Description of the OIB Strain

4.2.1: Identification and Taxonomic Classification

When the 16S rRNA gene was amplified from the OIB strain, the PCR product gave the expected band of approximately 1500 bp in size (Figure 4.1). This band was first sequenced with the forward primer. Analysis using BLASTN gave the first match in the list with 99% similarity to the uncultured *Neisseria* sp. clone EMP_C13 (accession number EU794238), and the first match in the list to a cultured bacterium was *Neisseria subflava* NJ9703 (accession number AF479578). Sequencing the amplified band with forward and reverse primers gave a read of 1365 bp (see nucleotide sequence below).

Analysis of the 1356 bp using BLASTN showed hits to many cultured *Neisseria* sp (Table C page 201). To show the relationships between the OIB strain and the matched database strains, a phylogenetic tree was created (Figure 4.2), based on the 16S rRNA gene sequence, 12 of which were chosen from the BLASTN database matches (Table 4.1). Four bacterial strains were also added to the tree, as three of them were used for the host range experiment (section 4.4), and in the case of *Neisseria gonorrhoeae* FA 1090, because most of the genomic A2 virus has highly significant matches to a phage in this strain. Based on the tree results, the closest match is to N. *subflava*, but bootstrap confidence on the N. *perflava* lineage is low, at 57%, so it could equally be most closely related to that. Indeed, the 73% bootstrap makes it possible that it could be most closely

related to any of the *Neisseria* sp. in this figure, including the pathogenic strains. The only exception is N. *lactamica* L 13AJ239305, which appears to be grouped alone.

Neisseria species are known to be among the bacteria that frequently exchange chromosome genes. Attempts were made to analyse about fifty isolates of human commensal *Neisseria* species, including the pathogenic N. *meningitidis* and N. *gonorrhoeae*, using specific genes such the 16S rRNA, recA, argF and rho genes. These were found to fall into five phylogenetic groups, supported by high bootstrap values; however, the phylogenetic relationships among these groups, based on the gene analyzed, were found to be varied (Smith *et al.*, 1999).

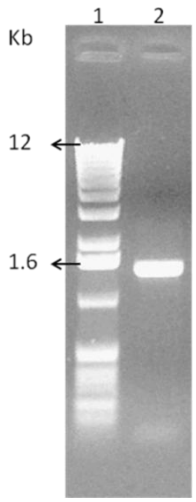


Figure 4.1: Amplified 16S rRNA from OIB strain
Lane 1: 1 kb marker
Lane 2: amplified 16S rRNA from OIB strain

The amplified 16S rRNA gene sequence (5'-3') from OIB strain

```
5′  1 TGGCGAGTGG CGAACGGGTG AGTAATATAT CGGAACGTAC CGAGTAATGG
   51 GGGATAACTA ATCGAAAGAT TAGCTAATAC CGCATATTCT CTGAGGAGGA
  101 AAGCAGGGGA CCTTCGGGCC TTGCGTTATT CGAGCGGCCG ATATCTGATT
  151 AGCTAGTTGG TGGGGTAAAG GCCTACCAAG GCGACGATCA GTAGCGGGTC
  201 TGAGAGGATG ATCCGCCACA CTGGGACTGA GACACGGCCC AGACTCCTAC
  251 GGGAGGCAGC AGTGGGGAAT TTTGGACAAT GGGCGCAAGC CTGATCCAGC
  301 CATGCCGCGT GTCTGAAGAA GGCCTTCGGG TTGTAAAGGA CTTTTGTCAG
  351 GGAAGAAAAG GCTGTTGCTA ATATCGACAG CTGATGACGG TACCTGAAGA
  401 ATAAGCACCG GCTAACTACG TGCCAGCAGC CGCGGTAATA CGTAGGGTGC
  451 GAGCGTTAAT CGGAATTACT GGGCGTAAAG CGAGCGCAGA CGGTTACTTA
  501 AGCAGGATGT GAAATCCCCG GGCTCAACCT GGGAACTGCG TTCTGAACTG
  551 GGTGACTAGA GTGTGTCAGA GGGAGGTAGA ATTCCACGTG TAGCAGTGAA
  601 ATGCGTAGAG ATGTGGAGGA ATACCGATGG CGAAGGCAGC CTCCTGGGAT
  651 AACACTGACG TTCATGCTCG AAAGCGTGGG TAGCAAACAG GATTAGATAC
  701 CCTGGTAGTC CACGCCCTAA ACGATGTCAA TTAGCTGTTG GGCAACTTGA
  751 TTGCTTAGTA GCGTAGCTAA CGCGTGAAAT TGACCGCCTG GGGAGTACGG
  801 TCGCAAGATT AAAACTCAAA GGAATTGACG GGGACCCGCA CAAGCGGTGG
  851 ATGATGTGGA TTAATTCGAT GCAACGCGAA GAACCTTACC TGGTCTTGAC
  901 ATGTACGGAA TCCTCCAGAG ACGGAGGAGT GCCTTCGGGA GCCGTAACAC
  951 AGGTGCTGCA TGGCTGTCGT CAGCTCGTGT CGTGAGATGT TGGGTTAAGT
 1001 CCCGCAACGA GCGCAACCCT TGTCATTAGT TGCCATCATT AAGTTGGGCA
 1051 CTCTAATGAG ACTGCCGGTG ACAAGCCGGA GGAAGGTGGG GATGACGTCA
 1101 AGTCCTCATG GCCCTTATGA CCAGGGCTTC ACACGTCATA CAATGGTCGG
 1151 TACAGAGGGT AGCCAAGCCG CGAGGTGGAG CCAATCTCAC AAAACCGATC
 1201 GTAGTCCGGA TTGCACTCTG CAACTCGAGT GCATGAAGTC GGAATCGCTA
 1251 GTAATCGCAG GTCAGCATAC TGCGGTGAAT ACGTTCCCGG GTCTTGTACA
 1301 CACCGCCCGT CACACCATGG GAGTGGGGGA TACCAGAAGT AGGTAGGGTA
 1351 ACCGCAAGGA GCCCG 3′
```

Table 4.1: Some of the significant matches to the 16S rRNA gene sequence of the OIB strain

Accession number	Description	Query coverage	E - value	Max identity	Source of isolate
EU794238.1	Uncultured *Neisseria* sp. clone EMP_C13 16S ribosomal RNA gene, partial sequence	100%	0	99%	Fecal
EF512007.1	Uncultured bacterium clone P1D1-725 16S ribosomal RNA gene, partial sequence	100%	0	99%	Endotracheal aspirate (human)
EF511998.1	Uncultured bacterium clone P1D1-762 16S ribosomal RNA gene, partial sequence	100%	0	99%	Endotracheal aspirate (human)
AJ786809.1	*Neisseria* sp. R-22841 partial 16S rRNA gene, isolate R-22841	100%	0	99%	Commercial nitrifying inoculum
DQ279353.1	*Neisseria* sp. TM10_4 16S ribosomal RNA gene, partial sequence	99%	0	99%	Tuber magnatum
EF512003.1	Uncultured bacterium clone P1D1-542 16S ribosomal RNA gene, partial sequence	97%	0	99%	Endotracheal aspirate (human)
AY138232.1	Uncultured *Neisseriaceae* bacterium Sto1-2 16S ribosomal RNA gene, complete sequence	100%	0	99%	Human stomach biopsy
EF511861.1	*Neisseria perflava* 16S rRNA gene (partial), strain U15	97%	0	99%	Upper respiratory tract of human
AJ239279.1	*Neisseria flavescens* 16S rRNA gene (partial), strain LNP444	97%	0	99%	Upper respiratory tract of human
EF511915.1	*Neisseria subflava* NJ9703 16S ribosomal RNA gene, partial sequence	97%	0	99%	Upper respiratory tract of human
AF310565.1	*Neisseria meningitidis* strain M7724 16S ribosomal RNA gene, partial sequence	100%	0	98%	
AF310417.1	*Neisseria meningitidis* strain M8172 16S ribosomal RNA gene, partial sequence	100%	0	98%	

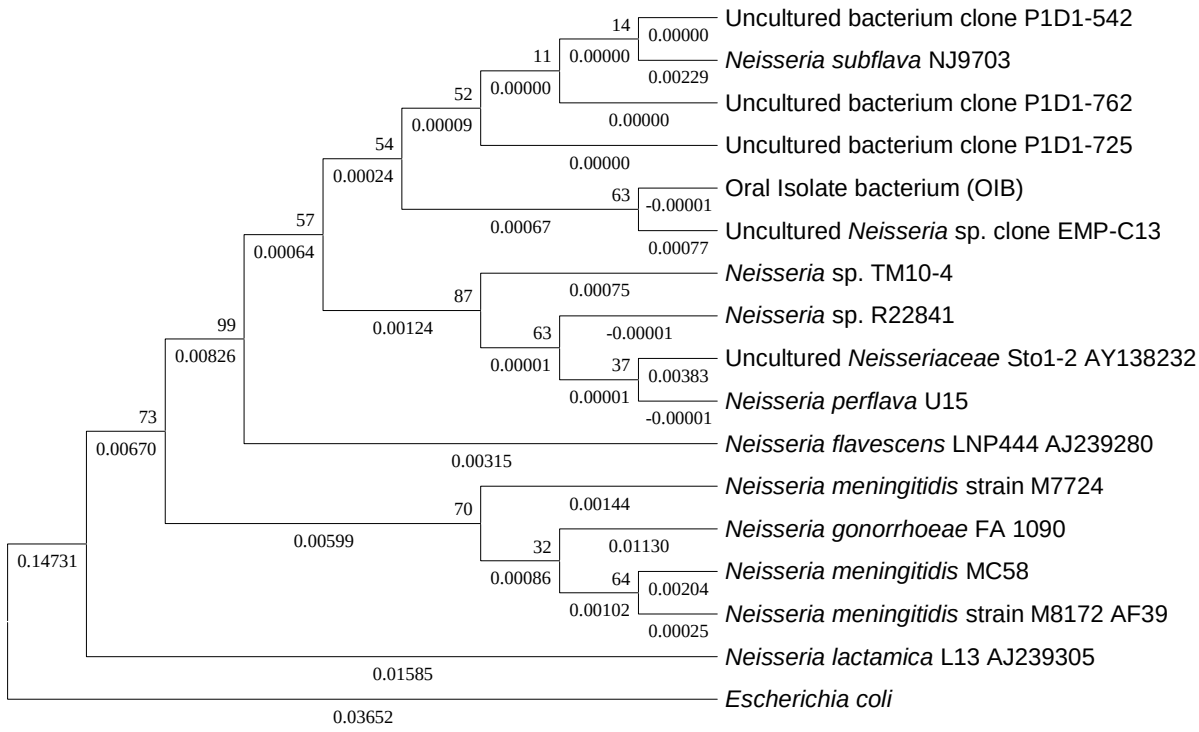


Figure 4.2: Phylogenetic tree showing evolutionary relationships of the OIB strain to 17 taxa.

"*The evolutionary history was inferred using the Neighbor-Joining method (Saitou and Nei 1987). The optimal tree with the sum of branch length = 0.25118560 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown above the branches (Felsenstein, 1985) (next to the branches). The evolutionary distances were computed using the Jukes-Cantor method (Jukes and Cantor 1969) and are in the units of the number of base substitutions per site. All positions containing gaps and missing data were eliminated from the dataset (complete deletion option). There were a total of 1312 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4"* (Tamura *et al*., 2007).

4.2.2: Comparison of the Phenotypes of the OIB Strain and *Neisseria subflava* (ATCC)

Based on the close relationship between OIB and *subflava* the latter was obtained from the ATCC. The OIB strain and *Neisseria subflava* (ATCC) showed some phenotypic differences, such as in colony morphology, when cultured on blood agar plates at 37°C for 24 hours (Table 4.2). The OIB strain requires further genotypic and phenotypic characterisation to identify its taxonomy. It may possibly be a pathogenic bacterium, or a commensal that plays an important role in human health.

Table 4.2: Comparison of OIB strain and *Neisseria subflava* (ATCC)

Features	Isolated Bacteria	Ordered *Neisseria subflava*
Colony colour	Yellow	White
Colony size	Bigger	Smaller
Growing	Faster	Slower
Shape	Entire, domed	Entire, domed
Negative control	Formed plaques	No plaques formed
Lawn	Formed	Formed
Gram stain	Negative	Negative
Hemolysis	α-hemolytic	Non- hemolytic

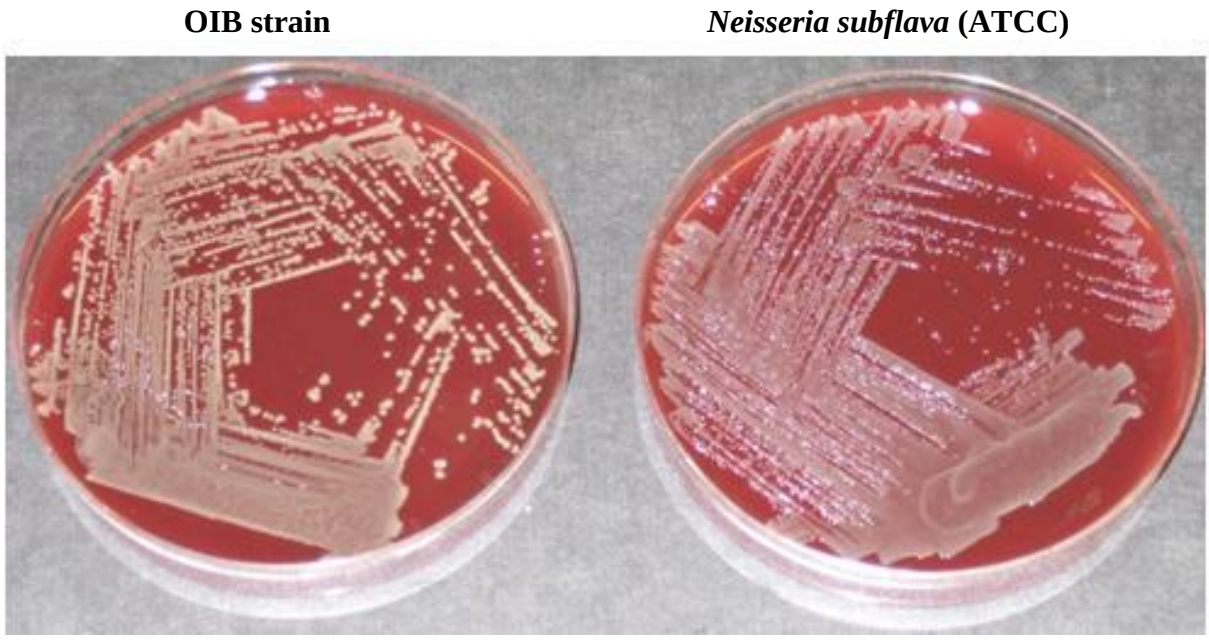


Figure 4.3: Colony morphologies of the OIB strain and *Neisseria subflava* (ATCC) bacteria on blood agar plates

4.2.3: Phage Typing of the OIB Strain and *Neisseria subflava* (ATCC)

When the 16S rRNA gene of the OIB strain had sequenced using the forward primer, *Neisseria subflava* NJ9703 was the first cultured bacterium in the matched list with 99% identity using BLASTN analysis. Thus it was ordered from the American Type Culture Collection (ATCC). It was then infected by the A1 and A2 viruses; however, no plaques were detected in the infected plates, nor did the negative control lawn show spontaneous plaques.

4.3: Description of the Two Isolated Viruses

4.3.1: Plaque Morphologies

4.3.1.1: Plaque Morphology of A1 Virus

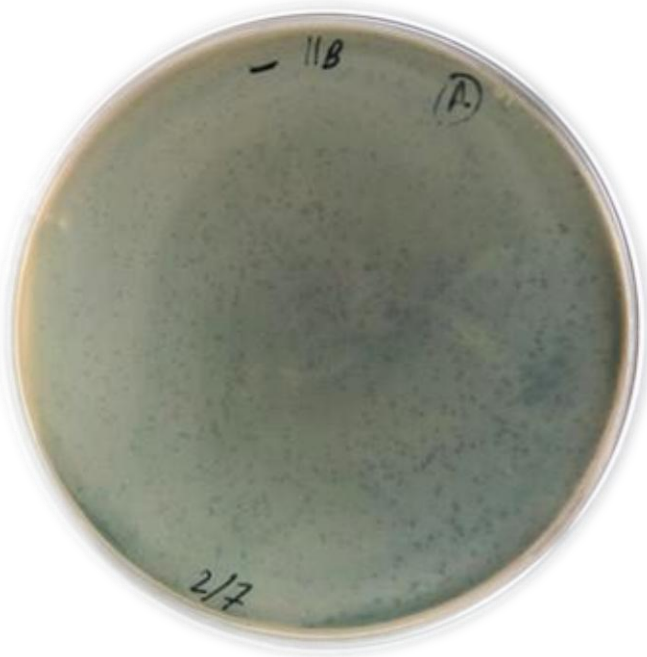
Plaques caused by the A1 virus occurred spontaneously on the soft-top agar after repeated passage of strain OIB. The sizes of plaque varied from 0.5 mm to 0.7 mm. These were visible on the soft tops after 8 h of incubation at 37°C. Some plaques were difficult to see because they were small and cloudy (total lysis had not occurred), while others had very clear lysis. The percentage of different plaque types varied from plate to plate. The total number of plaques in each plate also varied: some plates formed few while others had hundreds, even though the preparation of the plates was the same. Increasing the plaque number on the soft-top agar was impossible, as the plaques occurred spontaneously and there was no way to control their number. While some plates had a normal plaque distribution (Figure 4.4 I), others showed an unusual distribution; for example, in some plates plaques formed only on one side or only in the middle of the soft-top agar. Sometimes plaques formed a line from one side of the plate and spread to the other side, as shown in Figure 4.4 II. Previously it was thought that poor mixing could cause this; however, this was not the case, because the procedure was repeated many times with inversion and agitation of the soft top before pouring.

It is known that several genera of bacteria form autoplaques on confluent lawns, but some of the mechanisms that induce lysis remain unknown. Autoplaquing is a term used to distinguish the spontaneous occurrence of plaques on bacterial lawns where the host has

not been deliberately infected on a soft top agar as opposed to plaques that form on a sensitive strain after incubation with a virus (Breyen and Dworkin 1984). For example, two different strains of *Neisseria gonorrhoeae*, RUN5287 and RUN5290, form irregularly shaped autoplaques (medium lacking arginine causes these autoplaques). The cell density, the agar base and incubating temperature influence autoplaques, but phage induction agents such as UV, mitomycin C and ethylmethanesulfonate do not (Campbell *et al.*, 1985). It was found that the OIB strain often exhibited background plaques when plated from -80°C storage, grown in broth and used as a lawn; however, when repeatedly streaked from plate to plate and grown in broth, plaques of the A1 virus appeared. It is not clear if the A1 virus causes the plaques present on the soft-top agar, because this lysis could be due to some form of bacterial lysis and the A1 virus could be released in this process; but A1 is associated with these plaques.

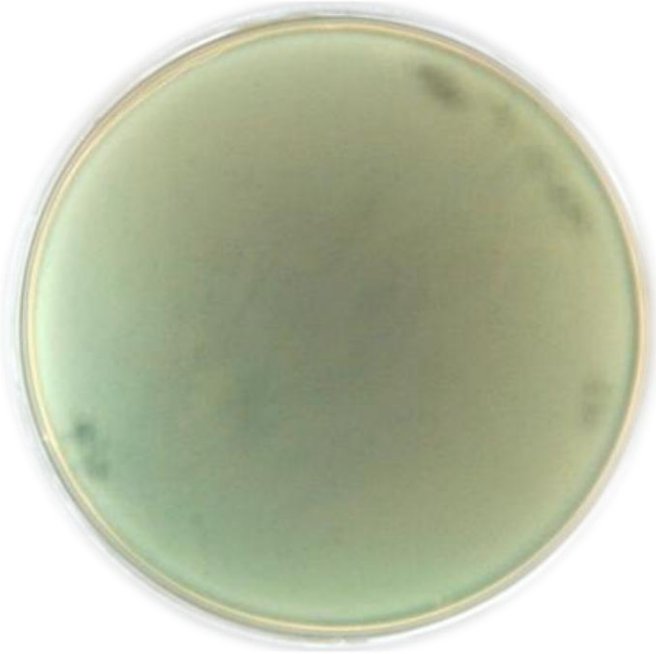
4.3.1.2: Plaque Morphology of A2 Virus

As the two viruses (A1 and A2) infect one host, the plaques of the A1 virus that occurred spontaneously on the soft-top agar could contaminate the plaques of A2 virus. It was found that the OIB exhibited background plaques when plated from -80°C storage, grown in broth and used as a lawn; however, when repeatedly streaked from plate to plate and grown in broth, plaques of the A1 virus appeared. So effectively strain O1B could be 'cured' of A1 infection by using and infecting single passage cultures with A2. The A2 virus seemed to be a typical lytic virus, as its titre was increased by the plaque assay method and the lysis formed by the virus was very clear, not cloudy. The size of the plaques on the soft-top agar varied from 1 mm to 1.5 mm (Figure 4.4 IV). Plaques caused by A2 virus were visible within 3 h of incubation on the soft tops at 37°C.

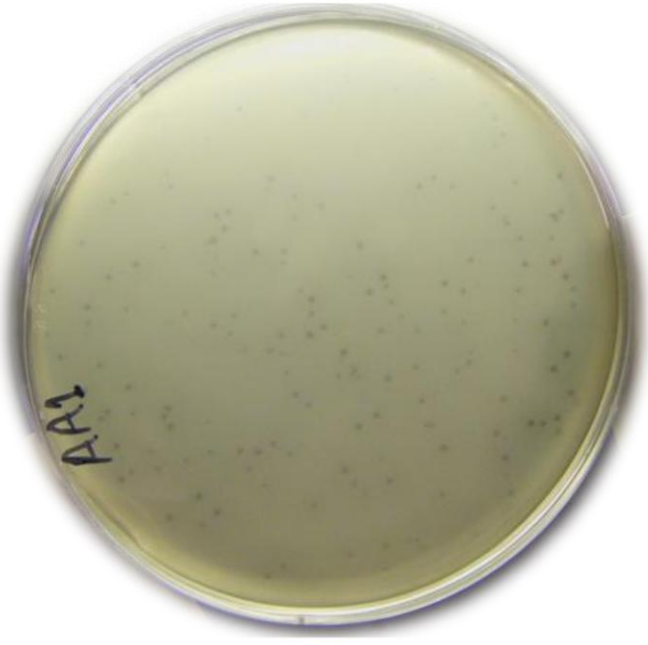


I
Plaque morphology of A1 virus, normal plaques.

II
Plaque morphology of A1 virus, strange plaques.



III
The OIB strain lawn



IV
Plaque morphology of A2 virus

Figure 4.4: Plaque morphology

4.3.2: Transmission Electron Microscopy Analysis of the Viral Particles

4.3.2.1: Virus A1

The A1 virus particles directly isolated from plaques were stained with uranyl acetate and visualized using the electron microscope (section 2.14). It appeared that the virus has an isometric head and no detectable tail, and the size of the virus particles varied from 32 nm to 58 nm (Figure 4.5). Figure 4.6 shows the A1 virus surrounding the host; more information is needed to confidently classify this virus. The difficulties in classifying it are the variation in the particle size and whether it contains an internal or external lipid layer. According to the International Committee on Taxonomy of Viruses (ICTV; http://www.ncbi.nlm.nih.gov/ICTV), the A1 virus may be a member of the *Tectiviridae* family. Viruses belonging to this family have internal lipids with a capsid size of about 60 nm. For example, Coliphage PR772 has a lipid membrane beneath the icosahedral shell with a genomic DNA size of 14,946 bp and a capsid size of about 63 nm (Lute *et al*., 2004). Another example of a member of the *Tectiviridae* family is Bam35, a bacteriophage that infects *Bacillus thuringiensis*, which has an internal lipid membrane and a linear genomic DNA with a size of about 15,000 bp. The phage Bam35 was found to be highly similar to the *Bacillus cereus* linear plasmid pBClin15 (Daugelavicius *et al.*, 2007, Stromsten *et al.*, 2003).

It is possible that the A1 virus is a plasmid- or satellite-like virus e.g., the satellite temperate phage P4, which requires the products of the helper phage P2 in order to grow lytically (Halling *et al.*, 1990). The capsid morphology of A1 virus is similar to that of the other viruses indicated, except that the size of the A1 virus varies. Therefore, the OIB strain was tested for the presence of plasmid using a commercial kit.

Three ml of overnight culture was extracted using a Promega Miniprep. Fifteen µl was run on 1% agarose gel electrophoresis and showed the presence of a high molecular weight band, >12,000bp (Figure 4.7 A). A larger volume, 100 ml, was then extracted using the Qiagen kit, which gave a higher concentration genomic DNA (Figure 4.7 B). No other experiments or sequencing were done to determine the origin of this band. So this strain contains a discrete dsDNA band that could be a plasmid or a virus genome.

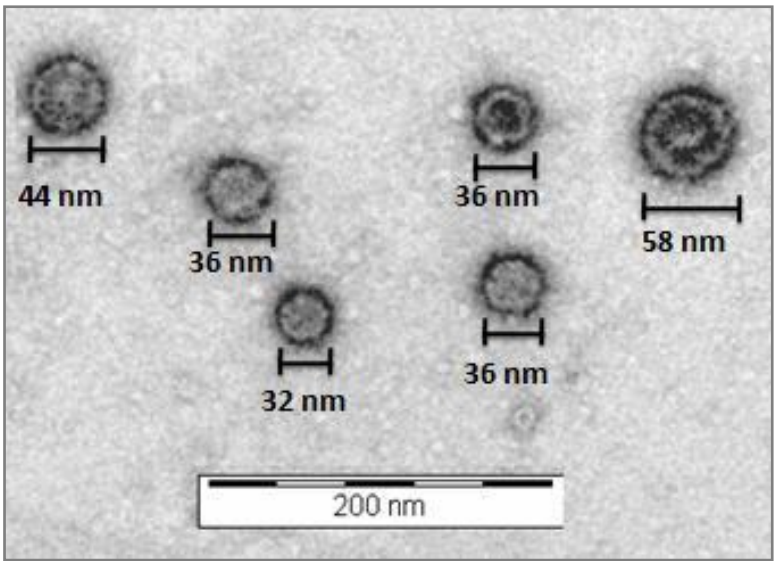


Figure 4.5: Electron micrograph of A1 virus
The size of the virus particles appeared to vary from 32 nm to 58 nm.

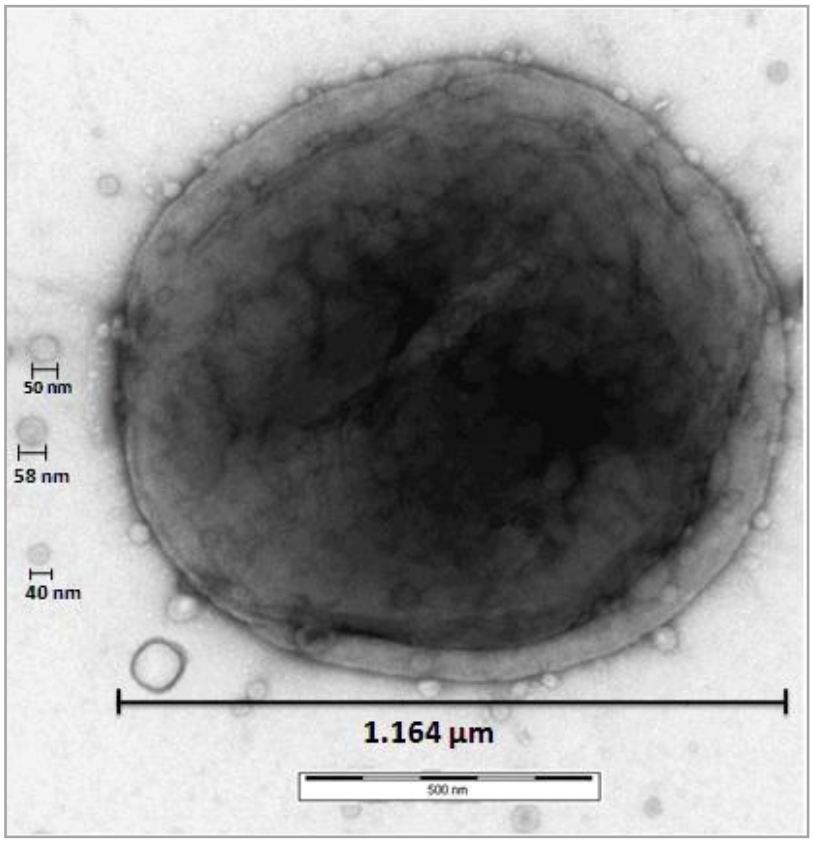


Figure 4.6: Electron micrograph of the A1 virus around the OIB strain

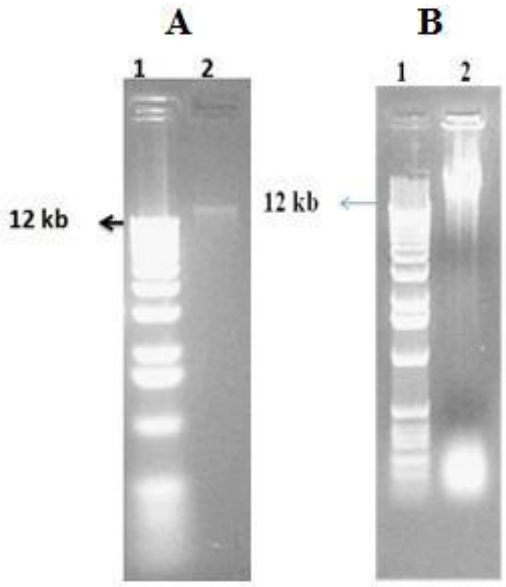


Figure 4.7: Plasmid preparation from the OIB strain
Lane A1: 1 kb marker
Lane A2: The plasmid extracted from 3 ml of overnight culture
Lane B1: 1 kb marker
Lane B2: The plasmid extracted from 100 ml of overnight culture

4.3.2.2: Virus A2

According to the ICTV (http://www.ncbi.nlm.nih.gov/ICTV), the A2 virus can be classified into the order *Caudovirales*, family *Siphoviridae*, based on the presence of an icosohedral head and a sheathed rigid tail with no base plates or tail fibres detected (Figure 4.8). The head size is 60±3 nm in diameter. It has a thick non-contractile tail 175 nm long, covered with sheath striations. No collar or connection neck is detectable. An inner tube was visible running from the head to the end of the tail. A few virus particles with a different morphology were detectable in the same sample taken from the one plaque. All the viruses, based on their morphology, belonged to the family *Siphoviridae* (Figure 4.9). The structural appearance of the isolated A2 virus particles were similar on 3 of the 4 occasions they were analysed by electron microscopy. On one occasion more than a single type of virus particle was detected. Since bacteria are known to harbour several types of phages, it is possible that the dissimilar phage structures seen could have been due to spontaneous activation of undetected latent lytic prophages in the OIB.

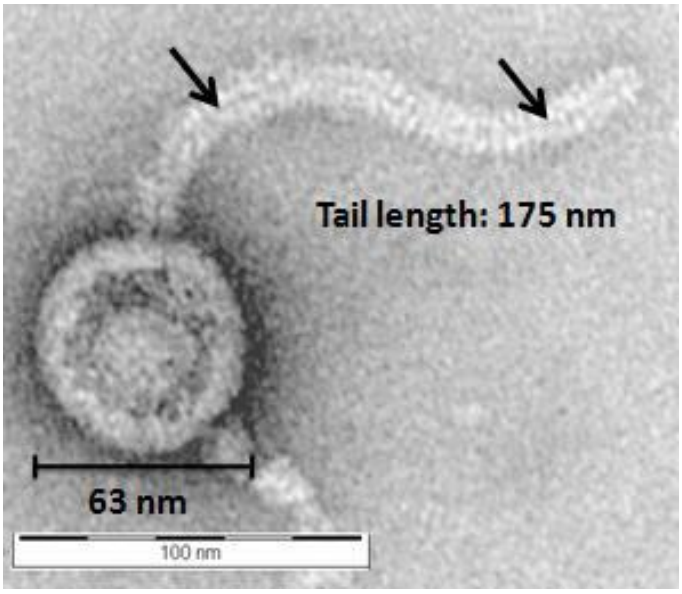


Figure 4.8: Electron micrograph of A2 virus
The black arrows show the inner tube, running from the head to the end of the tail.

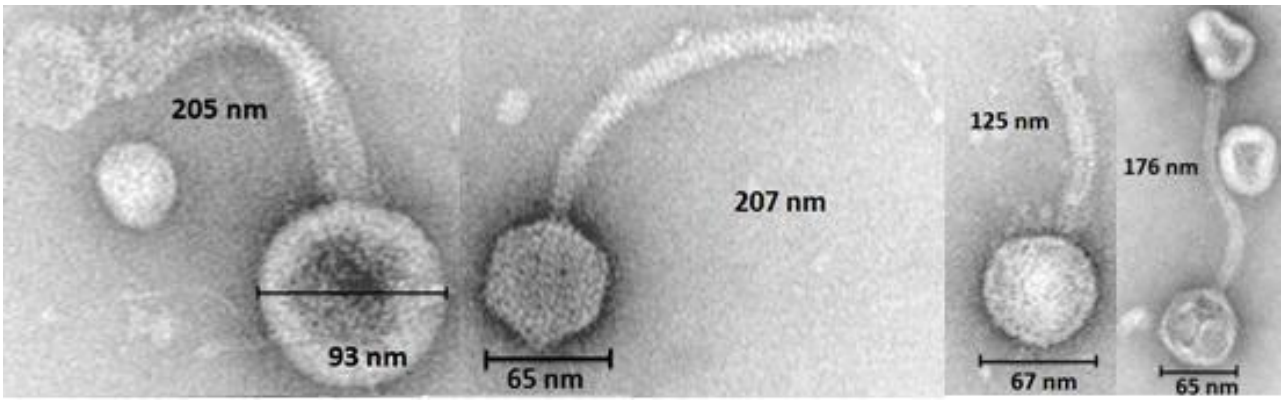


Figure 4.9: Electron micrograph of the other viruses present in the same plaque
The figure shows different sizes and lengths of capsid and tail that were detected in one plaque, which was caused by the A2 virus.

4.4: Further Host Range Studies of A1 and A2 Virus

Viruses A1 and A2 were tested for host range; neither appeared to lyse the *Neisseria meningitides* strains MC58 and 8047 or *Neisseria lactamica* (tested at the University of Leicester, Department of Genetics). It is of interest that no lytic phage has yet been isolated for *Neisseria meningitides* or *Neisseria gonorrhoeae* (Chanishvili *et al.*, 2001),

and presently no lytic virus has been detected for any *Neisseria spp.* The A2 virus could be the first lytic phage to be isolated that infects a species of *Neisseria*.

4.5: Single-Step Growth Curve for the A2 Virus

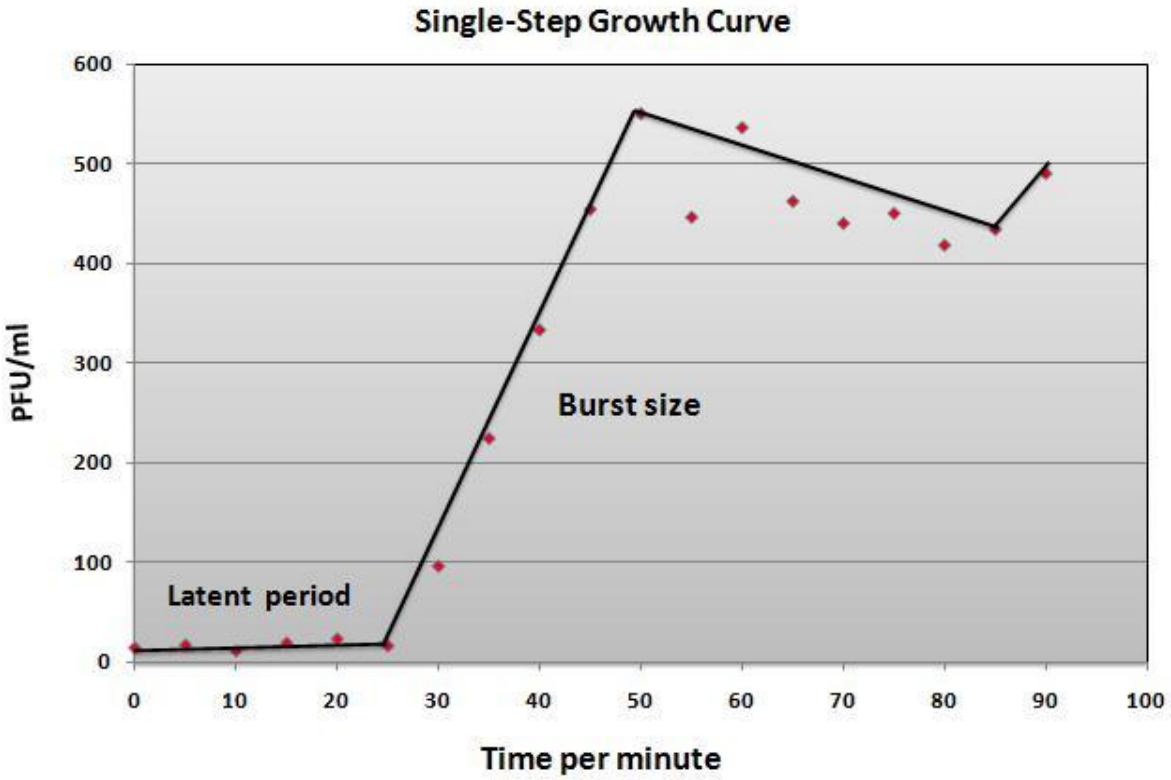
The latent period or eclipse and the burst size were determined for this virus using the single-step growth curve as described by Ellis and Delbruck, 1939. A very low multiplicity of Infection (MOI) was applied (section 2.6.5), to ensure host cells was infected with only one virus.

Calculation of the MOI:

The total number of phages was:
0.01 ml x (12×10^6) / ml = 12×10^4 PFU/ml
The total number of cells was:
0.5 x (145×10^7) / ml = 725×10^6 CFU/ml
Therefore the MOI was 6042 cells for each phage.

Sampling took place every 5 minutes up to 90 min (Figures 4.10 and 4.12) and 250 min (Figure 4.11). Based on single-step growth, as shown in the three figures, the phage demonstrated typical lytic phage characteristics with the bacterial host, i.e. a latent period lasting 25 minutes, followed by an increase in the number of virus particles. The latent period was observed to be 5 min less than the *E. coli* phage (Ellis and Delbruck, 1939) and 7 min longer than phage T4 which grows on *E. coli* (Hadas *et al.,* 1997). The average burst size of the three experiments (24, 25 and 27) was calculated as 25±2 virus particles per bacterial cell. There was no observed experimental variation when the single-step growth was repeated three times.

In the second single-step growth, the time was extended to determine the second burst size, but it did not level off, and the reason for this is not known. However, in the second and third experiments at time 50 to 85 min and 60 to 75 min respectively, the number of plaques on the soft-top agar could not be counted, due to contamination with other plaques, which were found to have similar morphology to the spontaneous plaques referred to above (Figure 4.4 II).

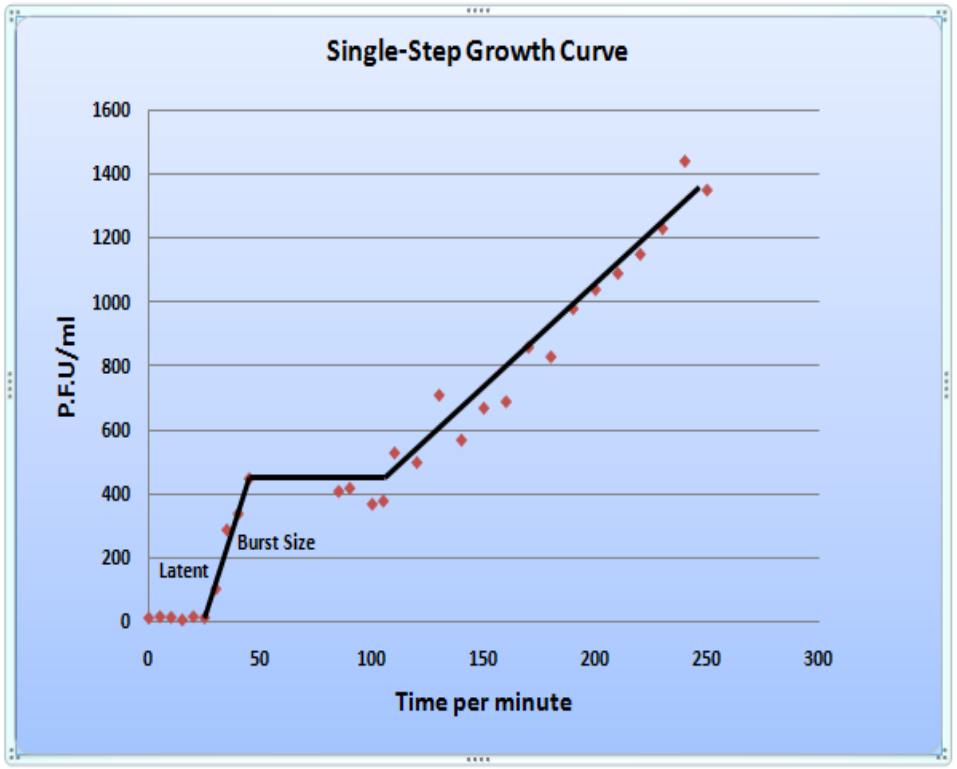


Time	Plaques no.
0	16
5	19
10	13
15	21
20	25
25	18
30	98
35	226
40	335
45	456
50	552
55	448
60	538
65	464
70	442
75	452
80	420
85	436
90	492

Figure 4.10: First A2 single step growth with the OIB strain.

This figure shows the latent period and one burst size, which were observed within 90 min. The latent period started at zero time and ended at 25 min of incubation. The average burst size was determined by dividing the average phage yield of the latent period by the average of the overall rise in phage numbers, as follows:

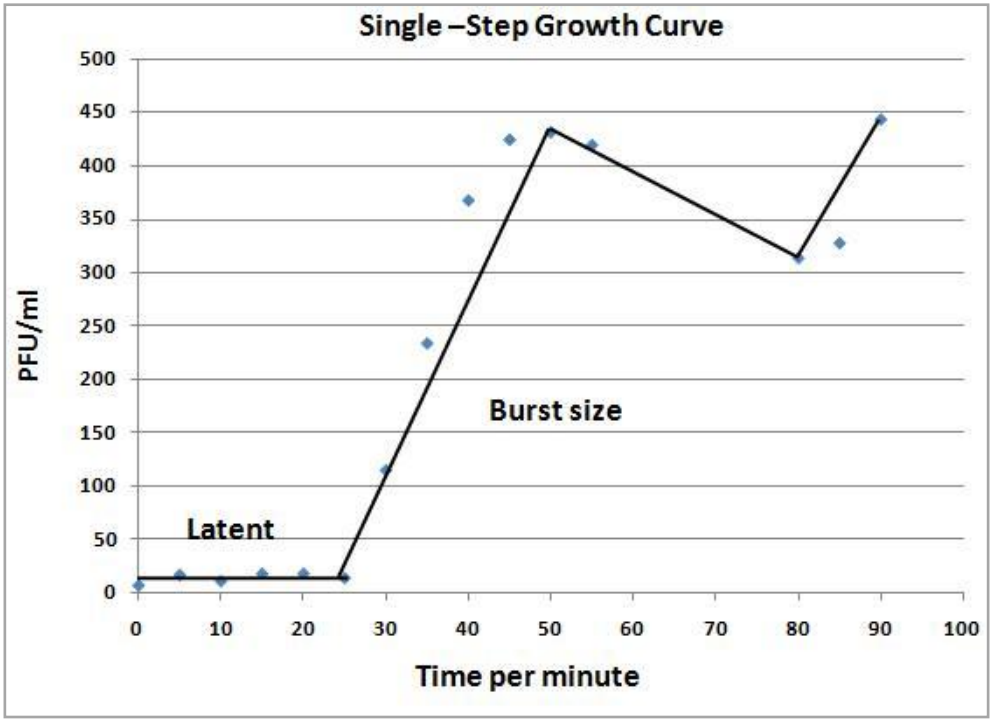
- Average number of plaques observed in the latent period:
  16 + 19 + 13 + 21 + 25 + 18 = 112 / 6 = 19 plaques
- Average burst size (time point 45- 80) = 450
  - ❖ Therefore: 450/19 = 24 particles per cell.



Time	Plaque number	Time	Plaque number
0	15	105	380
5	19	110	530
10	17	120	500
15	9	130	710
20	19	140	570
25	16	150	670
30	106	160	690
35	290	170	860
40	340	180	830
45	450	190	980
50		200	1040
55		210	1090
60		220	1150
65		230	1230
70		240	1440
75		250	1350
80		260	1860
85	410	270	2080
90	420	280	2310
100	370	290	2110

Figure 4.11: Second A2 single step growth with the OIB strain
This figure shows the latent period and two burst sizes, which were observed within 250 min. However, the second burst size did not level off, for unknown reasons which require further research. The latent period was 25 min.

- Average number of plaques observed in the latent period:
  15 + 19 + 17 + 9 + 19 + 16 = 95/6 = 16 plaques
- Average burst size (time point 85- 105) = 395
  - ❖ Therefore: 395/16= 25 particles per cell



Time	Plaques number
0	7
5	17
10	11
15	18
20	18
25	14
30	115
35	234
40	368
45	424
50	428
55	417
60	
65	
70	
75	
80	308
85	328
90	444

Figure 4.12: Third A2 single step growth with the OIB strain.

This figure shows the latent period and one burst size, which were observed within 90 min. The latent period started at zero time and ended at 25 min of incubation. The average burst size was determined by dividing the average phage yield of the latent period by the average of the overall rise in phage numbers, as follows:

- Average number of plaques observed in the latent period:
  7 + 17 + 11 + 18 + 18 + 14 = 85 / 6 = 14 plaques
- Average burst size (time point 428- 308) = 384
  - ❖ Therefore: 384/14 = 27 particles per cell.

4.6: Genome Characterisation, Type and Size

4.6.1: Genome Extraction and Nucleic Acid Characterisation of A1 Virus

Increasing the number of A1 virus particles was difficult. Many methods were applied to increase the titre of this virus. The first used broth media to increase the number of virus particles: a colony was picked from blood agar plate and grown in 10 ml of LB broth, which was incubated aerobically overnight at 37°C with shaking at 150 rpm in an orbital shaker. In the same growing conditions, 500 µl of this growth was added to 500 ml of LB broth containing 10 mM $MgSO_4$. The bacterial cells were then precipitated from the total sample, the virus particles in the supernatant were precipitated with PEG 6000 (section 2.9.2.1) and the viral genome was extracted with an equal volume of phenol: chloroform (section 2.10.1).

The second method used soft-top agar to increase the number of the virus particles. As indicated, the plaques of A1 virus occurred spontaneously on the host lawn. In this case, growing cells of the host were added directly to the molten soft-top agar with no infection step. Twenty to twenty-five soft-top plates were used, because the numbers of plaques varied from plate to plate, even under the same preparation conditions. The resulting plates had plaques ranging from a few to hundreds in number (a few plates were half lysed). The soft-top agars were scraped off from those plates which showed a high number of virus plaques (see sections 2.6.3 and 2.9.2.1 for virus particle precipitation). The viral genome was extracted repeatedly using the precipitation method with PEG 6000 and extraction with an equal volume of phenol: chloroform; however, all the results gave a low concentration of sheared viral genome (Figure 4.13).

