Studies on Phages Infective for Mycolic Acid Producing Bacteria that Cause Foams in Activated Sludge Systems

Submitted by

Zoe Anne Dyson

Bachelor of Computing/Bachelor of Applied Science (Honours)

La Trobe University

A thesis submitted in total fulfilment

of the requirements for the degree of

Doctor of Philosophy

School of Molecular Sciences

College of Science, Health and Engineering

La Trobe University

Bundoora, Victoria 3086

Australia

April 2015

Table of Contents

List of figures

List of tables

List of abbreviations used

aa	Amino acid
Abi	Abortive infection
ACN	Acetonitrile
ADP	Adenosine diphosphate
ATCC	American Type Culture Collection
BDH	British Drug House
bp	Base pairs
BLAST	Basic Local Alignment Search Tool
BOD	Biological Oxygen Demand
CCUG	Culture Collection University of Göteborg
CDS	Coding Sequence
CID	Collision Induced Dissociation
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CSH	Cell Surface Hydrophobicity
C-terminus	Carboxy-terminus
DAS	Dense Alignment Surface
DOC	Dissolved Organic Carbon
DNA	Deoxyribonucleic acid
DNAse	Deoxyribonuclease
ddH_20	Double distilled water
dsDNA	Double stranded Deoxyribonucleic acid
DSM/DSMZ	Deutsche Sammlung von Mikroorganismen und Zellkulturen

EBPR	Enhanced Biological Phosphorus Removal
EFM	Epifluorescence Microscopy
EPS	Exopolymeric Substance Production
E value	Expect value
FISH	Fluorescence *in situ* hybridisation
GALO	*Gordonia amarae*-like organisms
G+C content	Guanine+cytosine content
GI	GenInfo Identifier
HAV	Hepatitis A Virus
H-NS	Heat-stable nucleoid structuring
HTH	Helix-turn-helix
IHF	Integration Host Factor
Kb/Kbp	Kilobase pairs
LKB	Liljeholmen, Kema, and Bryggerierna
LIMS	La Trobe Institute of Molecular Sciences
MDS	Modification dependant systems
MS	Mass Spectroscopy
N	Nitrogen
NBRC	NITE Biological Resource Centre
NCBI	National Centre for Biotechnology Information
NITE	National Institute of Technology and Evaluation (Japan)
NSW	New South Wales
nt	Nucleotide
N-terminus	Amino terminus

ORF	Open reading frame
OTU	Operational Taxonomic Unit
P	Phosphorus
PAO	Polyphosphate Accumulating Organisms
PCR	Polymerase Chain Reaction
PDE	Polysaccharide Depolymerase Enzyme
PEG	Polyethylene Glycol
PFGE	Pulse Field Gel Electrophoresis
PFU	Plaque forming unit
PHA	Poly-beta-hydroxyalkanoate
PHAST	PHAge Search Tool
phageFISH	Phage Fluorescence *in situ* hybridisation
PTLO	Pine tree like organisms
PYCa	Peptone Yeast Extract Calcium
R2A	Reasoner's 2A
RAS	Return Activated Sludge
RM	Restriction Modification
RNA	Ribonucleic acid
RNAse	Ribonuclease
rRNA	Ribosomal Ribonucleic Acid
Sie	Superinfection Exclusion
SS	Suspended Solids
SSU	Small Subunit
TEM	Transmission Electron Microscopy

TMP	Tape Measure Protein
tmRNA	Transfer messenger RNA
tRNA	Transfer RNA
tRNA-Asn	Asparagine-tRNA
tRNA-Asp	Aspartic acid-tRNA
tRNA-Glu	Glutamic acid-tRNA
tRNA-Ile	Isoleucine-tRNA
tRNA-Lys	Lysine-tRNA
tRNA-Met	Methionine-tRNA
tRNA-Pro	Proline-tRNA
tRNA-Ser	Serine-tRNA
tRNA-Thr	Threonine-tRNA
tRNA-Trp	Tryptophan-tRNA
tRNA-Tyr	Tyrosine-tRNA
UCT	University of Cape Town
UV	Ultraviolet
VIP	Vegetative Insecticidal Protein
VLP	Virus Like Particles
WGS	Whole Genome Sequence

Summary

This study aimed to isolate phages infective for the Mycolata, which includes the genera *Gordonia, Mycobacterium, Nocardia, Rhodococcus, Skermania,* and *Tsukamurella.* Mycolata are the etiological agents of activated sludge foaming, so phages lytic for these were sought and extensively characterised, and their potential as biological control agents explored. A summary of the major outcomes follows:

- Three Tsukamurella phages TIN2, TIN3, and TIN4 that were genetically very similar to each other and to *Gordonia* phage GTE7 were isolated. Mass spectroscopy revealed TIN4 phage structural genes unexpectedly located within the DNA replication gene module. Phage TIN3 produced smaller plaques on average than that of TIN4 on the same strain of *T. inchonensis* potentially corresponding to mutations in its lysis gene module. All three phages are attractive choices for phage mediated biocontrol of activated sludge foaming

- The first ever phage lytic for *Skermania piniformis,* a major cause of foams, was isolated and genetically characterised. Its genome was very different to all other phage genomes sequenced, but only lysed some of the *S. piniformis* strains screened, compromising its value in foam biocontrol

- Nine phages infective for the genus *Gordonia* were isolated and characterised, with some obtained from environments other than activated sludge. Genome sequencing revealed evidence of spontaneous induction events occurring despite co-infection with lytic viruses and the presence of genes encoding toxins. Three of these *Gordonia* phages appear suitable for foam biocontrol

- Bioinformatic analysis suggested prophages are common among the Mycolata. Induction experiments recovered three temperate phages, GAL1 from *Gordonia alkanivorans*, GMA1 from *Gordonia malaquae*, and TPA4 from *Tsukamurella paurometabola,* but none from multiple *Gordonia amarae* strains. *In silico* analysis revealed phage defence systems capable of inhibiting phage infections in one strain of *Gordonia amarae*

Statement of authorship

Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis submitted for the award of any other degree or diploma. No other person's work has been used without due acknowledgment in the main text of the thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution.

Zoe Anne Dyson

April 30th, 2015

List of publications associated with this thesis

Dyson, Z. A., Tucci, J., Seviour, R. J. & Petrovski, S. (2015). Three of a kind: Genetically similar *Tsukamurella* phages TIN2, TIN3, and TIN4. *Appl and Enviro Microbiol* **81**, 6767-72.

Dyson, Z. A., Tucci, J., Seviour, R. J. & Petrovski, S. (2015). Lysis to kill: evaluation of the lytic abilities, and genomics of nine bacteriophages infective for *Gordonia* spp. and their potential use in activated sludge foam bio-control. *PLoS ONE* **10**, e0134512.

Dyson, Z. A., Tucci, J., Seviour, R. J. & Petrovski, S. (2015, *in press*). Isolation and characterisation bacteriophage SPI1, which infects the activated-sludge-foaming bacterium *Skermania piniformis*. *Arch Virol.*

Dyson, Z. A., Brown, T. L., Farrar, B., Doyle, S., Tucci, J., Seviour, R. J. & Petrovski, S. (2015, *submitted manuscript*). Locating and activating molecular 'time bombs': can Mycolata prophages be selectively induced *en masse* to biologically control activated sludge foaming? *Front Microbiol*.

List of publications of work not included in this thesis

Petrovski, S., Dyson, Z. A., Seviour, R. J. & Tillett, D. (2012). Small but sufficient: the *Rhodococcus* phage RRH1 has the smallest known *Siphoviridae* genome at 14.2 kilobases. *J Virol* **86**, 358-363.

Petrovski, S., Dyson, Z. A., Quill, E. S., McIlroy, S. J., Tillett, D. & Seviour, R. J. (2011). An examination of the mechanisms for stable foam formation in activated sludge systems. *Water Res* **45**, 2146-2154.

List of conference presentations

Dyson, Z.A., Petrovski, S., Tucci, J., & Seviour, R. J. (2015). Potential phage based methods to control foam in wastewater treatment systems. Oral presentation. *The Australian Society for Microbiology conference*. Canberra.

Dyson, Z.A., Petrovski, S., Tucci, J., & Seviour, R. J. (2014). Genomics of lytic and lysogenic phages of selected *Actinobacteria*. Oral presentation. *COMBINE symposium*. Melbourne.

Dyson, Z. A., Brown, T. L., Petrovski, S., Tucci, J., & Seviour, R.J. (2014). Induction and genomic characterisation of *Gordonia alkanivorans* prophage GAL1. Poster presentation. *COMBINE symposium*. Melbourne.

Dyson, Z. A., Brown, T. L., Petrovski, S., Tucci, J., & Seviour, R.J. (2014). Induction and genomic characterisation of *Gordonia alkanivorans* prophage GAL1. Poster presentation. *The Australian Society for Microbiology conference*. Melbourne.

Dyson, Z.A., Petrovski, S., Tillett, D., & Seviour, R.J. (2013). 454 pyrosequencing of Actinophage GTE10 reveals high levels of nucleotide sequence identity between phages of the Mycolata. Poster presentation. *BioInfoSummer conference*. Adelaide.

Dyson, Z.A., Petrovski, S., Seviour, R.J. & Tillett, D. (2013). Actinophages isolated for foaming Mycolata show high levels of nucleotide sequence identity. Poster presentation. *The Australian Society for Microbiology conference*. Adelaide.

Dyson, Z.A., Petrovski, S., Seviour, R.J., & Tillett, D. (2012). Characterisation and Application of Actinophages to Combat Activated Sludge Foaming. Poster presentation. *The Australian Society for Microbiology conference*. Brisbane.

Dyson, Z.A., Petrovski, S., Seviour, R.J. & Tillett, D. (2012). Characterisation and Application of Actinophages to Combat Activated Sludge Foaming. Poster Presentation. *Exploiting bacteriophages for bioscience, biotechnology, and medicine conference*. London.

Dyson, Z.A., Petrovski, S., Seviour, R.J. & Tillett, D. (2011). Genome sequence and characterisation of the *Gordonia terrae* bacteriophage GTE6 and its potential use as a foam bio-control agent. Poster presentation. *The Australian Society for Microbiology conference.* Hobart.

Dyson, Z.A., Petrovski, S., Seviour, R.J. & Tillett, D. (2011). Genome sequence and characterisation of the Gordonia terrae bacteriophage GTE6 and its potential use as a foam bio-control agent. Poster Presentation. *La Trobe Institute for Molecular Science conference.* Melbourne.

Petrovski, S., Dyson, Z.A., Seviour, R.J. & Tillett, D. (2011). Control of biological foaming with bacteriophages. Poster Presentation. *La Trobe Institute for Molecular Science conference.* Melbourne.

Dyson, Z.A., Petrovski, S., Seviour, R.J. & Tillett, D. (2010). *In situ* Detection of Bacteria using Fluorescently Labelled Bacteriophages. Poster presentation. *The Australian Society for Microbiology conference*. Sydney.

Petrovski, S., Dyson, Z.A., Seviour, R.J. & Tillett, D. (2010). Control of biological foaming with bacteriophages. Poster presentation. *The Australian Society for Microbiology conference*. Sydney.

Petrovski, S., Dyson, Z.A., Seviour, R.J. & Tillett, D. (2009) Bio-control of foaming in activated sludge using bacteriophages. Poster presentation. *Activated Sludge Population Dynamics 5 conference*. Aalborg.

Dyson, Z.A., Petrovski, S., Seviour, R.J. and Tillett, D. (2009). Isolation of novel, broad host range bacteriophages infective for *Gordonia* spp. and their potential use as foaming bio-control agents. Poster presentation. *Activated Sludge Population Dynamics 5 conference*. Aalborg.

List of awards received during PhD candidature

Best oral presentation, COMBINE Symposium, 2014:

Dyson, Z.A., Petrovski, S., Tucci, J., & Seviour, R. J. (2014). Genomics of lytic and lysogenic phages of selected *Actinobacteria*. Oral presentation. *COMBINE Symposium*. Melbourne.

Second place, 3 Minute Thesis Competition, School of Molecular Sciences, La Trobe University, 2014:

Dyson, Z.A. (2014). Phage vs. Foam. Oral Presentation. *La Trobe University 3 Minute Thesis Competition, School of Molecular Sciences Division.* Melbourne.

Best poster in Environmental Microbiology division, Australian Society for Microbiology conference, 2013:

Dyson, Z.A., Petrovski, S., Seviour, R.J. & Tillett, D. (2013). Actinophages isolated for foaming Mycolata show high levels of nucleotide sequence identity. Poster presentation. *The Australian Society for Microbiology conference*. Adelaide.

Best poster in Environmental Microbiology division, Australian Society for Microbiology conference, 2012:

Dyson, Z.A., Petrovski, S., Seviour, R.J., & Tillett, D. (2012). Characterisation and Application of Actinophages to Combat Activated Sludge Foaming. Poster presentation. *The Australian Society for Microbiology conference*. Brisbane.

Best poster, Exploiting bacteriophages for bioscience, biotechnology, and medicine conference, 2012:

Dyson, Z.A., Petrovski, S., Seviour, R.J. & Tillett, D. (2012). Characterisation and Application of Actinophages to Combat Activated Sludge Foaming. Poster Presentation. *Exploiting bacteriophages for bioscience, biotechnology and medicine conference*. London.

Acknowledgements

I have been very fortunate to have the opportunity to pursue and complete my PhD studies. I have many people to thank for their assistance, mentorship, and support throughout my candidature.

Firstly, I would like to thank my supervisor Professor Robert Seviour for his patience, advice, genuine enthusiasm, and mentorship throughout my PhD studies as well as for the many hours he has invested in spent encouraging, guiding, and assisting me to develop as a research scientist.

I would also like to thank my co-supervisor Dr. Steve Petrovski for his continued encouragement, guidance, and enthusiasm throughout my PhD, as well as the time he took to teach me valuable laboratory techniques. I would also like to thank my other co-supervisor Associate Professor Joseph Tucci for his support, enthusiasm, and for taking the time to offer his mentorship and assistance with my professional development as a researcher. I would also like to thank Dr. Daniel Tillett for technical advice and useful discussions. Thankyou to Dr. Stephen Doyle for taking the time to teach me prepare samples and sequence them with the Illumina MiSeq. Thankyou to Dr. Robert Glashier, Dr. Sabine Wilkens, Ms. Glenys Shirley, Dr. Alexander Fink, Dr. Eric Hanssen, and Dr. Christopher Adda for their assistance with Transmission Electron Microscopy.

I would like to say a very big thankyou to the staff at the La Trobe University Bendigo Campus for their support, especially Dr. Mike Angove, Dr. Sabine Wilkens, Dr. Michelle Gibson, Dr. Christine Kettle, and Ms. Elizabeth Seviour for their encouragement and advice throughout my PhD. Thankyou to the technical staff in the School of Applied Science and Pharmacy, especially to Dot Andison and Barb Udale who helped me solve many problems.

Thankyou to all my fellow PhD candidates at Bendigo, especially the members of the Biotechnology Research Centre, who I have had the pleasure of studying with over the last few years, for their help, advice, and friendship. I had a lot of fun working with you all.

Finally, thankyou to my friends and family who have listened, offered advice, and supported me throughout my studies. I would like to thank Carley Morrison, Lena Ackermann, Hayden Maggs, Jessica Cola, and the entire Cola family for their support. Thanks to Lachlan Speirs for many useful scientific and non-scientific discussions. Thankyou to the ladies who I met at Dragon City for their support, and the biggest ever thankyou to my mother Rhonda and my father Philip for their unwavering support, encouragement, advice, and love.

To my mother Rhonda and my father Philip.

“The field of bacterial viruses is a fine playground for

serious children who ask ambitious questions”

- Max Delbrück

1. Introduction

1.1. Activated sludge systems

Effective wastewater treatment is required in order to reduce levels of organic and inorganic compounds in the influent, and pathogens (Schwartzbrod *et al.*, 2002; Watanabe *et al.*, 1997; Wyn-Jones & Sellwood, 2001) in order to minimise incidences of waterborne diseases (Ashbolt, 2004; Schwartzbrod *et al.*, 2002; Theron & Cloete, 2002; Watanabe *et al.*, 1997; White & Godfree, 1985; Wyn-Jones & Sellwood, 2001). Many wastewater treatment systems exist but the most widespread are activated sludge systems which are found around the world (Bitton, 2005). The first pilot activated sludge wastewater treatment plant was developed by Lockett & Ardern in 1914 (Lindrea & Seviour, 2002; Seviour, 2010a). There are two phases in this system, first an aerated reactor followed by separation of the biomass (sludge) and liquid phases in a clarifier as shown in Figure 1.1 (Lindrea & Seviour, 2002; Seviour & Blackall, 1999). Initially, raw sewage influent consisting of faecal material, xenobiotics, heavy metals, surfactants, and microbes (Bitton, 2005; Kraigher *et al.*, 2008; Lindqvist *et al.*, 2005; Seviour & Blackall, 1999; Seviour *et al.*, 2010; Ternes, 1998) flows into an aeration basin/tank where it comes into contact with recycled biomass from the clarifier of mainly bacteria and protozoa organised as three-dimensional 'flocs' (Bitton, 2005; Lindrea & Seviour, 2002; Seviour, 2010a; Seviour & Blackall, 1999).

Aeration supplied either by diffusion or mechanical systems supplies oxygen to these microbial communities, keeping the flocs in constant contact with metabolisable substrates which are aerobically respired resulting in the formation of carbon dioxide, and substrates for cell growth (Bitton, 2005; Seviour, 2010a). The quality of final effluent from these systems relies entirely on the settling properties of the flocs in the clarifier, so that a clear supernatant is produced (Lindrea & Seviour, 2002; Seviour, 2010a). A portion of the biomass generated during this treatment process is recycled to inoculate the raw influent (as shown in Figure 1.1) and is termed return activated sludge (RAS). Thus, instead of the previous approach of disposing the biomass leaving the reactor, its recycling, containing the microbial populations best suited to treat the raw incoming sewage is the key feature of this process, permitting faster treatment of wastes in a smaller foot print (Lindrea & Seviour, 2002; Seviour, 2010a).

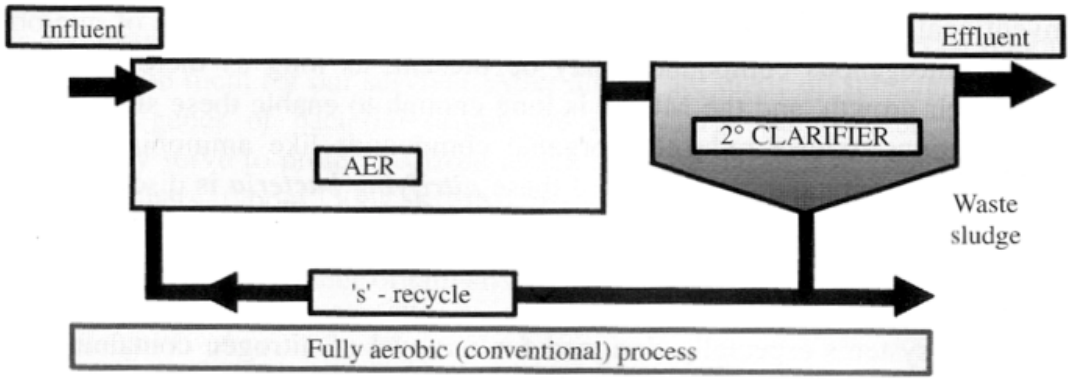

Figure 1.1 Flow diagram for conventional activated sludge wastewater treatment plant reproduced from Seviour *et al.* (2010).

AER indicates aerobic conditions, and 's'-recycle indicates return activated sludge (RAS).

These conventional plants remove organic carbonaceous material adequately, but not adequate levels of nitrogen (N) and phosphorus (P), both undesirable environmental pollutants, causing the eutrophication of water bodies receiving treated sewage from them (Seviour & Nielsen, 2010). The inherent self purification mechanisms fail and community homeostasis breaks down, where blooms of often toxic cyanobacteria develop (Seviour *et al.*, 2003). As discharge license requirements became more stringent, plants were developed, often empirically, which could reduce N and P levels to environmentally acceptable levels by imposing selective pressures encouraging certain bacterial populations (Orhon & Artan, 1994; Seviour *et al.*, 2003; Seviour *et al.*, 2010). Although it should be noted that some nitrogen removal does occur in conventional activated sludge systems, these do leave high residual phosphorus levels, thus more specialised systems have been designed such as the Ludzack-Ettinger plant configuration shown in Figure 1.2 (Daims & Wagner, 2010; Seviour *et al.*, 2003; Seviour *et al.*, 2010).

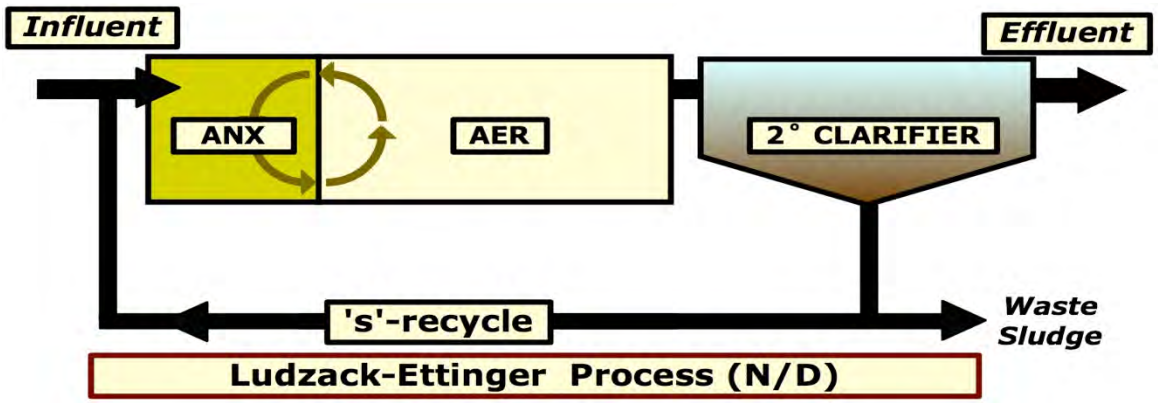

Figure 1.2 Flow diagram for Ludzack-Ettinger Nitrogen removal activated sludge plant configuration reproduced from Seviour *et al.* (2010).

ANX indicates anoxic conditions, AER indicates aerobic conditions, and 's'-recycle indicates return activated sludge (RAS).

Phosphorus removal is biologically achieved in activated sludge systems by encouraging the intracellular accumulation of phosphorus in polyphosphate accumulating organisms (PAO) such as *Candidatus* 'Accumulibacter phosphatis' (henceforth referred to as *Accumulibacter*) in high amounts as polyphosphate granules (McMahon *et al.*, 2010; Seviour *et al.*, 2003). This is achieved by manipulating the PAO metabolism; initially anaerobic (feed) conditions are used to allow for their preferential assimilation of volatile fatty acids such as acetate to build up intracellular carbon and energy reserves of poly-beta-hydroxyalkanoate (PHA) where polyphosphate is degraded to provide energy and phosphate is released into the bulk liquid (McMahon *et al.*, 2010). Then, aerobic (famine) conditions where extracellular soluble substrates are rare allows for carbon utilisation for both cellular replication and the uptake of external phosphate, which is converted to intracellular polyphosphate (McMahon *et al.*, 2010). These conditions are cycled to achieve optimum phosphate accumulation and to maintain dominance of the PAO in these systems (McMahon *et al.*, 2010; Seviour *et al.*, 2003). The subsequent removal of PAO cells from the process during sludge wasting also removes the phosphorus, and this process is referred to as Enhanced Biological Phosphorus Removal (EBPR), and a modified University of Cape Town (UCT) configuration of this type of plant is shown in Figure 1.3 (McMahon *et al.*, 2010; Seviour *et al.*, 2003; Seviour *et al.*, 2010).

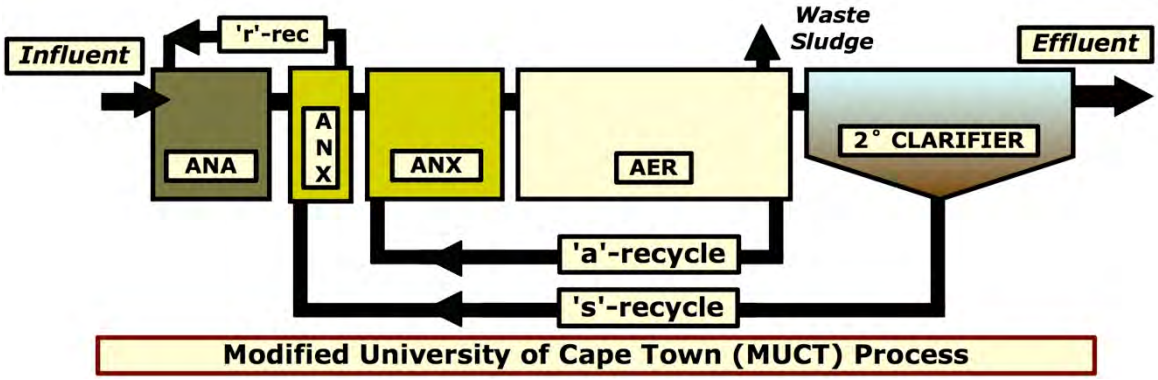

Figure 1.3 Flow diagram for modified UCT activated sludge Enhanced Biological Phosphorous Removal (EBPR) system reproduced from Seviour *et al.* (2010).

ANX indicates anoxic conditions, AER indicates aerobic conditions, ANA indicates anaerobic conditions, 's'-recycle indicates return activated sludge (RAS), and 'a'-recycle indicates aerobic RAS to the anoxic reactor.

However, despite their advancements and advantages over other treatment systems, activated sludge systems frequently succumb to solid-liquid separation problems caused by the proliferation of filamentous bacteria, giving rise to foaming and bulking (Seviour, 2010b), both of which are global problems.

1.2. Activated sludge bulking and foaming

Bulking occurs when filamentous bacteria required for good floc formation, extend from flocs into the bulk solution, sometimes forming inter-bridges between flocs, causing poor compaction and settling of sludge in the clarifier (Seviour, 2010b). Activated sludge foaming differs from the ephemeral white foams produced during plant start up, or in the presence of slowly biodegradable surfactants. It appears as a heavy, highly stable, viscous, grey to brown layer on the liquid surface of aeration basins as shown in Figure 1.4 (de los Reyes III, 2010; Jenkins *et al.*, 2003; Soddell, 1999). Severe foaming incidences, like that shown in Figure 1.4, can lead to the spread of this foam to other parts of the plant such as clarifiers, and into surrounding regions of the plant such as walkways and lawns (de los Reyes III, 2010; Jenkins *et al.*, 1993).

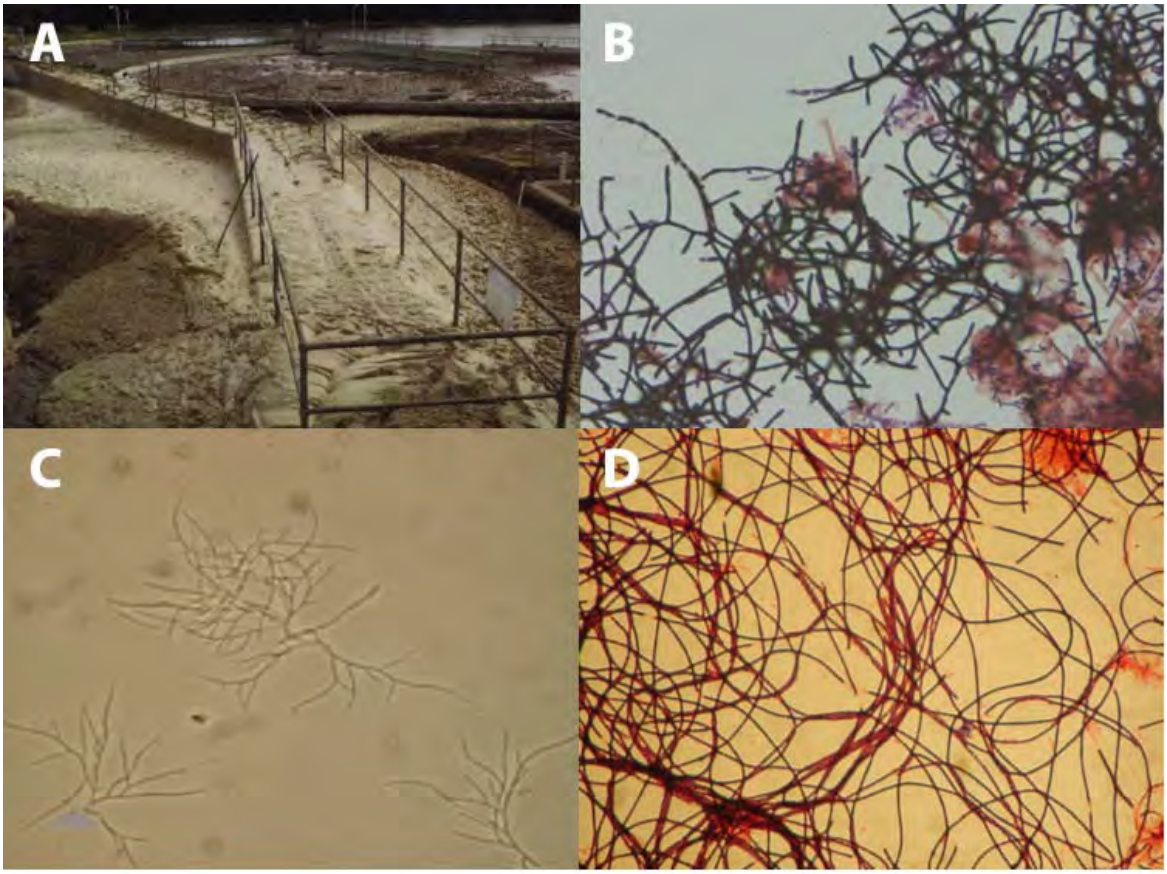

Figure 1.4 Severe foaming incident, *Gordonia amarae*-like organism (GALO) morphology, Pine Tree like organism (PTLO) morphology, and *Candidatus* 'Microthrix parvicella' morphology.

(A) severe foaming incident at an activated sludge wastewater treatment plant located in Bendigo, Victoria, Australia (image courtesy of Prof. R. J. Seviour); (B) Gordonia amarae-like organism morphology (image courtesy of Ms. E. Seviour); (C) Pine tree like organism (Skermania piniformis) morphology (image courtesy of Ms. E. Seviour); (D) Candidatus 'Microthrix parvicella' morphology (image courtesy of Ms. E. Seviour).

The first known report of activated sludge foaming was in 1969 where it was described as 'a heavy growth of floating solids' (Anonymous, 1969; de los Reyes III, 2010). Many later reports of foaming have described incidences in the United States, United Kingdom, Denmark, the Netherlands, France, the Czech Republic, Italy, Austria, Germany, Switzerland, South Africa, Saudi Arabia, Japan, Hong Kong, and Australia (Blackall *et al.*, 1991; de los Reyes III, 2010; de los Reyes III *et al.*, 1997; Mielczarek *et al.*, 2012; Seviour *et al.*, 1990; Seviour *et al.*, 1994; Soddell, 1999, 2002), demonstrating the widespread nature of this problem. Furthermore, the proportion of plants affected worldwide is also considerable (see Table 1.1). Surveys conducted

by Seviour *et al.* (1990), Blackall *et al.* (1991), and Seviour *et al.* (1994) indicated that between 51 to 92% of Australian activated sludge plants had undergone foaming events (Bitton, 2005) during their sampling periods.

1.2.1. Etiological agents of activated sludge foams

When foam forms, the bacterial communities responsible can be examined by microscopy and staining, or via their isolation by micromanipulation and subsequent *in situ* characterisation studies (Seviour & Nielsen, 2010). Methods involving their *in situ* identification with antibodies and fluorescence *in situ* hybridisation (FISH) or membrane DNA:DNA hybridisations (de los Reyes III, 2010) have also been used. They have revealed a wide foam biodiversity (Soddell, 2002), and the genera identified as 'foaming bacteria' include several *Actinobacteria* among which are *Candidatus* 'Microthrix parvicella' (henceforth referred to as *Microthrix*), and members of the Mycolata, a group of Gram-positive, filamentous, mycolic acid producing, hydrophobic bacteria. The Mycolata include members of the genera *Corynebacterium, Dietzia, Gordonia, Millisia, Mycobacterium, Nocardia, Rhodococcus, Segniliparus, Skermania, Tsukamurella*, and *Williamsia* (de los Reyes III, 2010). The hydrophobicity of these Mycolata is assumed to be because of the unique presence of long chain hydroxylated mycolic acids in their outer cell membranes (de los Reyes III, 2010; Soddell, 2002).

Microthrix is a Gram-positive, slow growing long, thin, unbranched, unsheathed, spaghetti-like, septate filament (Figure 1.4). It does not synthesise mycolic acids but instead accumulates large intracellular stores of triacyl glycerides, which impart hydrophobicity to the filaments (Rossetti *et al.*, 2005). While isolates have been obtained (Rossetti *et al.*, 2005), this organism is not easily maintained in pure culture. Therefore, this discussion will focus on the Mycolata (Blackall *et al.*, 1991; Seviour *et al.*, 1990; Seviour *et al.*, 1994), the foaming organisms used in this study.

The *Gordonia amarae*-like organisms (GALO) are recognisable by their characteristic right-angled branching filamentous morphology (Figure 1.4) and this morphotype is known now to embrace members of other Mycolata genera including *Dietzia*, *Gordonia*, *Nocardia*, *Rhodococcus*, and *Tsukamurella* (Soddell, 1999, 2002). The pine tree like organisms (PLTO) display a characteristic

acute angled branching morphology (Figure 1.4) and studies examining Australian isolates suggest that this morphotype, which can demonstrate marked differences *in situ* in its branching patterns, is associated exclusively with *Skermania piniformis* (Chun *et al.*, 1997; Seviour *et al.*, 2008; Soddell & Seviour, 1998).

In Australian activated sludge wastewater treatment plants surveyed by Seviour *et al.* (1994) located on the eastern coast of Australia *Microthrix,* GALO, and PTLO were the dominant filaments seen in foams, and similar results were obtained for Queensland plants surveyed earlier by Blackall *et al.* (1991). Findings by Seviour *et al.* (1994) also supported by those of Eikelboom (1991), which suggested that GALO and PLTO were more dominant in foams under warmer weather conditions, and thus that temperature may play an important role in determining foam communities. Furthermore, many Mycolata foam isolates obtained in pure culture grow under warmer conditions (Soddell, 2002). For instance, when 35 strains of *Gordonia amarae* were characterised by Lechevalier and Lechevalier (1974) only 6% grew at 10°C. Similarly, *S. piniformis* grew in temperatures up to 31°C (Soddell & Seviour, 1995).

A co-dominance of *Microthrix* and GALO was often observed, especially in Victorian plants (Seviour *et al.*, 1990). Interestingly, Seviour *et al.* (1990) noted a dramatic shift in filament dominance from *Microthrix* as reported by Blackall (1986), to GALO within a plant located in Helensvale (Queensland, Australia), but, the reasons for this were not elucidated. However, growth of *Microthrix can* occur when plant aeration is intermittent and pure culture studies have shown it to be microaerophilc (Rossetti *et al.*, 2005).

Table 1.1 Proportion of wastewater treatment plants affected by foaming in different nations

Region	Proportion of plants experiencing foaming [a]	No. plants sampled	Description of incidences [a]
Australia	51 %	129 [b]	51% experienced foaming. *Microthrix,* GALO, and PLTO most commonly observed (de los Reyes III, 2010; Seviour *et al.*, 1990; Soddell, 1999).
	92 %	46 [c]	92% of plants experienced foaming, 59% contained GALO and/or PLTO (Blackall *et al.*, 1991; de los Reyes III, 2010; Soddell, 1999).
	68 %	65	68% experienced foaming, *Microthrix* was most commonly observed. GALO and PTLO also observed (de los Reyes III, 2010; Seviour *et al.*, 1994; Soddell, 1999).
Denmark	50 %	N.D. [d]	*Microthrix* observed commonly (de los Reyes III, 2010; Wanner, 1994).
France	20 %	6000	Foam observed in 20% of plants, but this number is higher (87%) where plants operate with extended aeration. Fifty-eight plants were sampled and *Microthrix* was most commonly observed in these (45% of cases), in 14% of cases GALO were identified (Pujol *et al.*, 1991; Soddell, 1999).
Italy	50 %	167	Foaming problems observed, *Microthrix* most frequently observed, GALO also observed (de los Reyes III, 2010; Madoni *et al.*, 2000).
Netherlands	48 %	70 [e]	Foam caused by *Microthrix parvicella* occurs in spring when maximum population size is reached (de los Reyes III, 2010; Eikelboom, 1994).
	17 %		Foam in Autumn (de los Reyes III, 2010; Eikelboom, 1994).
South Africa	68 %	111	Foam observed in 68% of plants, *Microthrix* observed in 59% of foam samples, and GALO observed in 41% (Blackbeard *et al.*, 1986; Soddell, 1999).
United States	66 %	114	Experienced some form of foaming (de los Reyes III, 2010; Pitt & Jenkins, 1990; Soddell, 1999)

[a] For the duration of the study period; N.D. indicates no data; [b] Australian plants located in Queensland, New South Wales, and Victoria; [c] Australian plants located in Queensland; [d] BIO-DENITRO nutrient removal activated sludge systems; [e] Oxidation ditch plants surveyed

1.2.2. Mechanisms of stable foam formation

Foam is a dispersion of gas in a liquid, separated by thin liquid films called lamellae (Schilling & Zessner, 2011). Stable foam is a flotation event, being comprised of three elements; hydrophobic stabilising particles to reduce liquid drainage from the films, surfactants to stabilise the bubbles formed against coalescence, and gas bubbles (Petrovski *et al.*, 2011d; Schilling & Zessner, 2011; Soddell, 2002). In the activated sludge environment gas bubbles are provided by plant aeration systems required for their operation (Soddell, 2002). Slowly biodegradable surfactants can originate from both domestic and industrial influent, and also from the bacteria present in the system, as for example, with *Gordonia amarae* (de los Reyes III, 2010; Petrovski *et al.*, 2011d). Finally, activated sludge foams are further stabilised by the presence of hydrophobic bacteria (Petrovski *et al.*, 2011d; Soddell, 1999, 2002).

Hence, the selective enrichment of these bacteria in mixed liquor can lead to the spontaneous, often unpredictable, and rapid formation of foams (Soddell, 1999). Subsequently, it seems that foam formation occurs when the proliferation of foaming Mycolata have reached a particular mixed liquor cell density or threshold level, and such relationships have been postulated several times (Davenport *et al.*, 2008; Davenport *et al.*, 2000; de los Reyes & Raskin, 2002; de los Reyes III, 2010; Hiraoka & Tsumura, 1984; Jolis & Marneri, 2006; Mori *et al.*, 1992). Davenport *et al.* (2008) suggested that $2x10^6$ cells per mL of mixed liquor could be viewed as the universal threshold for the formation of all activated sludge foams, although this conclusion was based on several dubious assumptions.

Firstly, data obtained using FISH are likely to underestimate individual filament populations as FISH is limited to detecting those for which 16/23S rRNA sequence data are available (Amann & Ludwig, 2000; Kragelund *et al.*, 2007). Secondly FISH is limited to metabolically active cells, and does not detect dead or moribund cells themselves able to stabilise foams as demonstrated *in vitro* by Petrovski *et al.* (2011d). Davenport *et al.* (2008) also assumed that all Mycolata species have the same degree of Cell Surface Hydrophobicity (CSH), an assumption that is incorrect, since CSH is thought to reflect the mycolic acid chain lengths, which can differ markedly between individual Mycolata genera (de los Reyes III, 2010; Petrovski *et al.*, 2011d), and change in response to factors like temperature and nutrient status (Stratton *et al.*, 2002). When Petrovski *et al.* (2011d) studied 65 Mycolata strains in pure culture, they showed that individual strains differed in their levels of surfactant produced, and their CSH, and hence that foaming thresholds

are strain specific. Thus, foams of differing stabilities will be produced depending on which Mycolata strains are present in these microbial communities.

1.2.3. Effects of activated sludge foaming on plant operation

Stable activated sludge foams may occur that exceed one metre in height, and cause serious odour, aesthetic, operational, safety, and public health problems (de los Reyes III, 2010; Goddard & Forster, 1987). Public health concerns arise over their potential to disperse opportunistic bacterial pathogens in the aerosols these generate, including *Nocardia farcinica* (de los Reyes III, 2010; Jenkins *et al.*, 1993; Soddell, 1999; Stratton *et al.*, 1996). Safety concerns arise from reduced access to work areas, as well as slippery walkways and handrails (de los Reyes III, 2010; Goddard & Forster, 1987; Jenkins *et al.*, 1993). Foams also increase maintenance and operational costs as dried foam can complicate plant cleaning and overwhelm physical foam removal systems (de los Reyes III, 2010; Soddell, 1999). Consequently, the foam may enter the sedimentation tanks reducing the final effluent quality by increasing suspended solids (SS) and increasing biological oxygen-demand (BOD) levels (Bitton, 2005; de los Reyes III, 2010; Jenkins *et al.*, 2003; Soddell, 1999). Foams can also cover aeration basins to such an extent that the increased water head in the basin may prevent raw influent from entering it (de los Reyes III, 2010; Jenkins *et al.*, 1993). In sealed pure oxygen activated sludge systems which improve treatment efficiency by increasing the rate of oxygen transfer and hence its availability to the microbes present within it, foam can sometimes enter the oxygen compressors creating a fire hazard (Bitton, 2005; de los Reyes III, 2010). Thus, the need for a successful means of controlling foams is evident.

1.2.4. Current foam control methods

Given that stable foam formation requires air bubbles, surfactants, and hydrophobic bacterial cells, any successful control method must target one of these contributing factors (Petrovski *et al.*, 2011d; Soddell, 2002). Aeration cannot be stopped in activated sludge systems because it is required for waste treatment by aerobic organisms (Soddell, 2002). Similarly, surfactant levels and types cannot be controlled as these enter the plant as influent, and are also produced by

resident bacteria (de los Reyes III, 2010; Petrovski *et al.*, 2011d). Therefore, any effective methods for treating foaming must focus on eliminating the stabilising hydrophobic bacteria. Many control strategies have been proposed (Soddell, 1999, 2002), most are empirically based, or situation specific. Consequently none control all foams, at all times, at all plants, and this is mainly because of the lack of understanding of the bacteria involved in foaming. Current control methods are listed in Table 1.2 along with their known limitations. They include, but are not limited to, control of sludge age, mechanical scraping, using water sprays, biological selectors, reduction of pH, control of oil and grease levels entering plants, alteration of aeration regimes, and non-specific chemical treatments like the application of synthetic polymers (e.g. polyacrylamide cationic polymer), chlorination, iron salts (e.g. $FeCl_3$), hydrogen peroxide, and ozone (de los Reyes III, 2010; Kragelund *et al.*, 2010; Soddell, 2002).

Many of these methods are not specific to foaming bacteria, and many remain comparatively untested in activated sludge systems or show generally poor or inconsistent outcomes where testing has been carried out (Soddell, 2002). For example, a commercial product FEX-120 that is hypothesised to act as a substrate uptake suppressor, was effective against foams caused by *Gordonia* or *Skermania*, but did not perform well in plants with foaming problems caused by unknown Mycolata (Kragelund *et al.*, 2010). The failure of these methods to adequately control all foams highlights the need to seek a universal treatment method, and one that targets only the foaming bacteria. In an attempt to address this requirement Thomas *et al.* (2002) proposed the novel use of bacterial viruses also known as bacteriophages (phages), using phage therapy/biocontrol as a highly specific means of eliminating naturally known foaming bacteria based on the current understanding of the identity of the bacterial populations involved.

Table 1.2 Current methods for controlling foams and their limitations

Method	Limitations
Addition of anaerobic digester supernatant	Toxic agents are not often fully characterised, but are assumed to be competing microbes from the supernatant exhausting required nutrient sources. These are not successful in controlling all *Actinomycetes* foams, despite some promising results when screened against pure cultures of *Nocardia*. Additionally, previous occurrences of anaerobic digester foaming suggest successful application of this control method is unlikely (Bitton, 2005; de los Reyes III, 2010; Soddell, 1999; Soddell & Seviour, 1990).
Antagonistic microflora	While some predatory microbes have been observed, they have not been fully tested, and of those that have been, these were not effective against the foaming Mycolata (Bitton, 2005; Soddell & Seviour, 1990).
Antifoaming agents	Very expensive and often not consistently effective against all foams, possibly due to testing against less stable foams than those in activated sludge wastewater treatment plants (de los Reyes III, 2010; Kragelund *et al.*, 2010; Soddell, 1999, 2002; Soddell & Seviour, 1990).
Bioaugmentation	Commercial mixtures of microbes (composition often unknown) and enzymes, are often expensive and ineffective. Requires regular dosing for a lasting effect and does not appear effective in the laboratory. Some evidence does suggest a shift in foaming organisms from *Nocardia* spp. to *Microthrix* but without any overall decrease of the foam observed. Chemical composition remains unknown and very little data exist regarding their performance in activated sludge systems (de los Reyes III, 2010; Soddell, 1999, 2002; Soddell & Seviour, 1990).
Chlorination	Non-specific control method that requires careful monitoring to avoid over-chlorination. Not as effective as it is in bulking control. May not be sufficient for filament death as these are selectively retained in foams and not in the RAS lines. Excessive chlorination can cause floc dispersion and effluent deterioration. *Microthrix* appears to require a dose that is 10 to 100 times greater than that required for other filaments which could lead to over-chlorination (Bitton, 2005; de los Reyes III, 2010; Soddell, 1999, 2002; Soddell & Seviour, 1990).
Control of sludge age	Should wash out slow growing foamers but is not always successful as different foamers have different growth rates. For example, as *S. piniformis* and *Microthrix* both have slow growth rates they can be controlled this way, but control of GALO might possibly occur at the expense of nitrification processes where these organisms require a long sludge age (retention time prior to wasting). Hence, this method is not successful for all foams. Additionally, temperature

Method	Limitations
	and stability of foams already formed will effect how successful this control method is (Bitton, 2005; de los Reyes III, 2010; Soddell, 1999, 2002).
Biological selectors (contact zones)	Depends on a selective kinetic advantage between foaming and non-foaming populations. Not successful with all foam-formers e.g. anoxic selectors favour growth of floc formers at the expense of *Nocardia* spp., but are not as useful for *Microthrix* as it grows well anoxically. More successful for controlling bulking (Bitton, 2005; de los Reyes III, 2010; Jenkins *et al.*, 1993; Soddell, 1999, 2002).
Iron salts	Mixed results observed when dosed in mixed liquor. Useful for *Rhodococcus* scums, however, results varied *in vitro* amongst the Mycolata with some *Gordonia* and *Rhodococcus* strains not controlled, but growth of *G. amarae* inhibited. Costly, and masks the problem rather than treating it (Bitton, 2005; de los Reyes III, 2010; Soddell, 1999, 2002; Soddell & Seviour, 1990).
Manipulation of cell surface chemistry	Mixed results have been obtained from the addition of clay compounds, but these appear affective against foams produced by *G. amarae*. 3-hydroxyhexanoic acid inhibits cell wall Mycolic acid synthesis and has shown any promise as a foam control agent, however, production of it could be economically unsound (Bitton, 2005; de los Reyes III, 2010; Soddell, 2002).
Mechanical scraping	Collected foam can sometimes be recycled to the head of the plant, thus re-inoculating the mixed liquor with an enriched culture of foamers. May increase problems with aerosols where increased aeration is used to encourage foams to form on the surface where they can be selectively wasted (de los Reyes III, 2010; Soddell, 1999, 2002; Soddell & Seviour, 1990).
Ozonation	Suppresses *G. amarae* foams without affecting the rest of the biomass (de los Reyes III, 2010; Soddell, 2002).
Peroxide	Successful in controlling *G. amarae* foams (de los Reyes III, 2010).
Reduction of pH	Reduction in scum accumulation in some cases (Bitton, 2005).
Reduction of oil and grease levels	Reduction in scum accumulation in some cases (Bitton, 2005).

Method	Limitations
Synthetic polymers	Mixed results reported, however, cationic polymers have shown promising results with some *Nocardia* foams in California, and also some *Microthrix* foams. Mechanisms of action are yet to be fully elucidated. Inhibition of nitrification observed at higher concentrations and effects can vary based on cell viability (Bitton, 2005; de los Reyes III, 2010; Soddell, 2002).
Water sprays	Does not succeed in dealing with already established and/or heavy foams as it will not result in a complete mechanical collapse. Can dilute foams, however, this can create problems at the clarifier unless adequate scum traps are present to cope with the increased scum (Bitton, 2005; de los Reyes III, 2010; Soddell, 1999, 2002; Soddell & Seviour, 1990).

1.3. Bacteriophages

Bacteria are ubiquitous in the environment, with global populations estimated to be between 4×10^{30} to 6×10^{30} cells (Hendrix *et al.*, 1999; Whitman *et al.*, 1998). Phages are also abundant, and are present in any environment where bacteria occur. Phages are thought to outnumber their hosts by 5 to 10-fold in aquatic environments (Chibani-Chennoufi *et al.*, 2004; Ewert & Paynter, 1980; Hendrix, 2002; Hendrix *et al.*, 1999). Extrapolating these estimates across earth's entire biosphere suggests a total population of approximately 10^{31} individuals (Chibani-Chennoufi *et al.*, 2004; Ewert & Paynter, 1980; Hendrix, 2002; Hendrix *et al.*, 1999), making them the most abundant biological entities on earth (Ashelford *et al.*, 2002; Hendrix, 2002, 2003).

Phages were simultaneously and independently discovered by Frederick Twort and Felix d'Herelle in 1915 and 1917, respectively, as clear transmissible substances capable of passing through filtration devices that removed bacteria (Abedon *et al.*, 2011; d'Herelle, 1917; Twort, 1915). The term bacteriophage was coined initially by d'Herelle as a combination of bacteria and the Greek "phagin" meaning to devour, referring to the presence of phage infections (plaques) on bacterial lawns as an area devoid of bacterial cells. Earlier, albeit more controversial reports alluded to the potential action of phages prior to the work of Twort and d'Herelle. For example, Hankin (1896) reported antibiotic properties of the Ganges and Jumna rivers, although these are not widely acknowledged, and their legitimacy has been debated (Abedon *et al.*, 2011). Most phages observed to date consist of a protein coat, sometimes with a surrounding lipid envelope, that encapsulates the phage genome of either RNA or DNA, and the majority of dsDNA phages possess tails which aid their attachment to host bacteria at the beginning of the phage infection cycle (Ackermann, 1991, 1998, 2001, 2007; Withey *et al.*, 2005).

1.3.1. Phage taxonomy

At present phage taxonomy is based primarily on virion morphology, acknowledging only the basic viral genetic material (Nelson, 2004) and not its phylogeny. Phages were classified originally into six different morphotypes, grouped from A to F based on their morphology, and genome composition (Bradley, 1967), which were retained as genera when the first report from the International Committee on the Nomenclature of Viruses, later the International Committee on Taxonomy of Viruses, described the classification system which formed the basis of the

present system (Ackermann, 2003; Bradley, 1967). Since then many new phage groups have been added, and now phages are classified into one order, the *Caudovirales,* which contains the tailed phages, as well as thirteen families, and the "floating genus" *Salterprovirus* as shown in Figure 1.5 (Ackermann, 2007). The order *Caudovirales* is divided into three families based on phage tail morphology with the *Podoviridae* containing short tailed phages, the *Myoviridae* containing contractile tailed phages, and the *Siphoviridae* containing phages with long, flexible, non-contractile tails (Weinbauer, 2004). The *Caudovirales* are the most frequently observed group of bacterial viruses, as of the ~5,500 viruses observed under transmission electron microscopy to date, ~96 % have possessed tails (Ackermann, 2007; Fokine & Rossmann, 2014).

Despite the widespread use of this classification system, morphological and physiological characteristics are not strong keys for identifying phages, and the resultant phage groupings are not well structured, as those with little to no genome sequence similarity are frequently grouped together, perhaps because of the breadth of their genetic diversity (Chibani-Chennoufi *et al.*, 2004; Hendrix *et al.*, 1999; Lawrence *et al.*, 2002; Otawa *et al.*, 2007). For example, phages HK97 and L5 are presently both members of the *Siphoviridae*, while phage P22 which is closer genetically to HK97, than to L5, is placed into the *Podoviridae* because of its shorter tail morphology (Lawrence *et al.*, 2002).

As marked biological differences may result from few genetic differences, the current taxonomic system may not reflect meaningful phylogenetic relationships (Brüssow *et al.*, 2004b; Lawrence *et al.*, 2002). Consequently, alternative models have challenged the current system (Lawrence *et al.*, 2002; Lima-Mendez *et al.*, 2011; Pride *et al.*, 2006; Proux *et al.*, 2002; Rohwer & Edwards, 2002), including the phage proteomic tree model (Rohwer & Edwards, 2002). However, none of these have been adopted widely in the phage community. Doubt has been cast as to whether phages can be classified better using genomic data, as their genome sequences are extremely diverse and lack a common genetic element or marker sequence, such as the 16S rRNA gene targeted by universal Polymerase Chain Reaction (PCR) primers in the majority of prokaryotes (Bibby, 2014; Breitbart & Rohwer, 2005; Breitbart *et al.*, 2004; McNair *et al.*, 2012; Otawa *et al.*, 2007; Rohwer & Edwards, 2002; Rohwer & Thurber, 2009; Wommack *et al.*, 2012).

A number of genes have been proposed as markers among genetically related phages. These include the T7-like Podophage DNA polymerase and the *g20* capsid protein gene in

cyanophages, although neither is universal to all phages (Breitbart & Rohwer, 2005; Breitbart *et al.*, 2004; Fuller *et al.*, 1998; Otawa *et al.*, 2007; Rohwer & Edwards, 2002; Short & Suttle, 2005; Weinbauer & Rassoulzadegan, 2004). Phage taxonomy is further complicated by frequent genetic exchanges known to occur between phages and their hosts suggesting that they might instead exist as an indivisible taxonomic continuum (Brüssow & Desiere, 2001; Lawrence *et al.*, 2002; Rohwer & Edwards, 2002).

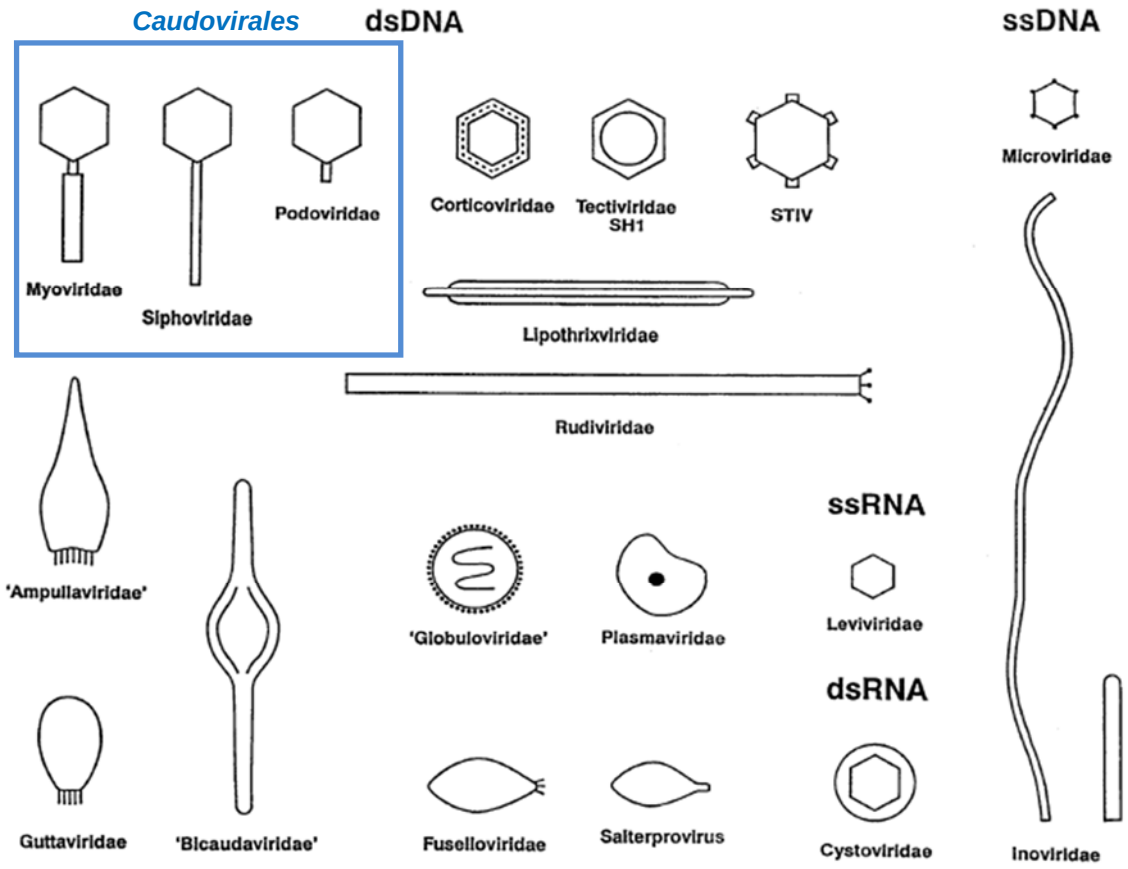

Figure 1.5 Current phage taxonomic system reproduced and modified from Ackermann (2007)

1.3.2. Prevalence of phages in activated sludge

Early studies conducted by Ewert and Paynter (1980) demonstrated the presence of phages in activated sludge systems, and they suggested abundances of 3.0×10^6 to 9.5×10^7 viral particles per mL using Transmission Electron Microscopy (TEM) to identify them based on their

morphology. However, such estimates usually underestimate viral abundances as these are difficult where phage concentrations are low because a higher magnification is required, and particulate matter can interfere with phage visualisation (Hennes & Suttle, 1995; Wommack & Colwell, 2000). Later epifluorescence microscopy (EFM) (Otawa *et al.*, 2007) suggested higher viral numbers of between $4.2x10^7$ and $3.0x10^9$ virus like particles (VLP)/mL, which then was the highest concentration of phages recorded for any environment, and supported by Wu and Liu (2009). This high number of phages might suggest that activated sludge is an ideal source for phage isolation, a notion supported by later metagenomic studies by Tamaki *et al.* (2012). In these fluorescence based viral abundance studies, it is unclear if the numbers stated are for total viruses, or relate only to phage.

In similar studies conducted on marine environments, the majority of VLP are assumed to be phages as bacteria are the most common prey there, and phage numbers correlate with bacterial abundance (with ratios between 5 to 10 times observed commonly). It is assumed that the same applies to estimates obtained for wastewater environments (Breitbart & Rohwer, 2005). Studies conducted by Otawa *et al.* (2007) revealed the majority of phage genome sizes were between 40 to 70 kbp in size using Pulse Field Gel Electrophoresis (PFGE), which also suggests that phages make up the majority of viruses. However, it is worth noting that epifluorescent studies of phage abundances could be potentially problematic as phage sizes are usually close to that to the resolution limit of most light microscopes. As all viral sized particles are counted overestimates could result, and no such test to eliminate these currently exists (Hennes & Suttle, 1995; Wommack & Colwell, 2000). However, the development of several new techniques for enumerating phage including counters like the Virus Counter 3100 (ViroCyt) which detects viral proteins and nucleic acid simultaneously, might allow more precise viral abundance estimates to be obtained, and techniques such as phageFISH (Allers *et al.*, 2013; Dang & Sullivan, 2014) might also permit better monitoring of specific phage populations *in situ*.

1.3.3. Origin and distribution of phage communities in wastewater

Ewert and Paynter (1980) demonstrated an increase in phage numbers in mixed liquor compared to those of plant raw sewage influent levels, suggesting that some phages are autochthonous to these systems and actively replicate there. These findings have been

supported by laboratory scale studies (Hantula *et al.*, 1991; Otawa *et al.*, 2007). Whether the majority of the viruses are naturally occurring there, or enter these systems with the influent is unknown. Metagenomic studies by Parsley *et al.* (2010) have suggested that these systems harbour large numbers of viruses. PFGE studies conducted by Otawa *et al.* (2007) showed that these had many different genome sizes, and also suggested that identical or very similar viruses might occur in geographically distant wastewater communities. This view is supported by studies where the same phages can be isolated from different treatment plants (Petrovski *et al.*, 2011a).