The Viraprep Lambda kit (section 2.9.2.2) overcame these problems by increasing the amount of viral genome and producing a viral genome band without shearing and with an apparent increase in size (Figure 4.14). One plate with confluent lysis and 10 ml of the buffer supplied with the kit were required to extract the viral genome, but obtaining confluent lysis was difficult, as the number of plaques could not be increased. However, the problem of shearing of the viral genome was solved, and about 80 ng of the viral genome was extracted (Figure 4.14), which was enough DNA to cut into small fragments

and clone into a vector. The size of the A1 viral genome was estimated at between 12 kb and 23 kb by comparison with the size marker.

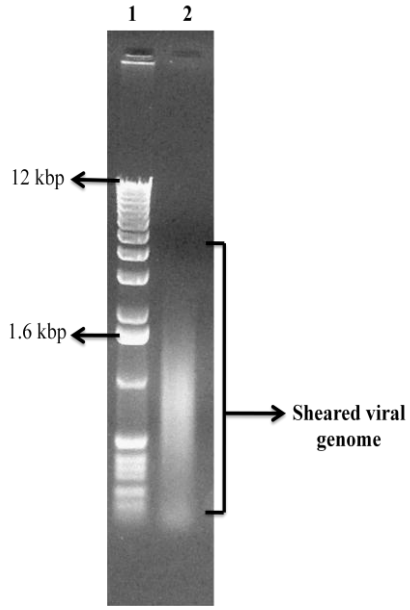


Figure 4.13: PEG precipitation of the A1 genome
Lane 1: 1 kb marker; lane 2: A1 genome precipitated by PEG 6000 and extracted with equal volume of phenol: chloroform.

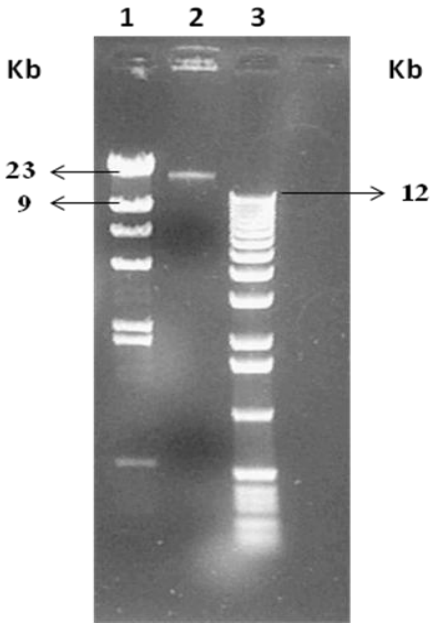


Figure 4.14: Extraction of A1 viral genome using the Viraprep Lambda kit
Lane 1: Hind III digest; lane 2: the A1 viral genome extracted by using the Viraprep Lambda kit, which was between 12 and 23 kb; Lane 3: the 1 kb DNA marker.

The A1 virus had a DNA genome, because it was digested by DNase, but not RNase. It was a linear genome, because it cut with Exonuclease III, and a double-stranded genome, as it cut with restriction endonucleases. Eight restriction endonucleases (four- and six-base cutters) were initially used to cut the A1 virus genome, but only two four-base cutters successfully cut the genome (Figure 4.15).

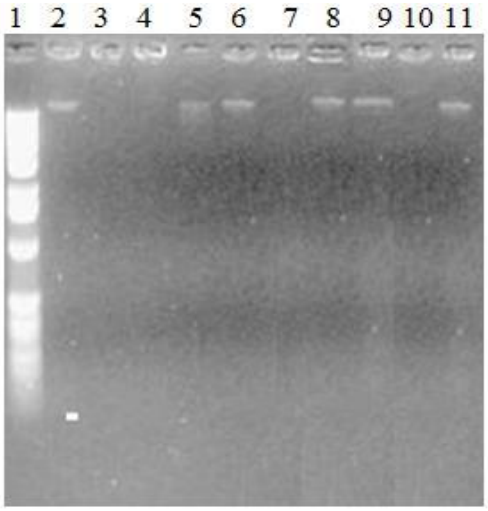


Figure 4.15: Digestion of the A1 virus genome with nucleases
Lane 1: 1 kb marker; lane 2: the A1 genome; lane 3: DNase digest; lane 4: Exonuclease III digest; lane 5: *Eco*RI; lane 6: *Bam*HI; lane 7: *Hae*III; lane 8: *Msp*I; lane 9: *Eco*RV; lane 10: *Sau*3A1; lane 11: *Xba*I. The *Not*1 enzyme was also tested, and did not cut but is not shown here.

4.6.2: Genome Extraction and Nucleic Acid Characterisation of A2 Virus

The nucleic acid of the A2 virus was easier to extract, because a high titre of plaques on the soft-top agar could be achieved. Fifteen plates of soft-top agar exhibiting confluent lysis were pooled into a 250 ml Sorvall tube. The virus particles were able to diffuse from the soft agar into the buffer (section 2.9.2.1) and were then precipitated using PEG 6000, followed by extraction of the virus particles with an equal volume of phenol: chloroform. Shearing of the viral genome occurred when it was electrophoresed on 1% agarose gels. This problem was solved by heating the extracted top aqueous layer at 70°C for 20 min immediately after the phenol: chloroform step, presumably by denaturing any nuclease activity. About 1-2 µg of the viral genome was extracted (Figure 4.16). The size of the A2 genome appeared to be the same as that of the A1 genome, which was immediately above the 12 kb marker (Figures 4.14).

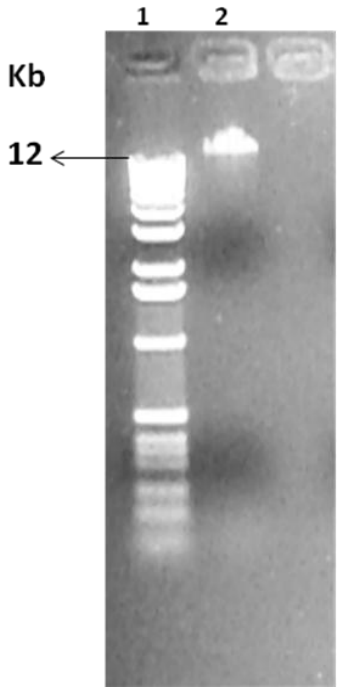


Figure 4.16: PEG precipitation of the A2 genome
Lane 1: 1 kb marker; lane 2: The A1 viral genome was precipitated with PEG 6000 and extracted with equal volume of phenol: chloroform, heated at 70°C for 20 min. 5 µl was run on 1% agarose gel electrophoresis. The virus band was set closely above the 12 kbp marker.

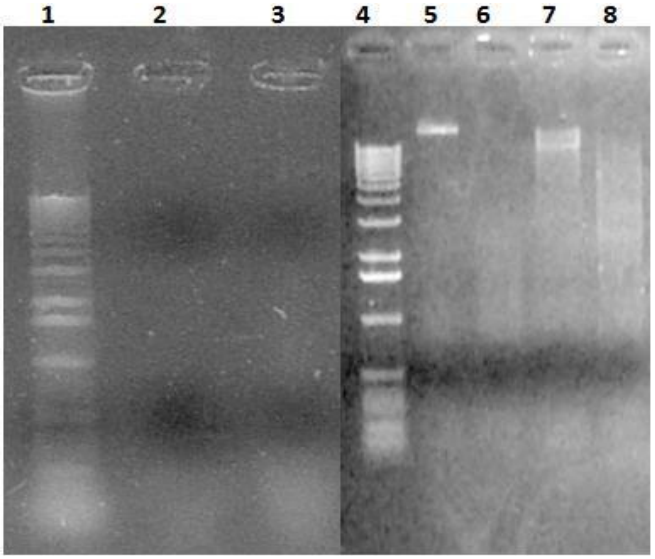


Figure 4.17: Digestion of the A2 virus genome with nucleases
Lane 1: 1 kb marker; lane 2: DNase digest; lane 3: Exonuclease III digest; lane 4: 1 kb marker; lane 5: the A2 virus genome; lane 6: *Dra*I; lane 7: *Eag*I; lane 8: *Mfe*I.

4.6.3: Pulse Field Gel Electrophoresis (PFGE) for A1 and A2 viruses

A low range PFGE molecular weight marker was used to find the exact size of the A1 and A2 viruses. The preparation method to increase the number of particles was different for the A1 and A2 viruses. For the A1 virus, one colony of the host on blood agar plate was inoculated into 20 ml of LB broth and incubated aerobically overnight at 37°C with shaking at 150 rpm in an orbital shaker. The sample was centrifuged to pellet the cells, then the virus was precipitated from the supernatant using PEG 6000 (section 2.9.2.1). The pellet was dissolved in SM buffer, and the PEG 6000 was removed by adding an equal volume of chloroform and mixing well, then centrifuged at 1200 x g for 10 min. The top layer was transferred to a fresh tube containing the A1 virus particles. In the case of the A2 virus, the particles were concentrated using soft-top agar (section 2.6.3). 40 µl of each concentrated virus preparation was run on a 1% PFGE agarose (BioRad) (section 2.10.2). Figure 4.18 shows that the A1 virus genome, band 2, appeared to be low in concentration and smeared. As indicated, the A1 virus is sensitive to chloroform, so the genome was degraded when run on gel. From this result, the size of the A1 viral genome could not be calculated. However, the A2 viral genome, band 3, appeared to be sharply defined, just under the 48.5 and above the 23.1 kb marker.

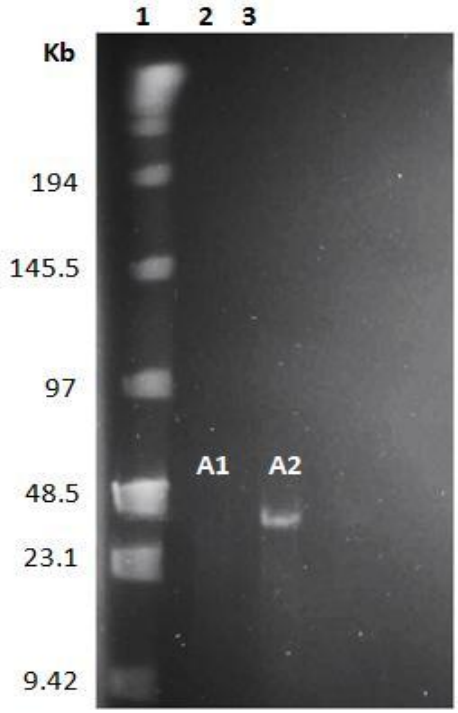


Figure 4.18: Pulsed field gel electrophoresis of the virus A1 and A2 genomes

4.7: Cloning and Sequencing the A1 Viral Nucleic Acids

Only two enzymes, *Hae*III and *Sau*3A1 (four-base cutters), were found to digest the A1 viral genome (Fig 4.15). The viral genome was sheared after cutting with these enzymes. Reduction of the viral genome to small fragments occurred, presumably because the genome has many recognition sites for these two enzymes. To obtain larger fragments, the incubation time of digestion was decreased from 30 min to 10 min at 37°C. The digested genome was run on a 1% agarose gel, and fragments between 1 kb and 5 kb were recovered from the gel slice using a gel extraction kit (section 2.10.5). The recovered fragments were cloned using the pGEM-T Easy vector (section 2.12). Twenty-four clones were sent to be sequenced by AGOWA in Germany (see Figure 4.20 for the strategy of sequencing a dsDNA viral genome).

4.7.1: Sequence Analysis of A1 Virus

The 24 sequences (see appendix for the nucleotide sequences) were searched first against the GenBank database using BLASTN analysis (Table 4.3). It was found that 8 of 24 sequences (2, 8, 10, 24, 25, 26, 35, and 45) matched significantly to bacterial genomes, except sequence 24, which matched to bacteriophage MM1, which infects *Streptococcus pneumonia*.

Table 4.3: Nucleotide Sequence analysis of the A1 virus clones using BLASTN

Sequence number	Sequence length	LMN	E-value	Identity (%)	Nearest match to GenBank using BLASTN
2	861	819	0.0	91%	*Neisseria meningitidis* serogroup C FAM18
8	574	571	0.0	95%	*Neisseria meningitidis* strain Z4756 UvrA
10	659	657	0.0	97%	*Neisseria gonorrhoeae* NCCP11945
24	616	368	2e-48	77%	Bacteriophage MM1, *Streptococcus pneumoniae*
25	552	550	0.0	93%	*Neisseria gonorrhoeae* NCCP11945
26	620	618	0.0	92%	*Streptococcus thermophilus* LMG
35	499	499	0.0	95%	*Neisseria sicca pilin*
45	244	238	3e-101	94%	*Neisseria meningitidis* serogroup C FAM18

LMN: length of matched nucleotides

Then ORFs and TBLASTX programs were used to try to identify the 24 sequnces of the A1 virus. Table 4.4 summarises the features of the 24 sequences received, all of which matched to bacterial genomes except sequences 1, 17, 42, 37 and 41. Sequence 2 was matched to a plasmid protein of *Streptococcus pneumoniae* CDC3059, sequence 17 to a hypothetical protein of Streptococcus phage SM1 and sequence 24 to a hypothetical protein of Streptococcus phage MM1. Sequences 37 and 41 had no matches to the database using ORF and TBLASTX analysis. No contig was formed when the 24 sequences were assembled using the Lasergene SeqMan version 7.0 program (DNAStar).

Since most of the sequences had matches to bacteria genomes, no more clones were then sent to be sequenced. The origin of these sequences will not be known until the viral genome is completed. A reason for the bacterial matches might be the contamination of the virus genome with bacterial genome because of host DNA also being extracted. The Viraprep does not contain DNase, as according to the manufacturer's protocol the host genome is absent from the sample, because washing with buffer removes it (see section 2.9.2.2).

The presence of sequences matching bacterial genomes which may or may not be contaminant DNA makes it difficult to characterise and identify the A1 viral genome. The identities of the 24 sequences and whether or not they belong to the A1 virus will not be known until the A1 virus genome has been completely sequenced.

Table 4.4: Sequence analysis of the A1 virus clones using ORFs and TBLASTX

Seq. number	Seq. Length	aa	E-value	Identity (%)	Predicted function (domain)	Significant matchs (accession number)
1	852	123	1e-32	54%	Plasmid protein (pfam09643)	S. *pneumoniae* CDC3059 (ZP_01833708)
2	861	242	9e-13	497%	Alcohol dehydrogenase class-III (pfam08240)	N. *meningitidis* MC58 (NP_274323)
3	401	107	7e-39	71%	Membrane-bound lytic murein transglycosylase B (COG2951)	N. *meningitidis* 053442 (YP_001599322)
6	526				Unknown	*Lactobacillus salivarius* UCC118
7	590	181	6e-12	36%	Ferripyoverdine receptor (COG4773)	*Campylobacter curvus* 525.92 (YP_001408909)
7r	499	166	1e-33	44%	ligand_gated_channel (cd01347)	N. *meningitidis* 053442 (NP_284378)
8	574	190	7e-104	99%	UvrA gene (cd03270)	N. *meningitides* (ABE99560)
10	659	196	3e-110	98%	NADH dehydrogenase subunit C (PRK06074)	N. *meningitidis* Z2491 (NP_282873)
11	612	199	5e-74	67%	rRNA (guanine-N1-)-methyltransferase (pfam08241)	S. *pneumoniae* SP9-BS68 (ZP_01821828)
17	532	83	2e-26	69%	Hypothetical protein	Streptococcus phage SM1 (NP_862893)
18	405	93	1e-25	100%	Glutamine synthetase (PRK09469)	N. *meningitidis* 053442 (P25821)
23	312				Putative membrane peptidase	N. *meningitidis* FAM18 (AM421808)
24	616	124	2e-41	67%	Hypothetical protein	S. pneumoniae bacteriophage MM1 (AAZ82420)
25	552	174	1e-94	97%	Glutaminyl-tRNA (cd00807)	N. *meningitidis* 053442 (YP_001599586)
26	620	188	6e-101	99%	Aspartyl/glutamyl-tRNA amidotransferase subunit B (PRK05477)	S. *thermophilus* LMG 18311 (YP_140043)
30	437	157	4e-17	64% 36%	Putative TonB-dependent receptor exported protein(cd01347)	*Pedobacter sp*. BAL39 (ZP_01886124)
32	424	139	8e-36	54%	Putative carbohydrate kinase (cl00192)	*Alteromonadales bacterium* TW-7 (ZP_01612681)
35	499				(pilE1) gene	*Neisseria sicca pilin* (DQ007936)
37	285	70	-----	-----	No significant match	
38	472	108	8e-27	54%	Sugar-binding cell envelope protein (COG1653)	*Streptococcus gordonii* (YP_001451160)
41	391	96	-----	-----	No significant match	
42	326	97	2e-27	64%	Spermidine synthase (PRK03612)	*Ralstonia metallidurans* CH34 (YP_585602)
44	406	119	3e-40	68%	Nucleotide transport and metabolism (COG0775)	S. *sanguinis* SK36 (YP_001035576)
45	245	53	1e-21	96%	Glutamate synthetase (PRK04308)	N. *gonorrhoeae* FA 1090 (YP_002002449)

Identity (%): the percentage of similarity

4.8: Cloning and sequencing the A2 Viral Nucleic Acids

The two restriction endonucleases, four-base cutters *Sau*3A1 and *Hae*III, were the first enzymes chosen to cut the A2 virus, because they had cut the virus A1. They did cut the A2 genome; however, it sheared when cut with these enzymes. The sheared genome fragments between 1 Kb and 5 Kb were excised from the gel using autoclaved scalpel blades. The gel slice was dissolved to recover the DNA fragments using a gel extraction kit (section 2.10.3). These fragments were then ligated into a cloning vector (section 2.12). The vectors with inserts were cut with *Eco*RI to check the size of the inserted fragments. Most of the cloned inserts appeared to be small: about 500 bp in length (Figure 4.19).

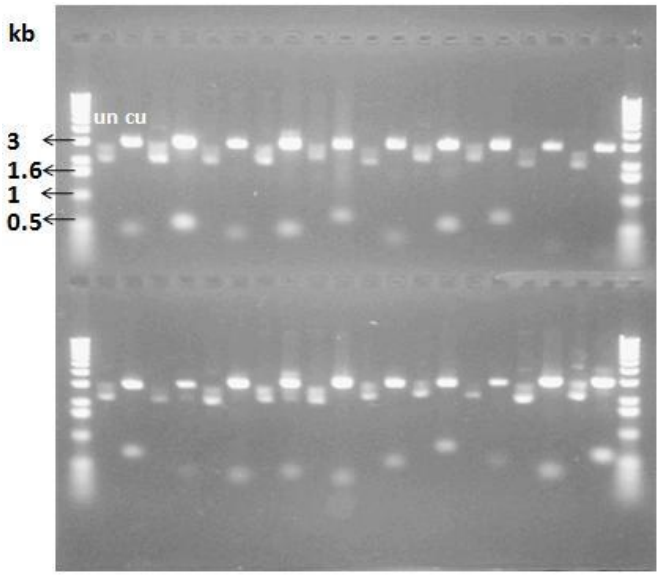


Figure 4.19: Cloned plasmid sequences excised with *Sau*3A1 and digested with *Eco*RI
The 1 kb marker was run on the left and right sides of the gel. Bands marked "un" are the plasmid with its insert uncut with enzyme, while bands marked "cu" are those where the insert was cut from its plasmid using the *Eco*RI enzyme.

Fifty clones that originated from *Sau*3A1 digests and 10 from a *Hae*III digest were sent to AGOWA in Germany to be sequenced. The sequences obtained showed that digests of the viral genome with a six-base cutter, *MFe*I, would produce larger overlap fragments and give the right size of viral genome to produce sharp bands. However, the viral genome sheared when it was digested with this enzyme. Thirty-six clones of this cut were sent to be sequenced.

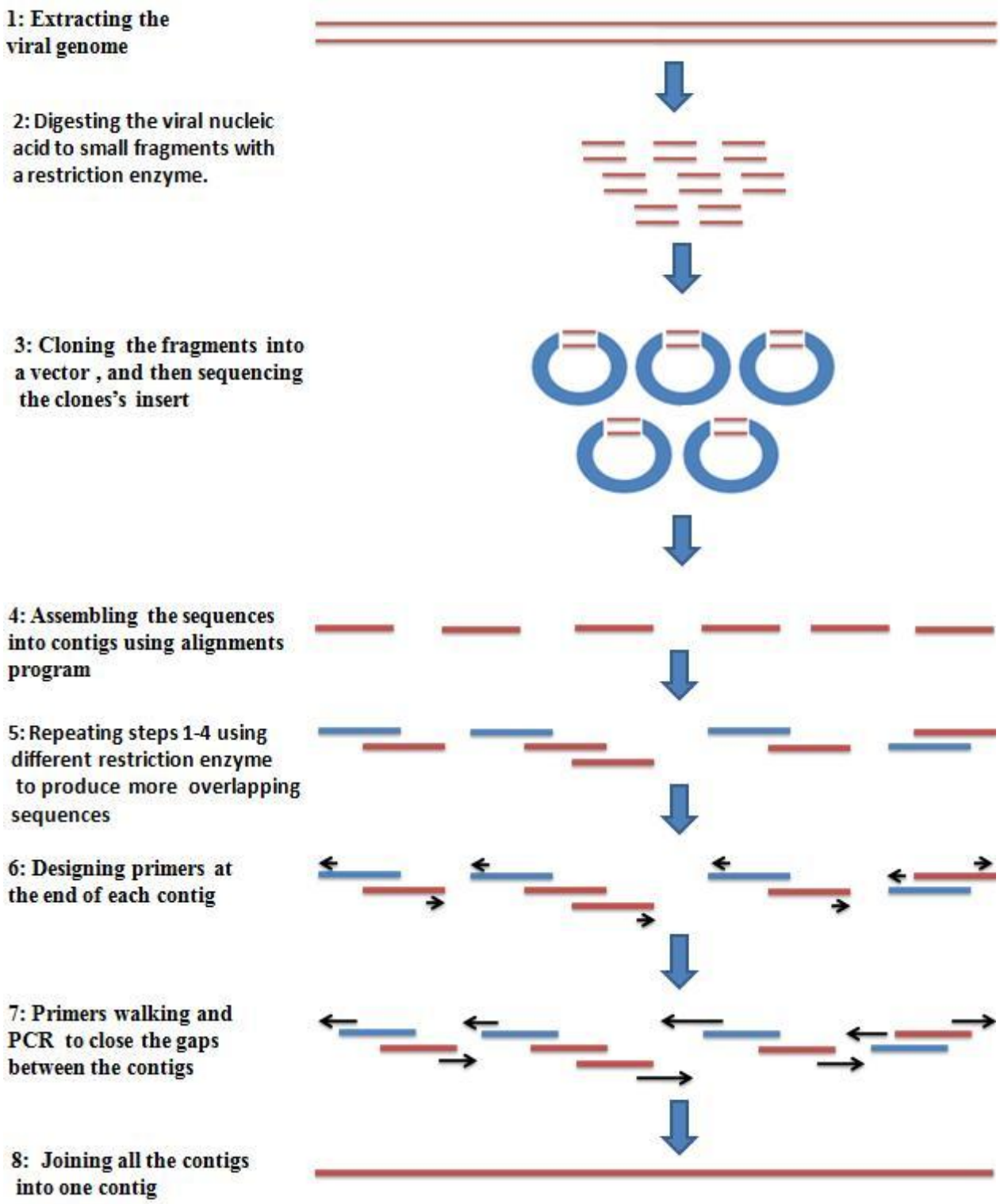


Figure 4.20: Strategies for sequencing dsDNA viral genomes

4.8.1: Assembling Sequences into Contigs

Assembly of the sequences into contigs (Figure 4.20) was done by the Lasergene SeqMan version 7.0 program (DNAStar). Ten small contigs were formed, with a few single sequences that did not overlap. Basically, the orders of assembled contigs were based on ORF and TBLASTX analysis. It was found that most of the contigs were significantly matched to a prophage in both the *Neisseria gonorrhoeae* FA 1090 and *Neisseria gonorrhoeae* NCCP11945 genomes (the N. g genome NCCP11945 sequence was just published in September 2008, and I found the prophage has an identical genome in the both strains) Twenty primers were designed: one for each end of each contig. PCR filled the gaps between the contigs, using the viral genome as a template. All except two gaps between the contigs were successfully filled; this resulted in three contigs, A, B and C, as shown in Figure 4.22 part I. Different primers at the end of each of these three contigs were designed to fill the gaps between them; unfortunately, no PCR products were obtained.

4.8.1.1: Extending the A2 Viral Contigs

Based on the sequencing results, two restriction endonucleases, six-base cutters *Dra*I and *Eag*I, were used to cut the viral genome; however, it was sheared and no bands were detected. The incubation time was therefore reduced from 30 minutes to 10 min at 37°C in an attempt to obtain larger fragments of the viral genome, which was now sheared with the *Dra*I enzyme but not completely sheared with *Eag*I enzyme; see Figure 4.17. This was because of the reduced incubation time and because the *Eag*I enzyme has 5 recognition sites against 12 for the *Dra*I. The viral genome was completely sheared when the incubation time was increased to 1 h (data not shown). The cut sample was electrophoresed on a 1% agarose gel to check the fragments size (Figure 4.17). The cut fragments were ligated into a cloning vector and the size of the inserts checked by cutting with *Eco*RI followed by agarose gel electrophoresis, as indicated. Sixty fragments were sent to be sequenced by AGOWA in Germany and PNACL at the University of Leicester. Of these, 55 sequences overlapped with the three contigs. The identities and analyses of these are summarised in Table 4.6. Contig B was extended 3 kb towards contig C, which was extended 2 kb towards contig B, as Figure 4.22 part II shows. (See appendix of chapter 4 for a list of the three contig nucleotide sequences and the primers which were used to walk between the contigs and PCR to fill the gaps).

No overlap occurred between the three contigs A, B and C. New primers at the end of these three contigs and three different polymerases Taq, Phusion High-Fidelity DNA and Herculase^{R} II Fusion DNA Polymerase were used to fill the gaps. PCRs were also set for the GC-rich content, as it was hoped to join two or more of the three contigs. In addition, sequencing directly from the entire genomic DNA using designed primers at the end of the contigs were tried, but no read was obtained. Unfortunately, all these efforts were resulted in no joining of the three contigs and some of the fourteen single sequences.

4.8.2: Sequence Analysis of A2 Virus

All three contigs had significant matches to other viral genes in the GenBank database. Table 4.6 shows the genes detected using GenMark and ORFs, followed by the structure and the function of the predicted genes. ORFs were often preceded by a sequence displaying similarity to the consensus ribosome-binding site (RBS), TAAGGAGGT (Shine and Dalgarno, 1974 and 1975). The ORF and TBLASTX analyses showed the strongest matches to a putative phage in the *N. gonorrhoeae FA 1090* gene, while other matches were to *Neisseria meningitides*. There was no significant match with the databases showing that the A2 virus belongs to the *Siphoviridae* family.

The multiple ORFs in the contigs had the same direction and order as the genes on the phage in the *N. gonorrhoeae FA 1090* genome (this will be shown below). In addition, in preparation for joining the 10 small contigs, most of the primers were designed at the end of each contig, confirming the direction and order of contigs A, B and C. A total of 32 ORF genes were identified, 17 of these being matched to phage proteins of unknown function and 13 to known genes or domains, while only 2 of the 32 had no significant match to the GenBank database (Table 4.6) (Figure 4.22). In total, of 161 clones which were sequenced, 14 single sequences did not overlap with the three contigs which had been developed. The identities and similarities of theses sequences are summarised in Table 4.7.

The estimated size of the A2 virus is 35 to 40 kbp, based on the PFGE result. However, only 24.424 kbp is accounted for by the total nucleotides obtained from the three contigs. If the 14 single sequences were to belong to the A2 virus and not to other prophages detected using electron microscopy, including the A1 virus, then only 3-9 kbp might be

missing from the total viral genome. As the 14 single sequences contain 7279 nucleotides plus the total number of nucleotides in the three contigs (24424), the total is 31703 nucleotides. Thus, the total number of all the sequenced nucleotides is close to the virus genomic size obtained by the PFGE. One conclusion which may be drawn is that one or both of the two gaps may have genes that cannot be cloned due to being lethal to *E. coli*; for example, the lysin gene (Paul *et al.,* 2002). It is known that five successive genes forming a cluster are usually found in viruses of Gram-negative bacteria: these are endolysin, holin, antiholin and Rz/Rz1 equivalents (Wang *et al.,* 2000). Another conclusion is that the gaps are still too large to be filled by PCR.

4.8.2.1: BLASTN Analysis

Significant similarities were found using BLASTN analysis of the three contigs of the A2 virus in different parts of the virus genome. In contig B (Table 4.5), four parts of the virus genome matched to *Neisseria gonorrhoeae* FA 1090 (accession number CP001050.1). In contig C, only one part of the virus genome matched to the same bacterium, and there were no significant matches to contig A.

Table 4.5: Summary of BLASTN analysis for the three contigs of the A2 virus

Contig B	NP	Length bp	E-value	Identity (%)	Gaps	Nearest similarity
1	2 – 3636	3634	0.0	85%	1%	Putative phage associated protein *Neisseria gonorrhoeae FA 1090*
2	3613-4744	1131	0.0	88%	1%	Putative phage associated protein *Neisseria gonorrhoeae FA 1090*
3	5772-10350	4578	0.0	76%	5%	Putative phage associated protein *Neisseria gonorrhoeae FA 1090*
4	11732- 14304	2572	0.0	76%	5%	Putative phage associated protein *Neisseria gonorrhoeae FA 1090*
Contig C	1111-2463	1352	0.0	82%	2%	Tail length tape measure protein *Neisseria gonorrhoeae FA 1090*
Contig A		No match				

NP: Nucleotide position on contigs of the A2 virus.

4.8.2.2: The GC-Content of the A2 Virus

The GC content can be used as an indicator of the replication direction in many prokaryotes; for example, it known that the lowest GC region of the whole genome indicates the origin of replication and the highest GC region indicates termination

(Grigoriev A 1998, Grigoriev A 1999). Contig A of the A2 virus had a lower GC content than contigs B and C. The ORFs of the A contig showed significant similarity to replication genes which may be the beginning of the A2 virus genome (Figure 4.22 part I). ORF analyses indicate that the structural order of the A2 virus genes is similar to that of a prophage in the *N. gonorrhoeae* A1090 genome. The ORF arrangement is divided into modules that are common to both bacterial and archaeal viruses. Genes relating to the establishment of cell infection are in the early region, those relating to DNA synthesis are in the middle region, and those relating to virus assembly and cell lysis are located in the late region (Brussow and Desiere 2001, Hendrix *et al.,* 1999).

The G+C content of Contig A was found to be 47.1 mol%. Contigs B and C were found to have higher values, at 50.6 and 49.7 mol% respectively. The GC content of these contigs is lower than that of four prophages (54-57%) and similar to that of one prophage (49%) found in the *N. gonorrhoeae* genome. These percentages are comparable to the average GC content (52.5%) of the N. *gonorrhoeae* genomic DNA (Piekarowicz *et al.,* 2007).

4.8.2.3: ORF Analysis of A2 Virus Genome

All predicted ORFs were searched for similarity against the GenBank databases using BLASTP analysis. ORFs 1, 6, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 26, 28 and 29 showed significant matches to hypothetical proteins, putative phage-associated proteins with unknown functions, all of which were matched to prophages in the *Neisseria gonorrhoeae* FA 1090 genome; see Table 4.6 for more details and Figure 4.22 part IV for the order of these ORFs in the virus genome. ORFs 3 and 24 showed no match or similarity against the databases. ORF 2 in contig A was matched to a phage replication protein which is found in the early genes that play an important role in the replication of viral genomes. Careful analysis showed that this gene has a 15 bp region that is directly repeated eight times (Figure 4.21 shows the repeated 15 bp sequence and Figure 4.22 part IV ORF 2). The presence of a series of repeats has been identified in many phage genes; for example, it was found that all lactococcal phages have a series of repeats in the replication gene. It has been proved that these are the origin of phage replication and essential for it (Ostergaard *et al.,* 2001).

The 15 bp repeat may be one of the several known types of genomic termini that occur in tailed phage genomes. Based on the length of the repeat, it could be either cohesive end or

a short exact direct repeat. In the case of cohesive ends, the two ends of the phage genomic DNA are single stranded identical in length and complementary to each other. After infection, the two ends anneal to each other to form a circular genomic DNA which serves as a DNA replication. The length of the cohesive ends varies between viruses, commonly, it is between 7 to 19 nucleotides long, and they can be at either 5'- or 3'-protruding strands (Hershey and Burgi, 1965; Ellis and Dean, 1985). Other viruses with linear dsDNA genomes require short exact direct repeats at the end of the genome to maintain genome integrity following DNA replication. Each round of replication would ordinarily result in loss of sequences at the 5' ends. These repeats allow concatemers formation and resolution with endonucleases to maintain genome integrity.

The function of terminal short direct repeats in linear dsDNA genome is to maintain genome integrity. Either by intramolecular base paring to form a closed circular DNA molecule which is then replicated by a rolling circle mechanism, as explained by the 'cos' sequences of lambda type phages, or to facilitate concatemerisation of multiple copies of the linear genomes as exemplified by phage T4.

Thus the number of the 8 repeats found at the end of linear genomes of A2 virus could mark the end and the beginning of the genome. The 8 repeated sets were also searched against the GenBank databases using BLASTN, BLASTP and BLASTX. Only a significant similarity was detected using BLASTP, which was over the total length of the repeated sets (39 aa) with E-value of 2e-07 and identity of 62% to a hypothetical protein of *Plasmodium falciparum* 3D7 (accession number: XP_001350291).

ORF 4 matched a putative replicative DNA helicase; most of the 12 known hexameric helicases have roles in DNA replication, recombination and transcription (Patel and Picha, 2000). ORF 5 was matched to a terminase gene, but it was not clear whether it was a large or small subunit. Terminase genes are responsible for ATP-dependent packing of concatameric DNA in phage capsids. The small subunit possesses DNA recognition specificity, while the large subunit has catalytic activity (Fujisawa and Morita, 1997). ORFs 7, 9, 13, 16, 18, 19 and 20 were matched to a variety of domains; see Table 4.5 for more details. ORF 27 was matched to a putative ATP binding protein. ORFs 30 and 31 were matched to a tail-length tape-measure protein; it has an important role in the assembly and determination of length of the phage tail (Katsura 1987; Pedersen et al.

2000). Finally, ORF 32 was matched to a putative integrase. It is an enzyme that may be divided into families, the serine and the tyrosine recombinase, based on their mode of catalysis. Integrase play variety of important roles such as mediating the recombination site between two DNA recognition sequences and the attachment site in the phage and the bacteria (Groth and Calos, 2004).

CTGCCTTTACTATCTATCAGTACTTGTTGAATATCAGTATTTACTAGTGTGGTCTCAGCC
TGATTAGGGTTAGCC-TGATTAGGGTTAGCC-TGATTAGGGTTAGCC-TGATTAGGGTTAGCC-
TGATTAGGGTTAGCC-TGATTAGGGTTAGCC-TGATTAGGGTTAGCC-TGATTAGGGTTAGCC
TGATTAGGCTCATCAAAAACAATGTAATCTGTTGTTCCGTTCCCATTCTTGCGGACTTGC

Figure 4.21: Diagram of the direct repeats of ORF 2
The replication gene identified in ORF 2, is placed in contig A, and has eight sets of direct repeats, each of 15 bp.

4.8.2.4: Detected Genes

The order of the known genes classes that were found in the three contigs A, B and C (Figure 4.22) are ORF 2: replication protein; ORF 4: putative replicative DNA helicase phage; ORF 5: terminase; ORF 7: prohead protease (peptidase); ORF 9: DNA replication, recombination and repair protein; ORF 27: putative ATP binding protein; ORF 30 and 31: tail-length tape-measure protein; ORF 32: putative integrase, and no lysis gene was detected such as endolysin or holin.

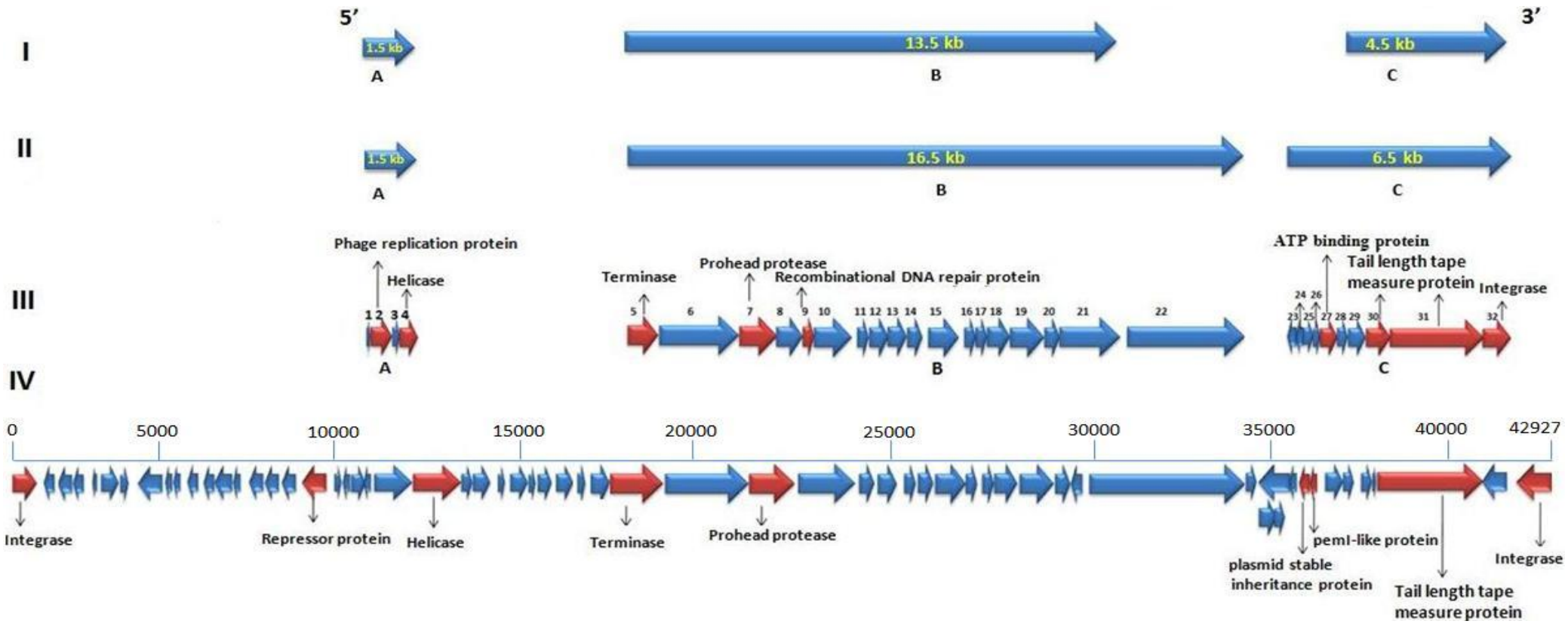


Figure 4.22: Genome annotation of the A2 virus

Representation of the direction and order of the three contigs A, B and C based on the ORF matches to a prophage in the N. *gonorrhoeae FA 1090* genome. Part I shows the three contigs obtained using the restrictions endonucleases *Hae*III, *Sau*3AI and *Mfe*I. Part II shows the extension was mad in contigs B and C using *Dra*I and *Eag*I restriction endonucleases. Part III shows the genes order in the three contigs and then these were compared to the genes order of the prophage in the N. *gonorrhoeae* FA 1090 part IV. The known structural genes are represented by red arrows and the unknown putative phage-associated protein genes by blue ones.

Table 4.6: A2 gene annotation using BLASTP and TBLASTX analysis

Contig name and size	ORF	Start-Stop Position	D	aa	E-value and identities %	Mr	pI	RBS sequences	Function or Similarity	Domain
A 1510 bp	A1	<1-114	+	37	3e-11 88%	10058	5.43	-----	Putative phage associated protein (PPAP) (Accession number (AN): YP_207633) *Neisseria gonorrhoeae FA 1090*	
	A2	126-782	+	218	1e-15 35%	40769	5.19	AGGA (-1)	Phage replication protein (AN: NP_833429) *Bacillus cereus G9842*	Domain like "DnaD"
	A3	833-1003	+	56	------	13524	5.38	-------	No match	
	A4	1006->1509	+	169	8e-25 42%	53742	5.16	AAG (-8)	Replicative DNA helicase (AN:YP_207635) *Neisseria gonorrhoeae FA 1090*	Domain like "DnaB"
B 16446 bp	B5	186-974	+	262	1e-142 93%	63868	5.11	-----	Terminase **(**AN: YP_207645**)** *Neisseria gonorrhoeae FA 1090*	Terminase Phage-related
	B6	971-3118	+	715	0.0 89%	174121	4.9	AGG (-7)	PPAP (AN: YP_207646) *Neisseria gonorrhoeae FA 1090*	
	B7	3157-4122	+	311	4e-112 63%	74794	5.09	GGAG (-6)	Prohead protease (AN: YP_207647) *Neisseria gonorrhoeae FA 1090*	Domain of Peptidase_U35
	B8	4164-4853	+	229	7e-93 91%	57016	5.12	GGAG (-7)	PPAP (AN: YP_207648) *Neisseria gonorrhoeae FA 1090*	

D = Direction of translation
aa= Number of amino acids
Mr = Molecular mass
pI = Calculated isoelectric point
Putative 5' upstream ribosome binding sequence (RBS) TAAGGAGGT.

Continuing Table 4.6

Contig name and size	ORF	Start-Stop Position	D	aa	E-value and Identities %	Mr	pI	RBS sequences	Function or similarity	Domain
	B9	5044-5424	+	227	7e-93 91%	55808	5.14	------	Recombinational DNA repair protein (AN: ZP_00135136) *Actinobacillus pleuropneumoniae*	COG3723, RecT
	B10	5421-6458	+	345	9e-119 90%	85582	5.05	GGA (-6)	PPAP (AN: YP_207648) *Neisseria gonorrhoeae* FA 1090	
	B11	6514-6843	+	109	1e-28 67%	26467	5.26	AGG (-10)	PPAP (AN: YP_207649) *Neisseria gonorrhoeae* FA 1090	
	B12	6846-7376	+	176	6e-64 67%	43523	5.18	------	PPAP (AN: YP_207650) *Neisseria gonorrhoeae* FA 1090	
	B13	7378-7857	+	159	2e-32 63%	39361	5.2	------	PPAP (AN: YP_207651) *Neisseria gonorrhoeae* FA 1090	PksG, 3-hydroxy-3-methylglutaryl CoA synthase
	B14	7857-8285	+	142	4e-70 86%	34625	5.22	AGG (-8)	PPAP (AN: YP_207652) *Neisseria gonorrhoeae FA 1090*	
	B15	8309-9082	+	257	1e-118 82%	63436	5.12	AAGGA (-6)	PPAP (AN: YP_207653) *Neisseria gonorrhoeae FA 1090*	
	B16	9142-9471	+	109	3e-46 87%	26850	5.27	AGGA (-7)	PPAP (AN: YP_207654) *Neisseria gonorrhoeae FA 1090*	PRK06975, bifunctional uroporphyrinogen-III synthetase/uroporphyrin-III C- methyltransferase

D = Direction of translation
aa= Number of amino acids
Mr = Molecular mass
pI = Calculated isoelectric point
RBS = Putative 5’ upstream ribosome binding sequence (RBS) TAAGGAGGT.

Continuing Table 4.6

Contig name and size	ORF	Start- Stop Position	D	aa	E-value and identities %	Mr	pI	RBS sequences	Function or similarity	Domain
	B17	9483-9749	+	88	1e-41 89%	21763	5.3	-----	PPAP (AN: YP_207655) *Neisseria gonorrhoeae FA 1090*	
	B18	9751-10350	+	199	1e-90 79%	49182	5.18	GGA (-9)	PPAP (AN: YP_207656) *Neisseria gonorrhoeae FA 1090*	CHP2217
	B19	10347-11213	+	288	1e-119 75%	71433	5.09	GAG (-9)	PPAP (AN: YP_207657) *Neisseria gonorrhoeae FA 1090*	BR0599 COG5449
	B20	11214-11657	+	147	3e-58 72%	36480	5.23	GGT (-7)	PPAP (AN: YP_207658) Neisseria gonorrhoeae FA 1090	PRK10838, predicted peptidase, outer membrane lipoprotein
	B21	11694-13340	+	548	0.0 79%	135111	4.95	GAGGT (-5)	PPAP (AN:YP_207660) *Neisseria gonorrhoeae FA 1090*	
	B22	13462-16440	+	1059	3e-166 74%	257357	4.82	-----	PPAP (AN:YP_207660) *Neisseria gonorrhoeae FA 1090*	
C 6468 bp	C23	6016-6240	-	74	4e-04 39%	12542	5.36	-----	PPAP (AN: YP_207860) *Neisseria gonorrhoeae FA 1090*	
	C24	5849-6001	-	50	-------	12643	5.36	-----	No match	

D = Direction of translation
aa= Number of amino acids
Mr = Molecular mass
pI = Calculated isoelectric point
RBS = Putative 5' upstream ribosome binding sequence (RBS) TAAGGAGGT.

Continuing Table 4.6

Contig name and size	ORF	Start- Stop Position	D	aa	E-value and identities %	Mr	pI	RBS sequences	Function or similarity	Domain
	C25	5466-5810	+	114	9e-28 53%	28006	5.28	AGGAGG (-4)	PPAP (AN: YP_207669) *Neisseria gonorrhoeae* FA 1090	
	C26	5269-5469	+	57	8e-05/ 55%	13741	5.40	AGGTGGTGG(-2)	PPAP (AN: YP_002002040) *Neisseria gonorrhoeae* NCCP11945	
	C27	4809-5282	+	157	4e-62 72%	38872	5.24	AAG (-8)	Putative ATP binding protein. (AN: YP_974952) *Neisseria meningitidis FAM18*	Domain of Peptidase M15_3
	C28	4449-4760	+	103	8e-23 47%	25253	5.32	GGA (-0)	PPAP (AN: YP_207664) *Neisseria gonorrhoeae* FA 1090	
	C29	3967-4452	+	161	3e-42 53%	40572	5.22	GGAA (-1)	PPAP (AN: YP_207663) *Neisseria gonorrhoeae* FA 1090	
	C30	2908-3636	+	296	3e-18 88%	24133	5.27	AAGG (-4)	Tail length tape measure protein (AN: YP_207672)	
	C31	916-3570	+	884	0.0 78%	214621	4.86	-----	Tail length tape measure protein (AN: YP_207672) *Neisseria gonorrhoeae FA 1090*	
	C32	< 2-769	+	131	1e-25	62544	5.01	-----	Putative Integrase (AN:YP_207674) *Neisseria gonorrhoeae FA 1090*	C-terminal catalytic and DNA breaking-rejoining enzymes

D = Direction of translation
aa= Number of amino acids
Mr = Molecular mass
pI = Calculated isoelectric point
RBS = Putative 5' upstream ribosome binding sequence (RBS) TAAGGAGGT.