Tamaki *et al.* (2012) have suggested that the activated sludge virome is more diverse than that of any other aquatic environment, but differs in its composition at different sampling points (Tamaki *et al.*, 2012). This trend may be because of viral inactivation as phage populations like coliphages infective for enteric bacteria such as *Escherichia coli* are known to become inactivated in activated sludge (Ewert & Paynter, 1980; Hantula *et al.*, 1991; Otawa *et al.*, 2007). As human pathogenic enteric viruses such as the Hepatitis A virus (HAV) and Poliovirus require costly tissue culture dependent techniques and expertise for detection, coliphages might make good indicators of their inactivation as surrogates (Arraj *et al.*, 2005; Mesquita & Emelko, 2012; Tree *et al.*, 2003). These coliphages are also useful as indicators of their enteric hosts so their presence could be used to detect a host bacterium posing a possible threat to public health (Hilton & Stotzky, 1973; Mesquita & Emelko, 2012), and as indicators of disinfection efficiency in water and/or wastewater treatment (Worley-Morse & Gunsch, 2015). Efficient viral disinfection by wastewater treatment systems is not only important in controlling the spread of human and animal pathogens, but also to minimise the environmental impact of their release. However, estimates by Tamaki *et al.* (2012) suggest that approximately 1.8 X 10^8 VLP are discharged from plants located in the United States annually having persisted in effluents after treatment.

1.3.4. Ecology and population dynamics of phages in wastewater

The impact and importance of the roles carried out by phages in ecosystems were largely overlooked until data emerged describing perhaps their most obvious role in the predation and mortality of bacteria. This appears to occur to the same level, on average, as protozoan grazing induced mortality which prompted studies on how phages might impact the microbial food web

(Fuhrman, 1999; Hambly & Suttle, 2005; Rohwer & Thurber, 2009; Suttle, 2005; Weinbauer, 2004; Weinbauer & Rassoulzadegan, 2004; Wommack & Colwell, 2000).

Marine environments are the best studied phage ecological systems, although it is assumed that the same principles apply to phages in other environments like soil and activated sludge. Thus, it is thought that phages in activated sludge have the same potential to contribute to horizontal gene transfer, increase the spread of infectious disease by horizontal gene transfer of virulence factors, affect biogeochemical cycles by contributing to dissolved organic carbon levels (DOC) through viral lysis of bacterial hosts which consume this, nutrient recycling by liberating nitrogen and phosphorus through cell lysis making these available to other organisms, metabolic capacity by horizontal gene transfer, and possibly driving speciation/microbial diversity by killing specific populations within a community and by modifying specific genetic properties of bacteria by exerting a selection pressure for phage resistance (Allers *et al.*, 2013; Buckling & Rainey, 2002; Fuhrman, 1999; Haerter *et al.*, 2014; Hambly & Suttle, 2005; Rohwer *et al.*, 2009; Stern & Sorek, 2011; Thingstad & Lignell, 1997; Weinbauer, 2004; Weitz & Dushoff, 2007). The exact mechanisms by which phage co-exist in systems with their bacterial hosts, and other phages with the same bacterial hosts without extinction occurring is still subject to debate in the absence of much experimental data. So additional studies are required to understand phage population dynamics in wastewater systems where functional populations potentially could be manipulated to generate more efficient treatment processes. A number of studies in wastewater systems appear to provide evidence for some current models of phage-host dynamics and these are described below.

1.3.5. Phage predation of wastewater bacteria

Similar to marine habitats, the available data clearly show that phage have an important predatory role in activated sludge systems where they appear to reduce bacterial population densities and thus contribute to community homeostasis (Barr *et al.*, 2010; Hantula *et al.*, 1991; Khan *et al.*, 2002a; Lee *et al.*, 2004; Lee *et al.*, 2006a; Lee *et al.*, 2007; Ogata *et al.*, 1980). Many studies (Hantula *et al.*, 1991; Lee *et al.*, 2006b, 2007; Otawa *et al.*, 2007) have reported seemingly steady state population abundances between phage and their host, without extinction of either occurring. Many models have been published in various systems on the mechanisms by which this might occur, these include 'kill the winner' dynamics where the relative levels of abundant bacterial populations are controlled by viral infection (Thingstad &

Lignell, 1997), nested networks where viral strains have overlapping host ranges such that the most permissive host is infected by the most specialist virus (Flores *et al.*, 2011; Jover *et al.*, 2013), spatial refugees which might escape phage predation in biofilms or other less accessible areas of reactors (Heilmann *et al.*, 2012; Schrag & Mittler, 1996), and formation of resistance through antagonistic co-evolution (Buckling & Rainey, 2002; Koskella & Brockhurst, 2014).

Because of the ability of phage to lyse their hosts, attempts have been made to correlate phage numbers to poor treatment plant performance, especially in studies with laboratory scale EBPR systems. It has been hypothesised that decreases in EBPR capacity may arise from the death of PAO caused by phages (Khan *et al.*, 2002b). However, some of these data appear suspicious. For example, phages ØP27, ØP35, ØP37, and ØP38 shown to be lytic for the PAO (*Microlunatus phosphovorus*) were reported to be infective for both Gram-positive and Gram-negative bacteria. This suggests that they may have become contaminated, as phage are rarely known to cross generic boundaries (Ackermann & Dubow, 1987). A second study by Khan *et al.* (2002a) also isolated phages that were reported to be able to lyse both Gram-positive and negative bacteria from activated sludge, but suggested that some of these plaques observed might have arisen from induction of temperate phages on these hosts, and the same might be true for the other study. Lee *et al.* (2004) confirmed the presence of phages in EBPR systems, but their estimates of the number of phages isolated seems an overestimation as it was only based on plaque size and host range. Lee *et al.* (2006a) later isolated two phages infective for *M. phosphovorus,* they said appeared to bring about a substantial decrease in *M. phosphovorus* in a laboratory scale reactor. This event coincided with a temporary increase in phage numbers. Barr *et al.* (2010) also hypothesised that poor phosphorus removal performance in a laboratory scale granular wastewater reactor resulted from phage lysis of *Accumulibacter* (a PAO). Addition of a suspected phage rich supernatant from a failed EBPR reactor to a healthy EBPR community appeared to reduce the abundance of *Accumulibacter* (Barr *et al.*, 2010). Subsequently, mass spectroscopy data revealed a high abundance of phage tail proteins and proteins showing homology to those associated with an *Accumulibacter* prophage. However, no appropriate phage could be isolated and purified, as *Accumulibacter* cultures cannot yet be grown axenically (Barr *et al.*, 2010), and so all the evidence is indirect.

Systems containing *Accumulibacter* are high density, clonal, and non-recombining (little evidence of recombination or genomic mosaicism) and hence likely to demonstrate the 'kill the winner' style of phage predation (Kunin *et al.*, 2008), suggested by Thingstad and Lignell (1997).

A rise in abundance of a particular bacterial species occurs prior to an increased abundance of phages infective for it, and this is similar to the bank model proposed by Breitbart and Rohwer (2005) whereby only the most abundant phages which correspond to abundant bacterial hosts are active, while phages of other populations are still present but in lower numbers. The two theories differ only in how populations of non-active phages are maintained in the community. In these models, the winner presumably out competes other bacterial species with similar metabolic traits for a particular niche, or rises in numbers in response to a change in environmental conditions, only to become susceptible to specific phages. It is then no longer the winner, and remains in low numbers until populations of phage infective for it decline. So the process can repeat and hence over time mediates bacterial population diversity (Breitbart & Rohwer, 2005; Shapiro *et al.*, 2010; Thingstad & Lignell, 1997).

Shapiro *et al.* (2010) provided the first *in situ* demonstration of this 'kill the winner' hypothesis in a full-scale membrane bioreactor treating industrial wastewater, where frequent changes in community composition were seen that did not appear to be linked causally to either reactor conditions or wastewater composition. They showed that phage counts conducted against bacterial isolates from the monitoring period changed their levels corresponding to alterations in the bacterial community relating to two Operational Taxonomic Units (OTUs), both related to *Alcaligenaceae.* It is not unreasonable to suspect that similar relationships and dynamics might be observed across other activated sludge populations including the foaming Mycolata, although metagenomic studies suggest that perhaps such dynamics are not always clear as community composition might also be affected by phage host range and their ability to switch hosts in different environments, as well as higher resistance to treatment processes (Tamaki *et al.*, 2012).

1.3.6. Phage-host coevolution in wastewater

Phage infecting populations are not only viewed as predatory, but can also be viewed as a selective pressure in driving host and phage evolution as a result of their coexistence in the same system (Bohannan & Lenski, 2000; Buckling & Rainey, 2002). However, these appear to be affected by nutrient levels available to the host bacteria (Bohannan & Lenski, 2000). Alternatively, the continued coexistence of phage and bacterial hosts could reflect some bacteria

serving as spatial refugees by growing in biofilms and on walls of culture vessels (Bohannan & Lenski, 2000; Heilmann *et al.*, 2012; Schrag & Mittler, 1996). A similar hypothesis could be made in regards to plant structure and floc composition in activated sludge. However, given the vigorous aeration this seems less likely to occur in full-scale systems than in laboratory scale reactors.

While spatial refugees may contribute to phage-host community structure they fail to explain existence of stable host-phage communities like those in batch cultures where the growth vessel is replaced daily (Bohannan & Lenski, 2000). Furthermore, phages able to penetrate biofilms have also been isolated (Sutherland *et al.*, 2004). The 'red queen' hypothesis (Stern & Sorek, 2011; Van Valen, 1973) suggests that phage and their bacterial hosts are engaged in an antagonistic co-evolutionary 'arms race' leading to continuous host adaptations and counter-adaptations by the phage, resulting in these bacteria and their phages together selecting for resistant bacterial mutants, which in turn select for more infective parasites (Buckling & Rainey, 2002). These adaptations explain why some phage infections are not always successful, despite a putative bacterial cell with an appropriate receptor site and the appropriate molecular machinery. However, it is suspected that these adaptations are not always symmetrical i.e. the bacterium may produce a mutation that cannot be overcome by the phage (Bohannan & Lenski, 2000), and hence this theory may not fully explain how a phage and its bacterial host might co-exist. The apparent resistance observed in some phage isolation studies from wastewater might be explained by this theory (Hantula *et al.*, 1991; Khan *et al.*, 2002a; Khan *et al.*, 2002b).

Known defensive adaptations by the bacterial host are numerous, and include, but are not limited to, blocking of phage adsorption sites, Superinfection exclusion (Sie) systems that block the entry of foreign DNA into host cells by prophage conferred resistance, restriction modification (RM) systems, and Clustered Regularly Interspaced Repeats (CRISPR) systems (Dy *et al.*, 2014; Kruger & Bickle, 1983; Labrie *et al.*, 2010; Samson *et al.*, 2013; Skennerton *et al.*, 2011; Sorek *et al.*, 2008). These defence systems are described, along with corresponding counter-adaptations observed in phages to evade these, in Table 1.3. Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) systems involved in phage defence have been detected in *Accumulibacter* populations in comparative genomics studies of Flowers *et al.* (2013). They demonstrated that genes associated with these systems were present in two clades of *Accumulibacter* and suggested they might provide a level of resistance to phage in these populations, potentially contradicting earlier studies by Barr *et al.* (2010) while supporting the

'red queen' hypothesis. Evidence of CRISPR regions were also present in bacterial metagenomic studies of EBPR systems conducted by Kunin *et al.* (2008), which were then linked to phage metagenomic data showing a direct link between bacterial host and phage virion interactions in two geographically separated (Australia and the United States) sludge bioreactors dominated by a single population of *Accumulibacter*. Variability and redundancy between the US and Australian gene cassette controlling exopolymeric substance production (EPS) possibly involved in phage defence by masking their attachment sites, also suggests phage-*Accumulibacter* interactions (Kunin *et al.*, 2008). Further evidence of phage-bacterial coevolution was seen from phage metagenomic studies of an EBPR reactor dominated by *Accumulibacter* by Skennerton *et al.* (2011) where phage EVP1 contained a putative homologue of a heat-stable nucleoid structuring (H-NS) protein was also present in *Accumulibacter*, and thought to regulate the expression of CRISPRs in *E. coli.* This may have a similar function in phage EVP1 allowing it to evade CRISPR/Cas systems, and explain how *Accumulibacter* in the EBPR reactor of Barr *et al.* (2010) might have succumbed to phage predation and EBPR hence failed. However, no genome sequence data are available from Barr *et al.* (2010) to confirm this and in the absence of any cultured *Accumulibacter*, the designation of phage EVP1 as an *Accumulibacter* phage has only been inferred (Skennerton *et al.*, 2011).

1.3.7. Phage mediated horizontal gene transfer events in wastewater

Pure cultures have been used to isolate phages from wastewater that infect bacteria associated with bulking including the *Nostocoida limicola II* morphotype of *Tetrasphaera jenkinsii* (Petrovski *et al.*, 2012a), *Sphaerotilus natans* (Choi *et al.*, 2011; Jensen *et al.*, 1998), and *Haliscomenobacter hydrossis* (Kotay *et al.*, 2011), where they could play a predatory role in population control. The two phages isolated by Jensen *et al.* (1998) (SN-X and SN-T) infective for *S. natans* appeared lysogenic, and the latter was capable of transducing ribosomal genes of *Pseudomonas aeruginosa and S. natans* (Beumer & Robinson, 2005) suggesting that such horizontal gene transfer events occur in wastewater. This could potentially influence data based on 16S/23S rRNA genes. Later studies (Del Casale *et al.*, 2011a, b) used phage metagenomes to track these horizontal gene transfer events across wastewater communities, and determine which bacteria were participating in them. They could demonstrate that transduction events occur across most bacterial groups in treatment plants involving members of phylogenetically distant phyla including members of the *Nitrospirae, Planctomycetes,* and *Gemmatimonadetes*.

However, these horizontal gene transfer events did not appear to coincide with marked changes in the total bacterial community composition based on the transduced small subunit (SSU) rRNA genes (Del Casale *et al.*, 2011a, b). Such transduced 16S gene sequences were observed across the *Actinobacteria,* including in species related to *Rhodococcus* (a member of the Mycolata), however, not all of these member genera appeared to participate equally (Del Casale *et al.*, 2011a, b).

The presence of prophages in the foaming Mycolata was described as far back as the 1970s, such as the studies of Crockett and Brownell (1972) who reported induction of phage ϕEC from a lysogenic strain of *Rhodococcus erythropolis* with Ultraviolet (UV) light and mitomycin C. This strain was isolated from wastewater but is yet to have its genome sequenced (Sunairi *et al.*, 1993). The presence of these transducing phages suggests horizontal gene transfer events may occur, but also that 'fitness enhancing' genes such as those associated with resistance to other phages can be spread among the population. In the absence of genome sequence data this is difficult to predict, however, investigations focusing on the genomics of Mycolata prophages in wastewater might determine if prophages are able to contribute to the prevalence of the foaming Mycolata in wastewater treatment plants worldwide in this manner. Petrovski *et al.* (2013b) has also isolated two *Rhodococcus* phages RER2 and RGL3, that appear to be temperate phages based on their genome sequences (discussed in detail later). No evidence of genes associated with lysogenic conversion were detected in either genome, but as only 33 to 37% of their genes could be assigned any putative function, it is possible that some of the genes for which no function has been ascribed may play a role in superinfection exclusion, or confer factors improving host fitness (Petrovski *et al.*, 2013b). Large percentages of ORFans have also been observed in wastewater virome metagenomic analyses (Del Casale *et al.*, 2011a, b; Tamaki *et al.*, 2012), suggesting a need for more pure culture/single virus sequencing projects to update and expand existing databases of reference phage genomes (Bibby, 2014).

Table 1.3 Bacterial host defence systems and phage evasive mechanisms

Host defence system	Host defence system mechanism	Phage evasive mechanisms
Abortive Infection (Abi) Systems	A wide variety of abortive infection (Abi) systems exist and few of these are fully understood, however, they are usually composed of either a single protein or a protein complex and can commonly be found on mobile genetic elements (e.g. prophages and plasmids). It is known that these systems typically target replication, transcription, or translation in order to terminate phage infection as these are all crucial steps of the phage replication process, however, this also leads to the death of the bacterial cell and is often termed "altruistic suicide" as it presumably prevents phage infection in non-infected cells. These can sometimes include Toxin-Antitoxin systems (Dy *et al.*, 2014; Labrie *et al.*, 2010; Samson *et al.*, 2013).	Although many different and varied Abi systems have been discovered, the mechanisms by which phages have adapted to overcome these remain to be fully elucidated but evidence exists suggesting that some lack or have mutated genes that are usually involved in activating these systems, recombining with prophages within the genome of the bacterium containing the Abi system, encoding proteins that inhibit antitoxin systems, encoding their own antitoxins, encoding proteins that neutralise toxins associated with antitoxin systems, or by preventing depolarisation of cell membranes when Abi systems are activated (Dy *et al.*, 2014; Labrie *et al.*, 2010; Samson *et al.*, 2013).
Adsorption blocking	Various adsorption blocking mechanisms exist including blockage of phage receptors by altering the structure of the cell surface, production of an extracellular matrix to mask sites, and the production of competitive inhibitors (Dy *et al.*, 2014; Labrie *et al.*, 2010; Samson *et al.*, 2013).	Enzymes to assist phage adsorption to a bacterium that has adopted a mucoid colony morphology to partially block this, and alteration of a phage receptor site on a host bacteria can result in phage tail protein/receptor binding protein modifications which alter their adsorption specificity accordingly (Dy *et al.*, 2014; Labrie *et al.*, 2010; Samson *et al.*, 2013).
Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) Systems	CRISPR-*Cas* system systems have only recently been described and the exact mechanisms of how they confer immunity to a bacteria from invading nucleic acid such as that of phage or plasmids is yet to be fully elucidated, though they appear provide adaptive immunity in prokaryotes (Labrie *et al.*, 2010). A CRISPR-*Cas* loci contains a number of 21 to 48 bp direct	Can be evaded by single-nucleotide substitutions in the protospacer (corresponding sequence in the phage genome to the spacer acquired by the host CRISPR system) or in the region adjacent this. Homologues of these H-NS proteins have been found to regulate the expression of CRISPRs in *E. coli* and might have a similar function in bacteriophage to help them evade

Host defence system	Host defence system mechanism	Phage evasive mechanisms
	repeats interspaced by non-repetitive spacer regions of 26 to 72 bp. When a bacterium possessing such a systems is challenged with a phage, immune mutants arise and genomic analysis reveals the presence of a newly added spacer region. Similarly, the loss or alteration of this spacer region results in sensitivity to the phage itself (Dy *et al.*, 2014; Labrie *et al.*, 2010; Sorek *et al.*, 2008).	CRISPR-*Cas* systems. Other mechanisms of evading CRISPR systems can include the presence of anti-CRISPR-*Cas* genes, and by possessing their own CRISIPR-*Cas* systems that produce viral crRNAs and Cas proteins forming a CRISPR-*Cas* interference complex that hijacks the host CRISPR-*Cas* system to interfere with other antiphage systems derived from phage-inducible chromosomal island-like structures (Dy *et al.*, 2014; Labrie *et al.*, 2010; Samson *et al.*, 2013; Skennerton *et al.*, 2011).
Restriction Modification (RM) Systems	If injection of phage DNA is successful the bacterium may attempt to destroy the phage nucleic acid by means of a Restriction-Modification (RM) system. In the case of an RM system un-methylated phage DNA is recognised by a restriction enzyme and rapidly degraded while host bacterial DNA is protected by methylation (Dy *et al.*, 2014; Labrie *et al.*, 2010; Samson *et al.*, 2013).	Evaded by accidental methylation of phage DNA by the host, possession of methylase genes with the same specificity of those in the bacterial host genome, production of inhibitor proteins, destruction of restriction endonuclease cofactors, stimulation of host modification functions, the presence of unusual bases in their genome, complex DNA topologies reducing access to restriction sites, protective DarA/B proteins that bind to phage DNA preventing its degradation, modification of restriction endonucleases, or the absence of the prerequisite restriction sites for such systems (Dy *et al.*, 2014; Kruger & Bickle, 1983; Labrie *et al.*, 2010; Samson *et al.*, 2013).
Superinfection Exclusion (Sie) Systems	Prevents phage DNA from entering a host cell via superinfection exclusion system proteins which usually achieve this goal by either blocking the injection site, inhibiting transfer of phage DNA into the host cells, or altering the conformation of the injection site (Dy *et al.*, 2014; Labrie *et al.*, 2010).	In some cases these appear to be limited to closely related phages e.g. the Sie system of T4 stops phage DNA translocation for other T-even phages (e.g. T2, T4, and T6), and might be overcome by more genetically diverse phages. In other systems blocking of adsorption sites was overcome by mutations to tail apparatus (Dy *et al.*, 2014).

1.3.8. Phage therapy in wastewater

Lytic phages were utilised soon after their discovery in attempts to control medically important populations of bacteria in a process known as phage therapy (Sulakvelidze *et al.*, 2001). Though this practice was not widely used medically in most countries because of the availability of antibiotics together with historically controversial scientific and political issues (Sulakvelidze *et al.*, 2001), we are currently experiencing a renaissance in phage therapy research to combat the increasing numbers of antibiotic resistant bacterial pathogens (Sulakvelidze *et al.*, 2001). Phage therapy/biocontrol is not limited to medical applications, but, can also be used to control other problematic bacteria including those infecting crops and contaminating food, as well as food preparation processes (Parracho *et al.*, 2012).

When selecting phages for use in phage therapy:

- their biology should be well understood
- phages should be obligatory lytic/highly virulent as well as polyvalent where possible
- storage conditions of phage isolates/mixtures should be validated to ensure their viability
- receptor sites should be identified so that cocktails of phages targeting multiple receptor sites can be composed to decrease the chances of resistance developing
- all preparations should be tested in an appropriate model system to ensure predictable behavior as these might behave differently *in vivo* (Carlton, 1999; Skurnik & Strauch, 2006; Weld *et al.*, 2004)

Phage therapy has also been suggested as a means of improving the quality of effluent from activated sludge plants by specifically targeting groups of problematic bacteria within these systems without effecting other desirable ones (Withey *et al.*, 2005). Potential phage therapy

targets within wastewater systems include, but are not limited to, pathogen control in combination with biological sludge stabilisation processes, improving sludge dewaterability by targeting producers of exopolysaccharides using phage carrying polysaccharide depolymerase enzyme (PDE) encoding genes, improving sludge digestibility by targeting undesirable populations to increase substrate availability to anaerobic microflora, control of non-polyphosphate accumulating bacteria to prevent them becoming out-competed, the removal of biofilms from filtration systems, and control of filamentous bacteria which lead to solids separation problems such as foaming and bulking (Choi *et al.*, 2011; Kotay *et al.*, 2011; Petrovski *et al.*, 2011b, c; Petrovski *et al.*, 2012a, b; Thomas *et al.*, 2002; Withey *et al.*, 2005; Zhang *et al.*, 2013).

Laboratory scale phage biocontrol assays conducted using *S. natans* (with SN-phage) and *H. hydrossis* (with HHY-phage), appeared to be promising for the control of bulking, as improved settling was observed as a reduction in both turbidity of supernatant and sludge volume index after phage exposure, although whether the relative abundances of the filaments fell was not stated (Choi *et al.*, 2011; Kotay *et al.*, 2011).

As mentioned above, the idea for applying lytic phages to control foaming bacteria came from the work of Thomas *et al.* (2002). They isolated seventeen phages from activated sludge infective for members of the Mycolata genera *Gordonia, Mycobacterium, Nocardia, Rhodococcus,* and *Tsukamurella*. Additional Mycolata phages have since been isolated, namely *Gordonia* phage GTE7 (Petrovski *et al.*, 2011b), *Rhodococcus* phages REQ1, RER2, RGL3, and RRH1 (Petrovski *et al.*, 2013a, b; Petrovski *et al.*, 2012c). Many phages infective for the genera *Mycobacterium* have also been isolated (Hatfull & Hendrix, 2015), however, these are not specific for only the foaming Mycolata. Some of these Mycolata phages have very broad host ranges (Table 1.4), as shown by phages GTE2 and RGL3, which can lyse bacteria from three closely related Mycolata genera (Goodfellow & Maldonado, 2006; Petrovski *et al.*, 2011c, 2013b). Such broad host ranges might suggest that their bacterial hosts share a common receptor site, although which are the receptors for any of these Mycolata phages remains to be determined.

Table 1.4 Mycolata phage isolation, host ranges, and foam control assays

Phage name	Isolation source	Number of genera lysed	Strains lysed	Other lytic effects	Results of foaming assays	Reference(s)
GRU1	Wastewater: Loganholme (Queensland, Australia)	2	*G. terrae* (Gter34^T), *G. rubropertincta* (Grub38^T), N. nova (Nnov47^T)		*G. terrae* (Gter34^T) reduced from a foaming score of 2 to 0, *G. terrae* (G232) reduced from a foaming level of 2 to 0, *G. rubropertincta* (Grub38^T) reduced from a foaming level of 1 to 0, *N. nova* (Nnov47^T) remained at foaming level 1a	Petrovski *et al.* (2012b), Thomas *et al.* (2002)
GRU2	Wastewater: Burwood beach (New South Wales, Australia)	*3*	*G. terrae* (Gter34^T), *N. otitidiscaviarum* (Noti14), *N. brasiliensis* (Nbra42^T), *R. globerulus* (Rglo35^T), *R. erythropolis* (Rery19, Rery29^T)		N.D.	Thomas *et al.* (2002)
GTE1	Wastewater: Loganholme (Queensland, Australia)	3	*G. amarae* (Gama9), *G. rubropertincta* (Grub48), *G. terrae* (Gter34^T), *N. brasiliensis* (Nbra42^T), *N. brevicatena* (Nbre43), *N. otitidiscaviarum* (Noti14, Noti15), *R. erythropolis* (Rery19, Rery29, J32), *R. globerulus* (Rglo35^T), *R. rhodochrous* (Rrhod3)		N.D.	Thomas *et al.* (2002)
GTE2	Wastewater:	3	*G. terrae* (Gter34^T), *N.*		*G. terrae* (Gter34^T) reduced from	Petrovski *et al.* (2011c),

Phage name	Isolation source	Number of genera lysed	Strains lysed	Other lytic effects	Results of foaming assays	Reference(s)
	Merrimac (Queensland, Australia)		*otitidiscaviarum* (Noti14), *N. brasiliensis* (Nbra42^T), *R. globerulus* (Rglo35^T), *R. erythropolis* (Rery19, Rery29^T)		a foaming score of 2 to 0, *N. brasiliensis* (Nbra42^T) remained at a foaming level of 1a, *N. otitidiscaviarum* (Noti14) reduced from a foaming level of 5 to 0, *R. erythropolis* (Rery29^T) reduced from a foaming level of 5 to 2, *R. erythropolis* (Rery19) reduced from a foaming level of 4 to 1, *R. globerulus* reduced from a foaming level of 5 to 1	Thomas *et al.* (2002)
GTE3	Wastewater: Burwood beach (Queensland, Australia)	3	*G. sputi* (Gspu48), *G. terrae* (Gter34^T), *N. brasiliensis* (Nbra42^T), *N. otitidiscaviarum* (Noti15), *R. globerulus* (Rglo35^T), *R. rhodochrous* (Rrhod3)		N.D.	Thomas *et al.* (2002)
GTE4	Wastewater: Edgeworth (New South Wales, Australia)	3	*G. amarae* (Gama44^T), *G. terrae* (Gter34^T), *N. brasiliensis* (Nbra42^T), *N. farcinica* (Nfar26^T), *N. transvalensis* (Ntra40^T), *R. erythropolis* (Rery19, J32), *R. fascians* (N.D.), *R.* globerulus (Rglo35^T), *R. rhodochrous* (Rrhod3, Rrhod11), *R. ruber* (Rrub33^T)		N.D.	Thomas *et al.* (2002)

Phage name	Isolation source	Number of genera lysed	Strains lysed	Other lytic effects	Results of foaming assays	Reference(s)
GTE5	Wastewater: Carrum (Victoria, Australia)	1	*G. terrae* (Gter34^T, G232), *G. rubropertincta* (Grub38^T)		*G. terrae* (Gter34^T) reduced from a foaming score of 2 to 0, *G. terrae* (G232) reduced from a foaming level of 2 to 0, *G. rubropertincta* (Grub38^T) reduced from a foaming level of 1 to 0	Petrovski *et al.* (2012b), Thomas *et al.* (2002)
GTE7	Bendigo (Victoria, Australia)	*2*	*G. australis* (18F3M), *G. amiticia* (Ben607), *G. malaquae* (A554^T, A448), *G. terrae* (Ben601, Ben602, Ben603, Ben604, Gter34^T), *N. nova* (Nnov47^T), *N. asteroides* (Nast23^T)		*G. terrae* (Gter34^T, Ben601, Ben602) foam was reduced from an intermittent stable foam to fragile bubbles of 1cm in height, *G. amicitia* (Ben607) remained the same producing 1cm fragile bubbles, *N. asteroides* (Nast23^T) foam was reduced from stable films >8cm in height to 1cm fragile bubbles	Petrovski *et al.* (2011b)
MPH1	Wastewater: Loganholme (Queensland, Australia)	1	*M. phlei* (Mphl2), *M. smegmatis* (Msme1)		N.D.	Thomas *et al.* (2002)
NAS1	Wastewater: Merrimac (Queensland, Australia)	1	*N. asteroides* (Nast4), *N. otitidiscaviarum* (Noti15)		N.D.	Thomas *et al.* (2002)

Phage name	Isolation source	Number of genera lysed	Strains lysed	Other lytic effects	Results of foaming assays	Reference(s)
NBR1/2	Wastewater: Merrimac (Queensland, Australia)	1	*N. otitidiscaviarum* (Noti25^T), *N. brasiliensis* (Nbra42^T)		N.D.	Petrovski *et al.* (2014) Thomas *et al.* (2002)
NBR3	Wastewater: Burwood beach (New South Wales, Australia)	2	*N. asteroides* (Nast23^T), *N. brasiliensis* (Nbra42^T), *N. farcinica* (Nfar26^T), *N. transvalensis* (Ntra40^T), *R. erythropolis* (Rery19), *R. globerulus* (Rglo35^T)		N.D.	Thomas *et al.* (2002)
REQ1	Wastewater: Nambour (Queensland, Australia)	1	*R. equi* (Requ28^T)		N.D.	Petrovski *et al.* (2013a)
RER1	Wastewater: Burwood beach (New South Wales, Australia)	1	*R. erythropolis* (Rery19, Rery29^T), *R. globerulus* (Rglo35^T)		N.D.	Thomas *et al.* (2002)
RER2	Nambour (Queensland, Australia)	2	*N. otitidiscaviarum* (Noti15), *N. carnea* (Ncar30^T), *R. erythropolis* (Rery19. Rery29^T), *R. globerulus* (Rglo35^T)		N.D.	Petrovski *et al.* (2013b)

Phage name	Isolation source	Number of genera lysed	Strains lysed	Other lytic effects	Results of foaming assays	Reference(s)
RGL1	Wastewater: Loganholme (Queensland, Australia)	3	*G. sputi* (Gspu48), *N. nova* (Nnov47^T), *R. equi* (Requ10), *R. erythropolis* (Rery19, Rery29^T, J32), *R. fascians* (N.D.), *R. globerulus* (Rglo35^T), *R. rhodochrous* (Rrhod3, Rrhod11)		N.D.	Thomas *et al.* (2002)
RGL2	Wastewater: Nambour (Queensland, Australia)	2	*N. brasiliensis* (Nbra42^T), *R. erythropolis* (Rery19, Rery29^T, J32*), R. globerulus* (Rglo35^T), *R. rhodochrous* (Rrhod3)		N.D.	Thomas *et al.* (2002)
RGL3	Wastewater: Carrum (Victoria, Australia)	2	*N. otitidiscaviarum* (Noti15), *N. carnea* (Ncar30^T), *R. erythropolis* (Rery19. Rery29^T), *R. globerulus* (Rglo35^T)		N.D.	Petrovski *et al.* (2013b)
RRU1	Wastewater: Loganholme (Queensland, Australia)	2	*N. brasiliensis* (Nbra42^T), *N. nova* (Nnov47^T), *N. otitidiscaviarum* (Noti15), *N. transvalensis* (Ntra40^T), *R. coprophilus* (Rcop18, Rcop41), *R. fascians* (N.D.), *R. ruber* (Rrub33^T)		N.D.	Thomas *et al.* (2002)
RRH1	Wastewater: Daylesford (Victoria, Australia), Werribee	1	*R. equi* (Requ10, Requ28^T), *R. rhodochrous* (Rrho39, Rrho46), *R. erythropolis* (Rery29^T), *R. globerulus* (Rglo35^T)		N.D.	Petrovski *et al.* (2012c)

Phage name	Isolation source	Number of genera lysed	Strains lysed	Other lytic effects	Results of foaming assays	Reference(s)
	(Victoria, Australia), Ballarat (Victoria, Australia), Nambour (Queensland, Australia)					
TPA1/2	Wastewater: Daylesford (Victoria, Australia), Ballarat (Victoria, Australia)	1	*T. paurometabola* (Tpau37^T, DMSZ20162, IMRU1520, IMRU1312, IMRU1505, IMRU1283, NCTC107411), *T. pulmonis* (DSM44142), *T. tyrosinosolvens* (DSMZ44234^T), *T. pseudospumae* (N1176^T), *T. spumae* (N1171^T, JC85)	Turbid lysis on *T. inchonensis* (DSMZ44067^T), but no phage replication	N.D.	Thomas *et al.* (2002) Petrovski *et al.* (2011a)

N.D. indicates no data available; ^{T} indicates type strain.

The majority of isolated Mycolata phages (GRU1, GTE5, GTE7, RER2, RGL3 and GTE2) could distinguish between different strains of the same species (Petrovski *et al.*, 2011b, c; Petrovski et al 2012b; Petrovski *et al.*, 2013b). For example, GTE2 could lyse only one of five strains of *G. terrae* (Gter34^{T}) (Petrovski *et al.*, 2011c). While such a feature is not unusual, and has been exploited historically in phage typing of pathogenic bacteria (Chirakadze *et al.*, 2009; Petty *et al.*, 2007). The reason(s) for such specificity among phages infective for the Mycolata have not yet been elucidated, but probably reflects receptor site distribution among strains, or the presence of one of the host defence mechanisms discussed above e.g. RM, CRISPR, Abi, or Sie systems (Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2013a, b; Petrovski *et al.*, 2012a, b; Thomas *et al.*, 2002).

Promising results were obtained from laboratory scale foaming assays conducted using phages GTE2, GTE5, GTE7, and GRU1 (Petrovski *et al.*, 2011b, c; Petrovski *et al.*, 2012b) as described in Table 1.4. These demonstrated a marked decrease in the level and stability of existing foams when phages were present. For example, *Nocardia otitidiscaviarum* strain Noti42 and *R. erythropolis* strain Rery29 both produced a stable foams between 5 to 10 cm in height before addition of phage GTE2, which eliminated stable foam formation in the case of the former and only allowed formation of unstable films in Rery29 (Petrovski *et al.*, 2011c). Whether these lytic phages might work better when applied as a cocktail with others was not examined in these experiments.

1.4. Foaming Mycolata phage genomics

1.4.1. Mycolata genome structures

All isolated phages infective for the foaming Mycolata belong to the *Siphoviridae* family, possessing characteristic long non-contractile tails and dsDNA genomes (Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012b, c; Petrovski *et al.*, 2013a, b, 2014; Thomas *et al.*, 2002). However, only twenty-one such phages have been isolated and characterised so far. So these observations are seen as preliminary still, and representatives from other families possibly may be found. Of the 21 Mycolata phages characterised for their morphology and host range, only ten have undergone genome sequencing (listed in Table 1.5 with their genomic features).

Table 1.5 Summary of Mycolata phage genome structures

Phage name	Sequence Topology	Genome size (bp)	G+C % mol	No. putative tRNA	No. putative ORFs	Putative ORFs with assigned function*	No. putative novel genes*	No. inverted repeats	No. direct repeats	No. palindromic sequences	No. putative *rho*-independent transcriptional terminators	Reference
GTE2	Linear	45,530	60.3	0	57 [a] (35 forward and 22 reverse)	19	19	N.D.	N.D.	N.D.	N.D.	Petrovski *et al.* (2011c)
GTE5	Circularly permuted	65,839	65.0	0	93 [b] (70 forward, 23 reverse)	24	60	78	N.D.	6	N.D.	Petrovski *et al.* (2012b)
GTE7	Circularly permuted	74,431	56.8	1 (tRNA-Asn)	102 [a] (31 forward, 71 reverse)	13	57	N.D.	N.D.	N.D.	N.D.	Petrovski *et al.* (2011b)
GRU1	Circularly permuted	65,766	65.5	0	95 [b] (69 forward, 26 reverse)	18	54	69	N.D.	10	N.D.	Petrovski *et al.* (2012b)
NBR1	Linear	46,140	67.5	0	68 [b] (40 forward, 28 reverse)	17	39	N.D.	N.D.	N.D.	N.D.	Petrovski *et al.* (2014)
RER2	Circularly permuted	46,586	65.4	3 (tRNA-Asn, tRNA-Trp, tRNA-	67 [b] (24 forward, 43 reverse)	25	21	N.D.	N.D.	N.D.	N.D.	Petrovski *et al.* (2013b)

Phage name	Sequence Topology	Genome size (bp)	G+C % mol	No. putative tRNA	No. putative ORFs	Putative ORFs with assigned function*	No. putative novel genes*	No. inverted repeats	No. direct repeats	No. palindromic sequences	No. putative *rho*-independent transcriptional terminators	Reference
				undet)								
RGL3	Circularly permuted	48,072	65.9	3 (tRNA-Trp, tRNA-Gln, tRNA-undet)	66 [b] (24 forward, 42 forwards)	22	20	N.D.	N.D.	N.D.	N.D.	Petrovski *et al.* (2013b)
REQ1	Linear	51,342	66.3	0	85 [b] (59 forward, 26 reverse)	10	63	N.D.	N.D.	N.D.	N.D.	Petrovski *et al.* (2013a)
RRH1	Circular	14, 270	68.3	0	20	6	8	N.D.	N.D.	N.D.	N.D.	Petrovski *et al.* (2012c)
TPA2	Circularly permuted	61,440	69.1	0	78 [a] (33 forward, 45 reverse)	15	34	95	N.D.	6	4	Petrovski *et al.* (2011a)

** At the time of publication; N.D. indicates no data; [a] predictions made with Glimmer 3; [b] predictions made with Geneious; undet indicates undetermined.*

These ten genomes range between 14,270 bp (RRH1) and 74,431 bp (GTE7) in size. Of these, phage RRH1 has the smallest phage genome isolated for any member of the family *Siphoviridae*, so far (Petrovski *et al.*, 2012c). Naked DNA from this phage could be transformed by electroporation into electrocompetent *R. erythropolis* cells and produce visible plaques (Petrovski *et al.*, 2012c). All phage genomes had high G+C mol % contents (Table 1.5) usually close to those of their hosts, suggesting that are all well adapted. These range between 60 mol % and 78 mol % for host members of the genera *Gordonia, Nocardia, Tsukamurella,* and *Rhodococcus* (Goodfellow & Maldonado, 2006) from which they were isolated. The majority of the genomes were circularly permuted in structure, although three linear genomes were also observed in phages GTE2, NBR1, and REQ1 (Petrovski *et al.*, 2011c, 2013a, 2014). Thirty per cent of these phages contained from zero to three tRNA species, where their presence is said to reflect codons that are highly used by the phage (Enav *et al.*, 2012). An alternative hypothesis suggested that tRNA are carried by phages to boost the expression of late genes encoding the structural proteins (Enav *et al.*, 2012; Weigele *et al.*, 2007). Some phages, such as *Mycobacterium* phages Wildcat and DS6A also encode tmRNA, which in DS6A are incomplete. They may function to inhibit full length host tmRNA function, which is thought to benefit the phage lytic cycle (Chunhong *et al.*, 2009; Hatfull, 2010).

Features of the genome annotation data for these foaming Mycolata phages are summarised in Table 1.6 and reveal their typical modular genome architecture where genes of similar function are clustered together. Thus, each genome is essentially an assemblage of modules containing a single or multiple genes (Breitbart & Rohwer, 2005; Hendrix, 2002; Hendrix *et al.*, 1999; Juhala *et al.*, 2000; Lawrence *et al.*, 2002; Pedulla *et al.*, 2003; Weinbauer & Rassoulzadegan, 2004). However, these modules are often not well defined and may have other genes inserted into them. These are known as MORONs (for more DNA) and can be derived from other phages or their host cells (Breitbart & Rohwer, 2005; Juhala *et al.*, 2000). Mycolata phage genomes are highly diverse and variable, subject to horizontal gene transfer, and contain many novel genes, thus making it difficult to discuss the numbers of gene combinations possible (Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012b, c; Petrovski *et al.*, 2013a, b, 2014). Determining these for phages infecting the foaming Mycolata is not feasible, as only ten have had their genomes sequenced to date. However, some generalisations can be made.

1.4.1.1. Mycolata phage sequence repeat structures

Phages TPA2, GTE5, and GRU1 contain a number of palindromes (six to ten) ranging in size from 15 to 49 bp, together with inverted repeat structures (between 69-95) that range in size from 16 to 110 bp (Petrovski *et al.*, 2011a; Petrovski *et al.*, 2012b). Repeat sequences can be indicative of repeated protein structures, promoters, replication origins and transposable elements (Mott & Berger, 2007), but none of these could be identified in any of these Mycolata phage genomes. In phage TPA2, six of the repeat structures identified are similar to each other, and may be involved in antisense translation regulation, a view further supported by the presence of a sequence highly similar to sigma -35 (TTGACA) and -10 (TATAAT) sequences (Petrovski *et al.*, 2011a). This is the first report of such sequences in phage genomes, and their presence there warrants further study to better understand how transcriptional processes relate to phage infection. Many of the palindromes found in these Mycolata phage genomes were located in intergenic regions and consequently might act as *rho*-independent terminators (Lesnik *et al.*, 2001).

1.4.1.2. Mycolata phage DNA packaging

In all Mycolata phages examined a DNA packaging module is located upstream of the virion morphogenesis/structural gene module. The lysis and DNA replication modules are located in this order downstream of these genes. The DNA packaging module is responsible for inserting phage dsDNA into the capsid in a highly condensed state, and usually contains both a small terminase gene translated immediately upstream from a large terminase gene, with both genes being translated in the same direction (Rao & Feiss, 2008). These two terminase genes function as a complex, where the small terminase determines the specificity of DNA binding, as it is involved in recognising viral DNA which contains specific *cos* or *pac* sites for this purpose (Fujisawa & Morita, 1997; Rao & Feiss, 2008). The large terminase is involved in binding/docking to the prohead connector (portal protein), DNA translocation through the portal channel, and mediating the cleavage of phage DNA packaged into the phage prohead, the preformed protein shell of the capsid made up of multiple copies of the major capsid protein monomer, as well as a second cleavage once the capsid is full, termed the headful mechanism (Fujisawa & Morita, 1997; Rao & Feiss, 2008). Although the large terminase subunit genes were identified in all the available Mycolata phage genomes so far, small terminase subunit genes appear to be absent

from those of phages TPA2, RER2, and RGL3 (Petrovski *et al.*, 2011a; Petrovski *et al.*, 2013b). In the case of phage GTE7 the small terminase gene seems to be a fused version of two genes from the phage ReqiDocB7, neither of which appear to have an assigned function in this phage (Petrovski *et al.*, 2011b).

1.4.1.3. Mycolata phage structural genes

Downstream of the packaging region is the structural gene module, usually beginning with capsid morphogenesis genes and ending with those involved in tail morphogenesis (Pedulla *et al.*, 2003), as is the case with the phages of the Mycolata (Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012b, c; Petrovski *et al.*, 2013a, b, 2014;). The structural gene module typically begins with the phage portal protein through which DNA is packaged into the phage prohead capsid, and later serves as a connector for tail attachment following tail assembly. It is through this that DNA exits the phage capsid during infection (Fujisawa & Morita, 1997; Rao & Feiss, 2008). Also present is the gene encoding the major capsid protein which forms the prohead and undergoes rearrangement during DNA packaging to form the mature capsid, with an increased DNA holding capacity (Brüsow & Hendrix, 2002). The genes involved in capsid assembly vary between phages but sometimes include scaffold proteins around which the prohead forms.

In phage T4 scaffold proteins are required during assembly but are later cleaved to smaller peptides by prohead proteases that are then expelled to accommodate phage DNA (Dokland, 2000; Fokine *et al.*, 2004; Morgan *et al.*, 2000). Prohead proteases can be encoded either separately from scaffolding proteins (usually upstream of them), or can be nested within them (Dokland, 2000; Fokine *et al.*, 2004; Morgan *et al.*, 2000). Other genes encoding capsid proteins such as coat proteins, head-joining proteins, and decorator proteins may also be present (Rao & Black, 2010). For example, the genome of phage T4 encodes decorator proteins Hoc or Soc. The Soc protein increase capsid stability (Fokine *et al.*, 2004; Steven *et al.*, 1992), while Hoc aids in binding phage to surfaces such as those of mucous or host *E. coli* cells, a feature likely to improve success for infections (Barr *et al.*, 2013; Fokine *et al.*, 2011; Minot *et al.*, 2012).

Many of these features are not identifiable by sequence homology comparisons in the 10 Mycolata phages genetically characterised to date, with the possible exception of a portal vertex protein identified in phages GTE5 and GRU1 (Petrovski *et al.*, 2012b) and a prohead protease in phages RRH1 and NBR1 (Petrovski *et al.*, 2014; Petrovski *et al.*, 2012c). Major capsid protein genes were identified in all Mycolata phages except RRH1 (Petrovski *et al.*, 2012c) and TPA2 (Petrovski *et al.*, 2011a), where they are probably present, but if unique in their nucleotide and translated amino acid sequences their identification becomes difficult.

Tail morphogenesis genes are usually located downstream of head morphogenesis genes and may include genes encoding the main tail subunit, tape measure protein (TMP), tail fibre(s), base plates, spikes, and any additional proteins required for tail assembly (Brüsow & Hendrix, 2002; Sandmeier, 1994). Typically, the TMP has an alpha-helical structure and determines the length of the virion tail shaft, but can sometimes be slightly longer. Thus it is often identified as the largest gene in the phage genome (usually greater than 2,000 bp) as is the case with the Mycolata phages (Pedulla *et al.*, 2003; Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012b; Petrovski *et al.*, 2013a, b, 2014). Here, the TMP is located downstream of the main/major tail protein, and these two genes are often separated by two other genes, the second of which often lacks an obvious start codon and is expressed via a conserved programmed translational frameshift using a ribosomal slippage sequence (Xu *et al.*, 2004). This feature is common in *Siphoviridae* phage genomes and is found in the majority of Mycolata phages whose genomes have been sequenced (Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012b, c; Petrovski *et al.*, 2013a, b). The phage NBR1 genome had a different structure, with eight genes located between the major tail and tape measure protein, six of which were confirmed to encode structural proteins, suggesting a more complex tail structure. Other tail genes are usually located downstream of the TMP. Of these, tail fibre genes are of particular interest as they are responsible for host specify of binding to receptor molecules on a host cell's surface and thus contribute to the specificity of the host ranges (Sandmeier, 1994). Some phages, including Mu, possess genes for multiple sets of tail fibres in combination with invertase systems allowing them to alter their tail fibres and hence host ranges (Kamp & Kahmann, 1981; Morgan *et al.*, 2000; Sandmeier, 1994). Phage Mu accomplishes this by inverting a region of its genome, the G-segment, that contains the two different sets of tail fibres, thus controlling which of these will be expressed and widening its host range, so that it can attach not only to *E. coli* strain K12, but also to *Citrobacter freundii, Enterobacter cloacae,* and *Serratia marcescens*, and *vice versa* (Kamp & Kahmann, 1981; Morgan *et al.*, 2000; Sandmeier, 1994). No such features have been recognised in any of the Mycolata phage genomes analysed to date.

1.4.1.4. Mycolata phage lysis module

The phage lysis module is typically located adjacent to the phage structural gene module. This is the case in Mycolata phage genomes, although exceptions were seen in phage GTE2 (Petrovski *et al.*, 2011c) where the lysis genes were in the structural module, and phages REQ1, RER2, and RGL3 where these were seen either within or following the DNA replication/maintenance gene module (Petrovski *et al.*, 2013a, b). In RRH1 phage a single lysis gene was located in the structural region (Petrovski *et al.*, 2012c) disrupting its expected modular genome architecture. The lysis module is responsible for the destruction of the host cell wall, allowing release of mature phage (Loessner, 2005; Wang *et al.*, 2000). This module typically encodes a holin upstream of one or more endolysins (phage-encoded peptidoglycan hydrolases) (Loessner, 2005). Endolysins target directly bonds in the peptidoglycan, degrading the rigid murein layer, and are synthesised at the end of the phage multiplication cycle. They are muralytic/mureolytic enzymes, which can act as glycolases, transglycolases, amidases, or endopeptidases (Loessner, 2005; Wang *et al.*, 2000; Young *et al.*, 2000). Of the Mycolata phages characterised so far these genes were shown to encode for N-acetylmuramoyl-L-alanine amidases, chitinases, lysozymes, and peptidases as described in Table 1.6 (Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012b, c; Petrovski *et al.*, 2013a, b, 2014).

Holins are small (< 150 amino acids), hydrophobic membrane proteins containing two or more transmembrane domains (Loessner, 2005). They are required for permeabilisation of the host cytoplasmic membrane where they insert, assemble into oligomers, and create lesions that allow endolysins access to the cell wall murein after a genetically programmed period of time lapses (Loessner, 2005). This delay in their activity ensures a perfectly timed release of progeny phage, and that cell lysis occurs only after virion assembly and maturation have been achieved (Loessner, 2005; Young *et al.*, 2000). Holin encoding genes were not identifiable in all the Mycolata phages (as described in Table 1.6). However, in Gram-positive bacteria including the Mycolata, the cell wall murein is often accessible without a holin, allowing endolysins to act as exolysins (accessible from without), resulting in 'lysis from without', an attribute that might explain the apparent absence of these genes in some Mycolata phages (Loessner, 2005). The DNA replication module was located downstream of the lysis genes in most of the Mycolata phage genomes.

1.4.1.5. Mycolata phage DNA replication genes

Phage DNA replication typically initiates at an origin of replication where primers are synthesised by primases, RNA polymerases, or Rep proteins (Kutter *et al.*, 2005). Phages like Lambda have one origin of replication, but some including T4, T7, and P4 have multiple origins, each of which may operate under different environmental conditions (Kutter *et al.*, 2005). Replication origins are usually flanked by AT rich regions that aid in the unwinding of DNA (Kutter *et al.*, 2005), but none of these could be identified in the Mycolata phages. The DNA replication/recombination/maintenance module sequences vary greatly among the dsDNA phages. Some have genes that encode all proteins required for DNA replication, as in phage T4 where all genes required for its replisome and most of the enzymes required for preparing host nucleotides prior to their incorporation into phage genomic DNA are encoded (Miller *et al.*, 2003).

Other phages encode fewer DNA replication genes and rely on the host bacterial cell to carry out their functions, which appears to be the case with RRH1 phage, whose genome only encodes a HNH endonuclease, and no other DNA replication genes are present (Petrovski *et al.*, 2012c). Presumably phage RRH1 relies upon the molecular machinery of the host cell it can infect for these functions (Petrovski *et al.*, 2012c). Such findings question what had been suggested by metagenomic studies, that genes relating to DNA replication are more abundant in phages from wastewater than other environments, possibly resulting from the continuous flow of nutrients into these systems, leading to higher biomass levels and hence increased viral infection rates (Tamaki *et al.*, 2012). Other Mycolata phages like GTE7 (Petrovski *et al.*, 2011b), fit this more general picture outlined above. This phage contains several DNA replication genes, demonstrating that this genome module can vary substantially between phages. The genome of phage GTE7 (Petrovski *et al.*, 2011b) encoded for a DNA methylase, which may help evade host RM systems (Labrie *et al.*, 2010), and these genes appear to be common features of wastewater viral genomes (Tamaki *et al.*, 2012).

1.4.1.6. Mycolata phage lysogenic maintenance genes

Genes related to the maintenance of a lysogenic lifecycle usually include excisionases, integrases, and their associated repressor proteins (Canchaya *et al.*, 2003). These are responsible for integration of the phage genome into its bacterial host chromosome, or alternatively it's circularisation like a plasmid, and excision of the phage genome when a lytic cycle is induced (Ptashne, 2004). Repressor proteins control the direction and products of transcription for each pathway and are organised as a cascade of regulatory proteins that switch expression of blocks of genes on and off (Little & Michalowski, 2010; Ptashne, 2004). Genomes of Mycolata phages GRU1, GTE5 (Petrovski *et al.*, 2012b), and GTE7 (Petrovski *et al.*, 2011b) all contain putative excisionase genes typical of lysogenic phages. However, no integrase genes were located, and so in their absence these phages are assumed to be obligatory lytic/highly virulent. Similarly, phage REQ1 contained a putative serine recombinase gene characteristic of some lysogenic phages, but again no detectable integrase genes were seen, so this phage is again likely to be lytic (Petrovski *et al.*, 2013a). Interestingly, of the ten Mycolata phages with sequenced genomes, only phages RER2 and RGL3 were thought to be lysogenic because both possess putative phage integrase genes (Petrovski *et al.*, 2013b). However, no corresponding excisionases or phage repressor proteins could be identified with them.

Only two probable lysogenic Mycolata phages, RGL3 and RER2, have been sequenced (Petrovski *et al.*, 2013b), but it is not known if these contribute any selective genetic advantages to their host bacteria by altering their phenotype for example by conferring traits like toxin production, phage resistance, antibiotic resistance, altered metabolic capacity, resistance to environmental stress, or growth factors (Breitbart *et al.*, 2007; Brüssow *et al.*, 2004a; Canchaya *et al.*, 2003; Chen & Novick, 2009; Dang & Sullivan, 2014; Juhala *et al.*, 2000; Labrie *et al.*, 2010; Miao & Miller, 1999; Paul, 2008; Wang *et al.*, 2010). Based on their genome annotations this does not appear to be the case (Petrovski *et al.*, 2013b).

Table 1.6 Summary of Mycolata phage genome annotations

Phage	Packaging genes	Structural genes	Lysis genes	DNA maintenance, replication, and recombination genes	Lysogeny genes	Other features of interest
GTE2	Small terminase [b] Large terminase (pfam03354)	Phage portal protein (pfam05133) Major capsid protein (pfam05065) [f] Five structural proteins identified [e] Major tail protein (pfam05345) [f] Conserved programmed translational frameshift detected 28 bp before the end of *orf12* (GGGGGAA slippery sequence) Two tail proteins identified [a] Putative lysin gene within structural module [a] Putative holin gene located adjacent to putative lysin gene and contains two transmembrane regions [c]	Lysin and holin appear to be encoded in the structural gene module	Transcriptional elongation (pfam10263) Archaeal Holliday Junction Resolvase (pfam01870) DNA Polymerase I (COG0749) dCMP deaminase (pfam00383) Thymidylate synthetase (pfam00303) Phage encoded dCTP pyrophosphatase (pfam08761) Thymidine monophosphate kinase (cd01672) Helicase (pfam00721, pfam00176) Primase (COG3598) Homing endonuclease (pfam01844)		Cutinase (pfam01083)

Phage	Packaging genes	Structural genes	Lysis genes	DNA maintenance, replication, and recombination genes	Lysogeny genes	Other features of interest
GTE5	Small terminase (pfam01844) [b] Large terminase (pfam03354)	Portal protein (pfam05133) Portal vertex protein (PHA02531) Main capsid protein [a] Major tail protein [d] TMP with lytic transglycolase motif with peptidoglycan hydrolase motif (pfam06737, pfam01576, pfam02029) Programmed translational frameshift detected (*orf28-orf29*) Three putative minor tail protein [a] Eight structural proteins identified [e] and three confirmed [f] and in GRU1 by homology	Two adjacent lysin genes resembling *lysA/lysB* system Chitinase/lysin [a] Lysozyme (pfam00182) Holin (PHA02531) adjacent to lysin genes with two transmembrane domains [c]	Primase (pfam08706, pfam09250) DNA Polymerase III (pfam07733, pfam02811) DNA Helicase (pfam00271, pfam00176) Regulatory protein (cd00569)	Putative DNA binding excisionase protein (cd04762)	Shares nucleotide sequence identity with phage GRU1 GRU1 contains a truncated version of *ofr11* (COG5434)
GTE7	Small terminase [b] that appears to be a fused version of two genes from phage	Portal protein (DUF935) Major capsid protein (03864) TMP (pfam01464, pfam10145,	Three putative lysin genes, two adjacent to each other, with a third in the DNA replication	Putative *lysB* gene located in structural module that is closely related to that of ReqiDocB7 that contains a	Excisionase (pfam00176)	DNA Methylase (pfam01555) Cutinase

Phage	Packaging genes	Structural genes	Lysis genes	DNA maintenance, replication, and recombination genes	Lysogeny genes	Other features of interest
	ReqiDocB7 Large terminase [a]	pfam12128, pfam01576) Major tail protein identified [b] Conserved programmed translational frameshift detected for *orf17/orf18* Putative base plate protein (PHA02579)	module. N-acetylmuramoyl-L-alanine amidase and PGRP (pfam01510) Adjacent to this amidase is a putative *lysA/*Peptidase gene similar to that of phage ReqiPine5 and ReqiDocB7 (pfam01551)	cutinase motif (pfam01083) DNA Methylase (pfam01555) Exonuclease (pfam00929) Helicase (pfam00176) Primase [a] DNA Polymerase III (pfamcd00140) von Willebrand factor [a] ATP Hydrolysing protein (pfam07728)		(pfam01083)
GRU1	Small terminase (pfam01844) [b] Large terminase (pfam03354)	Portal protein (pfam05133) Portal vertex protein (PHA02531) Main capsid protein [a] Major tail protein [d] TMP with lytic transglycolase motif with peptidoglycan hydrolase motif (pfam06737, pfam01576, pfam02029,	Two adjacent lysin genes resembling *lysA/lysB* system Chitinase (pfam03412) Lysin (COG3179, pfam00182) Holin adjacent to lysin genes with two transmembrane domains	Primase (pfam08706, pfam09250, COG3378) DNA Polymerase III (pfam07733, pfam02811, COG0587) DNA Helicase (pfam00271, pfam00176) Regulatory protein (cd00569) DNA recombination inhibitor	Putative DNA binding excisionase protein (cd04762)	Shares nucleotide sequence identity with phage GTE5 *orf11* is a truncated version of *orf11* in GTE5

Phage	Packaging genes	Structural genes	Lysis genes	DNA maintenance, replication, and recombination genes	Lysogeny genes	Other features of interest
		COG5412) Programmed translational frameshift detected (*orf30-orf32*) Phage tail protein [a] Eight structural proteins identified [e] and three confirmed [f] in GTE5 (and GRU1 by homology)	[c]	(PRK00409)		
NBR1	Small terminase (pfam05119) Large terminase (pfam03354) Small and large terminase separated by a putative gene.	Portal protein (pfam04860) Prohead protease (pfam04586) Major capsid protein (05065) Excinuclease (pfam01541) Putative major tail protein homologous to that of phage GTE5 [a] Tape measure protein (pfam01464, cd00254) Eight genes located between the major tail protein and the	N-acetylmuramoyl-L-alanine amidase (pfam01510) with two peptidoglycan recognition protein motifs (pfam08310) Putative holin adjacent to lysis gene and contains two transmembrane domains [c]	Putative excinuclease in structural gene module Helicase (pfam04851, pfam00270) Bi-functional Primase/Polymerase (pfam09250, cd01125) DNA Polymerase I (pfam00476, cd00007) Holliday Junction Resolvase (cd00529)		

Phage	Packaging genes	Structural genes	Lysis genes	DNA maintenance, replication, and recombination genes	Lysogeny genes	Other features of interest
		putative tape measure protein [b]; six confirmed as structural proteins [d]				
REQ1	Small terminase [b] Large terminase (pfam033354)	Phage portal protein (pfam04586) Putative major capsid protein (pfam04586, pfam05065, TIGR01554) directly adjacent to portal protein and appears to be two fused genes (prohead protease and major capsid) TMP with lytic transglycolase domain in C terminus with peptidoglycan hydrolase motif (pfam01454, COG5280) [b]	Two lysis genes in (and after) DNA replication module No putative holin genes identified	Recombination endonuclease (pfam02945) Peptidase with a Transglycolase domain (pfam06737), as well as peptidase motif (pfam01476), and a *lysM* like motif (pfam01551) N-acetylmuramoyl-L-alanine amidase (pfam01510) with two peptidoglycan recognition proteins (pfam08310) similar to that of phage TPA2 Resolvase (cd00338, pfam07508) Putative endonuclease [a] Bi-functional Primase/Polymerase (pfam09250) with RecA motif (cd01393)	Serine recombinase (cd0038)	

Phage	Packaging genes	Structural genes	Lysis genes	DNA maintenance, replication, and recombination genes	Lysogeny genes	Other features of interest
RER2	Large terminase (pfam03334)	Portal protein (pfam05133) Major capsid protein (pfam05065) [a] Major tail protein [a] Tape measure protein (COG5412) Conserved programmed translational frameshift detected (*orf12-orf13*) Two minor tail proteins identified [a] Putative structural protein identified as a homologue of a minor tail protein of phage GTE5[a]	Putative holin [c] Single lysin gene following DNA replication/maintenance module only	Deaminase (pfam00383) Putative ligase (pfam01653) DNA Polymerase I (pfam00476) Thymidylate synthase (pfam02511) Ribonucleotide reductase (pfam02867) Primase (pfam01807) Recombination endonuclease (pfam02945) Putative helicase (pfam03796) RecB (pfam12705) Antirestriction protein (pfam07275) Lysin (pfam01551, pfam01183)	Integrase (pfam00239, pfam07508) Putative excisionase (pfam10935)	Shares nucleotide sequence identity to phage RGL3 Antirestriction protein (pfam07275)
RGL3	Large terminase	Portal protein (pfam05133) Major capsid protein	Putative holin [c] with homology to that of	Deaminase (pfam00383)	Integrase (pfam00239,	Shares nucleotide sequence

Phage	Packaging genes	Structural genes	Lysis genes	DNA maintenance, replication, and recombination genes	Lysogeny genes	Other features of interest
		(pfam0565) [a] Major tail protein [a] Tape measure protein (pfam00606, pfam04513) contains additional domains not evident in that of phage RER2 despite them sharing a high degree of nucleotide sequence identity including that of herpesvirus glycoprotein B (pfam00606) and also a baculovirus polyhedron envelope protein family (pfam045103) Conserved programmed translational frameshift detected (*orf12-orf13*) Two minor tail proteins identified [a] Putative structural protein identified as a homologue of a minor tail protein of phage GTE5[a]	phage GTE5 Single lysin gene following DNA replication/maintenance module only	Putative ligase (pfam01653) DNA Polymerase I (pfam00476) Thymidylate synthase (pfam02511) Ribonucleotide reductase (pfam02867) Primase (cd01029) Recombination endonuclease (pfam03796) Putative helicase (pfam03796) RecB (pfam12705) Antirestriction protein (pfam07275) Lysin (pfam01551, pfam01183)	pfam07508) Putative excisionase (pfam10935)	identity with phage RER2 Antirestriction protein (pfam07275) Tape measure protein contains motif for herpesvirus glycoprotein B (pfam00606) and one for a baculovirus polyhedron envelope protein family (pfam045103)

Phage	Packaging genes	Structural genes	Lysis genes	DNA maintenance, replication, and recombination genes	Lysogeny genes	Other features of interest
RRH1	Small terminase [b], Large terminase (pfam03354)	Portal protein (pfam04860) Lysin gene located adjacent to portal protein Thee major and five minor structural proteins identified by mass spectroscopy Prohead protease and major capsid protein appear to be a fusion protein (Orf7)	Single lysis gene (pfam01471) with homology to other known lysins Holin gene identified adjacent to lysin [c]	No DNA replication genes present, only a putative HNH endonuclease (pfam08144)		
TPA2	Large terminase (pfam03237)	Head morphogenesis protein (pfam04233) TMP (pfam10145) with lytic transglycolase motif (pfam06737) in C terminus with peptidoglycan hydrolase motif [b] Major tail protein [f] Tail component (pfam04883) Six other structural proteins identified [a] Putative *rho*-independent transcriptional terminators	Single lysis gene (l*ysA*) in lysis module with homology to N-acetylmuramoyl-L-alanine amidase (pfam01510), peptidoglycan recognition proteins (pfam08310) No *lysB* gene detected *Orf54* is a putative novel holin gene overlapping the putative *lysA* gene and contains two transmembrane domains [c]	Holliday junction resolvase (cd00529) DNA Polymerase I (pfam01612, pfam00476) Primase (pfam09250) Helicase (pfam04851, pfam00271) Endonuclease [a] Putative *rho*-independent transcriptional terminators (after *orf56*, between *orf61*		Metallophosphoesterase (COG4186) Chimeric genes present (e.g. *orf33, lysA*) Evidence of complex recombination events (shuffling of

Phage	Packaging genes	Structural genes	Lysis genes	DNA maintenance, replication, and recombination genes	Lysogeny genes	Other features of interest
		flank tail morphogenesis genes (*orf22-orf49*) No translational frameshift detected but *orf28* and *ord29* overlap by 169 bp and possess their own initiation codons (possible alternative mechanism) Ten structural proteins detected by Thomas (2005) [e]		and *orf62*) 750 bp non-coding region located between *orf64* and *orf65*, as well as between *orf77* and *orf78*		genes)

[a] identified by homology; [b] identified by gene position and orientation; [c] identified by size; location and the presence of transmembrane regions; [d] confirmed with shotgun mass spectroscopy; [e] identified by SDS-PAGE analysis; [f] identified with N-terminal sequencing

1.4.1.7. Evolutionary ancestry of Mycolata phages

It has been suggested that all these phages share a common ancestry, and that their mosaic genomes result from horizontal gene transfer from a large common genetic pool, where access is not as uniform to other sources of DNA including that of bacterial host and co-infecting plasmids and phages. However, with the latter, this is presumably limited by host range specificity (Hendrix *et al.*, 1999).