Table 4.7: Features of the 14 single sequences that did not overlap with the three contigs of the A2 virus

Seq number	Seq length	aa	E-value Identity (%)	Nearest match to the GenBank (identified by ORF and TBLASTX)
Sa30	484	102	4e-28 68%	Phage terminase small subunit *Escherichia coli* RS218
Sa49	449	86	2e-20 64%	Putative phage associated protein, *Neisseria gonorrhoeae* FA 1090
sa11	714	66	1e-23 79%	Single-stranded binding protein, *Neisseria meningitidis* 053442
E11	729	147	4e-55 71%	Putative phage associated protein, *Neisseria gonorrhoeae* FA 1090
E17	681	142	5e-08 29%	Putative integron gene cassette protein, uncultured bacterium
Mf32	743	173	3e-45 51%	Putative replicative DNA helicase, *Neisseria gonorrhoeae* FA 1090
Mf23	368	54		No match
E37	428	141	7e-71 90%	Inner membrane insertion protein, *Serratia proteamaculans* 568
Sa52	441	122	1e-39 66%	DNA-damage-inducible protein d, *Haemophilus influenzae* PittHH
E14	220			ORFs no match, TBLASTX immunoglobulin heavy chain variable region
B16	348			Cloning vector
D6	229			No match
E03	792	62	3e-13 55%	Putative phage associated protein *Neisseria gonorrhoeae* FA 1090
Mf5	653	117	1e-21 43%	Putative phage associated protein *Neisseria gonorrhoeae* FA 1090

4.9: SDS PAGE Analysis and Proteomics

SDS-PAGE was used to separate the virus particle proteins based on their electrophoretic mobility, and trypsin was used to separate the peptides in order for them to be sequenced and characterised. The virus particles were prepared from soft-top agar plates and concentrated by CsCl gradient (section 2.8). The structural proteins were analysed by SDS–PAGE gel, as shown in Figure 4.23, and next to this figure is Table 4.8, which shows the predicted protein mass (kDa), the number of observed peptides, the matched score, the nearest similarity and the origin of the protein based on the GenBank database search. The viral capsid showed three major protein bands as well as many minor bands. Initially, a small piece was cut out of each of six bands (9, 10, 11, 12, 13 and 14) and sequenced at the University of Leicester (PNACL) (section 2.8.3). Comparisons of the sequences against the GenBank databases and the contigs of the A2 viral genome showed clear matches with the bacterial host proteins from the NCBI nr database, while only three bands (11, 12 and 14) of the six sequenced bands matched ORFs 6, 8 and 10 of the viral genome (Table 4.9). It was considered promising that three of the six sequenced protein bands matched the corresponding ORF genes, so the remainder of the ten protein bands were sent to be sequenced. Unfortunately, the annotation analysis of the ten sequenced protein bands were matched significantly to bacterial proteins rather than the A2 virus capsid proteins (Table 4.8), and no corresponding ORF genes were detected except one peptide of band 5 matched ORF 22.

The peptides of the four bands (5, 11, 12 and 14) which had corresponding ORFs are listed in Table 4.9. Bands 11 and 12 could be the major structural genes which are involved in the capsid and tail morphogenesis, based on the site order of the corresponding ORF genes, 8 and 10 (Figure 4.22). ORF 8 is set next to a prohead protease gene, which indicates that ORFs 8 and 10 are probably the major structural genes. These ORFs were matched to putative phage-associated proteins of unknown function in the *N. gonorrhoeae* A1090 genome.

Band 11 had five peptides that were matched to two corresponding ORFs (Table 4.9). Four of the five were matched to ORF 8 and one to ORF 10. It is also notable that the molecular mass of band 11 and ORF 8 were identical, at 57 kDa. This supports the contention that band 11 corresponds to ORF 8, not 10. However, band 12 had ten peptides matching three

corresponding ORFs. One of the ten peptides was matched to ORF 6, four to ORF 8 and five to ORF 10. It was found that the molecular mass of ORFs 6, 8 and 10 were 174.13, 57 and 85.59 kDa respectively, the only similarity being between ORF 8 and band 12, whose molecular weight was 52.6 kDa. It is not known whether one of the proteins bands was split in two, but the molecular weights indicate that they did not split. It was also found that four peptides in bands 11 and 12 that were matched to ORF 8 were the same peptides (see Table 4.10 for peptides and translated nucleotides). Despite other similarities, the sequence homology of the protein bands 11 and 12 were different. A search against the GenBank database showed that band 11 had a significant match to putative phosphate acetyltransferase of *Neisseria meningitidis MC58.* In contrast, a sequence of band 12 had a significant match to a putative phage-associated protein of *Neisseria gonorrhoeae* FA 1090.

In the case of bands 5 and 14, only one peptide from each matched relating ORFs of 22 and 10 respectively. The bands 5 and 14 held significant matches against the databases to putative isocitrate dehydrogenase and citrate synthase of *Neisseria meningitidis* FAM18. Therefore, these bands could play an important role in the capsid structure of the A2 virus.

Significant matches to the host proteins would not be surprising, several reasons could explain such matches, however, they might not be considered due to contamination impending from the host proteins. Firstly, it is likely that the A2 virus may have a lipid associated with its capsid. Secondly, as previously indicated the OIB strain holds the spontaneous A1 virus, which may be present in the sample, (it probably enters the cell by the membrane fusion and exists by the budding events) (Figure 4.6). All enveloped viruses are acquainted with these processes, thus, some of the host proteins would be packaged into the virus capsid.

In a previous study based on the GenBank databases search, analysis of 36 enveloped capsid proteins of influenza virus showed >95% significant matches to the host proteins (Shaw *et al.*, 2008). In the case of band 15 (Figure 4.23), it seems to be the major band protein, however, in accordance to the databases a match was found to an outer membrane protein of *Neisseria meningitidis* MC58; in addition, no corresponding ORFs were detected. As

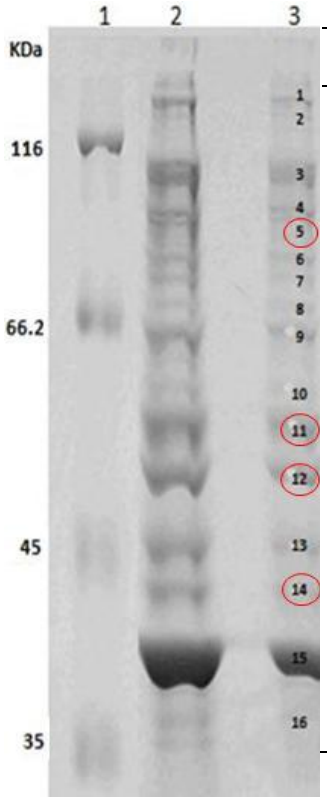
indicated above this protein may have significant roles in the A1 and A2 virus capsid, and similar protein may also be present in the cell membrane of the host.

Features of other bands, which significantly match bacterial proteins rather than viral proteins (or even had no corresponding ORFs), are listed in Table 4.9. These could be virus proteins; the following reasons support this hypothesis:

1- Despite bands 5, 11 and 4 significantly matching host proteins in the GenBank databases, they still have corresponding ORFs genes. Thus, this could be extended to other bands that had significant matches to host proteins but not to the ORFs.
2- The virus genome is not a completed sequence, thus some corresponding ORFs would be missing.
3- As indicated, similarities of the viral and the host proteins could be possible.
4- Contamination from the host proteins may be possible.
5- The GenBank databases may have much poorer coverage of viral than bacterial proteins, hence, matches were significantly more likely to be made to bacteria rather than virus.

Indeed, completing the analysis of these proteins would be greatly improved if the viral genome was a completed sequenced. Completing the genome will be challenged, and it will be one of the first future works.

Table 4.8: SDS-PAGE analysis of A2 virion proteins



kDa	No. of different observed peptides	Score	Function	Nearest match to the GenBank
128.2	41	1397	DNA-directed RNA polymerase beta chain	N. *meningitidis FAM18*
124	9	398	Ribonuclease E	N. *meningitidis Z2491*
107.6	25	968	Pyruvate dehydrogenase E1 component	N. *meningitidis FAM18*
98.3	37	1527	Elongation factor EF-2	N. *meningitidis Z2491*
94.8	26	983	Putative isocitrate dehydrogenase	N. *meningitidis FAM18*
85.1	15	510	Putative chaperone protein	N. *meningitidis FAM18*
76.8	14	565	Aspartyl-tRNA synthetase	N. *meningitidis Z2491*
71	15	654	30S ribosomal protein S1	N. *meningitidis MC58*
65.3	19		Chaperonin, 60 kDa	N. *meningitides MC58*
60.4	6	339	Inositol-5-monophosphate dehydrogenase	N. *meningitidis Z2491*
57	7	326	Putative phosphate acetyltransferase	N. *meningitidis MC58*
52.6	3	143	Putative phage associated protein	N. *gonorrhoeae FA1090*
45.5	9	672	Translation elongation factor Tu	N. *meningitidis MC58*
43	10	575	Citrate synthase	N. *meningitidis FAM18*
39.4	11	263	Outer membrane protein	N. *meningitidis* MC58
35.5	8	471	Alcohol dehydrogenase	N. *meningitidis Z2491*

Figure 4.23: SDS-PAGE analysis of A2 virion proteins

Band 1 is the protein molecular weight marker unstained standard. Molecular mass in kDa of the marker is shown on the left. Band 2 contains less dilution of the structural protein samples than band 3. The circled bands are directly related to the translated genomic sequence. Gel stained with Brilliant Blue G-Colloidal and destained as indicated in the materials and methods.

Table 4.9: Predicted molecular masses of A2 protein bands and corresponding ORFs

The score and the number of peptides were matched to the contig B of the A2 viral genome.

Band Number	Predicted molecular Mass (KDa)	Score	Matched peptides To sequence	Corresponding ORFs
1	128.2			
2	124.03			
3	107.64			
4	98.32			
5	94.79	41	1	22
6	85.15			
7	76.8			
8	71.01			
9	65.32			
10	60.39			
11	57.07	271	5	8 and 10
12	52.59	684	10	6, 8 and 10
13	45.48			
14	43.05	42	1	10
15	39.41			
16	35.52			

Table 4.10: Peptide sequences of A2 virus protein bands generated by SDS-PAGE

Bands	Peptide	Translated nucleotide sequence 5' to 3'
5	K.ALRNIIVQAR.Y	GCACTGCGGAACATCATCGTTCAAGCACGA
11	K.SAQANGEPLNK.G	TCAGCACAAGCGAACGGTGAGCCGTTGAATAAA
	K.AITNINVGNQR.A	GCCATCACAAATATCAACGTAGGTAATCAACGC
	K.SGVTPTPTAVVSAGAGK.I	TCAGGCGTTACACCTACTCCGACCGCTGTTGTTTCAGCTGGCGCGGGTAAA
	R.LPAYVQGVGNLLQVR.T	CTGCCTGCCTACGTTCAAGGCGTGGGCAACCTGCTGCAAGTGCGT
	K.GFTQPTSFTTGLQTYDLSAPSQK.L	GGTTTTACTCAGCCGACCAGCTTTACTACTGGTTTGCAAACCTATGACCTGTCCGCGCCGTCCCAAAAA
12	K.TPLSQGFISR.V	ACACCATTATCACAAGGATTTATTTCCCGT
	K.LIIGNGGAPLIK.L	CTGATTATCGGCAACGGCGGCGCGCCGCTGATTAAG
	K.AITNINVGNQR.A	GCCATCACAAATATCAACGTAGGTAATCAACGC
	R.GGVINHEMVER.N	GGTGGTGTTATCAATCACGAAATGGTTGAACGT
	R.LSPDTIYVNAR.D	TTGTCCCCTGACACCATCTACGTCAACGCGCGC
	K.LNVDVNNTANIK.A	CTGAATGTGGACGTGAACAACACCGCAAACATTAAA
	K.SGVTPTPTAVVSAGAGK.I	TCAGGCGTTACACCTACTCCGACCGCTGTTGTTTCAGCTGGCGCGGGTAAA
	R.QEYYQIEWPLR.T	CAAGAGTATTACCAAATCGAATGGCCGCTGCGT
	R.LPAYVQGVGNLLQVR.T	CTGCCTGCCTACGTTCAAGGCGTGGGCAACCTGCTGCAAGTGCGT
	K.GFTQPTSFTTGLQTYDLSAPSQK.L	GGTTTTACTCAGCCGACCAGCTTTACTACTGGTTTGCAAACCTATGACCTGTCCGCGCCGTCCCAAAAA
14	R.SLCELLL.	AGCCTTTGCGAACTTCTACTC

Four peptides in bands 11 and 12 were found to be similar and are presented in the same colours.

4.10: Summary of Result and Discussion

Two viruses that infect one host were detected. The 16S rRNA gene of the host matched an unknown *Neisseria* strain. The polygenetic tree 16s rRNA analysis showed that the nearest match to the OIB strain was the uncultured *Neisseria* sp. clone EMP_C13. This OIB could be a serotype of one of the *Neisseria* sp. More tests, such as genotypic and phenotypic characterisation, are required to identify its taxonomy.

4.10.1: A1 Virus

The A1 virus was found to have plaques of normal and strange morphology which appeared spontaneously on the soft-top agar plates. These varied in size from 0.5 to 0.7 mm (Figure 4.4 I and II). It is not known whether the appearance of the plaques was caused by the A1 virus or something related to the host. TEM analysis showed that the virus has an isometric capsid varying in size from 32 nm to 58 nm. Based on the ICTV taxonomy, structural analysis of the A1 virus suggests that it might be classified into the family *Tectiviridae*. This virus may contain an internal lipid, as it was very sensitive to an extraction method including a chloroform step. Similarities to Bam 35 and some other plasmid-like phages were found, such as the genome size and type, and the apparent lack of a tail. The A1 virus was also found to have similarities to other eukaryotic viruses.

It was found that the A1 viral capsid was very sensitive to methods using chloroform, which tended to cause the shearing of the viral nucleic acid. This was solved by using a commercial kit (Viraprep Lambda kit) and the genomic virus was set between 12 kb and 23 kb in comparison with the size marker (Figure 4.14). It appeared that the virus had a linear dsDNA after digestion with restriction endonucleases. Unfortunately, the PFGE did not provide a valid result, because the chloroform was applied to remove the PEG precipitate, so the exact size of the A1 virus is not known.

All of the 24 sequenced clones of the A1 virus were matched to bacteria genomes in the databases, except 4 sequences (see section 4.6.1). The extent of similarity and matches to bacteria genome will remain unknown until the complete sequence of the A1 virus is obtained.

4.10.2: A2 Virus

Plaque Morphology

A2 is a lytic virus forming clear and stable plaques on soft-top agar. The size of the plaques varied from 1 mm to 1.5 mm, and they were visible within three hours of incubation at 37 °C. As indicated, the host strain exhibited no spontaneous plaques when taken from the -80°C storage, plated and grown in broth and used as a lawn. This strategy was used to avoid contamination from spontaneous plaques that may have been caused by the A1 virus. Preventing the occurrence of spontaneous plaques on the IBO lawn facilitated the conducting of further experiments on the A2 virus. For example, a single-step growth curve was drawn successfully to predict the latent period of 25 min and the burst size of 24±2 virus particles per cell. A host range experiment was also conducted for the A2 virus; however, none of the bacterial strains used formed plaques on the soft-top agar.

TEM analysis

The TEM analysis showed A2 to be a tailed virus with no detectable base plates or tail fibres, belonging to the order *Caudovirales* and classified under the family *Siphoviridae* (Figure 4.8). A few virus particles with different morphologies were detected in the same sample taken from one plaque. All the viruses, based on their morphology, were classified under the family *Siphoviridae* (Figure 4.9). These could be prophages integrated into the host genome and were induced by the cell lysis caused by the A2 virus. Note that all these were from one plaque only. This indicates that the host may be a factory of prophages, and it is worth considering how these are assembled into the host genome.

Genomic Characterisation

The titre of the A2 virus was easily increased for genomic extraction using the plaque assay method; virus particles were precipitated using PEG 6000, then extracted using an equal volume of phenol: chloroform. As recommended, immediately after the extraction step the virus nucleic acid received heat treatment to avoid degradation of the genome. It was confirmed that the A2 virus had a linear dsDNA after digestion with restriction endonucleases. The A2 viral genome appeared to sit immediately above the 12 kb marker (Figures 4.16). However, analysis obtained from PFGE showed that the size of the viral genome was about 35 to 40 kbp (Figure 4.18), indicating that the remaining two gaps were wider than expected. If all 14 single sequence nucleotides (7279 bp) belonged to the A2 viral

genome and were added to the total of three contig nucleotides (24424 bp), this would give a total of 31703 nucleotides. If this were the case, then the gaps would not be large: their estimated length could be between 3 and 9 kbp.

The A2 virus genome was degraded when digested with restriction enzymes, although the virus genome had five recognition site cutters using the *Eag*I restriction enzyme. The reason for the genomic degradation is unknown; it should be noted that the experiment was conducted in sterile conditions and that the extracted virus genome was subjected to heat treatment to terminate any nuclease activity present in the sample.

Sequencing, Cloning and Assembling the Virus Genome

Initially, three restriction enzymes two four-base cutters and one six-base were used to cut the viral genome to be cloned and sequenced. The sequences obtained were assembled into 10 contigs using the Lasergene SeqMan version 7.0 program (DNAStar), then primers were designed at the end of each contig to fill the gaps, using PCR and primer walking. All the gaps were filled except two, resulting in the formation of three contigs: A, B and C (Figure 4.22 part I). Despite a great deal of effort, the remaining two gaps could not be filled.

The A2 viral genome was extracted again and cut with two different six-base cutter restriction enzymes, then fragments were cloned and sequenced. All the received sequences were assembled with the three contigs. Although an extension was observed, the gaps were still not filled (Figure 4.22 part II). A new set of primers were designed at the end of the extended contigs, and PCR was set to fill these gaps, but no PCR product was obtained. Three different polymerases Taq, Phusion High-Fidelity DNA and Herculase^{R} II Fusion DNA Polymerase were also used to try to fill the gaps. In addition, PCRs were set for the GC- rich content, as it was hoped that this would help to join two or more of the three contigs. Unfortunately, none of the gaps was filled, despite all these effort.

Sequence Analysis

Thirty-two ORF genes were identified using ORF and GeneMark analysis. The GenBank databases were searched for the identities of these ORFs using BLASTP and TBLASTX. Twenty one of the 32 ORFs were matched to putative phage-associated proteins of unknown function and 9 to known genes, while only 2 of the 32 had no significant match to the

GenBank Database (Table 4.6). The known genes detected were phage replication protein, putative replicative DNA helicase, terminase, putative ATP binding protein, tail-length tape-measure protein and putative integrase genes.

SDS PAGE analysis and proteomics of A2 virus

Sixteen protein bands were sequenced and compared against the GenBank database, the 32 predicted ORF genes of the A2 viral genome and the 14 single sequences. All 16 bands had significant similarities within the GenBank databases (Table 4.8). However, 4 of the 16 were matched to 4 corresponding ORFs (Table 4.9). The two proteins bands 11 and 12 were considered major bands because of their high density compared to other minor bands, and many peptides were matched to corresponding ORFs. However, the rest of the other proteins bands could play important roles in the structure of the A2 virus capsid, and greater analysis of these could be observed if the virus genome had totaly sequenced.

Chapter 5: General Conclusion and Discussion

Chapter 5

5. General Conclusion and Discussion

5.1: Over Aims and Objectives

This thesis consists of two sections; in the first section, a metagenomic analysis was adopted, in an attempt to measure the viral diversity in the dental plaque of the human mouth. The virus genomes were directly extracted and amplified using the sensitive nucleic acid amplification method, the MDA, to overcome the difficulties of cloning, sequencing and sample limitation. This was designed to characterise the virus population and to detect novel virus genes present within the human mouth.

The second section focused on isolating lytic viruses from the human mouth using bacterial lawns. The plaque assay method was used to detect plaques in the formed lawn, which is a clear area, due to the lysis from the host. Subsequently, detected lytic viruses were examined by transmission electron microscopy, in order to determine virus morphological structure. Attracted lytic virus was used to perform a typical single-step growth curve, in order to obtain more information about virus and host interactions. Sequencing, annotation and characterisation of the detected virus genes were also planned. Finally, Mass spectral technique was used to sequences virion associated proteins that were analysed by SDS–PAGE gel electrophoresis.

5.2: Overall Findings and Methodologies Used

5.2.1: Metagenomic Analysis of Virus Population in the Human Mouth

The MDA method was used to amplify the viral population present in a limited size (the dental plaque sample of the human mouth). The directly extracted viral genes (from the human mouth) were increased from, ng to μg genomic DNA using the indicated method; later, amplified samples were sheared, cloned and sequenced. However, the static analysis of the viral diversity obtained from the first volunteer appeared to be very low, based on the Shannon index calculation. The value of species diversity and evenness is 1.9, which is lower than the previous reported virus libraries. According to Chao 1, it appeared that the library

only contains the equivalent of five classes of viruses (if we assume all the clones sequenced are virus related); however, this obvious discrepancy is unclear. The only reason identified is the source of DNA used in this research was limiting. It could be increased considerably by using pooled samples obtained from 30 volunteers or by obtaining samples directly via a dentist. This could considerably influence the results, since an increase in sample size could result in different viruses being detected.

The unknown sequences compared with the GeneBank databases using TBLASTX were found to constitute more than half (55%). This percentage is still high, despite the GenBank' non-redundant database being increased by two-fold (Delwart 2007; Edwards and Rohwer 2005). The origin of these sequences are still unknown, they could be viruses if they were not artefact sequences that had been generated by the MDA method. As indicated in chapter 3, (section 3.1) in previous studies improvements were made to the MDA method, in an attempt to reduce artefact sequences.

Most of the annotated sequences in this library were found to be phages upon relating them to other published metagenomic viral libraries. The *Siphoviridae* and prophage members were strongly represented, forming 57% of the total viral matches in the dental plaque library. These were also identified strongly as 80% in the faecal (Breitbart *et al.*, 2003) and marine sediment (Breitbart *et al.*, 2004) libraries. However, they were found to constitute fewer than 50% of the total viral matches in the seawater libraries. On the other hand, the members of the *Myoviridae* and *Podoviridae* families were found to be as high as 83% of the total viral matches of the Chesapeake Bay Virioplankton library (Bench *et al.*, 2007). In addition to the low population found, many clones were significantly matched to a small region of putative phage in *Corynebacterium diphtheria* genome. Primer walking and PCR sets were used to fill the gaps between these which form a continuous contig of 11554 bp. It was concluded that the phage was actively lytic, with a higher titre in comparison to other virus identified in the sample.

5.2.2. Isolating Lytic Phages from the Human Mouth

Isolating lytic phages from the human mouth using the culture-based method is possible, but it requires much patience and careful examination of the plaques formed by lysis of the bacterial host. Knowledge of the bacterial effect on the soft top agar is also required, as many bacteria can form plaques due to their antibacterial activity (Hitch et al., 2004), their sensitivity or the leaking of the required elements in the growth medium. At the beginning of this study, these factors created significant doubt as to the origin of some of the observed plaques. It would be of interest to find a method that might help to distinguish between plaques caused by viruses and bacteria, as many of the plaques detected did not appear at the second or third times of propagation.

The two isolated viruses, A1 and A2, and their host (OIB) are unknown, and they have not been characterised. The OIB needs to be characterised by genotypic and phenotypic tests in order to identify its taxonomy. Few clones of the A1 virus were sequenced because of difficulties in increasing its titre and to avoid the degradation of its genomic DNA. It was found that the commercial kits provided the best means to extract the A1 viral genome, as chloroform is not used in any of these. On the other hand, more clones were sequenced for the A2 virus, which fell into three contigs and 14 single sequences. Much effort was made to complete the genome sequence of the A2 virus; however, while an extension was obtained, the virus genome remained incomplete. Therefore, in the future, other strategies will be required to fill the gaps in the A2 virus genome, such as using the MDA method, or amplifying the remaining gaps directly from the host by PCR. Alternatively, more sets of restriction endonucleases could be tried, in the hope of revealing the missing part of the virus genome. However, if I had a budget of about 5000 pounds, I would have used the new high throughput pyrosequencing technique, 454 sequence method. This method could generate large amounts of sequence information in a single run, which does not rely on cloning for the generated sequences (Hoper, et al., 2008).

The A2 virus particle proteins were analysed using SDS-PAGE gel electrophoresis. Four major protein band sequences were directly related to the translated genomic sequence, while the remaining bands were matched to bacterial proteins rather than viral proteins.

Contamination from the bacterial host proteins is unlikely to have happened, and a possible explanation of this result is described above.

The A2 virus is confirmed to be a lytic virus, as the latent period of 25 min and the burst size of 25±2 particles were determined by a single-step growth curve. It is also confirmed that the A2 virus is the first lytic virus that infects one of the *Neisseria* sp. This proves that detecting lytic phages from the human mouth for other pathogenic bacteria is possible, and they could be used as a second treatment option for bacteria which proved to be resistant to a variety of antibiotics. The analysis of the two viruses has just begun, and the interactions between these and the OIB strain will be clearer if the sequences of the two viruses are completed. Thus, completing these sequences is a worthy goal, which could expand knowledge of gene function and structure. It is also a worthy goal to know more about the effect of the OIB strain on human health.

References

Aas, J.A., Paster, B.J., Stokes, L.N., Olsen, I. and Dewhirst, F.E. (2005). "Defining the normal bacterial flora of the oral cavity." J Clin Microbio **43**:5721-32.

Abulencia, C.B., Wyborski, D.L., Garcia, J.A., Podar, M., Chen, W., Chang, S.H., Chang, H.W., Watson, D., Brodie, E.L., Hazen, T.C. and Keller, M. (2006). "Environmental whole-genome amplification to access microbial populations in contaminated sediments." Appl Environ Microbiol **72:**3291-301.

Achtman, M. and Wagner, M. (2008). "Microbial diversity and the genetic nature of microbial species." Nat Rev Microbiol **6**:431-40.

Ackermann, H.W. (2007). "5500 Phages examined in the electron microscope." Arch Virol ***152****:227-43.*

Ackermann, H.W. (2003). "Bacteriophage observations and evolution." Res Microbiol **154**:245-51.

Ackermann, H.W. (2001). "Frequency of morphological phage descriptions in the year 2000. Brief review." Arch Virol **146**:843-57.

Ackermann, H.W. (1998). "Tailed bacteriophages: the order caudovirales." Adv Virus **51**:135-201.

Amann, R.I., Ludwig W. and Schleifer, K.H. (1995). "Phylogenetic identification and in situ detection of individual microbial cells without cultivation." Microbiol **59**:143-69.

Ambrose, H.E. and Clewley, J.P. (2006). "Virus discovery by sequence-independent genome amplification." Rev Med Virol **16**:365-83.

Aviel-Ronen, S., Zhu, C., Coe, B. P., Liu, N., Watson, S.K., Lam, W., Tsao, M.S. (2006). "Large fragment Bst DNA polymerase for whole genome amplification of DNA from formalin-fixed paraffin-embedded tissues." BMC Genomics 7:312.

Bachrach, G., Leizerovici-Zigmond, M., Zlotkin, A., Naor, R. and Steinberg, D. (2003). "Bacteriophage isolation from human saliva." Lett Appl Microbiol **36**:50-3.

Bench, S.R., Thomas E., Hanson, T.E., Williamson,Kurt E., Ghosh, D., Radosovich, M., Wang, K. and Wommack, K.E. (2007). "Metagenomic characterization of Chesapeake Bay virioplankton." Appl Environ Microbiol **73**:7629-41.

Bergh, O., Borsheim, K.Y., Bratbak, G. and Heldal, M. (1989). "High abundance of viruses found in aquatic environments." Nature **340**:467-8.

Blanco, L., Bernad, A., Lazaro, J.M., Martin, G., Garmendia, C. and Salas, M. (1989). "Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication." J Biol Chem 264:8935-40.

Borsheim, K.Y. (1993). "Native marine bacteriophages." FEMS Microbiol. Ecol., vol. 102, pp. 141-159.

Bradley, D.E. (1967). "Ultrastructure of bacteriophage and bacteriocins." Bacteriol Rev **31**:230-314.

Breitbart, M., Felts, B., Kelley, S., Mahaffy, J.M., Nulton, J., Salamon, P. and Rohwer, F. (2004). "Diversity and population structure of a near-shore marine-sediment viral community." Proc Biol Sci **271**:565-74.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J.M., Nulton, J., Salamon, P. and Rohwer, F. (2003). "Metagenomic analyses of an uncultured viral community from human feces." J Bacteriol **185**:6220-3.

Breitbart, M. and Rohwer, F. (2005). "Here a virus, there a virus, everywhere the same virus?" Trends Microbiol **13**:278-84.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall,Anca M., Mead, D., Azam, F., and Rohwer, F. (2002). "Genomic analysis of uncultured marine viral communities." Proc Natl Acad Sci USA **99**:14250-5.

Breyen, S.A. and Dworkin, M. (1984). "Autoplaquing in Myxococcus strains." J Bacteriol **158**:1163-4.

Brussow, H. and Desiere, F. (2001). "Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages." Mol Microbiol **39**:213-22.

Campbell, L.A., Short, H.B., Young, F.E. and Clark, V.L. (1985). "Autoplaquing in Neisseria gonorrhoeae." J Bacteriol **164**:461-5.

Cann, A.J., Fandrich, S.E. and Heaphy, S. (2005). "Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes." Virus Genes **30**:151-6.

Cannon, R.D. and Chaffin, W.L. (2001). "Colonization is a crucial factor in oral candidiasis." J Dent Educ **65**:785-7.

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

Chanishvili, N., Chanishvili, T., Tediashvili, M. and Barrow, P.A. (2001), "Phages and their application against drug-resistant bacteria." Journal of chemical technology and biotechnology 0268-2575

Chao, A. (1984). "Non-parametric estimation of the number of classes in a population." Scandin J Stat **11**: 265-70.

Choi, B.K., Paster, B.J., Dewhirst, F.E., and Gobel, U.B. (1994). "Diversity of cultivable and uncultivable oral spirochetes from a patient with severe destructive periodontitis." Infect Immun **62**:1889-95.

Crawford, J.T. and Goldberg, E.B. (1980). "The function of tail fibers in triggering baseplate expansion of bacteriophage T4." J Mol Biol **139**:679-90.

Dabrowska, K., Switala-Jelen, K., Opolski, A., Weber-Dabrowska, B. and Gorski, A. (2005). "Bacteriophage penetration in vertebrates." J Appl Microbiol **98**:7-13.

Daugelavicius, R., Gaidelyte, A., Cvirkaite-Krupovic, V. and Bamford, D.H. (2007). "On-line monitoring of changes in host cell physiology during the one- step growth cycle of Bacillus phage Bam35." J Microbiol Methods **69**:174-9.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., Driscoll, M., Song, W., Kingsmore,Stephen F., Egholm, M. and Lasken, R. S. (2002). "Comprehensive human genome amplification using multiple displacement amplification." Proc Natl Acad Sci USA **99**:5261-6.

Dean, F.B., Nelson, J.R., Giesler, T.L. and Lasken, R.S. (2001). "Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification." Genome Res **11**:1095-9.

Delwart, E.L. (2007). "Viral metagenomics." Rev Med Virol **17**:115-31.

Dorigo, U., Jacquet, S. And Humbert, J.F. (2004). "Cyanophage diversity, inferred from g20 gene analyses, in the largest natural lake in France, Lake Bourget." Appl Environ Microbiol **70**:1017-22.

Downes, J., Hooper, S.J., Wilson, M.J., and Wade, W.G. (2008). "Prevotella histicola sp. nov., isolated from the human oral cavity." Int J Syst Evol Microbiol **58**:1788-91.

Duckworth, D.H. (1976). "Who discovered bacteriophage?" Bacteriol Rev **40**:793-802.

Edwards, R.A. and Rohwer, F. (2005). "Viral metagenomics." Nat Rev Microbiol **3**:504-10.

Ellis, D. M. and Dean, D. H. (1985). "Nucleotide sequence of the cohesive single-stranded ends of Bacillus subtilis temperate bacteriophage f105." J. Virol. **55**:513-515.

Ellis, E.L. and Delbrück M. (1939). "The Growth of Bacteriophage." J. Gen. Physiol. **22**:365-384.

Felsenstein, J. (1985). " Confidence limits on phylogenies: An approach using the bootstrap." Evolution **39**:783-791.

Fortuna, W., Miedzybrodzki, R., Weber-Dabrowska, B. and Gorski, A. (2008) "Bacteriophage therapy in children: facts and prospects." Med Sci Monit **14**:RA126-32.

Fujisawa, H. and Morita, M. (1997). "Phage DNA packaging." Genes Cells **2**:537-45.

Fullwood, M.J., Tan,Jack J.S., Ng,Patrick W.P., Chiu, K.P., Liu, J., Wei, C.L. and Ruan, Y. (2008). "The use of multiple displacement amplification to amplify complex DNA libraries." Nucleic Acids Res **36**:32.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L. and Field, K.G. (1990). "Genetic diversity in Sargasso Sea bacterioplankton." Nature **345**:60-3.

Grigoriev, A. (1999). "Strand-specific compositional asymmetries in double-stranded DNA viruses." Virus Res **60**:1-19.

Grigoriev, A. (1998). "Analyzing genomes with cumulative skew diagrams." Nucleic Acids Res **26**:2286-90.

Groth, A.C. and Calos, M.P. (2004). "Phage integrases: biology and applications." J Mol Biol **335**:667-78.

Hadas, H., Einav, M., Fishov, I. and Zaritsky, A. (1997). "Bacteriophage T4 development depends on the physiology of its host Escherichia coli." Microbiology **143**:179-85.

Halling, C., Sunshine, M.G., Lane, K.B., Six, E.W. and Calendar, R. (1990). "A mutation of the transactivation gene of satellite bacteriophage P4 that suppresses the rpoA109 mutation of Escherichia coli." J Bacteriol **172**:3541-8.

Hancock, H., Sigurdsson, A., Trope, M. and Moiseiwitsch, J. (2001). "Bacteria isolated after unsuccessful endodontic treatment in a North American population." Oral Surg Oral Med Oral Pathol Oral Radiol Endod **91**:579-86.

Haring, M., Rachel, R., Peng, X., Garrett,Roger A. and Prangishvili, D. (2005). "Viral diversity in hot springs of Pozzuoli, Italy, and characterization of a unique archaeal virus, Acidianus bottle-shaped virus, from a new family, the Ampullaviridae." J Virol **79**:9904-11.

Hegde, S.K., Kumar, K.B., Sudha, P. and Bhat, S.S. (2005). "Estimation of salivary bacteria capable of inhibiting and stimulating Streptococcus mutans and its correlation to dental caries and untreated carious teeth." J Indian Soc Pedod Prev Dent **23**:126-30.

Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E. and Hatfull, G.F. (1999). "Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage." Proc Natl Acad Sci USA **96**:2192-7.

Hershey, A. D. and Burgi, E. (1965). "Complementary structure of interacting sites at the ends of lambda DNA molecules." Proc. Natl. Acad. Sci. USA **53**:325-330.

Hitch, G., Pratten, J. and Taylor, P.W. (2004). "Isolation of bacteriophages from the oral cavity." Lett Appl Microbiol **39**:215-9.

Hoper, D., Hoffmann, B. and Beer, M. (2008). "Simple, sensitive, and swift sequencing of complete Avian Influenza H5N1 genomes." J Clin Microbiol, 1098-660.

Hosono, S., Faruqi, A.F., Dean, F.B., Du, Y., Sun, Z., Wu, X., Du, J., Kingsmore, S.F., Egholm, M. and Lasken RS (2003). "Unbiased whole-genome amplification directly from clinical samples." Genome Res **13**:954-64.

Hugenholtz, P., Goebel, B.M. and Pace, N.R. (1998). "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity." J Bacteriol **180**:4765-74.

Hutchison, C.A., Smith, H.O., Pfannkoch, C. and Venter, J.C. (2005). "Cell-free cloning using phi29 DNA polymerase." Proc Natl Acad Sci USA **102**:17332-6.

Jado, I., Lopez, R., Garcia, E., Fenoll, A., Casal, J. and Garcia, P. (2003). "Phage lytic enzymes as therapy for antibiotic-resistant Streptococcus pneumoniae infection in a murine sepsis model." J Antimicrob Chemother **52**:967-73.

Jikia, D., Chkhaidze, N., Imedashvili, E., Mgaloblishvili, I., Tsitlanadze, G., Katsarava,R., Glenn, M.J. and Alexander, S. (2005). "The use of a novel biodegradable preparation capable of the sustained release of bacteriophages and ciprofloxacin, in the complex treatment of multidrug-resistant Staphylococcus aureus-infected local radiation injuries caused by exposure to Sr90." Blackwell Publishing Ltd, Clinical and Experimental Dermatology **30**: 23–26

Jukes, T.H. and Cantor, C.R. (1969). "Evolution of protein molecules." In Munro HN, editor, Mammalian Protein Metabolism, pp. 21-132, Academic Press, New York.

Kanamaru, S., Leiman,Petr G., Kostyuchenko, V.A., Chipman,Paul R., Mesyanzhinov, V.V., Arisaka, F. and Rossmann, M.G. (2002). "Structure of the cell-puncturing device of bacteriophage T4." Nature **415**:553-7.

Katsura I 1987, "Determination of bacteriophage lambda tail length by a protein ruler." Nature 327(6117):73-5.

Keller, M. and Zengler, K. (2004). "Tapping into microbial diversity." Nat Rev Microbiol, **2**:141-50.

Kroes, I., Lepp, P.W. and Relman, D.A. (1999). "Bacterial diversity within the human subgingival crevice." Proc Natl Acad Sci USA **96**:14547-52.

Kumar, P.S., Griffen, A. L., Moeschberger, M.L. and Leys, E.J. (2005). "Identification of candidate periodontal pathogens and beneficial species by quantitative 16S clonal analysis." J Clin Microbiol 43:3944-55.

Kumar, P.S., Leys, E.J., Bryk, J.M., Martinez, F.J., Moeschberger, M.L. and Griffen, A.L. (2006). "Changes in periodontal health status are associated with bacterial community

shifts as assessed by quantitative 16S cloning and sequencing." J Clin Microbiol **44**:3665-73.

Kumar, S., K. Tamura, et al. (2004). "MEGA 3: Integrated software for molecular evolutionary genetics analysis and sequence alignment." Brief Bioinform **5**:150-63.

Kutter, E. and Sulakvelidze, A. (2005). Bacteriophages Biology and Applications, CRC Press.

Lamfon, H., Porter, S.R., McCullough, M. and Pratten, J. (2003). "Formation of Candida albicans biofilms on non-shedding oral surfaces." Eur J Oral Sci **111**:465-71.

Larsen, J.B., Larsen, A., Bratbak, G. and Sandaa, R.A. (2008). "Phylogenetic analysis of members of the Phycodnaviridae virus family, using amplified fragments of the major capsid protein gene." Appl Environ Microbiol **74**:3048-57.

Lasken, R.S., Egholm, M. (2003). "Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens." Trends Biotechnol **21**:531-5.

Lasken, R.S. and Stockwell, T.B. (2007). "Mechanism of chimera formation during the Multiple Displacement Amplification reaction." BMC Biotechnol **7**:19.

Leiman, P.G., Chipman, P.R., Kostyuchenko, V.A., Mesyanzhinov, V.V. and Rossmann, M.G. 2004, "Three-dimensional rearrangement of proteins in the tail of bacteriophage T4 on infection of its host." Cell **118**:419-29.

Lepp, P.W., Brinig, M.M., Ouverney, C.C., Palm, K., Armitage, G.C. and Relman, D.A. (2004). "Methanogenic Archaea and human periodontal disease." Proc Natl Acad Sci USA **101**:6176-81.

Logan, N.A., Lappin-Scott, H.M. and Oyston, P.C.F. (2006). Prokayotic diversity mechanism and significance.

Lute, S., Aranha, H., Tremblay, D., Liang, D., Ackermann, H., Chu, B., Moineau, S. and Brorson, K. (2004). "Characterization of coliphage PR772 and evaluation of its use for virus filter performance testing." Appl Environ Microbiol **70**:4864-71.

Lyons, T., Scholten, T. and Palmer, J.C. (1983). "Oral amoebiasis: the role of Entamoeba gingivalis in periodontal disease." Quintessence Int Dent Dig **14**:1245-8.

Macarthur, D.J. and Jacques, N.A. (2003). "Proteome analysis of oral pathogens." J Dent Res **82**:870-6.

Magurran, A.E. (2004). "Measuring Biliogical Diversity." Blackwell Publishing, Malden, MA.

Mamone, A. (2003). "Ge netic sequencing and variation - representational amplification of genomic DNA", Amersham Biosciences, Piscataway, NJ, USA.

Marcy, Y., Ishoey, T., Lasken,Roger S., Stockwell,Timothy B., Walenz, B.P., Halpern, A.L., Beeson, K.Y., Goldberg, S.M. and Quake, S.R. (2007). "Nanoliter reactors improve multiple displacement amplification of genomes from single cells." PLoS Genet **3**:1702-8.

Mikhailov, V.S. and Rohrmann, G.F. (2002). "Baculovirus replication factor LEF-1 is a DNA primase." J Virol **76:**2287-97.

Miles, A.A. and Misra, S.S. (1938). "The estimation of the bactericidal power of the blood." J Hyg **38**:732–749

Miller, C.S., Avdiushko, S.A., Kryscio, R.J., Danaher, R.J. and Jacob, R.J. (2005). "Effect of prophylactic valacyclovir on the presence of human herpesvirus DNA in saliva of healthy individuals after dental treatment." J Clin Microbiol **43**:2173-80.

Munson, M.A., Banerjee, A., Watson, T.F. and Wade, W.G. (2004). "Molecular analysis of the microflora associated with dental caries." J Clin Microbiol **42**:3023-9.

Muyzer, G., Waal, E.C. and Uitterlinden, A.G. (1993). "Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA." Appl Environ Microbiol **59**:695-700.

Nelson, D. (2004). "Phage taxonomy: we agree to disagree." J Bacteriol **186**:7029-31.

Ostergaard, S., Brondsted, L. and Vogensen, F.K. (2001). "Identification of a replication protein and repeats essential for DNA replication of the temperate lactococcal bacteriophage TP901-1." Appl Environ Microbiol **67**:774-81.

Pace, N.R. (1997). "A molecular view of microbial diversity and the biosphere." Science **276**:734-40.

Pagaling, E., (2007). "Microbial diversity of Chinese lakes." PhD thesis at the University of Leicester

Pagaling, E., Haigh, R.D., Grant, W.D., Cowan, D.A., Jones, B.E., Ma, Y., Ventosa, A. and Heaphy, S. (2007). "Sequence analysis of an Archaeal virus isolated from a hypersaline lake in Inner Mongolia, China." BMC Genomics **8**:410.

Parfitt, T. (2005). "Georgia: an unlikely stronghold for bacteriophage therapy." Lancet **365**:2166-7.

Paster, B.J., Boches, S.K., Galvin, J.L., Ericson, R.E., Lau, C.N., Levanos, V.A., Sahasrabudhe, A. and Dewhirst, F.E. (2001). "Bacterial diversity in human subgingival plaque." J Bacteriol **183**:3770-83.

Patel, S.S. and Picha, K.M. (2000). "Structure and function of hexameric helicases." Annu Rev Biochem **69**:651-97.

Paul, J.H., Sullivan, M.B., Segall, A.M. and Rohwer, F. (2002). "Marine phage genomics." Comp Biochem Physiol B Biochem Mol Biol **133**:463-76.

Peciuliene, V., Balciuniene, I., Eriksen, H.M. and Haapasalo, M. (2000). "Isolation of Enterococcus faecalis in previously root-filled canals in a Lithuanian population." J Endod **26**:593-5.

Pedersen, M., Ostergaard, S., Bresciani, J. and Vogensen, F.K. (2000). "Mutational analysis of two structural genes of the temperate lactococcal bacteriophage TP901-1 involved in tail length determination and baseplate assembly." Virology **276**:315-28.

Piekarowicz, A., Klyz, A., Majchrzak, M., Adamczyk-Poplawska, M., Maugel, T.K. and Stein, D.C. (2007). "Characterization of the dsDNA prophage sequences in the genome of Neisseria gonorrhoeae and visualization of productive bacteriophage." BMC Microbiol **7**:66.

Prangishvili, D. and Garrett, R.A. (2004). "Exceptionally diverse morphotypes and genomes of crenarchaeal hyperthermophilic viruses." Biochem Soc Trans **32**:204-8.

Raghunathan, A., Ferguson, H.R., Bornarth, C.J., Song, W., Driscoll, M. and Lasken, R.S. (2005). "Genomic DNA amplification from a single bacterium." Appl Environ Microbiol **71**:3342-7.

Rappe, M.S., Kemp, P.F., Giovannoni, S.J. (1997). "Phylogenetic diversity of marine coastal picoplankton 16S rRNA genes cloned from the continental shelf off Cape hatteras," N.C. Limnol, Oceanogr. **42**:811-826.

Rogers, A.H. (2008). Molecular and Oral Microbiology, Caister Academic Press.

Rohwer, F. (2003). "Global phage diversity." Cell 113:141.

Roszak, D.B. and Colwell, R.R. (1987). "Survival strategies of bacteria in the natural environment." Microbiol Rev **51**:365-79.

Saitou, N. and Nei, M. (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol Biol Evol **4**:406-25.

Sakamoto, M., Huang, Y., Umeda, M., Ishikawa, I. and Benno, Y. (2002). "Detection of novel oral phylotypes associated with periodontitis." FEMS Microbiol Lett **217**:65-9.

Sambrook**,** J., Fritsch, E.F**.** and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual. New York, Cold Spring Harbour.

Sarnow, P., Cevallos, R.C. and Jan, E. (2005). "Takeover of host ribosomes by divergent IRES elements." Biochem Soc Trans **33**:1479-82.

Shannon, C.E., MD Comput **14:**306–317

Shaw, M.L., Stone, K.L., Colangelo, C.M., Gulcicek, E.E. and Palese, P. (2008). "Cellular proteins in influenza virus particles." PLoS Pathog **4**:e1000085.

Shine, J. and Dalgarno, L. (1975). European Journal of Biochemistry **57**:221-230

Shine, J. and Dalgarno, L. (1974). Proceedings of the National Academy of Sciences of the United States of America **71**:1342-1346.

Short, C.M. and Suttle, C.A. (2005). "Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments." Appl Environ Microbiol **71**:480-6.

Skurnik, M. Strauch, E. (2006). "Phage therapy: facts and fiction." Int J Med Microbiol **296**:5-14.

Smith, N.H., Holmes, E.C., Donovan, G.M., Carpenter, G.A., Spratt, B.G. and Spratt, B.G. (1999). "Networks and groups within the genus Neisseria: analysis of argF, recA, rho, and 16S rRNA sequences from human Neisseria species." Mol Biol Evol, 16(6):773-83.

Society for General Microbiology (2008, August 11). New Bacterial Species Found In Human Mouth. ScienceDaily.

Speicher, K.D. and Kolbas, O., Harper, S. and Speicher, D.W. (2000). "Systematic analysis of peptide recoveries from in-gel digestions for protein identifications in proteome studies. " J. Biomol. Tech **11**:74-86.

Spratt, D. (2008). "Microbial life in the mouth." Microbiology Today **35**:1464-0570

Staley, J.T. and Konopka, A. (1985). "Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats." Annu Rev Microbiol **39**:321-46.

Stephen J.B. (2007). "Bacterial Diversity and Competition from the population to the community Level." PhD thesis.

Stromsten, N.J., Benson, S.D., Burnett, R.M., Bamford, DH., Bamford D.H. and Bamford, J.K. (2003). "The Bacillus thuringiensis linear double-stranded DNA phage Bam35, which is highly similar to the Bacillus cereus linear plasmid pBClin15, has a prophage state." J Bacteriol **185**:6985-9.

Suau, A., Bonnet, R., Sutren, M., Godon, J.J., Gibson, G.R., Collins, M.D. and Dore, J. (1999). "Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut." Appl Environ Microbiol **65**:4799-807.

Sullivan, M.B., Waterbury, J.B. and Chisholm, S.W. (2003). "Cyanophages infecting the oceanic cyanobacterium Prochlorococcus." Nature **424**:1047-51.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." Mol Biol Evol **24**:1596-9.

Tanaka, Y., Primi, D., Wang, R.Y., and Umemura, T., Yeo, A.E., Mizokami, M., Alter, H.J. and Shih, J.W. (2001). "Genomic and molecular evolutionary analysis of a newly identified infectious agent (SEN virus) and its relationship to the TT virus family." J Infect Dis **183**:359-67.

Torsvik, V., Goksoyr, J. and Sorheim, R. (1996). "Total bacterial diversity in soil and sediment communities a review." Appl Environ Microbiol **56**:776-781.

Torsvik, V., Goksoyr, J. and Daae, F.L. (1990). "High diversity in DNA of soil bacteria." Appl Environ Microbiol **56**:782-7.

Vickerman, M.M., Brossard, K.A., Funk, D.B., Jesionowski, A.M. and Gill, S.R. (2007). "Phylogenetic analysis of bacterial and archaeal species in symptomatic and asymptomatic endodontic infections." J Med Microbiol **56**:1473-5644

Wang, J., Hu, B., Xu, M., Yan, Q., Liu, S., Zhu, X., Sun, Z., Reed, E., Ding, L., Gong, J., Li, Q.Q. and Hu, J. (2006). "Use of bacteriophage in the treatment of experimental animal bacteremia from imipenem-resistant Pseudomonas aeruginosa." Int J Mol Med **17**:309-17.

Wantland, W., Wantland, E., Remo, J. and Winquist, D. (1958). "Studies on human mouth protozoa." J Dent Res **37**:949-50.

Weinbauer, M.G. (2004). "Ecology of prokaryotic viruses." FEMS Microbiol Rev **28**:127-81.

Weisburg, W. G., S. M. Barns, et al. (1991). "16S ribosomal DNA amplification for phylogenetic study." J Bacteriol **173**: 697-703.

Wilhelm, S.W., Carberry, M.J., Eldridge, M.L., Poorvin, L., Saxton, M.A. and Doblin, M.A. (2006). "Marine and freshwater cyanophages in a Laurentian Great Lake: evidence from infectivity assays and molecular analyses of g20 genes." Appl Environ Microbiol **72**:4957-63.

Williamson, K.E., Radosevich, M. and Wommack, K.E. (2005). "Abundance and diversity of viruses in six Delaware soils." Appl Environ Microbiol **71**:3119-25.

Xiang, Y., Morais, M.C., Battisti, A.J., Grimes, S., Jardine, P.J., Anderson, D.L. and Rossmann, M.G. (2006). "Structural changes of bacteriophage phi29 upon DNA packaging and release." EMBO J **25**:5229-39.

Zhang, T., Breitbart, M., Lee, W.H., Run, J.Q., Wei, C.L., Soh, S.W., Hibberd, M.L., Liu, E.T., Rohwer, F. and Ruan, Y. (2006). "RNA viral community in human feces: prevalence of plant pathogenic viruses." **PLoS Biol 4**:e3.

Appendix of Chapter 3

The total nucleotides observed by joining the four contigs and one sequence of the partial phage genome that had signinficant matches to the prophage in the *Corynebcterium diphtheria* genome.

```
   1 CAGGCCGATC CTGCTCCCGG CCGCATGGCG GCCGCGGGAA TTCGATTATA
  51 GGGGGAGATA GGGCCGCTGG TGCCGGATAC CCAAGCGGGT TGGTCAGGTA
 101 GGGGACCGTC GTCATCGTGT ACTACCAGGG GGCATCTGGC TATGCTGCTG
 151 ACGATGATGC GGCGGGCGCG CGCTAACGCG GCGACATTCA TGGCGACGTC
 201 ACGGGTTATT GATTCGGGTA GCAGGTCAGG GGTGCCTACT GTGATGAGGT
 251 GGTTTGGGTC AGCCCTTGAC GCCGCCCGTG AGAATCAAGT GATTCACCCT
 301 ATAGATGAAG GGATCGCCAC GGTGCTTCGA GCAGGCGCCT GGGCCCTCGA
 351 CACTTTGGAA AAACAGGACC GGCCTTATGG GCCGGCAAAG CTCATTCCGG
 401 CCATGACCGA GGCACTCACT GCGGCGCACA TGACGCCCGA AAGCCGGAAG
 451 CTGGAAAGCG AAGACCTAGC CAAACAGCTA TTCGAGGACT TAGCCGCCCT
 501 AGAAGGCCGA TGCCAGACTG AATGCCGCAA GTGGCTACCT GGCCGGGTAG
 551 AACCCCGCTA CCTAACCCCT ATCCCCGAGG GAGCCATAGT CGACCTCAGG
 601 GCGGTGAAGA AGGTTGCCGC AATCATGGGC CGGCAACCGA CGTTCTACCA
 651 GGTGGAAATC CTCGAACGCC TGGTAGCCAA GTGGCCTGAC GGCACCCCGG
 701 TTTTTACCAC CATTCTGGTG AGTTTCCCCA GGCAGACCGG TAAAACCACG
 751 TGCATCATGG ATTGGCTCAT GTATGTGGCC ATGACCCGCC CCTATCAAAA
 801 GCTCTGGTTC ACTGCCCAGA CCGGTATGGC GGCTAGGGAA CGTTTTCTTG
 851 CTGAGCTGGT AGAGCCCAGC AAAAAATATC TCGAACCGCT GGGGATCGTA
 901 GATACCAAGC TTGCCGCGGG GGCGACCAGA ACGGTAGTGG TGGCCACGGG
 951 TTCCCAAATC CGCCCAATGC CGCCGACCAG CCAGTACCTA CACGGTGGCC
1001 AAGGCGACAA AATCATCGCT GACGAACAAT GGGCTTTCAC CCAGAAGCAA
1051 GGGAAAGACC TCATGCAGGC GGTGCGCGCT ACCCAGCTGA CCAGGAATAA
1101 TAGCCAGATT GTGCAAATCA GCGCCGCTGG TGACGCCGAA TCCGACTATT
1151 GGCATGCCCG ATTGGCCAAA GCCATTGCCG AGCCTTCGCC CCGCGTGGCA
```

1201 GTAATCGACT ACGGGGTAGG CACTAGCGCT GATCCTCAGG AGGTCACTTC
1251 CTTCACCATC GAGGATGTTC TAGCCGCTCA CCCTGGTGTC GCCGCTGGTC
1301 TGTGCACCCG CGAAAAGGTC TTGGAGCCTT TGGAAAACGA GGACATGGAC
1351 TTCAACGAAT GGTTGCGCGC GTATGGCAAT GTGCGCTCAA AGAACACGCG
1401 CCAGAAGGCC ATTGATCTAG ACGCCTACCG CAGTATCACC ACCACGGTGC
1451 CGCTAGACGA CGGTCCGGTG ATGCTGGGGG TTGGCGTTTC CTGGGACGGG
1501 GCGACTACCG CCCTAGCTGC GGTAGGCACC ATCAACCAAG GCCGGGGCGT
1551 GGGTATCGAG ATCATCGACG CCCCGCCCTG GCCGGCAATG GGTCATCGAC
1601 ACTACCCAAG AACTAGTGCG CCGCGGTATC GCCACCGAGG TATGCGGCGA
1651 CGCCTACGGA CCCACGAAGC GCCTTGCCGA CCAGCTAGCT ATTGCCCTTC
1701 CTGAGCACTG GAAACCCCTA TCCACTGACG AGATGATCGC CGCCACCGAG
1751 GACTTTCTAC AGGCCCTCGA CCAGGAAGCC GATACTATGC CTATCCGAGT
1801 CCGCCGCTGT GCTGGTGTCG AATACGAGCT AGACGTCGCT GAGCTGCGGA
1851 ATGTCGGCGA GAAAGGAGGG ATGTTCAGCA GACGCAACAG CGCCGCTGGC
1901 ACCGCACGGC TAGAGGCCGG CCTAGCCGCC CTGGCCGGCT ATCAAATTCC
1951 CGAAACCACC GCCCCCGAAC CTTTTATTGG ATAAATTATG CGAAACCAAA
2001 AACGGAAGTC ACCAGCAATC GACGCCACCG AACACACTAT CCTCATCACA
2051 TGCGATAAAT GCGAATGGCG AGAAATGCAT GATGACCGGA ACGCCGCCTG
2101 GTACGCACTG GCACGGCACT TGAAAACCGG CCACGATGAC CCCTATGCCG
2151 CCAAAAGCGC CGCCCGAAAT ATCTACCGCA ACCACCACGA ATAGCCGTTT
2201 TGTCGCCCCC TTGCTGCATC ATTAGGGCAT GGGGTTCTTC GAGAAAGTAA
2251 GACAGGCACT CTCCCTCCCC GCCCTAGCGG CGGGGAGCCT CGAAGTGCCC
2301 TACGCCAGCG CCTGGGCTGA CCCAAACCAC CTCATCACAG TAGGCACCCC
2351 TGACCTGCTA CCCGAATCAA TAACCCGTGA CGTCGCCATG AATGTCGCCG
2401 CGTTAGCGCG CGCCCGCCGC ATCATCGTCA GCAGCATAGC CAGATGCCCC
2451 CTGGTAGTAC ACGATGACGA CGGTCCCCTA CCTGACCAAC CCGCTTGGGT
2501 ATCCGGCACC AGCGGCCCTA TCTCCCCCTA TCACCGCATG CTGTGGACAG
2551 TTGATGACCT GCTGTTTTAC GGGTGGTCGC TGTGGGCAGT AAAACGGAAC
2601 GGGGCCGGTG CTGTTGTCGC CGCCGATCAC GTGCTCTACG AACGCTGGGG
2651 TTTCACCCCC AATGGTGAGG TATATTTCGA GGGCGAAGAG ATATGCGCCC
2701 GATGATGTTA TCCTTATCCA ACCGGCTCTG ACCAGGGAAT CCTTCGCTAC
2751 CCGAGCTGCT ATCCGGCATG CCGCCCAGAT CAACGCCGCC GCCGCTTCCG
2801 CCGCCGCCAA CCCGGTTGCC CATACTGAGC TGCACCGGTC TGGGCAGTGA
2851 ACCAATGCAC TGATCCGGTG AAAATTGATG CGGCTAATAG AGGCATGGAA
2901 CCGCGGCCGG AACCGCCCGG GGGGCAGTGT CGGTTTCACA AACAGCAGCA

2951 TTGAAGCAAA GAGCCATGGC TCTTTCGAGG CTCACCTTTT GGTGGAAGGC
3001 CGGAATGCTG CTGCTATTGA TATCGCCCGC GTCTGTGGCA TACCAGCTAT
3051 CCTGTTGGAT GCTTCCCTAG CCGACTCTAG TATCCGCTAC TCCAATATGG
3101 ATGCGCGCAA CGTTGAATTG GTTGATTACT GCTTAGCCTC ATTTATGGCG
3151 CCGATTGCCG CCCGCCTAGG CATGGATGAT GTTGTTGCCC CCGGCCAGAG
3201 TGTGGAGTTT GACCTCGATC ATCTGACCCG CCTCGATCCT AACAGCATCG
3251 CCCCGCCTGA CGACGCCCAC CGACCACGCC CCACCCCCGC CACCAACGAA
3301 CTAACCCAGC TAATTAACTA ATTATGGATT TTCAAACACT AGAACCTGAT
3351 CTATATCAGT TGATGAACAA GCATTACACG CCCGGCCGAC CAGGCCCCAT
3401 CAAATACCTG GTGATCCACC ACAATGCAGG CGTCAATCTC AGCACCGCCG
3451 ATTGCTGGCG GATTTGGCAA GACCGCGAGG CTAGCGCCCA CTACCAGGTA
3501 GAAGTGGATG GGACCATTGG TCAGCTGGTC AACGATTGGG ACACCGCCTG
3551 GCACGCTGGT GACGCCGCCG CCAATTCCTA CTCGATTGGC ATTGAGCATG
3601 CTAACGTTGG TGGCGCCGCT GAAGATTGGC CTATCAGTCA GGAAACTATC
3651 ACCGCAGGCG CTCACCTGGT TGCCGCCCTG TGCCACGCCT ACGACCTGGG
3701 AAAACCCGCA TGGTTCAATA ACGTCTTTCC ACACTCATAC TTCTATAGCA
3751 CCAGTTGCCC CTACCAGCTG GCCGGTGCGG ACCGTGACCA ATACATGTCT
3801 TTGGCTGAAG AGTTTTACTT CAGCATGCAA GCAGGAAATA CACCACAAGC
3851 AGGGAAAATG ACGAATTTTA CTGAAGCTGA CCGGCAACTA CTCCGCGAGA
3901 ATAACGAGTT ACTGCGGGTG ATCCGCGACC AGATTTGCGG CCCTGGTAGT
3951 GGATTCCCTG GTTGGCCACA AACCGGTGGC CGGACCCTGG TTGATACGGT
4001 TGCCGCCCTA GGCGCTGCTC AGGGGATTGA TGGTTGCCGC GACACCAAGA
4051 AAGCCAAGTG ACCGTGAGCC TTCTTGATTC CACGCAAATT CGTTACCCGT
4101 GGCGTTCGGT AATCCGTAGC GTTGCCGTGG CCACTATCGC GCTGCTACCG
4151 GTGCTACCGG AAATCGCCAA GGCGGCGGGT GTGGAAACCG TGCCGCTGGT
4201 TGCTTCCACG CTGGGAATTG TGGCGGTTTT GCAGCGGATA ATCACGATCC
4251 CCGAGGTTGA TAGGTGGTTG ACCACCACGA TCAACATGGG CGCTAGGAAG
4301 CGCCAGGAAG AGATAGGAGA AGGAAATGCC GAGTGATCTA GAAACCATCA
4351 ATGGTGACGC GGCCCCCGCA ACCGTGTCAT GCAACGAATC CGAACGAATT
4401 ATGGAAGGCC TTGTCCTCCC CTGGGGTGAT ACTGGTGCTA CCGCTACTGG
4451 AAGTTACACG TTTCCCCGCG GTAGTCTCGA TATCCCTTCC AACATCGAGC
4501 GGGTAAAGCT ACTATCTGAG CATTCACGCC CTGGTCATCA GCCCAAGGCC
4551 ATTGGCCATG CTATCAGTGC CGAGAACACG CCTGAAGGCC TGGTTATGCG
4601 TTTTCAGCTG GGCAGTAGCG CCGCCGCCAC CGAAGCCCTC ACCAATGCCG
4651 CTGAGCACAT CATTGACTCT TTCAGCATCG AGGCGGTGGG TGTGCGCCGC

4701 ACCGGTGGCA CAATTGATTC TGCCCTGCTC AAAGCTGTTG CACTGGTGCC
4751 ATTCCCAGCG TTTGAGAAAG CCAAGGTATA CGCCGAGTCT GGCAACCTCG
4801 AAGAGAAAGA AACCACAGAA ATGACCCTAT CCGCTGAAGA TATTGCTGCT
4851 ATCGCCGCGA AAGTCACCGA GAACCTCAGT TCCACTACAG CCACTCCCCG
4901 GAATAAGATT CCGGCCGGTA TCCCAGGCGG TAAGGACGCC GCTAAGCCGG
4951 AAGTCATCAC CGCCGCCCAT GCCGCCGAGA CTATCTTGGG AATCCACACG
5001 GGTGAAATCC CTGATGATGA GATCCAAGCT GCCCTTGCCG ATATTAAGGG
5051 CTCAGATTCG ATTGTCACCC AGCCCAAGGC TTGGCTGGGG GAATTGTGGT
5101 CTGGTGTTGT TTACCAGCGC CGCATTATCC CGCTAATCGC CACCAAGGCA
5151 CTAACTGGCC CGGAAGGCTA TTGGTTTCCG CTGGAAGAAG GACGCCGATA
5201 GCGGAAAGCT GCTCAAGCCT GGTGTTGCCA AGTGGTCCGG CAATAAAACC
5251 GAGATTCCCA CGCAAAAAGC CCAGTGGGAA GAGGTGTCGA TGGATGCCCA
5301 GCCCTGGGCC GGTGGCAATG ACCTCGACCG GCAAATTTTC GATTTTAACG
5351 AGGCCGAAGC GCTGCTGGCC TACTGGCAAG CCATGAATGA GTCCTATGCT
5401 TTCGAGACCG ACCGCGATGC TGGGAAATTC CTGGTAGACC ACGCAACTGA
5451 TATCCCTGAG GTCGCCCAAG ATATTATCCG CGCTATCACC ATTGGCGCCA
5501 TTCGGGTTGA CGAGGCGGTG CATGTCCCCG CCGCCTACGC CATTGTCAAC
5551 CCCCGTGACC TCGAAAAGGT TCTGAAATAT TCGCAGCTGG ATGTTCCGCA
5601 CTACATGAGC CGACCCCGGT GTCTGAACCG GCAACATGGA CAACTTCTGA
5651 GTTTGTCGAG TCCGGCACCG CCATTGTTGG CTGCAAGGAC GCTACAACGT
5701 TTTTCGAGCT CCCCGGTTCC CCACTGCGTG CCGAAGCTGA GCATATCGCC
5751 CATGGTGGCC GAGATGTTGG CCTTTTCGGC TATACCGCGC ATATGCTCAA
5801 TCGTGCTGAG GGCCTGGTCA AGGTGCATTT TAACAATGCC TAAGGTGGAA
5851 GATTCAGAGG TTCCGGCCTG GCTTGGTGTT GATGCGGTGG GTGACGCCGC
5901 CGAACAGCAG GCGCTGAAGG GGATTGTGGC GGCGGTTAAC GCCACTGTGA
5951 CGGATTGGCA TGGTAACCCA GACTCTTGGT CCGACCGGAT TCATACTGGT
6001 GCCGTGATGC TTGCTGCCCA CCTGTGGCGG CGGCGTGCTA CGCCTGGTGG
6051 TGTAGCAGCC CTAACCGACG AAGGCACCGC CTACGTGCAG CGTCATGACC
6101 CCCAAGCCGC CATGCTGCTT GGTCTTGGCG GTTGGACTGC CCCGGCGGTG
6151 GGATAGATGA CGCCCGATAA TATCCCTTTG CACCTGGGCC GGTTAGCTCA
6201 GGAAATCACC CGCGATACCG GTATCCCCGC CACGATCCAC CCGAATTTGG
6251 TTAATGCACC CGGTGCGTGG GTCAGCCTAA AAGATCTTGA TTTGGAATCC
6301 ATGGCCCGCG GGGAAGTCAT GGCCACCGCA AGCGTCTATC TTGTTGCCGC
6351 CGATCTAGGC ACGACCCTAG CAGTCGAACA CCTGACCAGC ATGCTAGACG
6401 ATCTACTACG CCTGGTAGAG GGCCGATATC CCACAAACGT AGAGGTCACC

6451 ACCATCACAC TGCCGAGCTT TGGTCAGGTA CCGTTACCCG CTATTGAAGT
6501 TGAATACGAA CTGAAAGGAA CATGACGATG GCGAATGTCA ACACTCTGGA
6551 CAGCCGAATC TCGACCGGCC CCGGAAAATT GGTTTTCGGT AAGGCTGGTG
6601 CCCAGAACGA AGTTTTCCGC CCTGGTCACC AAGGCCGAGC TGAACCCCGC
6651 TGTAAATACC GAGGACGGTA AGCACGTCCT TTCGGGTGAT TATGCGCCCG
6701 GCAAAGACAC CATCACGTGG ACAATGGAGC TTACCTGTTT TATTAATCTC
6751 AAGAAAAATG GGATTTTTGA TTGGTGTTTT GCTAACCGCG GTAAGGAAGT
6801 TGAGTTTGAG TTCCGGCCGG TAGAGGGCGA GAAATCGGCA AAATTTACCG
6851 GCACGGTGAA AGTCCGCCCG TTGGGTGTTG GTGGTGAGGT GAATAAGGAG
6901 ATGAGTAAAG ACCTCACATT TCCGCTGGTT GGGGAGCCGG TTTTCACACC
6951 TGAGGAACCG CTAAGATTGT CCGGCCATGT GGATGTGTCC GCCGAGGTGG
7001 AGGGGTTGAA AAATCTTCGC CGCACTATTC GGCAAGCTGG TGGCGACACG
7051 AAGGATTTGC GCAATGCCAA TCTAGCCGCT GCGCAGACCA TCGTGCCTAT
7101 CGCTGCTGGT CTGGCGCCGA AGGTCACCGG CCGGTTGGCG GCGAGTATCA
7151 GGGCGGGTGC CACTCAGAAG GCCGGCATGG TCAGGGCCGG CCGGAAATTA
7201 ATACCCTACG CAAACCCTAT CCACTGGGGT TGGCCAAAGC GCCACATCGC
7251 ACCGAACCCG TGGATTGCTA CCGCCGCCGC CGCCAACGAG GAACTGTGGC
7301 TCAAAGTCTA TGAGCAGCAT ATTGACCGTA TTTTAGGAAA GATTGAAGGA
7351 AAGAAACGAT GAAACTTGTC ATTAATGTGA GGTACACCAG CGGTGAGGAG
7401 GTCACCGTAA CGCCTATTCT GTCTGACCAG GTTGCTTTTG AGCGTACCGC
7451 CCGCCTTCGT GATTGGGGCA CCGCAACCGA CAGCCCCTTA ACCTTTGCTG
7501 CTTTCTTGGC GTGGAAGGCG CTACAGCGCA CCGGCCAAAA CCGAATACAG
7551 TTTCGAGGAA TTTTTAGAGA GTGTCGAGGC CCCTAAGCCC AGTCTGGCGG
7601 TGAGATGGGT TTTAGCCCCC TACGGAGGCG ACGCCCTGCC GCGTCATAGC
7651 CCTATTGTCC GTGAACACGG GGGATTCCCG CCCTAGTGTG CTGCTGGCGG
7701 AAGACCCCGC ATGGATAGAT ACGATGCTGG AAGTCATGGC TGAGCAGGCG
7751 GAAGCAGCGA AAAAAGAGAT AAAAGAGGTA ACCGGTGGCA GGGAAAAAGA
7801 AGTCGGCAAT CCTGTCGGTC AACATTGTCA GTGACGCCAA CACGAAGGGC
7851 TTCTCTGAGG CGGCGCGCGC CGCCCAGAAG ATGGCGGCCG ATATCAACGC
7901 TTCGACTGCC CAAGCGGCCG GCATGGCCAC TAAGATAGGT GGGCTGACCA
7951 CCGGAATTAC CTCCCTTGTC TCTATTGCTG GTGGCGCCAT TGGTCAGGTT
8001 GCTGCTGGTG CCACTGCGCT AGCGGCGGTG GCCGGCCCCG CCCTAGGCGC
8051 CGTCGTGCTA GGCTTCGACG GTATCAAGGA AGCCGCCGAA GGGCTCAAGG
8101 AACCTTTCGA TGGCTTGAAG GAATCTGTAT CAGGCGAGTT TGCCGCGGCG
8151 TTGGAGGAGC CCTTCGAGAA CCTGGGTGGC CTTATTACCG ATTTAGAGGG

8201 GCCAATGGCT GGGCTTGGTG CCTCCGTGGG TAATCCATCA TGGGGGGCCT
8251 TGTTGATACG ATTGATCAGC AATCAAAGTG AATTAGAGAA GCTGATAGCT
8301 TCTGCTGGCC AGTTCACCGA CGCCATGGGG CCGGGATTGA ACACCCTGCT
8351 AGAGGGGGTT TTATCCATTG GCACCGGCCT AGACGGCATA GCAGGCGATT
8401 TTGGTGCTGC GTTCGGTGGC GTCCTCGAAA CCCTAGGCGA GAAGTTCCAG
8451 GAATATGCTT CCAATGGCGC CACTACTGCT CTGATTCAGG GCATGATCGA
8501 CGCCCTAGGC GGGTTGTCGG ATTTGATAGG CCCCTTGCTG GATTTGATTG
8551 TTGAGCTGGG CATTGCCCTA GGCCCTTCGT TTGGCGGCGT TCTGTCTGCC
8601 CTGGGTGGGA TTATCGCCCA GCTTGTGGAG CCGCTTTCTA CTATCGCCCA
8651 GGTGGCGGGT GTGGCGCTGG TTGATGCGTT AGTCGCTTTG TCGCCAATGT
8701 TTGGGCCGAT AGCTCAGGCA ATCGCCGATT TGGTGGTGGC GTTGGCGCCG
8751 TTGTTGCCGT CGATTGCTGA GCTGGTGGCG TTCCTGGGCA CAGCGTTGGC
8801 TGAGGCAATT AGTGCCGTGG CGCCGCTGGT TGGGGACATT TCCGCCCTGC
8851 TGGGTGAGGT GTTCCGCATA GCCATTGATG CCTTGACGCC GATTATGCCT
8901 GTCATTATCG AGCTGATTCA GACGCTGGCC GGTGTGGCTA GTGAACTGTT
8951 GCCGTCGATT GCTGAGCTGG CCAGTGTGTT GTTCCCGGCG TTTGCCCAAA
9001 TCATGGAGGC TATCGCCCCG ATTCTGGGTG ATATCGGTGC CCTGATTGGT
9051 GATGTGCTCC GCATGGCCAT TGAGGCAGTG ATTCCGCTGA TTCCGGTGAT
9101 CGTCGATACG ATCCGCATTC TGGCTGACGT GGTGGCCATG CTGATTCCGG
9151 TGATCGCTGA GGTAGCGCAG TTCCTTTTCC CCGCCCTAGC TGAGATTCTT
9201 CAGGTACGTC GCCCCATTGC TTCCTGATCT GGCTAATTTG ATAAAGTCCC
9251 TGATTGAGGC CTTATTGCCG ATTATTCCGC CCCTGATGCA GGTAGCGGAA
9301 GCCCTGTTCC CCGCCCTGGT CCGAATCATT GAGCTGATTA TCCCGATAAT
9351 CATTCAGGTG GCCGATATCT TCGTGCAGCT GGTGCAGGCG CTCACGCCGC
9401 TGCTACCGCC GTTGGCCGAT TTAATTACCG AGCTACTACC GCCGATTGTT
9451 GAGCTGATGG AGGCAATCGC CCCGGCAACC AGCGCTGTTG TCGGGATTGT
9501 CGGCAAACTC GCCGTCGCGC TGACCAAGGG CCTGGTGGAT GCGGTAATCG
9551 CCATTGGCGG TAAGCTGGGC TGGCTCAAGG ACCTGTTCTT TACGATCATC
9601 GACGTCATCA AAAAGGCATT TAATTGGATC ACCGACTTCT TGGATGCGGC
9651 CGATGGCGTT GGTGGGATTT TCGGCGGCGG CGGTAGTTTT GGCGGCGTAG
9701 GCGGCGGGGG CGGTGGTTTC GTTGGCGGCG GCGACGATGG GACGTTCCAT
9751 GGGGCCGGTG GCGGCGGTAT TGGCGCCGCC TTCCACAACC TACTAAACCG
9801 GCCACTACCA GCACCCCAGG TCATCAATAA TTTTGAGATT GTCATCAACG
9851 GCCCCATCGA CGCCCTAGAG ACCGGCCGGA AACTCCGCGA AATCCTCGAC
9901 TACTACGATG AGAGGATGAA ACGTTAATGG GTGTCATGGC GAACATGCTG

```
 9951 CAAATTTCAA TCTTCCCGCC GAACAGCCAA TGGAAACCTG AACTTACGTG
10001 CCGTCGTTGA CGGTCTCACC ATCAACTGGG GGCGCACAAA CCTTTACCGT
10051 GCCCCAGCCC AATCGCACGT GTCAATTCCA AATGCTGATG GAAGCACGTT
10101 ACTTTATCAC GGGTAATGCA GAAATGGGTC AATTCGGAAT TGATTATTAC
10151 GGCTAAACCC GCGAGTGGCG ATTTGGTGAT ATTTCAGGGC ATTATCGACG
10201 ATTTTAAAGT CACTCCGAAA GACACGAAAA TTGGTGATTA TATTGTTGAT
10251 TTTACCGCTA CTGAATCGCC TACCTGGTCA AATAAGCTCA ATGGATTATT
10301 TTATGATGCT AAAAACCTGG CCGTTGCGAT TTTAATTGCC TCGTTTAGGG
10351 CGTGTCCAAC GTGAATTAGG CACATTTATT GCCCTGGATA TAAATACAAG
10401 TTATTTGGCT GAACCACCCG AGAATCAAAT CAGTGTGAAA CAACTAGCTG
10451 AATCCCTTGT CTGGAGACCG GGGGCTTTCC CCGCCTGGTG CCCCGATTGG
10501 AAAAGGCTAG CGCCGACGGT TCACCAGCTG GACACCCCAG AGGGCGGCGC
10551 ACCGTGGGTA TTGTCCCCCA AGGTTTTAAT TGACTTGGAT CAGGGCATGT
10601 CATGGACTTC CGATAACACA CCCACGACCA TCTTGTATAG TGCTGGTGGC
10651 CTTTTCGGGA AAAGTAAATA CGCCCGCGAT ACCCGTGTCC TACGTGAAAC
10701 CCGCGATCAA TGGGATCGCG GTAACATCGT AGAGCTTGAT ACTCCGTATT
10751 GCCCAGATCA GGGCGGTATT CTCGGCTATG CCGAGAATCA TTCTGAGCTG
10801 GCGAAAGCCC AGCTTGGGAG CCCCCGCCGA ATCCGGCTTG ATACCCGCCG
10851 AAACGCCGAC TTCCTCAACA CCTACCTGGG GTGGGAATGC TGGGAGACTC
10901 CGAACCGGTA TATTCAGGTG ACAGGGGATA AGTGGGCAAC AAAGTATCAT
10951 GGTGAGCTGC TATTACAACA AACCTACTAC CCAATTGGCG GGACATTGAC
11001 GCTGTATCAC TGGGGTTTTA CCCATGATCT TTACTGTGCC TGGGGGCCAA
11051 CAGACGACGC TTTAACGCCC CCGCCGCCAC CGCCTCCCCC GCCGCCGCCG
11101 AAGCCTACCA CTTGGGCAAC CACCACGACT ACCTGGGCTA CTACTGCTGG
11151 AACTTGGAAA GGATAGATTA TTTCATGGCC ACCGCCGACC CCCGCAACGT
11201 TCAATACCTG AATGCTGACG GGAGTGACAC AATCTCTCAA TTTCCTTCGG
11251 TTCAACGCAA TAATGCCACT CGCTTATCAG AGGCGATTAC TACGAGCAGC
11301 GATACCGTGG CGCTTAATTC CACTTTCCGC AATGCTTCCG GCCTGATTCA
11351 AAGGATAGGG AAATTGCGGA TTTTAAGCCT TGAATTTCGC ACTTCTAGTG
11401 ATGCGGTAGC TGGTACCCGG CTTCTGGCGA ACAATCTTGC CACCGGTGAC
11451 CGGCCGGCAA AAACCATCTA CGCCGCTTTG GCCGGCGCTA ACGATTTGAA
11501 CGACGCTGTA GGTGTAAGGG CCCGCCTAGG CACCGACGGC ACCGTTACCT
11551 GTCC
```

Sequences of the Amplified Viruses Genes from the Human Dental Plaque.

>Contig 5
TAGGAGGTGTATCCAACGCACCCTGCAACGAATGCAGAGGGTGCCAGATTACCAGCGTTC
TTCAATGTTGAAGTCGTTTTGGATTCGGTTTCCTCGCCATTATAGGCTTCGAGGGAAGTG
CTGTTAGAGTCAGCGATGTTTGCGAGATTGTCAGACATGACAGTGAGCCTTTCTTAACTC
GTAGAGGTATCTCATATAGAGGGTGGTTTTTTGTACGAGATTTCAGAAAAAAATTAAACC
CCGGTTTTTAAGCCGAGGTTTAATCCTAGGGATTATTCCTTTGTGGAATGGTCAGTATGA
ACTGATCCGTTTGTCACTTACCGGAATAGCCGATCAACATAGGTTGCGCCGATGGTCAGT
GCGCCAATAGCCAGTACATAGGCTGCTGTGACTTGCACAGCGCCAATGACAAAGCCCTTA
GCGAACTGCTTTGCGACGCCAGCGTAGGTGATGGTCTCGTCGGTAACAGTGGTGTTAGCG
GTTTGGGTGTTCATGGTTTGAACCTTTCTTGTTCAGTAGGTGGGTATCTCCCATATACAG
CATGGTTTTTTCGTACGAATCCCATGCCGTATACTATACTCTACCGAACGCGATTGGATA
CATCGAGTGGCGTAGGACGGCTTCCTGACAGATTTGCAAATTACTCAACAAGAAGGCAGA
CGCACAAGGAAGCGTTCGACAAAGCAGTCGAAGCGATGGCACGCAAAATGGTTGCAGAAG
AAAATGTAAAACTACAAGCCCAATCTACTGAACCCGAAGAACAACCTCAAGAACTAGAAG
ATAAGGAGTAACACCCTATACCTCGGCTTAAAAACCGGGGTTTAATTTTTTCTGAAATCT
CGTACAAAAAACCACCCTCTATATGAGATACCTCTACGAGTTAAGAAAGGCTCACTGTCA
TGTCTGACAATCTCGCAAACATCGCTGACTCTAACAGCACTTCCCTCGAAGCCTATAATG
GCGAGGAAACCGAATCCAAAACGACTTCAACATTGAAGAACGCTGGTAATCTGGCACTCT
CTGCATTCGTTGCAGGGTGCGTTGGATACACCTCCTCTAATCACTAGTGAATTCGCGGCC
GCCTGCAGGTCGACCATATGGGAG
>Contig 6
TGACTGTGAGTGTGCGTGTTGTGTGTGGTGCGTGTGTGTGTGCGTGTTGTGTGTGCATAT
GTGTATGACTTTTTGCATGGTGTGTCTGGTGTGTGCATGTGTGTGTGCATGTGTGTGTAT
GACTGTGAGGGGAAAGGGGAGGACAGGGGAGGACACTGGGATGGTGTGTGTGTGAAGTGT
GTGAATGCGTGTGTGTTGTGTGTGAATGTGTGGTGTGTGTGGTGAGTGAATGTGTGTGTT
GTGTGTGGTATATGTGTGTGCATGTGTGTGAATGTGTGTGT
>seq 7
TGATTTAAGTGACGAGCAACTAAGAGAGCGAGGCTATACCGAAGAACAAATCGCCTCGAT
TCAAAAACTGGCTCAAACAGCTAATGGCGCAGCTCAAGACGTTAAGACTTTCACCCAGAT
GATTGGTACCATCAATGAATCAATTGGCTCGGGTTGGGCTCAGTCTTGGCGTATTATCAT
CGGTGACTTTGAGGAAGCCAAAGTTTTATGGACTCGGGTCAGTAAAGTCATTTCTGGTGC
TGTAGACCAATCAGCCAAAAGTCGTAACGAACTACTCCAAGGCTGGAGTGATGCTGGTGG
TCGTACCATCGTCATTGATTCTCTGGTACGAGTGTTCAACGGCTTGTACAGCATAGTAGG
GCAAGTAGGTCGTGCTTTCTCTCAGATATTCCCGCCCATGACAGTCGCTCAGGTGATGAA
GCTCACACGAAGTTTCGAGGACCTGTCCTACAAACTAACACCTTCTATCGAAACTGTTGT
GAATCTCGGCAGGACGTTCAAGGGTTTCTTTGCGTTACTACACATCGGTTGGTCGTTAAT
TAAAGCTGTCGGTTCTATGATAGCTAAGATTTTCGGTGCTTCTGGTGGTGCAGCTGGCGG
GTTGCTCTCAATGACAGCTTCGCTAGGCGACTTCCTGGTGAAGCTCGATGAGAGCATTCA
AAATGGGAAGCTGTTTGAGAACATCTTCGGTAAGGCCGGCACTTTCATAGCTGGTGTATT
CGAGTGGATCGGTAAATCGGTCGATGACACCAGCGTAGCCTGGGGTAAATTCCTAGACTT
CGTTCATAAAGTAACAAACGGTCTCATCAAGATTTTCAACATCCTCACCTCCGGTAAGTA
TCAAGACGGTACTTTCATGGGGGTTGAAAAGTGATGCTGGCTTCATCA
>seq 8
GGACGGAAACGTAGTACATGGCACCCGCACAGCGGGGCGAGAGCATGCACAGCTTCGGCG

TAGAGGCGCTGCACGGTGCGGATCTGGTCAGTTAGAGAGAGGTCGTTGGTCTGGGCCTGG
CTGGCAACGTTCCAGAAGCCCTCGGTGCTGTCACCGACGTCGGCGATGATGATGCGCTTG
TAGGAGCCTCTGAAGCGGGTGTCGTCAGCGATATCGTGGATAGCTCTACGGACGAGCCTC
ACCGTGTCCTCAGTGCCACCTCCCGCCTGGACCTTCCCGCATTGAAGATCCGCCAGGCAG
ACCACGAGCGTGTCTTCGTCATTCTTGATGAGCGGGGCGGGCTTCGGCAGGAGGGGCTCC
CGGAAGACCGGCTCGAGGTCCTCGTAGGAGAGGCGCTTGGCTTCAGCCATCTCGACGGCG
CCCGGCTTCCAGGTGATTTTCTCATAGGAGCCGTCGGGGAGGCGTATGGTCTTCCCTCGC
TGGACGATTGCGTCAACGGGGACGTCGCTGAAGAATGCGTCGTTCCCTTCGTTGGGTGCG
CCGCGTCTCTTGAGCTTGGCGCGATGGCGTCTCACGGTCGCCTCGGAGGTGTTGAACTTG
TCCGCGAGCTCAACGTTTGTGAGTCGCTGGTCCTCCGGGAGGAGGTCGTTCTCGATGATC
GCTTCATCAAGGGGGGTCATGCGTGTTGTGTCTTTCTGTCCTTGGGTGCGACTGCGGCCC
AGGGGGAACTTTTGGTCAACCCC
>seq 9
CCGGCCGCCATGGCGGCCGCGGAGAATTCGATTTGGAAGTTGGACAATCACATCCAATTT
CTTACTGCTGAGATCATCTTCAATCCCATCCATCATCCGAAGAGCTTTAGCAAGTCGTTG
AAACATGCTATTGGGCTCATTCATGATGTCATAGAGTGGGTTTTCAATGATCGCAACAGT
ACTCTTTGGCAAAGTGATTTGTTTATGTTGTCCGTCACGCTCGTCATACAAGTCAACCAT
TACGTGCTGAGGATACCAGTTGACGATTTCCCCGACACGTAAGGACTTTATGTCGAAACC
GCCAGATTTCAATGGCGAAATATCGGTCTCGACAGCAACCACAGCCGCAACACCCTTCTT
AAACAGTGTGATAGCTAGATCCTGTTTGAAAGCACGTGCGGCTTGGTCTGTGTTTGCTTC
GATGTTCAGGCAGTTCTGAAGTCCAGATGGTATGGTCTCTAGATAACCGCCGTTCTCATC
TGTACGAGCGTGTA
>seq 10
CACCTTCATGATGAAATCGCATGCTGGGCTTCGTCCCCGGGTGGAATTCCTTCCTCAGGT
TCCGCTCTATGAGTTTCCGGATGGATTTCTCGTTCTGCATGACCTCCCGACGGATCAAGG
GCTTCTCCTTGACGGGGAGATATCCGTCCTCAACTGCGGTAAGGCCCTGGTGGGCGGTGC
ACTCGGTGAGACGTCGACGGAGGGCGGATTTACCTGCGTAGACCTTGCATTCGACTCCGG
GCCGCACCAGCCGTGAAAGTACGGAAGGATCCCCCGAATATCGATAGTGGATGTGGATTC
CACCCCCCGATCGGCTGAGTTCAGCATAGGAGGGAACCCACCTGCGAGCCTCTTCCAGAC
ACTTGTCTCTGTCCTTGTCGAGGTCGATGTCGATGACGACGTCTTGCTCGGGTACGAGGA
CATAATGCTCCTTTCTAGTGTCCAAGTCCTTCAGTGTTGTCGTGACGTCGTCCCAACGTT
TTGCTGGGAGTCCGTTTTCATTGGCGTACTGTGCCGGACGGTCCTTGTAGAGCTCGTCGA
GATATGACGGCTGCTCTTTCATCT
>seq 11
TTCCAAGAAATGACGCAAGGTAAAGCAACCATCTGTCCTAGGTACTTCCTACGAGATGGG
TACTTACTTTGGAACACTTACCGAGAGTGAAGTTATTGATAAGAAGAATGACGCAWCTTA
TAGTCCGTTAGCATTTGATAGAGAATATCGTTCTATTTTTACTGGATCCTCTGAAGGGTC
GTTAGTAAGTGCTGATGAATTGTCTAAGACACGAACATTGACTAGACCTTTCACCAAAGC
AGATGATAAAGATATTAAGAATCCAAACGTACACTATGTACTATCCTATGACGTGGCGAG
AGCGAGTGGTAAATCAAGTGCCAACTCTGCTTTGGTCGTTGTTCGGATTGAAGATAGGGG
TAACGGAACTTACACGAAACAACTTGTGAACATTTTCACACAAGAGGGTGTCCACTTCGA
GAACCAAGCAAAATTCTTGAAGCGAAAAGTTGTGGAGTACAACGCAAGAATACTCTGTAT
TGATATTAACGGTATGGGTTGGGGGTTAGTTGAC
>seq 13
TATGGGAGGGCCGCTATAACGCGCTCCTCGAAGAGGCTGAGGCCGCCCGCCCGCCCGAGA
TCGACGGCGGCGGGCACGACCTCCAAGAAGGGTCGATCGCTATCGACGCAGACGGACTCC
CCTGGGGGAGGACGCTGTGGGGGTGGAGTGTCCTGCAACCGGATTTCCGCAGCAGAACGA
AACTACTCGCCCCAGAGAACGGTCCCTACACCATCGTCTACACCTCAAGGAAGGATTCAT
GAAACCTAAGTTCTGGCAGCGGCTCGGGAGCGTTGTCGGCCTCACCCGAGCATGGCGCGC
AATCATCGGAGGCTGACATGGGCAACAAGCACACCCCATACCGCGACCGCGCGGACGCCG
CCCTCGAAGCGCTCAACGAGAACATGTGGAGAGCAACTGACCACGCGCGATACCCGACAG
CCGCGCTTGTAGCCCTCACCCAGGCCGTCCTCGAAGCCAGTGAACAACTCCGAATCGCAA

ACCGGCTCACCATCGCCATCGCCGCAAAAGACCCCATCGACATCACCCCCGAAACCGCCA
CCACCCTCGGCATCAAGACTCAGGAGTAATCATGAACGGTTGGCCCGACAAGCCCAGCGC
CCTCGACAGGGCTGTGTCTAAGGTCAAGGACCTTGGCGAAACGCCCCTTACCCTCAAAGG
CTTCCCTCAGGAGTACGACCTGGCACTATTGCTGGAGTCGCTAGCCGCTCTTCAGGTAGC
CGTCCACAAGGGGCCGACACTCGCGAAGATCGTCCTCGTGTGCACCACGTGGGTGGACTT
CTTGTCCTTGGGCGGTGAATCACTCGATCATGTCAGGGAGCTCGCCGAATCTGACGCCGA
TCGGGCAGAGTTCACTGACATGGTGCGCATTGCTGGGGATA
>seq 14
CTCTGTGTGGTGTGCAAATGTGTCATGTGTATGTGTGTACATGTGCATGTGTGTTGTGTG
TGCATGTGTGAATGTGTGTGATGTGTGTGCGTGTTGTGTGTGCATATGTGTATGACTTTT
TGCATGGTGTGTCTGGTGTGTGCATGTGTGTGTGCATGTGTGTGTATGACTGTGAGGGGA
AAGGGGAGGACAGGGGAGGACACTGGGATGGTGTGTGTGTGAATGCGTGTGCTGTGTGAA
TGTGTGTTGTATGTGTGTGTGTGTGTGACTGTTGTGTGTGGTGTATGTGTATGTTGTGTA
TGTGTGTGACTGTGAGTGTGTGTTGTATGTGTGTGTTATGTGGT
>seq 15
TACAGGAGAGGGGGGTTATGCCTTACGCCGGAACCAGGCATTTATGATGACGTGGCATTG
CTGGATGTGGCTTCGATGCACCCTACGTCGATCGAAGAATTAAACTTATTCGGTCCATAC
ACAAAGGTGTTCAGCCAAATCAAGCAAGCACGTCTTGCAATTAAGCATGGGGATTTGGAA
GCAGCTCGGCAGATGCTGGGCGGCAAGTTGCGACCATTTTTGGAGGGTGGTAAGGACGAG
CTCGACAACCTGTCATATTCACTGAAGATTATTATCAACAGCGTCTATGGCTTAACGTCA
GCGAAATTCGACAACCCGTTCCGAGATATTCGTAACGTCGACAACATCGTAGCAAAGCGT
GGGGCGCTGTTTATGATCGACTTGAAGCATTTTGTGCAAGAACAGGGTTTCACGGTCGCG
CATATTAAGACAGACTCCATCAAGATTCCCAACGCCACTCCTGAGATTATTCAGGCTGTG
ACGGAGTTCGGTAAGAAATATGGCTATGAGTTCGAGCACGAAGACACGTACAAACGCATG
TGTCTTGTGAACGATGCTGTGTACGTCGCATATTCTACCGCGAAAGGAAAGTCGGTATAG
ACGAAATCGTCTTGAGGGTTCTTCTCGTCTCCGAAAATGATCTTAGCGACGTGCTTACGA
GTGACGTCATTAACAGGAAGTCCTGACAATTCCGACAGTAATAACCGTGCTTCGTAGTCA
GCCTTCCGGTCATGGAAGACTGCTTTCGGTTGCACGGACGTCGTTAGAACAATACTCAAT
CACCCTTGG
>seq 16
TGGAATTACCCCTTTGTCCGAGGGTGCTCCGGTTATCCAAGTGTCTGACGTCCTCTCTGG
AACCGGGAACGTTGACAGACCGCAGCAAGATGGTACTCGTGATATTCTGATGCAAGGCGA
TGATACACACCCCACGAAGATTGGGTCGGTCGCTTTCGGGTCAGCTCTAGCATACCGGAT
TGCTACTGCGGTCCAGAAGCTCGAGCCGTGGATGCAGGCAAAGGGCTCTGTTATCGCCGC
TCCCGAACCTGCTACTCCCCCGACTCCGGCTCCCGCTGGGCTCCCGATCATGGCATGGTT
GCAGAAGGGTTGGGACGACCCGAATAGGATCGCCTATTCATTTGATGACATCAAAGCGAT
CGCAGACAGTGGCATCGGCCAAGTCGCTCTGCCCATTCAAGCTACTGCTGATGCAGCAGA
CGCCGCAGTTGCCATCCCATCAAACTACAAGGAAGGCCGTAGCTTCTCGGATTATGGTCT
CGAGACGATCCGTAATGATGGAGTCAATGTTGCAGGGATGATTGCGGCACTCAATGTCCT
TGAGGCAAAGAATATTGCGGTCCTTCCCAATATTCGAAATGGGCTAATAGACTCGGGCGC
CCAATGGTATAAGTCATCCGATGGCAAGCTTCTACCTATTCTTTCTAGTCGTAGAAGTCT
ATACTTCACAATCCATGGTAGGGCTCAGAATAAGCTCCGTGAGATCATGAAGACGGATTA
TTCCGGATTCAAGCGAGTGGCAGACAGCACTGATGGCCCTGCTGACTGGCAGATCTCCGC
AGTTAAATACGCTAACCTCGGGCGTTCTCCCATCAGGAATGAATGGAGAAGCCTGGCGAC
AGGCCAAGAATGTCTATCCCCGAAGGGGTTTGGGG
>seq 17
GACATTGGAAGGCTCGTGCAGGGGACACTTGCCGTAACAGTCCTTATTGCGGCGCTTGCT
GTGGCTACCAATGTCGCAGGACGAACTCGAGGAAAGGGTGCCGGAGTCATCCTAGCCATG
TCCATCGCAATTATGGCTCTTGTCGGGGCTGTATACTTGCTTGGAAGCATGGACTTGGTG
AAACTTGCCCAGGGAATGATTGCTCTTGCAGCAGGACTTGCGATTATGGTATTTTCTATG
GCCGCTGCAGACCAGTTCCTGGAAGGGGCTCTGGCTCTTGCTATTGCTTCTATCAGCTTG
ATGATGTTCGCAGCAGCAATGGAGAGACTCAGCAAGCTGAGCTGGGGCGAGGTTGCGATC