Mycolata phage genomes contained several interesting features that might clarify their evolutionary history. These include apparent gene shuffling events (seen in phage TPA2), the presence of motifs for a putative archaeal Holliday junction resolvase (as in phage GTE2), herpesvirus glycoprotein B (phage RGL3), baculovirus polyhedron envelope protein (phage RGL3), cutinases involved in degrading plant cell wall cutin in phytopathogenic fungi and bacteria (GTE2), as well as what appear to be chimeric genes (as in phage TPA2) (Petrovski *et al.*, 2011a, c, 2013b). However, one of the most striking of these is seen in phages GTE5 and GRU1, both of which display high levels of nucleotide sequence identity to each other (84.2% pairwise identity), and phages RER2 and RGL3, which also share high levels of nucleotide sequence homology (52.3% pairwise identity) to each other (Petrovski *et al.*, 2012b; Petrovski *et al.*, 2013b). In both cases this suggests a close evolutionary relationship exists between them (Petrovski *et al.*, 2012b; Petrovski *et al.*, 2013c).

The modular theory of phage evolution attempts to explain the mosaic appearance of phage genomes (Lima-Mendez *et al.*, 2011). Regions of high sequence similarity are followed sharply by regions of no sequence homology, and the suggestion is that recently acquired genetic elements are those represented by shared high nucleotide sequence homology, having had less time to succumb to mutational drift (Hendrix *et al.*, 1999; Juhala *et al.*, 2000). The model accounts for when genes from two different phages might share high levels of expressed amino acid sequence homology, but, differ markedly in their encoding nucleotide sequences (Brüssow & Desiere, 2001). This is exemplified by the genome of phage GTE7, which is highly novel at the nucleotide sequence level, but with a shared high level of relatedness at the amino acid level to the phage ReqiDocB7. This appears to indicate a more distant evolutionary relationship (Petrovski *et al.*, 2011a). Hendrix *et al.* (1999) have reported this situation across a number of phages infecting a phylogenetically very diverse range of host bacteria.

In their study of genome sequences of 60 Mycobacterium phages, similar distant types of evolutionary relationships emerged and were subsequently illustrated by the phage clustering system devised by Hatfull *et al.* (2010), in their attempts to better classify genetically the many *Mycobacterium* infective phages they isolated. Under this system, *Mycobacterium* phages could be divided into nine clusters, and twelve sub-clusters, based on both their nucleotide, and amino acid sequences, with the latter reflecting, and revealing more distant relationships (Hatfull *et al.*, 2010). As members of the genus *Mycobacterium* are also Mycolata, (Goodfellow & Maldonado, 2006), attempts were made to see if this scheme was applicable to the phages RGL3, RER2, and TPA2 (Petrovski *et al.*, 2011a, 2013b) and to determine their relatedness, if any, to these. Only three of the ten phages isolated that are infective for the Mycolata, were integrated into this system. These are RGL3 (cluster A2), RER2 (cluster A2), and TPA2 (cluster B). The rest of these phages, all with unique genome sequences appear as singletons, as did phage GTE7 which is closely related to ReqiDocB7 (also a singleton) at an amino acid level.

1.5. Aims of this study

If lytic phages for the Mycolata are to be developed for the control of foaming in activated sludge plants, then it is clear that much more needs to be learnt about their ecology, phylogeny, host specificities, and genetic diversity. In particular, this study aims to:

- Isolate lytic phages for the most common foaming Mycolata i.e. *Gordonia amarae* and *Skermania piniformis*. Although three phages infective for *G. amarae* were reported by Thomas *et al.* (2002), there are concerns about the authenticity of the identity of the strains used in that study (Dyson and Petrovski, unpublished). Thus, more phage isolates for these and other Mycolata members were sought

- Identify which of the Mycolata might contain temperate phages by conducting *in silico* screening of Mycolata whole genome sequence data in GenBank and then attempting to induce lytic cycles in these and other activated sludge isolates with mitomycin C. Induced prophages were characterised to see if these are potentially useful for foam biocontrol

- Investigate the presence of possible phage resistance systems that might interfere with lytic infections and movement of these phages through the activated sludge environment

2. Three of a kind: Genetically similar Tsukamurella phages TIN2, TIN3, and TIN4

2.1. Abstract

Activated sludge foaming is a global problem caused by the proliferation of bacterial members of the Mycolata, including *Tsukamurella spumae, T. sunchonensis* (Cheong et al., 2003) and *T. pseudospumae* (Nam et al., 2003, 2004). Three broad host range *Tsukamurella* phages TIN2, TIN3, and TIN4 were isolated from activated sludge treatment plants located in Victoria, Australia using conventional enrichment techniques. Illumina and 454 whole genome sequencing of these *Siphoviridae* viruses revealed similar genome sequences for each, ranging in size between 76,268 and 76,964 bp. All three phages shared 74 % nucleotide sequence identity to the previously described *Gordonia* phage GTE7. Genome sequencing suggested that phage TIN3 had suffered a mutation in one of its lysis genes, compared to that of TIN4 phage, to which it is genetically very similar. Mass spectroscopy data show the unusual presence of a gene encoding a virion structural protein within the DNA replication module of phage TIN4, disrupting the characteristic modular genome architecture of *Siphoviridae* phages. All three phages had broad host ranges and appeared obligately lytic, making them suitable for consideration as a cocktail of phages to biologically control *Tsukamurella* forming stable activated sludge foams.

2.2. Introduction

The genus *Tsukamurella* consists of Gram-positive organisms in the *Actinomycetales* (Schwartz *et al.*, 2002), and currently includes eleven species (Munk *et al.*, 2011). Members of this genus are strictly aerobic, weakly/variably acid-fast, non-motile, non-spore forming, whose cell envelopes contain long chain unsaturated mycolic acids (Almeida *et al.*, 2010; Bouza *et al.*, 2009; Collins *et al.*, 1988; Nam *et al.*, 2003; Yassin *et al.*, 1995). They, together with other mycolic acid producing genera *Mycobacterium*, *Nocardia, Rhodococcus, Williamsia, Rhodococcus, Tsukamurella, and Gordonia* are often referred to collectively as the Mycolata, where these mycolic acids render their cells highly hydrophobic (de los Reyes III, 2010; Savini *et al.*, 2012).

This high hydrophobicity and the biosurfactants they produce implicates them as stabilising agents of foams in activated sludge plants (de los Reyes III, 2010; Petrovski *et al.*, 2011d). Several species including *T. spumae* (Nam *et al.*, 2004), *T. sunchonensis* (Seong *et al.*, 2003) and *T. pseudospumae* (Nam *et al.*, 2004) have been isolated from these. Stable foams complicate sludge management, impact negatively on plant aesthetics, and increase maintenance costs (de los Reyes III, 2010). Furthermore, several opportunistic *Tsukamurella* pathogens are known to occur, thus posing a potential health hazard to plant operators from aerosol foam dispersal (de los Reyes III, 2010).

Stable foams consist of air bubbles, surface active agents, and hydrophobic particles (the Mycolata cells). As it is not possible to eliminate air bubbles and detergents from the activated sludge milieu, any successful control strategy must be directed at the bacterial component. Currently, no universal method exists that can control such foams, a situation which reflects our incomplete understanding of their microbial ecology, and composition of foaming communities (de los Reyes III, 2010).

Thomas *et al.* (2002) proposed that a phage therapy biocontrol approach might be an attractive option for treating this serious problem after successfully isolating 17 phages infective for foaming Mycolata from Australian wastewater treatment plants. More Mycolata phages have been isolated and described since then (Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012a, b, c; Petrovski *et al.*, 2013b, 2014). Among them is TPA2 phage, infective for certain members of the genus *Tsukamurella,* and the first such phage whose genome has been sequenced (Petrovski *et al.*, 2011a). A second temperate *Tsukamurella* phage TPA4 has since been induced from *Tsukamurella paurometabola* strain CON55 (see section 5). Both phages have highly novel genome sequences.

As of March 2015, genomes of 228 phages infective for members of the closely related genus *Mycobacterium* have been sequenced, (NCBI, 2015). To better understand their genomics, Hatfull *et al.* (2010) devised a novel clustering system that claimed to reveal insights into their evolutionary inter-relationships. In this system TPA2 phage would be placed into their cluster B as its genome organisation is similar to that of *Mycobacterium* phage Rosebush (Petrovski *et al.*, 2011a). On the other hand, phage TPA4 failed to fall into any of the pre-existing clusters.

In this study we report the characteristics of three lytic *Tsukamurella* phages TIN2, TIN3, and TIN4. This study explores their ecological roles, and suitability for biological control of activated sludge foams formed by, or containing, *Tsukamurella* strains.

2.3. Materials and methods

2.3.1. Isolation and preliminary characterisation of phage TIN2, TIN3, and TIN4

The bacterial strains used and methods for their growth are described by Petrovski *et al.* (2011a) and in Table 8.1. Phage TIN2 was isolated and subsequently purified from an activated sludge sample collected from Bendigo, Victoria, Australia, as described by Petrovski *et al.* (2011a) using the enrichment pool described in Table 2.1 (section 2.6).

Subsequently, plaques were observed on *T. inchonensis* strain BEN701 lawn plates. Phages TIN3 and phage TIN4 were also isolated from activated sludge samples from the Eastern Treatment Plant at Carrum, Victoria, Australia in the same manner, on the same strain of *T. inchonensis*. TIN3 and TIN4 phages were then separated from each other based on their respective plaque sizes during the phage purification process, since the plaques formed by TIN4 phage were slightly larger (~0.50 mm) than those formed by TIN3 phage (~ 0.25 mm) as discussed later.

Host range determinations were performed as described by Petrovski *et al.* (2011a). Grids for visualisation of the virions were prepared as described in Petrovski *et al.* (2011a) on carbon and formvar coated grids (Electron Microscopy Sciences, Australia), stained with uranyl acetate, and examined using a Tenaci Fei T30 Transmission Electron Microscope.

2.3.2. Genome sequencing of phages TIN2, TIN3, and TIN4

Genomic DNA was extracted from phage TIN2 and sequenced using a Roche GS FLX genome sequencer and titanium chemistry, as described in Petrovski *et al.* (2011a). Genomic DNA was

extracted from phages TIN3 and TIN4 in the same manner and DNA sequencing libraries were prepared using an Illumina Nextera XT sample preparation kit as per the manufacturers instructions. The prepared DNA libraries were sequenced on an Illumina MiSeq as a 150 bp paired end run. Sequence reads were assembled for all three phages using CLC Workbench (v6.5.1).

2.3.3. Genome annotation

The genome Open Reading Frames (*orf*s) within the *de novo* assembled sequences were detected initially using Glimmer (v3.02), where ORFs with a minimum size of 90 bp were detected (Delcher *et al.*, 2007). All predicted start codons were inspected for the presence of putative ribosomal binding sites and corrected manually if required. Sequence similarity searches were carried out against the GenBank database as described by Petrovski *et al.* (2011c). The presence of tRNA and tmRNA were determined using ARAGORN (Laslett & Canback, 2004) and tRNAScan-SE (Schattner *et al.*, 2005). Transmembrane domains were predicted with the DAS Transmembrane Prediction Server (Cserzo *et al.*, 1997).

Phage DNA when analysed by gel electrophoresis gave results consistent with circularly permuted DNA genomes. Therefore, for consistency the genomes annotations were conducted starting with the DNA packaging operon.

2.3.4. Mass spectroscopy of phage proteins

To identify phage structural proteins, high titre plaque purified filtered (0.22 µM pore size) phage particles (PFU/mL > 10^{10}) were precipitated using $(NH_4)_2SO_4$ following precipitation with $ZnCl_2$ to remove remnants of PEG used for phage concentrations, and other contaminant proteins. Pellets were re-suspended in 8 M urea to a final volume of 100 µL prior to analyses at the Mass Spectroscopy and Proteomics facility at La Trobe Institute of Molecular Sciences (Melbourne, Australia). Peptides reconstituted in 0.1% formic acid and 2% acetonitrile (buffer A)

were loaded onto a trap column (C18 PepMap 300 µm i.d. × 2 cm trapping column, Thermo-Fisher Scientific (Scoresby, Australia)) at 5 µL/min for 6 min and washed for 6 min before switching the precolumn in line with the analytical column (Vydac MS C18, 75 µm i.d. × 25 cm, Grace Davison). The separation of peptides was performed at 300 nL/min using a linear acetonitrile (ACN) gradient of buffer A and buffer B (0.1% formic acid, 80% ACN), from 5% buffer B to 40% over 60 min. Data were collected on an hybrid quadrupole/time-of-flight MS (MicroTOF-Q, Bruker, Germany) with a nano-electrospray ion source in the Data Dependent Acquisition mode and m/z 150–2500 as MS scan range. Nitrogen was used as the collision gas. The ionisation tip voltage and interface temperature were set at 4200 V and 205°C, respectively. Collision Induced Dissociation (CID) MS/MS spectra were collected for the 3 most intense ions only. Dynamic exclusion parameters were set as follows: repeat count 2, duration 60 s. The data were collected and analysed using Data Analysis Software (Bruker Daltonics, Bremen, Germany).

2.3.5. Nucleotide sequence accession numbers

Nucleotide sequences for phages TIN2, TIN3, and TIN4 have been deposited with the GenBank database under accession numbers KR011062, KR011063, and KR0011064, respectively.

2.4. Results and discussion

2.4.1. TIN2, TIN3, and TIN4 phages are all polyvalent

TIN2, TIN3, and TIN4 phages were isolated from activated sludge samples using pooled enrichments consisting of the 23 Mycolata hosts listed in Table 2.1 (section 2.6). Similar enrichment pool techniques were used successfully in the isolation of phage TPA2 (Petrovski *et al.*, 2011a). All three phages were then isolated from plaques that appeared after phage infections of *Tsukamurella inchonensis* strain BEN701. Both small (~0.25 mm) and large plaques (~0.5 mm) appeared on lawn plates of *T. inchonensis* strain BEN701. These plaque morphologies

were then separated and phages from each subsequently purified and designated as TIN3 (small plaques) and TIN4 phages (large plaques). All three purified phages were then screened against the 94 Mycolata isolates (Table 8.1), which came mainly from activated sludge plants, held in the La Trobe University culture collection. They included species of the genera *Gordonia*, *Nocardia*, *Rhodococcus*, *Mycobacterium*, *Skermania,* and *Tsukamurella*. As well as lysing *Tsukamurella inchonensis* strain BEN701, all three phages lysed three other *T. inchonensis* strains(CON50^{T}, BEN702, and BEN704), as well as four strains of *Tsukamurella paurometabola,* namely strains CON53, CON54, CON55, and CON61 (Bendigo culture collection numbers).

The identical host ranges of phage TIN2, TIN3, and TIN4 suggested that they might be similar genetically. Polyvalency across different *Tsukamurella* genera has been reported for *Tsukamurella* phage TPA2, which lysed strains of *T. paurometabola*, *T. pulmonis, T. tyrosinosolvens, T. pseudospumae,* and *T. spumae* (Petrovski *et al.*, 2011a).

When screened against other *T. paurometabola* strains, including CON51, and CON52 no plaque formation occurred with any of these three phages. A similar level of stringent strain specificity was observed in other actinophages (Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012b, c; Petrovski *et al.*, 2013a, b, 2014). Possible explanations include an absence of a compatible receptor site in the resistant host cell, absence of appropriate molecular machinery to support infection, or possession of phage resistance systems like Clustered Regularly Interspersed Repeat (CRISPR) systems, Restriction Modification (RM) systems, and abortive infection (Abi) systems (Labrie *et al.*, 2010).

2.4.2. Phages TIN2, TIN3, and TIN4 are all members of the family Siphoviridae

Phages TIN2, TIN3 and TIN4 all display an isometric capsid and a long non-contractile tail, characteristic of phages belonging to the *Siphoviridae* (Figure 2.1). All three demonstrated very similar morphologies, with all having type B1 isometric capsids ranging from ~57 to ~69 nm in size, and tails ranging from ~450 to ~471 nm in length.

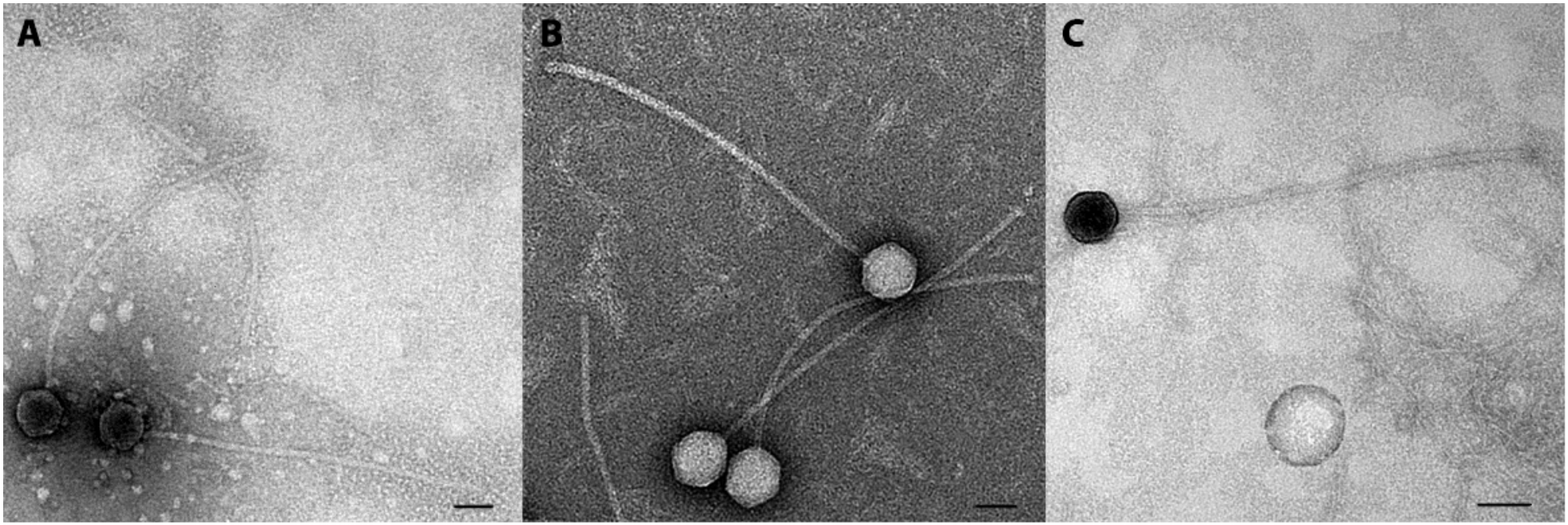

Figure 2.1 *Tsukamurella* TIN phage morphologies.

(A) phage TIN2; (B) phage TIN3; (C) phage TIN4. Scale = 50 nm. Negatively stained electron microscopy, 2 % uranyl acetate.

2.4.3. Phages TIN2, TIN3, and TIN4 are all genetically very similar

The genome sequence of TIN2 phage was obtained with Roche/454 pyrosequencing that generated 47,091 reads, with an average of ~224-fold average coverage following assembly. The genome sequences of phages TIN3 and TIN4 were generated with Illumina sequencing that gave 1,432,104 and 712,336 reads, respectively. When assembled, the genomes of phages TIN3 and TIN4 had an average of ~1,987-fold and ~900-fold average coverage, respectively.

All three phages possessed novel dsDNA genomes, of 76,964 bp, 76,269 bp, and 76,268 bp, respectively. Phages TIN2, TIN3, and TIN4 had G+C contents of 58.9 mol %, 59.3 mol %, and 59.3 % mol respectively, which is lower than those of their corresponding hosts, which range between 67 to 78 mol % (Goodfellow & Maldonado, 2006). All three phages contained a single tRNA-Asn that might aid in the acquisition of rare codons (Bailly-Bechet *et al.*, 2007; Enav *et al.*, 2012), but no tm-RNA were detected in any of these.

An alignment of all three phage genomes demonstrated that they are very similar to each other, and also to phage GTE7, as shown in Figure 2.2 and Table 2.2 (section 2.6) (Petrovski *et al.*, 2011b). Whole genome alignments showed that phage TIN2 shared high levels of nucleotide sequence identity with both of that of TIN3, and TIN4, with 89% shared identity for both. However, of the three phages, TIN3 and TIN4 are most similar, with TIN3 phage differing from TIN4 by a single short variable region of 20 bp long located in the phage lysis module in TIN3. In all three phage genomes genes are oriented in both forward and reverse orientations, with 107 to 109 putative *orf*s in total, with 34 arranged in a forward orientation in each. Only 21 to 23% of these putative genes could be annotated functionally in each, and a high number of ORFans of unknown function were present in all three genomes (Table 2.2 in section 2.6).

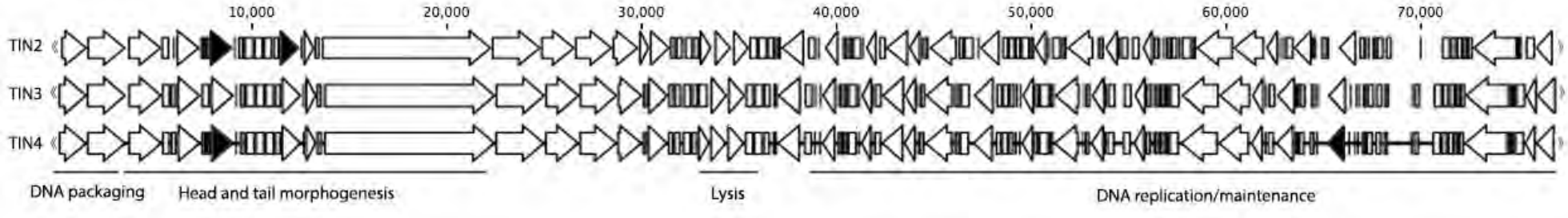

Figure 2.2 Genome map of phages TIN2, TIN3, and TIN4.

Black shaded genes indicate those identified as structural proteins with mass spectroscopy data.

Genomes of all three phages are highly modular in their genetic arrangement with genes of similar function clustered together. As indicated in Figure 2.2, these modules included genes related to DNA packaging, head and tail morphogenesis, cell lysis, and DNA replication/maintenance gene modules. In the absence of an obvious origin of replication in any of these phages, all *orf*s were numbered consecutively, starting with the small terminase gene, and in the same transcriptional direction as this gene (*orf1*).

2.4.4. Sequence repeats occur in all three Tsukamurella phage genomes

Several repeat structures were identified in these sequences, as summarised in Table 2.3 (section 2.6), and in Table 8.1 (section 8.2). All three contained 11 or 12 palindromes ranging from 16 to 54 bp in size. The majority of these were located intergenically, where they might act as *rho*-independent transcriptional terminators (Lesnik *et al.*, 2001). All three phage genomes contained 16 or 17 inverted repeat structures ranging in size from 16 to 68 bp. These repeat structures may be associated with repeated protein structures, promoters, replication origins (Mott & Berger, 2007), and transposable elements, but none of these were seen here.

A total of 17 to 19 direct repeat structures were also identified in each of the genomes, ranging in size from 16 to 246 bp. Similar repeat structures have been observed in *Tsukamurella* phages TPA2 (Petrovski *et al.*, 2011a) and TPA4 (see section 5) as well as other Mycolata phages (Petrovski *et al.*, 2012b), where similar hypotheses have been proposed concerning their possible functions.

2.4.5. Genome annotation of phages TIN2, TIN3, and TIN4 and their suitability for use in the biological control of foams

As mentioned above, all three phages are genetically similar to each other, and to the *Gordonia* phage GTE7 (Petrovski *et al.*, 2011b). Annotations of the TIN2, TIN3, and TIN4 genomes are summarised in Table 2.2 (section 2.6), and clearly demonstrate how similar they all are at the amino acid level.

The packaging modules of TIN2, TIN3 and TIN4 phages contain genes *orf1* and *orf2,* which encode the large and small terminase subunits, respectively. Typically these two genes function together as a complex, with the small terminase subunit determining the specificity of DNA packaging (Catalano, 2000), while the large terminase subunit mediates cleavage of the phage DNA packaged into the prohead (Fujisawa & Morita, 1997).

The phage structural module was located immediately downstream of the DNA packaging module in all three genomes. All contained within this a conserved translational slippage mechanism commonly seen in *Siphoviridae* phages (Xu *et al.*, 2004), located in the two genes immediately upstream of the putative tape measure protein encoded by *orf*18. These genes are thought to function in tail assembly (Xu *et al.*, 2004). The same slippage mechanism was also seen in phage GTE7 (Petrovski *et al.*, 2011b).

The phage lysis module is located after the structural genes in all three genomes, and typically consists of one or more phage lysin genes located adjacent to a phage holin gene. Together these lysins and holin are responsible for the release of progeny phage at the end of the replication cycle (Daniel *et al.*, 2007). Gene *orf*29 in all three phages appears to encode one of the phage lysis enzymes, as it shares between 74 to 77% identity to a lysozyme encoding gene from phage GTE7. *Orf3*0 in all three phages may encode holins, as their transcribed and translated proteins all share amino acid sequence similarities to the Gp54 protein of *Tsukamurella* phage TPA2, that was identified as a putative holin (Petrovski *et al.*, 2011a). However, Gp54 does not satisfy the listed criteria (Wang *et al.*, 2000) for a holin protein. These suggest that such a protein should be less than 150 amino acid residues in size and contain two or more transmembrane regions. As neither occur here its function remains unclear.

A second putative lysin is encoded by *orf31* in all three phages, and shares between 74 to 75 % of its protein amino acid sequence identity with a putative peptidase gene of phage GTE7. It also contains the peptidase motif pfam01551. *Orf*32 in all three might also encode a holin protein, as its translated protein is 152 amino acids in size, which is close to the 150 amino acids suggested by Wang *et al.* (2000) to be its likely size, and also contains between 4 and 5 transmembrane regions.

Immediately downstream of the phage lysis module in all three are the genes responsible for DNA replication/maintenance. More differences between the genomes of phage TIN2 and those of phages TIN3 and TIN4 were noticed in this module. Several motifs were present in the putative genes in this module in phage TIN2 that were absent from phages TIN3 and TIN4. For example, in the genome sequence of phage TIN2, *orf38* shared 49% identity to *Gp034* of phage GTE7 and encoded a motif for a prolipoprotein diacylglyceryl transferase (PRK131008). This motif was absent from the equivalent gene (*orf37*) in phages TIN3 and TIN4, which only shared 57% identity to product of *gp034* of phage GTE7. Orf40 in all three phages shares 86 to 87% amino acid sequence identity to the DNA methylase in GTE7 phage, and possesses the pfam01555 motif diagnostic for a DNA methylase. Thus, *orf40* might aid phage resistance to the host restriction modification systems involved in cleaving foreign DNA (Labrie *et al.*, 2010).

No genes like those encoding integrases, excisionases, and repressor proteins associated with lysogeny, were identified in these genomes suggesting all are obligately lytic phages. This would be a beneficial trait in any potential application for phage control of activated sludge foaming bacteria, since lysogenic phages may integrate into the host cell genome, where cell lysis could not occur unless triggered by an induction agent (Canchaya *et al.*, 2003).

2.4.6. Structural proteomics of TIN2, TIN3, and TIN4 virions

Structural genes appear to be located from gene *orf7* to *orf18* in all three *Tsukamurella* phages. Mass spectroscopy studies confirmed this function for genes *orf7, orf8,* and *orf15* in phage TIN2, and genes *orf7, orf8*, and *orf89* in TIN4 (Table 2.4 in section 2.6). As only data for the three most prominent structural proteins were obtained, this supports the annotation of *orf8* as the major capsid protein, and *orf15* as the major tail protein. Because phages TIN3 and TIN4 were in all aspects genetically very similar, TIN3 phage proteins were not analysed. The function of *orf89* as a structural protein gene in phage TIN4 is unusual, as it is located in the DNA replication/maintenance module, thus disrupting the otherwise expected modular structure of these genomes. This arrangement is likely to hold for TIN3 phage as well, where the same motifs in phage TIN4 (see Table 2.4 in section 2.6) are present in Orf89 (Table 2.4 in section 2.6). This is the first confirmed report of phage structural genes located outside the structural module in the wastewater foaming Mycolata phages, and it may be unique to the phages described

here. What functional attribute this gene plays in virion morphogenesis remains to be determined.

2.4.7. TIN3 contains mutations in the lysis module

The phage lysis module in TIN3 phage appears to extend from *orf29* to *orf32* and to contain putative holin encoding genes (*orf*30 and *orf*32), and two putative lysin genes (*orf*29 and *orf*32). Phage genomes of TIN3 and TIN4 only differ in their sequences within one gene (*orf29*) by 20 bp in TIN3, where it appears to result in a missense mutation. Subsequently, compared to that in phage TIN4, a truncated phage lysis/lysozyme may be synthesised. Whether this mutation in phage TIN3 contributes to reduced phage replication kinetics or rates of cell lysis remains to be determined.

Phages TIN3 and TIN4 had been separated as described earlier on the basis of plaque size and morphology generated after infection of the same *T. paurometabola* strain from the same wastewater sample. Thus, it could be argued that this 20 bp mutation may be the basis for this observed difference in plaque size where TIN3 plaques were smaller (~0.25mm) than those of phage TIN4 (~0.5 mm). It is possible that these two phages represent variants of the same phage, with phage TIN3 having undergone a genetic mutation, as one phage particle theoretically gives rise to one plaque. The second putative lysin gene is encoded by *orf31* in all three phages, so if *orf29* is rendered dysfunctional by mutation, *orf31* alone might facilitate host cell lysis.

2.4.8. Evolutionary relationships of phages TIN2, TIN3 and TIN4

The high degree of genome sequence similarity of the three *Tsukamurella* phages suggests a common evolutionary ancestry, most strikingly displayed with phages TIN3 and TIN4 where the only difference is in the putative lysin gene *orf29* in TIN3. Several putative indels were seen after comparing the genome sequences of these three *Tsukamurella* phages. For example their

putative tape measure protein genes (*orf18*) differed in size by only 9 bp, consistent with either a putative insertion or deletion events having occurred. Several putative substitution events were also noticed in *orf32* of these TIN phages. This gene in all three is the same length and has a similar sequence, but produces different best-match results after a BLAST search using GenBank database. Here, the closest relative of *orf32* is found in *Mycobacterium* phage Jolie1, while homologues in phages TIN3 and TIN4 are most similar to the homologue in phage 39HC.

The only sequenced lytic *Tsukamurella* phage genome published to date is that of phage TPA2, which contains several chimeric genes. However, none of these were seen in genomes of TIN2, TIN3, or TIN4 phages (Petrovski *et al.*, 2011a). Yet possible evidence for recombination events resulting in new modular arrangements of genes, like those seen in phage TPA2 (Petrovski *et al.*, 2011a), were also noticed here. A good example of this is in the genome sequences of phages TIN3 and TIN4, where the translated amino acid sequence of genes *orf54* to *orf57* in both show high sequence similarities to genes *gp051* to *gp054* in phage GTE7. Then following these genes in phages TIN3 and TIN4 are three novel genes (*orfs58* to *orf60*) and a GTE7 phage gene encoding a *gp056* protein homologue (Orf61). Furthermore, several genes in the phage GTE7 genome are absent from the genomes of phages TIN2, TIN3, and TIN4. For example, the third lysin gene of phage GTE7 (*gp040*) has no homologue in any of the three phages characterised here. This finding might indicate that some of these genes are accessory elements and superfluous to core phage functions, or perhaps that their roles are carried out by other genes in these TIN phages.

The only protein detected in the three *Tsukamurella* phages described here that was similar in its amino acid sequence to any of those present in the previously characterised *Tsukamurella* phage TPA2 was the *orf*30 that was a homologue of Gp054, the putative holin gene in phage TPA2. The lack of other genes homologous to those of TPA2 in these phages might suggest that phages infecting members of this genus are genetically diverse.

The high degree of nucleotide sequence similarity between these three *Tsukamurella* phages and the phage GTE7 infective for certain members of the genera *Gordonia* and *Nocardia* (Petrovski et al., 2011b), makes their host range differences surprising. The high level of genetic relatedness between *Actinophages* GTE5 and GRU1 (Petrovski *et al.*, 2012b), and RGL3 and RER3

(Petrovski *et al.*, 2013b), corresponds to them all having highly similar and/or identical host ranges.

The three *Tsukamurella* phages described here could not be assigned to any cluster in the scheme of Hatfull et al. (2010), and instead group with similarly unrelated phages called singletons. As several phages of this type exist that appear to be related evolutionarily, perhaps the *Mycobacterium* phage grouping should be extended to include a new cluster for phages ReqiDocB7, GTE7, TIN2, TIN3, and TIN4 (Petrovski *et al.*, 2011b; Summer *et al.*, 2011).

2.5. Conclusions

The isolation of three genetically very similar *Tsukamurella* phages has allowed a number of important insights into their genomics and evolution. These include identifying mutations within the lysis module of phage TIN3 alone. Mass spectroscopy data has identified the presence of a structural protein gene within the DNA replication module of phage TIN4, disrupting its expected modular genome architecture. Furthermore, TIN2, TIN3 and TIN4 phages are genetically quite different to the *Tsukamurella* phages TPA2 (Petrovski *et al.*, 2011a) and TPA4 (see section 5). The three phages described here have broad host ranges and appeared obligatory lytic. However, none of these lyse, under laboratory conditions, cells of the two common foaming bacteria *T. spumae* and *T. pseudospumae*, although phage TPA2 is lytic for both (Petrovski *et al.*, 2011a).

2.6. Appendix

Table 2.1 Strains used in the enrichment of phages TIN2, TIN3, and TIN4

Lab ID	Species
CON34[T]	*Gordonia terrae*
CON44[T]	*Gordonia amarae*
CON9	*Gordonia amarae*
G238	*Gordonia terrae*
G290	*Gordonia terrae*
BEN700	*Gordonia malaquae*
G255	*Gordonia terrae*
G232	*Gordonia terrae*
GOR9	*Gordonia terrae*
CON49[T]	*Gordonia sputi*
CON48	*Gordonia sputi*
CON65[T]	*Gordonia hydrophobica*
CON66	*Gordonia hydrophobica*
CON67	*Gordonia malaquae*
CON69[T]	*Gordonia desulfuricans*
CON71	*Gordonia polyisoprenivorans*
CON72	*Gordonia alkanivorans*
A554[T]	*Gordonia malaquae*
A448	*Gordonia malaquae*
CON38[T]	*Gordonia rubropertincta*
BEN701	*Tsukamurella inchonensis*
BEN703	*Rhodococcus erythropolis*
CON22[T]	*Gordonia aichiensis*

[T] *indicates type strain*

Table 2.2 Genome annotations of phages TIN2, TIN3, and TIN4

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN2-*orf1*	59..1402	448	putative small terminase [*Gordonia* phage GTE7]	57	0.0	Putative small terminase
TIN2-*orf2*	1395..3386	664	terminase large subunit [*Gordonia* phage GTE7]	98	0.0	Putative large terminase
TIN2-*orf3*	3502..5220	573	hypothetical protein GTE7_gp003 [*Gordonia* phage GTE7]	79	0.0	Unknown (pfam06074)
TIN2-*orf4*	5207..5632	142	hypothetical protein GTE7_gp004 [*Gordonia* phage GTE7]	49	2e-36	-
TIN2-*orf5*	5758..5964	69	hypothetical protein GTE7_gp006 [*Gordonia* phage GTE7]	85	6e-37	-
TIN2-*orf6*	5979..7187	403	hypothetical protein GTE7_gp007 [*Gordonia* phage GTE7]	71	0.0	Unknown (pfam09979)
TIN2-*orf7*	7221..7667	149	hypothetical protein GTE7_gp008 [*Gordonia* phage GTE7]	70	3e-72	Phage structural protein
TIN2-*orf8*	7681..8886	402	putative major capsid protein [*Gordonia* phage GTE7]	87	0.0	Phage major capsid protein E (pfam03864)
TIN2-*orf9*	8898..9089	64	hypothetical protein GTE7_gp010 [*Gordonia* phage GTE7]	76	2e-24	-
TIN2-*orf10*	9166..9459	98	-			

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN2-*orf11*	9466..9978	171	hypothetical protein GTE7_gp012 [*Gordonia* phage GTE7]	68	1e-73	-
TIN2-*orf12*	9966..10460	165	hypothetical protein GTE7_gp013 [*Gordonia* phage GTE7]	74	3e-62	Phage virion morphogenesis family (pfam05069)
TIN2-*orf13*	10457..10942	162	hypothetical protein GTE7_gp014 [*Gordonia* phage GTE7]	80	1e-91	-
TIN2-*orf14*	10956..11291	112	hypothetical protein GTE7_gp015 [*Gordonia* phage GTE7]	64	2e-21	-
TIN2-*orf15*	11295..12317	341	hypothetical protein GTE7_gp016 [*Gordonia* phage GTE7]	85	0.0	Phage structural protein (major tail)
TIN2-*orf16*	12511..13095	195	putative tail assembly protein [*Gordonia* phage GTE7]	81	2e-105	Putative tail assembly protein
TIN2-*orf17*	13077..13358	94	hypothetical protein GTE7_gp018 [*Gordonia* phage GTE7]	72	1e-14	Putative tail assembly protein translated by conserved translational slippage slippage sequence
TIN2-*orf18*	13467..22205	2913	phage tape measure protein [*Gordonia* phage GTE7]	53	0.0	Tape measure protein (pfam10145; pfam01576; PHA0135; pfam01464)
TIN2-*orf19*	22205..24736	844	hypothetical protein GTE7_gp020 [*Gordonia* phage GTE7]	73	0.0	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN2-*orf20*	24736..26484	583	hypothetical protein GTE7_gp021 [*Gordonia* phage GTE7]	78	0.0	-
TIN2-*orf21*	26484..28415	644	hypothetical protein GTE7_gp022 [*Gordonia* phage GTE7]	67	0.0	-
TIN2-*orf22*	28415..29767	451	-			-
TIN2-*orf23*	29782..30327	182	-			-
TIN2-*orf24*	30337..31377	347	hypothetical protein [*Mycobacterium abscessus*]	32	5e-17	-
TIN2-*orf25*	31374..31682	103	hypothetical protein ISGA_3314 [*Gordonia* sp. NB4-1Y]	34	5e-04	-
TIN2-*orf26*	31682..32047	122	hypothetical protein GTE7_gp025 [*Gordonia* phage GTE7]	71	1e-47	-
TIN2-*orf27*	32133..32570	146	hypothetical protein GTE7_gp026 [*Gordonia* phage GTE7]	58	8e-40	-
TIN2-*orf28*	32573..32899	109	hypothetical protein GTE7_gp027 [*Gordonia* phage GTE7]	53	5e-34	-
TIN2-*orf29*	32896..33525	210	lysozyme [*Gordonia* phage GTE7]	77	2e-118	Putative lysozyme
TIN2-*orf30*	33665..34528	288	hypothetical protein TPA2_gp54 [*Tsukamurella* phage TPA2]	55	2e-103	-
TIN2-*orf31*	34611..35456	282	putative peptidase [*Gordonia* phage	74	1e-137	Peptidase (pfam01551)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
			GTE7]			
TIN2-*orf32*	35453..35908	152	hypothetical protein Jolie1_013 [*Mycobacterium* phage Jolie1]	34	2e-09	Putative holin
TIN2-*orf33*	35896..36399	168	hypothetical protein GTE7_gp030 [*Gordonia* phage GTE7]	46	5e-37	-
TIN2-*orf34*	36392..36757	122	hypothetical protein GTE7_gp031 [*Gordonia* phage GTE7]	46	5e-21	-
TIN2-*orf35*	complement(36723..36878)	52	-			-
TIN2-*orf36*	complement(36865..37086)	74	-			-
TIN2-*orf37*	complement(37094..38290)	399	hypothetical protein GTE7_gp033 [*Gordonia* phage GTE7]	66	0.0	Nuclease (pfam12705)
TIN2-*orf38*	complement(38469..38891)	141	hypothetical protein GTE7_gp034 [*Gordonia* phage GTE7]	49	2e-20	Prolipoprotein diacylglyceryl transferase (PRK131008)
TIN2-*orf39*	complement(38958..39092)	45	-			-
TIN2-*orf40*	complement(39337..40059)	241	DNA methylase [*Gordonia* phage GTE7]	86	3e-153	DNA Methylase (pfam01555)
TIN2-*orf41*	complement(40254..40505)	84	-			-
TIN2-*orf42*	complement(40515..40769)	85	-			-
TIN2-*orf43*	complement(40769..41254)	162	hypothetical protein GTE7_gp041 [*Gordonia* phage GTE7]	53	2e-49	Nucleoside Triphosphate Pyrophosphohydrolase (cd11542)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN2-*orf44*	complement(41251..41415)	55	-			-
TIN2-*orf45*	complement(41454..42020)	189	hypothetical protein GTE7_gp043 [*Gordonia* phage GTE7]	69	1e-91	-
TIN2-*orf46*	complement(42135..42503)	123	hypothetical protein GTE7_gp045 [*Gordonia* phage GTE7]	60	3e-45	Unknown(pfam14359)
TIN2-*orf47*	complement(42493..43635)	381	DNA polymerase III beta subunit [*Gordonia* phage GTE7]	62	0.0	DNA Polymerase III Beta Subunit (cd00140)
TIN2-*orf48*	complement(43655..44272)	206	exonuclease [*Gordonia* phage GTE7]	69	3e-94	Exonuclease (pfam00929)
TIN2-*orf49*	complement(44269..44454)	62	DNA binding protein [*Gordonia* phage GTE7]	61	9e-19	DNA Binding Protein (pfam12728)
TIN2-*orf50*	complement(44438..44695)	86	-			-
TIN2-*orf51*	complement(44720..46147)	476	helicase [*Gordonia* phage GTE7]	75	0	Helicase (COG0553)
TIN2-*orf52*	complement(46218..46499)	94	-			-
TIN2-*orf53*	complement(46496..47014)	173	hypothetical protein GTE7_gp051 [*Gordonia* phage GTE7]	42	1e-37	-
TIN2-*orf54*	complement(47164..47319)	52	-			-
TIN2-*orf55*	complement(47319..48353)	345	hypothetical protein GTE7_gp052 [*Gordonia* phage GTE7]	68	1e-159	-
TIN2-*orf56*	complement(48469..48798)	110	hypothetical protein GTE7_gp053	63	2e-39	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
			[*Gordonia* phage GTE7]			
TIN2-*orf57*	complement(48776..49132)	119	hypothetical protein GTE7_gp054 [*Gordonia* phage GTE7]	36	4e-11	-
TIN2-*orf58*	complement(49132..49434)	101	-			-
TIN2-*orf59*	complement(49431..49892)	154	-			-
TIN2-*orf60*	complement(49899..50123)	75	-			-
TIN2-*orf61*	complement(50120..50758)	213	hypothetical protein GTE7_gp056 [*Gordonia* phage GTE7]	62	7e-87	-
TIN2-*orf62*	complement(50755..50964)	70	-			-
TIN2-*orf63*	complement(51020..51499)	160	hypothetical protein GTE7_gp057 [*Gordonia* phage GTE7]	41	1e-19	-
TIN2-*orf64*	complement(51549..51788)	80	hypothetical protein GTE7_gp058 [*Gordonia* phage GTE7]	70	6e-35	-
TIN2-*orf65*	complement(51802..53172)	457	-			-
TIN2-*orf66*	complement(53287..53454)	56	-			-
TIN2-*orf67*	complement(53451..53714)	88	-			-
TIN2-*orf68*	complement(53766..54434)	223	hypothetical protein GTE7_gp062 [*Gordonia* phage GTE7]	50	3e-73	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN2-*orf69*	complement(54415..54567)	51	-			-
TIN2-*orf70*	complement(54569..54961)	131	-			-
TIN2-*orf71*	complement(55130..55552)	141	hypothetical protein GTE7_gp065 [*Gordonia* phage GTE7]	78	2e-68	-
TIN2-*orf72*	complement(55671..56231)	187	hypothetical protein GTE7_gp066 [*Gordonia* phage GTE7]	36	5e-23	-
TIN2-*orf73*	complement(56244..56528)	95	-			-
TIN2-*orf74*	complement(56525..56701)	59	-			-
TIN2-*orf75*	complement(56694..56897)	68	-			-
TIN2-*orf76*	complement(56909..57166)	86	hypothetical protein GTE7_gp068 [*Gordonia* phage GTE7]	32	1e-04	-
TIN2-*orf77*	complement(57166..57495)	110	-			-
TIN2-*orf78*	complement(57497..57703)	69	-			-
TIN2-*orf79*	complement(57700..58236)	179	-			-
TIN2-*orf80*	complement(58226..58450)	75	-			-
TIN2-*orf81*	complement(58447..60351)	635	hypothetical protein GTE7_gp071 [*Gordonia* phage GTE7]	52	0.0	Cobalamin biosynthesis protein CobT VWA domain (pfam11775)
TIN2-*orf82*	complement(60354..61955)	534	ATPase family protein [*Gordonia* phage	62	0.0	AAA domain (pfam07728)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
			GTE7]			
TIN2-*orf83*	complement(62034..62675)	214	hypothetical protein GTE7_gp073 [*Gordonia* phage GTE7]	52	6e-74	-
TIN2-*orf84*	complement(62672..62833)	54	-			-
TIN2-*orf85*	complement(62830..63030)	67	hypothetical protein GTE7_gp074 [*Gordonia* phage GTE7]	44	5e-07	-
TIN2-*orf86*	complement(63040..63456)	139	hypothetical protein GTE7_gp075 [*Gordonia* phage GTE7]	37	1e-15	-
TIN2-*orf87*	complement(63459..64400)	314	hypothetical protein GTE7_gp077 [*Gordonia* phage GTE7]	53	3e-78	-
TIN2-*orf88*	complement(64397..64621)	75	-			-
TIN2-*orf89*	complement(64878..65186)	103	-			-
TIN2-*orf90*	complement(65191..65319)	43	-			-
TIN2-*orf91*	complement(65775..66695)	307	hypothetical protein GTE7_gp084 [*Gordonia* phage GTE7]	58	1e-104	-
TIN2-*orf92*	complement(66840..67151)	104	hypothetical protein GTE7_gp085 [*Gordonia* phage GTE7]	41	5e-09	-
TIN2-*orf93*	complement(67144..67260)	39	-			-
TIN2-*orf94*	complement(67270..67422)	51	-			-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN2-*orf95*	complement(67511..67732)	74	-			-
TIN2-*orf96*	complement(67732..68100)	123	-			-
TIN2-*orf97*	complement(68191..68523)	111	-			-
TIN2-*orf98*	complement(69932..70078)	49	-			-
TIN2-*orf99*	complement(71080..71541)	154	hypothetical protein GTE7_gp096 [*Gordonia* phage GTE7]	39	6e-29	-
TIN2-*orf100*	complement(71528..71842)	105	-			-
TIN2-*orf101*	complement(71839..72165)	109	-			-
TIN2-*orf102*	complement(72155..72409)	85	-			-
TIN2-*orf103*	complement(72406..72684)	93	hypothetical protein GTE7_gp097 [*Gordonia* phage GTE7]	54	9e-22	-
TIN2-*orf104*	complement(72681..74975)	765	putative primase [*Gordonia* phage GTE7]	44	0	Putative primase (pfam13148; PRK11633)
TIN2-*orf105*	complement(74972..75202)	77	hypothetical protein GTE7_gp099 [*Gordonia* phage GTE7]	56	5e-17	-
TIN2-*orf106*	complement(75398..75892)	165	hypothetical protein GTE7_gp101 [*Gordonia* phage GTE7]	43	1e-20	-
TIN2-*orf107*	complement(75893..76804)	304	hypothetical protein GTE7_gp103 [*Gordonia* phage GTE7]	67	3e-145	Unknown (COG4951)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN3-*orf1*	39..1493	485	putative small terminase [*Gordonia* phage GTE7]	59	0.0	Putative small terminase
TIN3-*orf2*	1474..3477	668	terminase large subunit [*Gordonia* phage GTE7]	80	0.0	Putative large terminase
TIN3-*orf3*	3593..5317	575	hypothetical protein GTE7_gp003 [*Gordonia* phage GTE7]	78	0.0	Unknown (pfam06074)
TIN3-*orf4*	5304..5729	142	hypothetical protein GTE7_gp004 [*Gordonia* phage GTE7]	49	2e-36	-
TIN3-*orf5*	5753..6061	103	hypothetical protein GTE7_gp006 [*Gordonia* phage GTE7]	82	4e-56	-
TIN3-*orf6*	6076..7281	402	hypothetical protein GTE7_gp007 [*Gordonia* phage GTE7]	70	0.0	Unknown (pfam09979)
TIN3-*orf7*	7315..7761	149	hypothetical protein GTE7_gp008 [*Gordonia* phage GTE7]	72	2e-72	Phage structural protein
TIN3-*orf8*	7775..8980	402	putative major capsid protein [*Gordonia* phage GTE7]	88	0.0	Phage major capsid protein E (pfam03864)
TIN3-*orf9*	8996..9187	64	hypothetical protein GTE7_gp010 [*Gordonia* phage GTE7]	77	6e-26	-
TIN3-*orf10*	9265..9549	98	-			
TIN3-*orf11*	9556..10068	171	hypothetical protein GTE7_gp012 [*Gordonia* phage GTE7]	68	5e-74	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN3-*orf12*	10056..10553	166	hypothetical protein GTE7_gp013 [*Gordonia* phage GTE7]	77	2e-60	Phage virion morphogenesis family (pfam05069)
TIN3-*orf13*	10550..11035	162	hypothetical protein GTE7_gp014 [*Gordonia* phage GTE7]	81	7e-93	-
TIN3-*orf14*	11049..11384	112	hypothetical protein GTE7_gp015 [*Gordonia* phage GTE7]	65	4e-21	-
TIN3-*orf15*	11388..12410	341	hypothetical protein GTE7_gp016 [*Gordonia* phage GTE7]	84	0.0	Putative phage structural protein (major tail)
TIN3-*orf16*	12602..13186	195	putative tail assembly protein [*Gordonia* phage GTE7]	81	6e-105	Putative tail assembly protein
TIN3-*orf17*	13268..13449	94	hypothetical protein GTE7_gp018 [*Gordonia* phage GTE7]	72	2e-23	Putative tail assembly protein translated by programmed conserved translational frameshift
TIN3-*orf18*	13557..22286	2910	phage tape measure protein [*Gordonia* phage GTE7]	53	0.0	Tape measure protein (pfam10145; pfam09486; COG5412; pfam09486)
TIN3-*orf19*	22286..24817	844	hypothetical protein GTE7_gp020 [*Gordonia* phage GTE7]	72	0.0	-
TIN3-*orf20*	24817..26565	583	hypothetical protein GTE7_gp021 [*Gordonia* phage GTE7]	67	0.0	-
TIN3-*orf21*	26565..28496	644	hypothetical protein GTE7_gp022	68	0.0	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
			[*Gordonia* phage GTE7]			
TIN3-*orf22*	28496..29767	424	-			-
TIN3-*orf23*	29776..30039	88	gp80 [*Bacillus* phage G]	46	3e-06	Glutaredoxin (pfam00462)
TIN3-*orf24*	30041..31081	347	hypothetical protein [*Mycobacterium abscessus*]	34	1e-19	-
TIN3-*orf25*	31078..31383	102	hypothetical protein ISGA_3314 [*Gordonia* sp. NB4-1Y]	36	1e-06	-
TIN3-*orf26*	31383..31748	122	hypothetical protein GTE7_gp025 [*Gordonia* phage GTE7]	72	3e-47	-
TIN3-*orf27*	31834..32271	146	hypothetical protein GTE7_gp026 [*Gordonia* phage GTE7]	55	4e-38	-
TIN3-*orf28*	32274..32600	109	hypothetical protein GTE7_gp027 [*Gordonia* phage GTE7]	58	1e-38	-
TIN3-*orf29*	32597..33106	170	lysozyme [*Gordonia* phage GTE7]	74	9e-42	Putative lysozyme
TIN3-*orf30*	33224..34084	287	hypothetical protein TPA2_gp54 [*Tsukamurella* phage TPA2]	58	5e-110	-
TIN3-*orf31*	34107..35003	299	putative peptidase [*Gordonia* phage GTE7]	75	3e-137	Peptidase (pfam01551)
TIN3-*orf32*	35000..35455	152	hypothetical protein 39HC_013 [*Mycobacterium* phage 39HC]	37	4e-12	Putative holin