GGCCTTGTCGCTTTGGCTGGTGGACTCTTCATCCTTCTCGCGGCAGGCTTCGTAGCGGGT
CTAGTGGCAGAAGGCCTGGTCATCCTGACCATCGTTCTGCTCGCACTTGGTCTGGCACTA
CTTCCGATCTCAATCGGTATGGCGGCCTTTGCAGCAGTTCTCGGTATCTGTGCTACCACC
GGTTCAGCAGCATTCCTAGTCCTTACCGAGGGACTGAAGCAGCTGGCTGCAATCCTTCCC
CAGGTTGCGATTGATATTGCTAATGCAGTAGCCAACTTCATCATCACTCTCGGACAGAAG
GCACCGGAGATGGGTGTGGCCATGGCACTTATTATTGCGGCAATCATCCAGGCCATTAAC
GATAACATTCCTGCCGTTGTGGCAATGATATTCAACCTGATCTCGGCTATGCTCACAGA
>seq 18
TGACGGCCTATGCGTATTGACAAACTTCAGTCAATGCGTACCAAGGTGAACACCGTTGAC
GAATGGCTTTCAGGCGATGTCAAAGAGGATGTTATCGGTGCTATCGAGCAGGGTGCTTCT
AAGAACGAGGATTATTTTATCCTCGCTATCTCTTCGGAGGGTACTGTTCGTAACAGTGTT
GGAGACACCATCAAGTTGGAACTCCAAGACATTCTGAAAGGTAATTACCTAGCGCCGATA
GCAACTCGCTTCAAAATCTTCTTGCGACGAGCACTTCGTTGTGCACCCGACAACTGGGGA
CGAGCCTTCCGGTGACCCCAGCGCATGCCCTTCACACCAAAGTGGTAGAGTTCTGTGTTT
GTCATGTATCCTCCTATTCAAAACTGTCAGGATTTAGTTTGTACGCAACGTAGGCGTCCA
TGAGGGCAGCCACGTTGTCAATCTTTTCTTCATAGCGTCGTTTTAGAATCTTGCGGTTAC
CGTTTGTATCCTCCATGACGATAGCATTGCCCATACAAAATGTCATGAGCGCCTCGTCAA
AGAGTAATGCACGATCTTCCGCCAGCTTCTTCAACTCGCCGAGTGGGACGGATTCCGTTT
TGGCACCTTGAATGACCTTTTCGACTCCCCACTCGCTGTTCTCGGTCGTCCACCGTTCGA
TAAATTCACGCGCATTATATGGGTCATACCCGACAGATCGAATGTCGTACTGACACTCTT
GGATATGAGCATCCAGATCATCATAGACTTGCATCATGTCTAACACGGTGCCGTCCATGA
CAATTAATGTGCCTTCCTCGATGAACTCTTCGTACTTTAG
>seq 19
TGAGGAGGACACAAAGCAGGCGCTCGGAATCGCTCCTGACTACACCCCCTTCGACCAGGA
GCTGATTCTTCACATCAACTCTGCGATAATGATGGTGGAGCAGCTCGGCCTGCCCCCGTT
TCTACCTCTATACAGGGAAAGAGACGTGGTCTGACTATCTCGGAGCCTCGGAGTTCTCCT
TCGAGGCGGTGAAGAGTCTCATCTTCCTCAGGGTTCGCCTGATATTCGATCCTCCGCAGA
ACTCCTTCGTCACAACGGCTATCGAGAAGCAGATCGAGGAGTACAACTGGCGAATCGTGA
TGCAGAAGGAGACGATTCATGACATTTAGCGAAGAGCTGGCCCATTTCGGAATCAAGGGG
ATGAAGTGGGGGGTTCGGAAGAACCGATCCGCTTCGGGAGGAGAGTCATCCAAGTCCCAG
TCGTTCCTGTCGAAGGATGAGCCCACCACGTCCGAGTGGCACAGGTCTCGTCGGTTCGAC
CATCTCTCGAACAAGGACCTCAAGGCGCGGATCAACCGCCTGGAGATGGAGAAGAAGTAC
CGAGACTTCACCATGACGAAGTCTGAGCGCCGACGCAAGGCCATGAAGAAGGCAATCGGC
GACATCCTCGCCAACACGGTGAAGAAGCAGGCGCAGCTCGTGGTCGACAAGCAGGTCTCG
GCTCTGATGAACGGCAAGATCTCCGCGTCTCTCCCAAAGCCGCCGAAGATGCCGAAGCCA
GTCC
>seq 20
TGAGAGAAATTGGAGGTGTCGGGAGCTGGGATTCTTTATGAGGCAGACAGCAGGCGGAAT
GCAGGCACGAGGGTGGAAAGGCTGGAAGGATGGCAGCGCAGGACTGGAGGATGACTGACC
AGAGACTGAGCTTGATGATGCCAAGAGGCAATTTTTGTGGAATGAAACACAAACAGACAT
TATCGTTCAAAGTAAACGATGATATTCATCTCCAGTTCATCCAAGCCTATAATTTGATAC
CTGGCTGCAGGGTAGGGTAACCGCAGGCAGGGGATTAAACTCTGTTGACCCTGTGACCTA
GTTGACATAGGGTCATACTGACAGCTAATTAATTGTCCTGCCTAATTAGAGCAAGTTTTG
TAATTGTTTTAAAAATTTGTCTCAGGTCCCCGTGGCCTGGGGGCTGGTCACCCTGTTGTG
GGACTGTTCTGCTGGGCCAGGAGAGCAGGGTCCCTGAAGCGGAGCATGCATTGGCTGCCA
CACACCTCTCAGGACTTCCCTGCAGCCCCCACGGTAATTGGCTGTCTCCTCTTCCTGGGC
TCACCCCTCCTTGCACTGAGCTTTCTCTCCTCCTATCCCAGTCTCAGGTTCAGTATTCAC
CCCAGGGCCAGGAGTGTGGGGAAGGGGCTAGAGCTGCACATGGACCCAAGGCGGGAGGGG
CGCAGGCCTGGGCAGGTGGGCAGCTTTGGACGAGGGACCAGCAAGCAGCTGCAG
>seq 21
ATGCCGTGACCAAGAAGGAGTTTCCTGGCGGGTCGAACATCAGTCGAACTTTGAGCGCGA
TGTACGACTTGATGGCCGCTTCATCGTCGATGTTCTCAAAGATCGACCAGTTGGAGTCCT

TCTGGACTGTGGTGAGGCATTTTGGCCCCAATTGTGCGAGATCCATCCGTGCAGAGTTAA
TGTACAGCAGGATCTGGTCATCAAAGGCGTCATATCCCGGGATAATCCCGAGAGCCTTCT
TGACGTCTTCAAGAATTGTTCCCATTAGATCCTCCAGGGAGCTTGATCATTGGGTTGACG
CTCAACAACTCGTGGTGTCAACCTCGATCGGTCTCCGAAGTGTATCGCGTTGTGGGTATT
CTTGGTTGTTGTGATGAGAAACTCTGGCTCAAGGATGTCTGGATTGAATTCCTCGAGATC
TTTGGGCTGGATTGGGTTCATGTGGTGAATCAGCGGCATGTACTTGATGTCCAGACCTTC
GATTCCGAGATCACAGGCTTCATCCCGAGCCAGAACAAAGTTCCTGACCTTCTTCCACTC
GGTTGAGGAGTAGAATCTTTGGTTCAGATATCTATCGAAGCCAAACGTGGCTGTTCCGAT
CTGCCCATCGAGTGATAGATAGTCGAATCGTTCTTCGAACGTTGGCAGTCGGATCAGATC
AGAGTACGTCCGTAACATCTCCCGCTCCAGAGTATGTACGG
>seq 22
TAGAGCAGCTTGGATTTGCGCCGGTTATAAAACCTTGCGAAAATGCTCGTGTTGTGCGAA
AATATAAGTAACGAACGCCGCCCCGACGTGGGAGTGTTCCGAAAGGAATATTCATGAGTC
ACGCGGCGTTTTACTTATACACGGGAGTACGTAAATGAACGTACGAACCCTCATTGCAAA
AATACATTTGGCGTATAAAGGCAAGCAGTCCAGCAAAGCGCCAAAGCCAAGCGAACCGAA
ATATGAAGTATATTTATCACTCGCTAATGATAATCAAGACCGCTGGGCTGAAGACCCCGA
CCATGATTGGGAAAGTTTATATGCTGGAGCGCTTACACTACCTATCGTAAACGGGCAGAT
AAGCCTGCCGGAAGAAACCTGCAAAGTAATATCCTGTACATACCAAAATAAGCAAGTGCC
TATCGTGTATTTCAAGCAGCGGTATGACGTAAAATCTGGAGCGTATATCACCGGTCCTAA
GGGCGAGAAAATACTACACCTTATCAATCCGGAAGATTACACCGGATCGGCGGTTGTAGA
GACGCTTGCATACCCGGAGGAGATGAAGAGAGAGAACGATACAGTCGCGTGTGACAATCC
GCGTTGGCTTGCTCTACAAACAGCAGCAATGCTTGCGCGA
>seq 23
TGGCGAGTCGCATGCTCCGGCCGCCATGGCGGCCGCGGGAATTCGATTGGATGATGGTGG
CGGCTGTTAGCTTGGCGAGGCTGGTGAATGCGTTCTCGAGGGTTGCTGGGTTGATGCCGT
TGCGTGTTGCGCTGGAGATTAGCTGGGCTGCTGTGGCGGGGTTGGTTTCGGTGTGGTCGT
CTTCGATGAGGGAGAGC
>seq 24
CAACAACTCTAGGAGGAAGCAACATGCTTTGGAACCGTATTCGCTGGTACTACTACGAGC
GGGTCTTCATGGGGTACAACCCCCACTACTGTCCGTTCAACACCATCTGTCAGTGATCTG
ACAAAGCCCCCGGGTCTGTAAAAGGGCCCGGGGGTCGCGTCGGAAACGTAGGGTATTGTG
AAGACCCTACTCTGAAAGGACACATCATGAACCCCGTTACTAAGCTCATCGTTCGCGGAG
TCCTCGAGACCTGCTCTGGAATGGTCATCACTCGCGCTCTCAAGCCCATTATCCAATCGG
CGAGCGGCCTTACAAAGGTTGCTCTCTGGGTTGGTGTATTCGGCCTGAGTTCTGCAGGTG
GTGCCCTCGCCAGCAAGGTCGTGATCGACTCTGTTGAAGATGGTCTTAAGATCGGCGACG
CGATTGTCGACGAAGAAGACTGATCTCTAGTTTATACCCCACTAACCTGGGGTATAGGCT
TTTCTGAAAGGAGCACACATGGGAAAGCTTGTACCGCACGAGGACCGTCTCACGATTGGA
GGCGAGTTTCTCAGTCTTAAGTATTGCTTCGAAGCATTTCGCCGAAGCATTGAGTGCGCA
GAGTCAAGCGTCTACGAGGACATACTGATCATCTGCAACTCAATTGACATCATCGAGCAC
CAGTACTCGAATGGGAACAGCGTGCTTGTGACGTATGATGATGTTCACAAGGTCATGATC
ATGCGGATCTTCATCAATGAGGGAGACCAGGTGATCAAGCCCATTTACATCTACACCACA
GGGAGTACCAGATCGCCTGTAACTTCATGCGCCAGGTCCTGTCTGCAAATCTCGAGCTGA
AAGAGGAGTGGCTCGTATGACTCAGCCTAGGCGAGAGATTGGTGGGAAAGTTCACGGAGA
ACTTCGAGTATACTGCTAACGGTAAA
>seq 25
TATGGGTGTAGCGCAACTATCGTTAGCTTGAAGGAAGTTCTCGCTGAGCAGATCCAAGCT
AGCAAATACAGGCAGCAGTTGTGGGCTCGCGGGGGTAAAGTCTCTGCTGTGCTTCAACGG
CCTGTTGATGCACCTCGATGGACTGATGGCCAGAGGGAAGCGTTTAGGGAAGACTGGTAC
GAGAAATATACCGGTTCTGGTAAGCGTGCTGGAGGGACGCCTATCCTTGAAGATGGGATG
ACCCTTAACCGTGTGGACTTCAGTGCAACTGATCAGCAGTACATCGAGGGCGTAAAGCTT
GCTTATTCGACTGTAGCTAACGCGTTCCACGTTAATCCCACAATGGTCGGTATTCTCGAC
AATGCTAATTACAGCAATGTTCGGGAATTCCGTAAAATGCTTTATGGGGATACTCTAGGT

CCGCTTATTGCGGAAATAGAATCTACTCTTAATGCATTCCTTATTCCTATTATGGGTGGA
GCTAAAGGC
>seq 26
TAGCATCCAGAACCAATATGTTCGCCATCTTGTTTGGTGAGAATGATTTCGGTATTGTTG
GGAATAGCAATTATACCAACAACATTTTGAACTTCTTCTATGTGCGATACGACAATGACA
AGACTGTTATTCTGACATCGGATAAGGCGTGTGTCATACGGAATGGTGTTGCCGAGAATT
ATAAAATAACGACTCTGACACAATATCTCGGCAATATGTTGACAACAGCAACGACCTTTG
ATCAGTTTCTCAAGGCGCCGATTTTGTTCCAAGGCTCACAGAATACGACAGTCCAAACGT
CTGTCGTATGGGGGATTCAGAATACGGGCATCAACGATATCGTGTTGTCAAAGAACCACG
ATGCCGTTTGTGTTCCGTATGGTAACAGCTATCCATTTCTTGCCTTTGCCATTGGAGAAA
ACAGATGACAGTGAACATATTCAGTTTTGAGATGGCAACGGCATATCTCAAAGGCGCCGA
GCCAAACAAACACCTGCTCTACACCCCGCAAGTCCTCGATTCGTATTATAAGCTGTTCAC
AACCGGGTTCGAGACCGAATTGAGTGTCGTTAATTGTGTCGAGCGCGATGGTCGCTATAT
CGCCCTTGTCGATAATACTGTGTTTGTTGCTACGCCCGGTTGCATGGTCAAGACAGACAG
AACGGCTCGAGTCTTGAAAATTGACAAACGGGATAATAAGTGTGAGTTGACCCTCGACAA
GGAGATAGCACCCGGCAAGCTAATAGTCTGTGGCCTCGGATTCGTTGCTGTCGGCAAAGA
CGAAACGGCGAGAACGATTACAGTTCGGGATACCATGTTCGGCAACGAGATGATGTTTCA
AACTCAACTTTC
>seq 27
GGTAGTAGGTGGGTAGGGCAGTAGCCTTAGGGAGAACCCCCAAGGGGTTCCCGGCTAGAT
GGCCTCCCCATTGGCGAAAATCCCATTTTTGACCGCTTCCATTGCCACCCCATGCCGAGT
AGCCGTTCCGCCTCCTTGCGGAACGGCTTCCATAGGTCGCACCACGGGTCAACAGGCCAC
GTTGACTGTGGCCCCCCGGGGCCGGACATACCGGGGGATTTATAGCGAGAAGCACATTAT
TCCGGCTAGCTGTTGTTTTGTTTTTCTGTTGTGTCCTGTACAGCTAC
>seq 28
TGGAGGATGGTGACCACGTACAAAGGCAAGACACTCGGGTATATTTTTGAGCATCGTCCT
TGTGTTCCTATCATCAAGGATAAGATCTACCTCGAAAAGGAATACAAGAACGGGTGGGAG
TTCGGAGACAAAGAAGCAGTAAACGAGTATATCGACTGGACTGAAGTTTTCGATACAATA
TTCAAAGTGGCTATCAGCAACTATGGGCACCGCGAGTTAGATTTTTTGAAGTTGCCCATC
CAACCCCTTACTATTGGTAAAAACATCGGAACAACATGGTATCAGCAGGTGTACGCCAGC
TTTGATCTGGGTTCGACGATGTGGGACGCTTTCAAACGCATCATGGAAAACAATACGTCC
ACATTACGTTTCAAAGCTGCGCCACACTACGGTGTAGGATGGTGGTCGTTTGATCCTTGG
CCGTCGACTCGTATCGTGTGTTATCACTATCGTGGCCAAGACAAATCGGACACGATTATT
TTCGACACTTTTGATGATTCGCTACTAAAAGGCCAAATCAGCATTAATTTGGAAGGCGTG
AAGAACGTATCCATATTATCTAACGAAGACCAAACCTATGTAGCACAGGCATTAAACGCT
CACTATTTTTTGGCCCAGGTAAGTTTCCACATGGAAAACATGGCCAAACAGTCGGATTAC
GATTGGCATAATACAGTCATGACCCATATCTATCGTGACC
>seq 29
AAGGGCAAATCCGCAATAGCTCTTATACTTTCTGGGTCCGTCCATAATCGTATACAAATT
CTGCTTATATATATATTCAGACCGTTGGGCTTCCTGATGCGCGCGTTCAATATCCATACC
CTGTGCGATTAAATCAAGTTCGCCAATTGCATATCTAAAGCGTTCCTCGCTCATATATCC
GTCAGAAGCCTTTGCAATGACAGTAGCACTGATTACT
>seq 30
TTGGAGGTGAATATTATGAGTTGGCGCTACTTCGGATGGCTTCTCGGAGGCATGTTCAGA
TGCCATTTCGATGACGGCTGGCCCTGCTAGAAGCGAATAGGTGCCTCGGAGCCCCCCAGT
GTGACGCTGGGGGGCCCTTGCCGTTTTGACGGCGCCATGCGGAACAGGCTATACTGTGGG
CATGGAAGACAATAGACACGGCGTCGCCGATATGGCCGCCTACCGACATGATTTCACTGA
GGCCACTAGGCAGGCGCATGAGCGCGCGAGCATCGTGGTCACCGCCGACCAGCTGGCCGC
CTACCGGTACATGGCCCGCACGGCGTGGGAGCGCCTCCGCCAAGAGGGCAGGAATAACGG
CTTGGTGTGGTTCGATCCCGAGCTGCGGGTAGTGCCCGCAGCCAAGCGCCACCACGGCCC
GTACACGGCCGAGCTGACCACCGTCGCCATCGGCGTCGTCCACGAAGCGTACGTCGATGG
CGACGGCAAGAATCGGGAACGTGTGGTGCCCGACCGGGCTCAGGTCTCGATCCAGAGTCC

CAAGGAGGCCGGCCGGTTCCCGGCCCTGTGGGTCTCGGATTTCAAAACCGCCGAAGGCCC
GACCGACGCCGATGTCCACTACTGCAATCGTGTAGACTGGCACGATCCGGCTCGTCAGAA
AGGAGAAAGACGATGAGCAAAAACAAGAATCCGCGACGCGGGGAGGGCATGAACGCCCGG
GCGCCCCTGACATCTCTGACCCGCTCCCACGCCCAGCGCGGCTGGGGGGGGTTCGACCGG
TGCATGACCTACGCGATGCAGGGCCTGGCCGTGTGCCTGCCGTCGAAATCCGCGGAGGGC
CGCACCACCGCCTACCAGGTCGCCA
>seq 31
CATGGGCCCGAATTTCGCTGCTCCCGGCCCGCCTGGTCGGCTCGTGGGAAATTCGCCCTG
TAGGGGTCAGAGGATCACAACCAGTAAAAGTACAGAAAAGAATCCTTCCCCTTTCCAGGC
AGGCAACTACCCCCATTCACTGCTTGCATTCAGGCAACACTGGAGAGTAGCCCTGGCCAG
AATCCTGCAGTTACCTTCCTGTTTAGTCACTGCCCATCAAGGGTCCACAGTTGAGGAGAG
AGGAGAGAGGAGAGAGGATAGAGGAGAGAGGAGAGAGAGATTCCGCTGTTTGGAACAGAA
AGGAAAATGAGAAAAATAGACCCCAAGCTTTGGGCTTACCTCCTGACTGGGTTGCCGAAA
TATGTTACCTAGGCTGGCCTTAAATCCTGGGCTCAAGA
>seq 32
TTGGGAAATTCGGCCTTGAGAGCAAAGTCGTCCTGCAGCGCCTCGTTCATGGCAAATCGG
CGGCCCCCAGGTGTGGGTACCTCGAGGCGATGGACAACCTCGGAGCGCAACAAAGACAGC
CCAATTTGGACTGCTCGGCGGTAGAGACGACGTCGTTCAGCGACGGCCGGGGTGGCGGCC
TCGACAAGGTGCTGGACGGCCCGGGCGGTGTCCTCGTCACGCTGCAACAGGTTGAGCAGC
ATCGCCTCGGTGATTCTCATGCGAGCATGGAGGGTCTCCGGGGTGGATTCGATGAGCTTA
GTGAAGG
>seq 33
GCCGCCAAGGGCGGCCGCGGGAAATTCGATTCTCGCCGAATGCTGAAGATTCCATTGACG
AATTTTGAGTCTACATTACCCGCGATAGCTGCTATGGTACAGCCGTTTGCACCCATGCTA
TCCCAAGCTTTAGGCGGCGGGTTAGCTAGTGCAGCAGCTGCTATCACAGGACCGACAGGA
TTAATCATCGCTGGTTTGTCTGCGTTATTTGTCTTCCTGAAAGATCAAGCGGGTAACCGT
AAGATCATGAAGACATTTATGTCGCTCTGGACCGGTCTGATTGATTTCCTG
>Con 34
CAGGGCCGGGTTCGCTGGCTCCCGGCCGCCTGGCGGCTCGCGGGTAATTCGATTAAGGGC
ATCTAAGCTGGAAGCAGTAGCGCGCTGGCTTTCAAGCAGGTGAGGCATATGGCGAAGGCA
TGGCCGACGGTATTGAGAACAACAAAGACCGCATCAAGACCCTCGTCGAGTATATTATGG
ACGAGCTTACGGAATAGCAACAGAAGCTGAAGACCAAGGTTGAAAATGTAGTGAACGTTT
TCGACGGTATCAGCAAGATTAAGAACTCTGTCACGAGTCTTACTGACGCTGCGAAGGACT
TTGTGCGCGCATTTAATCGCATGCGCAATGCTTCGAGCGMTCGGTCGTTCATACSGAACC
TTGGGTATATGCTGGATTACGGTTATTCAAATGGGTCACGGCGTTGGTGGGATTATTGAC
>seq 35
TAGGTAGGGTTGGCGCTCAGTATGCTCAGTGGGCTCAGAGCGTCATGCAGCCCAAGCAGA
TCGCTCTTAACAACAAGAACGATTTTCTCAACGCCTTCAACAACGACCCGAATCTCAAGC
GAGTCATTACTCGGGGAAACAACCTGGGCAGGGTCATGACGCCCGCCCAGAAGGCTGCCA
TTCGGAACGGGACGTTCGACGGCTTGTGGCTGGGCGACTACTGGCAGTACAACGATAATT
CCTGCAAGTGGATCATTGTCGACTTCGACAGATGGCTGGACTACCCGAACGGCGAGAATC
AGCACCGAATCACGGTCATGAGCGATCGGAACCTCGGGATCGACAATATCGGCGAGTCCG
GATGGTGCGAATACGGCTGGAACGGCACCAAGATGCGACGGGACTATGCTAACGGTATGG
TGCGTTTCTCCACGCTTACCCAGGTATTCGCTATGTCGGACTTCCGGACATTCCCAGTTA
TGGAGCCGCACGGCTACGAGAACACCGGAAACGCCTGGGAACGCACGGAGAAGGACTGGA
ACTGGGAGTATCCGCAACTAACCATTCCGTCCGAGTTCGAGATGTTCGGCTCATATCTTG
TGCACAACCGCATAAACGGCGACACTCACACTATCGGCCCGATCTCTCGTCAGTTCTCAT
ATTTCCGTGTTGGCAACCCGATTCCGACCCCGGGCGAGTCCTTCTGGCTCCGGGATCAGA
TCTCTAAGGACTACTTCGGCCTGTACTACGGCGACCAGCGTCGAGTCACTTGGGCCCAGT
GGACGGAGAAGTACGGGGTTCGCCCAATCGTTTCTATCGGAGGCTAAATGTCTCATACTG
TGGAGCTGGTGATCACCATATTC
>seq 36

CGGGTCAAACTTGCGTACTATTCTAGGCATGAGTGAACCACGATGGCGGCTATCGAAAGA
ACATGGCCTAACGATTGACGGTGTGATGGTTTGTACGCCGCTTATGGTCTGCGCTGATGG
CATTATTGTTGAAAACAACCCGTCTGGTTCTTCGTATCTACGTTTGACTATCTGCATGGA
CGAACCTATCGACGTGGCCTCCGATATTCCTTTCAATATTGGCGCACTGCAACCTGGAAA
AAATCGAGAATCATTGGCTGAGCTTGAGCCCTACGGGGTTCAAGCTCATTAGTCATTTTT
GAGGAGGAAATATGGACGAGTACACAGATGAAGAACTTGCCTTGATGGCTAAGGAAGCGT
GGGACGAGGCGTTTTCCAAGATGCCCCCGGTGCCTTACTTCGAGGCTTGCTTGGACGCGA
TAGAATCCGCCGCCGAGGTAGAGAACTACCCGAACGAGCAAGTCGAAACCCTGTACTACG
AGCTGGCTTTTGCTGTTCGGAAAACGGCCCTCGAAGGGCACGGGATCGACTACCTAGACA
ACGAGCTACGGGAAATGATGGAGCGGAAAGCCTCAAAGAAAGGGTGGTTTTTCCTAAATG
ACCACCTCAAAGATGCTCAGGCTGATGCGTTGCGTCTGGAGCAGGAGCGTATGCCGCTAG
GAGAAGAAAATGCCGATGACCAAAACTAATCTTGGCTGGGAGTTCCGGCCGGCAAGAGCC
GATAGCGGTATCTACTGCGACGTTTGCGGAAGAGTGTTCGCAAAGCCCGCGCCGCCGCCG
AACCAGGCCGGCAAGCGGGTTTGCCGAGACTGCCGACAAGCAGCGAAGGAACGAAAAGCA
GGAATGCTCTTCTAACGTTTGTTGGGTTTA
>seq 37
AGAGAGAAAGAGAATTCAGATCGATCGCGTCAGGGAGTATGGAGAGATTGAAAAATACGC
TACCTTTGACGGCAACGGATACGATAGATACAATCGTCGCTTGGGTTATATTCTAGAAAC
TAGGGTTTATTTTAGCGATTTTGATATGAGCTGGAAAGATGCTACTAGCGGTATTGCTAC
TGCGAAAAATCGAGACGGCGAAGAAGTGAGATTTATTGTAGAGGATGAAAACCTACAAAA
TTCAATTACAAAATTCGTAGAGGAGTAAAAAATGCAAAATTTAGCAAGTGGTTGGCAGAC
GAGGATTTTTCCTCGTCATTGTAATAGTTCTGATGAAAAAGGGCGTTATTATGTTCGTTT
TTGGCATTATGATATGGGGCGATATTGCTATGTTTATGATGGAGATTTTGTTTCTTGGCT
TGATTCAGAATCGAAAAGGCGAATTCAACGTTGCGTTTCTACAAAAAATAGTGAGAAATA
CTATATTTGTCGTAATTTGCCCCGGGATACTCCTCCTTTGTCAAACTACGATATAGACCA
CTTTGTCATTCTTGAAAATGGCAAAACGCTAGCTGTTCCTGTTGTTGGTGATGGTGCTTT
TGAAATTTCTGATGAAGAGTATCATCAAGAGTGTTGGATAGATTGATGACGCTCAAAGAA
GTATCACAGCTTACGGGTATTCCGTATCAGACTTTGTTGGGTTGGGAAAGAGCTGGCGGT
TATCGCCGAAATTTGGCGCGATTTCTGAAAAGTTGTGATCGTGCTACGCTTGAAAAGCAT
TTTGTCGAAGCTTTATCTGACATGGTAGTTGGGAATGGGGATATTGTCGTTGTCGAAAAA
AACGGCAATTTTTCAGTAGCTACTTATCCCACAGAGGGGCCGGGTTATATCTATCCTGCA
GAAATTTA
>seq 38
TAGGGGCGGCCTTCACTGTAACCCCTTCAGGGGTGACTTCCAGGGACCACTCCCCGTATC
GGGCGCGGGTTTTCATGTCGACACCCCAGACCCAGAGGCCTCTCTTAGAGCCGACCAGCT
CCAGGCCCGCTGTTGGGCGGACGGCTGAGGCGGATAACGTAAGGTCTTTGCGTTGCTTGT
CGGTCAGGGCGGGAAAGCTCTTAGTGACCACATACGTCCCCGTCAGTCTTTGCAAGGCCC
TCGTAGGCGCGATAGAATACAAGCACTCCGTTCGGTGAGTGGGTAGGTGCGCGGCCCA
>seq 39
TACATGGAGTTCACCTCCACTACATCTACGACGGAGACCCTACCGAACTGGCGAGGCTCT
ACGACGAAGATATTGAGATCAAGGTCTTCACCGGTGATTCCTCTTTGAGAAGAAAGGTCA
CCCACTGCAACAACATCCCGGTGGCTCATATCTCAGAAGGGCTGCCACTTAAGGAGAAGA
AAGTGATCAACAAGACCACCATGGCCAACGAGAAGAAGGTCAGGGAGCTTATTGAGCGCA
ACCTTCGGAAGGAGATCCACCCCTCGACCAAGCCCTCAGTCGACTTCATCGCCAAGATCC
TCCGTGACGCAAAGGAACAGGGGATGGTGTATGATGTCAAGGACCTTAAGCCTCGGATTC
TTGCGTTTGCGATGAACTCCACTCACCAGTCTGAGGCAGCCATCAAGACTGTCATGGAGA
TGCCGTTCACCAATGAAGATCCTGACAAGAAAGTCGTCGGGTTCCCCGCTGGCGAGTTGG
TATTCTTTGACTGTGAGGTGTTTCCGAACCTGTTCCTCGTGAACTGGAAGGTGAAGGGGA
ATCCTCAGGTGCACCGGATGATTAACCCCACCCCTGAAGAGATCGAGGCCCTCTGTGAGA
TGCGACTTGTTGGCTTCAACTGTCGGAAGTATGACAACCATATTCTCTATGCTCGTACGC
TCGGGTTCAACAACGCCAAGCTGTATGACTTGAGCAAGCGGATCATCGAGAACAGCGTCA
CCGCTGGGTTCGTTTGAAGGCGTACAACCTGTCCTACACGG

>seq 40
TATTGGAAAATACAGTAAAACTAACTTTCGAAGGAAAGGAATACGAGTTCCCGCTGGTAG
TGGGAACCGAAAACGAAAAAGCAATTAACATAGAGAAATTGCGCGCCCTCACAGGGCTGA
TTACCTTAGACTCTGGCTACAAGAATACTGGTTCTTGCAAGAGTGCTATTACTTTCTTGG
ACGGAGAAAAAGGTATCCTCCGCTATCGTGGCTATAACATAGAGGATTTAGCAGCCAAGG
CAGAGTTCTTGGAAGTGGCTTACCTGCTGATCTTCGGAGAACTCCCTACGGTAGAAGAGT
ATGACAAGTTCAAGAAAACCATTCACAAGTACACCTTGGTACATGAGGAAATGCGCCAAA
TACTCAATGGTTTCCCTAAGGCTGCCCACCCTATGGGGGTACTCTCAGCCGCCACCAGTG
CGCTGACTGCCTTTAATCCTGTTCCAGTAAATGTGAAATGCGAAAAAGATGTGTATGAAG
CAGTGTGTAAGGTGATGGCCAAAGTGACCATTATCGCCACTTGGGTATATCGTAGGCGTG
AAGGATTACCACTGAACTATTACAACAATGATTTGGGTTATATCGAGAATATTCTACAGT
TATTCTTCTCTATCCCTACCGAAAAGTATGAGATCAACCCAACCGTAGTTTCAGCTCTTA
ACAAACTACTTATTCTCCATGGCGATCATGAGCAAAACTGCTCTACTTCTACCGTACGCT
TGGTAGGCTCTTCCGAAGCGGGACTCTTCGCTAGTATCTCCTCTGGGGTATCTGCCCTCT
GGGGTCGCCTACATGGCGGGGCTAACCAAGCCGTAATTGAAATGCTCGAAGAGATTCACA
AGGACGGAGGCGATGTG
>seq 41
CTTCTTTGGTGTTAATGTAGACAGCAGTATTTGCGTGCGCCCGATCACGAAGCCAGTACT
GCTGGCCGGTAGTAATCATAGAGTGGTTAAAGAAGAACAGCGGGAGCTGGCCTAGATTTA
GCCCGTTTCCCCATGGGCCGGCGCCCATAAGACCTCGCCCAAACACCATTCCCTCGTCAA
GAAGCATGATGTCGGCATTATACCAGGTGTACTCGGTTGTCTGGCCATTATCGTTTATGT
TCTTGCTAACACGAGTCCAGCACCTAAGTACATTGGCCGAGCCGAATGCAGACTGCGCCA
TTCTCAGTGCCTGAGAGAGTCCGTTCCGGTTTACATCATAATCAATATAAGATCCCTTGA
ACGGCGTAGAACCATGGATCTGGCCCGCGTATAGTCCTTTGTCCGGAATTACCACAACAT
GGTTCTGATCGACAGGAGACTGCCCAAATCCCTTGAAGTAATTGAATGCAGCAATCCTCC
AGTTTACTCCACCATATGTCCAGTAGTCTCCAATATAGAGATCCGAGAATGTTCCATTTC
GAATCCCGGCTATATACCGCGACACATTAGTACCGAGCGAGGCTCCTCGATACATAGCAT
TATGCATTCCAGCATTACCGCGATTTGCCATCTTGAACAACGAAACGGGGTCGTCAAACA
TCATGGCATTGCTCGTAATACGGGCATCAAACTGTGTTGCCTTACCTTCTAGGGCCTCGA
TCTTGGAGTTCTGTGTGGCATCACTCGCCTTGAGGTTAGCAACATCGGTCGAGGTGTTTC
CACCAGCATTAGCCAGAGCATCACGAACAGAAGCAAACCATGTATCAAATTCGCCCTTAA
GCTTTGCCTCCAGAGCAGTGAGATTAATGGTGTTCACTGGACCACTGATATA
>seq 42
CTTTAATGAGGATCAGTTCTTGTAGCGAACGTCGATGACTTGCTTCTTGTCGTCGACCTC
GAACTTGCGGATGTATCCCGAAAGACGAACTCGGTGGCCATCATCTTTCACAATCTCCAG
GCTACCGTGCTCTGGCCAACGCATCTTTCCTG
>seq 43
GGGAAATTCGCCCTTATCCGGTGGAGGCCGAGACGGTTTTCCAACCGCTACGGCTTTCCC
TCTCGATTCAAATTGCGCTGCGCCGTGTTCTCCCCTGCGGCTGGGGACACGCTTTACCCC
TATTTCGATGAGACGGGCGACATGGTCGCTTTCTCTCGCAGCTTTGCCCGCAAAGACGAT
GAGGGGAACACGGTGGACTTCTTCGAGACCTACACCGCCGATGCGCATTATATGTGGCAG
AGTGGCGACAACGGGTGGACGCCCTCCGAGGGATACCCCCGCGCTGTGGCGATTGGGAAA
ATCCCCGTGGTGTACGCACGGCAGGACGAGACGGAAACGGCGATCGTGAACTCTCTCATA
GCGCGACTCGAGACGCTGCTGTCGAACTTCGCCGATACGAACGATTACCACGCCTCCCCG
AAGCTCTTTATCACGGGGCATATTCAAGGCTTTAGCAAGAAGGGCGAGGCCGGGGCCATC
ATTGAGGGCGATGAGGGGTCGACGATGAATTACGTCTCGTGGGCGCACGCCCCCGAATCG
GTGCGTCTGGAGATCGAGACGATCCTCAAAATGATCTACACCTTGACGCAAACGCCCGAC
ATCTCTTTCGATTCCGTGAAGGGTATCGGCGCCGTTTCCGGCATCGCCCTGAAGCTGCTA
TTCATGGATGCACATCTGAAGGTGCAGGACAAACGGGAGATCTTCGACGACTATCTGCAA
CGCCGCGCCAACATCATCAAGGCTTACATCGGCCTCTTTTCACCGCCGTTGGCCACCGTG
GCCGATGAGATGGAAATCACCCCCGAGATCACTCCTTACATGCTGACGAATGAGATCGAC
GAACTGAACTATTGGCTGACGGCAAATGGCAACAAGCCCGTGGTTTCGCAGGAGGAATCG

ATCGAAAGGCGGGCGTTT
>seq 44
ATTCGATTAGCGGAGGAGTGGGTTCCGGCAAGTCTTTTACGGCTTTGGAGTATTTTGTCA
GGAATGAAAAGGGTCGTAAGCTGTATATTATCACGACTGCTAAGAAACGCGATAGTCTTG
AATGGCGGAAGGACTGCAAGACGTATGGCGTTGAGGTGGAGAAGGTGGATTCTTGGAATA
ATATTGATAAGTACAAGAACGTGAAGGGGGCGTTTTTCATCTTTGATGAGCAGCGTGTCG
TCGGTCGAGGGCTTTGGGCGAAGCGGTTTGTGAAGATTGCGAAGAATAACCATTGGATTA
TGCTGAGCGCAACTCCGGGGGATACTTGGAAGGACTATACCGCTGTTTTTGTGGCTCGGG
GCTTTGTAAAGACATTTACGGAGTTTGACCGTGAGTATTGTATTGTGACGCGGTGGGGCG
GGTTTCCTAAAATTGAGGGCTACAGGCACGTTAAACGCCTCGAACAGTGGCGGGATGATG
TTTTAGTCGATATGCCTTTTTCAAGGCGTACAACGCGCTGTGAGAGGCGAATTTGGTGTC
ATTTTGAGGTGTCTGTGTATCAGGAGGCTTTTAAGAAACGTGTGGTGCCTTGGACAGGGG
AGCCTATGAAGAATGCTGCGCAGCTTGGGTATGTTTTGAGGAGGGTTTGTGGGACTGATA
AGAGTCGTATTGACGAGTCATTCTTCAGAATGATACATTCTGGAAACCGGGCCTTGATTC
GTTTTATCAATCCTGATTGGTATGTGCTTTCCTTCATGGAATCCTCCTAGGGGAAAGTAC
AGTAAAAATTAAAGGACAGAAAAATTGCCTCTTCTATCCATATACACCCTAGTTTTTCGT
ACGAATTGTACGA
>seq 45
CCTGGCCGAGTTCCTGCTCCCGGCCGCCTGGCGGCTTTGGGAATTCGCCCTATAGGCTGA
CCCACGCACCGGGTGCATTAACCAAATTCGGGTGGATCGTGGCGGGGATACCGGTATCGC
GGGTGATTTCCTGAGCTAACCGGCCCAGGTGCAAAGGGATATTATCGGGCGTCATCTATC
CCACCGCCGGGGCAGTCCAACCGCCAAGACCAAGCAGCATGGCGGCTTGGGGGTCATGAC
GCTGCACGTAGGCGGTGCCTTCGTCGGTTAGGGCTGCTACACCACCAGGCGTAGCACGCC
GCCGCCACAGGTGGGCAGCAAGCATCACGTGGACAATGGAGCTTACCTGTTTTATTAATC
TCAAGAAAAATGGGATTTTTGATTGGTGTTTTGCTAACCGCGGTAAGGAAGTTGAGTTTG
AGTTCCGGCCGGTAGAGGGCGAGAAATCGCAAAATTTACCGGCACGGTGAAAGTCCGCCC
GTTGGGTGTTGGTGGTGAGGTGAATAAGGAGATGAGTAAAGACCTCACATTTCCGCTGGT
TGGGGAGCCGGTTTTCACACCTGAGGAACCGTAAGATTGTCCGGCCATGTGGATGTGTCC
GCCGAGGTGGAGGGGTTGAAAAATCTTCGCCGCACTATTCGGCAAGCTGGTGGCGACACG
AAGGATTTGCGCAATGCCAATCTAGCCGCTGCGCAGACCATCGTGCCTATCGCTGCTGGT
CTGGCGCCGAAGGTCACCGGCCGGTTGGCGGCGAGTATCAGGGCGGGTGCCACTCAGAAG
GCCGGCATGGTCAGGGCCGGCCGGAAATTAATACCCTACGCAAACCCTATCCCACTGGGG
TTGGCCAAAGCGCCACATCGCACCGAACCCGTGGATTGCTACCGCCGCCGCCGCCAACGA
GGAACTGTGGCTCAAAGTCTATGAGCAGCATATTGACCGTATTTTAGGAAAAGATTGAAG
GAAAAAAACGATGAAACTTGTCATTAATGTGAGGTACCCCAGCCGGTGAAGAAGGTCACC
CGTAACCGCCTTTTCTGT
>seq 46
TGGGACCGAGTTTTATGCTCCCGGCCGCCATGGCGGTTTTTTGGAATTCGCCCATTGACA
TTTTGTCTACACAGCTTTTGCCAAAGCACTCCTGAGCCGAGGCTCGGCACTGTCGGGCAG
CCGTCATAGTCAGACCAGCAGGGGCCACTGGAGCTGTGGACGCAGCCTCGGGATAGATGC
AGCGGGGCCCCACGCCTCTTCCCCTGGTTTAAATGTTTATCCAAAAAAGAAAAACGTAGC
TTGAGATTTCACTCTGTGTGGCACGGACACGGTCACACACAACACACATACACAACACAC
ACTCACATGAATCACACACACACACGCACCACACACAACACGCACACTCACAGTCACACA
CGCACACACAACACACACATACACCACACACAAGTCACACACACACTCACATGCACCACA
CACATGCACACACCACACAACATGCACACTCAGTCACACATGTACACACAACACACACAT
ACACCACACACAACACAGTCATACACACACATGCACACACACATGCACACACCAGACACA
CCATGCAAAAAGTCATACACATATGCACACACAACACGCACACACATCACACACATTCAC
ACATGCACACACAACACACATGCACATGTACACACATACACATGACACATTTGCACACCC
CACAATACACACACACACACAACACACACT
>seq 47
CTGACACACCATGCAAAAAGCCATACACATATGCACACACAACACGCACACACATCACAC
ACATTCACACATGCACACACAACACACATGCACACTCAGTCACACATGTACACACAACAC

ACACAACACGCACACACATCACACACATTCACACATGCACACACAACACACATGCACATG
TACACACATACACATGACACATTTGCACACCACACAATACACACATACACCACAGCAAAG
TCACACACACACTCACAAACTCACATGCACCACACAATCATGCACACATGCTTACACAAA
ACACATGCACATGTACAC
>seq 48
CAGCACCCCGACTCTTACTTCGTATCGGTTTACGATGCAGAGCCCGGGGACATCGTCATC
TTCAACTGGGACGGCGGTGGTACGGACCACGTCGGATTCGTGGAGAAGAACCTTGGCGGG
GGTACACTCCAGACTATTGAGGGTAATACATCGTCTGGATCTTATGGGTCGCAGTCTGCG
GGTAACGGTGTGTGGCGCCGAGTTCGCAGTTGTTGCATGGAGTACGTGATCCGCCCTGCT
TACTCCGACTCTGGCGAGTCTTCGGCTCCTTCTGGTCCCGCAGACATCCGCGCTCTCCAG
CAGGCCGTCCACGCTACACCCGACAATGTGGCCGGACCCGACACTAGGGCTCGCTGTTAT
GCTCTGGCGGCTGCTTCCGCCTGGGGCGGGCGAACCTTTCCCTTTGGCGTTCAGTTCACG
CAGTCCGTCGTTGGTACTGAGCAGGACGGAATTTGGGGTGACGCCTCGGAGGAGGCTCAC
GATAACACCGTTGAGGCCGTCCAGGCCGCATTTGGATCTGAAGTCGACGGGATCTACGGT
CCCGACACAAACACTCGAGTGAACTCAGCGCTGGATCGCGCTGAGCAGCCTTAGGAGGCA
AATAAACGG
>seq 49
TGGCGAGAAATCGGCAAAATTTACCGGCACGGTGAAAGTCCGCCCGTTGGGTGTTGGTGG
TGAGGTGAATAAGGAGATGAGTAAAGACCTCACATTTCCGCTGGTTGGGGAGCCGGTTTT
CACACCTGAGGAACCGTAAGATTGTCCGGCCATGTGGATGTGTCCGCCGAGGTGGAGGGG
TTGAAAAATCTTCGCCGCACTATTCGGCAAGCTGGTGGCGACACGAAGGATTTGCGCAAT
GCCAATCTAGCCGCTGCGCAGACCATCGTGCCTATCGCTGCTGGTCTGGCGCCGAAGGTC
ACCGGCCGGTTGGCGGCGAGTATCAGGGCGGGTGCCACTCAGAAGGCCGGCATGGTCAGG
GCCGGCCGGAAATTAATACCCTACGCAAACCCTATCCACTGGGGTTGGCCAAAGCGCCAC
ATCGCACCGAACCCGTGGATTGCTACCGCCGCCGCCGCCAACGAGGAACTGTGGCTCAAA
GTCTATGAGCAGCATATTGACCGTATTTTAGGAAAGATTGAAGGAAAGAAACGATGAAAC
TTGTCATTAATGTGAGGTACACCAGCGGTGAGGAGGTCACCGTAACGCCTATTCTGTCTG
ACCAGGTTGCTTTTGAGCGTACCGCCCGCCTTCGTGATTGGGGCACCGCAACCGACAGCC
CCTTAACCTTTGCTGCTTTCTTGGCGTGGAAGGCGCTACAGCGCACCGGCCAAACCGAAT
ACAGTTTCGA
>>seq g 50
ATTGCTGGCATGTTCCGCACTACATGAGCCTGACCCCGGTGTCTGAACCGGCAACATGGA
CAACTTCTGAGTTTGTCGAGTCCGGCACCGCCATTGTTGGCTGCAAGGACGCTACAACGT
TTTTCGAGCTCCCCGGTTCCCCACTGCGTGCCGAAGCTGAGCATATCGCCCATGGTGGCC
GAGATGTTGGCCTTTTCGGCTATACCGCGCATATGCTCAATCGTGCTGAGGGCCTGGTCA
AGGTGCATTTTAACAATGCCTAAGGTGGAAGATTCAGAGGTTCCGGCCTGGCTTGGTGTT
GATGCGGTGGGTGACGCCGCCGAACAGCAGGCGCTGAAGGGGATTGTGGCGGCGGTTAAC
GCCACTGTGACGGATTGGCATGGTAACCCAGACTCTTGGTCCGACCGGATTCATACTGGT
GCCGTGATGCTTGCTGCCCACCTGTGGCGGCGGCGTGCTACGCCTGGTGGTGTAGCAGCC
CTAACCGACGAAGGCACCGCCTACGTGCAGCGTCATGACCCCCAAGCCGCCATGCTGCTT
GGTCTTGGCGGTTGGACTGCCCCGGCGGTGGGATAGATGACGCCCGATAATATCCCTTTG
CACCTGGGCCGGTTAGCTCACGAAATCACCCGCGATACC
>seq 51
CGATTCGATGGGTCTAGCAGTATCTCGTGTGATTCTCGCAACTGCTGTATATATGCTTGG
GTCTATGAACTCGGGTAAGGCATTGCAGGGTGTTATTGCCCTATCTGTTATTATCGGAGT
TCTAGCTGGATTCATGTATGTTTCCACAAAGAATCCATACATGGACCAAGGCGCTGCGGT
ACTGCCGTT