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN3-*orf33*	35443..35943	167	hypothetical protein GTE7_gp030 [*Gordonia* phage GTE7]	45	7e-38	-
TIN3-*orf34*	35936..36298	121	hypothetical protein GTE7_gp031 [*Gordonia* phage GTE7]	43	4e-19	-
TIN3-*orf35*	complement(36422..36646)	75	-			-
TIN3-*orf36*	complement(36643..37839)	399	hypothetical protein GTE7_gp033 [*Gordonia* phage GTE7]	67	0.0	Nuclease (pfam12705)
TIN3-*orf37*	complement(38018..38443)	142	hypothetical protein GTE7_gp034 [*Gordonia* phage GTE7]	57	1e-20	-
TIN3-*orf38*	complement(38511..38645)	45	-			-
TIN3-*orf39*	complement(38730..38933)	68	-			-
TIN3-*orf40*	complement(38911..39630)	240	DNA methylase [*Gordonia* phage GTE7]	87	4e-154	DNA Methylase (pfam01555)
TIN3-*orf41*	complement(39727..39972)	82	-			-
TIN3-*orf42*	complement(39983..40231)	83	-			-
TIN3-*orf43*	complement(40231..40716)	162	hypothetical protein GTE7_gp041 [*Gordonia* phage GTE7]	53	3e-48	Nucleoside Triphosphate Pyrophosphohydrolase (cd11542)
TIN3-*orf44*	complement(40713..40913)	67	-			-
TIN3-*orf45*	complement(40910..41476)	189	hypothetical protein GTE7_gp043 [*Gordonia* phage GTE7]	66	9e-88	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN3-*orf46*	complement(41437..41601)	55	-			
TIN3-*orf47*	complement(41591..41959)	123	hypothetical protein GTE7_gp045 [*Gordonia* phage GTE7]	64	1e-47	Unknown (pfam14359)
TIN3-*orf48*	complement(41956..43092)	379	DNA polymerase III beta subunit [*Gordonia* phage GTE7]	62	0.0	DNA Polymerase III Beta Subunit (cd00140)
TIN3-*orf49*	complement(43096..43722)	209	exonuclease [*Gordonia* phage GTE7]	68	2e-93	Exonuclease (pfam00929)
TIN3-*orf50*	complement(43719..43904)	62	DNA binding protein [*Gordonia* phage GTE7]	60	5e-19	DNA binding protein (pfam12728)
TIN3-*orf51*	complement(43888..44163)	92	-			-
TIN3-*orf52*	complement(44170..45591)	474	helicase [*Gordonia* phage GTE7]	77	0	Helicase (COG0553)
TIN3-*orf53*	complement(45662..45943)	94	-			-
TIN3-*orf54*	complement(45940..46446)	169	hypothetical protein GTE7_gp051 [*Gordonia* phage GTE7]	37	5e-29	-
TIN3-*orf55*	complement(46616..47659)	348	hypothetical protein GTE7_gp052 [*Gordonia* phage GTE7]	67	9e-162	-
TIN3-*orf56*	complement(47776..48105)	110	hypothetical protein GTE7_gp053 [*Gordonia* phage GTE7]	68	3e-39	-
TIN3-*orf57*	complement(48083..48439)	119	hypothetical protein GTE7_gp054 [*Gordonia* phage GTE7]	37	6e-11	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN3-*orf58*	complement(48439..48744)	102	-			-
TIN3-*orf59*	complement(48741..48932)	64	-			-
TIN3-*orf60*	complement(48932..49087)	52	-			-
TIN3-*orf61*	complement(49084..49719)	212	hypothetical protein GTE7_gp056 [*Gordonia* phage GTE7]	61	2e-86	-
TIN3-*orf62*	complement(49716..49976)	87	-			-
TIN3-*orf63*	complement(49969..50502)	178	hypothetical protein GTE7_gp057 [Gordonia phage GTE7]	36	4e-17	-
TIN3-*orf64*	complement(50486..50731)	82	hypothetical protein GTE7_gp058 [*Gordonia* phage GTE7]	72	1e-37	-
TIN3-*orf65*	complement(50745..52022)	426	-			-
TIN3-*orf66*	complement(52140..52310)	57	-			-
TIN3-*orf67*	complement(52307..52567)	87	-			-
TIN3-*orf68*	complement(52620..53315)	232	hypothetical protein GTE7_gp062 [*Gordonia* phage GTE7]	55	6e-89	-
TIN3-*orf69*	complement(53296..53448)	51	-			-
TIN3-*orf70*	complement(53450..53851)	134	hypothetical protein ISGA_4751 [*Gordonia* sp. NB4-1Y]	33	4e-09	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN3-*orf71*	complement(54258..54680)	141	hypothetical protein GTE7_gp065 [*Gordonia* phage GTE7]	77	1e-68	-
TIN3-*orf72*	complement(54799..55395)	199	hypothetical protein GTE7_gp066 [*Gordonia* phage GTE7]	34	2e-18	-
TIN3-*orf73*	complement(55408..55692)	95	-			-
TIN3-*orf74*	complement(55689..55862)	58	-			-
TIN3-*orf75*	complement(55859..56113)	85	-			-
TIN3-*orf76*	complement(56123..56380)	87	hypothetical protein GTE7_gp068 [*Gordonia* phage GTE7]	32	1e-04	-
TIN3-*orf77*	complement(56380..56607)	86	unnamed protein product [*Gordonia* phage GRU1]	45	1e-05	-
TIN3-*orf78*	complement(56604..57125)	174	-			-
TIN3-*orf79*	complement(57204..59114)	637	hypothetical protein GTE7_gp071 [*Gordonia* phage GTE7]	54	0.0	Cobalamin biosynthesis protein CobT VWA domain (pfam11775)
TIN3-*orf80*	complement(59118..60710)	531	ATPase family protein [*Gordonia* phage GTE7]	63	0.0	AAA domain (pfam07728)
TIN3-*orf81*	complement(60784..61410)	209	hypothetical protein GTE7_gp073 [*Gordonia* phage GTE7]	52	3e-73	-
TIN3-*orf82*	complement(61419..61619)	67	hypothetical protein GTE7_gp074 [*Gordonia* phage GTE7]	38	8e-04	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN3-*orf83*	complement(61628..61984)	119	hypothetical protein GTE7_gp075 [*Gordonia* phage GTE7]	37	3e-17	-
TIN3-*orf84*	complement(62047..62994)	316	hypothetical protein GTE7_gp077 [*Gordonia* phage GTE7]	48	2e-77	-
TIN3-*orf85*	complement(62991..63215)	75	-			-
TIN3-*orf86*	complement(63218..63535)	106	-			-
TIN3-*orf87*	complement(63784..64053)	90	-			-
TIN3-*orf88*	complement(64096..64224)	43	-			-
TIN3-*orf89*	complement(64663..65586)	308	hypothetical protein GTE7_gp084 [*Gordonia* phage GTE7]	60	9e-108	Phage structural protein
TIN3-*orf90*	complement(65731..65877)	49	hypothetical protein GTE7_gp085 [*Gordonia* phage GTE7]	47	4e-05	-
TIN3-*orf91*	complement(66035..66163)	43	-			-
TIN3-*orf92*	complement(66176..66331)	52	-			-
TIN3-*orf93*	complement(66425..66655)	77	-			-
TIN3-*orf94*	complement(66655..67017)	121	-			-
TIN3-*orf95*	complement(67112..67447)	112	-			-
TIN3-*orf96*	complement(67535..67780)	82	-			-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN3-*orf97*	complement(68891..69028)	46	-			-
TIN3-*orf98*	complement(69061..69363)	101	-			-
TIN3-*orf99*	complement(70028..70489)	154	hypothetical protein GTE7_gp096 [*Gordonia* phage GTE7]	40	4e-29	-
TIN3-*orf100*	complement(70486..70815)	110	-			-
TIN3-*orf101*	complement(70812..71135)	108	-			-
TIN3-*orf102*	complement(71132..71380)	83	-			-
TIN3-*orf103*	complement(71377..71655)	93	hypothetical protein GTE7_gp097 [*Gordonia* phage GTE7]	53	7e-24	-
TIN3-*orf104*	complement(71652..73904)	751	putative primase [*Gordonia* phage GTE7]	44	0.0	Putative primase (PRK11633)
TIN3-*orf105*	complement(73901..74125)	75	hypothetical protein GTE7_gp099 [*Gordonia* phage GTE7]	55	8e-16	-
TIN3-*orf106*	complement(74103..74234)	44	-			-
TIN3-*orf107*	complement(74227..74721)	165	hypothetical protein GTE7_gp101 [*Gordonia* phage GTE7]	47	1e-22	-
TIN3-*orf108*	complement(74718..75311)	198	endonuclease VII [*Streptomyces* sp. NRRL S-920]	31	1e-16	Endonuclease (pfam02945)
TIN3-*orf109*	complement(75298..76269)	324	hypothetical protein GTE7_gp103	68	6e-149	Unknown (COG4951)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
			[*Gordonia* phage GTE7]			
TIN4-*orf1*	39..1493	485	putative small terminase [*Gordonia* phage GTE7]	59	0.0	Putative small terminase
TIN4-*orf2*	1474..3477	668	terminase large subunit [*Gordonia* phage GTE7]	80	0.0	Putative large terminase
TIN4-*orf3*	3593..5317	575	hypothetical protein GTE7_gp003 [*Gordonia* phage GTE7]	78	0.0	Unknown (pfam06074)
TIN4-*orf4*	5304..5729	142	hypothetical protein GTE7_gp004 [*Gordonia* phage GTE7]	46	1e-34	-
TIN4-*orf5*	5753..6061	103	hypothetical protein GTE7_gp006 [*Gordonia* phage GTE7]	82	4e-56	-
TIN4-*orf6*	6076..7281	402	hypothetical protein GTE7_gp007 [*Gordonia* phage GTE7]	70	0.0	Unknown (pfam09979)
TIN4-*orf7*	7315..7761	149	hypothetical protein GTE7_gp008 [*Gordonia* phage GTE7]	72	2e-72	Phage structural protein
TIN4-*orf8*	7775..8980	402	putative major capsid protein [*Gordonia* phage GTE7]	88	0.0	Phage major capsid protein E (pfam03864)
TIN4-*orf9*	8996..9187	64	hypothetical protein GTE7_gp010 [*Gordonia* phage GTE7]	77	6e-26	-
TIN4-*orf10*	9265..9549	98	-			-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN4-*orf11*	9556..10068	171	hypothetical protein GTE7_gp012 [*Gordonia* phage GTE7]	68	5e-74	-
TIN4-*orf12*	10056..10553	166	hypothetical protein GTE7_gp013 [*Gordonia* phage GTE7]	73	1e-60	Phage virion morphogenesis family (pfam05069)
TIN4-*orf13*	10550..11035	162	hypothetical protein GTE7_gp014 [*Gordonia* phage GTE7]	81	7e-93	-
TIN4-*orf14*	11049..11384	112	hypothetical protein GTE7_gp015 [*Gordonia* phage GTE7]	65	4e-21	-
TIN4-*orf15*	11388..12410	341	hypothetical protein GTE7_gp016 [*Gordonia* phage GTE7]	84	0.0	Putative phage structural protein (major tail)
TIN4-*orf16*	12602..13186	195	putative tail assembly protein [*Gordonia* phage GTE7]	81	6e-105	Putative tail assembly protein
TIN4-*orf17*	13168..13449	94	hypothetical protein GTE7_gp018 [*Gordonia* phage GTE7]	72	2E-23	Putative tail assembly protein translated by conserved programmed translational frameshift
TIN4-*orf18*	13557..22286	2910	phage tape measure protein [*Gordonia* phage GTE7]	53	0.0	Tape measure protein (pfam10145; pfam02463; COG5412; pfam01464)
TIN4-*orf19*	22286..24817	844	hypothetical protein GTE7_gp020 [*Gordonia* phage GTE7]	72	0.0	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN4-*orf20*	24817..26565	583	hypothetical protein GTE7_gp021 [*Gordonia* phage GTE7]	78	0.0	-
TIN4-*orf21*	26565..28496	644	hypothetical protein GTE7_gp022 [*Gordonia* phage GTE7]	71	0.0	-
TIN4-*orf22*	28496..29767	424	-			-
TIN4-*orf23*	29776..30039	88	gp80 [*Bacillus* phage G]	46	3e-06	Glutaredoxin (pfam00462)
TIN4-*orf24*	30041..31081	347	hypothetical protein [*Mycobacterium abscessus*]	34	1e-19	-
TIN4-*orf25*	31078..31383	102	hypothetical protein ISGA_3314 [*Gordonia* sp. NB4-1Y]	36	1e-06	-
TIN4-*orf26*	31383..31748	122	hypothetical protein GTE7_gp025 [*Gordonia* phage GTE7]	72	3e-47	-
TIN4-*orf27*	31834..32271	146	hypothetical protein GTE7_gp026 [*Gordonia phage* GTE7]	55	4e-38	-
TIN4-*orf28*	32274..32600	109	hypothetical protein GTE7_gp027 [*Gordonia phage* GTE7]	58	1e-38	-
TIN4-*orf29*	32597..33226	210	lysozyme [*Gordonia* phage GTE7]	76	3e-117	Putative lysozyme
TIN4-*orf30*	33223..34083	287	hypothetical protein TPA2_gp54 [*Tsukamurella* phage TPA2]	58	4e-110	-
TIN4-*orf31*	34106..35002	299	putative peptidase [*Gordonia* phage	75	3e-137	Peptidase (pfam01551)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
			GTE7]			
TIN4-*orf32*	34999..35454	152	hypothetical protein 39HC_013 [*Mycobacterium* phage 39HC]	37	4e-12	Putative holin
TIN4-*orf33*	35442..35942	167	hypothetical protein GTE7_gp030 [*Gordonia* phage GTE7]	45	7e-38	-
TIN4-*orf34*	35935..36297	121	hypothetical protein GTE7_gp031 [*Gordonia* phage GTE7]	43	4e-19	-
TIN4-*orf35*	complement(36421..36645)	75	-			-
TIN4-*orf36*	complement(36642..37838)	399	hypothetical protein GTE7_gp033 [*Gordonia* phage GTE7]	67	0	Nuclease (pfam12705)
TIN4-*orf37*	complement(38017..38442)	142	hypothetical protein GTE7_gp034 [*Gordonia* phage GTE7]	57	1e-20	-
TIN4-*orf38*	complement(38510..38644)	45	-			-
TIN4-*orf39*	complement(38729..38932)	68	-			-
TIN4-*orf40*	complement(38910..39629)	240	DNA methylase [*Gordonia* phage GTE7]	87	4e-154	DNA Methylase (pfam01555)
TIN4-*orf41*	complement(39726..39971)	82	-			-
TIN4-*orf42*	complement(39982..40230)	83	-			-
TIN4-*orf43*	complement(40230..40715)	162	hypothetical protein GTE7_gp041 [*Gordonia* phage GTE7]	53	3e-48	Nucleoside Triphosphate Pyrophosphohydrolase (cd11542)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN4-*orf44*	complement(40712..40912)	67	-			-
TIN4-*orf45*	complement(40909..41475)	189	hypothetical protein GTE7_gp043 [*Gordonia* phage GTE7]	66	9e-88	-
TIN4-*orf46*	complement(41436..41600)	55	-			-
TIN4-*orf47*	complement(41590..41958)	123	hypothetical protein GTE7_gp045 [*Gordonia* phage GTE7]	64	1e-47	Unknown (pfam14359)
TIN4-*orf48*	complement(41955..43091)	379	DNA polymerase III beta subunit [*Gordonia* phage GTE7]	62	0	DNA Polymerase III Beta Subunit (cd00140)
TIN4-*orf49*	complement(43095..43721)	209	exonuclease [*Gordonia* phage GTE7]	68	2e-93	Exonuclease (pfam00929)
TIN4-*orf50*	complement(43718..43903)	62	DNA binding protein [*Gordonia* phage GTE7]	60	5.00E-19	DNA binding protein (pfam12728)
TIN4-*orf51*	complement(43887..44162)	92	-			
TIN4-*orf52*	complement(44169..45590)	474	helicase [*Gordonia* phage GTE7]	77	0.0	Helicase (COG0553)
TIN4-*orf53*	complement(45661..45942)	94	-			-
TIN4-*orf54*	complement(45939..46445)	169	hypothetical protein GTE7_gp051 [*Gordonia* phage GTE7]	37	5e-29	-
TIN4-*orf55*	complement(46615..47658)	348	hypothetical protein GTE7_gp052 [*Gordonia* phage GTE7]	67	9e-162	-
TIN4-*orf56*	complement(47775..48104)	110	hypothetical protein GTE7_gp053	68	3e-39	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
			[*Gordonia* phage GTE7]			
TIN4-*orf57*	complement(48082..48438)	119	hypothetical protein GTE7_gp054 [*Gordonia* phage GTE7]	37	6e-11	-
TIN4-*orf58*	complement(48438..48743)	102	-			-
TIN4-*orf59*	complement(48740..48931)	64	-			-
TIN4-*orf60*	complement(48931..49086)	52	-			-
TIN4-*orf61*	complement(49083..49718)	212	hypothetical protein GTE7_gp056 [*Gordonia* phage GTE7]	61	2e-86	-
TIN4-*orf62*	complement(49715..49975)	87	-			-
TIN4-*orf63*	complement(49968..50501)	178	hypothetical protein GTE7_gp057 [*Gordonia* phage GTE7]	85	4e-17	-
TIN4-*orf64*	complement(50485..50730)	82	hypothetical protein GTE7_gp058 [*Gordonia* phage GTE7]	72	1e-37	-
TIN4-*orf65*	complement(50744..52021)	426	-			-
TIN4-*orf66*	complement(52139..52309)	57	-			-
TIN4-*orf67*	complement(52306..52566)	87	-			-
TIN4-*orf68*	complement(52619..53314)	232	hypothetical protein GTE7_gp062 [*Gordonia* phage GTE7]	55	6e-89	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN4-*orf69*	complement(53295..53447)	51	-			-
TIN4-*orf70*	complement(53449..53850)	134	hypothetical protein ISGA_4751 [*Gordonia* sp. NB4-1Y]	33	4.00E-09	-
TIN4-*orf71*	complement(54257..54679)	141	hypothetical protein GTE7_gp065 [*Gordonia* phage GTE7]	77	1e-68	-
TIN4-*orf72*	complement(54798..55394)	199	hypothetical protein GTE7_gp066 [*Gordonia* phage GTE7]	34	2.00E-18	-
TIN4-*orf73*	complement(55407..55691)	95	-			-
TIN4-*orf74*	complement(55688..55861)	58	-			-
TIN4-*orf75*	complement(55858..56112)	85	-			-
TIN4-*orf76*	complement(56122..56382)	87	hypothetical protein GTE7_gp068 [*Gordonia* phage GTE7]	32	1.00E-04	-
TIN4-*orf77*	complement(56379..56606)	86	unnamed protein product [*Gordonia* phage GRU1]	45	1.00E-05	-
TIN4-*orf78*	complement(56603..57124)	174	-			-
TIN4-*orf79*	complement(57203..59113)	637	hypothetical protein GTE7_gp071 [*Gordonia* phage GTE7]	54	0	Cobalamin biosynthesis protein CobT VWA domain (pfam11775)
TIN4-*orf80*	complement(59117..60709)	531	ATPase family protein [*Gordonia* phage GTE7]	63	0	AAA domain (pfam07728)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN4-*orf81*	complement(60783..61409)	209	hypothetical protein GTE7_gp073 [*Gordonia* phage GTE7]	52	3.00E-73	-
TIN4-*orf82*	complement(61418..61618)	67	hypothetical protein GTE7_gp074 [*Gordonia* phage GTE7]	38	8.00E-04	-
TIN4-*orf83*	complement(61627..61983)	119	hypothetical protein GTE7_gp075 [*Gordonia* phage GTE7]	37	3.00E-17	-
TIN4-*orf84*	complement(62046..62993)	316	hypothetical protein GTE7_gp077 [*Gordonia* phage GTE7]	48	2.00E-77	-
TIN4-*orf85*	complement(62990..63214)	75	-			-
TIN4-*orf86*	complement(63217..63534)	106	-			-
TIN4-*orf87*	complement(63783..64052)	90	-			-
TIN4-*orf88*	complement(64095..64223)	43	-			-
TIN4-*orf89*	complement(64662..65585)	308	hypothetical protein GTE7_gp084 [*Gordonia* phage GTE7]	60	9.00E-107	Phage structural protein
TIN4-*orf90*	complement(65730..65876)	49	hypothetical protein GTE7_gp085 [*Gordonia* phage GTE7]	47	4.00E-05	-
TIN4-*orf91*	complement(66034..66162)	43	-			-
TIN4-*orf92*	complement(66175..66330)	52	-			-
TIN4-*orf93*	complement(66424..66654)	77	-			-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TIN4-*orf94*	complement(66654..67016)	121	-			-
TIN4-*orf95*	complement(67111..67446)	112	-			-
TIN4-*orf96*	complement(67534..67779)	82	-			-
TIN4-*orf97*	complement(68890..69027)	46	-			-
TIN4-*orf98*	complement(69060..69362)	101	-			
TIN4-*orf99*	complement(70027..70488)	154	hypothetical protein GTE7_gp096 [*Gordonia* phage GTE7]	40	4.00E-29	-
TIN4-*orf100*	complement(70485..70814)	110	-			-
TIN4-*orf101*	complement(70811..71134)	108	-			-
TIN4-*orf102*	complement(71131..71379)	83	-			-
TIN4-*orf103*	complement(71376..71654)	93	hypothetical protein GTE7_gp097 [*Gordonia* phage GTE7]	53	7.00E-24	-
TIN4-*orf104*	complement(71651..73903)	751	putative primase [*Gordonia* phage GTE7]	44	0	Putative primase (PRK11633)
TIN4-*orf105*	complement(73900..74124)	75	hypothetical protein GTE7_gp099 [*Gordonia* phage GTE7]	55	8.00E-16	-
TIN4-*orf106*	complement(74102..74233)	44	-			-
TIN4-*orf107*	complement(74226..74720)	165	hypothetical protein GTE7_gp101	69	1.00E-22	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
			[*Gordonia* phage GTE7]			
TIN4-*orf108*	complement(74717..75310)	198	endonuclease VII [*Streptomyces* sp. NRRL S-920]	31	1.00E-16	Endonuclease (pfam02945)
TIN4-*orf109*	complement(75297..76268)	324	hypothetical protein GTE7_gp103 [*Gordonia* phage GTE7]	68	6.00E-149	Unknown (COG4951)

[a] ORFs were numbered consecutively; [b] The most closely related gene (only if named) and the name of the organism; [c] Percentage identity is based on the best match when a BLAST P analysis is performed; [d] The probability of obtaining a match by chance as determined by BLAST analysis and only values less than 10^{-4} were considered significant; [e] Predicted function is based on amino acid identity, conserved motifs, and gene location within functional modules.

Table 2.3 Palindromes identified in the genome sequences of phages TIN2, TIN3, and TIN4

Phage-Palindrome number	Size (bp)	Coordinates	Sequence alignment	Position in genome
TIN2-P1	54	55581-55630	GCGTTCGTGCCCGTACCCGTTAATGGTACCA----CGGACACGGGCACGAACGC	Between *orf71* and *orf72*
		55630-55581	GCGTTCGTGCCCGTGTCCGT----GGTACCATTAACGGGTACGGGCACGAACGC	
TIN2-P2	40	33809-33847	CGGGCCTCGATCAGCGC-GATCAACGCTGATCGAGACCCG	Within *orf30*
		33847-33809	CGGGTCTCGATCAGCGTTGATCG-CGCTGATCGAGGCCCG	
TIN2-P3	40	55020-55059	GAATGACCCGAGCAGCGCTCCATTACTGCTCGGGTCATTC	Between *orf70* and *orf71*
		55059-55020	GAATGACCCGAGCAGTAATGGAGCGCTGCTCGGGTCATTC	
TIN2-P4	36	13362-13397	CACGAAGGGTAGTGCTTGCGAGCACTACCCTTCGTG	Between *orf17* and *orf18*
		13397-13362	CACGAAGGGTAGTGCTCGCAAGCACTACCCTTCGTG	
TIN2-P5	36	36772-36807	GAAAGCGGACCACGTCGTTTGACGTGGTCCGCTTTC	Within *orf35*
		36807-36772	GAAAGCGGACCACGTCAAACGACGTGGTCCGCTTTC	
TIN2-P6	32	71026-71057	AAGGCCCGGCTAGTTAGTACCAGCCGGGCCTT	Between *orf98* and *orf99*
		71057-71026	AAGGCCCGGCTGGTACTAACTAGCCGGGCCTT	
TIN2-P7	31	36493-36523	CGAGGCCTCCTCGCTTCGCGAGGAGGCCTCG	Within *orf34*
		36523-36493	CGAGGCCTCCTCGCGAAGCGAGGAGGCCTCG	

Phage-Palindrome number	Size (bp)	Coordinates	Sequence alignment	Position in genome
TIN2-P8	29	47046-47074	AGGGGGAGGCGTTTCCAGCGCCTCCCCCT	Between *orf53* and *orf54*
		47074-47046	AGGGGGAGGCGCTGGAAACGCCTCCCCCT	
TIN2-P9	28	65687-65714	CCCGCGCTACCTCATTCGGTAGCGCGGG	Between *orf90* and *orf91*
		65714-65687	CCCGCGCTACCGAATGAGGTAGCGCGGG	
TIN2-P10	18	71657-71674	CTCGAGAAGCTTCTCGAG	Within *orf100*
		71674-71657	CTCGAGAAGCTTCTCGAG	
TIN2-P11	16	39306-39321	TCCTTCGTACGAAGGA	Between *orf39* and *orf40*
		39321-39306	TCCTTCGTACGAAGGA	
TIN2-P12	16	62769-62784	TCACCCGCGCGGGTGA	Within *orf84*
		62784-62769	TCACCCGCGCGGGTGA	
TIN3-P1	42	13449-13490	ACACACGAAGGGTAGTGCTTGCGAGCACTACCCTTCGTGTGT	Overlapping the end of *orf17* by 1 nucleotide
		13490-13449	ACACACGAAGGGTAGTGCTCGCAAGCACTACCCTTCGTGTGT	
TIN3-P2	41	12437-12475	CCGCCCCTGGTCGTG--TTGTCAAGTCGCGACCAGGGGCGG	Between *orf15* and *orf16*
		12475-12437	CCGCCCCTGGTCGCGACTTGACAA--CACGACCAGGGGCGG	
TIN3-P3	40	33368-33406	CGGGCCTCGATCAGCGC-GATCAACGCTGATCGAGACCCG	Within *orf30*
		33406-33368	CGGGTCTCGATCAGCGTTGATCG-CGCTGATCGAGGCCCG	

Phage-Palindrome number	Size (bp)	Coordinates	Sequence alignment	Position in genome
TIN3-P4	40	53910-53948	GAATAGCCCGAGCAGCGATCCATA-CTGCTCGGGCTATTC	Between *orf70* and *orf71*
		53948-53910	GAATAGCCCGAGCAGT-ATGGATCGCTGCTCGGGCTATTC	
TIN3-P5	36	36315-36349	GGAAAGCGGACCGCGTCTAG-CGTGGTCCGCTTTCC	Between *orf34* and *orf35*
		36349-36315	GGAAAGCGGACCACG-CTAGACGCGGTCCGCTTTCC	
TIN3-P6	31	36037-36067	CGAGGCCTCCTCCTTGAAGGAGGAGGCCTCG	Within *orf34*
		36067-36037	CGAGGCCTCCTCCTTCAAGGAGGAGGCCTCG	
TIN3-P7	29	46476-46504	AGGGGGAGGCGTTTCCAGCGCCTCCCCCT	Between *orf54* and *orf55*
		46504-46476	AGGGGGAGGCGCTGGAAACGCCTCCCCCT	
TIN3-P8	22	53095-53116	ACTCGATCGAGCTCGATCCAGT	Within *orf68*
		53116-53095	ACTGGATCGAGCTCGATCGAGT	
TIN3-P9	16	36743-36758	TTGAGAAGCTTCTCAA	Within *orf36*
		36758-36743	TTGAGAAGCTTCTCAA	
TIN3-P10	16	50249-50264	TCGAGTGCGCACTCGA	Within *orf63*
		50264-50249	TCGAGTGCGCACTCGA	
TIN3-P11	16	64755-64770	CACCCGAGCTCGGGTG	Within *orf89*
		64770-64755	CACCCGAGCTCGGGTG	

Phage-Palindrome number	Size (bp)	Coordinates	Sequence alignment	Position in genome
TIN4-P1	42	13397-13438	ACACACGAAGGGTAGTGCTTGCGAGCACTACCCTTCGTGTGT	Overlapping the end of *orf17* by one nucleotide
		13438-13397	ACACACGAAGGGTAGTGCTCGCAAGCACTACCCTTCGTGTGT	
TIN4-P2	41	12385-12423	CCGCCCCTGGTCGTG--TTGTCAAGTCGCGACCAGGGGCGG	Between *orf15* and *orf16*
		12423-12385	CCGCCCCTGGTCGCGACTTGACAA--CACGACCAGGGGCGG	
TIN4-P3	40	33315-33353	CGGGCCTCGATCAGCGC-GATCAACGCTGATCGAGACCCG	Within *orf30*
		33353-33315	CGGGTCTCGATCAGCGTTGATCG-CGCTGATCGAGGCCCG	
TIN4-P4	40	53857-53895	GAATAGCCCGAGCAGCGATCCATA-CTGCTCGGGCTATTC	Between *orf70* and *orf71*
		53895-53857	GAATAGCCCGAGCAGT-ATGGATCGCTGCTCGGGCTATTC	
TIN4-P5	36	36262-36296	GGAAAGCGGACCGCGTCTAG-CGTGGTCCGCTTTCC	Between *orf34* and *orf35*
		36296-36262	GGAAAGCGGACCACG-CTAGACGCGGTCCGCTTTCC	
TIN4-P6	31	35984-36014	CGAGGCCTCCTCCTTGAAGGAGGAGGCCTCG	Within *orf34*
		36014-35984	CGAGGCCTCCTCCTTCAAGGAGGAGGCCTCG	
TIN4-P7	29	46423-46451	AGGGGGAGGCGTTTCCAGCGCCTCCCCCT	Between *orf54* and *orf55*
		46451-46423	AGGGGGAGGCGCTGGAAACGCCTCCCCCT	
TIN4-P8	22	53042-53063	ACTCGATCGAGCTCGATCCAGT	Within *orf68*
		53063-53042	ACTGGATCGAGCTCGATCGAGT	

Phage-Palindrome number	Size (bp)	Coordinates	Sequence alignment	Position in genome
TIN4-P9	16	36690-36705	TTGAGAAGCTTCTCAA	Within *orf36*
		36705-36690	TTGAGAAGCTTCTCAA	
TIN4-P10	16	50196-50211	TCGAGTGCGCACTCGA	Within *orf63*
		50211-50196	TCGAGTGCGCACTCGA	
TIN4-P11	16	64702-64717	CACCCGAGCTCGGGTG	Within *orf89*
		64717-64702	CACCCGAGCTCGGGTG	

Table 2.4 Summary of phage structural genes identified using mass spectrometry

Phage	Protein containing motifs	Amino acid sequence	Coverage	Present in other TIN phages?
TIN2	Orf7	SAQPGSYTVVSK	72.5 %	Orf7 of phages TIN3 & TIN4
		TVSSAAFPVEVVDGHDQK		-
		ILQEGDVLVALATGPNAGK		-
		VVPFQVGVAEDLATLVGVTK		Orf7 of phages TIN3 & TIN4
		DYFGWELNER		Orf7 of phages TIN3 & TIN4
		DVEAGVLVR		Orf7 of phages TIN3 & TIN4
		VAVTNAVADALR		Orf7 of phages TIN3 & TIN4
		KNLDILFF		Orf7 of phages TIN3 & TIN4
	Orf8	LKDHYDAADVQR	28.1 %	Orf8 of phages TIN3 & TIN4
		RPVDQQAQAPASGSYASTTHDPIGDINK		Orf8 of phages TIN3 & TIN4
		FLLSLVNSDKFILR		Orf8 of phages TIN3 & TIN4
		GVPATDIGTTGVALYR		-
		GLTTGLAPAIAESAESELAQK		Orf8 of phages TIN3 & TIN4
		VIFLPDDNAMSEYDSSPIGLGK		-
	Orf15	ELMIPDAEIGGGR	18.8 %	Orf15 of phages TIN3 & TIN4
		DVPDALLGPVSFGGDIEFYTR		Orf15 of phages TIN3 & TIN4
		VLYDGADIK		Orf15 of phages TIN3 & TIN4
		LGSFYLEDLTPK		Orf15 of phages TIN3 & TIN4

Phage	Protein containing motifs	Amino acid sequence	Coverage	Present in other TIN phages?
		EVNGSLTLR		Orf15 of phages TIN3 & TIN4
TIN4	Orf7	VLQEGDVLVALTTGPNAGK	31.5 %	Orf7 of phage TIN3
		TVSSAAFPVEKIDGFDQK		Orf7 of phage TIN3
		DYFGWELNER		Orf7 of phage TIN3
	Orf8	LEWLTMQSLSNGGIAYNDGNI	22.9 %	Orf8 of phage TIN3
		GLTTGLAPAIAESAESELAQK		Orf8 of phage TIN3
		VIFLPDDNAMAEYDSSPIGLGK		Orf8 of phage TIN3
		MLTSPHSMGNGAAGFYDWEQDTTDPWGK		Orf8 of phage TIN3
	Orf89	AVAEADADTGTVTEANLDAVSQAYR	22.7 %	Orf89 of phage TIN3
		EVLVDTADDEVADFEPIDTDLATR		Orf89 of phage TIN3
		QAAADAEHLPAVDAFTETVEK		Orf89 of phage TIN3

3. Isolation and characterisation of the bacteriophages SPI1, which infects the activated-sludge-foaming bacterium Skermania piniformis

3.1. Abstract

Foaming in activated sludge plants is a worldwide problem caused by the proliferation of members of the Mycolata bacteria. These include *Skermania piniformis,* which has been documented as a major causative foaming bacteria globally, and particularly in Australian treatment plants. Phage SPI1 is the first phage isolated and fully characterised genetically that is infective for this organism. It targets only seven of the nine strains of *S. piniformis* held in our culture collection, and none of the other 94 Mycolata strains of different genera, mostly isolated from wastewater, against which it was tested. This SPI1 phage is a member of the family *Siphoviridae* and has a highly novel, circularly permuted, dsDNA genome that is 55, 748 bp in size with a G+C content of 67.8 mol %. This phage appears to be lytic with no evidence of genes related to a lysogenic mode of existence.

3.2. Introduction

Skermania piniformis was referred to initially as the pine-tree like organism (PTLO) (Blackall & Marshall, 1989; Chun *et al.*, 1997) because of its distinctive acute angled branching morphology, which can vary among isolates in both the branching angles and interbranch lengths (Soddell & Seviour, 1998). It was first named formally as *Nocardia pinensis* (Blackall *et al.*, 1989), but was later reclassified as a member of a new genus *Skermania* (Chun *et al.*, 1997). *S. piniformis* is a common causative agent of stable foams and scums developing on the surfaces of aerobic reactors in activated sludge plants worldwide, and for reasons still unclear, particularly, those in Australia (Blackall *et al.*, 1991; Seviour *et al.*, 1990; Seviour *et al.*, 1994; Soddell & Seviour, 1998). As with other members of the Mycolata associated with foam stabilisation, *S. piniformis* has a strongly hydrophobic outer membrane of mycolic acids rendering its cells strongly hydrophobic (Chun *et al.*, 1997). Formation of stable foams is a global problem, impacting negatively on plant aesthetics, increasing maintenance costs, and complicating sludge management (de los Reyes III, 2010; Goddard & Forster, 1987; Jenkins *et al.*, 1993).

Foaming is a flotation event, and requires air bubbles, surface active agents, and hydrophobic particles, in this case the bacteria. Any control strategy should be directed at the bacterial component because neither air bubbles nor detergents can be eliminated from the process (Petrovski *et al.,* 2011d). Currently, no universal method is known that controls rationally all foams once formed, a situation probably reflecting our present poor understanding of the bacteria involved.

While previous studies (Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012b, c; Petrovski *et al.*, 2013a, b, 2014; Thomas *et al.*, 2002) have isolated phages infective for many members of the foaming Mycolata, no phage has been recorded yet for *S. piniformis*. In this study, we report for the first time the isolation and genomic characterisation of a lytic phage infective for *Skermania piniformis,* and assess its suitability for foam biocontrol.

3.3. Materials and methods

3.3.1. Bacterial strains used in this study

The bacterial strains used (Table 8.1) and methods for their growth are described by Petrovski *et al.* (2011a). The exception to this are bacterial strains of *S. piniformis* (NM40^T, NM41, NM101, NM109, NM168, J8, J20, J50, and J54, all code numbers for *S. pinensis* strains held in the La Trobe University culture collection) which were grown on homemade R2A medium (0.5 g/L Yeast extract (Oxoid, Adelaide, Australia), 0.5 g/L Proteose peptone (Difco, North Ryde, Australia), 0.5 g/L Casamino acid (Difco, North Ryde, Australia), 0.5 g/L Glucose, 0.5 g/L soluble starch (Difco, North Ryde, Australia), 0.3 g/L K_2HPO_4, 0.005 g/L $MgSO_4.7H_2O$, 0.3 g/L sodium pyruvate (BDH, Murarrie, Australia)) broth and agar (14 g/L) R2A (Oxoid, Adelaide, Australia) at 25°C. All remaining chemicals were obtained from Sigma (Sydney, Australia) unless otherwise noted.

3.3.2. Phage isolation, purification, and characterisation

Phages were isolated and subsequently purified from activated sludge samples as described by Petrovski *et al.* (2011a), using the enrichment pool containing the three *S. piniformis* strains listed in Table 3.1. Host range determinations were carried out as described by Petrovski *et al.* (2011a). Carbon and formvar grids (Electron Microscopy Sciences, Australia), for visualisation of virions were stained with uranyl acetate as described by Petrovski *et al.* (2011a) prior to examination using a JEOL JEM-2010HC Electron Microscope.

Table 3.1 Strains used for the pooled enrichment isolation of phage SPI1

Organism	Lab ID
Skermania piniformis	NM40^T
Skermania piniformis	NM41
Skermania piniformis	J20

^{T} *indicates type strain.*

3.3.3. Genome sequencing and annotation

The genome of phage SPI1 was sequenced as described by Petrovski *et al.* (2011a) and *de novo* assembly was performed using CLC workbench (v6.5.1). Genome Open Reading Frames (ORFs) within the assembled sequence with a minimum size of 90 bp (Delcher *et al.*, 2007) were detected initially using Glimmer (v3.02). All predicted 'start' codons were inspected manually for the presence of putative ribosomal binding sites and corrected as required. Sequence similarity searches were carried out against sequences held in the GenBank database, as described by Petrovski *et al.* (2011c). The presence of tRNA and tmRNA were determined by using tRNAScan-SE (Schattner *et al.*, 2005) and ARAGORN (Laslett & Canback, 2004). Transmembrane domains were predicted using the DAS Transmembrane Prediction Server (Cserzo *et al.*, 1997).

Phage DNA when analysed by gel electrophoresis gave results consistent with circularly permuted DNA genomes.

3.3.4. Nucleotide sequence accession number

The nucleotide sequence for *Skermania piniformis* phage SPI1 has been deposited in GenBank under accession number KR011061.

3.4. Results and discussion

3.4.1. Isolation and preliminary characterisation

An activated sludge sample containing a high number of PTLO morphotype filaments collected from Albury, NSW, Australia was screened for the presence of *S. piniformis* phages by multiple host enrichment (Table 3.1) and plaque plating as described by Petrovski *et al.* (2011a). Single plaques were obtained on lawn plates of *S. piniformis* strain NM40^T that were ~0.5 mm in diameter. This phage was plaque purified and named here as the SPI1 phage. To our knowledge this is the first *Skermania* phage to be described.

Phage SPI1 belongs to the *Siphoviridae* family, possessing the characteristic long, non-contractile tail (~239 nm) of members of the *Caudovirales* with a B1 isometric capsid (~60 nm) morphotype (Figure 3.1). When screened against 94 Mycolata strains (see Table 8.1) isolated mainly from activated sludge treatment plants, plaques were produced on seven of the nine *Skermania piniformis* strains screened including the type strain (NM40^T, NM101, NM109, NM168, J8, J50, and J54). This level of strain specificity has been noted for other actinophages (Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012b, c; Petrovski *et al.*, 2013a, b, 2014), and probably reflects the absence of a suitable receptor site, or action of bacterial phage resistance systems like Clustered Regularly Interspersed Repeat (CRISPR) systems, Restriction Modification (RM) systems, or

abortive infection (Abi) systems in these resistant strains (Labrie *et al.*, 2010). Mock infection controls were performed with each strain to ensure that cell lysis was not a consequence of spontaneous release of prophages from the susceptible host strains.

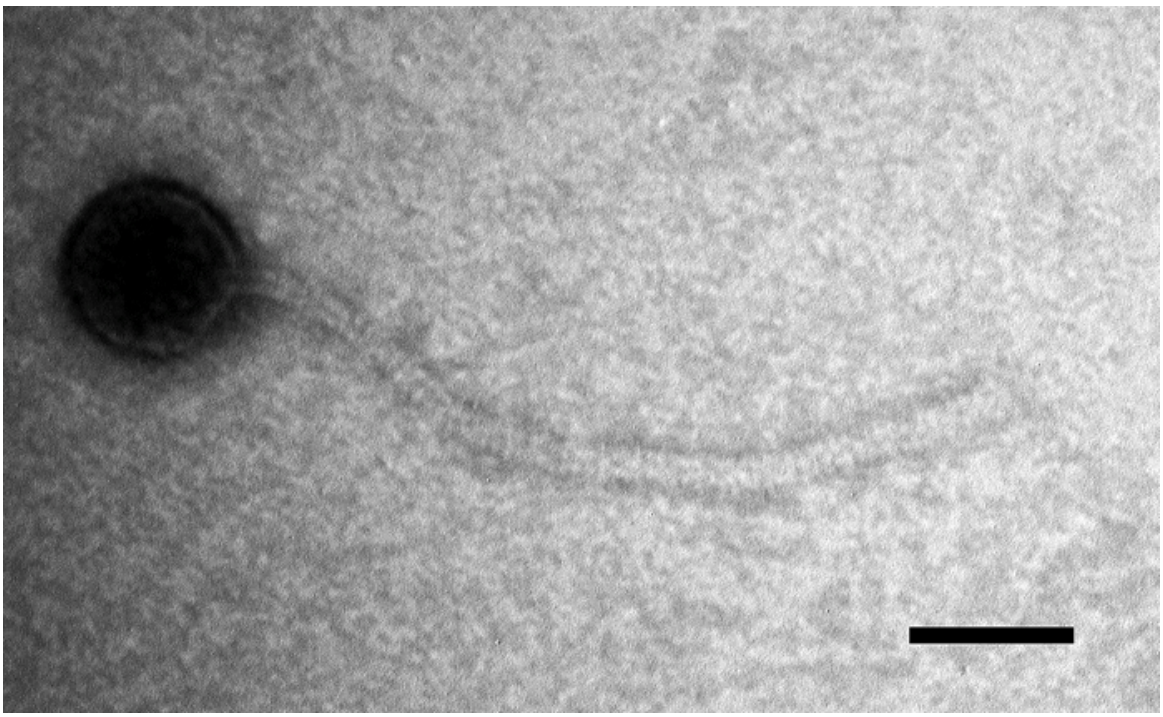

Figure 3.1 Electron micrograph of phage SPI1.

Scale = 50 nm. Negatively stained electron microscopy, 2 % uranyl acetate.

3.4.2. Genetic features of phage SPI1

The genome sequence of phage SPI1 was obtained using Roche/454 pyrosequencing. This generated 15,051 reads, with ~99-fold average sequence coverage following assembly. This phage possesses a novel, circularly permuted, dsDNA genome 55,748 bp in size. Phage SPI1 has a G+C content of 67.8 mol %, which is close to that of its host *S. piniformis* (67.5 mol %) (Chun *et al.*, 1997), suggesting SPI1 is well adapted to its host. At the DNA level this genome has a novel sequence, and therefore comparisons of putative open reading frames were based on their predicted expressed amino acid sequences. Analysis of this genome sequence revealed sixty-seven putative ORFs and detected no putative tRNA, or tmRNA. These ORFs are numbered

consecutively, with 51 oriented in a forward orientation, and 16 in a reverse orientation (Figure 3.2). While 47 ORFs showed statistically significant identity with those previously reported, only eighteen could be annotated functionally (Table 3.2). Eighteen ORFs (27%) exhibited no statistically significant identity to any hypothetical protein.

The SPI1 phage genome is modularly organised, as those in other actinophages are, consisting of DNA packaging, structural, cell lysis and DNA replication/maintenance modules (Figure 3.2). In the absence of a putative origin of replication, the ORFs were ordered from the beginning with gene *orf1* which is located immediately upstream of the packaging module discussed later, overlapping *orf2* the putative small terminase in the same transcriptional direction in what appears to be an operon-like structure.

3.4.3. Sequence repeats

A large number of repeat structures were observed in the genome sequence of phage SPI1, among which were twenty-five palindromes (Table 3.3), some of which were located in intergenic positions. These intergenic palindromes might function as *rho*-independent transcriptional terminators, although not all downstream flanking sequences displayed the T-rich region typical of those reported in *Escherichia coli* (Lesnik *et al.*, 2001). In *Streptomyces lividans* (Deng *et al.*, 1987) this T-rich region is not required for transcriptional termination, so what role, if any, these palindromic sequences might play in phage SPI1 remains to be determined. Both *orf12* and *orf22* contain four and two palindromes respectively, although again what functions these might play remain unclear.

Also present in the SPI1 genome are large numbers of other repeat sequences, with 110 direct repeats and 80 inverted repeats ranging from 15 to 347 bp in size (see Table 8.2 and Table 8.3 in section 8.2). Repeat sequences are associated commonly with repeated protein structures, promoters, transposable elements and replication origins (Mott & Berger, 2007), but neither could be identified in this genome. Some of these repeats are quite large in size and in number, compared to those seen in other actinophage genomes (Petrovski *et al.*, 2011a, c; Petrovski *et al.*, 2012b) suggesting that these might have functional roles as yet unknown.

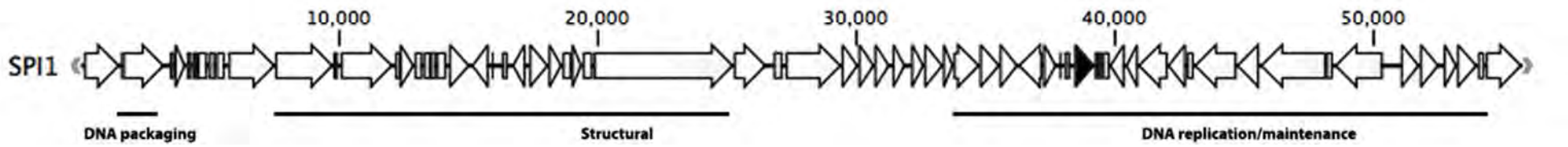

Figure 3.2 Genome map of phage SPI1.

Genes shaded in black indicate lysis module genes.

Table 3.2 Genome annotation of phage SPI1

ORF[a]	Coordinates	Size (aa)	Significant match[b]	% Identity [c]	E value [d]	Putative function (conserved motif) [e]
orf1	125..1504	460	-			
orf2	1501..1629	43	-			Putative small terminase
orf3	1626..3227	534	hypothetical protein [*Rhodococcus ruber*]	55	0.0	Large terminase (pfam03237)
orf4	3435..3623	63	hypothetical protein [*Rhodococcus rhodochrous*]	39	2e-06	
orf5	3620..4051	144	hypothetical protein TPA2_gp76 [*Tsukamurella* phage TPA2]	47	3e-09	
orf6	4097..4282	62	-			
orf7	4339..4503	55	-			
orf8	4508..4885	126	-			
orf9	4947..5195	83	-			
orf10	5232..5567	112	gp10 [*Mycobacterium* phage Pipefish]	39	2e-14	Unknown (pfam07098)
orf11	5702..7516	605	hypothetical protein [*Rhodococcus rhodochrous*]	47	6e-162	
orf12	7513..9762	750	hypothetical protein [*Rhodococcus rhodochrous*]	39	4e-145	Putative capsid protein (pfam04233)

ORF[a]	Coordinates	Size (aa)	Significant match[b]	% Identity [c]	E value [d]	Putative function (conserved motif) [e]
orf13	9762..9938	59	hypothetical protein TPA2_gp19 [*Tsukamurella* phage TPA2]	42	2e-04	
orf14	10048..12063	672	hypothetical protein [*Rhodococcus ruber*]	49	4e-174	
orf15	12149..12370	74	putative Gp13 [*Nocardia cyriacigeorgica*]	52	3e-11	
orf16	12322..12894	191	putative Gp13 [*Nocardia cyriacigeorgica*]	47	4e-36	
orf17	12909..13205	99	-			
orf18	13257..13538	94	holin [*Rhodococcus* phage ReqiDocB7]	40	2e-07	
orf19	13510..13746	79	-			
orf20	13743..14132	130	-			
orf21	14233..15042	270	hypothetical protein [*Rhodococcus ruber*]	70	5e-133	
orf22	complement(15110..15778)	223	-			
orf23	complement(15852..16001)	50	-			
orf24	complement(16278..16502)	75	-			
orf25	complement(16676..17197)	174	gp46 [*Mycobacterium* phage Acadian]	42	1e-26	
orf26	17337..18080	248	hypothetical protein [*Rhodococcus rhodochrous*]	54	2e-74	

ORF[a]	Coordinates	Size (aa)	Significant match[b]	% Identity [c]	E value [d]	Putative function (conserved motif) [e]
orf27	18082..18612	177	hypothetical protein [*Rhodococcus ruber*]	54	6e-51	
orf28	18626..18979	118	hypothetical protein [*Rhodococcus rhodochrous*]	38	1e-07	
orf29	18972..19376	135	tail assembly chaperone [*Mycobacterium* phage Jolie1]	47	1e-24	Putative tail assembly protein
orf30	19402..19797	132	hypothetical protein TPA2_gp32 [*Tsukamurella* phage TPA2]	39	3e-09	
orf31	19838..25240	1801	hypothetical protein [*Nocardia otitidiscaviarum*]	38	3e-180	Putative tape measure protein (TIGR01760; COG5412)
orf32	25249..26508	420	hypothetical protein [*Rhodococcus ruber*]	54	4e-163	
orf33	26804..27139	112	hypothetical protein [Rhodococcus ruber]	77	3e-18	
orf34	27255..29411	719	hypothetical protein [*Rhodococcus ruber*]	50	2e-115	
orf35	29414..30061	216	hypothetical protein [*Rhodococcus ruber*]	37	8e-38	
orf36	30071..30691	207	hypothetical protein [*Rhodococcus ruber*]	50	3e-65	
orf37	30688..31281	198	hypothetical protein [*Rhodococcus rhodochrous*]	48	2e-55	
orf38	31386..31886	167	hypothetical protein [*Rhodococcus rhodochrous*]	38	4e-27	

ORF[a]	Coordinates	Size (aa)	Significant match[b]	% Identity [c]	E value [d]	Putative function (conserved motif) [e]
orf39	32109..32612	168	-			
orf40	32609..33346	246	putative uncharacterized protein [*Rhodococcus* sp. AW25M09]	45	3e-10	
orf41	33333..33746	138	hypothetical protein [*Gordonia paraffinivorans*]	33	1e-04	
orf42	33743..34762	340	putative uncharacterized protein [*Rhodococcus* sp. AW25M09]	38	2e-36	CRISPR associated RAMP superfamily protein Csf2 (cd09706)
orf43	34759..35541	261	hypothetical protein [*Mycobacterium smegmatis*]	32	3e-13	
orf44	35523..36242	240	hypothetical protein [*Rhodococcus imtechensis*]	37	1e-23	Phosphoadenosine phosphosulfate reductase family (pfam01507)
orf45	complement(36247..37182)	312	MULTISPECIES: ribonucleotide-diphosphate reductase subunit beta [*Rhodococcus*]	89	0.0	Ribonucleotide reductase (pfam00268)
orf46	37216..37737	174	hypothetical protein [*Rhodococcus rhodochrous*]	31	2e-06	
orf47	37803..37958	52	-			
orf48	complement(38040..38258)	73	hypothetical protein [*Gardnerella vaginalis*]	52	3e-11	Dehydrogenase (PRK08324)
orf49	38423..39205	261	hypothetical protein [*Rhodococcus* sp. 29MFTsu3.1]	63	8e-92	Lysin (pfam13539)

ORF[a]	Coordinates	Size (aa)	Significant match[b]	% Identity [c]	E value [d]	Putative function (conserved motif) [e]
orf50	39202..39486	95	hypothetical protein [*Micromonospora parva*]	48	2e-05	Putative Holin
orf51	39515..39790	92	hypothetical protein PBI_BERNAL13_1 [*Mycobacterium* phage Bernal13]	42	3e-06	
orf52	complement(39778..40371)	198	RuvC-like resolvase superfamily protein [*Rhodococcus* phage E3]	40	6e-32	
orf53	complement(40401..40883)	161	gp48 [*Mycobacterium* phage Daisy]	48	2e-24	
orf54	complement(40880..42031)	384	gp49 [*Mycobacterium* phage Arbiter]	36	2e-54	Nuclease (pfam12705)
orf55	complement(42028..42738)	237	putative uncharacterized protein [*Rhodococcus* sp. AW25M09]	33	7e-08	
orf56	complement(42775..43041)	89	-			
orf57	complement(43057..44694)	546	helicase [*Tsukamurella* phage TPA2]	52	1e-160	Helicase (COG1061)
orf58	complement(44691..45566)	292	-			
orf59	complement(45615..48164)	850	primase [*Tsukamurella* phage TPA2]	40	2e-171	Primase/polymerase (pfampfam09250; pfam13481)
orf60	complement(48158..48418)	87	-			DNA binding (cd00569)
orf61	complement(48488..50353)	622	gp58 [*Mycobacterium* phage Acadian]	47	4e-163	DNA polymerase I - 3'-5'

ORF[a]	Coordinates	Size (aa)	Significant match[b]	% Identity [c]	E value [d]	Putative function (conserved motif) [e]
						exonuclease (COG0749)
orf62	51044..51766	241	-			
orf63	51783..52529	249	-			
orf64	52705..53217	171	hypothetical protein [*Nocardia cyriacigeorgica*]	36	3e-14	
orf65	53214..54002	263	-			
orf66	54033..54275	81	type B dihydrofolate reductase DfrB6 [*Salmonella enterica* subsp. enterica serovar Infantis]	55	4e-09	R67 dihydrofolate reductase (pfam06442)
orf67	54335..55735	467	hypothetical protein [*Corynebacterium falsenii*]	33	6e-28	Pentapeptide repeats (pfam13599;pfam13599)

[a] ORFs were numbered consecutively; [b] The most closely related gene (only if named) and the name of the organism; [c] Percentage identity is based on the best match when a BLAST P analysis is performed; [d] The probability of obtaining a match by chance as determined by BLAST analysis, only values less than 10^{-4} were considered significant; [e] Predicted function is based on amino acid identity, conserved motifs, and gene location within functional modules.

Table 3.3 Palindrome sequences identified in the genome of phage SPI1

Palindrome number	Size (bp)	Coordinates	Sequence alignment	Position in genome
P1	63	54251-54313	GTTGGTGAGCGTCCCTCCACTATAGCACGGTGACCGTGCTATAGTGGAGGGACGCTCACCAAC	Overlapping the 3' end of *orf66*
		54313-54251	GTTGGTGAGCGTCCCTCCACTATAGCACGGTCACCGTGCTATAGTGGAGGGACGCTCACCAAC	
P2	46	12070-12115	GGGGAACCGGGCGGGCGGCACGATCGCTTCGCCCGCCCGGCTCCCC	Between *orf14* and *orf15*
		12115-12070	GGGGAGCCGGGCGGGCGAAGCGATCGTGCCGCCCGCCCGGTTCCCC	
P3	40	51772-51811	GTAACGGTCGCATGTGGCTTACGGCACATGCGACCGTTAC	Overlapping the 5' end of *orf63*
		51811-51772	GTAACGGTCGCATGTGCCGTAAGCCACATGCGACCGTTAC	
P4	40	15057-15096	AACGGGCACGGTATCCGAAAGGACGGATACCGTGCCCGTT	Between *orf21* and *orf22*
		15096-15057	AACGGGCACGGTATCCGTCCTTTCGGATACCGTGCCCGTT	
P5	38	9359-9396	GCGTACCCGCTCCAGGAGCCAGCCCGGAGCGGGTACGC	Within *orf12*
		9396-9359	GCGTACCCGCTCCGGGCTGGCTCCTGGAGCGGGTACGC	
P6	37	50457-50493	GAACGGGCTCCCCCTCTCTCCGCGGGGGAGCCCGTTC	Between *orf61* and *orf62*
		50493-50457	GAACGGGCTCCCCCGCGGAGAGAGGGGGAGCCCGTTC	
P7	31	3349-3379	CGATACTAGCACGGCGACCGTGCTAGTATCG	Between *orf3* and *orf4*
		3379-3349	CGATACTAGCACGGTCGCCGTGCTAGTATCG	

Palindrome number	Size (bp)	Coordinates	Sequence alignment	Position in genome
P8	31	6204-6234	TCTTTCGGGTGTGGGACCCGCACCCGAAAGA	Within *orf11*
		6234-6204	TCTTTCGGGTGCGGGTCCCACACCCGAAAGA	
P9	30	9497-9526	TGGTCGCGTCCGGAGCGACGGACGCGACCA	Within *orf12*
		9526-9497	TGGTCGCGTCCGTCGCTCCGGACGCGACCA	
P10	30	8497-8526	GCCGTCATCGGTACACCGACCGATGACGGC	Within *orf12*
		8526-8497	GCCGTCATCGGTCGGTGTACCGATGACGGC	
P11	29	16078-16106	CACTATAGCACGGTCACCGTGCTATAGTG	Between *orf23* and *orf24*
		16106-16078	CACTATAGCACGGTGACCGTGCTATAGTG	
P12	29	1654-1682	ACGCTCAGCGGTACCGCGCCGCTGAGCGT	Within *orf3*
		1682-1654	ACGCTCAGCGGCGCGGTACCGCTGAGCGT	
P13	25	7910-7934	CGGACGCGCTCGACGAGCGCGTCCG	Within *orf12*
		7934-7910	CGGACGCGCTCGTCGAGCGCGTCCG	
P14	23	4278-4300	AGTAGCACGGTGACCGTGCTACT	Overlapping the 3’ end of *orf6*
		4300-4278	AGTAGCACGGTCACCGTGCTACT	
P15	22	46302-46323	CACGCTCGCGCGCGAGCGTG	Within *orf59*

Palindrome number	Size (bp)	Coordinates	Sequence alignment	Position in genome
		46323-46302	CACGCTCGCGCGCGAGCGTG	
P16	22	15378-15399	TCTTCCGCGAGCTCGCGGAGGA	Within *orf22*
		15399-15378	TCCTCCGCGAGCTCGCGGAAGA	
P17	20	15197-15216	CAGCGTGGCATGCCACGCTG	Within *orf22*
		15216-15197	CAGCGTGGCATGCCACGCTG	
P18	20	40924-40943	GGGCCAAGATATCTTGGCCC	Within *orf54*
		40943-40924	GGGCCAAGATATCTTGGCCC	
P19	20	48547-48566	GACCGCCCGCGCGGGCGATC	Within *orf61*
		48566-48547	GATCGCCCGCGCGGGCGGTC	
P20	18	43017-43034	TATCTATCCGGATAGATA	Within *orf56*
		43034-43017	TATCTATCCGGATAGATA	
P21	16	2618-2633	CCGGACCCGGGTCCGG	Within *orf3*
		2633-2618	CCGGACCCGGGTCCGG	
P22	16	7154-7169	GGTCCGCCGGCGGACC	Within *orf11*
		7169-7154	GGTCCGCCGGCGGACC	

Palindrome number	Size (bp)	Coordinates	Sequence alignment	Position in genome
P23	16	23992-24007	GGTCATCGCGATGACC	Within *orf31*
		24007-23992	GGTCATCGCGATGACC	
P24	16	45388-45403	GTGTCTCCGGAGACAC	Within *orf58*
		45403-45388	GTGTCTCCGGAGACAC	
P25	16	46832-46847	ACATCGACGTCGATGT	Within *orf59*
		46847-46832	ACATCGACGTCGATGT	

3.4.4. DNA Packaging Module

The gene *orf3* appears to encode the large terminase enzyme involved in packaging of phage DNA (Rao & Feiss, 2008). A putative small terminase subunit gene *orf2*, was identified upstream of *orf3* and is transcribed in the same direction based on its location. Orf3 shows amino acid sequence similarity to a hypothetical protein from *Rhodococcus ruber* and several other phage terminase proteins. Of these, the terminase of *Mycobacterium* phage LizLemon is the most similar, with which it shares 49% amino acid sequence identity. The majority of the terminase proteins similar to Orf3 appear to be large terminase subunits. Furthermore, Orf3 contains a motif pfam03237 commonly seen in terminase enzymes, supporting this hypothesised role. Typically these two genes (small and large terminases) function together as a complex, with the small terminase subunit determining the specificity of DNA packaging (Catalano, 2000), while the large terminase subunit mediates cleavage of the phage DNA packaged into the prohead (Fujisawa & Morita, 1997).

3.4.5. Structural gene module

Genes *orf12* to *orf31* appear to represent the structural module of phage SPI1 that is located typically adjacent to the packaging module in phage genomes. Orf5 shares amino acid sequence identity to the gp76 protein of *Tsukamurella* phage TPA2, while Orf10 contains a motif pfam07098 (of unknown function) and an amino acid sequence similar to several hypothetical proteins, especially to Gp10 of *Mycobacterium* phage Pipefish. The gene *orf12* contains a motif for a Phage Mu F like protein (pfam04233) commonly found in *Caudovirales* phages, and is probably necessary for viral head morphogenesis (Ratcliff *et al.*, 1979), consistent with it being a minor head protein there. Gene *orf12* also shares 39% of its translated amino acid sequence with a hypothetical protein from *Rhodococcus rhodochrous*, while Orf13 shares 42% amino acid sequence similarity with the Gp19 protein from *Tsukamurella* phage TPA2 which appears to be a putative structural protein (Petrovski *et al.*, 2011a).

Orf14 appears to be a chimeric gene, since the N-terminal region of the translated amino acid sequence shares sequence similarity with a hypothetical protein from *R. ruber* (49% identity) and

the C-terminal region is most similar to a hypothetical protein from *Streptomyces seoulensis* (58% identity). The N-terminal region of Orf14 also shows amino acid sequence similarity to a capsid protein from the *Mycobacterium* phage KayaCho (48% identity). Together with its position downstream from other phage structural proteins, and in the absence of any detectable protein motifs, these observations support its designation as a phage structural protein. Orf15 shares amino acid sequence identity with the gp13 protein of *Nocardia cyriacigeorgica* strain GUH-2 (52% identity), as does Orf16 (47% identity). Further examination of both Orf15 and Orf16 fails to show any substantial amino acid sequence homology between these two genes, suggesting that perhaps a nonsense mutation has occurred in what was previously a single gene creating two smaller truncated versions of the translated protein. This suggestion appears reasonable, as the length of gp13 of *N. cyriacigeorgica* is 257 amino acid residues, which is comparable to the combined size of both Orf15 and Orf16 (265 residues).

Orf21 is most similar to a hypothetical protein from *Rhodococcus ruber* (70% identity), although it also shares amino acid sequence homology with the main tail structural protein from *Tsukamurella* phage TPA2 (65% identity) (Petrovski *et al.*, 2011a), and on this basis we propose it may carry out the same or similar function in phage SPI1.

Orf28 was most similar in sequence to a hypothetical protein from *R. rhodochrous* (38% identity), but was also similar to the gp30 protein of *Tsukamurella* phage TPA2 (52% identity), located in its structural module, thus suggesting that *orf28* could also be a structural gene in SPI1 phage. *Orf29* and *orf30* are located upstream of the putative tape measure protein gene (*orf31*), although no putative ribosomal slippage sequences indicative of a conserved translational frameshift mechanism (Xu *et al.*, 2004) could be identified.

Orf29 shares 47% amino acid sequence similarity with a putative tail assembly chaperone protein from *Mycobacterium* phage Jolie1, while Orf30 shares 39% amino acid sequence homology with the Gp32 protein of phage TPA2. Orf31 contains 1801 amino acid residues, and is the largest protein encoded by the phage SPI1 genome, making it a strong candidate for the putative tape measure protein, and thus typical of phages in the *Siphoviridae* family (Pedulla *et al.*, 2003). This suggestion is supported by the presence of a motif (TIGR01760) diagnostic for a tape measure protein.

3.4.6. Host cell lysis module

A phage lysis gene module consists typically of one or more lysis genes and a holin gene. Together these are responsible for host cell lysis and the release of phage progeny at the end of the replication cycle (Daniel *et al.*, 2007). A putative holin gene *orf18*, was located within the putative structural gene module. Its translated amino acid sequence shares 40% similarity to a holin gene identified in *Rhodococcus* phage ReqiDocB7 (Summer *et al.*, 2011). This finding was surprising given that holin genes are located typically adjacent to lysin genes in dsDNA phage genomes (Daniel *et al.*, 2007; Lu *et al.*, 2010) and *orf18* was not in the vicinity of a recognisable lysin gene. Furthermore, this gene contained one putative transmembrane region, thus only partially satisfying the criteria for such a gene outlined by Wang *et al.* (2000) that suggests phage holin proteins should contain two or more transmembrane regions and be less than 150 amino acid residues in size.

A higher level of amino acid sequence homology supported the annotation of genes *orf49* and *orf50* as components of the lysis module of phage SPI1, despite their location within the DNA replication module. Such an unusual placement of lysis genes within the DNA replication module has been seen in the genome of *Gordonia* phage GTE7 (Petrovski *et al.*, 2011b) and *Rhodococcus* phage REQ1 (Petrovski *et al.*, 2013a). Therefore, this arrangement may be a more common feature of this group of phages than presently considered. Orf49 shares 63% amino acid sequence identity with a hypothetical protein from *Rhodococcus* sp. strain 29MFTsu3.1, and also contains a pfam13539 diagnostic of a D-alanyl-D-alanine carboxypeptidase phage lysin protein. Phage holin genes are typically smaller than 150 amino acids, and located adjacent to phage lysin genes, and are usually encoded in the same direction (Wang *et al.*, 2000). *Orf50* satisfies these criteria and also its product possesses two transmembrane regions, supporting its annotation as a phage holin (Wang *et al.*, 2000).

3.4.7. DNA replication/maintenance module

The DNA replication/maintenance module of phage SPI1 appears to be encoded by genes *orf42* to *orf66.* Many of the putative genes there appear novel, but some share homology with other

actinobacterial genes. For example, *orf48* encodes a protein which shares 52% amino acid sequence identity to a hypothetical protein from *Gardnerella vaginalis* and contains a motif for a short chain dehydrogenase (PRK08324).

Orf52 shares 40% translated amino acid sequence identity to a RuvC-like resolvase protein found in *Rhodococcus* phage E3 (Salifu *et al.*, 2013), suggesting that this gene might provide a mechanism for DNA recombination events (Sharples *et al.*, 1999) in phage SPI1. Putative Holliday junction resolvase genes have been found in the genome sequence of several other actinophages including TPA2 (Petrovski *et al.*, 2011a) and GTE2 phages (Petrovski *et al.*, 2011c).

Gene *orf57* appears to encode a helicase as its product shares 52% amino acid sequence identity with the helicase of *Tsukamurella* phage TPA2 and contains the motif (COG1061) for a helicase. Gene *orf59* also appears similar to a gene from TPA2 phage as its translated amino acid sequence shares 40% identity with a primase gene from this phage. Furthermore, its N-terminal region contains a motif (pfam09250) diagnostic for a bifunctional DNA primase/polymerase (Lipps *et al.*, 2004), and it's C-terminal region contains a motif (pfam13481) for a AAA domain found in many presumed DNA repair proteins (Ammelburg *et al.*, 2006). While no statistically significant matches could be identified for *orf60,* based on either its nucleotide and translated amino acid sequence, its product did contain a HTH DNA binding motif (cd00569) suggesting it has a possible regulatory function. Orf61 shares 47% amino acid sequence identity with the gp58 protein of *Mycobacterium* phage Acadian that encodes a DNA polymerase I-3'-5' exonuclease and polymerase. Orf61 also possesses a motif (COG0749) for a DNA polymerase I 3' - 5' exonuclease and polymerase, suggesting it might function in this capacity in phage SPI1.

3.4.8. Evolutionary events that contribute to the SPI1 genome

Analysis of the phage SPI1 genome suggests a number of complex recombination events may have contributed to its evolution. Gene *orf42* appears to encode a putative CRISPR associated RAMP superfamily protein Cfs2 indicated by the presence of a cd09706 motif, which might be present as a result of co-evolution of virus and host, and function in protecting the phage from CRISPR defence, however, this remains to be confirmed experimentally (Sakamoto *et al.*, 2009).

The presence of a Holliday junction resolvase is supportive of the potential for homologous recombination events in its ancestry, as these resolve Holliday junction structures formed during recombination (Sharples *et al.*, 1999). It is also possible that homologous recombination events may have influenced the order of genes in SPI1 as *orf5,* which appears to be a homologue of the Gp76 protein of phage TPA2, is located upstream of *orf30* which itself appears to be a homologue of the Gp32 protein of phage TPA2. Several genes with no similarity to TPA2 genes are interspersed between them (Petrovski *et al.*, 2011a) and similar recombination events have been observed in the genome sequence of phage TPA2, which might suggest that despite the conserved modular arrangement of genes within phage genomes, they can recombine to form new modular arrangements (Petrovski *et al.*, 2011a). A chimeric gene, *orf14*, was seen in the genome of phage SPI1, and as similar mosaic gene structures have been reported in other actinophage genome sequences including TPA2 phage, these may be a common feature of actinophage genomes (Petrovski *et al.*, 2011a). Similarly, in *orf45* a small region of nucleotide sequence homology with *Nocardia brasiliensis* strain ATCC 700358 is observed. It is possible that this gene is either highly conserved for functional reasons, or was perhaps recently acquired by horizontal gene transfer events.

Though the genome of phage SPI1 is highly novel at a nucleotide sequence level, an examination of its putative gene products suggests that 34% of its genes share statistically significant sequence similarities to those from the cluster B phages in the clustering system of (Hatfull *et al.*, 2010). This cluster also includes *Tsukamurella* phage TPA2 (Petrovski *et al.*, 2011a).

3.5. Conclusions

Given that PTLO are implicated as the etiological and/or stabilising agents of activated sludge foams worldwide, and that all currently available isolates fall into a single species *S. piniformis,* the isolation of at least one phage infective for this species may contribute to developing a successful phage therapy cocktail for its control in wastewater treatment plants. We propose that *Skermania* phage SPI1 appears to be such a phage suitable for this application. Furthermore, no excisionase, integrase, or repressor protein genes could be identified in its genome sequence, which is encouraging as phage SPI1 appears to be obligately lytic/highly virulent. However, it does not appear able to lyse all of strains of *S. piniformis,* and so by itself

may not be appropriate for control of all *S. piniformis* foams. Consequently, more *S. piniformis* phages with different host ranges are required before these important practical questions can be answered.

4. Lysis to kill: evaluation of the lytic abilities and genomics of nine bacteriophages infective for Gordonia spp. and their potential use in foam biocontrol

4.1. Abstract

Nine bacteriophages (phages) infective for members of the genus *Gordonia* were isolated from wastewater and other natural water environments using standard enrichment techniques. The majority were broad host range phages targeting more than one *Gordonia* species. When their genomes were sequenced, they all presented as highly novel dsDNA *Siphoviridae* phages, ranging from 17,562 to 103,424 bp in size, and containing between 26 and 126 genes, many of which were novel. Many of these phage genomes diverged from the expected modular genome architecture of other characterised *Siphoviridae* phages, and contained unusual lysis gene arrangements. Whole genome sequencing also revealed that infection with lytic phages does not appear to prevent spontaneous prophage induction in *Gordonia malaquae* lysogen strain BEN700. Novel TEM sample preparation techniques were developed to view both attachment and replication stages of phage infection. Of the nine phages investigated, three appeared suitable for use in phage therapy cocktails for potential use in the biological control of activated sludge foaming caused by hydrophobic *Gordonia* cells.

4.2. Introduction

Many isolates of members of the actinobacterial genus *Gordonia* have been cultured from wastewater treatment plants (de los Reyes III, 2010) where they probably play a key role in degrading some of the more recalcitrant influent substrates (Arenskotter *et al.*, 2004; Drzyzga, 2012). They include *Gordonia amarae,* an organism with a characteristic right-angled branching morphology, and among the first foam forming bacteria isolated and cultured (de los Reyes III, 2010; Klatte *et al.*, 1994; Lechevalier & Lechevalier, 1974). Other *Gordonia* species and members of closely related genera share this distinctive morphology, and so in the absence of more precise identification, those with it are commonly referred to as *Gordonia amarae*-like organisms, or GALO (de los Reyes III, 2010).

Members of the *Corynebacteriales*, which include *Gordonia*, *Nocardia*, *Rhodococcus*, *Tsukamurella* and *Mycobacterium*, are often referred to collectively as the Mycolata because they alone synthesise long chain hydroxylated mycolic acids, organised as an exocellular outer membrane (Goodfellow *et al.*, 2012). Their presence renders these cells highly hydrophobic. In activated sludge, high levels of these Mycolata stabilise foams formed on the surface of aeration tanks and clarifiers (de los Reyes III, 2010). Formation of these stable foams is a global problem that impacts negatively on plant aesthetics, increases maintenance costs, and complicates sludge management (Soddell & Seviour, 1990). Some of the Mycolata in these foams are opportunistic pathogens, thus posing a potential health hazard to plant operators from their aerosol dispersal (de los Reyes III, 2010; Soddell, 1999; Soddell & Seviour, 1990).

Formation of these stable foams requires air bubbles, surface active agents, and hydrophobic particles, in this case the Mycolata cells (Petrovski *et al.*, 2011d). A successful control strategy must be directed at the hydrophobic bacteria because neither air bubbles, nor detergents can be eliminated from the activated sludge process (Petrovski *et al.*, 2011d). Current foam control strategies are not effective universally, and no single method reliably controls all foams. This probably is a reflection of how little is known about the microbial ecology of these causative bacteria (de los Reyes III, 2010). It was Thomas *et al.* (2002) who proposed that phage therapy that exploits the natural lytic cycles of Mycolata phages would be an attractive and environmentally friendly approach to selectively control their population levels without affecting other desirable bacteria in these systems.

Currently (February 2015), 228 phages targeting members of the genus *Mycobacterium* have had their genomes sequenced (see section 5), but only four lytic *Gordonia* phage genomes sequences are available. These are for phages GTE2 (Petrovski *et al.*, 2011c), GTE7 (Petrovski *et al.*, 2011b), GRU1, and GTE5 (Petrovski *et al.*, 2012b). Also induced and characterised were two temperate *Gordonia* phages GAL1, and GMA1 from *Gordonia alkanivorans*, and *Gordonia terrae*, respectively (see section 5). All *Gordonia* phages isolated so far have highly novel genomes (Petrovski *et al.*, 2011b, c; Petrovski *et al.*, 2012b). Yet with such a small sample size, it is not sensible to comment on the general characteristics of *Gordonia* phages and draw conclusions from these as to their suitability or otherwise for foam biocontrol. Therefore, more *Gordonia* lytic phages are needed, including those from habitats other than activated sludge plants.

This study set out to increase the small existing library of *Gordonia* phages, and to characterise them in terms of their host ranges, morphologies, and genomics. Nine phages infective for members of this genus were isolated and their suitability for use in *Gordonia* foam biocontrol was investigated.

4.3. Materials and methods

4.3.1. Isolation and preliminary characterisation of Gordonia phages

Host strains (Table 8.1) held in the La Trobe University culture collection, which were used in this study and the techniques for their growth are those described by Petrovski *et al.* (2011a). All phages were isolated and subsequently purified from water samples collected from a variety of locations using enrichment pools of multiple host strains, as described in Table 4.1 and by Petrovski *et al.* (2011a). Phage host range specificity determinations were also carried out using a ten-fold dilution series plated onto bacterial lawns as described by Petrovski *et al.* (2011a).

4.3.2. Transmission electron microscopy of virion morphology

Grids for visualisation of virions were prepared with the negative stain uranyl acetate, as described by Petrovski *et al.* (2011a). Both carbon and formvar coated grids were used (Electron Microscopy Sciences, Australia), with the exception of phage GTE6 which was examined on grids coated with formvar only. Prepared grids were subsequently examined with a JEOL JEM-100CX, JEOL JEM-2010HC, or a Tenaci Fei T30 Transmission Electron Microscope, as detailed in Table 4.2.

4.3.3. Transmission electron microscopy to show phage attachment

A single colony of *Gordonia malaquae* (CON67) from a streak plate incubated at 30°C for 3 days was carefully removed and the cells added to 20 µL of high titre GMA6 phage lysate (> 10^{10} PFU/mL). This mixture of phage and bacteria was left to stand for 10 min to allow phage infection before they were adsorbed onto the surface of a carbon/formvar coated 200 mesh copper grid (Electron Microscopy Sciences, Australia). Grids were then washed twice in sterile double-distilled water (ddH_2O), and then negatively stained with 2% (w/v) uranyl acetate for 2 min. Excess liquid was absorbed onto filter paper and the grid was allowed to air dry. These were then examined under a JEOL JEM-2010HC Electron Microscope.

4.3.4. Transmission electron microscopy of phage virion assembly

A 1 mL aliquot of *Gordonia terrae* (CON34^T) incubated at 30°C for 3 days was added to 20 mL of PYCa broth and 100 µL of high titre phage GTE6 suspension (> 10^{10}). This mixture was allowed to stand for 10 mins to allow phage infection before further incubation at 30°C for 3 days. A 1.5 mL aliquot was then centrifuged (3,000 x g for 30 min) and the supernatant discarded. Pelleted cells were re-suspended and fixed in 2.5% (v/v) glutaraldehyde in 0.1 M phosphate buffer (pH 6.8-7.3), and incubated at 4°C overnight, then harvested (14,000 x g for 5 min) and washed in the same phosphate buffer 3 times, with 10 min between washes. Cells were then post-fixed in 1% osmium tetroxide in 0.1 M phosphate buffer for 90 min and washed three times in sterile ddH_2O. They were then dehydrated through an acetone series of increasing concentrations (30%, 50%, 70%, 90% and 100%) for 10 min each, prior to a final washing with 100% acetone with a molecular sieve (ProSciTech, Australia) (10 min). Dried specimens were then infiltrated with Spurr epoxy resin (ProSciTech, Australia), initially with 50% resin, 50% dehydrated acetone, and incubated overnight at room temperature. The following day, the mixture was replaced by 100% Spurr epoxy resin with a further incubation of 1-2 h, before finally being replaced by fresh Spurr expoxy resin, and polymerised at 65°C overnight. Thin sections (100 nm) were cut with a glass knife on an LKB Microtome and post-stained with uranyl acetate and lead citrate. Sections were placed on 200 mesh copper grids and examined under JEOL-JEM 2010HC electron microscope.

4.3.5. Mass spectroscopy

To identify phage structural proteins, purified phage particles >10^{13} PFU/mL were precipitated using $(NH_4)_2SO_4$ following precipitation with $ZnCl_2$ to remove any residual polyethylene glycol from the previous phage precipitation step. Pellets were re-suspended in 8 M urea to a final volume of 100 µL prior to transfer to the Mass Spectroscopy and Proteomics facility at the La Trobe University Institute of Molecular Sciences. Here peptides reconstituted in 0.1% formic acid and 2% acetonitrile (buffer A) were loaded onto a trap column (C18 PepMap 300 µm i.d. × 2 cm trapping column, Thermo-Fisher Scientific) at 5 µL/min for 6 min and washed for 6 min before switching the precolumn in line with the analytical column (Vydac MS C18, 75 µm i.d. × 25 cm, Grace Davison). The separation of peptides was performed at 300 nL/min using a linear acetonitrile (ACN) gradient of buffer A and buffer B (0.1% formic acid, 80% ACN), starting from 5% buffer B to 40% over 60 min. Data were collected on an hybrid quadrupole/time-of-flight MS (MicroTOF-Q, Bruker, Germany) with a nano-electrospray ion source using Data Dependent Acquisition mode and *m*/*z 15*0–2500 as MS scan range. Nitrogen was used as the collision gas. The ionisation tip voltage and interface temperature were set at 4200 V and 205°C respectively. Collision Induced Dissociation (CID) MS/MS spectra were collected for the 4 most intense ions. Dynamic exclusion parameters were set as follows: repeat count 2, duration 60 s. The data were collected and analysed using Data Analysis Software (Bruker Daltonics, Bremen, Germany).

4.3.6. Genome sequencing of Gordonia phages

Genomic DNA was extracted from phages GTE6, GMA2, and GMA6 and sequenced using a Roche GS FLX genome sequencer and titanium chemistry, as described in Petrovski *et al.* (2011a). Genomic DNA extracted from all other phages in the same manner was prepared with an Illumina Nextera XT sample preparation kit as per manufacturers' instructions. The prepared DNA libraries were sequenced on an Illumina MiSeq as a 150 bp paired end run.

4.3.7. Genome annotations

The genome open reading frames (ORFs) were screened initially using Glimmer (v3.02), where ORFs with a minimum size of 90 bp were detected (Delcher *et al.*, 2007). All predicted start codons were inspected for the presence of putative ribosomal binding sites and corrected as necessary. Sequence similarity searches were carried out against the GenBank database, as described by Petrovski *et al.* (2011c). The presence of tRNA and tmRNA were determined using both ARAGORN (Laslett & Canback, 2004), and tRNAScan-SE (Schattner *et al.*, 2005). Transmembrane domains were predicted with the DAS Transmembrane Prediction server (Cserzo *et al.*, 1997).

Phage DNA when analysed by gel electrophoresis gave results consistent with circularly permuted DNA genomes. Therefore, genome annotations were conducted starting with the DNA packaging operon.

4.3.8. Nucleotide sequence accession number

The nucleotide sequence for all phages have been deposited GenBank under the following accession numbers; GTE6 (KR053200), GTE8 (KR053201), GMA2 (KR063281), GMA3 (KR063279), GMA4 (KR053199), GMA5 (KR053198), GMA6 (KR063280), GMA7 (KR063278), and GRU3 (KR053197).

4.4. Results and discussion

4.4.1. Isolation and host range characterisation

All *Gordonia* phages isolated to date were obtained from wastewater, with most coming from activated sludge plants on the east coast of Australia (Petrovski *et al.*, 2011b, c; Petrovski *et al.*,

2012b; Thomas *et al.*, 2002). While most phage isolates described in this study were also from wastewater (Table 4.1), an additional two phages GMA4 and GTE8, were obtained from puddle sediment (Reservoir, Australia), and creek water (Bendigo, Australia), respectively. This might suggest that for obtaining phages similar to these, enrichments from more diverse environments may be rewarding.

Only one of these, phage GMA4, was monovalent being infective for only *Gordonia malaquae* strain BEN700. Phage GMA3 had a similar host range, as it could also infect the same *G. malaquae* strain (BEN700), but, also lysed *G. terrae* (G238). All other phages examined lysed multiple species of *Gordonia*, with phage GMA7 lysing eleven strains from four different *Gordonia* species i.e. *G. terrae* (CON34^T, GOR9, G238), *G. rubropertincta* (CON38^T), *G. malaquae* (CON59, CON60, A554^T, A448, BEN700), and *G. hydrophobica* (CON65^T, CON66). As well as phage GTE8 lysing three strains of *Gordonia terrae* (CON34^T, GOR9, G232) and one of *G. rubropertincta* (CON38^T), it also lysed *Nocardia asteroides* (CON12). Phages able to lyse members of both of these genera have been reported before. They include phage GRU1, which targets *Nocardia nova* strain CON47^T and *Gordonia terrae* strains CON34^T and G232, and also *Gordonia rubropertincta* strain CON38^T (Petrovski *et al.*, 2012b). This outcome might reflect the close phylogenetic relationship of these host bacteria.

Many of these phages overlap in their host range. For example, phages GTE6, GMA2, GMA6, and GMA7 all lyse the same four strains of *G. malaquae* (CON59, CON60, A554^T, and A448), a property which might make them useful additions to any phage cocktail designed to target foaming caused by *G. malaquae,* especially if they target different host receptor sites.

Interestingly, most phages exhibited some level of host strain specificity, as with phage GMA5, which lysed only two strains (G238, G232) of the eight *G. terrae* strains tested. A similar situation has been seen for other *Gordonia* phages including GTE2 phage (Petrovski *et al.*, 2011c) which lysed only one of five *G. terrae* strains. No phages infective for *G. amarae* were obtained in this study, an outcome that will be discussed later.

Table 4.1 Isolation and preliminary characterisation of nine *Gordonia* phages

Phage	Sample	Strain	Lab ID	Enrichment pool members	Host range
GMA2	Activated sludge, Kyneton, Victoria, Australia	*G. malaquae*	A448	See GTE8	*G. terrae* (CON34^T, GOR9, G238), *G. malaquae* (CON59, CON60, A554^T, A448), *G. hydrophobica* (CON65^T, CON66)
GMA3	Wastewater, Glenelg, South Australia, Australia	*G. malaquae*	BEN700	See GTE8	*G. terrae* (G238), *G. malaquae* (BEN700)
GMA4	Puddle water and sediment, Reservoir, Victoria, Australia	*G. malaquae*	BEN700	See GTE8	*G. malaquae* (BEN700)
GMA5	Activated sludge, Carrum (Eastern Treatment Plant), Victoria, Australia	*G. malaquae*	BEN700	See GTE8	*G. rubropertincta* (CON38^T), *G. terrae* (G238, G232), *G. malaquae* (BEN700)
GMA6	Activated sludge, Bendigo, Victoria, Australia	*G. malaquae*	CON67	See GTE8	*G. malaquae* (CON59, CON60, CON67, A554^T, A448, BEN700), *G. terrae* (G238)
GMA7	Activated sludge, Werribee, Victoria, Australia	*G. malaquae*	CON60	*G. terrae* (GOR9, G232, G238), *G. malaquae* (A554^T, A448, CON60, BEN700), *T. paurometabola* (CON61)	*G. terrae* (CON34^T, GOR9, G238), *G. rubropertincta* (CON38^T), *G. malaquae* (CON59, CON60, A554^T, A448, BEN700) *G. hydrophobica* (CON65^T, CON66)
GRU3	Wastewater, Inverell, Queensland, Australia	*G. rubropertincta*	CON38^T	See GTE8	*G. rubropertincta* (CON38^T), *G. terrae* (GOR9, G232)

Phage	Sample	Strain	Lab ID	Enrichment pool members	Host range
GTE6	Activated sludge, Nambour, Queensland, Australia	*G. terrae*	CON34^{T}	*G. terrae* (CON34^{T}, BEN601, BEN604), *G. sputi* (CON48, CON49^{T}), *G. amarae* (CON44^{T}, CON9)	*G. terrae* (CON34^{T}, GOR9), *G. malaquae* (CON59, CON60, A554^{T}, A448), *G. hydrophobica* (CON65^{T}, CON66)
GTE8	Bendigo creek water, Bendigo, Victoria, Australia	*G. terrae*	G232	*G. terrae* (CON34^{T}, G238, G290, G255, G232, GOR9), *G. sputi* (CON48, CON49^{T}), *G. amarae* (CON44^{T}, CON9), *G. hydrophobica* (CON65^{T}, CON66), *G. desulfuricans* (CON69^{T}), *G. polyisoprenivorans* (CON71), *G. alkanivorans* (CON72), *G. malaquae* (A554^{T}, A448, BEN700, CON67), *T. inchonensis* (BEN701), *R. erythropolis* (BEN703) *G. aichiensis* (CON22^{T})	*N. asteroides* (CON12), *G. terrae* (CON34^{T}, GOR9, G232), *G. rubropertincta* (CON38^{T})

^{T} *indicates type strain*

4.4.2. Gordonia phage morphology

All phages examined here by TEM had both the isometric type B1 capsids (~37 to ~63 nm in diameter) and long, non-contractile tails (~85 to ~474 nm long) characteristic of members of the viral family *Siphoviridae.* Further details are provided in Figure 4.1, and Table 4.2. Phage GMA3 was not examined by TEM, but based on its genome sequence containing a gene encoding a long tape measure protein and its dsDNA genome, it too is likely to be a member of the *Siphoviridae* (Pedulla *et al.*, 2003). With TEM, the morphology of phage GMA6 was not as expected of a *Siphoviridae* member, since its tail appeared to be uncharacteristically thick and rigid (Figure 4.1G). So to resolve this concern, phage GMA6 virions were exposed briefly to *G. malaquae* strain CON67 (it's host) and then examined by TEM. Images showed clearly (Figure 4.2) that its phage tail does not appear to contract during attachment to host confirming it to be a member of the *Siphoviridae.* Furthermore, TEM shows that host cell attachment can involve simultaneously many phages. Whether superinfections with more than one phage genome successfully invading the host cell of *G. malaquae* can occur, was not explored. We could also visualise the presence of post replication mature phage progeny within the host cell, by first infecting *G. terrae* strain CON34^{T} with phage GTE6 in PYCa broth medium, before the embedding, sectioning, and staining stages for TEM sample preparation. Uninfected controls were carried out which did not show similar bodies inside the cells, and capsids were consistent in size with that of phage GTE6 (62 nm). Thus, Figure 4.3 shows mature GTE6 virions inside the host cells, prior to cell lysis and release of progeny phage.

Table 4.2 *Gordonia* phage virion measurements

Phage name	Capsid diameter (nm)	Tail length (nm)
GMA2[b]	~61	~386
GMA4[b]	~54	~244
GMA5[b]	~37	~85
GMA6[b]	~62	~143
GMA7[c]	~63	~474
GRU3[b]	~43	~93
GTE6[a]	~48	~152
GTE8[b]	~56	~239

[a] electron micrographs obtained using a JEOL JEM-100CX; [b] electron micrographs obtained using a Tenaci Fei T30; [c] electron micrographs obtained using a JEOL JEM-2010HC.

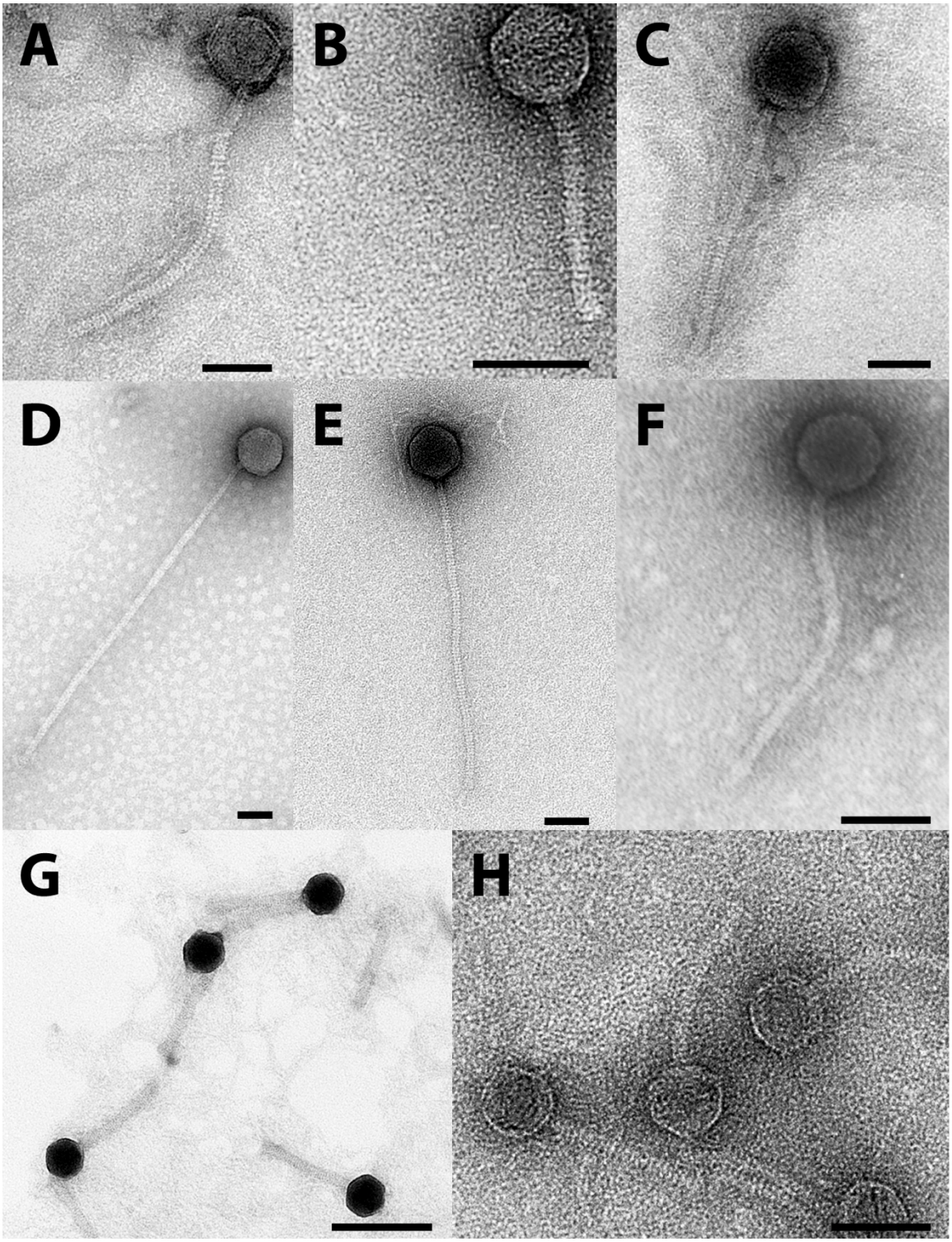

Figure 4.1 *Gordonia* phage morphologies.

(A) phage GMA4; (B) phage GRU3; (C) phage GTE8; (D) phage GMA7; (E) phage GMA2; (F) phage GTE6; (G) phage GMA6; (H) phage GMA5. Negatively stained electron microscopy, 2% uranyl acetate. Scale = 50 nm.

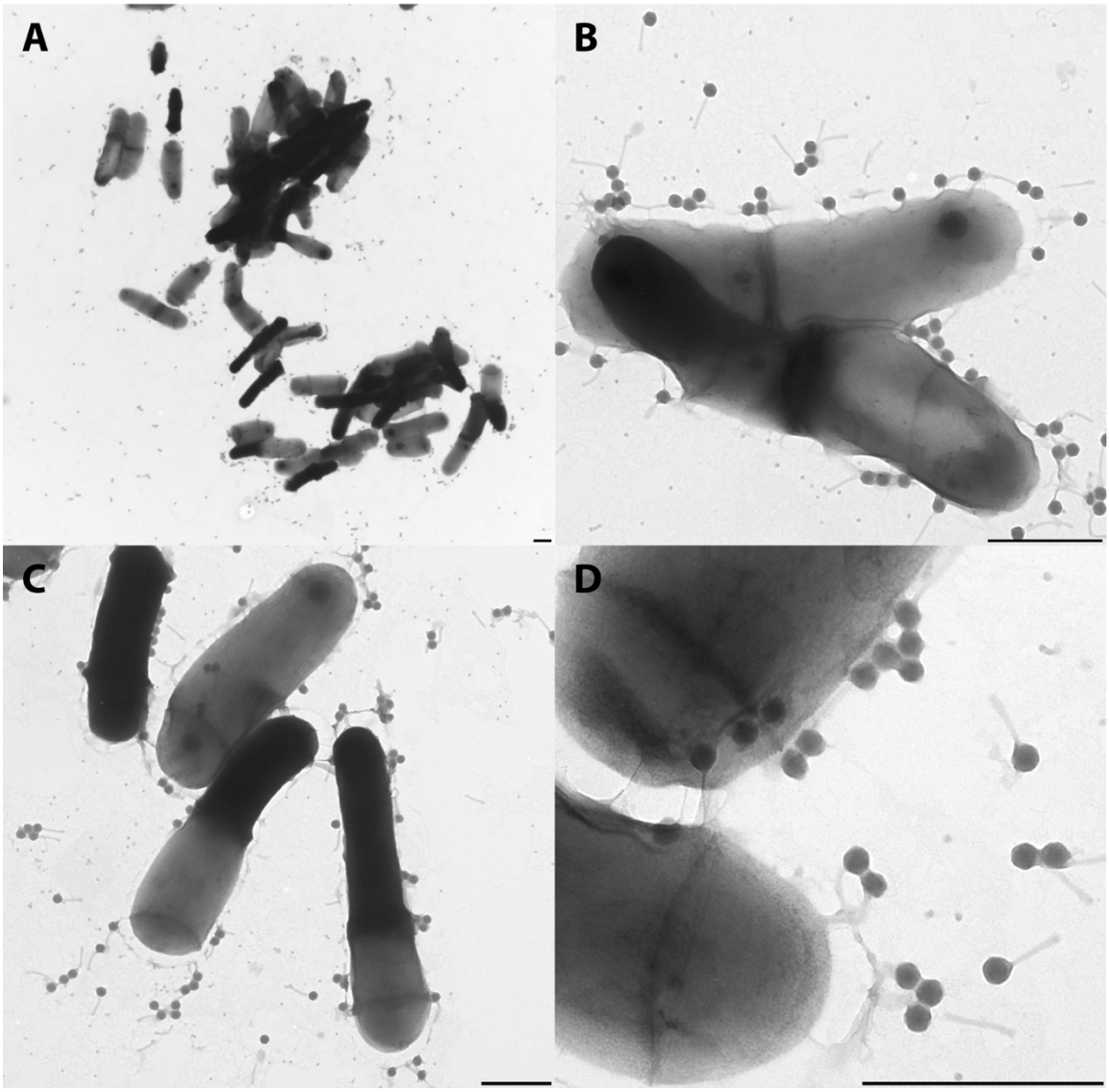

Figure 4.2 Attachment stage of the phage infection cycle between phage GMA6 and host bacterium *Gordonia malaquae* strain CON67.

(A-D) Attachment of virions to host cells. Negatively stained electron microscopy, 2% uranyl acetate, 20 minutes post infection. Scale = 500 nm.

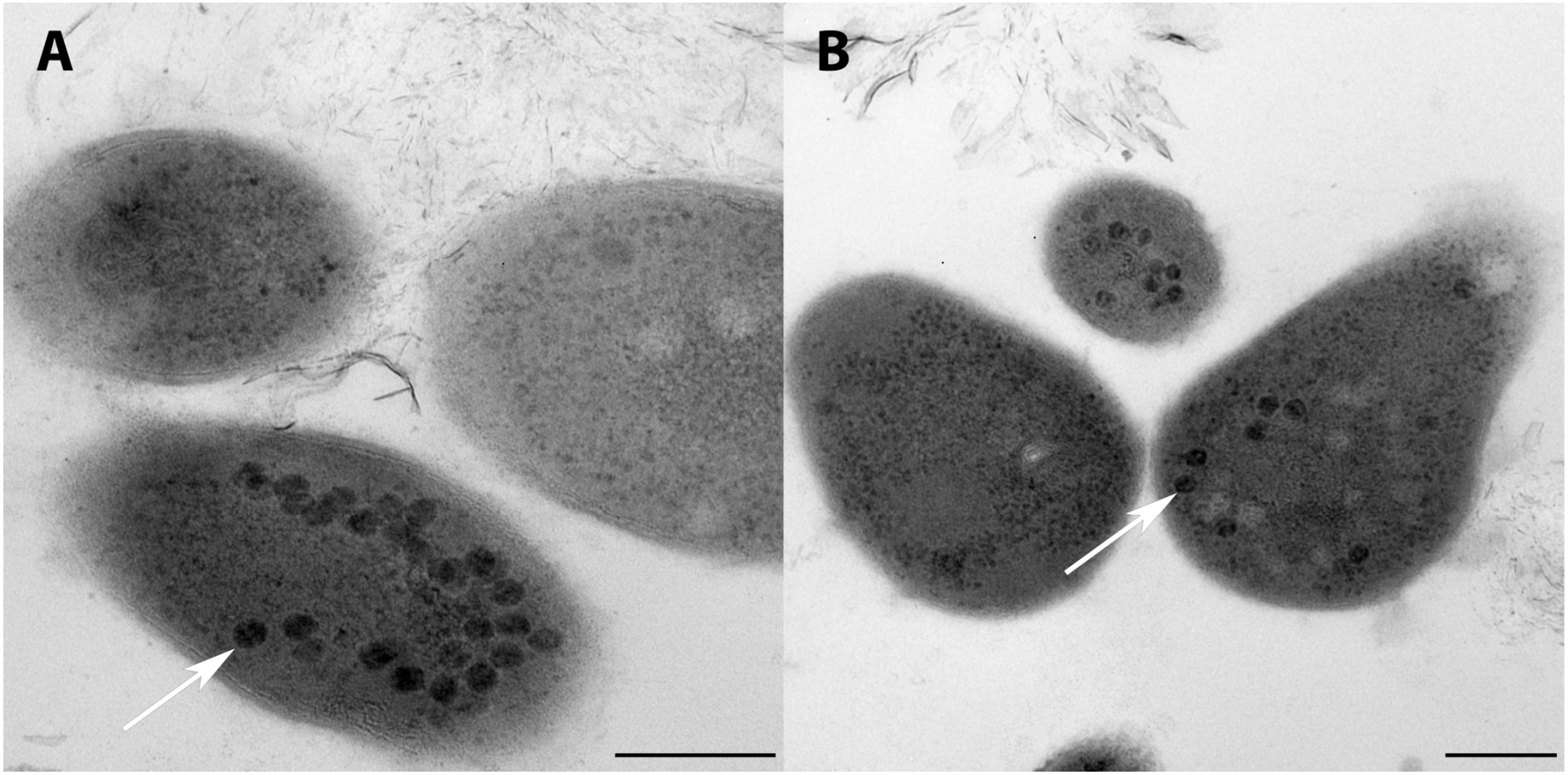

Figure 4.3 Replication of phage GTE6 inside *G. terrae* strain CON34^{T} cells prior to cell lysis.

(A-B) infected and non-infected cells. Arrows indicate assembled phage capsids inside bacterial cells. Negatively stained electron microscopy, 2% uranyl acetate, 3 days post initial inoculation of bacteria with phage. Scale = 200 nm.

4.4.3. Genome structure and organisation of Gordonia phages

Genomes of the assembled *Gordonia* phages ranged from 17,562 to 103,424 bp in size, and they contained between 27 to 127 putative *orf*s (**Table 4.3**), arranged mostly in the modular architecture commonly seen in the *Siphoviridae* phages (**Figure 4.4**). All contained putative genes orientated in both forward and reverse orientations, with the one exception being phage GTE6, where all its genes were in a forward orientation. Only between 22 and 50% of the putative genes identified in the nine *Gordonia* phage genomes could be annotated functionally (**Table 4.3**, and **Table 4.5** in section 4.6). The G+C mol % contents of all phages ranged from 51.3 to 67.8 mol % (**Table 4.3**) and for the majority this value was close to that of the corresponding host cells (Goodfellow *et al.*, 2012).

Most of these phage genomes contained no putative tRNA, and no tmRNA could be identified in any of them (**Table 4.3**). Of those phages where tRNA were seen, phage GMA7 contained 1 putative tRNA-Asn, GMA4 contained 1 putative tRNA-Try, and phage GMA2 contained a cluster of 16 putative tRNA from (nt 57,883 - 60,154) (**Table 4.6** in section 4.6). Such tRNA clusters have been observed previously in *Mycobacterium* phages where they appear to be important in late lytic growth, where they may compensate for degradation and inadequacy of host tRNA (Pope *et al.*, 2014).

All the phages characterised here had genomes having little or no sequence similarity to any other phage sequence in the GenBank database. However, there were two exceptions. The GMA7 phage genome sequence was highly similar to that of GTE7 phage (97% identity) (Petrovski *et al.*, 2011b), while phage GTE8 was highly similar to both GTE5 (81% identity) and GRU1 phages (83% identity) (Petrovski *et al.*, 2012b). Of the remainder, phages GMA4 and GMA5 were genetically similar to each other (77% nucleotide sequence identity).

Table 4.3 Summary of nine *Gordonia* phage genomes

Phage name	Average coverage (fold)	Total read count	Genome size (bp)	G+C content (mol %)	No. putative tRNA	No. putative genes	No. putative genes in forwards orientation	No. functionally annotated putative genes	No. novel genes	No. palindromes	No. direct repeats	No. inverted repeats
GMA2 [a c]	1,212	336,750	103,424	53.4	16	126	42	42	62	7	22	10
GMA3 [b e]	1,200	677,981	77,779	51.3	0	104	32	27	47	16	18	8
GMA4 [b e]	1,981	716,641	45,537	66.4	1	68	61	22	11	6	40	31
GMA5 [b f]	6,793	930,480	17,562	66.4	0	28	24	14	4	11	28	13
GMA6 [a c]	247	55,269	83,324	58.2	0	115	109	38	68	1	20	3
GMA7 [b c]	1,603	947,843	73,419	56.6	1	101	32	23	5	18	14	5
GRU3 [b e]	520	89,131	17,727	66.5	0	26	23	12	6	3	42	16
GTE6 [a c]	915	141,321	56,982	67.8	0	86	86	23	49	3	252	87
GTE8 [b c]	1,605	777,336	67,617	66.0	0	94	67	23	22	5	48	36

[a] sequenced using 454; [b] sequenced using Illumina; [c] reads assembled using CLC workbench (v6.5.1); [d] reads assembled using CLC workbench (v7.5.1); [e] reads assembled using Spades (v3.1.0); [f] reads assembled using ABySS (v1.3.7), [g] putative genes excluding tRNA, [h] genes with no statistically significant (10-4) E values against any known sequence in the GenBank database.

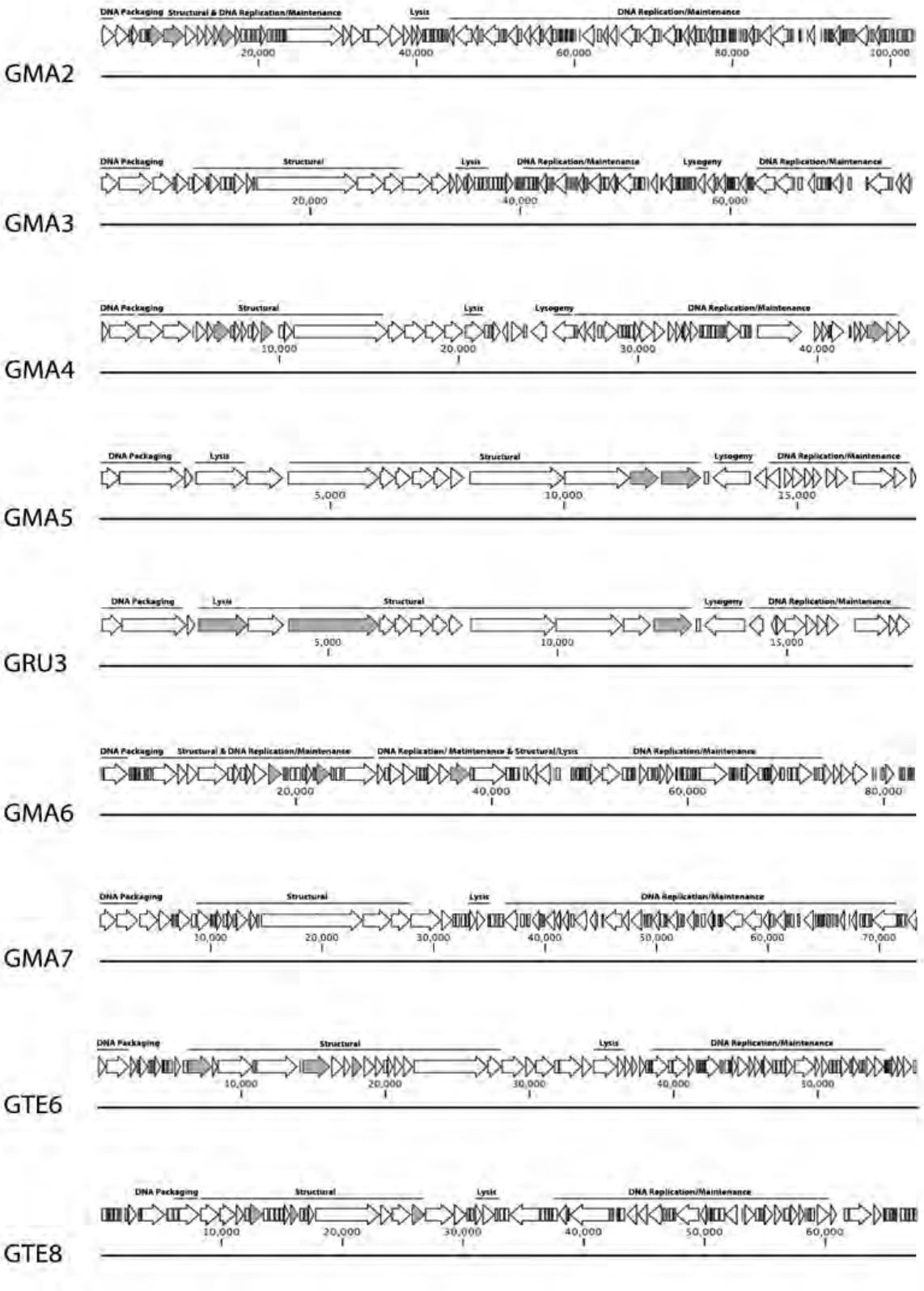

Figure 4.4 Genome map of nine *Gordonia* phages

Genes shaded grey indicate virion structural genes identified with mass spectroscopy data.

4.4.4. Evidence for spontaneous prophage induction events in genome assemblies

High coverage sequencing (~1200-fold to ~6793-fold) of phages GMA3, GMA4, and GMA5 all revealed the presence of a second contig of ~41 kb in size of average lower coverage (between ~17-fold to ~227-fold), as shown in Table 4.4. This was thought initially to represent contamination of the DNA samples used for sequencing. However, closer examination showed that this ~40 kb contig, although incomplete, was identical in its sequence to the Mycolata prophage GMA1, shown to be a spontaneously inducing temperate phage in *G. malaquae* (BEN700) (see section 5). All three phages (GMA4, GMA5, and GMA3) containing this additional contig had all been cultured on this strain.

This observation adds to the evidence (see section 5) that this lysogenic *G. malaquae* strain could tolerate co-infection with these three phages, as well as the GTE2 phage (Petrovski *et al.*, 2011c). Whether these phages interact while co-infecting is unknown. However, it seems that replication of these lytic phages in this *G. malaquae* BEN700 strain does not prevent the spontaneous induction of temperate phage GMA1.

Table 4.4 Coverage of phage GMA1 in the assemblies of phages GMA3, GMA4, and GMA5

Phage sequenced	Length of GMA1 contig	Average coverage of GMA1	Total GMA1 reads
GMA3	41,106	~17-fold	5,097
GMA4	40,897	~227-fold	70,089
GMA5	41,106	~32-fold	10,135

4.4.5. Sequence repeats in Gordonia phage genomes

Repeat structures have been previously reported in genome sequences of several Mycolata phages (Petrovski *et al.*, 2011a; Petrovski *et al.*, 2012b). All nine phage genomes examined here contain between 1 to 18 palindromic sequences of between 14 to 98 bp in length (Table 4.7 in section 4.6). Some of these are located in what appear to be intergenic areas, which might support their roles as putative *rho*-independent transcriptional terminators (Lesnik *et al.*, 2001). They also contained 14 to 252 direct repeats ranging in length between 14 and 425 bp (Table 8.4 in section 8.2). Also seen in these genomes were 3 to 87 inverted repeats that were 56 to 14 bp long (Table 8.4 in section 8.2). Sequence repeats may indicate repeated protein structures, promoters, replication origins and transposable elements (Mott & Berger, 2007), but neither of these could be identified in any of these phages, and so their roles here remain unknown.

4.4.6. Gordonia phage DNA packaging modules

In *Siphoviridae* phage genomes, the large terminase subunit protein usually functions in a complex with a small terminase subunit protein, and together these act to mediate cleavage of the phage DNA at specific sites prior to packaging into the prohead (Catalano, 2000; Fujisawa & Morita, 1997). The gene encoding the large terminase subunit was identified in all nine *Gordonia* phages examined here by either amino acid sequence homology to other known terminase genes (GMA4, GMA3, GMA5, GMA7, and GTE8), and/or in some cases the presence of diagnostic motifs. These were pfam03237 for GMA2, GRU3, GTE6 and GMA6 phages, pfam03354 for GMA5 and GTE8 phages, and cd01335 for GMA3 phage, as detailed Table 4.5 (section 4.6). In *Siphoviridae* phages the small terminase gene is typically located upstream, and is transcribed in the same direction as the large terminase (Catalano, 2000; Fujisawa & Morita, 1997). In all phages except GMA2 this pattern could be recognised, and in some cases supported by amino acid sequence homology to other known small terminases as described in Table 4.5 (section 4.6).

4.4.7. Gordonia phage structural protein genes and virion proteomics

Phage structural protein genes typically are located adjacent to the DNA packaging module, usually beginning with head morphogenesis genes, followed by tail morphogenesis genes (Pedulla *et al.*, 2003). Some deviations from this usual gene arrangement were seen in these nine *Gordonia* phages. For example, in phage GMA6, of seven genes identified between the terminase genes (*orf2* and *orf3*) and the putative portal protein gene (*orf11*), only one could be assigned a putative function as encoding a nucleoside triphosphate pyrophosphohydrolase (*orf8*). This allocation was based on the presence of a cd11541 motif. The location of this gene seemingly involved in DNA maintenance was between the structural and packaging modules, a place markedly different to the typical modulated genome architecture of *Siphoviridae* phages, where genes of similar function typically are clustered together (Hatfull, 2008). Furthermore, *orf14*, within the structural gene module of phage GMA6, appears to encode a HNH endonuclease, a decision based on its amino acid sequence homology to the diagnostic pfam01844 motif. Gene arrangements in phage GMA2 suggest that the structural module of contains a gene encoding a putative a DNA methyltransferase (*orf21*).

In all nine phages the tape measure proteins were identified as the longest gene in their genomes, which is the usual pattern in *Siphoviridae* phages (Pedulla *et al.*, 2003), as shown in Table 4.5 (section 4.6) and Figure 4.4. The only exception to this was that of phage GRU3 where Orf6 encoding a putative phage head protein was slightly larger in size (657 aa) than that of its tape measure protein (Orf12), which was 622 aa in size. In most of the *Gordonia* phages (GMA2, GMA3, GMA4, GMA6, GMA7, and GTE8), the two genes preceding this were identified as encoding putative tail assembly proteins, the latter of which appeared to be translated using a conserved programmed frameshift, a common feature of *Siphoviridae* phages (Xu *et al.*, 2004). Usually, the gene immediately upstream of these is that encoding the putative major tail protein, as is the case here. The same gene arrangement has also been observed in Mycolata phages RER2, RGL3, GTE2, GTE5, GTE7, and GRU1 (Petrovski *et al.*, 2011b, c; Petrovski *et al.*, 2012b; Petrovski *et al.*, 2013b). In all Mycolata phages, tail length and tape measure protein are proportional, i.e. a longer tail is observed with a longer tape measure protein.

Mass spectroscopy data as detailed in Table 4.8 (section 4.6), suggested several phage structural genes are located outside the phage structural gene module as shown in Figure 4.4. For

example, in phage GMA4 a structural protein gene (*orf66*) was seen in the DNA replication module, with a translated protein sequence homologous to a hypothetical protein from *Aeromicrobium marinum,* but also encoding a motif for a phage related tail fibre protein (COG5301)**.** Similarly, in phage GMA6 a structural protein gene (*orf43*) was located within the lysis and DNA replication gene modules. These unusual arrangements of structural genes have been observed in the *Tsukamurella* phage TIN4, where one structural gene was identified in the DNA replication module (see section 2). It may be that gene arrangements like these are a common feature of Mycolata phage genomes.

4.4.8. Gordonia phage lysis modules are diverse

Lysin genes could be identified in all nine *Gordonia* phages, but their locations and numbers varied, and as with many already discussed, they often appeared to disrupt the usual and expected modular genome architecture of *Siphoviridae* phages. Phages GMA5 and GRU3 both contain a D-alanyl-D-alanine carboxypeptidase encoding gene (*orf5* in both phages) showing amino acid sequence homology to a hypothetical protein in *Gordonia soli,* that was located within what appears to be the phage structural module, as shown in Table 4.5 (section 4.6) and Figure 4.4. A similar phage lysin motif and position within the structural gene module has been observed previously in *Rhodococcus* phage RRH1, and the amino acid sequences of these two translated genes are highly similar (Petrovski *et al.*, 2012c).

Phage genomes of GMA7 and GTE6 both appeared to contain their lysin genes adjacent to their phage structural proteins (*orf28* to *orf29,* and *orf38*, respectively), and unusually, contained additional lysin genes in their DNA replication gene modules (*orf41* and *orf58*, respectively). The same pattern was reported for phage GTE7 (Petrovski *et al.*, 2011b), to which phage GMA7 is genetically similar at a nucleotide level (97% identity, 95% coverage).

Phages GMA2, GMA3, and GMA6 also had unusual lysin gene arrangements, with higher numbers of putative lysin genes than the usual lysin A and B arrangement seen elsewhere (Payne *et al.*, 2009). Phage GMA2 was highly unusual as it possessed four putative lysin genes (*orf35 to orf38*), identified by either their amino acid homologies or presence of diagnostic

motifs including the pfam13529 (peptidase), pfam01510 (N-acetylmuramoyl-L-alanine amidase), and cd02619 (peptidase) motifs in Orf35, Orf36, and Orf38, respectively. Phage GMA6 also had four lysin genes (*orf34, orf37, orf40,* and *orf45*), many of which were separated by genes associated with DNA replication/maintenance and virion morphogenesis. A similar pattern was seen in phage GMA3, which contained three putative lysin genes (*orf22, orf24,* and *orf26*) separated by a putative nuclease gene (*orf25*), again associated with DNA replication/maintenance (Table 4.5 in section 4.6, and Figure 4.4).

Orf45 in GMA6 phage is a more complex lysin gene than any of those seen in all other Mycolata phages including those from earlier studies. It alone encoded an unusually high number of different lysin motifs. These include an N-terminal BacA motif of a bacterial lysin from *Enterococcus faecalis* (cd06418), an N-acetylmuramoyl-L-alanine amidase motif (pfam01510) downstream of this, a peptidase motif (pfam01551) further downstream, and an additional C-terminal motif (pfam13810) of unknown function (Table 4.5 in section 4.6).

Holins could not be identified in phages GMA4, GMA5, and GRU3 by nucleotide or amino acid sequence homologies, nor by the criteria of Wang *et al.* (2000), which state that their expressed products should be less than 150 amino acid residues and contain two or more transmembrane regions. If holins are present in these two phages it would seem that they are novel in their gene locations and/or amino acid sequences.

4.4.9. DNA replication/maintenance genes

DNA replication modules in all other Mycolata phage genomes sequenced so far are arranged in a modular genome architecture, where genes functioning in DNA replication/maintenance are located adjacent to phage lysin genes (Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012b; Petrovski *et al.*, 2013a, b, c, 2014). In GMA4, GMA7, and GMA2 phages this region contains putative DNA-methylase encoding genes. Of these, GMA2 appears to possess at least two (*orf21* and *orf51*). If functional, they may play a role in protecting their DNA from host cell restriction attack (Labrie *et al.*, 2010). Such enzymes have been identified in other *Gordonia* phages including GAL1 (section 5), GMA1 (section 5), and GTE7 (Petrovski *et al.*, 2011b) phages.

Metagenomic studies by Tamaki *et al.* (2012) have suggested that methylase genes are more prevalent in phages within the activated sludge environment, from where most of the Mycolata phages have derived. Glycosyltransferase encoding genes are also seen in many phage genomes (Markine-Goriaynoff *et al.*, 2004) including that of GMA2 (*orf4* and *orf22*), and appear to have similar functions to phage methylases where they help protect phage DNA from digestion with restriction endonucleases from host RM systems (Labrie *et al.*, 2010). However, these genes can also be involved in other functions including serotype conversion in temperate phages (Markine-Goriaynoff *et al.*, 2004), and so their purpose here remains to be fully elucidated.

4.4.10. Lysogeny and lysogenic conversion genes

Genomes of phages GMA3, GMA4, GMA5, and GRU3 all contain putative genes that are homologues of phage integrase genes (*orf76*, *orf29*, *orf17*, and *orf17* respectively) based on their product amino acid sequence similarities to those of known phage integrases, and the possession of the integrase specific motif pfam00589. If functional, their presence suggests the capability for a lysogenic lifecycle as well as a lytic one.

The phage GMA4 genome appears to encode several moron genes that may confer a selective advantage to its host lysogen. For example, this phage encodes a gene associated with phage resistance (*orf34*) (Henthorn & Friedman, 1995). The N-terminal region of Orf34 contains a Rha motif (pfam09669) that is thought to interfere with further phage infection of bacterial host strains lacking the integration host factor (IHF) (Henthorn & Friedman, 1995). This IHF regulates expression of the *rha* gene, and so may confer resistance to further phage attack to any bacterial host infected by it in a lysogenic cycle (Henthorn & Friedman, 1995). The temperate *Gordonia alkanivorans* phage GAL1 (see section 5) also contains a gene with this motif, suggesting it too might be a more common feature of *Gordonia* temperate phage genomes than presently believed.

4.4.11. Unexpected features of the Gordonia phages

As mentioned, all nine *Gordonia* phages sequenced in this study had highly novel genomes, with high percentages of ORFans (5 to 59%) (Table 4.5 in section 4.6), for which no statistically significant identifications could be made against sequences held in GenBank. Despite this, their gene products encode motifs suggestive of their putative function. For example, both phages GMA2 and GMA6 both contained a cd00233 motif in in their Orf14 and Orf13 putative proteins, respectively. These genes appear to encode a VIP2 family actin-ADP-ribosylating toxin with high specificity against the major insect pest, corn rootworms, and sharing a statistically significant sequence similarity with enzymatic components of other binary toxins, including *Clostridium botulinum* C2 toxin, *C. perfringens* iota toxin, *C. piroforme* toxin and *C. difficile* toxin (Han *et al.*, 2001).