Sequences of the Second Volunteer

> B1
AAAGGATCGACCAGATATGGGCTCAAGTGCACGTCGTAGCTCGCCGGCTGCAGGCTGTCC
TCATCGAATGGACTCACCAACGGTCTGTCGTTTCCATACTCCAGTGCGAACTCGCGAATG
TCTCCATCAGATAACATTGAACGACAGCTCCTTCCTGTCCAATATGTCCCGGTCGATCTT
GAACGCCACGACCGGCTTGTTAGCGTTCTTGTCTTCTTCGCTCGCTTGTGCGCACACGCA
CAGGACGAGGTCCGCACCGCCGCGGTTGACGAACACGTCACCCACAGTGAACGGCACACG
TTCGGGGTGGATGCGCACACACAAAGCGCCGTTCGTCACGGCGTCGTAGTCATCGTCCAA
ATGCAGGCACAACCCGCCATCCGTCGCCGACACGCTCTGCACACGAGATCCGACGAGGGC
TTCCTTCAACGCTGCGACGTACACCGGCAGCGGGTCGGTCGGCTGTGCGAGCTCGATCGT
GACGACCTTGCGGCCCTCCACATCGAACAGCCGCCACGCATCGGTTCCGTCCGGGCGCTG
CACACGCCTCATAGATCCCAGCGGGGCGGGGTCGCTGTACTCGCAATATTCGGAGATGCG
GTACTCCGAATTCTTAAACGCCGCCGCCGCCCCGTTGTCGAAGAACAAACCGACACCGTC
GAACATAAAGACAGCGCAGTCCAATGTCCACCCGGCCCTCACAAACGGGTCAAGCAGTTC
GAATGCATCGTTCGTTAGCGTAATGTCAGTCTTAGTCGTCGCCATCTTCGTCCTTGAAAA
GCAGTTCGACCGCGCTGCCCATTTCCGTCGCG
> B2
GGTAGGCGCTGGCTTGTTCGTGCTGCTCTGACACCATTTTCTGCAGGGTTGCTTGTACCC
ATAGTTGCAGTGAGGTGAGGGAGTCAACAATGATCCAGTCGAAATAGTCTGGTCGGTCTT
CAATGTCGTTGAGGTAGCCTTCCAGCTGCTCCCAGCTGGTGATTTTCGCCACATTGGTTT
GTCCACCACCACGGGCTGCGGAACCAATGTTTTCCCCGGGGAACGATAGGATGAGGTTTC
GTTTTCCGTGGTCTAGGCCTTGCCCTGCGAAGTAGGTTTTGCCTACGCCGGGTCGGCCGT
AGAGTAGCACATTGATGCTGGTGATTTTGTCGTCGACGGGTTCGACGTCATCAAATAAGC
TCACGATGTTTCCTTTCTGTTTAGTTGGGTTTATGGTTCTTCCGGTTCCGGCTCCGGTTC
ATCATGGTTGTCAGGTTCACTAGGTTTACTAGGGTCATCGTGGTCATTGCGACCATCAGA
ATTATTAGAGTTGTTATCATTTTTCAAATCGCTCCGCATTGCTTCCGCGTCTTCAGCTTG
AAGCAGTGCCATGATTTCCTCGAAGCTTAATCCTTCGGCTTCTTCGCGGGTTTTGCGTTC
AATTTCTTCAACTTCGCGTTCGATGATTTCTTCCGCAGTTGGCTTCTTCACGTCGGGAAT
AGTTTTCTCAACGTAAAATTCAGAGTAGAGAACTTCTAGCTGTTCCGCGGTGAAATATTC
AGTGGGGGGTGGTGCCTTCTAAAACGATGACACGGTAGTTGTGTG
> B3 + B4
CTATGGAGCCCGTGAGCTCGGCCGAGCCGAACAAACCGTTGCCCCGCTTCAGGAGCTCGC
GCAGGTCGAGCTGGAGGCGGCAGCACATCGAGCGGATCATGCCCGGGTCGAGCTCGGAGT
TGATGAAGTTCTGGAAGTAGGGCAGGCCGTACTTCGCGGTCATATCGAACAGGGCGCGGG
CGTTGTCCGACTCCCAGTCGAAGTCCTTCGTCATGTTGTAGGTCGGGATAGGGAAGGTGA
AGACGCGGCCGTCGGCGTCGCCTTCCATCATGACCTCGATGTAGGCGCGGTTGATCGTGT
CCATCTCGGCCTGGAGGTCCCCGTACGTGAAGTCGCACAGTTCACCGCCTATAAGCGGAT
GGTTGTCTTTGATATCCTCGGGGCACGTCCAGTCGAACGTCAGATTCGTGAAGGGGCATT
GGCTGCCCCAGCGGCTAGGGACGTTGAGGTTGAAGATGAGCTCCTGCATCGACTGCTTGA
CCTCCGCATAGTCCAACTTGTCGAGCCGGATGAACGGCGCCATGTACGTGTCGAAGGACG
AGAAGGCCTGGGCCCCCGCCCACTCGTTCTGCAGCGTGCCGAGGAAGTTGACGATCTGAC
CGCAGGCCGACCTGAAATGGCGCGGAGGATCGGAGGCGATGGCCCCCGCGATTCCGTTGA
AGCCCTCCTCCAAGAGTCTTCTGAGAGACCAGCCCGCACAATAGCCCGCGAGCATGTCGA
GGTCGTGGATATGGTAGTCGCCGTTTCTGTGTGCGGCTCCTTCTTCTTCGCTGTACACCT
TCGACAGCCAATAGTTCGCGATCGTCTTGCCGGCGGCGTTGAGAATGAGGCCGCCGACGG
AGTAGCCCTGGTTCGCGTTCGCGTTGACGCGCCAATCCGCCTGCTCCACGTACTCTTCCA
CTGTGGAGATCGGGTCGATGTTAACAGTCAAAATCTCGTCCTTTCTACGATGGGTCTTCG
ATTATACAGCCACCGGGGCCTTGATGGCTGGGTGATGCTTGTACCCGGCCGATGCGTAAA
TGTCGCTCATCTGGTAGTCGAATATCGATGGCGCCTTCTTGAGGTTGAGCTCAGGGAAAG
GGTAGGGTTTGCGCCCGAGCTGTTCTCGCACAGCAACCACGTGGTTCTTGTATATGTGGC

AGTCACCACCCGTCCAGATGAATTCGCCTACGTCATAGCCTGTCTGCTGTGCGACCATGT
GCGTTAACAGGGAATACGAAGCAATATTGAAGGGCACACCCAAGAAGAGATCGGCACTAC
GCTGGTACAACTGACATGAA
> B5
CAGCGAGTTGTAGTAGTTGGACGGCCCTGAGGATGCAGCGATGCGGAATGGCGAGTAGTC
GGACGTCACCATACCGCGCACAGTTACGATGATCTGGTTGTGGTTCTTCGGGTCGAGGCG
CACAGACAGGCTGCCGCCCTGCCCAAGCCACTGGGATGCTGTGATAGGCAATCCGTCGTT
GCCGGCAACGCAGTAGGCTGTATACTCCAACCCCGATGTGTCCTTTGCCGGGATGTAGTC
TTTGCACTGCGTCACCCATGGCGTCATGGCCTCGATCACGTAGGCGTCGAGCGTGATCGT
CTGCTCCACGGTCTTCCGGGCATCCACTTGGATGATCGTGTCTTTAGACTCCTTGCTCAA
CGGCAAGTACTCGTTGTAGGCGTAACGCATGGGTCTATACGTTGTTTTCACAGTCTTAGT
GGACTGTGCGAGATCCACGCTGTAGCTCATACCTGTCACATTGTTCATGTGTTCCTTCAG
GAAGTTATTGTCGCGAAGGAACAACAGGTTCGAGTTCTGGCGAATCATGTACACGTTATG
CACGGCGCACAGTGTATTCAGGTAGTCCCATACGTTGAACGACCCACCTGGAGCCATAAT
AATCGGATTGTACTGGTCAGACTTAATGAAGCCGTCCACGTACACTTTGTCATAGTCACA
CAGCTTGAACAGCTCGACGACTACGTTGCGGAAGTTATTGTACTGCGTGGGGATGACCTT
CACCTGCTTGAGCTTATAGCACAAGTCGTCGACGGTCACGGT
> B6
GAAGAAGGACGAAGAGGACGAAGAGGACGCCAAGAAGGCCAAGGAGGAAGAAGAGGACAA
GGCCAAGAAGGCTAAGGAGGCCATCCTTGCTCTCGCCGACTCCGACCTTCCCGAGGTTTC
CCGTGTGCGGGTCGCCGAGGCCATCGCCCGCGGATATGACGCCAAGACGATCCTGGACCG
CGAAACCAAGCTCGTCGAATCCATCCGCGAGAGCCTGTCGGGCGGCTTCGCCCCCGAGCA
CGTGCCCTCCGGTAAGAGCGCCGACGACTTCGAAGCCGAATTCGCCAAGCTGACCTGGTA
AGGAGACTACCGCATGGCACAGAATCACGTCAAGGGCGGGGACACCTACGAAGTCCAGGT
GGACGCCGCCGTCAAGTCGGGCGACGTCGTCGCCGTCGGCAAGGTCGGAGCCGTTGCGCT
CACCTCCGCCACGAATCACTAGTGAATTCGCGGCCGCCTGCAGGTCGACCATATGGGAGA
GCTCCCAACGCGTTGGATGCATAGCTTGAGTATTCTATAGTGTCACCTAAATAGCTT
> B7
TGATGGGCCCGGAGGACACGTATACGACGCACACGGGGTCGTCAACACTGGCCAGCGACG
AGGAGCAGATAAACAACCCGTTGATCAACACCCCGTCGCTGGCCATTGACAACAGCCTCC
GTACAGTGTGGGAGAAGTCGGGATCGATCCCGACGATCACGCTCACATCGCCCAACCTTG
AGAGCCGCACAACCTCGATGACGGGCAACGACTTTTTCCTTACATCGGGGTCGGCGTTCG
ACTACGGCGGGGACCGATTCATGACGACGCACGTCGATATGAACAACCAGGAGATCACGG
TGACGGCTACATCGCGAATCACTTGCGACGAATTCTCCAATAATATCGATACAGGTGTTT
CACTAGCCGATTACGAGGCCAAGATCCCGAAGACGATCTACAACGTTTTCCAATTCAACC
AACCGCACAAGGAGTACAAGCCGGAATGATACCCAACAAGAACCTCGGCGCCGGCGACAC
GTGGGGCACGTGGGTGCAGGACGAGATCTCTTCTATCAACACAGGCCTCAACAACCTAGG
GATAGGGGGGTGTGCGCAACTCCCTGAACGGGCTGATGAACAACATCGACAATACTAACA
ACAAGCTGTCATTCCGGTCCCTCACAGGGGACTTGCGTCAGCTCGGGCCCAATTCTAGCG
AGGTTCTGCTTTCAGAGAGCGTACTCAACTACTCCGAGAGCGGAAAGGGTTATCTGAATT
TCTTCTTCTTCGGAAGTGGGCGTTACGTGAACACGAATGCATCCGACGCTTTCCGTTCGA
AAATGCAACTTGTTATCCGGACTGCTTGGA
> B8
ATGGCAGATCGGCCAGAAGATCCACTGGCCGGAGGGCTACAAGCTGTGGGCGGGGTGCCG
CGGGGATGGAACGATCCAGATAAACGACACGGCGGTGACGGCGGTCAGCGAGTTCGACGG
CCGCTACGTCACGACGCAGATCCCGACGAACAACGTGAGCAATCCGTGGGGCGAGGTCCA
GATGTGGGCGAGCTCTCGTATTTCGAGTATCTGTGTGCGAGCCTACCCCGAGACACAGGT
GAAGACGGTCAACGACGTTCCCGACAACTACGGGCCGTTCCTGCCCGGCATGGGCTACGG
CGCACTCCAACAGAAGGAGCCGTATTCTATACAGGAGTACAGCGCGGCGATCGACGGCTA
CGAGGTGGCCGTGACGGCCACGTTCGTTGAGAAGGTGCTGCTGTGAGTATCGCACCCGAG
CCTTTCGAATACAGGACGGACCGCTCGCTGGAATCGTTCTCCGCACAATGGGACCGCACG
TCATACAGCGTACCGGGCGGCACCAAAGGCTACCCCGTGATGACGTTGACGGACAGGTTC

TTCAAGCCGGCGGACGTGTCGACGACGTGGACGAACAAGTACCCTGTGTCGGGGGTATAC
GAGTTCCGCGGGGATGTGCGAACGTTCACGTCCAACTATTCGACGAACACCGTGACCGTC
GACGACTTGTGCTATAAGCTCAAGCAGGTGAAGGTCATCCCCACGCAGTACAATAACTTC
CGCAACGTAGTCGTCGAGCTGTTCAAGCTGTGTGACTATGACAAAGTGTACGTGGACGGC
TTC
> B9
ATGGGCTTCCCGAAGGGGTGGGTGACCGACGCGGCCGACGTGTCGTACACAGGCAAACTG
GAGGCACTTGGGAACGCATGTACGCCCCAACAGGCCTCCGAGGCCTTCTACCATGGGCTC
CTGCAACACAAGATCAAGTCATTTACCGTGCCGTTCCCAGAAGAACTCGACCGCGTCCCG
CGCTCTCTCCTGTGCCAGGTCACGGATGCGATTACGCACAATATCATCGCCTTCACCATA
GATAAAAAGCAGTTCGGAGTGTCGGCTGCCTGCAACTGGCCCTGGCCCGTCGATACGAAC
GTTCACGTTAGCTGGGTGGCACTCGGCTGACAGCCACGCAAGCGCACAGAAGAAGCCCCC
CATCGAGCGGGGGCTTCTTCCTACTCACCTGCCTTATAGCGTCTCCACCACCGGTGGATG
TCTGTGTTCGGCGTGTACAGCCAGCTCGGGCCGATTATATTGAACAGCACGTCGATGAAC
CTGTGCGATCCGTTTCCTTGGCCGTTCCAAGGATGGGACGAGAACGGGTTGTCTGCAT
> B10
CATCGACGCTGAAAAAGATGTACTCGCACCGGACGAAGAAGCCCTGGAGAACATCGGCCA
GAAGATTATGGCTTACCAGGAAGAGCAGGAGCGACTGGCAAAAGAGGAAGCGGCACGCGT
TGACGCTATTTGCGCCAAGTTCGCTACCAACGCCAAATCACTACGCAGCCAGAAAGCCTG
CGATGAGCGAGGTGCTGAACTGAAGCAGGTTTTTGCCGAGCTACCTGAAGCCGATCAGAA
TCACGCCGAAATCAAGCTGGAAAGCATATCAGGCATCGCTAAAGCCGTATCGTGTTGGAA
GTTTCCCAGACGCTGCCTCAATGTGTCCTAATTGCGGTAAGTTCGTGAACGGCGTCAATC
CATATGACGATGGCGAGACGTGGATGGAGCAATAAATGTCAACTAAACCACTAATTTTGT
GGACATAGAGAAAGGAGATGTCAATGATTTACGAAGTAGAAGTTAGGCAAACTACTGCAG
GTAGAATTCTCATTAAGGCTAACAGCCTAGAAGAGGCAGATGAAGCTGCTAATCGATATA
TACAAGACGAATATAACCTTACCTCAATATATTTTGATGATATTTTAGACTGTGATGTTT
GGGATGTGTCAGAGGCAGATATCGTTGATAATGACGCAGAGATTATTAATGCGGAGGACG
TGCTATGACAAAACTAAAACTCAACGACGTTGTGCAGTTCAACGAAAACCACAAATGGTG
TGGTGCTTTGGGAATCGTAAATGAGGTCAAAGAACTCGAAAACGACACAAAGTACCTGAT
TGGCGTGCCGATTCCCGAAACGGCTAGTGTT

Sequences of the Third Volunteer

> C1
TGCCATTTGGGCTGACTTCGTGATCGTGAAGGCCAGCGAGGATGACGACTATGTGAACCC
CTACATGGTTTCCCAGGCTAACGCCACACTGGGGGCTTCGAAGCGGCTCGGCTTCTACCA
CTTCGCCCGCCCCGGTGACGCGGCCGCCCAAGCCCGAATGTTCGTCGCCACCGTCGGGGC
GTTCCGCAGCAAGGCCACCCTGTGGCTGGACTGGGAGGACAATGCGGTTCCGCAGGGGCC
GGGCTGGGCGAAGGCCTTCCTGGACACTGTGAAGTCCCTGACGGGCTCCACGCCGGGCAT
CTACATGAACGGCTCCGCACTCAACGGCTACGACTGGACTGCAGTGGCCGCCCAGTAATC
ATTAATCTCCAGGAGGGCGGCATGCCCCCACGGCACCTTATAGGTGTGG
> C2
AGGTAACTATTAATCGTCGCATTAATCTGTATTGTCCTCGTGAGGTCAATATTTTAGGTG
ATGACACTGAGGACACGCCGATAGTGGAAATGACTTTAGTGGACGGTGGCGGCTTCGTAT
ATGGCACACCAATGCCGCTTAATTTCCCTGTCACCAAGGCAAATGGGTCTACCATGATGC
AAAAACTCATGTCAGAAAAACTTGGAATGACGCTAGAGCCAGTTCCGCACTGATGCGGTA
GATGCCTATGATGAAATCCAAATCGCGTTTGATGGTGGGGACTGGGGTTATGCCACGCTG
GAGAAGTACTTGGGTACGAAGTCTGCCCAGGTACTGACCAAGTTCGCTAGCCAAGCAGGT

GAAACATACCGTAAGGTTACCGGGACCATAAGCCAGTTCCGCACTGATGCGGTAGATGCC
TATGATGAAATCCAAATCGCGTTTGATGGTGGGGACTGGGGTTATGCCACGCTGGAGAAG
TAC
> C3
TAACATCAGTCGGCGGGTCGCAGCCAATCGACCTCCGGGACCCCTACTTTGGCCTTCCTT
ACATCATTAGGGTGTCCTGATGCCTGGGCCTACTAAGCCATTCCTCTCGCCTGAGGGGGC
GCGGGGTGGCCAGTACGTAACCGTCCCGGCGTTCGCCTCCCCCGGGCACTCGTCCCCGTC
GAACACGCGGGACGCCCCCGGGTCGACGATTGTCTACTCCCCGAAGGGGTGGCGGTGGGA
GGAGGCCGGGGATGACTACTCCAAGACGGTCTCTAAGCTCACTGCCGCCACGATGGAGTC
CGCTGTTCGTCGCATCAAGACGTCCATGGGTGAGGTGTTCTACATTCGTGGCACCTCTGA
CACTGTGCCGCCTTTTGGGGGATCCTCGGTGGGGGATACGTGTCGAGTGCAGGATGCCCA
GACTCTCGACATTGTTGCTGAGTGGAAGTGGGATGGTGCCTCCTGGGAGCGTATGCGGGT
CACGAGCGAGCAGATCAGCAACCTCGATGTGGGTAAGCTGACCGCCGGCGCCGCGAACAT
TGCCGAGATCACGGCCCGGAAGATCGCCTCCGACGTCGGCCGCTTCCTGGAGATCACCAC
CGACCAGCTCACCGTCACCGGGAACGCCT
> C4
TGTCAGTGATCGGGGCGCTCCCACCCGACAGTCCGGAGAGGTCAGTGAGATTCACCTCCA
CCGGCTCCACGTCCTTCCACACGCCCTCAACAGCCTCGAAGACGACGGTAGTGTCGATCG
CCCACTCCCCGTACCGCCACGTCGGCTGAGACACGCTCACGAGCCGCACACGGGCCTCTC
TGGGGCTAACCCCGGCGGGCCTGTGCTGGAGTGCGCTCAAGCCGCCAACGCCCCTCAGGC
GAGCCATGAGGGCGTGCCAGTTAGTGTCCAGGGAGGCTCGTGAATCACCCTCAACCATGA
ACGCCACCGTCACCTTGAACGTCCCGAACTTAGAGGCAGCCCCATCAATGACGCCACTCC
GGGAGGGGACCTCAGTCGACGTAAGGCGCGGCTCCGGGACGGCAGGCAGGAGGGTTCCCA
GCATGACCCTCCACTTACCGGGCTGGTCCAGGTCTACCCCATTCAGGTGGTACTCACTAC
TCATGCTCTAATCCTAGATACTCGACGCCAGCCTGATAGCGTCCGCCACGTCGTCGCGGA
TCTTCGAGTCCCGCTGCGCCTGCGGGTAGTTATTGGTGATGTTCACGGTAGTGCCACCGG
ACACCCTGCCGCCCTGC
> C5
TACCGGAGGGATACTTGGAAGGACGGGACCTACCACTGGTCCCTTCGGTACTGCCCAGAC
TGCTGGCTCATCCTCGACGAGGTACAAGCCAGCACACACCGCACCTACCCACTCACCGAA
CGGAGTGCTTGTATTCTATCGCGCCTACGAGGGCCTTGCAAAGACTGACGGGGACGTATG
TGGTCCTGTAACCTCGTCTCTTCCATCGCCATGTGAGGTATTTGGCGTAGGGGCGCCATG
TGCACCGGGCTGCAATGAACTTGCCTTGCACGACATACGGTTCTTCGGTGCTCACTTGCT
CCTCCTGCATGCTCCGCATAGGGCAGCCTCTTCTCCAACCTTCCAGCCGAGGGTTCGGGC
GGTGGTTTTGATGGCGGATTCGACGGCGACCCATGGTTTTGTCCTGGGGTGGGCCTGCTC
GATGCGGGCTGTGCCGCAGCGGGCGCATGTGATGTGGGCGCGCCACTGTGGTCCGTTGGC
TTTGATGTTTACCATTCGGGTCCTCCAAGGATGCGGTAGT
> C6
TTCCCTGGCTGCCGCTTATGCCGCCGTCGACCAGGCCTTCAAGACCTACAAGAAGAAGGT
CGAGTCGAAGTTTGGTAAGGAAGCTGTTCTGGACGCCCTCGTCTCCACTGCGGAGGAGGA
CCTCACCAAGAACGAGCCCACTCTTGAGGCGATCGCGGCAGTAGATGATGTCTCGCCATA
CGGTGTCATCTTCGACAGCTCGAACCACAACTGGTCCGCAGATGAGGACCTTGCGATGCT
CCACCTCAAGTGCCAGCAGCAGTATGCAAATGATATTCTACAGACTCGTGGGCACATCTT
CCTCAACGAGGTCTACAAGATGCTCGGATTCCCGCACACTCCTGCTGGTGCCGTGACTGG
CTGGGTTAAGGGTAACGGCGATGACTTCATTGACTTCAATATCTATGACGGTATGTTCGA
GGGTGAGGACTCGAATGGGCGCACTGTCACCAAGTGGGCTCTCGACTTCAATGTCGATGG
CGTGATGTGGGACAAGATCTGAGGCGAATATGCTTGATCGAGTTCTCGCATTTGGAGCCG
GAGTTATCGCCGGCGGAGTGGGCGTATATGTCGTACTTGCTCGCAAGTTCGAGCGAGACT
TC
> C7
TAAGATGTTTTCACAGGCAAGGCTGATGCTCGTCAGATGCTGAACTACCAGAACCGAACG
GTTCGCCCAGTTCTGAAGGCGATCACTGATGCCCTCACCCGGACATTCCTCACGAAGACT

GCCCGAACGCAGAAGCAGCGGATCATGGCGATCGAGGATCCGTTCCTCAACGTCCCGCTC
GAGGAGATGTCCACGCTGGTCGACTCCGTCAAGCGCAACGAGATCGGTACAGCCAATGAG
CTTCGCCCGAAGTTCGGCTGGCCGCAATCAGAGGAAGAGACGGCAGACCAGTTGGTGAAC
TCCAACATCAATCCGGCAACCGAGATGGAGCCGATAGGTGAAGAGCCTATCGAGGAGATC
CCAGCTGCCGATGTCCCAATTTCCGAACTGATGGAGAGTAGTCAAAATGGCAGTTAAGTG
CGACTTCTCCGGCTACGCCACCAAGAACGACGTTCGGTGCTCGGATAACAAGATCATCCG
GCATGGCGCATTTGCGGCGTATGACGGGAAGACTGTACCTCTGGTCTGGCAGCACAA
> C8
TACAGACGCATTTACAGACGCCATCCGGGAGGAGTGGCCCACGCGACACCTGGTTACCAG
GATGGATCCCTGCCAGGATTTCTACGACGAGTCCGCCAGGCTGAAAATCAGGCGCTTGAT
GCGGTCAGTAGCGAAAGAACGGCGGATGCACTACCAGGTCATCAGGTCGCCGCTCGACCG
GACAGCGGGGCAGACGACCTACCTGGGCAGCCCGTCATCCAGCTACAGGATGCGGTTGTA
CGACAAAGGATGGGAACGTTTCGCCCACGTCCAGGCGCGCCTAGGACGTGCCAACGTTGA
CCCTGCATCGATGACGTTCACGACCAGCGACGGGACGCTGATACGGCCCGCCGATTGGAT
CCGGGCCGAGCTACAGGCCAGGCCAGAGGGCGAGGAGGCCAGGAGGGCGGCGGCAGCCGC
CACGCCAGAAGAGGCCTGGGGGTTCAGCGATTGGACCCACGATTTAGCCCGGCGGGCGTT
CGCCCTGGACC
> C9
TGAGGCTACTTACGACCGTATGTGTCTCGTAAACAAGGCCGTGTATGTCGACTACGAGGA
TGGGAAGTGGAGTGCTACCGGCGCCCAGTTCCAGCACCCCTACGTCTTCAAGGAGCTCTT
CTCGAAGGAGGAGCTGGATATTCGAGACGTGGCGGAGACCAAGAGCGTCACTACCGCTCT
GTATCTCAACAACGGCACAGAAGAGAAGCCAGAGATGGAGTTCGTCGGTAAGACCGGCGC
CTTTGTCCCCGTGAACCGTGGAGGCGGGATCCTTCTCCGCGAGAAAGATGGCAACTACCA
TGCCGCATCAGGCAGTACCGGTTACAGGTGGGTACAGTTCGAGTCCTTCAAGGAAGCCCA
CACAGACGACTGGAAGGAGTACGTCGCTTGGGACTACTTCGAAGGTCTTGCTGACGCTGC
AAAGGCTGCGGTGGGAGAATTCGGCGACTTCAAGGCCTTCACCCTTGGAGCTTGACGCCT
ACGATTGGAATTTCTATGTCGATGACTGAGAACGACTGGCAAGCATACTTCGAGAAGTCT
ATCGATGAACGCGATCCAGTTCTCGATGATCCCATCATCTACAAGGTGAATGAGGATCAC
TTCACTCTCA
> C10
CGCTCCCGCTGGTATTGCTCTTCACCCATGATTGACAGCTCGACTTCTCGGACCCAATCC
CACGAAATCAGATAGCCTAAGGACGGATTGGACTTTTGAACAGCTTCACGGCTTCGCCAA
TCTATCTTTTTGATATCCACGGACCATTCAAAAAATGCCAAATGCTTATTCTCCTCTGGG
CGCTCTGTAGCATCCTCACGAAGGTTTTTCAGCACGGTGGAGTAGTCGAAACCGGTTGAC
GACGTGAACCACACCTGGGCGTTTTCACGAGTAACCATGACCGGCAACAGGTCGGAAATC
AGCTCCTCCGACACTGCGAAAGCTTCGTCAATGATTACCAGGTCTCCCTGTAACCCACGC
CCTGAGGTGCGCACTCGGGATAGGAAATCCAGCCGCCGACCGTCCTTGAGGATGATTGCC
GTTTCCC
> C11
TAGACCAAGACGACACGGAGTGGGTCAAGCGCGGGCCGTGGTGGCACCTCAACGACGGCG
ACCGCAGGCTACTCGGCACTGAGCTCAAGCGACTCAGTGAGTACCTGTACGTGCTTCGCC
CGTACCGCCCGTACACGTACCCCAGATAGGGGGTTCCCAACGGAATAGGGGGTTCCCAAC
GGAATAGGGGGTTCCCAAGAAAGGAAGGAAACCATGGCAGACACCCCAACAGTCCACCAA
GCCCTGAACAAGGTCATGGAGGACGTCCAGGCAGTCAAGAAAGACAGCAAGAACCAGGCA
CAGAGATTCAACTTCAGGGGAATCGACGCGGTAATGAACGCGGTCGGCCCCGCACTACGC
AAGCACGGCGTCACCATCCTCCCAGAGGACGTGGACGTGCACCGATCAAACGGCACCACC
GCCAACGGCAAGCAGACCGCCGAGGTAGTCATCAAG

Table A: Primers were used to fill the gaps between the developed contigs of the metagenomic analysis study.

Primers' names	Primer sequence 5' to3'	Reverse Complement
O1F	GGTTCACTGCCCAGACCGGT	
O1R		GGTGTCGTACGCCACTATCC
O2F	CCCTGGTCAGAGCCGGGGAT	
O2R		CCACTAGACCACGATGGCG
O3F	CCAGCGGTGAGGAGGTCACC	
O3R		GCCGCTGCTACCCTGCAAGG
O4F	CGGATCCGCCCGCCGTTAGC	
O4R		GCCGTGCCAGTGCGTACCAG
74F	GCTTCCCTAGCCGACTCTAG	
10F	GGTTGCCGCTAGACGACGGTC	
10R		CTGACGATGATGCGGCGGGC
124R		CAGGCATGGGGAGACCTTGC
030R		GGCACGATGGTCTGCGCAGC
74F2	CGATTGCAACCACCGCGGCG	
114F	CCCGCTCGATGTTGGAAGGG	
114R		CTACATGAGCCTGACCCCGG
114F2	GGATTCGTTGCATGACACGG	
6R		CGGCGGACACATCCACATGG
267F	GGCACAGCGTTGGCTGAGGC	
43R		GGAACGTCCCATCGTCGCCG

Appendix of Chapter 4

Contigs of the A2 virus

Contig A 1510 bp

```
   1 GACTTGGAAA GCAGGGGATT TGTTTTCAAT AAACCGAGAT TCAAGGTTGG
  51 AAATTGTAAG AATCCAGTCG CGCATTATTC GATAGCTAAA TCAGGAATTG
 101 AGCCGAGGGC GTAGAGGAGT AGGAAATGGC AATTATTCGG GCAAAACGTG
 151 AACACAATTA CACGGTAATT AATAACAAGG TTTTCCAAAG AAATCAGCTT
 201 AGTTGGCAGG CCATGGGGAT GTTGAGTTAC CTGCTTTCAA AACCTGACGA
 251 TTGGTTGGTT GTTGTGAATG AATTAATCAG CGTGACAAAA GATACCGCGA
 301 AGCCAACAGG CAACAACGGC GTTTACAACA TTCTGAAAGA GCTGAAAGAA
 351 AAGGGATTTG TGCAAGTCCG CAAGAATGGG AACGGAACAA CAGATTACAT
 401 TGTTTTTGAT GAGCCTAATC AGGCTAACCC TAATCAGGCT AACCCTAATC
 451 AGGCTAACCC TAATCAGGCT AACCCTAATC AGGCTAACCC TAATCAGGCT
 501 AACCCTAATC AGGCTAACCC TAATCAGGCT AACCCTAATC AGGCTGAGAC
 551 CACACTAGTA AATACTGATA TTCAACAAGT ACTGATAGAT AGTAAAGACG
 601 GAGATGTAGG GGACGAGGTT TCTGAAACAT CTGAGTGTGC GAAACCAATC
 651 ACTGACAACA TCTTCGGCGA TTTCCAGATT ACTGACGACT GGGAGCCTAA
 701 AGACAAAAAG GCGTTTGATG GAAAACGATT CGCCGGTCAC AAATCCCAAG
 751 TCTTGACGAC AAGCGAATCA AAGACGCGCT GATTGAGTTT ACCGGCTACT
 801 GGGGAGCAAG AGGCGATACG CAAACGCAGG CGATGTGGGA ACACAAGTTT
 851 TTTCAATCAC TGACACGCCT GAAAGCCAAA GGCGAATTGG GAGCGGCAAA
 901 GCAAGACCCA TCACATAAAC GCTTCGAGAC AAGAACGGCA GACGGTATGC
 951 CGGTAGTGAT GGGCAAACAG GCAAGAGAGC TTAGACCACT AGGGAAGTTT
1001 TGAAAATGAC TGAGCAATTT GAAATTTTGG CAAGTTTGGA AGCAGAGCAG
1051 TCTGTACTGG GCGCAATCCT GATTGACAAC GATTCTGCAA ACTTCCTGAC
1101 AGACCTAAAG CCAAGTGATT TTTTCAGCAA CCAAAACGGC CTGATTTTCA
1151 AAACCGCCAT GGCGATGATT TCAGACGGCC TGCCGGTAGA TGTGATTACG
1201 CTTGATGCTG AACTTGGCAA GCGTGGATTA AGCGAAGAAA CAGGCGGCCT
1251 TGCCTACCTG ATTGACCTGC AACAAAACAC ACCGTCAGCA GCGAACGTTA
1301 GCCGATATGC ACGGCTGGTG TCAGAAAGCG CGGCAGAGCG TGAATTGCGA
1351 TTCGCTGCTG AACAAATCGA AAGACTGGCG ACAGAACGCG ATGGCCGCTC
1401 AATCGCCGAC AGACAGGCTG AAGCGGTTGC CCTGTTAGAC AAAATCAGCG
1451 GCACAGCGGC AGGCCAAAGC GAGGAAATGA GCTACGAAGA TGCAATCAGA
1501 GCAACACTA
```

Contig B 16446 bp

```
   1 TCGGTACGAT TACAGGTTTT GGCGCTGGTA AGCACCGTGA GGGCTTTGGA
  51 GGCGCCCTAA TTTTGGACGA CCTCCATAAG GCTGACGAAG CCCGAAGCGA
 101 GGTTAGACGG CAGAATGTCA TTGACTGGTT TCAAAATACG CTTGAATCTC
 151 GCAAAAACAG CATTGATACG CCTATTGTCG TGATTATGCA GAGGTTGCAT
 201 GAGAAAGACA TTGCAGGCTG GTTGCTTGAT GGTGGTAATG GCGAAGAGTG
 251 GGAACACCTT TGCCTGCCTG CTATCCAAGA CGACGGCACG GCATTGTGGC
 301 CTGAAAAGCA CGATATTGAA ACATTGCGCC AAATGGAACA AGCCGCGCCG
 351 AATGTGTTTG CCGGGCAGTA TTTACAAAAA CCTGCGCCGC CTGATGGCGG
 401 TACGTTCAAG CCTGACAATA TCCAATTTGT TAAGGCATTG CCTGCTGGTA
 451 ATATTCGATG GGTTCGTGCG TGGGACTTGG CGTCCACTGC GAACGATGGC
 501 GACTATACAG CAGGCGGTAG GCTTGGCGTA ACTGAAGATG GGCGGTACAT
 551 CATCGCCAAT ATCGTGCGCG GTCAGTATGG AGCGGACGAG CGGGATAGGA
 601 TATTGAAAAA CACGGCGCAA AAAGACGGCG TGAAAACTAA AATATCCATT
 651 CCGCAAGACC CTGGACAGGC TGGTAAATCG CAAACACTAT ATCTAACCCG
 701 TCAATTGGCG GGTTTTTCTG TATCTGCCAG CCCCGAATCG GGCGATAAGG
 751 TTACACGCGC CGAGCCGTTC GCGGCACAGG TCAACATCGG TAATGTGATG
 801 TTGCTAGATG ATGGCACATG GGACACAGCC GCGCTGATTT CAGAAATGCG
 851 GATGTTCCCA AACGGTCAGC ATGACGACCA AATCGACTGT TTAAGCCGCG
 901 CGTTTGGCGA GTTACTAGAC ACCCGAACGG GCATGATTGA TTACCTGCGT
 951 TCGCAGGTTG AGGCAAATAA ATGAGTAAAA AGACACCATT ATCACAAGGA
1001 TTTATTTCCC GTGTGGCCGC TGGTGTCCGT TACGCCTTTA CCGGCAACGC
1051 GGACGGGTGG TTTGACGCGG GCGAGCCTTT AGCCCCTGTC GCACAACAGG
1101 CAGAGGGTCG GCGGTTCGAT TATGAGCCGT TCTACAACGT AGGTCATTCC
1151 AAGCCGCGCG AACGTGAGGC AATAGGCTTC ACACAATTAC GCGCCCTTGC
1201 CGATAACTAT GATGTGTTGC GGTTGGTTAT TGAGACACGC AAAGACCAAA
1251 TGGAGTGCCT GAAATGGACA ATCCAAAAGC GCGACGTTGA ATCAACGGCA
1301 AAAAACGAAT CACAGCGCAA AGACCGAAAG GTCGATGAAG CGATCGCATT
1351 TTTCAAATCG CCCGATAAAG AACACACTTG GGCGGACTGG CTGCGCATCT
1401 TGCTGGAAGA CCTGTTCGTC ATTGACGCGC CTTGTATCTA TCCACGCAAA
1451 ACATTGGGCG GCGACTTGTA CGCCCTTGAA GTGATAGACG GCGCAACGAT
1501 TAAGCGCGTA CTGGACAATA CAGGCCGTCT GCCATTGCCG CCTGATACAG
1551 CTTATCAGCA AATCTTGCAC GGCATGGCGG CGGTTGATTA CACGGCGGAC
1601 GAATTGATTT ACCGCTCACG CAATAACCGA AGCTATAAAG TCTATGGCTA
1651 TTCGCCTGTT GAGCAAATCA TCATGACTGT GAATATCGCC CTAAAGCGGC
1701 AGCTTCACGC GCTGGAATAC TACACGGCTG GCAGTGTTCC CGATGCTTTG
1751 ATCGGCGTGC CTGAAACATG GGCGGCTGAT GATATTAAAC GATTCCAAGA
1801 GTATTGGGAT TTACTGCTGT CAGGCGAGAC GGCGGAGCGG CGCAAAATGC
1851 GTTTCGTTCC TGGTGAGTTA GCCAGAAACT TTAAAGAGAC GAAGCAGCCG
1901 CCGCTGAAAG ACGTTTATGA TGAGTGGCTG GCACGTGTCG TTTGCTTTGC
1951 GTTTAGTGTT GAGCCTACGC CGTTCGTGGC ACAGGTAAAC CGAAGCGTAG
2001 CAAAGACGAG CCGTGAACAA TCACTTTCAG ACGGCATGAG TAGTCTGAAG
2051 AACTGGGTTA AAGCCCTGAT TGATGACGTG CTTGCCCGTT ACATGGATAT
2101 GGCGGCTTAT GAGTTTGTTT GGCAGGAAGA GGAATCGCTC AATCCGAAAG
2151 AACAGGCTGA AATCTACGCC ATCTACAAAA ACGCAGGTAT CTTAACCGCT
2201 GATGAAATCC GCGCCGAACT GGGTAAAGAG CCGTTACCGG AGCAGGATAA
```

2251 TCCCGATCCG AATCAGCAAG ACGGCCAACA GCCTGAAGAA CAGCCGAATC
2301 AAGAGGCTGA AAAGCTGGGA AAGTCGGAAA GCCCGATGAG CGAAGACGAA
2351 GCCGCCGCGC TTATTGAGGC TTATTTGCTG ACGCGCGTTG ACGGCTTAGC
2401 TGAACAGATT GCCGCGCTGA TTGCTGGGGC GGTTGTTGAC TGGCAGGCCG
2451 ATGACCTGAC CACCGAACTG AATCGGGTGG CTAAAATCAT TACAGACGGT
2501 TTGGACTTTG GCGAATGGTC GGGCTTGTCC GATGTGGTCG AGCCGATAAT
2551 CAGGCGAGCG GCGGAAGACG GGGCGGTTGC CGCCTTGTTG CATGTTATGC
2601 CTGACCCTGC TGTCGGTATG GTTACGAATA TTCGCAGCCG TGCCGTCAAG
2651 TGGGCGCATG AACGCGCCGC CGAAATGGTC GGCATGAAGT GGGTGGGCGG
2701 CGAGCTTATC CAAAACCCTG CCGCTGAATG GCAAATCACA GAGGGAACGC
2751 GCGAAATGAT ACGCGCCCAA GTGGTCGAGG CTATGCGAAA CGGCGACAGT
2801 GTGCAGGAAT TAGCAGGCCG TCTGAAAGAA TCTCACGCTT TCAGTAATAC
2851 CCGCGCCCGA ACTATTGCCC GAACTGAGAC GGCGATGGCG GACGGCATGG
2901 GTAATCTGAT AGGCTGGGAA GAGACCGGGC TTGTTTCCGG CAAGCAGTGG
2951 CTGACAGCTG AAGACGATAA AGTGTCAGAG GTTTGCAATA CCAATGGGGA
3001 TATGGGCGTT ATTGGGCTAC ATGAGCATTT TGCGCATGGC TCACTGACGA
3051 TTCCAGGGCA CCCGAATTGC AGATGTACGG TTATCCCTGT TTTGGCAGAG
3101 GATATGCCTA AATCTTGATT CCTTTTGGGT AAAGTGAGTG TGTTTGCCAC
3151 CTCTTTGTGG GGCGGCTTTT TTTTTTGGAG CAACGAATGG CGAAGTTATA
3201 CGCGGAAATT GCCAAGATGG AGGCGCAGGA CGACGGAACT GTCAAAGTTT
3251 GGGGGTATGC CTCAAGTGAA GCGGTCGATT CGGACGGCGA AATTATCGCG
3301 GCAGAAGCAA TGAAAGCGGC CATTCCCGAT TATATGAAGT TTGGCGCGGT
3351 GCGTGAAATG CACGGCTCAA ACGCGGCGGG GACGGCTATT GAGATTAATG
3401 TAGAAAACGA TGGGCGCACA TTCTTTGGGG CGCATATCGT TGACCCTGTT
3451 GCCGTGACGA AAGTCAAAAC AGGCGTTTAC AAAGGCTTTT CAATCGGCGG
3501 CAGTGTTACC GCCCGCGATG AATTGAACAA GTCGCAAATC ACGGGTTTGA
3551 AGCTGACAGA AATCAGCCTT GTTGACCGCC CTGCAAATCC TGATGCGGTG
3601 TTTACTTGCT ACAAAGCCGA GAAGCCGAAA GACGAGCCGG TCACTAAATC
3651 AATGTGGCAA GTCAAATCAC TGGCTGATGT ATTGATGTCG ATGAAATGGC
3701 TGATTGAGGA CGCAGCATAC GACAACATCG ATGAAGCTGT TATCGCGCAA
3751 ATCAAAGAAT CAGCAGGGAG CCTTGCCGAA TCACTGAAAG CGTTGACAGT
3801 AAGCGAATCC GATAGGCTGG TCGATGGTTT GGCAGCCAAA GCCGATAAAT
3851 CAGACGACCT TGCCAAAGCC GAATCAGTGG ACGAACTGGC AAAAGCGCAG
3901 GACGCGCTGA AAAAATCGAA TGATGCACTT GCTAAAGCAC AGGCGGAAAT
3951 CGAAAGCCTG AAGAAACAGG CAGCGCCGCC GAAAGGTAGT ACGAAAGCTA
4001 TCAGCAAGGC AGAAGATAAC GGCGAAGACC CTTTAAATGG TTTTCAGCCG
4051 ATTGTAAAGA ATGACGGTTC GCTTGATGAC GTGGCAACAC TCGTCAAGGC
4101 AGCGCAAACA GGCCGTCTGT AACACCGCTT ACAGGCGGGT TTTTTATTAT
4151 CAGGAGCGAT AAATGAACGT GAATCAACTC ACACAAGAAA CAATTGAGCT
4201 GATGAAGTCA GCACAAGCGA ACGGTGAGCC GTTGAATAAA GGTTTTACTC
4251 AGCCGACCAG CTTTACTACT GGTTTGCAAA CCTATGACCT GTCCGCGCCG
4301 TCCCAAAAAC TCTATCCGGT ATTGACCCCG TTGCGTAACC GTATCCCACG
4351 CGTGGGCGGC GGTCGCGCCA TCGGCTCAAA CTGGAAGGCC ATCACAAATA
4401 TCAACGTAGG TAATCAACGC GCCGGTATCA GCGAAGGTAA ACGCGGTGGT
4451 GTTATCAATC ACGAAATGGT TGAACGTAAC GCGCAATTCC GCGCCATCGG
4501 CTTGGAAAAC CAAGTAACCT TTGAAGCTGA CTATGCCGCG CGTGGCTTCG
4551 AGGACGTGAA AGCGTTGGCG GTTGCCCAAA CCTTGCAAGC CACTATGATT

4601 GCTGAAGAAA TGATTTTGTT GGGTGGTAAC ACCAGCCTGA AATCAGGCGT
4651 TACACCTACT CCGACCGCTG TTGTTTCAGC TGGCGCGGGT AAAATCAGCA
4701 GCAGCACCTT GTCTGTAATC TGCGTGGCTT TGGGCTTGCA TGGCAGCAAC
4751 GCTGAATCTT CCAATCAACA ACAGCTTAGG CTTTGCGTAT ATCGTCCCTT
4801 TTCAAAATCG AAAAGAGAAC GTAACAGAAG CGCAGTTTCA GCTTGGTTAT
4851 AAAGGTTTCA TCCAGCTTGC ACAGCGAAGC GGACAGTTTA AGCGAATCAA
4901 CGCCTGCCCT GTTTATGACA CAGATGTAGA AGAAGATGTT TACCAACGCT
4951 TGACATCTCT CATCCCACGC AAACCAAGCG GACAAATCAT CGGCTATATC
5001 GCCTATTTTC AGCTTTTAAA CGGCTATGAA GCGAATCTGA CAATGACGAT
5051 GGAAGAACTG GAAGCACATG CCAAACGATA CAGCCAAACG TATAAGCGAG
5101 GCTTTGGCGT ATGGGCTGAC AACTTCGAGG CAATGGCGAA GAAAACAGTT
5151 ATCAAGCTGT TGCTTTCCCA GCAGGCACCA CTGTCAATCG AAATGCAAAA
5201 GGCGGTTTTA GCCGACCAGG CAATCGTGAA AGACGTGGAG GCAGAAGAGT
5251 TTGAATATAT CGACAACCAA CCCATGCCAG CAGAAACACC AAAAATGGCA
5301 GTTTCCGATG AAATGTTTGA GCAACTCAAA GAAAACATCA GCACCGGCGA
5351 TATTGATATT CAGACAGTCT TAGACAGTTA CGACTTGTCG GAAGAGCAGA
5401 AAGCGGAATT GGATAAATTA TGAAAATCAG ATGTTCATCA ATTCACAAAA
5451 TCATCGGCGA ACCAAAAAGC AAAGCCGAAA AAGAAGCCAA CGGATTAACA
5501 CAGACAGCCA AGTCTTACGT TATTGAACGA CTGAAAAACG AATATTCAGG
5551 CTTCGAGAGT TTCACAGGCA GTAAGGAAAC CGAAAAAGGG TTATTACTTG
5601 AAAATGAAGC AATACGATGC AGCGGCCTGA TTCGTGGCTT GATGTACAAG
5651 AAAAACACTG AACGGCGCGT CAATGATTGG ATTACAGGCG AATGCGATAT
5701 TTACGATCCG AAGCGTAAAA CAATTATTGA CACAAAATGC TCATGGGACA
5751 TCGGCACACA TCCATTCTTT CACTGGGGCT CTGCCGGTTA TGAAAAACTG
5801 GGGGCAATCA CCACAGCCGC CAAAGTGGAA ATTTTGGCAG ACGCTGAAGG
5851 CACTCAAACA GCCGCTTCTT TACCGTCTGA AGACAATTCC ACTTCTATCT
5901 TGGAATTTGA CGGCTTGCTG ACCCAAATCG CCCTGCCTGA TTCCGGCGCA
5951 TATTGGGCGG ATAATAAAGG CAGCGGCTTG ACCTCAGACG GTGCAGGCGG
6001 CGTGTATGAG TTTGAAGAAG CCTTTGCGAA CTTCTACTCT AAATATCGCT
6051 TGTCCCCTGA CACCATCTAC GTCAACGCGC GCGATTTAGC CTCTTTGACT
6101 AAGCTGATTA TCGGCAACGG CGGCGCGCCG CTGATTAAGC TGAATGTGGA
6151 CGTGAACAAC ACCGCAAACA TTAAAGCTGG TGTCGTTGTC GGTTCGTACC
6201 TGAACAAAAT CACAGGCGAC GAATTGAACA TCGTAGTACA CCCAAACCTG
6251 CCTGCCGGTA CTTACCTGTT CTATTCAAGC CGTCTGCCTG CCTACGTTCA
6301 AGGCGTGGGC AACCTGCTGC AAGTGCGTAC GCGCCAAGAG TATTACCAAA
6351 TCGAATGGCC GCTGCGTACC CGTATGTATG AATATGGCGT TTACGCTGAC
6401 GAAGTGTTGC AAGGCATGTT TATGCCTGCC TTTGGTATGA TTACCAACGT
6451 GGGTTAAGCC TAATCAGGCC GTCTGAATTT TCGGACGGCC TCTTTCTTTT
6501 GGAGATTTTG AAATGACTGA AATGGTTAAA TTACAAGCCC CTGAAGGCTT
6551 TACCGATGTT TCCTTTGGTG GCCAAAGCTA CGAAGTGGAC GAGAACGGCA
6601 TTGTTGAAGT ACCTGTTGAA TCAGCGCAAT TCCTGTATCA GTTCGGCTTT
6651 GGCAACGTGG CTGAAGAGCC TGCCGAAGCT GAAAAGGCTG AAAAAGCCGA
6701 AACCAAACGC GGACGCAAAG CCAAAACCGA ACAGCCGGTA GAACAGCCAG
6751 CCGAGCAGGC TGAAACTGTT GAAGCGGTAG AGCCTGCCGA AGCCGAACAA
6801 TCCGAAACTG AACAGCCTGC CGAAGCTGAA AAGGCTGAAT AACGATGACC
6851 GCCCTTGTCT CACTTGATTT GCTTAAGCAA CGGCTGGGTG TTACCCATGA
6901 CAAGCAGGAC GCGTATTTCA AAACCTTGCT TGATGGCGTG TCGGCGGCGG

6951 TTGAGGCTTT TATCGGTCGA AAACTTGAAG CGGCGGATTA TGTCGAGCGA
7001 TACAACGGCA ACGGCAAGAA TCGCCTTGTG CTGGAGCAAT GGCCTGTTAT
7051 TTCCGTGTCG TCTGTGAAAA TCAACGGGCG CGCGGTAGAT GACTGGGACT
7101 TTGATAACTG GCTGTTGATT CGTCATGCCT GTTTTGCGCA GGGAATCCGA
7151 AATGTCGAGG TATCGTACCG TGCAGGCTAC GAAACCATGC CTGCCGATAT
7201 TCAGGAAGCC GTCTTGATTA TCGCAACGCA ACGCTTGAAC GAAATCGAGA
7251 ACAAGGGCGT GCAGAGTAAA AGCCTTGCAG GGGAAACAAT ATCCTTTTCG
7301 AGCTTTAGCC AGTCGGGCGG TATTCCGCCG TCCGCTTACG CCATCTTGAC
7351 GGAATACAAG CGAAAGGCCG TCTGAAATGC TGAATGTTGA GTTTATCGGA
7401 GGCGACGCAA TCGCGGCTGT CTTGAAAGCT TATTCTGACG GCGTGCAGTC
7451 GGCTGTTGAG AAGTCAATCG GTCGGTCGGT TTTGAAGTTG CAACGTGAAG
7501 TCATGCAAAA CCGCCTGTCT GGGCAGGTGC TGAATGTACG GACTGGCAAT
7551 CTTCGCCGCT CACTACGTCA GCAGGTAACC AGTTCGGGCG GTTTGGTCTT
7601 TGGCGAGGTC GACACGAATG TCCGATACGG GGTGGCGCAT GAATATGGCT
7651 TTGGCGGAAA AGTCAACGTT AAAGCTTCAA TGAGGCACAT ACCTCAAGCT
7701 TTCCGCAGGC CGCTGAAATC GCCGCGTTAT GATCACATTC CCGCCCACTC
7751 TCCCAATGTG AAGCTTCCTG AACGGGCGTT TTTACGGCCG GCCTTGCGCG
7801 ATATGAAGCC GGATATTGAA GCAAATTTGC AAAAATCTAT TGAAAGGGCA
7851 TTGTGATGAA TCGTGAAGCG ATTTATTCCG CGCTGTGGGC AAAGCTTGAG
7901 GCTTTAGACG GTTTTACAAC CAAGAGCCGC AAGCTGCTGC ACTGGAACGA
7951 CGTGAAAGGC TACGACCAAC CGGCGTTATT TATGGCTCAA GGCGATATGC
8001 AGGCGGTAAC GACAACAGGG CAAGAGACGA AATGGCTGTT GCGCGTTGAT
8051 GTATATCTGT ATGTACAGAC GGCAGGCGAG CCGCCAGCGC CCATCATGAA
8101 TCCGCTGATT GACGCGGTGT GCAATGCCGT GAACGCTGTA CACCCAATCA
8151 CTGGGAAGAC GGCTTTAGCG GTCGATGGCG CGGACGTTGA GTATTGCCGC
8201 GTTGAGGGTA CGGTAGAAAC AGACGAGGGA ACGCTTGGCA ATCAGGCCGT
8251 CTGTATTATC CCGATTATGA TTTGCGCCGC TTAGTCGGCA ATTAGAAAGG
8301 AAATGTCATG CAGTTGACGT TTGGTAGCGG CGAGGTATTC GCCGAAATGA
8351 TTACGGATGC CTATGGCAAC CGTGTGCAAA ACGCAACGCC TGTGCGAATC
8401 ATGGGCTTGC AAGAAATGTC TGTTGACTTG TCGGCAGAGT TGAAAGAGTT
8451 TTACGGCCAA AACCGCTTTG CGCTGGCTGT TGCTCAAGGT AAGGTCAAAG
8501 TTTCAGGCAA ATTTAAAGGC GCGTTAATCA ACGGCCTGAC GCTGAATACT
8551 TTGTTCTTTG GTGCTGAGTT TGCAACCGGA ACAATGAAAG CCCTGTTTGC
8601 TGATACCGCT GGCAAAGCCG TGCCTGCTTC AGGTGCTTAC ACTGTTCAAG
8651 TGACTGCTCC AAATGGTGGC CGATTTGTTG AGGATGCTGG TGTGATGGGT
8701 GAGGACGGCA CGGCTTATAT CAAAGTAGCC AGCAACCCGG CAGCAGGTCA
8751 ATACACGGTT TCCAATACCG GCCTTTACAC GTTCCACGAG GGCGCTAAAG
8801 GCAAAACGGT GTTTCCAAGC TTTACCTACA CGGTATCTAT GCCGTCAGCC
8851 AAGAAAATTG AGCTGACTAA TATGGCGATG GGTAACACGC CGACCTTTAA
8901 ACTGAAATAC CTGACGCAGT TTAAAGGCAA AAAAGCCTTG TTGGAACTGG
8951 AAAGCGTAAC CAGTGGCAAA CTGGGCTTGT TCTCGACCAA AAACGATGAT
9001 TTCTCCGTGC CTGAAATTGA CTTTACTGCC TCAACCGACG AAGCAGGCTT
9051 TAAAGTCGGT ACGTTGTGGA TTCAAGAGTA ATAATGCAGG CCGTCTGAAA
9101 ATGACGGCCT TTTTCATTTA CCCCGAAAAA GGAAAACAAA ATGACCGTAC
9151 GAATTAAAGG CGTAACCGTT GAACTGAACG GCACTGAATA TGTTATTCCT
9201 CCGATCGCGT TGGGCGCATT GGAACAGTTG CAAAGCCAAA TTGGTGCATT
9251 TGACGGCAAT GTGCAAGATG CAAAACAAAT CTCTACCGTT ATCGATTGCG

```
 9301 CCTATGCCGC CATGCTCCGC AATTACCCTG ATATGACACG CGAAGAAGTG
 9351 GCTGATTTGA TTGATATTGG CAACATGAAC GAAGTATTCG CCGCTGTAAT
 9401 GGACGTTTCC GGCTTGAAAC GCAAGGAACA GGAAGCCGCG CAAGCGGGGG
 9451 AAGCTCAGGC GGCGGTTTAA GTTTCGGCGC AATGATTGCC CACGTCTGCG
 9501 CCTCAACTGG GTGGACGTGG GATTATGTTG CCGACAACTT GGATTTGCCG
 9551 CGAATCGGGC ATTTAAATGA CTATTGGCGC GAACATCCGC CTGTACATAT
 9601 CTTGGTAGCC TCATACATGG GCATTAAGCC GTCATCTAGC CCTGTACAGA
 9651 GCGAAACAGA CGAGGCAGAG GCCATCGGTA TGCTTGGCGG CGGCGAGCTG
 9701 TCAGAGGACG AATTTAACGC ATTACTGAAA GCGAAGGGGA TTATTTGATA
 9751 TGAGTAACGC AGTTTTCCCA ACGTTCCCCG GCTTGAAGTG GGGGCGTAAA
 9801 AGAACGGCTG TTTGGAGTAC CAATATTCAA AAGTCAGCTT CAGGGCGTGA
 9851 GATTCGCAGC GCGTACTATA CCTATCCGCA GTGGAAGTTT TCACTGTCGT
 9901 TTGAAGTGTT GAGAACAAAA GCCTCAATCA ATGAGCTTGA GAAGCTGGCA
 9951 GGCTTTTTCA ATGAGCGGCG TGGCAGTTTC GACAGCTTTT TGTACGAAGA
10001 CCCGGCGGAC AACAAGGTAA CAGACCAGCT TATCGGGAAT GTCGTTCAGG
10051 GTGTAACGAG ATACCAGCTT GTGCGCAATT ACGGCGGTTT TACCGAGCCT
10101 GTTTTAGCGG TTAAAGGCGT GCCGACGGTT AAAGTTGGCG GCGTTGCTTT
10151 GACACATGGC CGTGATTTCT CGATAGACAA TAACGGCGTA TTGGTTTTGA
10201 ACACACCGCA AACGCCCGGC AGACCCATCA CATGGACAGG CGGTTTTTAT
10251 TTTCGCGTCC GCTTCACGTC TGATACGGTG GATTTTGAAA ACTTTATCGG
10301 CCATCTGTGG AACGCGAAGA AAATCGAGTT TACGAGTTTG AAATTATGAA
10351 AAGTGCAAGC GCTGAATTAA TGAATCTGCT TCACAACGAA GACAGGTTTC
10401 TGATGGCCGA TTTGTTCACG ATTACTTTAT CGAACAGTCA AGTATTGCGT
10451 CATACGAATT TTGACAAGCC TGTTACATGG CAGGGGAATC AGTACGAGGC
10501 TTACAAGCTG ATTATCAAAC GCGGCGCGAC AAGAACGGCG GTAGGGCTTG
10551 ATGTTGATTC CAATACGTTG CAAATCGCCG CCGAGCCAAG TTATCGACTT
10601 GAGGGCTTAC AGTGGGCAGA GGCCGCGCTT GGCGGTGCTT TAGACGGTGC
10651 AAGGGTGGTT ATCGAGCGTG TCTTTTTCCG CGATTTCCTC ACGCCAAATC
10701 CTGAGCCTGT TGGTACGGTT ATCATCTTTT CCGGCCGCGT GTCGGATGTA
10751 TCGGGTAGCC GTTCGTCCGT CAAGGTTGAT GTTAAATCGG ATATTGAATT
10801 GCAGAACGTA TCAAGCCCAC GCAATATCTA TCAGGCCGGT TGCATGAGAA
10851 CGCTTTATGA CGGTGGTTGT AAGGTCAACC GCGAGAAATT CACAGTAAAT
10901 GGCCGCGTAA CCGCAAACAG CACGACCGGA ACAGAACTGA CTTGCAACCT
10951 GACACAGGCG AATGGGTGGT TTAATCAGGG CGTTATCAAG TTCACAAGCG
11001 GTCTTAATGT AGGGCTGACA CGCACCGTCA AGGAACATAA GGACGGCACG
11051 CTGTCTTTTG CCTTGCGCTT ACCGCACCCT CCACGCGCCG GAGATGTGTT
11101 CAAAATCTAT CCGGGCTGCG ATAAACGACA AAGCACTTGT AAAGACAAGT
11151 TTGACAACAT CGTGCATTTT CGCGGTTTTC CTTATATCCC ATCTGCTGAT
11201 ACGGTGGTTT AAATGAGGCC GTCTGAAATG GATTTGAGAA AACGAATTGT
11251 CGAAGAGGCT TATTCGTGGC TTGGAACGCC GTACCATCAT CAAGCGATGG
11301 TAAAGGGTGC TGGTGTAGAT TGCGCGATGA TTCTTGTCGC AATCTATCGG
11351 GAGGCTGGCT TGCTTCCTGC TGATTTTGAC CCACGGCCAT ATCCTCAAGA
11401 CTGGCACTTG CACCGCGACG AGGAGCGTTA TCTTGGCTGG GTTTTAAAAG
11451 TCTGCCATGA GACCGACACG CCGCAACCGG GCGACGTTGT CGTCTGGAAG
11501 TTTGGGCGCA CGTTTTCACA TGGCGCGGTT TATGTTGGCG ACAACAAAAT
11551 TATTCACAGC TACATCGGGC GCGGTGTGGT TTTGGACGAA TTGGATCAGG
11601 CCGAACTTTT AGGCCGTCCG ATGAAATTTT TTACTTTTGG GCGGCCTTTG
```

11651 GTATGAAGCG CATTTAATAG ATTGATTTTA TAGAGGTTAC TCATGGGCGG
11701 TAAGGCTTCC ACTATTTCAA ATTCTGAACA ACGGATTTTA TCCCTACAGG
11751 TCCAACAATC ATCTCAAGGC TTGACCCTGC CTGTTGTTTA CGGTCGGGCG
11801 CGTGTTGCTG GCAATTTGAT TTGGTACGGC GACTTTACCA CTATTGAGAC
11851 CAAGACAACG ACTCGACAAG GCGGTAAGGG CGGCGGTGGC GTAAAACAAG
11901 AGGATATTTC CTATACCTAC GAAGCCGCCG TCATGATGGC CTTGTGCGAG
11951 GGCGAGATTA AGGGAATCGG GCGTATTTGG CGAGACAAAG AAAAGTTTGA
12001 ATCGCTTTCC CAATTACGCC TGAATCTTGC AAAAGGCGGC GATGAGCAGC
12051 CGACTTGGAC GCATTTGCAA CAGCCGAAAC ACCAAGCGCA GGCCATCAAC
12101 TATTCCGGCA CGGCTTATAT TTACAGCCCG AATTACGAAC TGACAAAATC
12151 AGCGCAGATT TATAGCCATA ATTTCGAGGT TATCGGGAAA ATGGGGTATT
12201 CGTCCTCAAT CCCTGATGCA AATCCAAGCG AAATTATTCG CGATATGCTG
12251 ACGAATCAGA ACTACGGTTG TGGATTCCCT GCTGAAAACT TGGGCGACAC
12301 GAGCGTTTAC GGCGTTTATT GCCGCGCGGC AGGTATCTTT TTAAGCCCTG
12351 TTTACAGTGA GCAGACCGAG GCGCAACAAA ACATTTCCGA GCTGTTGGAG
12401 CAGACCAATA GCGCGGCAGT GTTTTCTCAA GGCCGTCTGA AAATTGTCCC
12451 TTATGGCGAC GTGAAGCTGT CAGGAAACGG CGCGGCCTAT GTGCCTAACC
12501 TGACACCTGT TTACGACTTA ACCGATGACG ATTTTATCGT CTCAGGCGCG
12551 GAAGACCCTT TAAAGGTTGA GCGCAAAACC AACGCTGACG CTTACAACCA
12601 AATACAGGTT GAATATCTCG ACCGTGCGAA TGACTACAAT ATCGCCGTGG
12651 CCGAAGTGAA AGACCAGGCG AATATTGAGC AATACGGCCT GCGCCCTAAA
12701 GATGCCGTGA AGATGCACGG AATCTGCGAC GCTAAAGTCG CAAACCATGT
12751 AGCGCAACTG CTTTTACAGC GTGCTTTGTA CGTCCGCAAC GAATATGAGT
12801 TTAAGCTTGG TTGGAAATAC TGCCTGCTTG AGCCTATGGA CTTGGTAACG
12851 CTGACAGACG AGGGTTTGGG GCTGGATAAA ACGCCTGTCC GAATCATTGA
12901 GATTGAGGAG GACGAAGAGG GCGTTTTGAC CGTCAAAGCC GAAGATTTCC
12951 CAATGGGCGC GGCATCGGCT ACGGCTTACC CTACGCAGCC GTCATTAGGT
13001 TATTCCGCCG ATTACAACAA ATCGCCAGGC AACGCCCATG CGCCTGTTAT
13051 CTTTGAAGCG CCTTTGCAGT TGACTGGCGG CGAGCCTCAA ATTTGGCTTG
13101 CAACCGCTGG CGGCGATATG TGGGGCGGCG CCGAAGTGTG GATTTCGACC
13151 GATGGCGACA GCTACACACG AATCGGGGCG ACCAACAAGA AAGCGCGTTT
13201 TGGCTCACTG TCCGCGCCTT TGGCAAGTGG TGCGGTTTTC GACCGTGCCA
13251 ACACTCTGAA TGTTGAAATT TACCGCCGGG CAAATGACAG GCGGAACGGA
13301 GCAGGACAGC CGCGATTTGC TGACCTTGTG TTACGTTGAC GGCGAGTTTT
13351 TGGCCTACGA GACTGCCGAG TTGAAAGGCG TGGGACGTTA TACGCTGGGC
13401 AACCTGACGC GCGGTGCGTA TGGCTCAAAC ATCGACCGAC ACAGCGCAGG
13451 CAGCCAGTTT GTGCGTATCG ATGAAGCGAT GTTCAAATAC GCCGTCCCTG
13501 CGAACTGGGT AGGACGCACG GTTTGGGTTA AGCTGGTGTC TTTCAACGTC
13551 TTTGGCAGTG GTGTTCAGGA GCTTGCAGAA GTGCCGGCTT ATTCCTACAC
13601 CATCAAGGGC GCACCGCTTG GGCAGATTCA AAATTTACGC CTCACATCAT
13651 CTTGGGCATA CGGCAAAGAA GCCGTTATTG CTTGGGATAA ATTGGGCGGT
13701 GCTGACACCT ACGATGTGGA AGTCTATGCA GGCAATACGC AAAAACGACT
13751 GCGAAGCTTG AGCGGTATTG TTGATAACGG TTTTACCTAC ACGCAAGCCG
13801 ACATGAAAGC TGACGGCGGA CAAGTGCGTG ATGTTGTCTT TAAAGTTCGT
13851 GGACGCGCGG TTACTGGGAA AACTGGCAAC TGGGCTCAAG TGGCTGCACA
13901 AAATCCTCAG CTCAAACCAT TGCAAGGTAT TGAGGTTGAC AGCGGTTTGC
13951 TTCAGGCGTT TTTCAAATGC GCTATGCCGT CTGAAGAGGA TTTCGCAGGT

14001 ATTGTTATTT GGGTGTCTGA AAATCAGGCC GTCCCAACAA CAGACGCGAA
14051 TAAAGCCTAC GATGGCGCGG AAACATTTGT TTCCATCACG AAATGCAACG
14101 GAAAGGATTT GCAACAGGGT AAAACCTATT ATTTACGCGC CGCCGGTTAT
14151 GACAGCTTCG GCAAAGACGG CATGCACGTC AGCAACAGTA TTGCTTTTAC
14201 CGTTGCCGAC GTGTCAGTCA CAGATTTAAC GGAAAGCAAT CTGAACAAGG
14251 CTTTGCGCGA CAAAATCGCA CTGATTGACG GAAACGGCGC AGGCAGTGTG
14301 AACGCACGAA TCGCAGCCGA AGCGCAAGCA CGGGCAGCGG TCACCCGCAC
14351 GGCAGAAGAC GCTAAAGCCG CAGCAAAAAA AGCCGCAGAC GACCTGACTG
14401 CAAAGGCCGC TGAACTTGGT AACAAGGTTG CAGTAGTCGA GCGAGTGAAT
14451 AACGAGCAGG CGCAGCAAAT CAGGACGGTT ACAGCAGCAC AAGGCACGAC
14501 CGCCGCAGGC TTGGAGGCCG AAAAGAAAGC ACGGGCAGAC GGCGACAGGG
14551 CGGAAGCTGC GGCGCGTGAA ACGTTGGCGG GTCGTGTATC TACGGCTGAG
14601 GGCAACATCA CGCGCGAAAC ACAAGCGCGG GTCACAGCCA TCAACGCCCA
14651 AACCGCTGCA ACGGAAGCCT TGAAAACGCG GGTCGGCAAT ACTGAAAGCA
14701 GTATCACAGC ATTGCGCGAA ACCGTTAATC AGAAAGACAG TGCGAGGTCG
14751 TCTGAAATCC AAACACTGAC CGCGAAGATT GACGATGTTT CGGTTGGTGG
14801 TCGCAACTAT GCCCTATCAA CAGGAACGCC CGGCAAAGTG CTGACAGTGA
14851 GCGGGAATAA TCAGACCAAG AATGTCAACA TCGACGTTTC GTCTGCTTTG
14901 GATCTGAAGC AAGGCGACAG CCTGATTATC TCGTGCGACA TTGAGCTGAC
14951 AAACGCTACA TCGCCATACG GTAAACCATA TCCACGAATC GGCGCGGAAT
15001 TTTCCGTGAC CTATGCCGAC AACTCAATCG GTTATTTTGC TGCATGGTAC
15051 GACGAGGCGA TAAACGGCAC GACCAAAACG CTGAAGCAGC GGATTGTCGC
15101 CAAGCACACG GTCGCCAAAG AGGTTAAGGC ACTGCGGAAC ATCATCGTTC
15151 AAGCACGATA CCAAACATCG GAAGCTATCA AGGTTTCCAA TGTGAAACTG
15201 GAACGCGGAA CGGTGGCAAC CGACTGGACG CCTGCGCCTG AAGACAATGA
15251 CGGTTTGCAG GAAATCCGCA GTACGGTTCA GGTAGTTCAG ACGACCTTAA
15301 CCAAGGCGAC AGGTGACATC AAATCGCTTG GCGAACGTAT CACAACAGTG
15351 CAAAGCAAAG CTGACGGCAA TACAGCGGCA GTGCAAGCCC ACGCCCAAAG
15401 TATCAACGGG CTTGGAGGCG CAATACACTG TCAAAGGGTT GACGTTAACG
15451 GCAAGGTAGC TGGCTACGGC TTGGCAACCA CGCCAAAAAA CGGCACGCCT
15501 GAAAGCAAGT TTATTGTAAA TGCTGACCGA TTCGGCGTTG GTTCGACTGG
15551 CAAGGCTGAC GTGTTCCCAT TTGTGGTTGA TACGCAGAAA AACCGTGTCG
15601 GCGTGAACGG CGAACTGGTG GTAAACGGTA AGGCGATTAT CGATAGATTG
15651 AACGCTGGGG ATATTCACGG CGATAAAATC ACGGCAAACT CGCTGAACGC
15701 AAACCGCCTG ACAGCAGGAA GCGTTACTGC GCGTGAAATG GCGGCTGGTA
15751 GTATTACCGC TGAAAAACTG GCAGCAAACT CTGTTACAGC CAATAAGATG
15801 AACGTCAACG AACTGTCGGC GATTTCGTCG AATTTGGGCA GTATCAACGG
15851 CGGCAGCTTG GATATCGGCA GCGGTAACTT TACCGTAACG CCAGACGGTA
15901 TGCTTGAGGC TAAAAATGCT GTCATTTATG GCAGGATTGA GGCGGAATCA
15951 GGTTATTTCA ACGGCACGGT CAAGGCATCG CACATTGAGG GCGACGTCCT
16001 GCGACTGCAT CGCATGAATA AGATTAATGC CAACACTTGG GAAGTCAAAA
16051 TTCAGGCGGA TGAAGTGCCA ACGTTGATGC GCCCTGATTT CAAAATATAC
16101 ACAACCAATG AGGAGCGTCT TCGTTATGGA TCCTTAATCG GACACCCTTC
16151 AAGTGTTTCT AGGTTTGAGT TGAATGGGGT GGTTATGCCA AAAACCACAT
16201 TTCACACATC AGAAATTTTA GGCGGAAAGC TCTTTGGCAG TGAGAGTGAA
16251 GTATTTCTCA GCCTAAGAAA AAGGCAAATA CATTCATTTA ATTGGATGAT
16301 GCTCCGACGT GATGCGGTTA ACACGATCCG TGTTACCTTA GGTGAAAATG

```
16351 AGACTTTCGA TATGGATGTT CCTGTATTGA TGACTTCATA CCTTGCACAA
16401 AGTGACCCTG AATATAAGGA GTTGATGGGT AATCACTAGT GAATTC
```

Contig C 6468 bp

```
   1 TAGGGCGGAA ACATACCGGT TGAATCGTGC GTTACGATCC GATTTAGCCG
  51 ATATAAAAGT CAGCGACCTG CGCCCACATC ATTTTGCCCA ATGGCGCGAC
 101 AATCGGAAAA AAGAAGTCCA AGAAGCCACG GTCAGGCGTG AGCTTGAAAC
 151 ACTGTCGGCC GTCTGCCAAA TAGCGGTCAA AGAATGGGGG CTTTTGCCGT
 201 CAAATCCTCT ATTGCAAATC AGACGGCCTG GGAAAGGTAA GGCGCGGAAC
 251 TACATACCGC CTGACGATAT TGTCTTGGCT GTCGTGCGAG AGCTTGGTGT
 301 AGCTGACGGC GTACCGATAA TCACAGTAAA ACAGCGTATC GGCTTGGCTG
 351 TCTTGTTTGC GATTGAGACG GCCATGCGCG CCGGGGAAAT CTGCAACATG
 401 GCTTGGCGTG ATGTGCATTT GAGCCGGCGC GTGGTACATT TGCCAATGAC
 451 AAAAAACGGT AGCAGTCGAG ACGTGCCGCT GTCTAAAAAG GCTATGGCGA
 501 TATTGGATAG ATTGCCACGC TCTGAGAGTG GGTCTGTGTT TGATATAAGC
 551 TCCCACACGC TTGACGTGAT GTTCAGACGT GCGAGGGCAA AGGTTGAGGG
 601 GGCTGAGGGA TTCCACTTTC ACGATACGCG CCATAAAGCC CTTACGCGCA
 651 TTGCGGCTAG GGTTGAGCCT ATGCAGCTTG CAAAAATCAG CGGCCATAGG
 701 GATTTACGCA TATTGTTGAA TGTGTACTAT AACCCTGATA TTGGCGAACT
 751 TGCCGATTTG CTGGATTAAA AAAAACCGCC TGTTACGGCG GTTTCTCTAT
 801 TTCTGACGGC GGCGGCGGAT AAATTCGTGT ACTTCTGCTT CCGGCCATAA
 851 GAATTTACGC GGCGAGATAA TAAAAGGCTT CGGAAAGCCT GCCTGTTTGC
 901 ACGTTTGATT AACGAATGGA GCAGTTTAAG GCGAAGCTGG CAGAAGTTAA
 951 CCAACTGTTG AATCTTGGCG CAATTGACGT AGAAACCTAC GAACGCAAGG
1001 TTCATCAACT GAACAGCGAG CTTGAGCAGA CGGACGGCAA GGCTTCGGCG
1051 GCGGCTGGTG GGTTGGGCAA AATCGGATCC GTTTTGGCAG GATTCGCCTC
1101 ACTGTCATTT GCCAAGTCCA TGCTTGACAC TGCCGACGCC ATGCAGTCAA
1151 TCAATGCACA AGTCAGACAG GTTGTGTCGT CTGAAAGCGA GTATTTGGCC
1201 GTACAACGTC AGTTACTGGA TGTAGCCAAT AGTACGCGTG CCTCATTGGA
1251 ATCAACAGCA AATCTGTACG TTTCTACAAG CCGCGCCTTG AAAGACTACG
1301 GCTACACGCA ACAGGAGATT TTGACCTTTA CCGAGGCGAC CAATAACGCT
1351 ATGGCTATTG GTGGCGTACA GGCGCAACAA CAGGCCGCCG CGCTTATGCA
1401 GTTATCGCAG GCTTTGGGTA GTGGTGTATT GCAGGGCGAT GAATTTAAAT
1451 CTATTTCCGA AGCCGCGCCG ATTCTGCTTG ATACGATTGC GGAATATATG
1501 GGCAAATCAC GAGCTGAGAT TAAAAAGCTT GGCAGTGAAG GGCAATTGAC
1551 GGCGGATGTG ATTTTTAAAG CCATATCCGG TGCGTCTGAG AAATTCGGCG
1601 AGCAGGCGGC CAAAATGCCT ATGACGATGG GTCAGGCTTT GACGGTGTTT
1651 TCAAACAACT GGCAAAGCAT GGTTTCCAAG CTGCTGAACG ACAGCGGCGC
1701 AATGTCTGGA ATTGCAGCCG TTATTAAACT TATTGCAGAT AATCTGAATT
1751 TGGTCGTTCC GATTGTCGCA GGTTTTGCCG TTGCTGTTGC GGCCGCAGTT
1801 GCGCCCACGC TGGCCTTGAA TGTGGCATTG CTGGCAAATC CGTTTGGGAT
1851 TGTGGCTGTT GCAATCGGCG CAGTTATCGG CCTTATTGCC CAATTTGGCG
1901 ATGAAATAGA CGTTTTCGGC GATGGTTGGT CGAATTTGTC TGATGTGATT
1951 CAGGCCGTCT GGCAAGTCAT CACGGAAACC ATCGGCGAGG CTGTCGATAC
2001 TGTTAAATCA TGGTTCGGCG AGTTGACGGC ATGGGTTGAT GAGAGTGTCG
```

2051 GCGGATGGTC GGCGGTATTT GACCGCGTGA TGAGCTTAAT CTCAaGCaCT
2101 ATCGGGGCGT ACGTCAACGT CTATATCAAC ACATTCGCAA CCGGCTGGAT
2151 GTTGATCAAA GAACCCGCCA ACAATATGCC GCATTTCTTT GCCAATCTTG
2201 GCAAGGTTAT TGGCAACGTG TTTATTTCCG CGATTGAGTG GATGGTAAAC
2251 AAGGCAGTCG GCATGATTAA CAGCATGATT GACTTTGCCA ACAAACCCGC
2301 GTCAAGGGTC GGCGTTTCGG GCATTGAAAA GCTGAATAAC GTCCAAATGG
2351 AACGGATGAA TGATGGCGGG CTTGGCGGTC AAATCGCTGA CAGTATGACT
2401 AAAGACCGTG CCGGAGCAAT GGCAAGCGCC ATCAAGGAAC GCGCGGCAAA
2451 CATTCACGAA GCCAAAGCAA TGAGAGGCGC ACGAGGTGGT GGCGGTGGCG
2501 GCGGTTCTGC CAAAGCTCAC GCGCCTGCCG GTGGCGGTGG TGGCTCAGGT
2551 CGTAAAGGTG GTGGTGGACG TAAAGGCGGC GGAAAGGGGC ATGCAGGCGG
2601 CTCAGGAGCG GCGCAAGACC CGATGCAAGG CTGGGAAGAG GAAATCAAAG
2651 CCCAAAAGCT TGCACACCGT GAAATGCAGC GCGAAACGCT CACGCACCAA
2701 GAATGGGATT TAGCGCGTGA GGCTGCCTAC TGGCGCGAAA AACTGGCAAC
2751 GGTTGACGCT GGTAGTAAAA CAGGCTTAAA GCTGCGTGAA AAAATCCTGA
2801 CCCTTGAAGA CCAGTTATCG AAGCAGTCAA CTGAAGCGAA AATGAATCAG
2851 GTGGCTGAAT GGGAGAAACT GGACAAGCAT AAGCTTGAGA TGGAGAAAGA
2901 CGCGGCAGAC CAGGCACTGT CTGAGGGCCG TATCTCGCAA CTCGAACGCC
2951 TAGACTTGGA AATCGAGTTT GAAAACCGCC GCTATCAAAT TGCCTATGAC
3001 GCATTGCAAG AACGGATCGC ACTTGCTGAA CAAAACCCGA CTTATAGTCA
3051 GACGGCCATT GATAAGCTTA AAGCACAAAT GGGGGAGCTT GGGCAAGGCC
3101 ATGAACGGAC CCAGGCGAAA AATGAGGGCA AACCCCAAAG CCAACCCCGG
3151 AAAAACCCCC CGAACGTCAT GGAAATGCTT CAGGACGGGG GCAAAAACGT
3201 TTGGCAAAAA GCGCAACAAC AGATGGGGCA GGCGTTTTCA GCCATGCTCC
3251 CCCGTACCCA AAATTTCCGC ACGGCCATGA AAAACTTTTT CCAGAATATG
3301 GGGCAAACCT TTATTCAAAA AATGGTTACA AAACCCCTGA TGGGCATGAT
3351 GCAGCGTATG GTTCAGGAAT CGGCGATTTA CAAAATGATT TTCAGCACTA
3401 AAGATACCCT GGAAACGGCG GCGGCGGCCA AGACTGGGGC AACCAAGTCC
3451 ACAGAAACGA CGACAGTTGT TGGCAAAAAT GCCGTTCAAG CGGCGTCAGG
3501 GGCTGCATCT TCACAAGCTG GATTCCTTAT GTTGGCCCGA TTCCTGCTGT
3551 TGCTGCGATG GCAGCAATGA AGGCGGCATG ATGGGAATAA TGGGCGGGGG
3601 TGGCGGTTCT CAACGACCCA CGACACACAA CACGGATTCC ATCGGCGGCA
3651 GGTGGTTGGG ATATTCCAGC CGGCATTAAC CCTCCTCACT CAGTTGCACG
3701 AAAACGAGAT GGTTTTGCCT GCAGAACACG CCCAAACAAT CCGCGAAATG
3751 GCAGGGCAAT CAGGCGGCGA CGACAGCACG ATTATCATCA ATTCAACCGG
3801 CGGCGACTTT ATCCACAAAA AGGACTTGGC GAAACTTTTG AAACAGATGA
3851 AACGTGATTT CAAATTTGTT TAACGGTCAG GCCGTCTGAA TGGTTCAGAC
3901 GGTTTTTTTC TTAAACGGAC TTTTTGATAA TTTTGTCAAA AAGTCCTTTT
3951 TCTTTGGAGA TGGAAAATGA GTAACTCATT GAAATGGGTT AAATATGTGT
4001 TGGAATGGCG TTTTCTGCCT GTACGTTTTC AGAAGTGGCT GTTTGGCACT
4051 GGGACGCGCG TTGTCGAGTT TGCCAGTGGG CTGTCGATGA TTGGTTATGC
4101 CGCCGTGTTT GCGTTTTCGC CTGTCGATAT TTATGAATGG CCGATTTACT
4151 ACAAATTCAA AACGATACCC GAATCTATCC TTATTCCAGT TTTCGGTGGT
4201 ATCGGCTTGT TGCAGTTGGC GGCGATGTAC TGGCAGACGT GCCGCGGAAA
4251 CGTCTTTTCA GGCTATTTAT TGCTGGTGTC GGCGTTCATT TGGTACTTGA
4301 CCGCTCAGGC GTTTTGGGGA GCGTTTCCGC CTGCTCACAC GGGCATGGAC
4351 ATTCCGCCGA TTCTGTCATT CTTGTGCCTA CTGGCTGGCA ATAACTCACT

```
4401 TAAATTTTTA TTTTCGGGCA GAAAACTGAA AGACGGCCTA AAGGGGGAAT
4451 GATGGATTTT TTCCAATTTG GCTACCTGTT TGCCATAGGG GGCGGTATTG
4501 TCGGCAGCGC GTGGTCGAGT ATTAAAGACC ACGACAAAAT CGTTTCGAGT
4551 GTTTTTGAGG CGGTTGTATC GGCAGTGGCA GCGGCGGCGG TAGTGGAACG
4601 GTTTTTGATG GTTAACCAAG TTTGGACTTG CGCTATTGCC GGCGCTTTTG
4651 TCGGCATTTT GACAGGTCAT GCCATGGATA CCGTTAAAAC CTTGGCTCCC
4701 GGGATGATGA AGAAGTGGGC GAAGAAAACG GCTGACAAAT TTATCGATAG
4751 TAAAGAGTAA CAACAGGTCG TCTGAAATCA GACGGCCTTT TTAATGGAAG
4801 AGTAAAAAAT GCAAATTACT GAACATTTTA GTTTAAAAGA GTTGACACGA
4851 AGCGAAATGG CTCGCCGTGC CGGTATCGAT AACACTCCAT CTGAAGCTGA
4901 GCTTGAGAAT GTCAAGTACA CGGCAGAGCA GTTGGAAAAA GTTCGTGCAT
4951 ACGTTGGTCG TGCGATTATT GTGACTTCTT GTTTTCGAAA TGAGCGCGTG
5001 AATAAGTTGG TGGGCGGTGT GCCTACGTCG GCCCATCGTT TTGGTTTGGC
5051 TGCCGATTGT GATGCGGTCG GATTGACTTC TTTAGCTTTC GGCAAATTGA
5101 TTCTCAAGAT GCGCGATGAA GGGAAAATTA AGTTCGACCA GCTTATCTTG
5151 GAATACCCTG AACGTGGTGA TGGCGCGTGG ATTCATCTTG GTTTCCGACG
5201 AAATTCGCCA ATGCGAAACC AAGTTTTCAC GCTCACGCGC GTAAAAGGGA
5251 AGGTTGTACG CATTGACGGT TTGCGTGTTT AAGTAGGTGG TGGCTATGAC
5301 TCCTATCGAA TTTTGTGATG CTCGAATTAA AGAATGGGAA GCCAAGCTCA
5351 AGCAAGCGAG TGAGCAAGCG GATTTGAAGG GCTTTGAACA CGCTGAGCGA
5401 GAGCTTAAGA ATTATCAGCA AATGCGCGAA ATTGAGCGTG CCAAGCTTGG
5451 TAAAGAGGAG GCTGTATGAT TGCCGGAATG TTGAAAAATT GGCGGTTTAT
5501 GTTGGCCTTG GTTGTTTGTG TTTCGGTTGT TTTCGCATGG CAATACGACC
5551 ATGCGGCCCA ATACAAGCGC GGCCGTGAAT CAATGGCGGC GGAAATTTCA
5601 GGCCGTCTGA AAGATGCCGC GATTGAAAAA GCGAAACAAG ACCGCGAATT
5651 GTCCGCTGCA TATCAGACCG GCAAAGCCGT GCGTGAAGAG AAAGAAAGGG
5701 TGCGATATGT TGAAGTCCAA AAGATTGTCG AAAAGCCTGT TTTTCGGAAT
5751 ACCTGTGTTG ATTCTGATGG GCTGTCAATC ATCAACGCCG CCATTGCCGA
5801 CGGCAATTAA ACCGCCTGCC GACCTTGTGC AGCCATGCCC GAACCTGCCT
5851 AAACTTGAGG GCGGTACCGG CGCGGACGTG TTGCCATGGT CGTTGCAAGT
5901 AATCGGCTTG TACAATGACT GCAAAGCGCG GCACAAGGCG TTATCTGATA
5951 CTCTTCAATA AAACAAAGGC CGTCTGAGAT TTCAGGCGGC CTTTTTCTCA
6001 TTCCAACTGT CAACTTTAAG TTGATGATTC AAGGTATTCC AACAATTTCA
6051 GCCATTTAGT GTGAGGCATG TTGGCGTAGC TTTTCAAATT CGGGCTTGTT
6101 TCCCATTTTT GGGCGGTTTT CAGTGTCGAT TCCGTGATGT CGGCAACGTT
6151 CTGCTGAGTT AGTCCATATC TTCGGCGCAA TGCCTTTAAA TTGTTTGGCG
6201 TGTAGCCCAA TTCCGCATTT TCAATCATAC AAATCCCCAT CTTCTGACAA
6251 CATCAAATCG TCTGCTGTGA TTTGGTCGCA TACTTTTTTA AGACTATTCA
6301 GCCATTCTTC GGTAATTTCA GGGCGGTTAT TGTCTAAAAA GATAAAGTCC
6351 GACAGCCAAA GCCCGATATT CGAGCGGTAG GCGAGGCGGC CAGTCTTGCT
6401 ATCCTCTATC AAATCGCCGT TGACAGGCTT TGCCGGTATT GTTCCCGTGT
6451 ATCTGTCGTC TAGTGAT
```

The 14 single sequences of the AA1 virus that did not overlap with the three contigs: (Total 7279 nucleotides)

>Sa30
TAGCATTTTATGTTAAACCTTACTCAGAACATCTTCAGCAATCTTGCGAAATTCTTCAGC
ATTAAGCCGTACCGCCGGTGTCATACTGCCATCAGCTTGATTTAACAGTCAAGTTCCGAT
TTGTCGCTCCATTTCCCACGTTGGCGATTCTGATCAGCCAAAAAATAGCAGCAGGCGTGT
CAGGCGGGTAGTATTTCGTCATCGGGGTTTGAATAATTTCCCCGCCAACTACCCGTATAT
CTACATCAGGGGCTTCATATCCCATTGCACGCTGATACAGGCGGTCAGCAACATTTGCAT
CCGCCAACATCTTCCCCTTTTTTATGGACTCCAAAAATTCGGGAAATTCGTTTTTCCAGT
TATTAACTGTCGATACTCCGACATTAAAAAAATCAGCCATGTCCTCGTCTGTTGCGCCAA
GCAAGCATAATTTATAAGCTTGCTCAGCATATTCATGCTTGTATTTTGTCGGACGCCCGA
TAGA
>Sa49
TTACTCAGACAGCGACTTGAGCAAGACGACCAACGAAGCGCAGGTGGCGATTGAGCAGTT
GCAAGAATTAGACGACCTATCGCCGAACCAACGCCGTGAGATTCTGAAGCCTGTTCTCCG
ATTGACCGAATATCAGGAAGCATACGGCGAAATTATCCCCGAGCGTACCGTTTCAAAAAT
CGGCCTGTATTGTGCAAGCGTACAAATCGCACTGTACACAGCGAGCCTATACACGCGGCC
GAAACGTTATACTGGCACTTGCACCGCGACGAGGAGCGTTATCTTGGCTGGGTTTTAAAA
GTCTGCCATGAGACCGACACGCCGCAACCGGGCGACGTTGTCGTCTGGAAGTTTGGGCGC
ACGTTTTCACATGGCGCGGTTTATGTTGGCGACAACAAAATTATTCACAGCTACATCGGG
CGCGGTGTGGTTTTGGACGAATTGGATCA
>Sa11
TGATCCGAAGCGTAAAACAATTATTGACACAAAATGCTCATGGGACATCGGCACACATCC
ATTCTTTCAAGAAGAAGCTGAAGCAAAAGCAGAAAAGGCCGGATATGGCTGGCAAATGCA
AGGCTACATGTGGCTTTTCGATTGCGAAAAGGCTGATATTGATTTTTGGATTTTCCCAAC
GCCCGAAGAGCTTTTAAAGCCTTATGACGATGTAGCCAATTTGGTTGAAGCGGTTGAGCG
TCTGCCGTTTGAAAAACGACTGACAACAATCACAGTGTACCGTGACGAAAACGCAATCAA
CCAAATCAAGCGAAAAGCAGAGGCGTGCTTTGAGTATGCCGAGAAATTAAAACAGGAATT
TGAGAAAGGTAAACAATGCTGAATAAAGTAATCCTCATTGGTCGCTTGGGCCGTGACCCT
GAAGTGCGCTATATGCCGAACGGCGAGGCCGTCTGTAACTTTTCCGTAGCTACAAGCGAA
AGCTGGAAAGATAGCAACGGGCAGAAGCAGGAGCGTTCCGAATGGCATAACGTGACCATG
TACCGCAATTGGCAGAGATTGCAGGGGAATACCTGAAAAAAGGCAGCCAAGTGTATTTAG
AGGGGAAAATCAAATCCCGCAAGTACACCGACAAGAACGGCGTGGAACGCACGGCTTACG
AGATTATCGCCAATGAGATGAAGATGTTGGGCGGCAATGCACAAACACCAGCGC
>E11
TACACTTGATAAACAAGGAAATACAAAATGGAAATTAAACAATTTGAAGTAAACAGCCCT
TCCGCAATGCTGATGATGCTTGCAAATATTTTGGCAGACATCGAAAGCAAAAAAGAGGCA
AAAGAAGAAGAACCATTGCCGCCTGTAACAGTTACAGAGGCCAAAGGCATTAATGACTAC
GCCATCGGCAAAGAAGTGATTATCCGCACATATTCCGCAGGCGTTTGGTTTGGTGTGTTG
AGCCAAAAAGCAGGCAATGAAGTGATTCTGACGAAAGCACGCCGTATGTACAAATGGTGG
GCGAAAGAATCAATCAGCCTGTCAGGTGTTGCACGACACGGCATCAAGCAGGACGACAGC
AAGATTTGCGGTGAGCTTGATTCCGTATGGCTTGAGGCGATTGAGATTATCCCAGTAACC
GGCAACTCGGCTGAATCAATTCGTACCGCGCCGGAGGTTGCTCAATTATGAGTTACCTAG
ATAAACCGTTGAGATACAGCTACGGCTACGGCAGCGGCGACGGCTACGGCAGCGGCTACG