Furthermore, phage GTE6 genome appeared to contain a gene (*orf12*), encoding a host cell surface-exposed lipoprotein since its expressed amino acid sequence shares homology with the pfram07553 motif. Such motifs are usually involved in superinfection exclusion, acting at the stage of DNA release from the phage head into the cell. These motifs have been associated with Superinfection exclusion (Sie) systems in temperate phages, where they interfere with co-infections from other phages (Ali *et al.*, 2014; Labrie *et al.*, 2010; Samson *et al.*, 2013). So the presence of such a motif in what appears to be an obligatory lytic phage is unexpected. Equally unexpected is that *orf21* in phage GTE6 encodes a putative Epstein-Barr nuclear antigen (Orf21), showing 35% amino acid sequence similarity to that of *Saccharomonospora* phage PIS 136 in this region (Bajpai *et al.*, 2012). Whether this homology reflects a similar function for the pair, or a distant evolutionary relationship between them is unknown. The *orf4* gene of phage GMA7 also appears to encode an unexpected motif (cd12820) normally associated with a putative adhesion virulence factor, forming a matrix on the bacterial outer membrane, which mediates binding to collagen and epithelial cells (El Tahir & Skurnik, 2001). What role if any it has in GMA7 phage is unknown.

4.4.12. Evolutionary ancestry of Gordonia phages

From the data presented here, it is clear that phages GTE8, even though isolated from creek water and GMA7 which came from activated sludge are genetically very similar to phages GTE5/GRU1, and GTE7, respectively. It is reasonable to suggest that these similarities reflect a closely shared ancestral past. Similar comments apply to phages GMA5 and GRU3. Despite not sharing nucleotide sequence identity with phage GTE7 DNA, the expressed amino acid sequences of phage GMA3 expressed genes are highly similar to it. Nine of the genes of phage GMA3 were most similar to those from GTE7, while 23 other genes were most similar to those from phage ReqiDocB7, to which phage GTE7 genome is closely related at an amino acid level (Petrovski *et al.*, 2011b; Summer *et al.*, 2011). As a similar closeness was not reflected at the nucleotide sequence level, one suggestion might be that more distant evolutionary relationships exist between phages GTE7, ReqiDocB7, and GMA3.

All individual *Gordonia* phage genomes sequenced here were unique, but given the close genetic relationships between the Mycolata host genera, attempts were made to classify these nine *Gordonia* phages according to the system of Hatfull *et al.* (2010) designed to show evolutionary relationships between the mycobacteriophages. It was not possible to place the nine *Gordonia* phages characterised here into any of the pre-existing clusters of mycobacteriophages. For example, while phage GMA7 is highly similar to phage GTE7 at a nucleotide sequence level, and phage GMA3 contains genes encoding several putative proteins also similar to those of phage GTE7, none could be grouped with other mycobacterium phages. Instead they emerged as singletons since none of the existing clusters embraced them.

4.4.13. Suitability of these phages for use in foam biocontrol

Of the nine phages examined in this study, phages GMA3, GMA5, GMA5 and GRU3 contained putative integrase genes, suggesting that they may undertake a lysogenic lifecycle. If these genes are functional, then these three phages are probably undesirable candidates for standard phage therapy for activated sludge foam control. However, they may be suitable for prophage induction therapy, an alternative to lytic phage control of foaming (see section 5).

Of the remaining five phages examined here, phages GMA2 and GMA6 both appear to contain a putative VIP2 family actin-ADP-ribosylating toxin gene. Consequently, neither phage would be assessed as being suitable for biocontrol strategies. The release of these phages into the environment may result in the spread of these undesirable genes and a possibly increase the virulence of other host bacteria.

All other *Gordonia* phages GMA7, GTE6, and GTE8 appear to be obligately lytic with no indication of integrase genes. Of these, phages GMA7 and GTE8 seem particularly attractive as both have impressive broad host ranges. For instance, GMA7 targets eleven strains of *Gordonia* species including those of *G. terrae, G. malaquae, G. rubropertincta, and G. hydrophobica*. Similarly, phage GTE8 targets several species including *G. terrae* (CON34^T, GOR9, G232) *and G. rubropertincta* (CON38^T), as well as *Nocardia asteroids* (strain CON12) (Table 4.1). Furthermore, phage GMA7 contains a putative DNA methylase gene (*orf38*) containing a pfam01555 motif. If this gene is functional, then this phage may evade cleavage by host defence RM systems (Labrie *et al.*, 2010; Samson *et al.*, 2013) and thus become an even more powerful addition to any phage therapy cocktail.

4.5. Conclusions

Nine phages infective for members of the genus *Gordonia* were isolated from wastewater and natural water environments, several of which had broad host ranges. Methods for phage visualisation of attachment and replication stages of phage infection using TEM were successfully applied and may prove to be useful in future studies looking at the mechanisms of phage infection. Whole genome sequencing of these nine *Gordonia* phages revealed that their genomes were all highly novel, failing to cluster with mycobacteriophages based on both nucleotide and amino acid sequence similarity. Some of these phages are less modular in their genomic architecture than those characterised previously, and some contain higher numbers of lysin genes than Mycolata phage genomes examined previously. Of the nine phages examined, three broad host range phages GMA7, GTE6, and GTE8 appear highly virulent and hence potentially suitable candidates for phage therapy cocktails to control activated sludge foaming.

4.6. Appendix

Table 4.5 Genome annotation of nine *Gordonia* phages

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA2-*orf1*	138..1817	560	ATP-binding protein [*Kitasatospora mediocidica*]	44	4e-90	Terminase (pfam03237)
GMA2-*orf2*	1862..3583	574	-	-	-	-
GMA2-*orf3*	3580..3978	133	-	-	-	-
GMA2-*orf4*	3971..4774	268	hypothetical protein [*Geobacter bremensis*]	24	4e-08	Glycotransferase (pfam04724)
GMA2-*orf5*	4814..5041	76	-	-	-	-
GMA2-*orf6*	5078..5734	219	hypothetical protein [*Xanthomonas vasicola*]	30	6e-10	Mu protein F (pfam04233)
GMA2-*orf7*	5731..5916	62	-	-	-	-
GMA2-*orf8*	5906..6217	104	sporulation protein [*Bacillus* sp. 72]	34	5e-05	Unknown (pfam07098)
GMA2-*orf9*	6225..7979	585	hypothetical protein [*Frankia alni*]	31	9e-51	Phage structural protein
GMA2-*orf10*	7981..10632	884	hypothetical protein [*Rhodococcus fascians*]	38	5e-15	Phage structural protein
GMA2-	10696..12123	476	hypothetical protein [*Streptomyces* sp. NRRL F-	55	1e-18	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf11			5135]			
GMA2-*orf12*	12206..13207	334	hypothetical protein [*Frankia alni*]	50	8e-53	-
GMA2-*orf13*	13226..14263	346	-	-	-	-
GMA2-*orf14*	14328..15059	244	-	-	-	VIP2; A family of actin-ADP-ribosylating toxin (cd00233)
GMA2-*orf15*	15052..15183	44	-	-	-	-
GMA2-*orf16*	15295..16959	555	mycobacteriophage protein [*Frankia alni*]	40	6e-64	Phage structural protein
GMA2-*orf17*	17011..17706	232	MULTISPECIES: hypothetical protein [*Actinomycetales*]	45	9e-60	-
GMA2-*orf18*	17722..18366	215	hypothetical protein [*Frankia alni*]	33	1e-10	-
GMA2-*orf19*	18455..19078	208	hypothetical protein [*Streptosporangium amethystogenes*]	29	2e-08	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA2-*orf20*	19143..19571	143	-	-	-	-
GMA2-*orf21*	19657..20244	196	glycosyltransferase/methyltransferase [*Mycobacterium* phage Llama]	56	9e-60	Methyltransferase (pfam13578)
GMA2-*orf22*	20237..20953	239	hypothetical protein [*Mycobacterium* sp. UM_RHS]	53	5e-84	Glycosyl transferase (pfam00535)
GMA2-*orf23*	21038..21652	205	glycosyltransferase [*Mycobacterium* phage CaptainTrips]	37	5e-30	-
GMA2-*orf24*	21753..22253	167	-	-	-	-
GMA2-*orf25*	22256..22675	140	hypothetical protein [*Streptomyces* sp. SPB74]	42	2e-09	Putative major tail protein
GMA2-*orf26*	22675..23118	148	-	-	-	Putative tail assembly protein
GMA2-*orf27*	23100..23402	101	-	-	-	Putative tail assembly protein translated by conserved programmed translational frameshift
GMA2-	23414..30640	240	unnamed protein product [*Rhodococcus* phage	39	1e-100	Tape measure protein (pfam03280; pfam13514;

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf28		9	REQ2]			COG5412)
GMA2-*orf29*	30651..31595	315	hypothetical protein [*Rhodococcus fascians*]	34	2e-46	-
GMA2-*orf30*	31614..33245	544	hypothetical protein [*Rhodococcus fascians*]	49	6e-171	-
GMA2-*orf31*	33254..33889	212	hypothetical protein [*Rhodococcus fascians*]	49	9e-12	-
GMA2-*orf32*	33886..36579	898	hypothetical protein [*Gordonia sihwensis*]	27	3e-55	-
GMA2-*orf33*	36581..38293	571	hypothetical protein [*Gordonia soli*]	70	1e-42	-
GMA2-*orf34*	38290..39237	316	-	-	-	-
GMA2-*orf35*	39327..40022	232	hypothetical protein [*Streptomyces* sp. SM8]	43	3e-40	Lysin/Peptidase (pfam13529)
GMA2-*orf36*	40019..40747	243	hypothetical protein [*Nocardia otitidiscaviarum*]	41	7e-39	Lysin (pfam01510)
GMA2-	40766..41137	124	N-acetylmuramoyl-L-alanine amidase	49	1e-08	Putative lysin

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf37			[*Rhodococcus opacus*]			
GMA2-*orf38*	41134..41802	223	papain cysteine protease family protein [*Mycobacterium xenopi* 3993]	50	1e-64	Lysin/Peptidase (cd02619)
GMA2-*orf39*	41812..42174	121	holin [*Mycobacterium* phage Milly]	36	1e-15	Putative holin
GMA2-*orf40*	42184..42681	166	hypothetical protein [*Rhodococcus fascians*]	36	2e-11	-
GMA2-*orf41*	42741..43205	155	hypothetical protein GTE7_gp030 [*Gordonia* phage GTE7]	35	3e-10	-
GMA2-*orf42*	43216..43593	126	-	-	-	-
GMA2-*orf43*	complement(43696..44082)	129	-	-	-	-
GMA2-*orf44*	complement(44075..44773)	233	thymidylate synthase, flavin-dependent [*Corynebacterium striatum* ATCC 6940]	57	1e-84	Thymidylate synthase (pfam02511)
GMA2-*orf45*	complement(44898..47093)	732	hypothetical protein [*Salinispora pacifica*]	30	2e-24	Cobalmin biosynthesis (pfam06213)
GMA2-	complement(47109..478	251	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf46	61)					
GMA2-*orf47*	complement(47995..48456)	154	unnamed protein product [*Gordonia* phage GRU1]	33	7e-11	-
GMA2-*orf48*	complement(48474..50507)	678	ATPase AAA [*Amycolatopsis thermoflava*]	40	3e-54	AAA protein (pfam07728)
GMA2-*orf49*	complement(50611..51057)	149	WhiB family transcriptional regulator [*Corynebacterium callunae*]	37	5e-07	Whib (pfam02467)
GMA2-*orf50*	complement(51091..51465)	125	-	-	-	-
GMA2-*orf51*	complement(51538..52830)	431	M2.BsmFI [*Geobacillus stearothermophilus*]	22	5e-06	C-5 cytosine specific methylase (pfam00145)
GMA2-*orf52*	complement(52827..53375)	183	deoxycytidine-triphosphatase [*Bacillus subtilis*]	35	1e-05	dUTPase (pfam08761)
GMA2-*orf53*	complement(53525..54511)	329	-	-	-	-
GMA2-*orf54*	complement(54513..55700)	396	-	-	-	Glycosyltransferase (COG0438)
GMA2-	complement(55700..560	132	hypothetical protein GTE7_gp062 [*Gordonia*	32	2e-04	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf55	95)		phage GTE7]			
GMA2-*orf56*	complement(56092..57162)	357	-	-	-	-
GMA2-*orf57*	complement(57205..57738)	178	-	-	-	GIY-YIG (cd10443)
GMA2-*orf58*	complement(60616..60747)	44	-	-	-	-
GMA2-*orf59*	complement(60935..62569)	545	RNA-binding protein [*Streptomyces carneus*]	53	0.0	TROVE (pram05731)
GMA2-*orf60*	complement(62877..63482)	202	hypothetical protein [*Vibrio* phage VpKK5]	31	1e-04	Unknown (pfam05037)
GMA2-*orf61*	complement(63565..64314)	250	hypothetical protein [*Caldanaerobius polysaccharolyticus*]	23	7e-11	AAA protein (13479)
GMA2-*orf62*	complement(64418..65413)	332	-	-	-	Nuclease (pfam12705)
GMA2-*orf63*	complement(65506..65715)	70	-	-	-	-
GMA2-	complement(65812..677	655	possible DNA helicase [*Aeromicrobium marinum*]	26	2e-47	Helicase (COG0553)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf64	76)					
GMA2-*orf65*	complement(67844..68347)	168	-	-	-	-
GMA2-*orf66*	complement(68344..70128)	595	gp44 [*Mycobacterium* phage Bxz2]	31	3e-68	DNA polymerase I (COG0749)
GMA2-*orf67*	complement(70132..70752)	207	hypothetical protein [*Arthrobacter* sp. 135MFCol5.1]	33	1e-10	Nuclease (cd10443; cd00283)
GMA2-*orf68*	complement(70884..71072)	63	-	-	-	-
GMA2-*orf69*	complement(71187..72845)	553	-	-	-	-
GMA2-*orf70*	complement(72861..73346)	162	cell division protein DedD [*Streptomyces* sp. NRRL F-2580]	37	2e-19	Deamimase (pfam00383)
GMA2-*orf71*	complement(73339..73629)	97	(2Fe-2S)-binding protein [*Rhodococcus opacus*]	36	3e-10	-
GMA2-*orf72*	complement(73712..74554)	281	methyltransferase [*Streptomyces* sp. NRRL S-340]	43	3e-23	Thymidylate synthase (pfam00303)
GMA2-	complement(74572..753	264	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf73	63)					
GMA2-*orf74*	complement(75371..75958)	196	hypothetical protein ACD_80C00175G0001 [uncultured bacterium (gcode 4)]	38	1e-04	Endonuclease (pfam01844)
GMA2-*orf75*	complement(76020..76604)	195	-	-	-	-
GMA2-*orf76*	complement(76664..77386)	241	hypothetical protein [*Gordonia soli*]	38	5e-15	-
GMA2-*orf77*	complement(77394..77513)	40	-	-	-	-
GMA2-*orf78*	complement(77510..78064)	185	-	-	-	-
GMA2-*orf79*	complement(78057..78731)	225	unnamed protein product [*Rhodococcus* phage REQ2]	44	3e-40	Phosphoesterase (COG4186)
GMA2-*orf80*	complement(78728..79051)	108	-	-	-	-
GMA2-*orf81*	complement(79048..79308)	87	-	-	-	-
GMA2-	complement(79524..798	102	hypothetical protein [*Mycobacterium abscessus*]	43	1e-08	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf82	29)					
GMA2-*orf83*	complement(79814..80008)	65	-	-	-	-
GMA2-*orf84*	complement(80010..80351)	114	hypothetical protein [*Rhodococcus fascians*]	69	3e-45	-
GMA2-*orf85*	complement(80348..80551)	68	-	-	-	-
GMA2-*orf86*	complement(80749..81003)	85	-	-	-	-
GMA2-*orf87*	complement(81000..81263)	88	-	-	-	-
GMA2-*orf88*	complement(81347..81784)	146	MULTISPECIES: hypothetical protein [*Micrococcineae*]	53	2e-38	-
GMA2-*orf89*	complement(81781..81942)	54	-	-	-	-
GMA2-*orf90*	complement(82002..82721)	240	hypothetical protein [*Gordonia malaquae*]	64	4e-30	-
GMA2-	complement(82725..832	190	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf91	94)					
GMA2-*orf92*	complement(83392..83565)	58	-	-	-	-
GMA2-*orf93*	complement(83562..84947)	462	DNA primase [*Fervidicella metallireducens* AeB]	32	8e-11	Bifunctional DNA primase/polymerase (pfam09250)
GMA2-*orf94*	complement(85005..87089)	695	hypothetical protein [*Escherichia coli*]	25	5e-10	Unknown (pfam13148)
GMA2-*orf95*	complement(87170..87367)	66	hypothetical protein [*Hellea balneolensis*]	48	2e-08	HTH DNA binding (pfam12728)
GMA2-*orf96*	complement(87394..87714)	107	-	-	-	-
GMA2-*orf97*	complement(88763..88972)	70	-	-	-	-
GMA2-*orf98*	complement(89408..89671)	88	-	-	-	-
GMA2-*orf99*	complement(89680..90450)	257	hypothetical protein TPA2_gp53 [*Tsukamurella* phage TPA2]	48	3e-20	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA2-*orf100*	complement(91125..91364)	80	-	-	-	-
GMA2-*orf101*	complement(91583..91927)	115	-	-	-	-
GMA2-*orf102*	complement(91934..92176)	81	-	-	-	-
GMA2-*orf103*	complement(92154..92477)	108	hypothetical protein [*Mycobacterium marinum*]	51	1e-15	-
GMA2-*orf104*	complement(92474..92803)	110	hypothetical protein HMPREF1211_07474 [Streptomyces sp. HGB0020]	39	1e-06	-
GMA2-*orf105*	complement(92784..93677)	298	-	-	-	Chromosome segregation ATPase (COG1196)
GMA2-*orf106*	complement(93680..93946)	89	hypothetical protein PBI_LLAMA_56 [*Mycobacterium* phage Llama]	42	2e-04	-
GMA2-*orf107*	complement(93939..94211)	91	gp49 [*Mycobacterium* phage PMC]	49	2e-13	-
GMA2-*orf108*	complement(94208..94429)	74	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA2-*orf109*	complement(94426..94866)	147	-	-	-	-
GMA2-*orf110*	complement(94967..95407)	147	gp056 [*Rhodococcus* phage ReqiDocB7]	34	3e-05	-
GMA2-*orf111*	complement(95493..95606)	38	-	-	-	-
GMA2-*orf112*	complement(95579..97174)	532	ADP-ribosylation/Crystallin J1 [*Mycobacterium rhodesiae*]	44	6e-26	ADP-ribosylglycohydrolase (pfam03747)
GMA2-*orf113*	complement(97225..97428)	68	-	-	-	-
GMA2-*orf114*	complement(97531..97956)	142	-	-	-	-
GMA2-*orf115*	complement(97985..98689)	235	-	-	-	Nucleotideyltransferase (pfam01909)
GMA2-*orf116*	complement(98704..98919)	72	-	-	-	-
GMA2-*orf117*	complement(98912..99322)	137	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA2-*orf118*	complement(99358..99921)	188	-	-	-	Whib (pfam02467; pfam07900)
GMA2-*orf119*	complement(99959..100168)	70	-	-	-	-
GMA2-*orf120*	complement(100181..100528)	116	-	-	-	-
GMA2-*orf121*	complement(100688..100930)	81	-	-	-	-
GMA2-*orf122*	complement(100931..101578)	216	hypothetical protein BADFISH_56 [*Mycobacterium* phage Badfish]	28	6e-09	-
GMA2-*orf123*	complement(101652..102011)	120	-	-	-	-
GMA2-*orf124*	complement(102069..102752)	228	-	-	-	-
GMA2-*orf125*	complement(102784..103044)	87	-	-	-	-
GMA2-*orf126*	complement(103105..103344)	80	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA3-*orf1*	105..1739	545	putative small terminase [*Gordonia* phage GTE7]	30	2e-16	Putative small terminase subunit
GMA3-*orf2*	1740..4886	1049	TerL [*Rhodococcus* phage ReqiDocB7]	48	1e-112	Large terminase subunit (cd01335)
GMA3-*orf3*	4995..6896	634	Mu gp29-like protein [*Rhodococcus* phage ReqiDocB7]	38	1e135	Unknown (pfam06074)
GMA3-*orf4*	6899..7189	97	-	-	-	-
GMA3-*orf5*	7211..8362	384	gp010 [*Rhodococcus* phage ReqiDocB7]	42	3e-70	-
GMA3-*orf6*	8376..8843	156	gp011 [*Rhodococcus* phage ReqiDocB7]	43	2e-32	-
GMA3-*orf7*	8860..10044	395	putative major capsid protein [*Gordonia* phage GTE7]	46	3e-108	Phage major capsid protein (pfam03864)
GMA3-*orf8*	10068..10358	97	-	-	-	-
GMA3-*orf9*	10402..11406	335	gp015 [*Rhodococcus* phage ReqiDocB7]	32	1e-12	Unknown (pfam07030)
GMA3-*orf10*	11406..11906	167	hypothetical protein GTE7_gp013 [*Gordonia* phage GTE7]	31	6e-04	-
GMA3-*orf11*	11907..12410	168	gp017 [*Rhodococcus* phage ReqiDocB7]	33	8e-18	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA3-*orf12*	12429..12743	105	-	-	-	-
GMA3-*orf13*	12749..13774	342	gp019 [*Rhodococcus* phage ReqiDocB7]	49	2e-104	Putative major tail protein
GMA3-*orf14*	13894..14499	202	pre-TMP frameshift protein [*Rhodococcus* phage ReqiDocB7]	35	4e-25	Putative tail assembly protein
GMA3-*orf15*	14481..14756	92	truncated pre-TMP frameshift protein [*Rhodococcus* phage ReqiDocB7]	36	1e-07	Putative tail assembly protein translated by conserved programmed translational frameshift
GMA3-*orf16*	14859..24419	3187	tape measure protein [*Rhodococcus* phage ReqiDocB7]	34	5e-174	Tape measure protein (pfam10145; pfam12889; COG5412; pfam01464)
GMA3-*orf17*	24422..27004	861	gp023 [*Rhodococcus* phage ReqiDocB7]	47	0.0	-
GMA3-*orf18*	27013..28752	580	gp024 [*Rhodococcus* phage ReqiDocB7]	50	0.0	Putative phage tail protein (pfam13550)
GMA3-*orf19*	28749..31493	915	hypothetical protein [*Gordonia sihwensis*]	28	1e-58	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA3-*orf20*	31495..33216	574	hypothetical protein [*Gordonia soli*]	64	1e-37	-
GMA3-*orf21*	33213..33800	196	hypothetical protein [*Gordonia soli*]	39	1e-12	-
GMA3-*orf22*	33881..34582	234	twin-arginine translocation pathway signal [*Mycobacterium phlei*]	48	1e-59	Lysin (cd06418)
GMA3-*orf23*	34586..34939	118	-	-	-	-
GMA3-*orf24*	34939..35589	217	hypothetical protein [*Rhodococcus fascians*]	65	6e-69	Putative lysozyme
GMA3-*orf25*	35593..36075	161	hypothetical protein FG87_22005 [*Nocardia* sp. W9851]	36	6e-24	Nuclease (pfam13392)
GMA3-*orf26*	36068..36415	116	putative peptidase [*Gordonia* phage GTE7]	46	7e-12	Putative lysin
GMA3-*orf27*	36415..36849	145	hypothetical protein GTE7_gp026 [*Gordonia* phage GTE7]	48	2e-21	Puative holin
GMA3-*orf28*	36944..37429	162	hypothetical protein [*Rhodococcus fascians*]	26	2e-04	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA3-*orf29*	37413..37865	151	hypothetical protein [*Salinispora arenicola*]	29	1e-05	-
GMA3-*orf30*	37862..38239	126	hypothetical protein [*Nocardia araoensis*]	29	1e-04	-
GMA3-*orf31*	38249..38593	115	-	-	-	-
GMA3-*orf32*	38590..39348	253	hypothetical protein [*Rhodococcus opacus*]	41	1e-45	PE-PPE (pfam08237)
GMA3-*orf33*	complement(39394..39585)	64	-	-	-	-
GMA3-*orf34*	complement(39572..39685)	38	-	-	-	-
GMA3-*orf35*	complement(39682..39822)	47	-	-	-	-
GMA3-*orf36*	complement(39826..40110)	95	-	-	-	-
GMA3-*orf37*	complement(40113..40325)	71	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA3-*orf38*	complement(40325..40837)	171	hypothetical protein GTE7_gp041 [*Gordonia* phage GTE7]	47	5e-36	Nucleoside Triphosphate Pyrophosphohydroplase (cd11542)
GMA3-*orf39*	complement(40830..41243)	138	hypothetical protein GTE7_gp101 [*Gordonia* phage GTE7]	48	2e-30	-
GMA3-*orf40*	complement(41240..41542)	101	hypothetical protein F989_02392 [*Acinetobacter parvus* NIPH 1103]	39	6e-07	-
GMA3-*orf41*	complement(41529..41789)	87	hypothetical protein PBI_HAWKEYE_80 [*Mycobacterium* phage Hawkeye]	48	2e-14	-
GMA3-*orf42*	complement(41794..42519)	242	hypothetical protein [*Rhodococcus fascians*]	48	3e-58	Thymidylate synthase (pfam02511)
GMA3-*orf43*	complement(42516..42854)	113	-	-	-	-
GMA3-*orf44*	complement(43209..44399)	397	gp033 [*Rhodococcus* phage ReqiDocB7]	30	2e-55	Nuclease (COG2887)
GMA3-*orf45*	complement(44400..44606)	69	-	-	-	-
GMA3-*orf46*	complement(44610..44762)	51	gp037 [*Rhodococcus* phage ReqiDocB7]	43	6e-04	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA3-*orf47*	complement(44762..44971)	70	-	-	-	-
GMA3-*orf48*	complement(44973..45185)	71	transcriptional regulator [*Mycobacterium abscessus*]	40	3e-08	WhiB transcription factor (pfam02467)
GMA3-*orf49*	complement(45224..45505)	94	-	-	-	-
GMA3-*orf50*	complement(45516..46070)	185	gp042 [*Rhodococcus* phage ReqiDocB7]	33	2e -10	Holliday Junction Resolvase/RusA (pfam05866)
GMA3-*orf51*	complement(46067..46372)	102	-	-	-	-
GMA3-*orf52*	complement(46369..46569)	67	-	-	-	-
GMA3-*orf53*	complement(46542..47669)	376	DnaN [*Rhodococcus* phage ReqiDocB7]	23	5e-20	DNA Polymerase III beta subunit (COG0592)
GMA3-*orf54*	complement(47666..48004)	113	-	-	-	-
GMA3-*orf55*	complement(48010..48480)	157	hypothetical protein [*Gordonia otitidis*]	48	1e-27	Unknown (pfam10686)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA3-*orf56*	complement(48480..49106)	209	oligoribonuclease [*Corynebacterium genitalium*]	38	3e-29	Nuclease (pfam00929)
GMA3-*orf57*	complement(49106..49348)	81	-	-	-	HTH DNA binding protein (pfam12728)
GMA3-*orf58*	complement(49351..50868)	506	helicase [*Rhodococcus* phage ReqiDocB7]	49	4e-162	Helicase (COG0553)
GMA3-*orf59*	complement(50871..51248)	126	-	-	-	-
GMA3-*orf60*	complement(51248..51466)	73	-	-	-	-
GMA3-*orf61*	complement(51469..51882)	138	hypothetical protein [*Gordonia sihwensis*]	36	7e-12	Unknown (pfam05305)
GMA3-*orf62*	complement(52079..52285)	69	-	-	-	-
GMA3-*orf63*	complement(52397..53080)	228	gp051 [*Rhodococcus* phage ReqiDocB7]	40	2e-21	-
GMA3-*orf64*	complement(53329..53586)	86	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA3-*orf65*	complement(53600..54439)	280	-	-	-	-
GMA3-*orf66*	complement(54429..54884)	152	hypothetical protein GTE7_gp056 [*Gordonia* phage GTE7]	29	7e-04	-
GMA3-*orf67*	complement(54868..54996)	43	-	-	-	-
GMA3-*orf68*	complement(54993..55214)	74	-	-	-	-
GMA3-*orf69*	complement(55207..55536)	110	-	-	-	-
GMA3-*orf70*	complement(55533..55790)	86	-	-	-	-
GMA3-*orf71*	complement(55777..55950)	58	-	-	-	-
GMA3-*orf72*	complement(55952..56224)	91	-	-	-	-
GMA3-*orf73*	complement(56241..56648)	136	PREDICTED: centromere protein F-like [*Musa acuminata* subsp. malaccensis]	28	6e-04	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA3-*orf74*	complement(56641..56823)	61	-	-	-	-
GMA3-*orf75*	complement(56823..57620)	266	recombinase XerC [*Thermococcus* sp. AM4]	34	4e-22	Integrase (pfam00589)
GMA3-*orf76*	complement(57695..58531)	279	-	-	-	-
GMA3-*orf77*	complement(58541..58870)	110	-	-	-	-
GMA3-*orf78*	complement(58867..59571)	235	gp064 [*Rhodococcus* phage ReqiDocB7]	34	3e-10	-
GMA3-*orf79*	complement(59664..59888)	75	hypothetical protein [*Rhodococcus* sp. UNC363MFTsu5.1]	40	1e-08	-
GMA3-*orf80*	complement(59888..60112)	75	-	-	-	-
GMA3-*orf81*	complement(60109..60351)	81	-	-	-	-
GMA3-*orf82*	complement(60344..60877)	178	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA3-*orf83*	complement(60883..61623)	247	hypothetical protein [*Gordonia otitidis*]	37	1e-29	-
GMA3-*orf84*	complement(61626..61793)	56	-	-	-	-
GMA3-*orf85*	complement(61762..62001)	80	-	-	-	-
GMA3-*orf86*	complement(61991..62260)	90	-	-	-	-
GMA3-*orf87*	complement(62260..64299)	680	vWFA [*Rhodococcus* phage ReqiDocB7]	36	5e-42	von Willebrand factor (pfam13519)
GMA3-*orf88*	complement(64299..65969)	557	ATPase family protein [*Gordonia* phage GTE7]	46	2e-121	AAA protein (pfam07728)
GMA3-*orf89*	complement(66035..66172)	46	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA3-*orf90*	complement(66406..66882)	159	PREDICTED: RNA-binding motif protein, X-linked-like-2 [*Pan troglodytes*]	39	2e-04	-
GMA3-*orf91*	complement(67333..68085)	251	-	-	-	-
GMA3-*orf92*	complement(68228..68671)	148	gp091 [*Rhodococcus* phage ReqiDocB7]	35	4e-11	-
GMA3-*orf93*	complement(68717..69091)	125	gp162 [*Mycobacterium* phage Wildcat]	46	8e-12	-
GMA3-*orf94*	complement(69177..69422)	82	-	-	-	-
GMA3-*orf95*	complement(69410..69646)	79	-	-	-	-
GMA3-*orf96*	complement(69678..70583)	302	-	-	-	-
GMA3-	complement(70650..707	39	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf97	66)					
GMA3-*orf98*	complement(71182..71544)	121	hypothetical protein A306_06092 [*Columba livia*]	58	9e-13	-
GMA3-*orf99*	complement(72890..73144)	85	-	-	-	-
GMA3-*orf100*	complement(73141..75102)	654	gp102 [*Rhodococcus* phage ReqiDocB7]	29	2e-44	Primase (TIGR01391)
GMA3-*orf101*	complement(75099..75464)	122	hypothetical protein COCSUDRAFT_57208 [*Coccomyxa subellipsoidea* C-169]	35	4e-09	-
GMA3-*orf102*	complement(75728..76318)	197	-	-	-	-
GMA3-*orf103*	complement(76315..77079)	225	gp105 [*Rhodococcus* phage ReqiDocB7]	30	6e-23	-
GMA3-*orf104*	complement(77179..77301)	41	-	-	-	-
GMA4-*orf1*	79..492	138	hypothetical protein LIKA_5 [*Streptomyces* phage Lika]	60	2e-44	Putative small terminase subunit
GMA4-*orf2*	489..2087	533	terminase [*Streptomyces auratus*]	56	0.0	Putative large terminase

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
						subunit
GMA4-*orf3*	2101..3495	465	hypothetical protein [*Gordonia malaquae*]	97	0.0	Portal protein (pfam05133)
GMA4-*orf4*	3492..5039	516	hypothetical protein [*Gordonia malaquae*]	99	0.0	Capsid maturation protease (PRK14694)
GMA4-*orf5*	5343..5924	194	hypothetical protein [*Gordonia malaquae*]	94	7e-113	Unknown (pfam14265)
GMA4-*orf6*	5937..6311	125	hypothetical protein [*Gordonia malaquae*]	98	8e-73	-
GMA4-*orf7*	6326..7234	303	hypothetical protein [*Gordonia malaquae*]	93	0.0	Phage structural protein
GMA4-*orf8*	7238..7465	76	hypothetical protein [*Gordonia malaquae*]	92	1e-31	-
GMA4-*orf9*	7458..7847	130	hypothetical protein [*Gordonia malaquae*]	96	2e-81	Phage protein (pfam09355)
GMA4-*orf10*	7847..8170	108	hypothetical protein [*Gordonia malaquae*]	99	1e-68	-
GMA4-*orf11*	8224..8490	89	hypothetical protein [*Gordonia malaquae*]	76	6e-33	-
GMA4-*orf12*	8487..8873	129	hypothetical protein [*Gordonia malaquae*]	98	3e-83	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA4-*orf13*	8955..9626	224	hypothetical protein [*Gordonia malaquae*]	92	1e-142	Putative major tail structural protein
GMA4-*orf14*	9946..10263	106	hypothetical protein [*Gordonia malaquae*]	99	3e-67	Putative tail assembly protein
GMA4-*orf15*	10245..10748	168	hypothetical protein [*Gordonia malaquae*]	98	1e-89	Putative tail assembly protein translated by conserved programmed translational frameshift
GMA4-*orf16*	10768..16050	1761	hypothetical protein [*Gordonia malaquae*]	97	0.0	Tape measure protein (pfam05701; COG5412; pfam01464)
GMA4-*orf17*	16043..16957	305	hypothetical protein [*Gordonia malaquae*]	99	0.0	-
GMA4-*orf18*	16957..18117	387	hypothetical protein [*Gordonia malaquae*]	98	0.0	Unknown(pfam14594)
GMA4-*orf19*	18117..19190	358	hypothetical protein [*Gordonia malaquae*]	97	0.0	-
GMA4-*orf20*	19192..20256	355	hypothetical protein [*Gordonia malaquae*]	96	0.0	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA4-*orf21*	20328..21401	358	hypothetical protein [*Gordonia rubripertincta*]	58	1e-120	Lysin (pfam01510; pfam08310)
GMA4-*orf22*	21398..21685	96	hypothetical protein [*Gordonia malaquae*]	93	2e-56	Glutaredoxin (pfam00462)
GMA4-*orf23*	21682..21882	67	DNA polymerase I [*Leifsonia xyli*]	46	8e-07	-
GMA4-*orf24*	21879..22325	149	hypothetical protein [*Gordonia malaquae*]	95	1e-91	-
GMA4-*orf25*	complement(22399..22749)	117	hypothetical protein [*Gordonia malaquae*]	84	1e-46	-
GMA4-*orf26*	22913..23575	221	hypothetical protein [*Mycobacterium colombiense*]	37	7e-06	Unknown (pfam05305)
GMA4-*orf27*	23590..23799	70	-	-	-	-
GMA4-*orf28*	complement(23993..24892)	300	hypothetical protein [*Rhodococcus pyridinivorans*]	38	4e-33	-
GMA4-*orf29*	complement(25217..26449)	411	phage integrase family protein [*Rhodococcus pyridinivorans*]	43	8e-82	Integrase (pfam00589)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA4-*orf30*	complement(26442..26648)	69	hypothetical protein [*Gordonia malaquae*]	74	8e-28	Unknown (pfam11662)
GMA4-*orf31*	complement(26645..27079)	145	unnamed protein product [*Rhodococcus* phage REQ2]	64	3e-54	Unknown (pfam06114)
GMA4-*orf32*	complement(27088..27585)	166	ribosomal protein S13 [*Corynebacterium falsenii* DSM 44353]	37	1e-20	Ribosomal protein S13
GMA4-*orf33*	27730..27987	86	hypothetical protein [*Corynebacterium ulcerans*]	45	7e-09	HTH DNA Binding (pfam01381)
GMA4-*orf34*	28046..28852	269	hypothetical protein [*Mycobacterium abscessus*]	64	4e-59	Anti-repressor/Rha regulatory protein (pfam03374; pfam09669)
GMA4-*orf35*	28849..29106	86	-	-	-	-
GMA4-*orf36*	29103..29303	67	DNA-binding protein [*Streptomyces sclerotialus*]	41	4e-08	HTH DNA Binding (pfam12728)
GMA4-*orf37*	29315..29482	56	-	-	-	-
GMA4-*orf38*	29479..29760	94	hypothetical protein [*Mycobacterium avium*]	39	8e-12	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA4-*orf39*	29757..30101	115	gp54 [*Mycobacterium* phage Charlie]	38	4e-13	-
GMA4-*orf40*	30065..30856	264	hypothetical protein [*Rhodococcus opacus*]	42	2e-55	-
GMA4-*orf41*	30857..31489	211	hypothetical protein [*Rhodococcus* sp. 29MFTsu3.1]	33	3e-11	-
GMA4-*orf42*	31658..32065	136	gp58 [*Mycobacterium* phage Dori]	52	2e-39	-
GMA4-*orf43*	32065..32415	117	gp82 [*Mycobacterium* phage Bxb1]	38	5e-12	-
GMA4-*orf44*	32408..32542	45	-	-	-	-
GMA4-*orf45*	32539..32883	115	-	-	-	-
GMA4-*orf46*	32880..33362	161	hypothetical protein [*Gordonia malaquae*]	98	1e-35	-
GMA4-*orf47*	33410..33589	60	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA4-*orf48*	33582..33890	103	hypothetical protein [*Gordonia malaquae*]	95	7e-46	-
GMA4-*orf49*	33848..34048	67	MULTISPECIES: hypothetical protein [*Streptomyces*]	55	6e-05	-
GMA4-*orf50*	34041..34310	90	hypothetical protein [*Gordonia alkanivorans*]	40	7e-09	-
GMA4-*orf51*	complement(34307..34519)	71	hypothetical protein [*Gordonia malaquae*]	97	5e-41	-
GMA4-*orf52*	34584..34763	60	hypothetical protein EN35_20025 [*Rhodococcus qingshengii*]	34	1e-04	-
GMA4-*orf53*	34736..34858	41	-	-	-	-
GMA4-*orf54*	34872..35675	268	DNA methylase N-4 [*Corynebacterium aurimucosum*]	73	2e-138	DNA Methylase (pfam01555)
GMA4-*orf55*	35668..35970	101	hypothetical protein [*Gordonia malaquae*]	100	3e-56	-
GMA4-*orf56*	35967..36167	67	hypothetical protein [*Gordonia malaquae*]	100	6e-39	DNA binding (cd00569)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA4-*orf57*	36164..36322	53	-	-	-	-
GMA4-*orf58*	36626..39130	835	hypothetical protein [*Mycobacterium avium*]	47	0.0	Primase (pfam08706)
GMA4-*orf59*	39801..40205	135	-	-	-	-
GMA4-*orf60*	40216..40572	119	hypothetical protein [*Rhodococcus equi*]	53	6e-27	-
GMA4-*orf61*	40576..40734	53	-	-	-	-
GMA4-*orf62*	40731..41486	252	hypothetical protein [*Rhodococcus* sp. UNC363MFTsu5.1]	42	6e-51	-
GMA4-*orf63*	41745..41891	49	hypothetical protein [*Gordonia malaquae*]	55	8e-08	-
GMA4-*orf64*	42024..42347	108	hypothetical protein [*Nocardia farcinica*]	33	5e-06	-
GMA4-*orf65*	42338..42712	125	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA4-*orf66*	42724..43833	370	hypothetical protein [*Aeromicrobium marinum*]	47	1e-59	Phage related tail structural protein (COG5310)
GMA4-*orf67*	43834..4449043840..44490	219 217	HNH homing endonuclease domain protein [*Mycobacterium* phage Hamulus]	445	4e-37	Endonuclease (pfam13392; pfam07463)
GMA4-*orf68*	44497..45141	215	hypothetical protein [*Amycolatopsis taiwanensis*]	29	2e-06	-
GMA5-*orf1*	93..500	136	hypothetical protein [*Gordonia neofelifaecis*]	72	9e-43	Small terminase subunit
GMA5-*orf2*	466..1860	465	putative phage terminase protein [*Gordonia neofelifaecis*]	68	0.0	Large terminase subunit (pfam03354)
GMA5-*orf3*	1869..2066	66	hypothetical protein [*Gordonia neofelifaecis*]	52	3e-09	-
GMA5-*orf4*	2108..3199	364	hypothetical protein [*Gordonia neofelifaecis*]	68	5e-158	Portal protein (pfam04860)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA5-*orf5*	3196..3984	263	hypothetical protein [*Gordonia soli*]	51	7e-67	Lysin - D-alanyl-D-alanine carboxypeptidase (pfam13539)
GMA5-*orf6*	4085..6046	654	hypothetical protein [*Gordonia neofelifaecis*]	63	0.0	-
GMA5-*orf7*	6050..6391	114	unnamed protein product [*Rhodococcus* phage RRH1]	46	3e-19	-
GMA5-*orf8*	6391..6732	114	hypothetical protein [*Gordonia neofelifaecis*]	63	2e-35	-
GMA5-*orf9*	6746..7207	154	hypothetical protein [*Gordonia neofelifaecis*]	76	3e-77	-
GMA5-*orf10*	7204..7551	116	hypothetical protein [*Gordonia neofelifaecis*]	47	4e-22	Phage protein HK97/gp10 family- possibly tail morphogenesis (TIGR01725)
GMA5-*orf11*	7567..7860	98	hypothetical protein [*Gordonia neofelifaecis*]	60	7e-30	-
GMA5-*orf12*	7975..10002	676	TP901 family phage tail tape measure protein , putative [*Gordonia neofelifaecis*]	54	1e-176	Tape measure protein (COG5412)
GMA5-*orf13*	9999..11417	473	unnamed protein product [*Rhodococcus* phage RRH1]	25	2e-23	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA5-*orf14*	11420..12025	202	unnamed protein product [*Rhodococcus* phage RRH1]	30	1e-15	Phage structural protein
GMA5-*orf15*	12085..12933	283	bacteriophage protein [*Mycobacterium thermoresistibile*]	41	2e-59	Phage structural protein (pfam08237)
GMA5-*orf16*	complement(13003..13107)	35	-	-	-	-
GMA5-*orf17*	complement(13184..13993)	270	integrase [*Gordonia neofelifaecis*]	69	2e-122	Integrase (pfam00589)
GMA5-*orf18*	complement(14074..14349)	92	hypothetical protein [*Salinispora arenicola*]	62	5e-24	-
GMA5-*orf19*	complement(14346..14633)	96	putative DNA-binding protein [*Gordonia neofelifaecis*]	49	2e-16	HTH DNA binding domain (pfam12844)
GMA5-*orf20*	14712..14891	60	hypothetical protein [*Tomitella biformata*]	51	1e-07	HTH DNA binding domain (pfam12728)
GMA5-*orf21*	14885..15145	87	hypothetical protein [*Gordonia neofelifaecis*]	43	2e-11	-
GMA5-*orf22*	15142..15327	62	hypothetical protein [*Gordonia neofelifaecis*]	49	2e-06	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA5-*orf23*	15324..15518	65	hypothetical protein [*Gordonia malaquae*]	69	2e-20	Hin/HTH DNA binding domain (cd00569)
GMA5-*orf24*	15607..15828	74	-	-	-	-
GMA5-*orf25*	15816..16079	88	-	-	-	-
GMA5-*orf26*	16193..17065	291	DNA polymerase III subunit epsilon [*Gordonia neofelifaecis*]	53	1e-51	DNA polymerase III subunit epsilon (COG0847)
GMA5-*orf27*	17062..17340	93	unnamed protein product [*Rhodococcus* phage RRH1]	62	8e-31	HNH endonuclease (pfam01844)
GMA5-*orf28*	17424..17558	45	-	-	-	-
GMA6-*orf1*	21..206	62	-	-	-	-
GMA6-*orf2*	203..373	57	-	-	-	Putative small terminase subunit
GMA6-*orf3*	370..2850	827	large terminase subunit [*Methanobacterium* phage psiM2]	38	2e-54	Large terminase subunit (pfam03237; PRK14715; smart00306)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA6-*orf4*	2942..3256	105	-	-	-	-
GMA6-*orf5*	3281..3460	60	-	-	-	-
GMA6-*orf6*	3457..3711	85	-	-	-	-
GMA6-*orf7*	3711..4031	107	-	-	-	-
GMA6-*orf8*	4028..4612	195	gp40 [*Mycobacterium* phage Che12]	47	2e-24	Nucleoside Triphosphate Pyrophosphohydrolase (cd11541)
GMA6-*orf9*	4609..4851	81	-	-	-	-
GMA6-*orf10*	4844..5236	131	-	-	-	-
GMA6-*orf11*	5324..7891	856	hypothetical protein [*Streptomyces* sp. Amel2xE9]	42	3e-79	Portal protein (pfam04860; pfam04233)
GMA6-*orf12*	7888..8862	325	hypothetical protein [*Meiothermus chliarophilus*]	31	1e-16	RNA ligase (pfam13563)
GMA6-*orf13*	8862..9977	372	-	-	-	VIP2; A family of actin-ADP-ribosylating toxin (cd00233)
GMA6-	9974..13003	101	hypothetical protein [*Tomitella biformata*]	44	1e-08	HNH endonuclease

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf14		0				(pfam01844)
GMA6-*orf15*	13000..13599	200	hypothetical protein [*Rhodococcus fascians*]	56	1e-11	-
GMA6-*orf16*	13599..14291	231	hypothetical protein [*Gordonia soli*]	37	1e-28	-
GMA6-*orf17*	14291..14863	191	hypothetical protein [*Segniliparus rugosus*]	48	9e-12	-
GMA6-*orf18*	14860..15507	216	hypothetical protein [*Gordonia neofelifaecis*]	55	5e-37	-
GMA6-*orf19*	15585..17039	485	hypothetical protein [*Streptomyces rimosus*]	32	3e-35	Prohead protease (pfam04586)
GMA6-*orf20*	17179..18546	456	capsid protein [*Streptomyces* sp. PRh5]	36	9e-73	Phage capsid structural protein (pfam05065)
GMA6-*orf21*	18597..18785	63	-	-	-	-
GMA6-*orf22*	18785..19093	103	-	-	-	-
GMA6-	19099..19305	69	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf23						
GMA6-*orf24*	19372..19947	192	-	-	-	-
GMA6-*orf25*	19951..20526	192	-	-	-	-
GMA6-*orf26*	20625..20996	124	phage protein, HK97 gp10 family [*Sideroxydans lithotrophicus*]	37	3e-05	Virion morphogenesis protein (pfam05069)
GMA6-*orf27*	20993..21688	232	-	-	-	-
GMA6-*orf28*	21703..22032	110	-	-	-	-
GMA6-*orf29*	22036..23484	483	hypothetical protein [*Kribbella catacumbae*]	43	4e-50	Phage tail sheath structural protein (pfam04984)
GMA6-*orf30*	23527..23934	136	phage tail protein [*Algoriphagus marincola*]	15	5e-38	Putative major tail protein (pfam06841)
GMA6-*orf31*	24035..24604	190	hypothetical protein [*Kribbella catacumbae*]	26	6e-05	Putative tail assembly protein

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA6-*orf32*	24586..24801	72	-	-	-	Putative tail assembly protein translated by conserved programmed translational frameshift
GMA6-*orf33*	24842..28261	1140	peptidase M23 [*Staphylococcus* sp. URHA0057]	26	3e-29	Tape measure protein (COG5412)
GMA6-*orf34*	28264..29061	266	hypothetical protein [*Kribbella catacumbae*]	23	1e-04	Lysin (pfam01476)
GMA6-*orf35*	29073..29516	148	hypothetical protein [*Streptomyces albus*]	41	2e-20	Endonuclease (pfam13392)
GMA6-*orf36*	29518..30879	454	unnamed protein product [*Rhodococcus* phage REQ2]	63	9e-49	-
GMA6-*orf37*	30895..32241	449	hydrolase Nlp/P60 [*Gordonia rhizosphera*]	51	2e-29	Cell wall hydrolase (COG0791)
GMA6-*orf38*	32254..32694	147	-	-	-	-
GMA6-*orf39*	32691..33092	134	-	-	-	-
GMA6-	33089..33412	108	hypothetical protein [*Kribbella catacumbae*]	33	6e-09	Lysozyme (pfam04965)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf40						
GMA6-*orf41*	33414..34571	386	hypothetical protein [*Kribbella catacumbae*]	39	3e-68	Unknown (COG3299)
GMA6-*orf42*	34564..35736	391	hypothetical protein [*Kribbella catacumbae*]	29	1e-19	-
GMA6-*orf43*	35754..37736	661	hypothetical protein [*Kribbella catacumbae*]	32	7e-04	Phage structural protein
GMA6-*orf44*	37736..38137	134	-	-	-	-
GMA6-*orf45*	38130..41393	1088	hypothetical protein [*Mycobacterium* sp. URHD0025]	57	1e-120	Lysin - Peptidase (cd06418; pfam01510; pfam01551; pfam13810)
GMA6-*orf46*	41390..41899	170	hypothetical protein [*Corynebacterium argentoratense*]	38	1e-06	Putative holin
GMA6-*orf47*	41896..42207	104	-	-	-	-
GMA6-*orf48*	42191..42718	176	membrane protein [*Rhodococcus* sp. JVH1]	35	9e-06	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA6-*orf49*	complement(42746..42877)	44	-	-	-	-
GMA6-*orf50*	complement(43215..43670)	152	unnamed protein product [*Synechococcus* phage S-CBS2]	37	8e-19	Recombination endonuclease (pfam02945)
GMA6-*orf51*	complement(43667..44602)	312	helicase DnaB [*Caldicellulosiruptor kronotskyensis*]	27	1e-23	DNA Primase (COG0358)
GMA6-*orf52*	complement(44611..46011)	467	DNA helicase [*Acidothermus cellulolyticus*]	34	6e-28	Replicative helicase (COG03050)
GMA6-*orf53*	complement(46243..46431)	63	-	-	-	-
GMA6-*orf54*	complement(46511..46978)	156	-	-	-	-
GMA6-*orf55*	48033..48371	113	-	-	-	-
GMA6-*orf56*	48391..48603	71	-	-	-	-
GMA6-*orf57*	48596..48976	127	hypothetical protein [*Mycobacterium genavense*]	31	2e-06	Unknown (pfam05305)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA6-*orf58*	48979..49254	92	-	-	-	-
GMA6-*orf59*	49254..49418	55	-	-	-	-
GMA6-*orf60*	49469..49927	153	-	-	-	-
GMA6-*orf61*	49979..51178	400	ATPase AAA [*Thioalkalivibrio thiocyanodenitrificans*]	33	5e-22	AAA protein (pfam07728)
GMA6-*orf62*	51221..53134	638	von Willebrand factor A [*Pelobacter propionicus*]	21	7e-04	Von Willlebrand factor/Cobalmin biosybthesis (pfam06213;pfam13519)
GMA6-*orf63*	53274..53819	182	-	-	-	-
GMA6-*orf64*	53816..54250	145	-	-	-	-
GMA6-*orf65*	54263..54556	98	hypothetical protein CRB1_33 [*Mycobacterium* phage CRB1]	42	4e-08	-
GMA6-	54553..54771	73	hypothetical protein PBI_RHYNO_66	45	9e-06	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf66			[*Mycobacterium* phage RhynO]			
GMA6-*orf67*	55029..55838	270	-	-	-	-
GMA6-*orf68*	55843..56382	180	-	-	-	-
GMA6-*orf69*	56385..56627	81	-	-	-	HTH DNA binding (pfam13411)
GMA6-*orf70*	56628..56978	117	-	-	-	-
GMA6-*orf71*	57069..57743	225	-	-	-	-
GMA6-*orf72*	57807..58559	251	-	-	-	-
GMA6-*orf73*	58715..58966	84	-	-	-	-
GMA6-*orf74*	59155..59451	99	-	-	-	-
GMA6-	59453..59935	161	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf75						
GMA6-*orf76*	59922..60326	135	-	-	-	-
GMA6-*orf77*	60323..60514	64	-	-	-	-
GMA6-*orf78*	60511..60807	99	-	-	-	-
GMA6-*orf79*	60811..61245		hypothetical protein [Streptomyces albus]	41	1e-20	Endonuclease (pfam13392)
GMA6-*orf80*	61160..64054	965	DNA polymerase III alpha subunit [*Halanaerobium saccharolyticum*]	32	1e-117	DNA polymerase III (COG0587)
GMA6-*orf81*	64168..64497	110	HNH domain protein [*Mycobacterium* phage Goku]	49	6e-23	Endonuclease (pfam13392)
GMA6-*orf82*	64501..64749	83	-	-	-	-
GMA6-*orf83*	64766..64912	49	-	-	-	-
GMA6-	64909..65130	74	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf84						
GMA6-*orf85*	65127..65333	69	-	-	-	-
GMA6-*orf86*	65333..65557	75	-	-	-	-
GMA6-*orf87*	65491..65952	154	-	-	-	-
GMA6-*orf88*	65953..67116	388	recombinase RecA [*Hirschia maritima*]	38	5e-60	Recombinase (pfam00154)
GMA6-*orf89*	67162..67716	185	hypothetical protein [Sphingobium chungbukense]	44	6e-23	Endonuclease (pfam13392)
GMA6-*orf90*	67738..67881	48	-	-	-	-
GMA6-*orf91*	67874..68185	104	-	-	-	-
GMA6-*orf92*	68182..68454	91	-	-	-	-
GMA6-	68445..69272	276	hypothetical protein TCA2_4616 [*Paenibacillus*	29	9e-08	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf93			sp. TCA20]			
GMA6-*orf94*	69278..69442	55	-	-	-	-
GMA6-*orf95*	69442..69879	146	MedDCM-OCT-S33-C31-cds10 [*Candidatus* Actinomarina minuta]	42	4e-17	-
GMA6-*orf96*	70127..70558	144	-	-	-	-
GMA6-*orf97*	70563..71153	197	-	-	-	Holliday junction resolvase (PRK00039)
GMA6-*orf98*	71158..72720	521	hypothetical protein [*Microbacterium* sp. UCD-TDU]	47	2e-34	Nuclease (pfam02195; pfam14386)
GMA6-*orf99*	72859..73206	116	-	-	-	valyl-tRNA synthetase (PRK14900)
GMA6-*orf100*	73320..73802	161	-	-	-	Metalloenzyme protein (cd08070)
GMA6-	73830..74732	301	hypothetical protein [*Streptomyces sulphureus*]	38	1e-35	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf101						
GMA6-*orf102*	74814..75557	248	-	-	-	-
GMA6-*orf103*	75571..76686	372	hypothetical protein [*Aeromicrobium marinum*]	42	8e-41	Phage-related tail fibre protein (COG5301)
GMA6-*orf104*	76794..78356	521	-	-	-	-
GMA6-*orf105*	78755..78997	81	-	-	-	-
GMA6-*orf106*	79073..79291	73	-	-	-	-
GMA6-*orf107*	79682..80032	117	-	-	-	-
GMA6-*orf108*	80146..81057	304	-	-	-	-
GMA6-*orf109*	81475..81708	78	-	-	-	-
GMA6-	81714..81839	42	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf110						
GMA6-*orf111*	81832..82194	121	gp58 [*Mycobacterium* phage Pipefish]	54	3e-28	-
GMA6-*orf112*	82328..82540	71	-	-	-	-
GMA6-*orf113*	82537..82827	97	putative regulator [*Tsukamurella* phage TPA2]	61	1e-06	-
GMA6-*orf114*	82827..82988	54	-	-	-	-
GMA6-*orf115*	83005..83148	48	-	-	-	-
GMA7-*orf1*	42..1499	486	putative small terminase [*Gordonia* phage GTE7]	99	0.0	Putative smalll terminase subunit
GMA7-*orf2*	1492..3459	656	terminase large subunit [*Gordonia* phage GTE7]	100	0.0	Putative large terminase subunit
GMA7-*orf3*	3574..5289	572	hypothetical protein GTE7_gp003 [*Gordonia* phage GTE7]	99	0.0	Unknown (pfam06074)
GMA7-*orf4*	5276..6424	383	hypothetical protein GTE7_gp004 [*Gordonia*	63	6e-157	YadA-like, left handed beta roll protein

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
			phage GTE7]			(cd12820)
GMA7-*orf5*	6434..6661	76	hypothetical protein GTE7_gp005 [*Gordonia* phage GTE7]	48	9.e-09	-
GMA7-*orf6*	6743..6994	84	hypothetical protein GTE7_gp006 [*Gordonia* phage GTE7]	99	5e-55	-
GMA7-*orf7*	7009..8187	393	hypothetical protein GTE7_gp007 [*Gordonia* phage GTE7]	99	0.0	-
GMA7-*orf8*	8221..8700	160	hypothetical protein GTE7_gp008 [*Gordonia* phage GTE7]	98	3e-110	-
GMA7-*orf9*	8712..9914	401	putative major capsid protein [*Gordonia* phage GTE7]	99	0.0	Phage major capsid protein (pfam03864)
GMA7-*orf10*	9931..10125	65	hypothetical protein GTE7_gp010 [*Gordonia* phage GTE7]	100	6e-36	-
GMA7-*orf11*	10200..10454	85	hypothetical protein GTE7_gp011 [*Gordonia* phage GTE7]	96	7e-35	-
GMA7-*orf12*	10464..10979	172	hypothetical protein GTE7_gp012 [*Gordonia* phage GTE7]	100	2e-121	-
GMA7-	11072..11467	132	hypothetical protein GTE7_gp013 [*Gordonia*	98	8e-86	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf13			phage GTE7]			
GMA7-*orf14*	11464..11949	162	hypothetical protein GTE7_gp014 [*Gordonia* phage GTE7]	99	3e-113	-
GMA7-*orf15*	11962..12267	102	hypothetical protein GTE7_gp015 [*Gordonia* phage GTE7]	99	3e-63	-
GMA7-*orf16*	12271..13290	340	hypothetical protein GTE7_gp016 [*Gordonia* phage GTE7]	99	0.0	Major tail protein
GMA7-*orf17*	13487..14077	197	putative tail assembly protein [*Gordonia* phage GTE7]	98	2e-137	Putative tail assembly protein
GMA7-*orf18*	14059..14337	93	hypothetical protein GTE7_gp018 [*Gordonia* phage GTE7]	97	5e-39	Putative tail assembly protein translated by conserved programmed translational frameshift
GMA7-*orf19*	14475..23615	3047	phage tape measure protein [*Gordonia* phage GTE7]	97	0.0	Tape measure protein (pfam10145; COG1196; COG5412; pfam01464)
GMA7-*orf20*	23615..26158	848	hypothetical protein GTE7_gp020 [*Gordonia* phage GTE7]	99	0.0	-
GMA7-	26160..27899	580	hypothetical protein GTE7_gp021 [*Gordonia*	99	0.0	Tail protein (pfam13550)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf21			phage GTE7]			
GMA7-*orf22*	27899..30667	923	hypothetical protein GTE7_gp022 [*Gordonia* phage GTE7]	98	0.0	-
GMA7-*orf23*	30667..31671	335	hypothetical protein GTE7_gp023 [*Gordonia* phage GTE7]	96	0.0	-
GMA7-*orf24*	31671..31964	98	hypothetical protein GTE7_gp024 [*Gordonia* phage GTE7]	98	2e-61	-
GMA7-*orf25*	31966..32343	126	hypothetical protein GTE7_gp025 [*Gordonia* phage GTE7]	100	2e-81	-
GMA7-*orf26*	32429..32848	140	hypothetical protein GTE7_gp026 [*Gordonia* phage GTE7]	100	6e-95	-
GMA7-*orf27*	32848..33171	108	hypothetical protein GTE7_gp027 [*Gordonia* phage GTE7]	97	1e-69	-
GMA7-*orf28*	33168..33803	212	lysozyme [*Gordonia* phage GTE7]	99	3e-155	Lysin (pfam01510)
GMA7-*orf29*	33800..34627	276	putative peptidase [*Gordonia* phage GTE7]	99	0.0	Lysin/Peptidase (pfam01551)
GMA7-	34824..35057	78	-	-	-	Putative holin

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf30						
GMA7-*orf31*	35045..35515	157	hypothetical protein GTE7_gp030 [*Gordonia* phage GTE7]	98	5e-108	-
GMA7-*orf32*	35508..35840	111	hypothetical protein GTE7_gp031 [*Gordonia* phage GTE7]	97	1e-67	-
GMA7-*orf33*	complement(35837..36241)	135	hypothetical protein GTE7_gp032 [*Gordonia* phage GTE7]	99	2e-88	-
GMA7-*orf34*	complement(36225..36341)	39	-	-	-	-
GMA7-*orf35*	complement(36338..37540)	401	hypothetical protein GTE7_gp033 [*Gordonia* phage GTE7]	99	0.0	Nuclease (pfam12705)
GMA7-*orf36*	complement(37757..38170)	138	hypothetical protein GTE7_gp034 [*Gordonia* phage GTE7]	100	8e-84	-
GMA7-*orf37*	complement(38252..38365)	38	hypothetical protein GTE7_gp035 [*Gordonia* phage GTE7]	100	3e-16	-
GMA7-*orf38*	complement(38788..39507)	240	DNA methylase [*Gordonia* phage GTE7]	99	1e-174	DNA methylase (pfam01555)
GMA7-	complement(39504..397	89	hypothetical protein GTE7_gp038 [*Gordonia*	97	2e-48	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
orf39	70)		phage GTE7]			
GMA7-*orf40*	complement(39767..40000)	78	hypothetical protein GTE7_gp039 [*Gordonia* phage GTE7]	99	-	-
GMA7-*orf41*	complement(39997..40896)	300	lysinB protein [*Gordonia* phage GTE7]	100	0.0	Lysin/cutinase (pfam01083)
GMA7-*orf42*	complement(40898..41383)	162	hypothetical protein GTE7_gp041 [*Gordonia* phage GTE7]	99	4e-113	Nucleoside Triphosphate Pyrophosphohydrolase (cd11542)
GMA7-*orf43*	complement(41380..41505)	42	hypothetical protein GTE7_gp042 [*Gordonia* phage GTE7]	100	3e-20	-
GMA7-*orf44*	complement(41535..42110)	192	hypothetical protein GTE7_gp043 [*Gordonia* phage GTE7]	100	4e-140	-
GMA7-*orf45*	complement(42103..42255)	51	hypothetical protein GTE7_gp044 [*Gordonia* phage GTE7]	98	3e-26	-
GMA7-*orf46*	complement(42269..42634)	122	hypothetical protein GTE7_gp045 [*Gordonia* phage GTE7]	99	9e-82	Unknown (pfam14359)
GMA7-*orf47*	complement(42624..43763)	380	DNA polymerase III beta subunit [*Gordonia* phage G TE7]	99	0.0	DNA poymearse III beta clamp (COG0592)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA7-*orf48*	complement(43964..44584)	207	exonuclease [*Gordonia* phage GTE7]	100	6e-149	Exonuclease (pfam00929)
GMA7-*orf49*	complement(44581..44766)	62	DNA binding protein [*Gordonia* phage GTE7]	98	3e-35	HTH DNA binding (pfam12728)
GMA7-*orf50*	complement(45045..45296)	84	hypothetical protein GTE7_gp049 [*Gordonia* phage GTE7]	100	4e-52	-
GMA7-*orf51*	complement(45296..46927)	544	helicase [*Gordonia* phage GTE7]	99	0.0	Helicase (COG0553)
GMA7-*orf52*	complement(46931..47533)	201	hypothetical protein GTE7_gp051 [*Gordonia* phage GTE7]	92	1e-130	-
GMA7-*orf53*	complement(47654..48700)	349	hypothetical protein GTE7_gp052 [*Gordonia* phage GTE7]	99	0.0	-
GMA7-*orf54*	complement(48834..49136)	101	hypothetical protein GTE7_gp053 [*Gordonia* phage GTE7]	100	2e-66	-
GMA7-*orf55*	complement(49133..49471)	113	hypothetical protein GTE7_gp054 [*Gordonia* phage GTE7]	100	9e-78	-
GMA7-*orf56*	complement(49477..49665)	63	hypothetical protein GTE7_gp055 [*Gordonia* phage GTE7]	97	3e-33	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]				% identity [c]	E_0 value [d]	Putative function [e]
GMA7-*orf57*	complement(49665..50288)	208	hypothetical phage GTE7]	protein	GTE7_gp056	[*Gordonia*	99	1e-151	-
GMA7-*orf58*	complement(50297..50746)	150	hypothetical phage GTE7]	protein	GTE7_gp057	[*Gordonia*	96	1e-101	-
GMA7-*orf59*	complement(50784..51029)	82	hypothetical phage GTE7]	protein	GTE7_gp058	[*Gordonia*	100	8e-53	-
GMA7-*orf60*	complement(51045..51902)	286	hypothetical phage GTE7]	protein	GTE7_gp059	[*Gordonia*	88	0.0	-
GMA7-*orf61*	complement(51895..52239)	115	hypothetical phage GTE7]	protein	GTE7_gp060	[*Gordonia*	99	1e-73	-
GMA7-*orf62*	complement(52243..52506)	88	hypothetical phage GTE7]	protein	GTE7_gp061	[*Gordonia*	98	6e-53	-
GMA7-*orf63*	complement(52668..53402)	245	hypothetical phage GTE7]	protein	GTE7_gp062	[*Gordonia*	97	1e-177	-
GMA7-*orf64*	complement(53399..53545)	49	hypothetical phage GTE7]	protein	GTE7_gp063	[*Gordonia*	98	3e-22	-
GMA7-*orf65*	complement(53545..53802)	86	hypothetical phage GTE7]	protein	GTE7_gp064	[*Gordonia*	99	2e-55	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA7-*orf66*	complement(53977..54399)	141	hypothetical protein GTE7_gp065 [*Gordonia* phage GTE7]	99	2e-94	-
GMA7-*orf67*	complement(54513..55064)	184	hypothetical protein GTE7_gp066 [*Gordonia* phage GTE7]	99	2e-131	-
GMA7-*orf68*	complement(55074..55277)	68	hypothetical protein GTE7_gp067 [*Gordonia* phage GTE7]	99	6e-39	-
GMA7-*orf69*	complement(55489..55728)	80	hypothetical protein GTE7_gp068 [*Gordonia* phage GTE7]	99	5e-50	-
GMA7-*orf70*	complement(55725..55952)	76	hypothetical protein GTE7_gp069 [*Gordonia* phage GTE7]	99	2e-44	-
GMA7-*orf71*	complement(55949..57772)	608	hypothetical protein GTE7_gp071 [*Gordonia* phage GTE7]	97	0.0	Cobalmin biosynthesis von willebrand factor (pfam11775)
GMA7-*orf72*	complement(57807..59429)	541	ATPase family protein [*Gordonia* phage GTE7]	98	0.0	AAA protein (pfam07728)
GMA7-*orf73*	complement(59511..60140)	210	hypothetical protein GTE7_gp073 [*Gordonia* phage GTE7]	98	6e-151	-
GMA7-*orf74*	complement(60152..60526)	125	hypothetical protein GTE7_gp075 [*Gordonia* phage GTE7]	100	8e-86	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]				% identity [c]	E_0 value [d]	Putative function [e]
GMA7-*orf75*	complement(60584..60712)	43	hypothetical phage GTE7]	protein	GTE7_gp076	[*Gordonia*	95	1e-20	-
GMA7-*orf76*	complement(60699..61589)	297	hypothetical phage GTE7]	protein	GTE7_gp077	[*Gordonia*	97	0.0	-
GMA7-*orf77*	complement(61570..61806)	79	hypothetical phage GTE7]	protein	GTE7_gp078	[*Gordonia*	97	7e-48	-
GMA7-*orf78*	complement(61803..61985)	61	hypothetical phage GTE7]	protein	GTE7_gp079	[*Gordonia*	98	2e-33	-
GMA7-*orf79*	complement(61982..62356)	125	hypothetical phage GTE7]	protein	GTE7_gp080	[*Gordonia*	100	2e-85	-
GMA7-*orf80*	complement(62617..62874)	86	hypothetical phage GTE7]	protein	GTE7_gp083	[*Gordonia*	95	2e-51	-
GMA7-*orf81*	complement(63157..64095)	313	hypothetical phage GTE7]	protein	GTE7_gp084	[*Gordonia*	98	0.0	-
GMA7-*orf82*	complement(64219..64500)	94	hypothetical phage GTE7]	protein	GTE7_gp085	[*Gordonia*	94	2e-42	-
GMA7-*orf83*	complement(64497..64619)	41	hypothetical phage GTE7]	protein	GTE7_gp086	[*Gordonia*	98	7e-18	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]				% identity [c]	E_0 value [d]	Putative function [e]
GMA7-*orf84*	complement(64592..64843)	84	hypothetical phage GTE7]	protein	GTE7_gp087	[*Gordonia*	100	4e-53	-
GMA7-*orf85*	complement(64858..65061)	68	hypothetical phage GTE7]	protein	GTE7_gp088	[*Gordonia*	100	4e-27	-
GMA7-*orf86*	complement(65115..65279)	55	hypothetical phage GTE7]	protein	GTE7_gp089	[*Gordonia*	53	1e-08	-
GMA7-*orf87*	complement(65276..65578)	101	hypothetical phage GTE7]	protein	GTE7_gp090	[*Gordonia*	40	4e-14	-
GMA7-*orf88*	complement(65690..65980)	97	hypothetical phage GTE7]	protein	GTE7_gp092	[*Gordonia*	34	2e-05	-
GMA7-*orf89*	complement(66094..66327)	78	-				-	-	-
GMA7-*orf90*	complement(66414..66899)	162	hypothetical phage GTE7]	protein	GTE7_gp091	[*Gordonia*	59	8e-48	-
GMA7-*orf91*	complement(67218..67400)	61	-				-	-	-
GMA7-*orf92*	complement(67467..67901)	145	-				-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GMA7-*orf93*	complement(68439..68672)	78	hypothetical protein GTE7_gp095 [*Gordonia* phage GTE7]	96	3e-44	-
GMA7-*orf94*	complement(68674..69123)	150	hypothetical protein GTE7_gp096 [*Gordonia* phage GTE7]	99	8e-100	-
GMA7-*orf95*	complement(69120..69380)	87	hypothetical protein GTE7_gp097 [*Gordonia* phage GTE7]	97	2e-53	-
GMA7-*orf96*	complement(69377..71620)	748	putative primase [*Gordonia* phage GTE7]	98	0.0	Putative primase (pfam13148)
GMA7-*orf97*	complement(71617..71850)	78	hypothetical protein GTE7_gp099 [*Gordonia* phage GTE7]	95	9e-46	-
GMA7-*orf98*	complement(71825..71938)	38	hypothetical protein GTE7_gp100 [*Gordonia* phage GTE7]	95	4.e-07	-
GMA7-*orf99*	complement(71935..72309)	125	hypothetical protein GTE7_gp101 [*Gordonia* phage GTE7]	99	7e-84	-
GMA7-*orf100*	complement(72341..72520)	60	hypothetical protein GTE7_gp102 [*Gordonia* phage GTE7]	100	2e-36	-
GMA7-*orf101*	complement(72517..73353)	279	hypothetical protein GTE7_gp103 [*Gordonia* phage GTE7]	97	0.0	Unknown (COG4951)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE6-*orf1*	65..625	187	-	-	-	Putative small terminase subunit
GTE6-*orf2*	622..2220	533	hypothetical protein [*Streptomyces sulphureus*]	40	5e-108	Large terminase subunit (pfam03237)
GTE6-*orf3*	2277..2726	150	-	-	-	-
GTE6-*orf4*	2723..2923	67	-	-	-	-
GTE6-*orf5*	2923..3546	208	-	-	-	-
GTE6-*orf6*	3543..3758	72	-	-	-	-
GTE6-*orf7*	3774..3965	64	-	-	-	-
GTE6-*orf8*	3969..4424	152	-	-	-	-
GTE6-*orf9*	4474..4629	52	-	-	-	-
GTE6-*orf10*	4626..4880	85	unnamed protein product [*Gordonia* phage GTE5]	52	2e-08	-
GTE6-*orf11*	4914..5237	108	-	-	-	-
GTE6-*orf12*	5362..5802	147	membrane protein [*Rhodococcus equi* 103S]	70	1e-38	Host cell surface-exposed lipoprotein (pfam07553)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE6-*orf13*	5949..6287	113	hypothetical protein [*Streptomyces violaceusniger*]	37	2e-10	Unknown (pfam07098)
GTE6-*orf14*	6302..7978	559	hypothetical protein [*Salinispora pacifica*]	30	6e-47	Phage structural protein
GTE6-*orf15*	7975..8361	129	-	-	-	-
GTE6-*orf16*	8361..10763	801	capsid maturation protease [*Mycobacterium* phage Bernal13]	43	3e-20	Head morphogenesis (pfam04233)
GTE6-*orf17*	10760..10981	74	-	-	-	-
GTE6-*orf18*	11027..13879	951	gp12 [*Rhodococcus* phage ReqiPine5]	56	1e-38	RNA ligase (pfam13563)
GTE6-*orf19*	14009..14155	49	-	-	-	-
GTE6-*orf20*	14261..16150	630	hypothetical protein [*Streptomyces sulphureus*]	33	6e-57	Phage structural protein
GTE6-*orf21*	16242..17072	277	Epstein-Barr nuclear antigen 1 [*Saccharomonospora* phage PIS 136]	35	8e-13	Putative Eppstein-Barr nuclear antigen 1
GTE6-*orf22*	17086..17577	164	-	-	-	-
GTE6-*orf23*	17681..18376	232	hypothetical protein [*Salinispora tropica*]	39	7e-44	Phage structural protein
GTE6-*orf24*	18480..19277	266	hypothetical protein [*Mycobacterium abscessus*]	35	5e-30	-
GTE6-*orf25*	19274..19837	188	hypothetical protein [*Salinispora tropica*]	28	3e-04	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE6-*orf26*	19842..20135	98	-	-	-	-
GTE6-*orf27*	20119..20550	144	hypothetical protein [*Streptomyces sulphureus*]	45	6e-18	Putative tail component (pfam04883)
GTE6-*orf28*	20594..21208	205	-	-	-	-
GTE6-*orf29*	21222..21899	226	-	-	-	-
GTE6-*orf30*	21946..27066	1707	hypothetical protein [*Gordonia* sp. KTR9]	33	3e-111	Tape measure protein (pfam06737; COG5280)
GTE6-*orf31*	27069..28046	326	tail protein [*Mycobacterium intracellulare*]	27	1e-11	Putative tail protein
GTE6-*orf32*	28048..29667	540	hypothetical protein [*Mycobacterium abscessus*]	30	2e-73	-
GTE6-*orf33*	29712..30401	230	-	-	-	-
GTE6-*orf34*	30398..31678	427	hypothetical protein ISGA_1789 [*Gordonia* sp. NB4-1Y]	51	4e-91	-
GTE6-*orf35*	31697..32035	113	-	-	-	-
GTE6-*orf36*	32032..33609	526	hypothetical protein [*Rhodococcus equi*]	38	4e-40	-
GTE6-*orf37*	33622..34326	235	hypothetical protein [*Mycobacterium thermoresistibile*]	33	2e-20	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE6-*orf38*	34411..36045	545	hypothetical protein [*Gordonia rhizosphera*]	62	6e-112	Lysin (pfam01510)
GTE6-*orf39*	36058..36567	170	hypothetical protein [*Rhodococcus equi*]	42	8e-18	Putative holin
GTE6-*orf40*	36569..37060	164	unnamed protein product [*Gordonia* phage GRU1]	60	3e-48	-
GTE6-*orf41*	37057..37491	145	hypothetical protein [*Gordonia sihwensis*]	37	2e-22	-
GTE6-*orf42*	37596..38033	146	-	-	-	-
GTE6-*orf43*	38035..38283	83	hypothetical protein [*Rhodococcus opacus*]	45	2e-08	-
GTE6-*orf44*	38280..38465	62	hypothetical protein GTE7_gp102 [*Gordonia* phage GTE7]	41	6e-04	-
GTE6-*orf45*	38462..39616	385	hypothetical protein [*Mycobacterium abscessus*]	35	3e-44	Recombinase (pfam09588)
GTE6-*orf46*	39613..39906	98	-	-	-	-
GTE6-*orf47*	39899..40960	354	recombinase RecT [*Mycobacterium avium*]	47	2e-93	Recombinase (pfam03837)
GTE6-*orf48*	40982..41434	151	-	-	-	-
GTE6-*orf49*	41493..41639	49	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE6-*orf50*	41632..41817	65	hypothetical protein [*Allofustis seminis*]	35	1e-04	HTH DNA Binding (pfam12728)
GTE6-*orf51*	41823..42017	65	hypothetical protein [*Allofustis seminis*]	35	1e-04	HTH DNA Binding (pfam12728)
GTE6-*orf52*	42014..42178	55	-	-	-	-
GTE6-*orf53*	42181..43176	332	unnamed protein product [*Rhodococcus* phage REQ3]	45	5e-18	-
GTE6-*orf54*	43173..43394	74	hypothetical protein 32HC_68 [*Mycobacterium* phage 32HC]	39	9e-04	-
GTE6-*orf55*	43476..43847	124	-	-	-	-
GTE6-*orf56*	43844..44005	54	-	-	-	-
GTE6-*orf57*	44002..44412	137	-	-	-	Putative protein serine/threonine phosphatase (PRK14559)
GTE6-*orf58*	44409..45188	260	putative cutinase [*Gordonia* phage GTE2]	36	1e-36	Putative cutinase/lysin
GTE6-*orf59*	45185..45565	127	-	-	-	-
GTE6-*orf60*	45562..46041	160	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE6-*orf61*	46038..46202	55	-	-	-	-
GTE6-*orf62*	46212..46838	209	RuvC [*Mycobacterium* phage Bernardo]	33	2e-14	Putative Holliday Junction Resolvase
GTE6-*orf63*	46835..47095	87	-	-	-	-
GTE6-*orf64*	47088..47441	118	-	-	-	-
GTE6-*orf65*	47434..47742	103	-	-	-	-
GTE6-*orf66*	47732..48301	190	-	-	-	-
GTE6-*orf67*	48354..49775	474	-	-	-	-
GTE6-*orf68*	49772..50149	126	-	-	-	-
GTE6-*orf69*	50250..50750	167	-	-	-	-
GTE6-*orf70*	50747..51001	85	-	-	-	-
GTE6-*orf71*	51011..51286	92	hypothetical protein E3_0905 [*Rhodococcus* phage E3]	47	7e-12	-
GTE6-*orf72*	51283..51579	99	-	-	-	-
GTE6-*orf73*	51576..52175	200	-	-	-	-
GTE6-*orf74*	52172..52378	69	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE6-*orf75*	52375..52836	154	gp87 [*Mycobacterium* phage Anaya]	47	1e-33	Polynucleotide kinase (PHA02530)
GTE6-*orf76*	52833..53069	79	-	-	-	-
GTE6-*orf77*	53066..53278	71	-	-	-	-
GTE6-*orf78*	53343..53906	188	-	-	-	-
GTE6-*orf79*	53903..54592	230	DNA polymerase III subunit epsilon [*Rhodococcus* sp. P27]	41	3e-48	DNA Polymerase III epsilon subunit (pfam00929)
GTE6-*orf80*	54589..54744	52	-	-	-	-
GTE6-*orf81*	54741..54896	52	-	-	-	-
GTE6-*orf82*	54893..55081	63	-	-	-	-
GTE6-*orf83*	55078..55488	137	-	-	-	-
GTE6-*orf84*	55485..55895	137	-	-	-	-
GTE6-*orf85*	55892..56542	217	-	-	-	-
GTE6-*orf86*	56599..56850	84	hypothetical protein [*Streptomyces* sp. AW19M42]	36	7e-05	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE8-*orf1*	51..518	156	unnamed protein product [*Gordonia* phage GRU1]	51	3e-40	-
GTE8-*orf2*	515..745	77	-	-	-	-
GTE8-*orf3*	742..1041	100	unnamed protein product [*Gordonia* phage GRU1]	71	8e-41	-
GTE8-*orf4*	1049..1438	130	unnamed protein product [*Gordonia* phage GRU1]	35	9e-11	-
GTE8-*orf5*	1435..1704	90	unnamed protein product [*Gordonia* phage GRU1]	50	6e-18	-
GTE8-*orf6*	1966..2262	99	unnamed protein product [*Gordonia* phage GTE5]	59	1e-17	-
GTE8-*orf7*	2259..2990	244	unnamed protein product [*Gordonia* phage GRU1]	53	4e-74	-
GTE8-*orf8*	3094..3345	84	unnamed protein product [*Gordonia* phage GTE5]	89	2e-45	-
GTE8-*orf9*	3345..5333	663	hypothetical protein [*Rhodococcus* sp. p52]	58	0.0	-
GTE8-*orf10*	5351..5530	60	unnamed protein product [*Gordonia* phage GTE5]	91	1e-19	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE8-*orf11*	5527..5994	156	unnamed protein product [*Gordonia* phage GTE5]	86	4e-93	-
GTE8-*orf12*	5991..6422	144	terS gene product [*Gordonia* phage GTE5]	88	5e-88	Putative small terminase subunit (pfam01844)
GTE8-*orf13*	6443..8170	576	terL gene product [*Gordonia* phage GTE5]	81	0.0	Large terminase subunit (pfam03354)
GTE8-*orf14*	8211..9800	530	unnamed protein product [*Gordonia* phage GRU1]	75	0.0	Portal protein (pfam05133)
GTE8-*orf15*	9797..11185	463	unnamed protein product [*Gordonia* phage GTE5]	72	0.0	-
GTE8-*orf16*	11182..11877	232	unnamed protein product [*Gordonia* phage GRU1]	57	6e75	-
GTE8-*orf17*	11895..12320	142	unnamed protein product [*Gordonia* phage GRU1]	69	1e-95	Head decorator protein (pfam02924)
GTE8-*orf18*	12360..13430	357	unnamed protein product [*Gordonia* phage GTE5]	74	0.0	Major capsid structural protein (pfam03864)
GTE8-*orf19*	13433..13885	151	unnamed protein product [*Gordonia* phage GRU1]	54	4e-40	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE8-*orf20*	13932..14387	152	unnamed protein product [*Gordonia* phage GTE5]	75	3e-78	-
GTE8-*orf21*	14387..14779	131	unnamed protein product [*Gordonia* phage GTE5]	76	9e-66	-
GTE8-*orf22*	14772..15137	122	unnamed protein product [*Gordonia* phage GTE5]	76	1e-52	-
GTE8-*orf23*	15134..15649	172	unnamed protein product [*Gordonia* phage GTE5]	81	4e-97	-
GTE8-*orf24*	15678..16361	228	unnamed protein product [*Gordonia* phage GRU1]	89	1e-146	Phage structural protein
GTE8-*orf25*	16446..16829	128	unnamed protein product [*Gordonia* phage GRU1]	53	4e-33	Major tail protein
GTE8-*orf26*	16862..17176	105	unnamed protein product [*Gordonia* phage GTE5]	56	1e-34	Putative tail assembly protein
GTE8-*orf27*	17158..17667	170	unnamed protein product [*Gordonia* phage GTE5]	66	4e-61	Putative tail assembly protein translated by conserved programmed translational frameshift
GTE8-*orf28*	17751..23138	179	unnamed protein product [*Gordonia* phage	74	0.0	Tape measure protein

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
		6	GRU1]			(COG5412; pfam06736)
GTE8-*orf29*	23138..24079	314	unnamed protein product [*Gordonia* phage GRU1]	78	4e-172	-
GTE8-*orf30*	24084..25697	538	unnamed protein product [*Gordonia* phage GRU1]	85	0.0	-
GTE8-*orf31*	25738..26865	376	unnamed protein product [*Gordonia* phage GTE5]	77	0.0	Phage structural protein
GTE8-*orf32*	26866..29232	789	unnamed protein product [*Gordonia* phage GRU1]	83	0.0	Carbohydrate binding domian (pfam02015)
GTE8-*orf33*	29229..30287	353	unnamed protein product [*Gordonia* phage GRU1]	69	7e-112	-
GTE8-*orf34*	30263..30565	101	unnamed protein product [*Gordonia* phage GRU1]	78	2e-49	-
GTE8-*orf35*	30569..30913	115	unnamed protein product [*Gordonia* phage GTE5]	78	1e-57	-
GTE8-*orf36*	30971..31603	211	unnamed protein product [*Gordonia* phage GRU1]	89	1e-137	Lysin - Peptidase (pfam13529)
GTE8-*orf37*	31603..32625	341	unnamed protein product [*Gordonia* phage	80	0.0	Lysin - Chitinase

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
			GRU1]			(pfam00182)
GTE8-*orf38*	32622..32933	104	unnamed protein product [*Gordonia* phage GTE5]	57	2e-33	Putative holin
GTE8-*orf39*	33002..33436	145	unnamed protein product [*Gordonia* phage GRU1]	61	1e-55	-
GTE8-*orf40*	33411..33839	143	unnamed protein product [*Gordonia* phage GTE5]	81	6e-77	Portal vertex protein (PHA02531)
GTE8-*orf41*	complement(33836..36385)	850	unnamed protein product [*Gordonia* phage GRU1]	71	0.0	Bifunctional primase/polymerase (pfam09250; COG3378)
GTE8-*orf42*	complement(36397..36801)	135	unnamed protein product [*Gordonia* phage GRU1]	44	6e-35	-
GTE8-*orf43*	complement(36758..37105)	116	-	-	-	-
GTE8-*orf44*	complement(37102..37374)	91	unnamed protein product [*Gordonia* phage GRU1]	68	5e-38	-
GTE8-*orf45*	complement(37371..37847)	159	unnamed protein product [*Gordonia* phage GTE5]	51	1e-40	DNA Binding (pfam12728)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE8-*orf46*	complement(37878..38609)	244	unnamed protein product [*Gordonia* phage GRU1]	56	2e-88	-
GTE8-*orf47*	complement(38606..38830)	75	unnamed protein product [*Gordonia* phage GRU1]	46	1e-10	-
GTE8-*orf48*	complement(38830..42138)	1103	unnamed protein product [*Gordonia* phage GTE5]	74	0.0	DNA Polymerase II alpha subunit (COG0587)
GTE8-*orf49*	complement(42135..42431)	99	-	-	-	-
GTE8-*orf50*	complement(42428..42589)	54	hypothetical protein [*Rhodococcus fascians*]	44	5e-06	-
GTE8-*orf51*	complement(42710..42940)	77	unnamed protein product [*Gordonia* phage GRU1]	71	7e-28	-
GTE8-*orf52*	complement(42947..43411)	155	unnamed protein product [*Gordonia* phage GRU1]	37	1e-17	-
GTE8-*orf53*	complement(43535..44425)	297	unnamed protein product [*Gordonia* phage GRU1]	52	2e-61	-
GTE8-*orf54*	complement(44458..45279)	274	unnamed protein product [*Gordonia* phage GTE5]	61	3e-99	AAAprotein (pfam13479)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE8-*orf55*	complement(45335..46534)	400	unnamed protein product [*Gordonia* phage GRU1]	71	0.0	-
GTE8-*orf56*	complement(46531..46842)	104	-	-	-	-
GTE8-*orf57*	complement(46880..47056)	59	unnamed protein product [*Gordonia* phage GTE5]	76	9e-09	-
GTE8-*orf58*	complement(47053..47388)	112	unnamed protein product [*Gordonia* phage GTE5]	66	1e-44	-
GTE8-*orf59*	complement(47385..47633)	83	-	-	-	-
GTE8-*orf60*	complement(47633..49561)	643	unnamed protein product [*Gordonia* phage GRU1]	76	0.0	Helicase (COG0553)
GTE8-*orf61*	complement(49612..50187)	192	unnamed protein product [*Gordonia* phage GTE5]	34	5e-12	-
GTE8-*orf62*	complement(50184..50369)	62	unnamed protein product [*Gordonia* phage GRU1]	49	6e-07	-
GTE8-*orf63*	complement(50412..50618)	69	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE8-*orf64*	complement(50618..50743)	42	unnamed protein product [*Gordonia* phage GRU1]	63	6e-10	-
GTE8-*orf65*	complement(50740..51213)	158	unnamed protein product [*Gordonia* phage GRU1]	83	5e-95	-
GTE8-*orf66*	complement(51213..51692)	160	-	-	-	Cytosolic phospholipase (cd7201)
GTE8-*orf67*	complement(51696..52721)	342	unnamed protein product [*Gordonia* phage GRU1]	65	2e-146	-
GTE8-*orf68*	53037..53228	64	-	-	-	-
GTE8-*orf69*	53292..54173	294	-	-	-	-
GTE8-*orf70*	54170..54565	132	unnamed protein product [*Gordonia* phage GRU1]	52	5e-34	-
GTE8-*orf71*	54601..54894	98	-	-	-	-
GTE8-*orf72*	54913..55671	253	-	-	-	-
GTE8-*orf73*	55760..56416	219	-	-	-	-
GTE8-*orf74*	56418..56585	56	unnamed protein product [*Gordonia* phage GRU1]	64	9e-14	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE8-*orf75*	56582..57067	162	gp85 [*Mycobacteriophage Astro*]	30	1e-06	-
GTE8-*orf76*	57064..57753	230	unnamed protein product [*Gordonia* phage GRU1]	47	6e-51	-
GTE8-*orf77*	57782..58288	169	-	-	-	-
GTE8-*orf78*	58285..58428	48	-	-	-	-
GTE8-*orf79*	58496..58765	90	unnamed protein product [*Gordonia* phage GTE5]	57	1e-12	-
GTE8-*orf80*	58798..59136	113	-	-	-	-
GTE8-*orf81*	59310..60392	361	hypothetical protein GOALK_093_00330 [Gordonia alkanivorans NBRC 16433]	37	9e-23	HNH endonuclease (pfam01844)
GTE8-*orf82*	60392..60949	186	unnamed protein product [*Gordonia* phage GRU1]	68	1e-72	-
GTE8-*orf83*	61537..61938	134	-	-	-	-
GTE8-*orf84*	61935..63851	639	unnamed protein product [*Gordonia* phage GRU1]	78	2e-07	-
GTE8-*orf85*	63971..64687	239	unnamed protein product [*Gordonia* phage GRU1]	69	7e-104	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GTE8-*orf86*	64780..64983	68	unnamed protein product [*Gordonia* phage GTE5]	75	4e-23	-
GTE8-*orf87*	64976..65326	117	-	-	-	-
GTE8-*orf88*	65323..65595	91	-	-	-	-
GTE8-*orf89*	65585..65725	47	-	-	-	-
GTE8-*orf90*	65718..65939	74	-	-	-	-
GTE8-*orf91*	66184..66642	153	unnamed protein product [*Gordonia* phage GRU1]	84	1e-90	-
GTE8-*orf92*	66639..66941	101	-	-	-	-
GTE8-*orf93*	66938..67309	124	unnamed protein product [*Gordonia* phage GTE5]	33	2e-07	-
GTE8-*orf94*	67306..67617	104	-	-	-	-
GRU3-*orf1*	65..502	146	hypothetical protein [*Corynebacterium diphtheriae*]	36	2e-09	Putative small terminase subunit
GRU3-*orf2*	495..1916	474	putative phage terminase protein [*Gordonia neofelifaecis*]	54	1e-133	Large terminase subunit (pfam03237)

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GRU3-*orf3*	1924..2109	62	hypothetical protein [*Gordonia neofelifaecis*]	48	4e-08	-
GRU3-*orf4*	2171..3247	359	hypothetical protein [*Gordonia neofelifaecis*]	66	7e-148	Portal structural protein (pfam04860)
GRU3-*orf5*	3244..4032	263	hypothetical protein [*Gordonia soli*]	51	2e-65	Lysin - D-alanyl-D-alanine carboxypeptidase (pfam13539)
GRU3-*orf6*	4133..6103	657	hypothetical protein [*Gordonia neofelifaecis*]	60	0.0	Caudovirales prohead protease structural protein (pfam04586)
GRU3-*orf7*	6107..6448	114	unnamed protein product [*Rhodococcus* phage RRH1]	45	1e-19	-
GRU3-*orf8*	6449..6787	113	hypothetical protein [*Gordonia neofelifaecis*]	58	5e-35	-
GRU3-*orf9*	6800..7261	154	hypothetical protein [*Gordonia neofelifaecis*]	75	6e-76	-
GRU3-*orf10*	7270..7608	113	hypothetical protein [*Gordonia neofelifaecis*]	51	2e-24	Phage protein (TIGR01725)
GRU3-*orf11*	7630..7923	98	hypothetical protein [*Gordonia neofelifaecis*]	68	1e-39	-
GRU3-*orf12*	8095..9960	622	TP901 family phage tail tape measure protein, putative [*Gordonia neofelifaecis*]	58	0.0	Tape measure protein

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GRU3-*orf13*	9957..11411	485	hypothetical protein [*Nocardia otitidiscaviarum*]	31	7e-41	-
GRU3-*orf14*	11426..12031	202	hypothetical protein [*Nocardia otitidiscaviarum*]	48	5e-53	-
GRU3-*orf15*	12090..12938	283	bacteriophage protein [*Mycobacterium thermoresistibile*]	42	2e-60	PE-PPE structural protein (pfam08237)
GRU3-*orf16*	complement(13012..13116)	35	-	-	-	-
GRU3-*orf17*	complement(13195..14088)	298	integrase [*Gordonia neofelifaecis*]	70	2e-124	Integrase (pfam00589)
GRU3-*orf18*	complement(14162..14485)	108	putative DNA-binding protein [*Gordonia neofelifaecis*]	54	7e-28	HTH DNA binding (pfam12844)
GRU3-*orf19*	complement(14648..14773)	42	-	-	-	-
GRU3-*orf20*	14761..14949	63	hypothetical protein [*Rhodococcus fascians*]	55	9e-08	HTH DNA binding (pfam12728)
GRU3-*orf21*	14946..15404	153	-	-	-	-
GRU3-*orf22*	15404..15646	81	hypothetical protein [*Gordonia neofelifaecis*]	38	6e-05	-
GRU3-*orf23*	15646..15870	75	-	-	-	-

ORF [a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E_0 value [d]	Putative function [e]
GRU3-*orf24*	15858..16127	90	-	-	-	-
GRU3-*orf25*	16472..17227	252	DNA polymerase III subunit epsilon [*Gordonia neofelifaecis*]	56	1e-155	DNA polymerase III subunit epsilon (COG0847)
GRU3-*orf26*	17220..17402	61	-	-	-	-

[a] ORFs were numbered consecutively; [b] The most closely related gene (only if named) and the name of the organism; [c] Percentage identity is based on the best match when a BLAST P analysis is performed; [d] The probability of obtaining a match by chance as determined by BLAST analysis and only values less than 10^{-4} were considered significant; [e] Predicted function is based on amino acid identity, conserved motifs, and gene location within functional modules.

Table 4.6 Putative tRNA detected in *Gordonia* phage genomes

Phage	No. tRNA	tRNA present	Coordinates	Size (bp)	G+C (mol %)
GMA2	16	tRNA-Thr(tgt)	Complement(57883..57955)	73	52.1
		tRNA-Glu(ttc)	Complement(58076..58153)	78	59.0
		tRNA-Ser(gct)	Complement(58157..58240)	84	58.3
		tRNA-Ser(tga)	Complement(58319..58403)	85	54.1
		tRNA-Lys(ctt)	Complement(58406..58478)	73	53.4
		tRNA-Lys(ttt)	Complement(58481..58552)	72	58.3
		tRNA-Glu(ctc)	Complement(58733..58806)	74	48.6
		tRNA-Leu(tag)	Complement(58894..58978)	85	55.3
		tRNA-Asp(gtc)	Complement(58993..59068)	76	57.9
		tRNA-Tyr(gta)	Complement(59077..59160)	84	58.3
		tRNA-Pro(tgg)	Complement(59301..59375)	75	49.3
		tRNA-Ile(gat)	Complement(59483..59557)	75	52.0
		tRNA-Met(cat)	Complement(59632..59706)	75	54.7
		tRNA-Asn(gtt)	Complement(59744..59816)	73	53.4
		tRNA-Gln(ttg)	Complement(59821..59895)	75	62.7
		tRNA-Trp(cca)	Complement(60083..60154)	72	51.4
GMA4	1	tRNA-Try(gta)	5132..5213	82	62.2
GMA7	1	tRNA-Asn (gtt)	13362..13435	74	47.3

Table 4.7 Palindromes detected in the genome sequences of nine *Gordonia* phages

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA2-P1	48	57783-57830	GAGCTGGAGATGATCTCAATCGTAGATGATAGAAATCATCTCCAGCTC
		57830-57783	GAGCTGGAGATGATTTCTATCATCTACGATTGAGATCATCTCCAGCTC
GMA2-P2	47	60886-60930	AGAGGATGCCGATAGACGCGACGATATCACGTCTATCGGC--CCTCT
		60930-60886	AGAGG--GCCGATAGACGTGATATCGTCGCGTCTATCGGCATCCTCT
GMA2-P3	43	62824-62866	TCACGTCTTCGGTCTAAAGTAAACCCTTAGACCGAAGACGAGA
		62866-62824	TCTCGTCTTCGGTCTAAGGGTTTACTTTAGACCGAAGACGTGA
GMA2-P4	39	50539-50577	AATCTACGTTGAACGAGCTTGTTTCGTTCAACGTAGATT
		50577-50539	AATCTACGTTGAACGAAACAAGCTCGTTCAACGTAGATT
GMA2-P5	32	47948-47979	CTCTACACGCGAGCCGAAGCTCGCGTGTAGAG
		47979-47948	CTCTACACGCGAGCTTCGGCTCGCGTGTAGAG
GMA2-P6	16	15010-15025	AGAATCACGTGATTCT
		15025-15010	AGAATCACGTGATTCT
GMA2-P7	16	22496-22511	CGTAGCCGCGGCTACG
		22511-22496	CGTAGCCGCGGCTACG
GMA3-P1	98	10371-10467	AAGAGACCACTTAGACAGGATGCAATCTAAAATGGCCGATGTATGGAGCAATTCGTATTTTGACGATTTTAGTG GGGA-CCTGTCTAAGTGGTCTTTT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		10467-10371	AAAAGACCACTTAGACAGGTCCCCA- CTAAAATCGTCAAAATACGAATTGCTCCATACATCGGCCATTTTAGATTGCATCCTGTCTAAGTGGTCTCTT
GMA3-P2	66	52315-52378	GCGATCCCCGATAATTGCTAGGCAC--TTGGGGCCTTAAAGGGGAATGACAATTATCGGGGATCGC
		52378-52315	GCGATCCCCGATAATTGTCATTCCCCTTTAAGGCCCCAA--GTGCCTAGCAATTATCGGGGATCGC
GMA3-P3	51	58726-58774	CCGAAGGCGCAGGCTTTGGCAGATTAAATCTGTTA--GCCTGCGCCTTCGG
		58774-58726	CCGAAGGCGCAGGCTAA--CAGATTTAATCTGCCAAAGCCTGCGCCTTCGG
GMA3-P4	50	33805-33854	TGCCAGTCCACGGTAAATCCCCTTTGCTGGGGGCATACCGTGGACTGGCA
		33854-33805	TGCCAGTCCACGGTATGCCCCCAGCAAAGGGGATTTACCGTGGACTGGCA
GMA3-P5	48	13807-13853	GGATTATCCCCATGTGCCAAG-TGCAGATTGGCACATGGGGATAATCC
		13853-13807	GGATTATCCCCATGTGCCAATCTGCAC-TTGGCACATGGGGATAATCC
GMA3-P6	45	36860-36904	CAAAAGGGCGTCGTACGGACATTTTACTCGTACGACGCCCTTTTG
		36904-36860	CAAAAGGGCGTCGTACGAGTAAAATGTCCGTACGACGCCCTTTTG
GMA3-P7	44	67284-67327	ACAAAAAATCCCCTACCGCGTGGATGCGGTAGGGGATTTTTTGT
		67327-67284	ACAAAAAATCCCCTACCGCATCCACGCGGTAGGGGATTTTTTGT
GMA3-P8	44	51946-51987	TAAGGCTAAGCCGGGAGAAATTAAT--CTCCCGGCTTAGCCTTA
		51987-51946	TAAGGCTAAGCCGGGAGA--TTAATTTCTCCCGGCTTAGCCTTA
GMA3-P9	42	27139-27180	ATCCTGGTCGGAGGAAATGTTAATGTTTCCTCCGACCAGGAT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		27180-27139	ATCCTGGTCGGAGGAAACATTAACATTTCCTCCGACCAGGAT
GMA3-P10	39	39354-39392	AAAAATACCCGGCACCATGAATTGGTGCCGGGTATTTTT
		39392-39354	AAAAATACCCGGCACCAATTCATGGTGCCGGGTATTTTT
GMA3-P11	36	71487-71522	ATGAATGCATATGCATCACTATGCATATGCATTCAT
		71522-71487	ATGAATGCATATGCATAGTGATGCATATGCATTCAT
GMA3-P12	36	68183-68218	GTCGACCAGCGAAATTTCCTGTTTTCGCTGGTCGAC
		68218-68183	GTCGACCAGCGAAAACAGGAAATTTCGCTGGTCGAC
GMA3-P13	26	72043-72068	GTACCGATAACACTTGTTATCGGTAC
		72068-72043	GTACCGATAACAAGTGTTATCGGTAC
GMA3-P14	26	72532-72557	AGAAAAAAATATATATATTTATTCCT
		72557-72532	AGGAATAAATATATATATTTTTTTCT
GMA3-P15	16	6982-6997	CTCGTTTATAAACGAG
		6997-6982	CTCGTTTATAAACGAG
GMA3-P16	16	20089-20104	TTGACAATATTGTCAA
		20104-20089	TTGACAATATTGTCAA
GMA4-P1	56	22342-22395	TGAACG--GAAATGCGCCCCAACCTCTTCGGAGGTTGGGGCGCATTTCTGCGTTCA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		22395-22342	TGAACGCAGAAATGCGCCCCAACCTCCGAAGAGGTTGGGGCGCATTTC--CGTTCA
GMA4-P2	46	7033-7075	CGCGTCGAACCA---GGTGCCCGCGGGCACCGCGTGGGTCGTCGCG
		7075-7033	CGCGACGACCCACGCGGTGCCCGCGGGCACC---TGGTTCGACGCG
GMA4-P3	22	8283-8304	CGCGGCGATCATGATCGCCGCG
		8304-8283	CGCGGCGATCATGATCGCCGCG
GMA4-P4	20	10274-10293	CGAACTCGTCGACGAGTTCG
		10293-10274	CGAACTCGTCGACGAGTTCG
GMA4-P5	16	14873-14888	TCGGCGGATCCGCCGA
		14888-14873	TCGGCGGATCCGCCGA
GMA4-P6	16	33433-33448	GCTGGCCGCGGCCAGC
		33448-33433	GCTGGCCGCGGCCAGC
GMA5-P1	42	8066-8107	CCGCCGCAGCGGCCCCGGCCGCCGCCGGGGCCGCTCAGGCGG
		8107-8066	CCGCCTGAGCGGCCCCGGCGGCGGCCGGGGCCGCTGCGGCGG
GMA5-P2	41	12951-12990	GACGCCCCGCGCCA-TCCTTGGGGGAGTGGCGCGGGGCGTC
		12990-12951	GACGCCCCGCGCCACTCCCCCAAGGA-TGGCGCGGGGCGTC
GMA5-P3	37	15912-15947	GCCGTTGAGGACCGCCGTGACG-CGATCCTCAACGGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		15947-15912	GCCGTTGAGGATCGC-GTCACGGCGGTCCTCAACGGC
GMA5-P4	32	14428-14459	CACGCCCCGTCACCATCCGGTGACGGGGCGTG
		14459-14428	CACGCCCCGTCACCGGATGGTGACGGGGCGTG
GMA5-P5	30	6072-6101	ACCCGGCGACCGTTGACGCGGTCGCCGGGT
		6101-6072	ACCCGGCGACCGCGTCAACGGTCGCCGGGT
GMA5-P6	16	16425-16440	CTCAGCGGCCGCTGAG
		16440-16425	CTCAGCGGCCGCTGAG
GMA5-P7	14	1507-1520	GTCGCGATCGCGAC
		1520-1507	GTCGCGATCGCGAC
GMA5-P8	14	3061-3074	GGACGACGTCGTCC
		3074-3061	GGACGACGTCGTCC
GMA5-P9	14	6311-6324	GCGACGTACGTCGC
		6324-6311	GCGACGTACGTCGC
GMA5-P10	14	6580-6593	AACGGCATGCCGTT
		6593-6580	AACGGCATGCCGTT
GMA5-P11	14	13965-13978	GCCGCGATCGCGGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		13978-13965	GCCGCGATCGCGGC
GMA6-P1	58	58978-59034	TGAGTAAAGCCGGGGTAGGTG-CGCGTGGGGATGCGTAATCTACTCCGGCTTTACTCA
		59034-58978	TGAGTAAAGCCGGAGTAGATTACGCATCCCCACGCGCAC-CTACCCCGGCTTTACTCA
GMA7-P1	90	62906-62992	TCAATAGATCAGGTGGCCGGTGGGAAATGCT--- AACGGTTACCAGCGAAAGCATTTCGATCAGTGCCCACCGGCCACCTGATATATTGA
		62992-62906	TCAATATATCAGGTGGCCGGTGGGCACTGATCGAAATGCTTTCGCTGGTAACCGTT--- AGCATTTCCCACCGGCCACCTGATCTATTGA
GMA7-P2	83	14335-14413	TAATACACATAAAGGGTAGTGCTTG-AATAGCACTACCCTTTATGTGTATT--- ATTGGCATCACTACCCTTTATGTGTATTA
		14413-14335	TAATACACATAAAGGGTAGTGATGCCAATAA---TACACATAAAGGGTAGTGCTATTCA- AGCACTACCCTTTATGTGTATTA
GMA7-P3	52	38198-38246	ACCAACGGGTAGCGTATCA---TGCCGACACATTCATACGCTACCCGTTGGT
		38246-38198	ACCAACGGGTAGCGTATGAATGTGTCGGCA---TGATACGCTACCCGTTGGT
GMA7-P4	51	14335-14385	TAATACACATAAAGGGTAGTGCTTGAATAGCACTACCCTTTATGTGTATTA
		14385-14335	TAATACACATAAAGGGTAGTGCTATTCAAGCACTACCCTTTATGTGTATTA
GMA7-P5	49	10130-10178	CAAGAAATGCCCCGCCTGGACTACCAAAGCCCAGTCGGGGCATTTCTTG
		10178-10130	CAAGAAATGCCCCGACTGGGCTTTGGTAGTCCAGGCGGGGCATTTCTTG
GMA7-P6	48	53858-53904	AAAAATCAGGCACGAGCAGTGATCCAT-ACTGCTCGTGCCTGATTGTT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		53904-53858	AACAATCAGGCACGAGCAGT-ATGGATCACTGCTCGTGCCTGATTTTT
GMA7-P7	47	64164-64210	TGCATTCGTTCGTGAACTAACTATAACTCGGTCCACGAACGAATGCA
		64210-64164	TGCATTCGTTCGTGGACCGAGTTATAGTTAGTTCACGAACGAATGCA
GMA7-P8	38	47553-47590	CAATTGAAGGGGCTGCATCGGTGCAGCCCCTTCAATTG
		47590-47553	CAATTGAAGGGGCTGCACCGATGCAGCCCCTTCAATTG
GMA7-P9	38	63072-63109	GTGCACCTCCCCCAACCCCGAGAATAGGGGAGGTGCAC
		63109-63072	GTGCACCTCCCCTATTCTCGGGGTTGGGGGAGGTGCAC
GMA7-P10	36	3460-3494	GCGAACTAGCGGTGTGCTAG-ACACCGCTAGTTCGC
		3494-3460	GCGAACTAGCGGTGT-CTAGCACACCGCTAGTTCGC
GMA7-P11	36	65620-65655	CACTAACTACATGGTACCACGCCCATGTAGTTAGTG
		65655-65620	CACTAACTACATGGGCGTGGTACCATGTAGTTAGTG
GMA7-P12	32	30692-30723	CCCCAGCACCGATCCCAGAATCGGTGCTGGGG
		30723-30692	CCCCAGCACCGATTCTGGGATCGGTGCTGGGG
GMA7-P13	30	236-264	AAACTTTC-ACGGATATCCGTGGAAAGCTT
		264-236	AAGCTTTCCACGGATATCCGT-GAAAGTTT
GMA7-P14	26	69409-69434	CGGACACCAGCTCGAGCTGTCGTCCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		69434-69409	CGGACGACAGCTCGAGCTGGTGTCCG
GMA7-P15	24	7474-7497	CTCTGGCGAAACGTTTCGGCAGAG
		7497-7474	CTCTGCCGAAACGTTTCGCCAGAG
GMA7-P16	16	24291-24306	GCCTTGTCGACAAGGC
		24306-24291	GCCTTGTCGACAAGGC
GMA7-P17	16	63264-63279	CACCCGAGCTCGGGTG
		63279-63264	CACCCGAGCTCGGGTG
GMA7-P18	16	66284-66299	CGGATGATATCATCCG
		66299-66284	CGGATGATATCATCCG
GRU3-P1	34	14234-14267	ACATGCCCCGTCACCATCCGGTGACGGGGCATGT
		14267-14234	ACATGCCCCGTCACCGGATGGTGACGGGGCATGT
GRU3-P2	14	9980-9993	CGTCGAGCTCGACG
		9993-9980	CGTCGAGCTCGACG
GRU3-P3	14	13976-13989	GCCGCTATAGCGGC
		13989-13976	GCCGCTATAGCGGC
GTE6-P1	42	18413-18454	CGAATGCCGCACGACCGGTTCACGCTGGTCGTGCGGCATTCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		18454-18413	CGAATGCCGCACGACCAGCGTGAACCGGTCGTGCGGCATTCG
GTE6-P2	22	21210-21231	CGAGGGCCACGCGTGGCGCTCG
		21231-21210	CGAGCGCCACGCGTGGCCCTCG
GTE6-P3	20	41544-41563	GTCAACCGGCGCCGGTTCAC
		41563-41544	GTGAACCGGCGCCGGTTGAC
GTE8-P1	45	32958-33002	CCCGGACCGGCAGGTGGACTCCCCTCATCCCCCTGCCGGTCCGGG
		33002-32958	CCCGGACCGGCAGGGGGATGAGGGGAGTCCACCTGCCGGTCCGGG
GTE8-P2	40	63863-63900	TAGCGCCGGGTGTCTGG-GTACGCCAG-CACCCGGCGCTA
		63900-63863	TAGCGCCGGGTG-CTGGCGTAC-CCAGACACCCGGCGCTA
GTE8-P3	36	43435-43470	AGCGGGTGGACCGACCCCTCAATCGGTCCACCCGCT
		43470-43435	AGCGGGTGGACCGATTGAGGGGTCGGTCCACCCGCT
GTE8-P4	28	8229-8256	CAATTGGTACGCGCGAACGTACCAATTG
		8256-8229	CAATTGGTACGTTCGCGCGTACCAATTG
GTE8-P5	16	26433-26448	GTTCACCGCGGTGAAC
		26448-26433	GTTCACCGCGGTGAAC

Table 4.8 Virion structural proteins identified by mass spectroscopy

Phage	Protein containing motifs	Amino acid sequence	Coverage
GMA2	Orf9	VYTEWQDDAWGYYDAISEIK	30.4%
		LYPALSIDADGVPISTSNY	
		GEQEKDIAGELSLPEEITDEVVK	
		HEELVAELFSGHGGQSSLLR	
		TVGDIELPR	
		ESDQFLVER	
		LLQGIDLPK	
		YSNAIQIDESLYK	
		TDPAASANAGYDNFAVSAATWR	
		GDVIAPTDR	
		NAQNDPVSPTESATTVNPGNPVGGR	
	Orf10	KYGVPIGSPINAAK	29.4%
		LSSNESNANSDLSW	
		TPSKDSVSAFIK	
		HYLLTGEK	
		QVTPDDLFEDPSWEHVK	
		HVVYETPHPDTK	
		QGSGSTFSESFQQYESAAK	
		KTLGDGTHPVVGAK	
		AIDADGNVGVITDVYQTYTK	
		SKLPDVLIGDSSHAWGTK	
		ERPMFFPSPDKHEK	
		YGLSLTPTQNK	
		ISDFTFGSSK	
		YNSHKPEYR	
		YGGFSGMSGIQMDK	
		VIEQINSNAGKEYYEPLVK	
		SSAVTAEVPGQFK	

Phage	Protein containing motifs	Amino acid sequence	Coverage
		SEVGDYDAYTPHITAVTGQQAR	
	Orf16	LDDLMSLDAGNREELR	37.7%
		ADITAVYEEADADAPEVDEMIVAMR	
		ALNNFETMLETFER	
		SVATLIASIGEER	
		DALPTFGATR	
		YIAPPVLGAYNDAISLWTAANDANPTNPTK	
		FAERTLLNKISAASTK	
		VTTSWNQGAAR	
		DLLLAITR	
		HRIPRPVQLR	
		ALAPEWAR	
		ISWHMDDTFTSQSNNAALNDLPSSIK	
		DSDLVGTNDYMTFVETFEGIAK	
GMA4	Orf7	IFANGGGTSGGAV	30.0%
		VYDQLTK	
		GDRDVQQVAPGAEFPIVDFER	
		SAFQTVYGNQWK	
		DVLANWNVDMVASNQVPAGTAWVVAER	
		LEKPLSTETWR	
	Orf13	KDGGFEFGADR	24.6%
		DISETESLGYASPTRR	
		DVQKEDVTINFALQEFKR	
		SGPQETLFGR	
	Orf66	SAAWMQLPQNLTSLEGTTK	41.3%
		LDEFAAPTTAVSMGTQR	
		ITNLATGTGAADAVTK	
		AQLDAVAAVANAAASGIAIK	
		TNITLTGAQTIDGVAVVAGD	

Phage	Protein containing motifs	Amino acid sequence	Coverage
		VLVAGQTSAAANGIYLAAAGAWS	
		MVAGSTGEIITAGNGLTK	
		ASGGITVDGTGIAVDSTIAR	
GMA5	Orf14	RPGIYNVTAVWPWAANATGR	9.90%
	Orf15	TIWNVQHVNYPAR	21.6%
		LIESKPGTFAVLGYSQGGAIASRIGQELLTGR	
		TAVEHSAFHLNYWGAR	
GMA6	Orf20	DIYSGLLTGSVNPRGFGSVQFDPAVPR	48.0%
		VRNLFPVAATSANLIDYFR	
		VLGFAENGGNGNAR	
		AAADGIAAPAGTATDTFGLKPK	
		FESAQAPVR	
		TIAHWEAAHR	
		SATLSVIANYPGTGFVLHPHDWEDIELQK	
		ANGDGQYMLVTNVAVGATTSVWR	
		QPVVETPAIAEGSWLTGAFGIGAQLYDR	
		IAEQHADFFVR	
		NAIAILAEER	
		LALAVKRPESFVKGTFV	
	Orf29	NGPSTPLR	60.9%
		SASGQAFFSGLAER	
		GPSDAAILIR	
		GLADYESVFGK	
		RPAYGYLYDTVK	
		TFFDEGGEQAYVTR	
		VVGPDATKGTIVLVDR	
		ATPTPANTLTFDAASAGAWSGDLK	
		IAVEDGSIADSVK	
		GEPVEVQNNLR	

Phage	Protein containing motifs	Amino acid sequence	Coverage
		TPAQIAQR	
		RIALLSHNDGATK	
		TELAQTVTAVDDDSAGLFAPWIQVNDGAGGIR	
		SLSSDEQNYGFLSAR	
		LVVESEKR	
		LEDYVFAPIDSK	
		NQLLSAINAELVGIVEPMR	
		QAGGLYEQIDANGQQIDPGYMIETGNTVNSAQSLANNEVR	
		LSPTGALVSLDIVK	
	Orf43	LPQWSSGTTDSPSR	45.8%
		YYVDTDGVIYR	
		TEGNDPALWIDLPAGTTADAIR	
		SANASGVTVYGAMAQSGNLIETK	
		TSTSADLTQITASGDVNTLGR	
		LMAGSLTPQDAQLYIQNSGTTRPALLAR	
		TNTTLPNANTAVAVENQSGATQLLK	
		AINAADNVAVNSNYTATNSGTSSLIQAQLVFR	
		QQPGDTSAGSSLGVVAGTAGADGLAPER	
		FGVNDAK	
		FVANEPDWAPVVVR	
		GAVSQAADLLLAQDSDENKVAGINYR	
		SLATTGTGLNAFSGPITSAGEIQGTNLR	
		VIQGSGNNAGVLSQIK	
		INNNLELK	
GRU3	Orf4	ATAMSVPAIKR	33.7%
		GGTVLSDQPTWIDR	
		TNGPVSPYHR	

Phage	Protein containing motifs	Amino acid sequence	Coverage
		GTGGAVIAADR	
		GENGGVAFTSNGVEVR	
		EHGTFDAHLLVDGR	
		VLSLPGSALDATVDK	
		ASLNYETQEGK	
		AADLIDYGLSAYMSPITAR	
	Orf6	ALAVPGVEAWLR	14.6%
		TVSGLVLPWNK	
		TTLGPLVIRPGGTR	
		SDGLWMDFAFAETPDADAAIAQVK	
		VTTLHAQHTK	
		ENNLMNAAAFIR	
		ALMAAGMSEADAR	
	Orf15	TIWNVQHVNYPAR	24.0%
		DTDSSIPGVALTGEGISGEPLGGFR	
		FGNSHTDYGNTSKR	
		TAVEHSAFHLNYWGAR	
GTE6	Orf14	WQDEVWSLR	18.1%
		VTEGVVGELAQQLFGNLPDVEQK	
		LVWSVHSSSELLGSQAGQYQITDGVTPR	
		LASGGGLLLVTQDVEILNK	
		ENLIALLSQFGQQPGAAVEILR	
	Orf20	ELLGESDRSEVDALLNEAIEK	24.4%
		ADIASLQIQFPEDLVASGASDDGQVLRR	
		QVDFGAIYAGGMAGR	
		VNAISLAR	
		FKVVLPIWYKPTFR	
		TGVDNALAVTDAQIEGWFSDR	
		ALVYPEGTIVR	

Phage	Protein containing motifs	Amino acid sequence	Coverage
		GRGDIINLEAVYDSVGLTTNDFLR	
		LFMEESLAIAWR	
	Orf23	RWGYHLFPFIGAARL	16.8%
		DGTGWGVGPYNVTLDDATTPAPAK	
GTE8	Orf18	IPLFEEDRIR	33.9%
		AVALQINLKR	
		LQFVGNGQNFPVDFGR	
		RVTDVPDGAVAPPSAIDALFGER	
		DINNQTIIKDLVPR	
		DGIIALDGEGDAAVAGSSELGSTMWGK	
		MWVSAHAIAMPVLVNYSSFMK	
	Orf24	KPLAGIIGVAPEDLELDADFK	28.9%
		SDGVVFSADTETSDVESWGALEPTR	
		SDQNWNSEDALVHGMTITAK	
	Orf31	AFIEVTNGEGELLLPR	30.3%
		LAPDLILEEDQDADITPQLPAGLSDVDR	
		GLAGDAILQANDFAAPEGGAEGMTDGTTLVWDSTVGK	
		IIAQITVPEQPFAWHPR	
		GPGHTISTFIDNYYPR	

5. Locating and activating molecular 'time bombs': can Mycolata prophages be selectively induced en masse to biologically control activated sludge foaming?

5.1. Abstract

Little is known about the prevalence, functionality, and ecological roles of temperate phages infecting members of the mycolic acid producing bacteria, the Mycolata. While many lytic phages infective for these organisms have been isolated, and assessed for their suitability for use as biological control agents of activated sludge foaming, no studies have investigated how temperate phages might be induced for this purpose. Bioinformatic analysis using the PHAge Search Tool (PHAST) on Mycolata whole genome sequence data in GenBank for members of the genera *Gordonia, Mycobacterium, Nocardia*, *Rhodococcus,* and *Tsukamurella* revealed that prophages were present.