GCGACGGCGACGGCTACGGCAGCGGCGACGGCTACAGCTACGGCTACAGCTACGGCTACG
GCAATGGCTACGGCAGCGGCGACGGCTACAGCTACGGCTACAGCTACGGCAGTGCCAACG
GTAGCGGCTACGGCTAAATTTTGATACAGCCCCTGCAACAGGGGGCTTATTTAAGCGGCT
GGAAGCGGC
>E17
TATTTAATAACACTCTTTATCACAAAAGGAAAACAACATGAAGAAACTGTTACTGACTGC
TATTGTCGCAGGATTACTGACTGCTTGTGCGGCGGCAATTGAGCCAAGTCAACAACAATT
AGCCGCTGCGACCTATCCCGCCCCAATGCCGCCAAGCCAGTTTGAGAAAGCTGTAAAAGA
ATGGGCGGTCGATAACCTTGTTGACCCTGATTCTATGAATATTCGCAGCGTTGATACAAC
GCCAGCGCGTAAAGGTTGGATTGCGGTTTGTACGAAAATTGACCCGTCAATGGGTAATTG
CATGACGCGTATGTTTTACTTTGGCCATATCTTCAATGCGCGTATTAATGCGAAAAATCA
GCATGGCGGATATACAGGCTTTAAAGACTACGCCTTTGTTGTGCGTGGCGACCAAATTAG
CTATGGCGTTGAAACTGAAAAAATTTCTAATATGAAATTGTTCTAATCTGTTGACATGAT
GTTACCTTTTATGTACAATTATGCTATAGTTTGGAAATAGCTATATAAACCGCCTTTACA
GGGCGGTTTTTGCATTTCAAGATAGCCTGTGATTCAGGCAGAAAGTCAATAAAACGCGGT
GCAAGTGAAACGCGTTTGCCCAGCCTAGATGGTTGTCATGCATGACAGACTATAAAGCGG
TTCTTGCACTTCGCCCTATGA
>Mf32
TGATTACGGCCGATTGAAAAAATACCGCATGACCCGTGAAGAGCATAACAGCTTCGCAGA
ATACTTGAACGAGGCTAGAGACTGGAAGTTTGTGATTGATACTGAGATGGTCGGAATTGA
AGCAATCGCCGCGCGATGCCGATTAGAGAAGCGCAAATCAGGGCTTGATGTGTTGGTTGT
TGACCATTTGCACTTGATGCCGCGCAAAGGCGTGAATGAGGTTGCTGAACTTGATGATAT
TACGGCACGACTGAAACGCTTGGCAATGGAACTGCAAATTCACGTCTTGCTTGTTGCACA
GCTTAACCGAGCGACAGAGAAACAGGCAGATAAACGGCCAAGCCTTGCAGACCTTCGAGG
AAGCGGCGGCATTGAGCAAAACGCAAACTTAGTGCTGATGCCATACCGTGAGGGCTACTA
CGATTCAGACGCACCGCAAGAGACGGCTGAATTAATTATCGCAAAAAACCGAGACGGCGA
ACGTGGCGTTTTAGACCTAGTTTGGCAAGGGCAGTATCAACGGTTTTGCGAGTATGAGTA
TTACAGCTAGGAGTGGTAAATGCGTGAAACCTGTTATCACTGCCTACACGCAGATTTTAA
AGCCGAAGCAAACGGAACAATGCGCGGATTTGCAAGATGCACCAAAGCAAAGACGGTCGA
AAATAAAGCAAGTTACTACTTCGGCGGTTTCAAATGCGACAAGGGCGAATTTAAACCGGC
AAATGCCGAAACGATGAAAGTAC
>Mf23
TGCATGAAATACGCAATCAGAACAGTTTTGGCCGTGTCAGCAATCGCAATCGCGGCTTGT
AGCTTTTCCGGTAAAGCCGAGAAACCAGTAGAACCGGAAACCATCAGCCAAGTGCAAGAA
ATTGAACAGACATACGAAACCATGCCGGATGAAGTAAAGGTCATGGGAGACGCGGAGATT
AAGCCATGCAACACATTTTAACAGGCTGCTATGTACCAAACCCAAAGGCCATAGCCATAT
CAGCAAGCATTGCGGCGCGCGGATGAGAAACGAAAACGGCGTTTGGCACGTTTGGCAAGA
CGTTGAGACATTCGTCCGAAACACTCACTCGCTGATTCTGAAGAAAAACGCAGGGCGCAA
GAAAGTCA
>E37
CCCGCTGACCTACACCGATAAAGACGGCGCGGTCTACACCAAGACCTTCGTGCTCAAACG
CGATCACTATGCGGTCGGCGTGGATTACAGCATCGACAACAAAGGCGCGACGCCGCTGGA
GCTGACCCTGTTCGGCCAGCTGAAGCAGACCACCGAATTGCCTAAGCACCGCGATACCGG
CAGCAGCAACTTCGCGCTGCACACCTTCCGCGGCGCGGCGTACTCGTCCAGCGATGACAA
GTACCAGAAATACGCATTCGACAAAGACGAAAACCTGAGCGTCACCACCAAAGACGGTTG
GGTCGCCATGCTGCAACAGTACTTCGCCACCGCGTGGGTACCGGCGACCAAGGGTGACAA
CACCTTCTACACCGCCAAACCAGGCGACAACCTGTCCACCATCGGCTTCAAATCTACGCC

GGTCGTGA
>Sa52
TCGTTAATTTCTTCTGCCAGCTTGGCAAAATAGGCTTGCGCCGCCGCCACGCGCGGATTG
CTGATATTTCCATTCATCACGGTTAAGTAACAGGCAAACCGTGTCAATTTAATATCATTG
TCGCAATTATGTGAGGCCGTCTGAATAAAGTTTTCAGTGATCGGAATATCCAACTGAAAG
CAGACGGAATGGGCGCGGTTGATTGCTTTTAAAATTGCCTGCATATCATTGTAGCCAAGC
ATCATCGCTAAGTCTGAAGCATACCAAAAGGTGTTGTCATCGGCTTGGGCAAAGCTGTCA
AATTCTACTGTTGATTGAGGGGAAAAGACGGCAAGCTGGTGTGTCATATCGCAATTAAAT
TTAGTCATTTTTTAAGCTCTTATTTTACCATAAAATATTGTTTTATAATGATGATTTACA
CTTCCATCAAGGCGGGGAAAG
>E14
TAACAAATTATGCGGACTCCGTGAAGGGACCATTCACCCTCTCCTCAGACCACGCCCAGA
ATACCGAGGATCTTCTGATGAACAACCTGAAACCTGACGACTGGACGTCTATTACTGAAA
AACTGACGGTTTGGGGACCGAACAATAGGGGTCAGGAAATTCACACCACCTTCTCCTAAG
CGACAGGAGAAATAAAACTCCTCTCAAAAGAGGATAAGAA
>B16
AGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGT
GCGGTATTTCACACCGCATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGT
TAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCC
GCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACA
AGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACG
CGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTA
>D6
TGTTTCAGACTTGGCTATGCCTTTCAGACGGACTTTTACGCGATACTTTCCGTTGCGCTT
TTCGATGGTTGCCATTGGTATATTATTGGGACACTGAGGGGACACGGCATTATATGTTAT
AATCAATCCTAATCAATCATAATGTATTGAGTGGCGCAGATATAAGATTGATTTATATGT
ATAATCTATCAAAATCAATCATAATCTACTATAATCGATTAGCTGTTTA
>D03
TAGGGCAAATGTTCTTTCCCTGCCTTACCCGATCAGTACTAACCGATACTGGAAAACCTT
TCGGAATCGGCAAGTGTTAAGCAAGGAAGCGAAAGCATACAAGCTTTGCGTTTCTCACGC
GGCAGAAAGAGCAGGATTTAGGCCGTCTGAAAAGGACGTAATCCTTTTTGTCAGCCTAGT
GCCAAAGATGAACAAGGACGGCACATCAAGCAAGGTAATTTTAGACCTTGATAACTGCCT
AAAGGTCGCTGTCGATGCCTTGCAAGGTGTTGTCTATCACAATGACAACCAAGTCAAATC
TATTTTATCAACATATTCGAGCGAGCCAAGAGAGGGCGGCGGGCTTGATATAGGAATTGC
GGAGATTTCGAAATGAGCATTTTTAATCTGATTGTTTACGCCTTGAGCGCGGCGGTTTGT
TTGTGTTTGACGTTCAAGATTGCCGTTTTCGCATATCTGACCTACGGCGAGCGTGTGTCA
TGGTGGTGGGGCAAAGAAGATGACCGTATAAAGGTGGGTTGGTCAGATTTCGCGTTGTGC
GCCGTGTGGTTGTTTGATTTGAGTATTGATTTCCTGCTTCGATTTACGGAGGAATTGCTG
AAGTAATGAGCGCGATCAGAAAAGCGGCCAAAGGGGAAGATTGCACACTCAATATCGCAG
GGGTGTGCAATTACAACCCTGAAACAGTGGTTTTGTGCCACTTCCCAAGTGAGACACATG
GCATGGGGCTGAAGAGTAACGATTTATCGGCAGGCTTTGGGTGTAGTGCTTGCCATGATG
TGATAGACGGCC
>Mf5
GCCACTGATGACGACCATCTACAACAATCTAGGGGCGTGGTTAGAGGCTAATGCGGAGCT
TGAAGTATGTATCAGGCCGTATAACTCAAAGCGAAGCATAGAGCAGAACAGGAGACTATG
GAAAATCTACGGCGAACTGGCAGATAAAGCGTGGGTAAATGGCAGGCGATACAGTGCAGA
GACATGGCACGAGTATTGCAAAGGAATGTTTTTAGGCTTTGAGTTAAAGGCCATGCCGGA

CGGCACGGAGATTAAAACGCCGATAAGCACGACAACGCTAAACACGGCGGAAATGACAGA
CTATCAAAACCGCTTGCAGGCGTGGGCGGCAGGGAACTTCGGGTTAATTTGGGAATTTTA
AAGGGGCGGGAAATGTATTACACGGTTGAGCAGGTTTTGGCGGATGTTTATAAAATCAGG
GGCGTGAGAATGGAGCCGTTGAACAACACGGCTTCAGTTTGCGCTTGGTGTGAAAGCAAA
GGTGTTACCGGCGGTGGTGGAGATTTGACGCAGGCAGAAACCCACGCAAACGCCGCGATG
ATTATCAGCAAAATTGAGCGCGTATTGAATCGGTACGAGCTTGCAGTAGTGGAGTGTAAA
TACAGCGAGGACTTGAGTGGAATCGTTGATATTACCGCCTATATTGAGCAACA

Sequences of A1 Virus:

>1
TCCCGTGTTATCCAGCGAAAAAGAATCTTATGAAGAATATTCAAATTCTGTAATAAGCAG
TATCTCAAAAATCATTGCATGTGTATTTTTAATATGGTTAGCTAGTTTGTTTATTTGAGG
AGGTGCAAGATGAATCCAAAATATAGAGCGTGGAATTCAGAAACAAAAGAAATTGAAGTA
TTTAATACTTACGAAGAAATTAGTGAATTATTTTTAGTGTTAAGTGCAGATGATGGTTTT
TATTCAATCATGCAATCAACAGGGCTATTTGACAAGAATGGCAAGGAAGTTTTTGTCGGA
GATATTATTAAATGTACAAGAGGTTGTCTTCATGAAGTATATATAGAAAAAGAATATGGT
GGTACATATGTAGGCGGAATGCCAGCTATATATCTAAAAGGTATAAGAGAGGGTTATGCT
TGGACTGGTGATGAGGAAATCATTGGTAACATTTACGAAAATCCCGAATTGCTGGAGGTG
CAAGATGACACGACCAAATAGATACCCATACTCAAAAAGTCAGTGGGAAGAAGAAATAAC
GTTAGTATGCTTGGGTGATGACCACCTTAAATTGAGAACAGAGCGTAATAGAATTACAGG
AGAAACGAAACAATGACTTGGGGCGAAAGAATGGTAGAAATTAACTGCAGGGAGTATGAA
ACTCTTACAATAGGACTCGAAGATGACAATTGGGAGTACTTGAAGACTGGTAGAGAGTTG
TTATTCCCTTGTAAAGAAATCAAGCGAGTAATCGTGTTGAAACATCGATAAGGAGGAAGC
ATGACGAATAATGTAAAGCTGATGAACGCTAATTTTGCATTTTTGATCTTTGTTTTAATC
GTGGTATGCGTC
>2
TCCCATCAGTCGATCATGCTACCTCCGGCGCAAAATTTCAAAAATGCTTGCTTTATAGTG
AATTCCAAAGTATAAGCTGTTGTTCTCTTTCTTAGGCTAGATAACTTTAATTAATGAATT
AAAAGGCAAAAATAATGGAAAATAGACAAACCAATTCAACCATCAAATCTCGTGCGGCGG
TGGCATTCGCACCAAATCAACCCTTACAAATTGTGGAAATCGACGTAGAAATGCCACGCA
AAGGCGAGGTGTTAATCCGCAATACCCATACGGGCGTGTGCCATACTGATGCATTTACGT
TATCAGGAAGCGATCCTGAAGGTGTATTCCCTGTGGTGCTTGGGCATGAAGGTGCGGGTG
TGGTTGTTGCTGTGGGCGAAGGTGTGTCAAGCGTAAAACCGGGTGATCACGTTATTCCTC
TTTATACCGCTGAATGTGGCGAATGTGAGTTTTGCCGTTCAGGTAAAACTAACTTGTGCG
TCTCAGTGCGTGATACACAAGGTAAAGGTTTAATGCCGGACGGCACGACGCGTTTTTCTT
ATCAAGGTCAGCCGATTTATCACTATATGGGCTGTTCGACTTTCAGTGAATACTCAGTTG
TTGCCGAAGTTTCATTGGCGAAAATCAACCCGGAAGCGAATCACGAACAAGTATGTTTAC
TTGGCTGCGGCGTTACTACAGGTATTGGTGCGGTGCATAACACAGCAAAAGTGCAAGAAG
GCGACTCTGTGGCGGTGTTTGGCTTGGGGGCGATTGGTTTGGCTGTGGTGCAAGGTGCGC
GTCAAGCCAAAGCCGGTCGCATTATCGCTATTGATACCAATCCTTCAAAATTCGAGCTGG
CAAAACAGTTTGGTGCAACTG
>3r
TCCTCTGTCGCCAATTATATGAAACAGCATGGCTGGCAGACCGGCGGTAAGATAGTTGTG

CCGGTCAGTCTGACGATTACGCCGCACTTGCAGGCGATTATCGATGAGAAAACTGCTTTG
ACACGTACTGTCTCAGATTTCAAAGCTTTGGGTGTCGTGCCTCAAGCTGCTGTTGCGGAT
AATGAAAAGGCTGTATTGTATGCTTTGGAAACCAGCCCGGGCGTATTTGAATACTATTTG
GGCTTGAATAACTTCTACACAGTATGGCAATACAACCACAGCCGCATGTATGTAACAGCG
GTGCGCGATATTGCGAATGCAATCAATAACAATGGCCTGTGAGCCATAAGAAAACCACCC
TTCGGGGTGGTTTTTAAGTATTTATGACTTTGCAGAACAAT
>6
TAAGCGACTAAAGTCGCAACAACTGTGTCAATTGCACCAATTTAGTAAAAAAATCTAAAA
AAGAACACCTTGAAAGGTGCTCTTTGTAAATATAGTAATTCTTTCGAATTAACGTTTACT
AAATTGTGATGCTTTACGAGCTTTCTTAAGACCTGGTTGGGGCACCACTTATTTCATCTA
ATTTCAAAGTGCTTTAAAATCAACGTTTTCATAACTTTTCAAAAAATAAAATTTTAACTA
ATTTCAATCAGTTTCATTAAAATAAGGTAAACTTTCATAAAAAATAAGCGAAAATCGAAC
CAGAATATCAAAAATTTACCTTATTTAGTTTTTATAGAAGTTAAATAATAGGTCATTACG
TTTCAACAGAGACGTGAAAACCTTGATTTAACGCGATTCTTGAATTTTCGAATTTCAATG
TATTGCAATATAAATCAACCGAATCACTACCTTTTTCACTACCTTTTTTGAAATAAGAAT
AATGATTCGGATATATAAAAGAAGAGGAACTTTATTGTGAAGCTCC
>7f
TCCTCTATCAACGCAAACACACCAAGCCCGACGACTGGGGCGTGGCATTGAACCTGCCGC
GCGATACCTATTTGGGCTACAACTGGAACAAAGGCATTTACGACAAAGCGAACGCATTCG
CCGAAGTAGAACACTATTTCAACGACAACTGGCGTTATACCGGCAAGCTGGATTACAACT
ACAACGAAAATATCAAGAAAAACAGCGGCATTTACAATACGTCCACGTCCTACGCAGGCT
ACACGCCCGGCGGCGCATTAGCCTCTGGCTGGTTGAGCCGTTACGACAATGATGAAAAAC
AACTGACTTTCAAAAATAATTTGAACGGCAAATTTGAAATTATAGGCGTGCCGCAGGAAA
TCTTTACCGAATACACTTACACGCACACCAAAAACAACGGCAGCCGCCGTCAATACAACC
CGGGCGTTTCCTTCGACCCGATGACCTTTACCGGTAACGAAATCGCCGAACCCGCCGACT
GGTACGCCACACCTTATCAGATGTATTGGGAAACCCATTCCAAACGCACCACCCACGCGC
TGCTTTTGGGCTTCCGTTTCAATATGTTGAAAGAAAAACTGCACATCATG
>7r
TCCATAGTAGTTGTTCGCGCCTTTGCTGCTGACTTTGTTGTTCTCGTAGTAGTAACGGTT
GAACAGGTTGGTCGAAATCAGTGCAAGTTTCAGATTTTTCGATGCCTGATAATCAATGTT
GGCATTGAATACCGCATGCCCGCCGCGCTTCACGCCGTAGAGGCTGGCGGTTTTGCTTTG
TGCGGTTACGCCGCCGCCGATGGTCCATTTTTTGCCGTCAAAAGGCAGGTTGTAGCTGGT
GTGGAGGCGCAGCATGTGTTTGGGCGTGTGCAGGCTGAAATTGGTGCCTGCGGGGACGGT
GCCTTTGCTTTCCGCTTCAAGATATTTGGACTGGTTGAACGTATAACCCGCAAACAGTTT
CCAATCTTCGCTCAGGTTGCCGGATATTTCGGCTTCGATGCCGCGGCTGCGGACTTTGCC
GACGGGTTCCCAATACCATTTTTGGTCTGCCGTATCCCAAACCTGCACCGCGCGGTTTTT
CTGTTCAATGTCAAACAGA
>8f
TCCTGTCGCCCGCGATTTCTATCGAGCAAAAATCTACCAGCCACAATCCGCGTTCCACCG
TCGGCACGGTTACGGAAATCCACGATTACCTGCGTCTTTTATACGCCCGTGTCGGTACGC
CGTATTGTCCCGAACACAATCTGCCGCTGTCGAGCCAGACCGTATCGCAGATGGTCGATG
CTGTGTTGAAGCTGCCGGAAGACACGCGCGTGATGATTCTTGCGCCGGCTGTGCGCGAGC
GTAAGGGCGAGTTTGTCGATTTCTTTGCCGACTTGCAGGCGCAGGGTTTTGCGCGGGTGC
GCGTGGACGGCGAAGTCTATCAGTTGGACGAAGTGCCGAAGCTGGAAAAAAACATCAAGC
ACAATATCGACGTGGTCATCGACCGCGTGAAAGTGAAGGCGGACATCAAGCAGCGGCTGG
CGGAAAGTTTTGAAACCGCGCTGCGCCACGGCAACGAGCGCGCGCTAGCGATGGAGATGG
ACAGCGGCGAAGAACATTGGTTTTCCGCGCGGTTTGCCTGCCCCGTTTGTTCGTACAGCC

TGCCCGAATTGGAGCCGCGCCTCTTCTCGTTCAA
>10
TCCTGATTCAGCTCCAACAAAAAATCAAGCGCACTTCCACCATCGCGCGTGACGAGTAAG
GAGAGGACGATATGGCAAGCATTCAAAACTTATACGAGACCGTCGTCGGCGTGCTTGGCG
ATCAGGCAAGCAAAGTCATTTCAGCTTTGGGCGAGATTACCGTCGAGTGTCTGCCCGAAC
ACTATATTTCAGTCATGACCGCATTGCGCGACCATGAAGATTTGCATTTCGAGCTTCTGG
TTGACTTGTGCGGCGTCGATTACAGCACTTACAAAAACGAAGCATGGCAGGGCAAACGCT
TTGCCGTCGTCAGCCAGCTGCTTTCCGTTAAAAATAATCAACGCATCCGCGTACGCGTCT
GGGTTTCAGACGACGACTTCCCCGTAGTCGAATCCGTAGTCGATATTTACAATAGCGCGG
ATTGGTACGAACGCGAAGCCTTCGATATGTACGGCATCATGTTCAACAACCATCCCGACC
TGCGCCGCATCCTGACCGATTACGGCTTTGTCGGACATCCGTTCCGCAAAGACTTCCCGA
TTTCCGGCTATGTCGAAATGCGTTATGACGAAGAGCAAAAACGCGTGATTTACCAACCTG
TTACCATTGAGCCGCGCGAGATCACGCCGCGTATCGTCCGTGAGGAGAATTACGGTGGA
>11
TCCATTCTTTTGATTTGGCGAAATTTGGCTATGTCAACCTGGCTCCCCAGATCAAACAAT
CAAAGGACTATGACAAAGAAAATTTCCAGAATCGCCAGCTAATCTTGGAAGCAGGTTTTT
ATGAGCCGATTCTAACAGCGATCGGACAAAAAATTCCGACCAGTGATGCCAGAATTCTAG
ATATCGGTTGCGGAGAAGGCTATTACTCCCGAAAACTACAAGAAGCCTATCCCAAGGCTA
CTTTCTATGCCTTTGATCTTTCCAAGGAATCTGTGCAACTAGCTGCCAAGAGTGACGCTA
GCTGGAAGGTCAATTGGTTTGTAGGAGACTTGGCTCATTTACCCATTCAATCCAAAAGTA
TGGAGGTCATTTTGGACATCTTCTCCCCGGCTAATTATGCTGAATTTGAGCGTGTCTTAA
AGGATGAAGGGGTGATCATTAAGGTTGTCCCAACCTCTTCTCATCTGAAAGAAATTCGTC
AGTTGGCCCAAGAGCAATTGACCAAGCAATCCTACTCCAACCAGGAAATTTTGGAGCACT
TTGAAGACCACTGCCAGATCCTCTCATCCGAAACGGTTAGCCTCACCAAGAGTTTAACTC
CTGAAGAACGCC
>35
CCGTCTGAAAATGATTTCAGACGGCATTTATTTTCGTAATGTGATCATTCAGGCTCTTAA
GGACCTTGTTTTGGAGAGTTTTGTTCTTTTGGTTTGTTCGCTTTCCGCGTTTGTATTTGG
CTTTGTTTGGGGTGTTTTCTTTGCAATGTATTTGTGTGTTTTTCTTAGGGTATATGCTTG
CTTCTTATCCGTGTGAGTTGTCACGGCACATGGCGGGAACCATTATCGGATCCAGCGTTC
CGCCGCATTGCCAGTTGAGCAACCAAGTCGAATCGTTATGATTTCGATAAGGGGTCAGGG
TCAGGACTTTGCCTTTGTTTTGCCCTTTATCGAAAGGAATGCTGATTTTGCCTGTATCTG
CTTCTACAGTTACGCCACCGTTTTGGATGTATGTGCCGCCGATGCGTTGAGCGGTAACAC
CGATTGTGCCTTGTTTGCTGACATCCTCTGCATTTGGCAGGCGGTTATTTACTGCAGAAG
CGATGCTGATTTCGGTTTT
>17
TCCTTGAATTGGTAGACCTTGGAAAAGTTGGCGAAACCTACCCAGCACAAGCGATCTTCG
CGATTGAGGACCCGAACCATATCGAGAAATACAGCGAAGGCAAGCGCGTGATGGTCCAAG
TGAATCAGCCGTTTACGTATCAAGGCGAAACGCTTGACCAACTCGCAAATCTTGAGCAAA
ACGGGAAGATTGGCATTTGGAAATGGTCCGAGCCAAAACCAGCGGGAGAGCTTGAAACTC
AACCCGTCCAGTAAGCTAGAGCGTTAATAGGGGGGTGGTTTAATTGGATCTATTAGCACT
AGTAGACAAGCTGACTCCAGTTTTAGTCGTGATTATTCCAAGTTATTTCTCGTTTAAGAG
TACAAAGACCACTAAAGAAGCTGACAAACGTCTTGAGGGTTTATCGAATAAAATCGATAC
CCTCGAAAAGTCAGTCTCAAACGTGGAAGAGATTGGGAAAGATAACAGAAAGAATCTGAC
TATGATCGGGAAAGGCTTACAACGGCTCCAGCGTTTTCGATTGCAGGAAAAT
>18
TCCTTTGATTTCCTATTTAAAATGTTTCAGCTGTTTTACAGGCTGTAATACATTTCAAAT

TCCAGCGGGTGCGGCGCCATGCGGATACGGCGGACATCTTCCTCTTTAAAGGCGATGTAG
CTGTCGATCCAGTCTTTGCTGAACACGCCGCCGCGCAGGAGGAATTCATGGTCGGCTTTA
AGAGCAGCCAGTGCTTCTTCCAAAGAAGCGCAAACGGTAGGCACCAATGCATCTTCTTCC
GGTGGCAGGTCGTACAGGTTTTTATCGGCAGGATCGCCCGGATGGATTTTGTTTTGAATA
CCGTCCAAACCGGCCATCAACAAAGCGGCAAATGCCAAGTATGGGTTGGCGGTTGGATCA
GGGAAACGCGCTTCGATACGGCGTGCTTTGCTGCTGTTCACAGAC
>23
TCCACAACCAACGTCTGCTGGTTAATGAAAACCAAACCGTTAAACGCGGGCAAACCATTG
CGCACATGGGTAATACCGATGCATCTCGTACCCAGTTGCACTTTGAAATCCGTCAAAACG
GCAAACCGGTCAATCCGGCAAATTATGTTGCCTTCTAAGCTGATGCAAGCAACCTAATTT
CAATTCAATCCCTGCCTAAATGTCATTTAGGTAGGGATTGTTTGTACCGGTTTCCCATGA
AATTTCATCTTCCCTTTCTTTATACCCTGCTATTGGGCGGTTGCGCCGCGTTTTTACCTT
CATTGGACGAAC
>24
CCAAAGGATCTCTTAAATTAAGAGGAAAAAGAGAAAATATTGCATCAGCATTGAAAGAAA
TGCTATTAAGCGACACTGTAACATTAGAAGTTAAATATGATGGTGCTCTACTTATATTCA
GCAGTACAGCTCCCTATTTTTACATTAATAAAACTAGACGAGCGTTTATTGATCAAAAAC
AAATAGAAGGTTGGCTTGAAGAAGAATTTTGTACCATCGAACTTGATAATTTTGAACAAG
CATGGAGCGCTATCCCAGAAAATTATCAAGAAATTTCAAGTAAGTTTGATGTTGATATTA
AAATCTTTACGTTTGAGATGGGTATGGAATTTACACAGGAAATTGAAATTTCCAAAGGTA
AAATTATCAAGAATGTTTGTCGGGAATACGATGATTATCAATGGGAAGTCCCTTTTAGCA
ATTTGGGAGGATAAAATGACAGAAGGGATCAAATTGACGTTGTTATGACTTGGATTGAAG
AACATTTTGCCAGAGAATATCCAGAGATTAAATCTATACAAGACATCTGGAATAAGGACG
ATTTAGGCGGATACGAAACACAGCGGTATTCGAGGGAATTGAATAAAGTGATTATCACGA
ACGACTTGACCGCTAT
> 25
TCCACGTTGCCGGTTGGGATGATCCGCGTATGCCGACCATTTCCGGTATGCGCCGCCGCG
GCTACACGCCCGAAGGCTTGCGCCTGTTTGCCAAACGCGCCGGTATTTCCAAATCTGAAA
ACATCGTTGACATGAGCGTGTTGGAAGGTGCGATTCGTGAAGAGTTGGAAAACTCTGCGC
CACGCATGATGGCGGTTTTGAACCCGCTCAAAGTGACCCTGACCAACTTTGAAGCCGGTA
AAACCCAAAGCCGCCGTGCCGCATTCCATCCGAACCACGAGGAAATGGGCGATCGCGAAG
TACCTATTTCACAAACCATTTACATCGAAGCCGACGACTTTGCCGAAAATCCGCCTAAAG
GCTTCAAACGCCTGATTCCCGGCGGCGAAGTGCGTTTGCGCCATGGCTATGTCATCAAAT
GCGATGAAGTAGTAAAAGACGCAGCGGGCAATGTGGTTGAACTCAAATGCAGTATCGACC
ACGATACCTTGGGCAAAAATCCCGAAGGCCGCAAAGTTAAAGGCGTGATTCACTGGGTGT
CTGCCGAACACG
>26
TCCTTGTCTGCATTCCGCTTCCCTGACTTGAAGTCGGCTACGGCTGCTTCGTTGTCTGCG
AAGACTTGGTGGATGATAGGGATCAAGACTCCAGGATCTGAAATCTGTACCAATCCAGCT
TTTTCGACGTATTCACGCGCACCGCCACCGTTTTTCGCCAAGTGAACGAAGACCTTCTTG
GCAATCTTAGAAGAAATGGTTCCATCTTCGATGATGGCAATCATTTCGACCAAGTTTTCT
GGTGTCAATTGGATTTCTTCCAGTGTCTTACCTTCTGCATTCAAGAATTGTGCTACTTCA
CCTTGGAGCCAGTTAGACACTTGCTTAGCATCGCCACCAAGGGCTACAGCTGCTTCAAAG
AAGTCAGAAGTAACTTTCGTCGCTGTCAATTGGTTGGCATCATAGTCAGACAAGCCAAGC
TCTGCCACATAGCGAGCACGACGGTCTTTTGGAAACTCTGGCAACTCCGTACGCATTTCC
TCAATCCACTCATCTGAGATTTCAAACAAGGGAAGGTCTGGTTCTGGGAAGTAACGGTAA
TCCGCTGCTCCTTCTTTGACACGCATAAGAATGGTGCTCTTATTCGCTTCATCATAACGA

CGTGTTTCTTGACGAATGAC
>30
TCCTGCGTTATGAACGTACGGAAACGGATGCCGGTTTTTCCAAAGCAGGAAATATCAATA
TAGCCGGTTATAATACCTTTGTTCCCTCCCTGATCCTTTCCCATGTTTTTGCCAATGAAC
AAACGCTCAAAATAAGTTATGCCAAAAGGATGCAGCGACCGGGTTACCGCTGGTTAAACC
CTTATGTGAATGCCAGCGACCCCAAAAACATAACGACAGGAAACCCCTACCTGGCTCCCG
AGATCGCTCATAATATCGACCTGACCTACAGCAAATCTTTTGAAAAAGGCAGCGCCCTGA
ATATCGTCCTTTTTTACAATCGTTCCAATCAGGATATACAACCCTATATCACCTATTATC
CCAGCTACCAGATCGGCGATTCGGTATATACAAATGTATCCGTTAATAAACCGGTGAATG
TCGGTTCCGAGAATAATTTCGGGTTGAATATTTATGGCTCCGTACCCCTGACC
>32
TCCAAGAATGAAAGTAACCAGTTTCGGCGAAGTGCTGTGGGACGATTTTCCAAACGGAAA
GGTATTGGGTGGTGCACCCCTAAATGTTGCTGTCCGTTTGCAGTCTTTGGGTATCGATGC
TGCCATCATCAGTCGTCGCGGCGATGATGCCGATGGCGAAGAATTACTACGTCAAATTCA
AAGTAAAAACGTCAATACTGATTATTTGCAGGTTTGTCACGAATGTGCGACCAGTTTGGT
GAAAGTACACTTGGACAAGTCTGGCAGTGCGTCTTATGAGATTGTTTATCCTTGCGCGTG
GGACAGGATTGCGGTTGATGATGCGGCAATCAAACGTGTGGTAGAATCCGATGCTTTCGT
TTTTGGCAGCTTGGCAACCCGTGATGAAGTTTCGCGTAAAAGCTTGGCAGTTTTGCTGAA
AGAA
>37
CCCAAGAGTTACAAGAGGTTGCCGATCAGAACGGAGTAGGAGGATAACATGATTAGAGCA
AGATTATTTAAAGATGAGAAAGGATTGGGGAATCCTGTACCCCCTGCTGAGAGGGTGCGA
AATTTTATTAATAATACTTCGGTCGATGTGATAAACGTTGTTATTACTTCAACGAAAACA
TATAAACCAACTTCTCACGAGGAAGGAAGCTGGGCAGATGAAATACTTTTGATATATCGA
CAGGAGGAAGAAAATGGATAGAATTTTAGAATTGGAAGGGTGGGA
>38
TCCAAACTCTTCTTGTCAGAAGGAACCGTGCGCAATTACTTATCAACGATTTTGGCGAAA
TTAAACCTACGAGATCGAACGCAATTAGCGATCTGGTCAGTCCAAACAGGAGTGACGAGA
CGTGATTTTTCTAAAGGAAATACAGAATGAAATTGAAGCGAAAGCATCTTTGTTGGGGGA
TCGGTCTCCTTGTTGTGCTCTGTGCAAGCGTAGGTTTTTATTGGTGGAAACAGCAACCGA
CCGTACTCCATATCGGCGTTTATGCAGGTTCTAGCTGGGATGTTCCAACTAGTCAACGTT
CCCATGCCTTGGATCGTGCGATTCAAAAATTTGAGAAGTCCCATCCCCACGTCCGTGTAG
AGTATGAAAATGGGATTCCGCAGTCAGATTATTCAGATTGGCTCTCTGAAAAGATTGTTT
CAGGAAAGACGCCGGATGTGTTTATGGTCTCTGAGCAAGATCTTTCCTTATT
>41
TCCATTTTTTTTAGCGAGATAGTCGCATAAATGCGACAATCTGGCAGGATAAATAGAAAG
GAGCAAGATATGAGACCTAAGAGATATCCATATAAACAAAAACCACCCTTTCCTTCAACT
AAAAGAGTGGAGAAAGCAATCAGCGAGCTTGAAGCACTGAAAGAACACTATCTCAGCTTG
ACTGATGATATGAGACCTAGAGCAAAAGCACTAGTCAGTGAAGAATCGGACTATGTTACT
GATTATGATCTTGAGATTGTTTCAACCGATTTAAAACTTCATTTTCGTGAGCTGCTAACA
TTTTTCGAACAATGTCCTTAACTTCACGGACTTTTACTGGTTCAAAAGTATGATTATCAT
CACGAACTAGCTCGATAAGTTCATCAATCAT
>42
TCCTTCGGTTTATCCGCCGCCCCGTTCCGCACCCTGCTCTACGCTTTTGTACTGATTGTC
GGCATGGTTGTCGGCATGGAAATTCCTTTGGTCATGCGCGTGTTAAACCAAAAAGGTGCG
GAATTTAAAGAACTCGTTTCCAAAGTCCTGACCTTCGACTACTTGGGCGCACTCGCCGTC
TCGTTATTATTTCCGCTCCTGCTCGCCCCCAAACTCGGCATGGCGCGTTCAGCCTTGTTG

TTTGGCATTTTCAACGCCGCCGTCGCCTATCTGACCGCGCGTGTATTCAAATCCGAATTA
CCCCGCTACCACGCCATCCGTACGCG
>44
TCCTATCATGATGTGGATGTGACAGCATTTGGCTATAAGTATGGTCAAATGGCAGGTCAA
CCACTTTACTTTGAATCCAGTCGCTATTTTGTATCAGAAATGAACAAAGTTCTGGAAGAA
GAGGCTGCGACCACACATGTGGGCTTGATTACAACAGGAGATAGCTTTATCGCAAGTGAA
GAAAAAGTCGCAGCGATTCGGGAGCATTTCCCAGAAGTCTTGGCAGTTGAGATGGAAGGG
GCAGCCGTTGCCCAAGCCGCACACGCAGCTGGACGTCCTTTCATGGTCATTCGAGCTATG
AGTGATACGGCAGATCACGTAGCCAATATCTCTTTTGATGAATTTATCGTAGAGGCTGGC
GAACGCTCTGCTCGAACCCTGATTACCTTCTTGAAGAGATTGGTGT
>45
TCCTTACAAAGCTTTGAACGCTTCGATAAACACTTCCGAACGGTGCGCATAGCCTTTGAA
CATATCGAAGCTCGCACAAGCCGGGCTGAGCAACACGATATCACCCGCTTCGGCTTGGGC
ATATGCCGTCTGAACGGCTTCTTCCAAAGTGGCGCAGTCGGTCATATTCAGACCGCAGCC
GTCCAAATCGCGGCGGATTTGCGGCGCATCGACACCGATCAGGAACACACCTTTTGCCTT
GC

TableB: Primers were used to fill the gaps between the contigs of A2 virus

Primers' names	Primer sequence 5' to 3'	Reverse Complement
MF21 F	GGGCGGCGACTTGTACGCCC	
B20F	CACAAGCGAACGGTGAGCCG	
Sa10F	GCCGGTAGAACAGCCAGCCG	
MF 13 F	GCCCACGTCTGCGCCTCAAC	
Sa33 F	GAGGCCGCGCTTGGCGGTGC	
SA 18 F	GGGTCTTCCGCGCCTGAGAC	
MF 15 F	CCGTGTCGGCGTGAACGGCG	
SA 50 F	CGCCTGCTCACACGGGCATG	
MF 17 F	GTGTCGGCGGATGGTCGGCG	
Mf 16 F	CGGCGGCGGCTGGTGGGTTG	
Sa36 F	CCAATACAAGCGCGGCCGTG	
SA 60 F	CAGCGGCACGTCTCGACTGC	
Sa39F	GGTGTTACCGGCGGTGGTGG	
E5F1	CGGCCGCCTGCAGGTCGACC	
E8F	CCAACTGCTCTGCCGTGTAC	
Sa32F	GGCCGCCTCGCCTACCGCTC	
D10F	GCTGAACTTGGCAAGCGTGG	
Sa46F	CTAACCTCGCTTCGGGCTTC	
E19F	TTGTTTGGCGTGTAGCCCAA	
E1F	AAACCTCGTCCCCTACATC	
MF10R		CGGCCTGCCAGTCAACAACC
MF20R		CAGCGTTGCTGCCATGCAAG
Mf1 R		GCACCTGAAGCAGGCACGGC
Sa33R		GCCAAGCGCGGCCTCTGCCC
MF31R		CTCGTCGCGGTGCAAGTGCC
B17R		GCGACCGTGTGCTTGGCGAC
Sa50R		CCATGCCCGTGTGAGCAGGC
Sa32R		GAGCGGTAGGCGAGGCGGCC
Sa60R		GACGTGTAGCAGATCGGCGC
B10R		GCCGCGAACGGCTCGGCGCG
Mf 17 R		GCCTGACCCATCGTCATAGG
Mf 16 R		CAACCCACCAGCCGCCGCCG
6aR		GTGTCGGCGGATGGTCGGCG
6aF		CGCCGCCTGATTGCCCTGCC
Sa60F		CAGCGGCACGTCTCGACTGC
Sa50R2		GTCGGCAGCGCGTGGTCGAG

Continuing of table B:

Primers' names	Primer sequence 5' to 3'	Reverse Complement
6aR3		GCACTGGGACGCGCGTTGTC
E7R1		GCGGCTGTCCTGCTCCGTTC
E5R1		CCCCATGCCATGTGTCTCAC
E7R2		CGGCCACGGCGATATTGTAG
Sa53R		CACATTGCGAGAGTGGGCGC
MF33R		CCGCCGCTTCCTCGAAGGTC
E5R		GACCGGCCGTCTATCACATC
Sa46R		CTAACCTCGCTTCGGGCTTC
E19R		TACGCCAACATGCCTCACA
E1R		ATCTGAGTGTGCGAAACCAA

Table C: Databases significant matches to the 16S rRNA gene sequence of the OIB strain

Accession	Description	Query coverge	E-value	Max ident
EU794238.1	Uncultured Neisseria sp. clone EMP_C13 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF512007.1	Uncultured bacterium clone P1D1-725 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511998.1	Uncultured bacterium clone P1D1-762 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511994.1	Uncultured bacterium clone P1D1-709 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511992.1	Uncultured bacterium clone P1D1-529 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511980.1	Uncultured bacterium clone P1D1-708 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511956.1	Uncultured bacterium clone P1D1-517 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511878.1	Uncultured bacterium clone P1D1-543 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
AM697049.1	Uncultured bacterium partial 16S rRNA gene, isolate BF0001B075	100%	0.0	99%
EF511922.1	Uncultured bacterium clone P1D1-538 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511911.1	Uncultured bacterium clone P1D1-741 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511959.1	Uncultured bacterium clone P1D1-558 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511943.1	Uncultured bacterium clone P1D1-499 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511938.1	Uncultured bacterium clone P1D1-705 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511937.1	Uncultured bacterium clone P1D1-527 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511935.1	Uncultured bacterium clone P1D1-718 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511897.1	Uncultured bacterium clone P1D1-727 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
AJ786809.1	Neisseria sp. R-22841 partial 16S rRNA gene, isolate R-22841	100%	0.0	99%
EF511986.1	Uncultured bacterium clone P1D1-549 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511983.1	Uncultured bacterium clone P1D1-710 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511925.1	Uncultured bacterium clone P1D1-495 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
EF511953.1	Uncultured bacterium clone P1D1-681 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
AM420191.1	Uncultured Neisseria sp. partial 16S rRNA gene, clone 502D04(oral)	100%	0.0	99%
AM697371.1	Uncultured bacterium partial 16S rRNA gene, isolate BF0002C068	100%	0.0	99%
AM697129.1	Uncultured bacterium partial 16S rRNA gene, isolate BF0002B079	100%	0.0	99%
EF511988.1	Uncultured bacterium clone P1D1-484 16S ribosomal RNA gene, partial sequence	99%	0.0	99%
AM697034.1	Uncultured bacterium partial 16S rRNA gene, isolate BF0001B060	100%	0.0	99%
AY963348.1	Uncultured bacterium clone AH55 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
DQ279353.1	Neisseria sp. TM10_4 16S ribosomal RNA gene, partial sequence	99%	0.0	99%
AM420196.1	Uncultured Neisseria sp. partial 16S rRNA gene, clone 502G08(oral)	100%	0.0	99%
EF512003.1	Uncultured bacterium clone P1D1-542 16S ribosomal RNA gene, partial sequence	97%	0.0	99%
EF511936.1	Uncultured bacterium clone P1D1-550 16S ribosomal RNA gene, partial sequence	97%	0.0	99%
DQ409137.1	Neisseria sp. J01 16S ribosomal RNA gene, partial sequence	100%	0.0	99%
AY138232.1	Uncultured Neisseriaceae bacterium Sto1-2 16S ribosomal RNA gene, complete sequence	100%	0.0	99%
EF511971.1	Uncultured bacterium clone P1D1-512 16S ribosomal RNA gene, partial sequence	97%	0.0	99%
EF511861.1	Uncultured bacterium clone P1D1-711 16S ribosomal RNA gene, partial sequence	97%	0.0	99%
EF511923.1	Uncultured bacterium clone P1D1-730 16S ribosomal RNA gene, partial sequence	97%	0.0	99%
EF511903.1	Uncultured bacterium clone P1D1-716 16S ribosomal RNA gene, partial sequence	97%	0.0	99%
AJ239295.1	Neisseria perflava 16S rRNA gene (partial), strain U15	97%	0.0	99%
AJ239279.1	Neisseria mucosa 16S rRNA gene (partial), strain M5	97%	0.0	99%
L06168.1	Neisseria flavescens 16S ribosomal RNA	100%	0.0	98%
EF511915.1	Uncultured bacterium clone P1D1-500 16S ribosomal RNA gene, partial sequence	97%	0.0	99%
AJ239280.1	Neisseria flavescens 16S rRNA gene (partial), strain LNP444	97%	0.0	99%
AM420167.1	Uncultured Neisseria sp. partial 16S rRNA gene, clone 501C06(oral)	99%	0.0	98%
AF479578.1	Neisseria subflava NJ9703 16S ribosomal RNA gene, partial sequence	96%	0.0	99%
AM697198.1	Uncultured bacterium partial 16S rRNA gene, isolate BF0001C039	99%	0.0	98%
AY831725.1	Neisseria cinerea 16S ribosomal RNA gene, partial sequence	99%	0.0	98%
AF310565.1	Neisseria meningitidis strain M2786 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310564.1	Neisseria meningitidis strain M2788 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310547.1	Neisseria meningitidis strain M26 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF398310.1	Neisseria meningitidis strain M7724 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF398308.1	Neisseria meningitidis strain M7509 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
EF511868.1	Uncultured bacterium clone P1D1-534 16S ribosomal RNA gene, partial sequence	97%	0.0	99%
AF382294.1	Neisseria meningitidis strain M7890 16S ribosomal RNA gene, partial sequence	100%	0.0	98%

Continuing of table C:

Accession	Description	Query coverge	E-value	Max ident
AF310561.1	Neisseria meningitidis strain M2783 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310560.1	Neisseria meningitidis strain M2795 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310544.1	Neisseria meningitidis strain M4015 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310347.1	Neisseria meningitidis strain M6304 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF398275.1	Neisseria meningitidis strain M812 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF399845.1	Neisseria meningitidis strain M8368(F10) 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF398311.1	Neisseria meningitidis strain M7931 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382292.1	Neisseria meningitidis strain M7591 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382291.1	Neisseria meningitidis strain M7590 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310471.1	Neisseria meningitidis strain M7107 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310467.1	Neisseria meningitidis strain M7150 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310463.1	Neisseria meningitidis strain M7187 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310462.1	Neisseria meningitidis strain M7184 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310442.1	Neisseria meningitidis strain M7322 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310430.1	Neisseria meningitidis strain M7089A 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310429.1	Neisseria meningitidis strain M7089B 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310425.1	Neisseria meningitidis strain M7115 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310422.1	Neisseria meningitidis strain M7123 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310421.1	Neisseria meningitidis strain M7104 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310420.1	Neisseria meningitidis strain M7105 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310419.1	Neisseria meningitidis strain M7106 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310417.1	Neisseria meningitidis strain M7149 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310408.1	Neisseria meningitidis strain M7034 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF310407.1	Neisseria meningitidis strain M7035 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF398315.1	Neisseria meningitidis strain M8172 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF398314.1	Neisseria meningitidis strain M8171 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382272.1	Neisseria meningitidis strain M8068 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382271.1	Neisseria meningitidis strain M8073 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382270.1	Neisseria meningitidis strain M8074 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382268.1	Neisseria meningitidis strain M8047 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382267.1	Neisseria meningitidis strain M8045 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382260.1	Neisseria meningitidis strain M8012 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382259.1	Neisseria meningitidis strain M8063 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382258.1	Neisseria meningitidis strain M8037 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382255.1	Neisseria meningitidis strain M8049 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382254.1	Neisseria meningitidis strain M8065 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382253.1	Neisseria meningitidis strain M8064 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382252.1	Neisseria meningitidis strain M8028 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382250.1	Neisseria meningitidis strain M8034 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382246.1	Neisseria meningitidis strain M7887 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382244.1	Neisseria meningitidis strain M7895 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382243.1	Neisseria meningitidis strain M7903 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382241.1	Neisseria meningitidis strain M7857 16S ribosomal RNA gene, partial sequence	100%	0.0	98%
AF382240.1	Neisseria meningitidis strain M7854 16S ribosomal RNA gene, partial sequence	100%	0.0	98%