Subsequent prophage inductions using mitomycin C were conducted on 21 Mycolata strains isolated from activated sludge foams. This led to the isolation and genome characterisation of three novel *Caudovirales* temperate phages, namely GAL1, GMA1, and TPA4, induced from *Gordonia alkanivorans*, *Gordonia malaquae,* and *Tsukamurella paurometabola*, respectively. All possessed highly novel dsDNA genome sequences. Investigation of the genome sequence of *Gordonia amarae,* a common foaming organism globally, suggested that Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), Restriction modification (RM), and Abortive Infection (Abi) systems may interfere with phage infection and diffusion throughout activated sludge foam communities.

5.2. Introduction

The availability of next generation DNA sequencing has resulted in a dramatic increase in the number of draft and fully annotated bacterial genome sequences in public databases. Yet a surprisingly neglected area of genomics is that of phages, especially considering that current estimates place their global abundance at 10^{31} (Chibani-Chennoufi *et al.*, 2004; Hendrix, 2002). To date (March 2015), 1,388 phage genome sequences have been deposited in GenBank (NCBI, 2015), and of these 228 infect members of the genus *Mycobacterium,* making their genomes the

most extensively characterised of all phages.

Based on genomic attributes, these mycobacteriophages have been grouped into clusters and sub-clusters based on their proposed evolutionary relationships (Hatfull *et al.*, 2010). The complete genome sequences of several other phages infective for members of the Mycolata, a distinct evolutionary lineage in the *Actinobacteria,* have been obtained (Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012b, c; Petrovski *et al.*, 2013b, c, 2014). Their hosts include members of the genera *Gordonia*, *Nocardia*, *Rhodococcus,* and *Tsukamurella.* Some of these phages share both nucleotide and amino acid sequence similarity with each other, and with some mycobacteriophages (Petrovski *et al.*, 2011b; Petrovski *et al.*, 2012b; Petrovski *et al.*, 2013b; Summer *et al.*, 2011).

Bacterial whole genome sequence data have revealed novel insights into their evolutionary origin, including several genetic features that appear to be of phage origin (Canchaya *et al.*, 2003; Casjens, 2003; Lawrence *et al.*, 2001). Putative silent phage infections can occupy up to 20% of a bacterial chromosome, and these can play important roles in conferring both virulence and resistance to phage infection to their host (Canchaya *et al.*, 2003; Casjens, 2003; Lawrence *et al.*, 2001).

Some prophages can also be defective, having suffered genetic mutations making their lytic cycle impossible (Canchaya *et al.*, 2003; Casjens, 2003). These cryptic prophage infections have been suggested as mechanisms for reducing cell metabolic burden caused by synthesis of additional phage DNA, while also defusing a dangerous molecular 'time bomb' capable of destroying an entire bacterial population should these be induced into their lytic cycle (Canchaya *et al.*, 2003; Lawrence *et al.*, 2001; Paul, 2008). The activation of temperate phages into their lytic cycle can occur under several inducing conditions. These include environmental perturbations (Brüssow & Kutter, 2004; Canchaya *et al.*, 2003; Lawrence *et al.*, 2001; Paul, 2008; Weinbauer, 2004). Eventual host cell lysis renders any genetic benefit conferred from such an infection ephemeral, while prophages remain intact (Brüssow & Kutter, 2004; Canchaya *et al.*, 2003; Lawrence *et al.*, 2001; Paul, 2008).

To our knowledge, most Mycolata phages completely sequenced and described are lytic phages

that have been isolated by their ability to form plaques on lawn cultures of their bacterial hosts (Crockett & Brownell, 1972; Petrovski *et al.*, 2011a, b, c; Petrovski *et al.*, 2012b,c; Petrovski *et al.*, 2014; Summer *et al.*, 2011; Sunairi *et al.*, 1993). These phages have possible use as biocontrol agents to limit proliferation of their hosts, including destabilisation of foams in activated sludge reactors and controlling animal bacterial pathogens, including those resistant to antibiotics (Petrovski *et al.*, 2011b, c; Petrovski *et al.*, 2012b; Summer *et al.*, 2011; Thomas *et al.*, 2002; Withey *et al.*, 2005). However, this conventional approach to phage therapy is faced with several challenges.

Bacterial host defence systems, many of which might be identifiable in whole genome sequence data, can prevent phage infections in several ways. These include Superinfection Exclusion (Sie), Clustered Regularly Interspersed Short Palindromic Repeats (CRISPR), Abortive infection (Abi), and Restriction Modification (RM) systems (Labrie *et al.*, 2010). The RM systems cleave invading foreign DNA including that of phages (Labrie *et al.*, 2010), using restriction endonucleases acting at specific recognition sites, while host bacterial DNA is protected from this by chemical modification, usually involving methylation (Labrie *et al.*, 2010). Alternatively, CRISPR systems also protect the host bacterium by acting as an 'immune system', targeting invading phages and other foreign DNA (Labrie *et al.*, 2010). CRISPRs usually contain between 21 to 48 bp direct repeats interspersed by non-replicative spacers matching the invading phage DNA sequence, that are usually 26 to 72 bp long, and flanked by between 4 to 20 *cas* genes (Labrie *et al.*, 2010). Furthermore, some bacteria are equipped with Abi systems that provide phage resistance by halting the infection cycle, usually by targeting a crucial step in phage replication (Labrie *et al.*, 2010). Abi systems can target replication, transcription, or translation, and results in a seemingly altruistic death of the cell, together with the infective phage (Labrie *et al.*, 2010). Furthermore, resident prophages that contain Sie systems can block the entry of other foreign phage DNA into the host cell and confer immunity to some phages onto the cell (Labrie *et al.*, 2010).

Prophage induction therapy could be particularly useful for phage biocontrol where lytic phages cannot be isolated readily for a particular bacterium, either because of fastidious host cell growth requirements or the phage defence mechanisms mentioned above, but where complete prophages are present in the host (de los Reyes III, 2010). Conventional phage therapy requires several stages. These would include phage isolation, purification, characterisation, mass production, determination of appropriate cocktail mixes of phages, and development of an

appropriate delivery method (Carlton, 1999; Loc-Carrillo & Abedon, 2011; Thiel, 2004). Thus, selective induction of temperate phages *en masse* across a group of problematic organisms like the Mycolata is attractive as an alternative means of controlling their growth, and likely to be more cost and time effective.

However, before this approach can be considered, several key questions need to be addressed. These include (i) Do foaming Mycolata members contain prophages? (ii) Is it possible to induce temperate Mycolata phages to activate their lytic cycles? (iii) What role might these prophages play in enhancing the fitness of their lysogen hosts? (iv) Are these Mycolata phages similar to other previously isolated lytic phages, and what can we learn about their evolutionary ancestry? (v) What potential phage defence systems might inhibit the diffusion of Mycolata phages, both lytic and lysogenic, within activated sludge? This study attempts to begin to address these questions.

5.3. Materials and methods

5.3.1. Mycolata bacterial, plasmid, and prophage sequence analyses

Bacterial and plasmid whole genome sequence (wgs) data were obtained from GenBank by searching for the genera of interest by name (*Gordonia, Mycobacterium, Nocardia, Rhodococcus,* and *Tsukamurella*), and those present are described in detail in Table 5.2 (section 5.6) and Table 5.3 (section 5.6). Putative prophage DNA sequence data were detected using PHAST (Zhou *et al.*, 2011), which was preferred over other programs including Prophage Finder (Bose & Barber, 2006), Prophinder (Lima-Mendez *et al.*, 2008), and Phage finder (Fouts, 2006) because of its faster run times, ability to process a wider variety of file types (annotated or un-annotated) of differing completeness (draft or finished), and increased sensitivity (Zhou *et al.*, 2011).

Table 5.1 Summary of Mycolata strains used in Mycolata prophage induction studies

Species	Lab ID	Other ID	Species genome sequences available [a]	Putative complete prophage(s)	Putative incomplete prophages(s)[b]	Mitomycin C induction detected	Spontaneous induction detected
Dietzia maris	CON27^T	Dmar27, DSMZ 43672	0	NA	NA	-	-
Gordonia alkanivorans	CON72		1	1	0	+	-
Gordonia amarae	CON44^T	Gama44, DSMZ 43392, NBRC 15530	1	0	1	-	-
Gordonia amarae	CON9	Gama9, UQCC2810	1	0	1	-	-
Gordonia amarae	BEN371		1	0	1	-	-
Gordonia amarae	BEN374		1	0	1	-	-
Gordonia amarae	BEN381		1	0	1	-	-
Gordonia amarae	BEN386		1	0	1	-	-
Gordonia amarae	BEN389		1	0	1	-	-
Gordonia desulfuricans	CON69^T	213E, NCIMB 40816	0	NA	NA	-	-
Gordonia polyisoprenivorans	CON71		1	0	2	-	-
Gordonia malaquae	BEN700		1	0	1	+	+

Species	Lab ID	Other ID	Species genome sequences available [a]	Putative complete prophage(s)	Putative incomplete prophages(s)[b]	Mitomycin C induction detected	Spontaneous induction detected
Tsukamurella paurometabola	CON55		1	0	1	+	+
Nocardia brasiliensis	CON42 [T]		1	0	0	-	-
Nocardia brevicatena	CON43		1	1	1	-	-
Rhodococcus erythropolis	BEN703		0	NA	NA	-	-
Rhodococcus sp.	J27		NA	NA	NA	-	-
Rhodococcus sp.	J71		NA	NA	NA	-	-
Rhodococcus sp.	J72		NA	NA	NA	-	-
Millisia brevis	J81		0	NA	NA	-	-
Millisia brevis	J82 [T]	DSMZ 44463	0	NA	NA	-	-

[a] Bacterial strains do not match with wgs data in all cases; [b] Includes those that scored a completeness level of questionable (Q) when screened using PHAST; [T] indicates type strain.

5.3.2. Bacterial strains and media

Bacterial strains used in this study are listed in Table 5.1, Table 8.1, and in Petrovski *et al.* (2011d), and methods for their storage and cultivation were those described by Petrovski *et al.* (2011d). Three strains (J27, J71 and J72) isolated in this laboratory and belonging to a proposed new species of *Rhodococcus* (Soddell, unpublished) were grown on homemade R2A medium (0.5 g/L Yeast extract (Oxoid, Adelaide, Australia), 0.5 g/L Proteose peptone (Difco, North Ryde, Australia), 0.5 g/L Casamino acid (Difco, North Ryde, Australia), 0.5 g/L Glucose, 0.5 g/L soluble starch (Difco, North Ryde, Australia), 0.3 g/L K_2HPO_4, 0.005 g/L $MgSO_4.7H_2O$, 0.3 g/L sodium pyruvate (BDH, Murarrie, Australia)) broth and agar R2A + 14 g/L agar (Oxoid, Adelaide, Australia) at 25°C. All remaining chemicals were obtained from Sigma (Sydney, Australia) unless otherwise noted.

5.3.3. Phage induction procedures

Bacterial strains were grown to stationary phase with shaking in either R2A or PYCA broth (Petrovski *et al.*, 2011d), and then exposed to mitomycin C in attempts to induce any potential prophages. Rapid screening of multiple concentrations of mitomycin C (0.2, 0.5, 1.0, 2.0, 5.0, 10.0, 15.0 and 20.0 µg/mL) were carried out on 1 mL aliquots of each of the stationary phase bacterial cultures. These cultures were exposed to the range of mitomycin C concentrations overnight, then combined and filtered (0.22 µM pore size) to remove any bacterial cell debris. Filtrates were subjected to DNAse/RNAse pre-treatment to remove any bacterial DNA prior to polyethylene glycol (PEG) precipitation to concentrate phage particles. Phage DNA extraction/purification involved proteinase K digestion, and phenol-chloroform-isoamyl alcohol purification, as detailed by Petrovski *et al.* (2011a). The DNA extracted from mitomycin C treated cells was screened by agarose gel electrophoresis, where presence of a suitably sized (>= 14 kbp) DNA band suggested detection of a putative *Caudovirales* prophage.

Where such a band was observed, separate assays using a 1 mL aliquot of stationary phase bacterial culture and mitomycin C at each of the aforementioned concentrations were then conducted to determine the optimum concentrations of mitomycin C for future prophage

induction experiments. Similar experiments were conducted on either solid PYCa or R2A media lacking mitomycin C, to screen for any spontaneous prophage inductions on solid media. Thus, lawn plates of the bacterial strains were incubated at 30°C for three days prior to recovering cells into suspension with either PYCa or R2A broth. DNA extraction and agarose gel electrophoresis to detect putative prophages were performed as described above.

5.3.4. Sequencing and annotation of induced phages

Phage DNA sequencing libraries were prepared using an Illumina Nextera XT sample preparation kit following the manufacturer's instructions. The prepared DNA libraries were sequenced on an Illumina MiSeq as a 150-bp paired end run and sequence reads were assembled using CLC Workbench (v6.5.1). Open reading frames (*orf*s) within the *de novo* assembled sequences were detected using Glimmer (v3.02), for *orf*s with a minimum size of 90 bp (Delcher *et al.*, 2007). All predicted start codons were inspected manually for the presence of putative ribosomal binding sites and corrected if required. Sequence similarity searches were performed against genome sequences in GenBank, as described by Petrovski *et al.* (2011c). The presence of tRNA and tmRNA was determined using ARAGORN (Laslett & Canback, 2004) and with tRNAScan-SE (Schattner *et al.*, 2005). Transmembrane domains were predicted with the DAS transmembrane prediction server (Cserzo *et al.*, 1997), as described in Petrovski *et al.* (2011c).

Phage DNA when analysed by gel electrophoresis gave results consistent with circularly permuted DNA genomes. Therefore, for consistency the genomes annotations were conducted starting with the DNA packaging operon.

5.3.5. Preliminary induced phage characterisation

Induced phage host ranges were determined by plating a 1:10 dilution series of phage onto host bacterial lawn plates (Table 8.1), as described by Petrovski *et al.* (2011c). Grids used for TEM visualisation of virions were prepared on a carbon coated formvar 200 mesh grid (Electron

Microscopy Sciences, Australia), negatively stained with uranyl acetate as described by Petrovski *et al.* (2011d), and examined with a Tenaci Fei T30 Transmission Electron Microscope (TEM).

5.3.6. Nucleotide sequence accession numbers

The nucleotide sequences for induced prophages GAL1, GMA1, and TPA4 were deposited in GenBank under accession numbers KR053194, KR053195, and KR053196, respectively.

5.4. Results and discussion

5.4.1. Prophages are prevalent in Mycolata species

All available whole genome sequence data from *Gordonia, Nocardia, Rhodococcus, Tsukamurella, Mycobacterium,* and their corresponding plasmid sequences were downloaded from the GenBank database (NCBI, 2015) on 26th February 2014 (bacterial wgs data), and 1st July 2014 (plasmid wgs data). In total 259 bacterial genomes and 49 plasmid sequences were obtained. The bacterial sequences screened came from Mycolata isolates from many different sources including clinical samples, contaminated soil, activated sludge, faeces, and animals, as described in Table 5.2 (section 5.6).

When these sequences were screened for putative prophage regions, PHAST predicted that 83% of the 259 bacterial genomes may have posses/carry genes of phage origin (Table 5.2 in section 5.6). Twenty-six per cent of these sequences appeared to contain putative complete intact prophage genomes, based on the PHAST predictions (Table 5.2 in section 5.6) that ranged from 5.4 to 135.5 kbp in size. Some Mycolata genomes contained up to nine putative PHAST predicted prophage regions, and up to 4.26% of complete Mycolata genomes were occupied by such regions.

As one possible application for prophage induction therapy is to control Mycolata activated sludge foaming in wastewater treatment plants, the four Mycolata isolated from wastewater whose genomes had been sequenced were examined closely. These were *Gordonia amarae* strain NBRC 15530, *Gordonia malaquae* strain NBRC 108250, *Gordonia sihwensis* strain NBRC 108236, and *Rhodococcus ruber* strain Chol-4. Of these only the *R. ruber* strain appeared free of genes of phage origin.

Several of these putative prophage regions showed sequence similarities at an amino acid level to previously characterised Mycolata phages. For example, *Mycobacterium tuberculosis* strain CCDC5180 contained a putative remnant prophage region that encoded putative capsid proteins similar to those seen in *Rhodococcus* phages REQ1 and RRH1 (Petrovski *et al.*, 2013a; Petrovski *et al.*, 2012c), and a small terminase subunit encoding gene was similar to that in the *Nocardia* phage NBR1 genome (Petrovski *et al.*, 2014). It was also possible to trace the same prophage sequence across genomes of multiple strains of the same host organism. For example, *Mycobacterium abscessus* strains 5S-0921 and 5S-1212 both contained an identical intact prophage sequence of 32.7 kbp in length.

Forty-nine plasmid sequences were also analysed from the genera listed above, as Kanda *et al.* (1989) have demonstrated that prophages can integrate into plasmids. Five Mycolata plasmids were predicted by PHAST to contain intact prophage genomic sequences (Table 5.3 in section 5.6). All were examined manually and the genes detected were mainly homologues of integrases, recombinases, and translocases, suggesting that most likely these were not from prophages. Similar homologues of these genes were observed in some of the bacterial genomic data, making it difficult sometimes to discern if these are true genes of phage origin, as these may also be associated with other mobile genetic elements like plasmids and transposons.

The three bacterial species in our culture collection predicted from this *in silico* screening to contain complete prophage genomes within their chromosomes were *G. alkanivorans, G. malaquae,* and *N. brevicatena*. Therefore, these three species whose genomes had not yet been sequenced were used in attempts to induce any possible prophages in their genomes (Table 5.1).

5.4.2. Detection of induced prophages in G. alkanivorans and G. malaquae isolates

When *G. alkanivorans* strain CON72*, G. malaquae* strain BEN700*,* and *N. brevicatena* strain CON43 from the La Trobe University culture collection were exposed to a range of concentrations of mitomycin C, potential prophages GAL1 and GMA1 were induced from *G. alkanivorans*, and *G. malaquae,* respectively. No prophages were detected from this strain of *N. brevicatena,* possibly because of an over-prediction by PHAST, a loss of inducibility, absence of the prophage, or resistance to mitomycin C at the concentrations used. Phage GAL1 was only inducible at 20 µg/mL mitomycin C, whereas spontaneous induction of phage GMA1 was also detected in the absence of mitomycin C, with *G. malaquae*. However, exposure to mitomycin C did appear to result in a higher level of GMA1 phage induction as evidenced by a higher intensity DNA band when subjected to agarose gel electrophoresis [data not shown]. When GAL1 phage was examined by TEM, the presence of a *Siphoviridae* phage possessing the characteristic long non-contractile tail (~450 nm) and isometric capsid (~75 nm diameter) was seen (Figure 5.1).

DNA was isolated from both phages GAL1 from *G. alkanivorans*, and GMA1 from *G. malaquae*, and sequenced using the Illumina platform. Their genomes were assembled and data revealed ~927-fold and ~2599-fold average coverage, respectively. GAL1 and GMA1 phages had genome sizes of 49,979 bp and 41,207 bp, and contained 82 and 68 putative ORFs, respectively. Their genomes did not appear to contain any tRNA or tmRNA.

Despite the GAL1 phage being induced from a different strain of *G. alkanivorans* (CON72) than that sequenced previously and deposited in GenBank, analysis of its genome sequence revealed an almost identical alignment to the putative prophage detected in *G. alkanivorans* strain NBRC16433 contig GOALK93 (49,954 bp). This high level of similarity between these two sequences suggests the predicted prophage in strain NBRC16433 might be functional. PHAST identified putative integration sites *attR* (nt 3577189-3577200) with the sequence TCGGCGTACGTG, and *attL* at locus (nt 3524537-3524548) with the sequence (TCGGCGTACGTG), which might also be shared between the two phages. Twenty-four per cent of the genome of phage GMA1 also shared 98% nucleotide sequence identity with a PHAST predicted putative incomplete prophage from *G. malaquae* strain NBRC 108250, suggesting an evolutionary relationship between them.

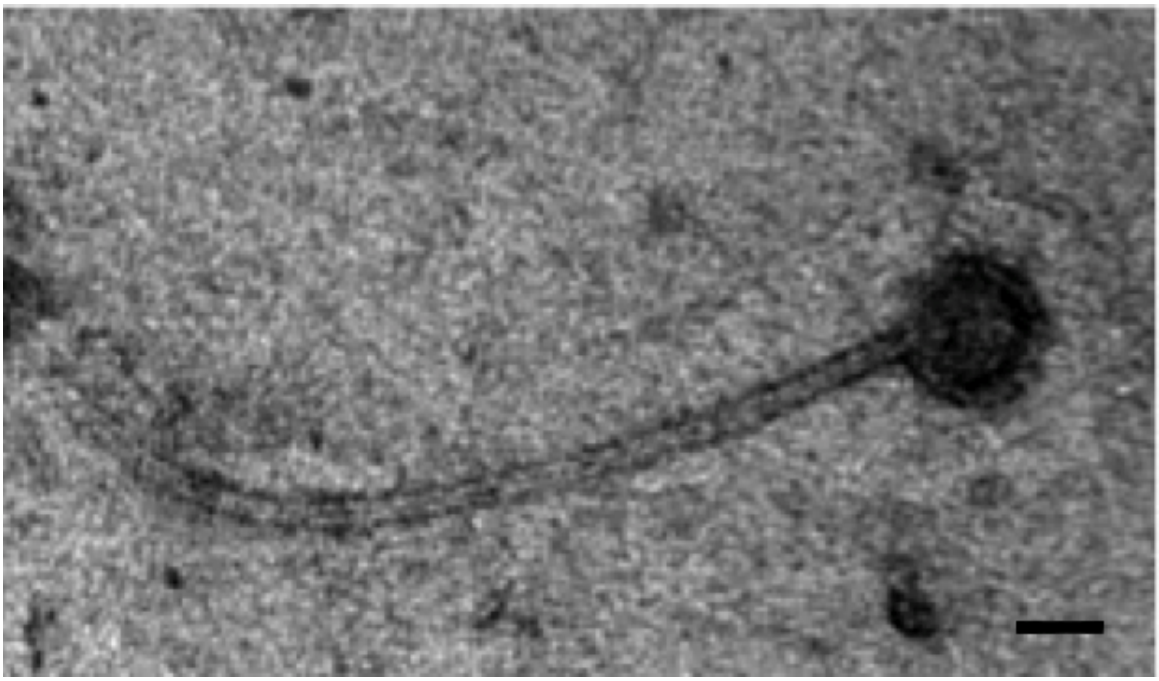

Figure 5.1 Morphology of temperate Actinophage GAL1.

Scale = 50 nm. Negatively stained electron microscopy, 2% uranyl acetate.

5.4.3. Activating the lytic cycle of prophages in other foaming Mycolata isolates

An additional 18 Mycolata strains (Table 5.1) whose genomes had not yet been sequenced and were isolated by our group from foaming wastewater treatment plants, were also exposed to mitomycin C to see if prophages were present there. These strains included seven *G. amarae* strains (Table 5.1), the most common foaming organism in Australian activated sludge treatment plants (Blackall *et al.*, 1991; de los Reyes III, 2010; Seviour *et al.*, 1990; Seviour *et al.*, 1994). In one strain of *T. paurometabola* (CON55), spontaneous prophage induction was observed, making it difficult to determine the impact of mitomycin C on it, although, DNA bands of a higher intensity were seen from DNA extracted from exposed cells on agarose gels [data not shown], suggesting an increase in the level of prophage induction. This phage is referred to as phage TPA4.

5.4.4. Infection properties of induced prophages

The induced phages GAL1, GMA1, and TPA4 were then characterised. In host range studies, none of the three phages formed plaques on any of the 94 Mycolata screened, including the hosts from which they were isolated. This finding may suggests that these phages will only enter their lytic cycles in response to a signal triggering the SOS response (Little & Michalowski, 2010). It is clear from work carried out here that *G. alkanivorans* (CON72) cells are lysed in order to liberate GAL1 phage after exposure to mitomycin C. Furthermore, an increase in titres of both TPA4 and GMA1 phages was seen with *T. paurometabola* (CON55) and *G. malaquae* (BEN700) after mitomycin C exposure, as indicated by increased intensities of phage DNA bands on agarose gels (see above). Phage GAL1 was only inducible using 20 µg/mL of mitomycin C.

Despite *T. paurometabola* strain CON55 carrying the temperate phage TPA4, it allowed a co-infection with a different lytic phage TPA2 (Petrovski *et al.*, 2011a) resulting in plaque formation. This outcome suggested that the TPA4 temperate phage does not confer immunity on this host to phage TPA2. A similar co-infection of *G. malaquae* strain BEN700 was successful with phage GTE2 (Petrovski *et al.*, 2011c).

5.4.5. Temperate phage TPA4 is a novel phage

The DNA obtained from the TPA4 phage was sequenced using the Illumina MiSeq platform, as with the GAL1 and GMA1 phages. Its genome was 56,212 bp in size, with ~1,749-fold average coverage, and containing 84 ORFs (Figure 5.2 and Table 5.4 in in section 5.6). No putative tRNA or tmRNA could be recognised. It shared a small region (1,825 bp) of nucleotide sequence identity (75%) with *Kineococcus radiotolerans* strain SRS30216. However, no nucleotide sequence similar was observed to *T. paurometabola* strain DSMZ 20162, or to the PHAST predicted prophage region found in its genome, or to phage TPA2 also infective for this strain (Petrovski *et al.*, 2011a).

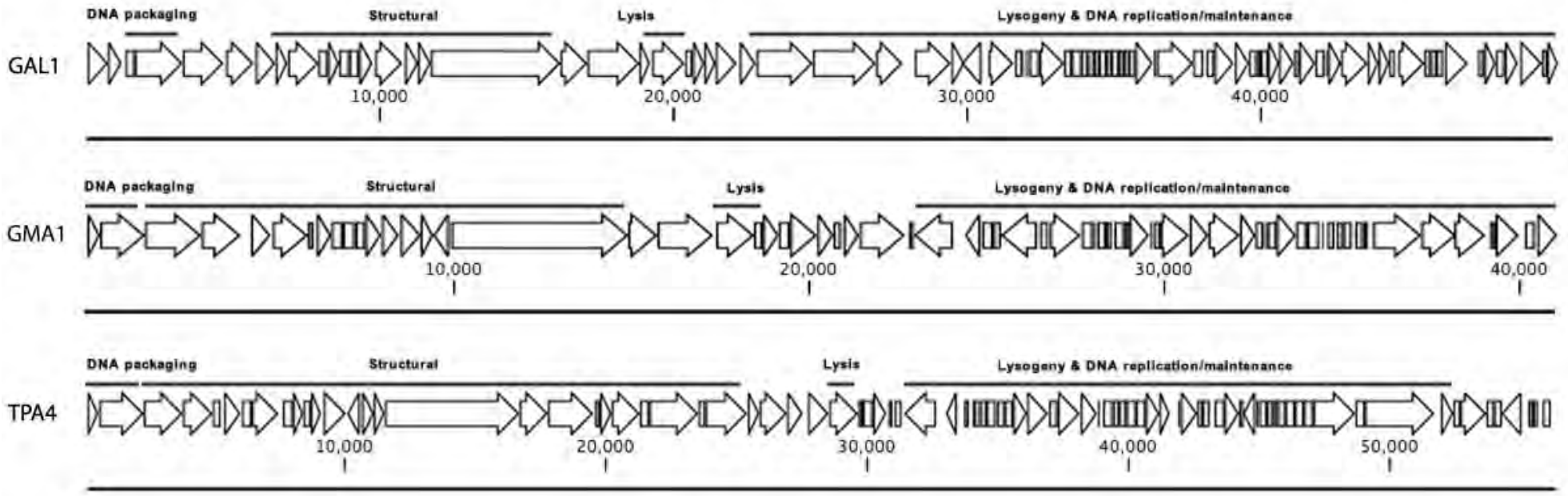

Figure 5.2 Genome maps of temperate Actinophages GAL1, GMA1, and TPA4

5.4.6. Sequence repeats are common within the prophage genomes

Several repeat structures were seen in genomes of all three prophages induced in this study. All three contained between 4 to 13 palindromes, ranging in size from 16 to 86 bp (see Table 5.5 in section 5.6). Some were located in intergenic regions where they might act as *rho*-independent transcriptional terminators (Lesnik *et al.*, 2001). Between 30 to 179 direct repeats were also found in each of the genome sequences of all three phages (Table 8.5 in section 8.2), and these ranged in size between 15 to 193 bp. Furthermore, between 11 and 103 inverted repeats ranging from 15 to 60 bp in size were also detected in all three phage genome sequences (see Table 8.5 in section 8.2).

Repeat sequences can indicate repeated protein structures, promoters, replication origins and transposable elements (Mott & Berger, 2007), but none of these were identified in any of the prophages induced and sequenced here. What functional roles these direct and inverted repeats might play, if any, remains to be determined. However, repeat sequences similar to these have been reported in lytic Mycolata phages (Petrovski *et al.*, 2011a, c; Petrovski *et al.*, 2012b) so their presence is not restricted to Mycolata prophages.

5.4.7. Summary of features of induced prophage genome sequences

All three temperate phages GAL1, GMA1, and TPA4 had the characteristic modular arrangement of *Siphoviridae* phages, encoding DNA packaging genes, structural protein genes, lysis genes, DNA replication genes, and genes associated with lysogeny (Figure 5.2).

5.4.7.1. Phage lysis genes

In GAL1, GMA1 and TPA4 phages, an N-acetylmuramoyl-L-alanine amidase motif was identified for products of genes *orf22* (pfam01510), *orf20* (pfam01510), and *orf31* (pfam01519), respectively. This motif identified these as lysis genes, presumably activated at the induction of

their lytic cycles. In the genomes of phages GAL1 and GMA1 these lysis genes were followed immediately by the genes encoding putative holin proteins. These holins, identified by amino acid sequence homology were similar to other putative phage holin proteins. They fulfilled the criteria outlined by Wang *et al.* (2000), who suggested that these proteins should be less than 150 amino acid residues in length and contain two or more transmembrane regions. In phage TPA4, no putative holin genes could be identified in the vicinity of the phage lysin gene, suggesting that if present, it is either in an unusual location or has a highly novel amino acid and nucleotide sequence.

5.4.7.2. Phage lysogenic conversion and maintenance genes

Integrases were identified in all three phage genomes by their amino acid sequence similarities. These were Orf31 in GAL1, Orf30 in GMA1, and Orf37 in TPA4. Their presence suggests an ability of all three to enter a lysogenic replication cycle (Groth & Calos, 2004). In phage GAL1, a BRO family protein (pfam02498) motif was seen in the N-terminal region of Orf39. The C-terminus contained a phage antirepressor KilAC motif (pfam03374) that may be involved in inactivating phage repressor proteins upon prophage induction (Lemire *et al.*, 2011). Phage GMA1 also contained a BRO family protein motif (pfam02498) at the N-terminal end of Orf36 that appears to be encoded by a chimeric gene, as the N-terminal region of its translated protein sequence showed similarity to that from *Mycobacterium avium*, and the C-terminal to that encoded by a gene from *G. alkanivorans*. This is the same region present in contig GOALK93, which corresponds to the Orf34 of phage GAL1. No putative excisionase genes could be found in any of these three phage genome sequences, suggesting that if present, they are novel.

A number of genes encoding putative proteins were identified in the GAL1 prophage genome that might confer enhanced fitness to *G. alkanivorans*. These included the *orf34* of phage GAL1 whose encoded product contains a motif (pfam09669) for a Rha family regulatory protein. In lamboid phage phi 80, this protein is thought to interfere with phage infection in bacterial strains that do not possess host integration host factor (IHF) (Henthorn & Friedman, 1995). The *orf47* gene of phage GAL1 encodes a protein with high amino acid sequence similarity to a putative beta subunit of a protocatechuate 3,4-dioxygenase, involved in degradation of aromatic compounds (Frazee *et al.*, 1993). Simiarly, *orf46* of phage TPA4 appeared to encode a Cas4-like

protein which could function to modulate the CRISPR-Cas system of the host, however, this remains to be experimentally confirmed.

5.4.8. Evolutionary insights into Mycolata prophages

All three phages differed to each other at the nucleotide sequence level, and their genomes encoded a high proportion of proteins (between 7 to 24%) for which no statistically significant matches in GenBank could be found. However, some amino acid sequence similarity (75% identity, 49% coverage) was seen between the products of *orf34* of GAL1 phage, a putative Rha protein, and *orf36* of GMA1 phage that encodes a putative BRO family protein. Furthermore, both phage GMA1 and TPA4 contained a number of genes that encoded *G. malaquae* protein homologues. Phages GAL1, GMA1, and TPA4 are each unique, and could not be integrated into the clustering system of Hatfull *et al.* (2010) for Mycobacteriophages because of the high novelty of their genomes at both a nucleotide and amino acid sequence level. They were placed instead in the category of singletons, sharing no obvious evolutionary relationship with any other characterised *Mycobacterium* phages. Other lytic actinophages such as *Gordonia* phage GTE2 could not be assigned to any clusters either, and thus is also a singleton (Petrovski *et al.*, 2011c).

Despite phage GMA1 not fitting into any of these clusters, it possessed several genes whose encoded products had amino acid sequence similarities to those in *Mycobacterium* phages Mutaform13, Ramsey, Gumball, PMC, and Dori. This finding infers some shared evolutionary ancestry between these phages and the GMA1 phage. Furthermore, the GMA1 phage genome contained many genes that possessed translated amino acid sequences with similarities to several hypothetical proteins from *Rhodococcus* spp., *G. malaquae,* and *G. sihwensis.*

5.4.9. Phage defence systems are present in the foaming Mycolata

Gordonia amarae has been implicated in activated sludge foaming events worldwide (de los Reyes III, 2010), and no lytic phages for this organism have been isolated and completely characterised. So the presently unpublished draft genome sequence for the type strain (NBRC

15530^T) in the GenBank database (accession no. NZ_BAED00000000, deposited 2011) was examined for genes potentially associated with phage defence systems. The putative remnant prophage region in this genome identified by PHAST (Table 5.2 in section 5.6) contained no genes suggestive of Sie systems. As RM systems consist typically of a modification enzyme and a restriction endonuclease (Furuta & Kobayashi, 2000), several putative nuclease and methylase genes were identified in this genome sequence, and these are listed in Table 5.6 (section 5.6). Contigs GOAMR_51 and GOAMR_25 both contained genes associated with Modification Dependent Systems (MDS), i.e., those encoding a McrC protein and a McrBC protein, respectively (Labrie *et al.*, 2010; Loenen & Raleigh, 2014). MDS are specific for either methylated or hydroxymethylated DNA (Labrie *et al.*, 2010). These are thought to have arisen as a result of the proposed evolutionary arms race between phages and their bacterial hosts (Labrie *et al.*, 2010), possibly by phages acquiring methylase genes that allow them to evade classical RM systems (Samson *et al.*, 2013).

Several Cas genes associated with CRISPR systems were located adjacent to each other in contig GOAMR_28 (Table 5.6 in section 5.6). If functional, these might suggest that *G. amarae* contains a CRISPR system. However, as these genes were located on a contig boundary, further support for this suggestion, by locating the characteristic repeats and spacer regions associated with these systems was not possible (Westra *et al.*, 2012). Genes that might be associated with abortive infection systems including toxin-antitoxin systems could also be identified in the *G. amarae* PHAST predicted putative prophage region. These included a single abortive infection protein encoding gene in contig GOAMR_03, which might be involved in cleaving essential cellular components (Samson *et al.*, 2013; Westra *et al.*, 2012).

The presence of phage defence systems in bacterial hosts is challenging for lytic phage based methods of biological control, as these presumably limit rates of successful phage infection and spread. However, where foaming Mycolata have been shown to contain complete functional prophages, the possibility of selectively inducing them *en masse* is an attractive idea. Clearly adding mitomycin C to operating full-scale wastewater treatment plants is not economically viable or safe. However, other prophage inducing agents including UV irradiation may be worth consideration, as these are used widely in activated sludge plants to minimise spread of pathogenic bacteria.

5.5. Conclusions

In silico wgs data analyses suggest that prophages occur commonly in the Mycolata, identifying many potentially useful molecular 'time bombs', to be activated *en masse* to target these organisms and hence biologically control any problems they cause including the formation of stable foams in wastewater treatment plants. Preliminary induction studies have suggested that some of these lysogenic phages are functional and can be induced into their lytic cycles. Their genomes reveal genes that may provide their host cells with advantageous attributes, which may be ecologically important. Prophage induction therapy could be an attractive approach in targeting hosts such as *G. amarae*, with multiple phage defence systems where these are infected with temperate phages, or where the bacterium in question is difficult to culture in a laboratory making lytic phage isolation difficult.

5.6. Appendix

Table 5.2 Putative prophages detected in publically available GenBank Mycolata wgs data using PHAST

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Gordonia aichiensis* NBRC 108223** [d]	Human sputum, Japan	NZ_BANR01000000	1	N.A.	3917277-3992148	74.8	Y	63	65.53
***Gordonia alkanivorans* NBRC 16433** [d]	Tar and phenol contaminated soil, Germany	NZ_BACI01000000	1	N.A.	3524537-3578378	53.8	Y	77	63.84
***Gordonia amarae* NBRC 15530** [d]	Abnormal foam on activated sludge, USA	NZ_BAED01000000	1	N.A.	299556-307233	7.6	N	10	65.67
***Gordonia amicalis* NBRC 100051** [d]	Garden soil, Russian Federation	NZ_BANS00000000	0	N.A.					
***Gordonia araii* NBRC 100433** [d]	Human, Japan	NZ_BAEE00000000	2	N.A.	1392152-1398883	6.7	N	8	65.30
					3895971-3904239	8.2	Q	10	66.31
Gordonia bronchialis	Sputum,	NC_013441	2	0.91%	402787-428773	25.9	Y	22	63.97

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
DSM 43247 [f]	Human with cavitary disease of both upper lungs				1230465-1251773	21.3	N	15	63.38
***Gordonia effusa* NBRC 100432** [d]	Human, Japan	NZ_BAEH00000000	0	N.A.					
***Gordonia hirsuta* DSM 44140 (NBRC 16056)** [d]	Packing material of a bio-filter	NZ_BANT00000000	0	N.A.					
***Gordonia* sp. KTR9** [f]	Soil from an explosives testing facility	NC_018581	2	1.18%	1517766-1570506	52.7	N	48	66.10
					3874442-3885732	11.2	N	7	62.93
***Gordonia malaquae* NBRC 108250** [d]	Sludge of a wastewater treatment plant, Taiwan	NZ_BAOP00000000	2	N.A.	3536151-3573410	37.2	Y	45	65.84
					4006158-4019973	13.8	N	23	65.79
***Gordonia namibiensis* NBRC 108229** [d]	Soil, Namibia	NZ_BAHE00000000	1	N.A.	3539990-3546762	6.7	N	7	68.73
***Gordonia neofelifaecis* NRRL B-59395** [d]	Faeces, China	NZ_AEUD00000000	2	N.A.	441982-458783	16.8	N	23	67.86
					4198346-4216288	17.9	Q	28	67.94

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
Gordonia otitidis **NBRC 100426** [d]	Human, Japan	NZ_BAFB00000000	3	N.A.	4243486-4248947	5.4	N	7	66.64
					4404497-4430217	25.7	N	20	63.48
					4658711-4673172	14.4	N	9	63.34
Gordonia paraffinivorans **NBRC 108238** [d]	Water sample from oil-producing well, China	NZ_BAOQ00000000	0	N.A.					
Gordonia polyisoprenivorans **NBRC 16320** [d]	Fouling tire water, Germany	NZ_BAEI00000000	2	N.A.	1472324-1482288	9.9	Q	9	63.67
					2332996-2341150	8.1	Q	10	62.19
Gordonia polyisoprenivorans **VH2 (DSMZ 44266)** [f]	Earth of rubber tree plantation, Vietnam	NC_016906	2	0.51%	1456870-1472745	15.8	Y	19	62.43
					1564965-1578231	13.2	N	12	63.57
Gordonia rhizosphera **NBRC 16068** [d]	Root, rhizosphere of mangrove trees, Japan	NZ_BAHC00000000	0	N.A.					
Gordonia rubripertincta **NBRC 101908** [d]	Soil, Japan	NZ_BAHB00000000	1	N.A.	594440-601941	7.5	Q	6	62.48

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
Gordonia sihwensis **NBRC 108236** [d]	Wastewater-treatment bioreactor, Korea	NZ_BANU00000000	6	N.A.	274177-286634	12.4	N	17	67.92
					285658-319974	34.3	Q	20	67.48
					1653943-1693179	39.2	Y	59	64.32
					3362150-3376889	14.7	N	17	64.22
					3615805-3635013	19.2	N	13	66.99
					3666347-3680269	13.9	N	7	70.68
Gordonia soli **NBRC 108243** [d]	Soil, Taiwan	NZ_BANX00000000	4	N.A.	4903248-4944450	41.2	Y	41	67.99
					4980403-4999965	19.5	N	32	68.15
					5225127-5255781	30.6	Q	36	66.33
					5328847-5359588	30.7	N	37	67.35
Gordonia sputi **NBRC 100414** [d]	Sputa of patients with pulmonary disease	NZ_BAFC00000000	1	N.A.	2579797-2589231	9.4	N	11	60.46
Gordonia terrae **NBRC 100016** [d]	Soil, Japan	NZ_BAFD00000000	1	N.A.	3548749-3558477	9.7	N	8	61.87
Mycobacterium	N.D.	NC_021282	5	4.26%	772815-828935	56.1	Y	41	63.62

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***abscessus subsp. bolletii* 50594[f]**					822655-845576	22.9	N	30	63.22
					1573263-1644763	71.5	Y	96	60.23
					1854758-1863909	9.1	N	10	65.94
					3433590-3486892	53.3	Y	75	63.04
***Mycobacterium abscessus* ATCC 19977**	N.D.	NC_010397	1	0.91%	1782165-1828386	46.2	Y	39	59.91
***Mycobacterium abscessus* 3A-0119-R [d]**	Sputum, Human	NZ_AKUX00000000	8	N.A.	109885-164646	54.7	Y	66	63.91
					407161-452128	44.9	Y	63	63.86
					1453952-1463715	9.7	N	15	59.09
					1477659-1520948	43.2	Y	50	60.47
					1577669-1595325	17.6	N	20	61.29
					2991215-3035148	43.9	Y	61	64.67
					3933832-3955847	22	N	26	64.07
					3952174-3968978	16.8	N	18	63.97
***Mycobacterium abscessus* 3A-0122-R [d]**	Sputum, Human	NZ_AKUY00000000	5	N.A.	514740-555566	40.8	Y	57	63.78
					1513575-1523338	9.7	N	16	59.09

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
					1537283-1580457	43.1	Y	52	60.47
					1651790-1669446	17.6	N	20	61.28
					3968888-4053927	85	Y	121	64.23
***Mycobacterium abscessus* 3A-0122-S [d]**	Sputum, Human	NZ_AKUZ00000000	9	N.A.	241418-281439	40	Y	56	63.78
					1065906-1075669	9.7	N	15	59.09
					1077967-1111407	33.4	Y	36	60.41
					1181409-1199065	17.6	N	20	61.28
					1482620-1493497	10.8	N	17	61.24
					2611951-2647689	35.7	Y	39	64.70
					3479109-3496213	17.1	N	29	63.37
					3769725-3793752	24	N	30	63.72
					3790079-3806883	16.8	N	18	63.97
***Mycobacterium abscessus* 3A-0731 [d]**	Sputum, Human	NZ_AKVA00000000	8	N.A.	109901-165856	55.9	Y	66	63.90
					1079592-1114497	34.9	Y	47	63.94
					1423191-1432954	9.7	N	14	59.09

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
					1446899-1490073	43.1	Y	49	60.47
					1546794-1564450	17.6	N	20	61.28
					3156535-3187637	31.3	Y	39	64.70
					4019064-4061249	42.1	Q	63	64.01
					4057576-4074380	16.8	N	18	63.97
***Mycobacterium abscessus* 3A-0810-R [d]**	Bronchial alveolar lavage, Human	NZ_AKUP00000000	9	N.A.	109837-165791	55.9	Y	67	63.90
					407636-446855	39.2	Y	58	63.75
					1509148-1518911	9.7	N	16	59.09
					1521209-1554840	33.6	Y	38	60.41
					1611561-1629059	17.4	N	21	61.30
					1753320-1782585	29.2	N	16	62.98
					3045844-3087591	41.7	Y	59	64.73
					3919086-3962282	43.1	Q	63	64.06
					3958609-3975422	16.8	N	18	63.96
Mycobacterium	Sputum,	NZ_AKVB00000000	7	N.A.	109877-165831	55.9	Y	67	63.90

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***abscessus* 3A-0930-R** [d]	Human				407847-449145	41.2	Y	58	63.79
					1506729-1516492	9.7	N	15	59.09
					1529492-1572666	43.1	Y	50	60.47
					1629387-1647043	17.6	N	20	61.29
					3887720-3931599	43.8	Y	58	63.81
					5244818-5273535	28.7	Y	39	64.91
***Mycobacterium abscessus* 3A-0930-S** [d]	Sputum, Human	NZ_AKVC00000000	8	N.A.	109889-165843	55.9	Y	67	63.90
					1118306-1128069	9.7	N	15	59.09
					1142014-1185188	43.1	Y	50	60.47
					1241909-1259565	17.6	N	20	61.28
					2206800-2248098	41.2	Y	58	63.78
					2990072-3004926	14.8	N	23	64.15
					3835701-3879144	43.4	Y	57	63.85
					5235152-5249987	14.8	Q	25	65.58
***Mycobacterium abscessus* 4S-0116-R** [d]	Sputum, Human	NZ_AKVD00000000	1	N.A.	3229344-3281457	51.2	Y	73	63.21

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Mycobacterium abscessus* 4S-0116-S [d]**	Sputum, Human	NZ_AKVE00000000	1	N.A.	2294181-2347641	53.4	Y	75	63.15
***Mycobacterium abscessus* 4S-0206 [d]**	Sputum, Human	NZ_AKUT00000000	1	N.A.	3236939-3290399	53.4	Y	74	63.15
***Mycobacterium abscessus* 4S-0303 [d]**	Sputum, Human	NZ_AKTU00000000	1	N.A.	3218556-3273405	54.8	Y	77	63.09
***Mycobacterium abscessus* 4S-0726-RA [d]**	Sputum, Human	NZ_AKTV00000000	1	N.A.	3155792-3210641	54.8	Y	77	63.09
***Mycobacterium abscessus* 4S-0726-RB [d]**	Sputum, Human	NZ_AKTW00000000	1	N.A.	2746347-2799807	53.4	Y	74	63.15
***Mycobacterium abscessus* 5S-0304 [d]**	Sputum, Human	NZ_AKTX00000000	4	N.A.	2613910-2642403	28.4	N	29	63.20
					2630693-2663407	32.7	Y	45	63.54
					3151707-3193917	42.2	Q	62	63.92
					3193933-3206308	12.3	N	18	64.66
***Mycobacterium abscessus* 5S-0421 [d]**	Sputum, Human	NZ_AKTY00000000	5	N.A.	2885728-2914221	28.4	N	29	63.20
					2902511-2935225	32.7	Y	45	63.54
					3423525-3465735	42.2	Q	62	63.92

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
					3465751-3478261	12.5	N	18	64.46
					4449383-4474591	25.2	Q	21	60.46
***Mycobacterium abscessus* 5S-0422** [d]	Sputum, Human	NZ_AKTZ00000000	4	N.A.	3777961-3806454	28.4	N	29	63.20
					3794744-3827458	32.7	Y	45	63.54
					4315696-4357906	42.2	Q	62	63.92
					4357922-4370432	12.5	N	18	64.46
***Mycobacterium abscessus* 5S-0708** [d]	Sputum, Human	NZ_AKUA00000000	5	N.A.	2659111-2687604	28.4	N	29	63.20
					2675894-2708608	32.7	Y	45	63.54
					3196846-3239056	42.2	Q	62	63.92
					3239072-3251582	12.5	N	18	64.46
					4242372-4267580	25.2	Q	21	60.46
***Mycobacterium abscessus* 5S-0817** [d]	Lung autopsy, Human	NZ_AKUB00000000	4	N.A.	2227716-2256209	28.4	N	29	63.20
					2244499-2277213	32.7	Y	45	63.54
					2765451-2807661	42.2	Q	62	63.92
					2807677-2820187	12.5	N	18	64.46

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Mycobacterium abscessus* 5S-0921** [d]	Lymph node biopsy, Human	NZ_AKUQ00000000	5	N.A.	3561816-3590309	28.4	N	29	63.20
					3578599-3611313	32.7	Y	45	63.54
					4097750-4139960	42.2	Q	62	63.92
					4139976-4152351	12.3	N	18	64.66
					5143650-5168858	25.2	Q	21	60.46
***Mycobacterium abscessus* 5S-1212** [d]	Sputum, Human	NZ_AKUC00000000	5	N.A.	2387334-2415827	28.4	N	30	63.20
					2404117-2436831	32.7	Y	45	63.54
					2925068-2967278	42.2	Q	61	63.92
					2967294-2979804	12.5	N	18	64.46
					3970653-3995861	25.2	Q	21	60.46
***Mycobacterium abscessus* 5S-1215** [d]	Sputum, Human	NZ_AKUD00000000	3	N.A.	543915-599790	55.8	Y	78	64.32
					2682084-2710577	28.4	N	29	64.20
					2698867-2731581	32.7	Y	45	63.54
***Mycobacterium abscessus* 6G-0125-R** [d]	Sputum, Human	NZ_AKUE00000000	3	N.A.	66052-120303	54.2	Y	68	64.38
					802919-841094	38.1	Q	36	60.12

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
					4557643-4575474	17.8	N	10	60.80
***Mycobacterium abscessus* 6G-0125-S [d]**	Sputum, Human	NZ_AKUF00000000	4	N.A.	547349-565279	17.9	N	11	60.78
					1197620-1251871	54.2	Y	69	64.38
					1915250-1925824	10.5	N	21	58.21
					1935012-1973358	38.3	Q	35	60.11
***Mycobacterium abscessus* 6G-0212 [d]**	Sputum, Human	NZ_AKUR00000000	3	N.A.	420248-438178	17.9	N	10	60.78
					1070515-1124766	54.2	Y	68	64.38
					1807596-1845939	38.3	Q	36	60.11
***Mycobacterium abscessus* 6G-0728-R [d]**	Sputum, Human	NZ_AKUS00000000	3	N.A.	548967-566897	17.9	N	10	60.78
					1199246-1253497	54.2	Y	69	64.38
					1936410-1974756	38.3	Q	36	60.11
***Mycobacterium abscessus* 6G-0728-S [d]**	Sputum, Human	NZ_AKUG00000000	3	N.A.	230291-248122	17.8	N	10	60.80
					929020-967363	38.3	Q	36	60.11
					1548133-1602384	54.2	Y	69	64.38
Mycobacterium	Sputum,	NZ_AKUH00000000	4	N.A.	550282-568212	17.9	N	10	60.78

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***abscessus* 6G-1108** [d]	Human				1197986-1252237	54.2	Y	70	64.38
					1915504-1926078	10.5	N	21	58.21
					1935269-1973612	38.3	Q	36	60.11
***Mycobacterium abscessus* 47J26** [d]	Sputum sample from CF patient, Human, United Kingdom	NZ_AGQU00000000	3	N.A.	1760597-1782369	21.7	N	27	60.53
					1776946-1799973	23	N	14	60.48
					2823468-2877820	54.3	Y	74	64.10
***Mycobacterium abscessus* 9808** [d]	Wound of exudate of postoperative patients, Human, China	NZ_ANAR00000000	2	N.A.	4294238-4344432	50.1	Y	70	64.06
					5059013-5122817	63.8	N	76	63.64
***Mycobacterium abscessus* subsp. bolletii BD (DSMZ 45149)** [d]	Broncho-alveolar liquid, France	NZ_AHAS00000000	1	N.A.	4299517-4371046	71.5	Q	85	63.68
***Mycobacterium abscessus* M24** [d]	Broncho-alveolar lavage sample from	NZ_AJLY00000000	7	N.A.	187423-202814	15.3	N	18	62.04
					413950-443162	29.2	Y	40	63.58

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
	patient with lung infection, Human, Malaysia				1055788-1112837	57	Y	62	63.53
					2302724-2320245	17.5	N	17	55.50
					2936437-2952326	15.8	Q	20	64.72
					4658246-4674037	15.7	N	20	61.14
					5328152-5343522	15.3	N	22	61.60
***Mycobacterium abscessus* M93** [d]	Sputum of a patient with lung infections, Human, Malaysia	NZ_AJGF00000000	4	N.A.	3300104-3338963	38.3	Y	43	63.84
					3352404-3373412	21	N	41	63.17
					4133118-4185793	52.6	Q	63	62.85
					4188228-4207912	19.6	N	25	61.88
***Mycobacterium abscessus* M94** [d]	Sputum of a patient with lung infections, Human	NZ_AJGG00000000	2	N.A.	1138860-1202845	63.9	Y	72	64.35
					3656790-3750803	94	Q	100	58.06
***Mycobacterium abscessus* M115** [d]	Sputum sample from a patient with lung infection, Human,	NZ_AJLZ00000000	3	N.A.	1531354-1554355	23	N	17	62.29
					3069333-3155505	86.1	Y	100	63.19
					4808660-4867675	59	N	81	63.91

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
	Malaysia								
***Mycobacterium abscessus* M139** [d]	Sputum sample from patient with lung infection, Human, Malaysia	NZ_AKVR00000000	2	N.A.	2893239-2942215	48.9	Y	58	64.78
					5034033-5046263	12.2	Y	16	54.62
***Mycobacterium abscessus* M148** [d]	Sputum sample from patient with lung infection, Human, Malaysia	NZ_AKVV00000000	9	N.A.	211977-247242	35.2	N	51	64.33
					409941-428623	18.6	N	32	56.40
					451544-473404	21.8	N	9	61.92
					535393-556169	20.7	N	25	64.98
					1624001-1648412	24.4	N	21	63.39
					2722788-2758053	35.2	N	21	63.30
					3371898-3389510	17.6	N	19	64.87
					4166366-4194351	27.9	N	40	63.75
					4696682-4717506	20.8	N	37	63.24
Mycobacterium	Sputum	NZ_AKVT00000000	3	N.A.	1897722-1912440	14.7	N	25	64.10

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***abscessus* M152** [d]	sample from patient with lung infection, Human, Malaysia				4781005-4806057	25	N	31	63.95
					4806073-4818805	12.7	N	17	64.37
***Mycobacterium abscessus* M154** [d]	Sputum sample from a patient with lung infection, Human, Malaysia	NZ_AJMA00000000	1	N.A.	1245061-1267131	22	N	23	61.29
***Mycobacterium abscessus* M156** [d]	Sputum sample from patient with lung infection, Human, Malaysia	NZ_AKVU00000000	5	N.A.	442783-481816	39	N	34	63.87
					2195989-2235859	39.8	Y	48	63.84
					2367196-2389266	22	N	23	61.29
					4435202-4487053	51.8	N	62	64.12
					5009680-5020998	11.3	Q	18	55.38
***Mycobacterium abscessus* M159** [d]	Sputum sample from a patient with lung infection, Human,	NZ_AJSD00000000	0	N.A.					

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
	Malaysia								
***Mycobacterium abscessus* M172** [d]	Sputum sample from a patient with lung infection, Human, Malaysia	NZ_AJSE00000000	9	N.A.	1558332-1580402	22	N	25	61.29
					1837893-1866432	28.5	N	33	63.87
					2714489-2730476	15.9	N	21	61.16
					2734183-2748188	14	N	25	58.30
					2792748-2809309	16.5	N	25	56.80
					2813622-2846708	33	Y	21	58.58
					3381549-3436748	55.2	N	74	64.04
					3425180-3476134	50.9	Y	58	63.48
					4473837-4530262	56.4	Q	67	63.77
***Mycobacterium abscessus* CF** [d]	N.D.	NZ_CAHZ00000000	1	N.A.	3000231-3051357	51.1	Y	62	64.60
***Mycobacterium africanum* GM041182**	N.D.	NC_015758	1	0.17%	1154358-1161894	7.5	N	11	65.25
***Mycobacterium avium* 104**	N.D.	NC_008595	4	2.02%	746437-795760	49.3	N	64	66.79
					1178359-1197714	19.3	N	23	66.53

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
					1429012-1464399	35.3	Y	24	66.79
					2934861-2949403	14.5	N	13	69.50
***Mycobacterium avium* subsp. avium ATCC 25291** [d]	Liver of diseased hen	NZ_ACFI00000000	2	N.A.	2474391-2498352	23.9	N	25	69.48
					4417447-4426114	8.6	N	7	70.18
***Mycobacterium avium* subsp. paratuberculosis JQ5** [d]	*Camelus dromedarius*	NZ_AHAZ00000000	0	N.A.					
***Mycobacterium avium* subsp. paratuberculosis JQ6** [d]	*Camelus dromedarius*	NZ_AHBA00000000	1	N.A.	3161127-3192743	31.6	N	22	67.28
***Mycobacterium avium* subsp. paratuberculosis K-10**	Faeces, animal, United States	NC_002944	0	N.A.					
***Mycobacterium avium* subsp. paratuberculosis MAP4**	Faeces, animal, United States	NC_021200	1	0.65%	4258270-4289551	31.2	N	10	68.20
***Mycobacterium avium* subsp. paratuberculosis S5** [d]	Terminally sick Jamunapari goat, India	NZ_ANPD00000000	1	N.A.	2875542-2885533	9.9	N	8	69.20

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
Mycobacterium avium **subsp. paratuberculosis S397** [d]	Suffolk breed of sheep with Johne's disease, United States	NZ_AFIF00000000	0	N.A.					
Mycobacterium bovis **AF2122/97 (ATCC BAA-935)**	Tuberculous cow with caseous lesions in lung and bronchomediastinal lymph nodes, England, United Kingdom	NC_002945	1	0.48%	1753159-1773871	20.7	N	22	66.10
Mycobacterium bovis **BCG str. Korea 1168P**	N.D.	NC_020245	1	0.18%	1158781-1166317	7.5	N	11	65.21
Mycobacterium bovis **BCG str. Mexico**	N.D.	NC_016804	0	N.A.					
Mycobacterium bovis **BCG str. Pasteur 1173P2**	N.D.	NC_008769	0	N.A.					

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
Mycobacterium bovis **BCG str. Tokyo 172**	N.D.	NC_012207	0	N.A.					
Mycobacterium canettii CIPT 140010059	N.D.	NC_015848	2	0.56%	1175143-1191934	16.7	N	11	64.40
					3132808-3141183	8.3	N	9	66.18
Mycobacterium canettii **CIPT 140060008**	N.D.	NC_019950	2	0.91%	2359232-2379642	20.4	Y	21	65.94
					3425584-3445682	20	N	9	64.61
Mycobacterium canettii **CIPT 140070002 [d]**	N.D.	NZ_CAOL00000000	0	N.A.					
Mycobacterium canettii **CIPT 140070005 [d]**	N.D.	NZ_CAOM00000000	1	N.A.	3932695-3940991	8.2	N	8	67.63
Mycobacterium canettii **CIPT 140070007 [d]**	N.D.	NZ_CAOO00000000	1	N.A.	912000-943297	31.2	Q	40	63.57
Mycobacterium canettii **CIPT 140070008**	N.D.	NC_019965	1	0.38%	2379516-2396325	16.8	N	21	65.73
Mycobacterium canettii **CIPT**	N.D.	NC_019951	1	0.24%	1511064-1521712	10.6	N	10	65.16

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
140070010									
***Mycobacterium canettii* CIPT 140070013** [d]	N.D.	NZ_CAON00000000	0	N.A.					
***Mycobacterium canettii* CIPT 140070017**	N.D.	NC_019952	0	N.A.					
***Mycobacterium chubuense* NBB4** [f]	Creosote-contaminated soil	NC_018027	0	N.A.					
***Mycobacterium colombiense* CECT 3035 (DSMZ 45105)** [d]	Blood of a HIV positive patient, South America	NZ_AFVW00000000	2	N.A.	346288-381299	35	N	13	68.96
					3425806-3442217	16.4	N	17	69.13
***Mycobacterium fortuitum* subsp. fortuitum DSM 46621** [d]	Cold abscess, Human	NZ_ALQB00000000	3	N.A.	2258394-2279587	21.1	N	21	66.07
					2965733-2997490	31.7	N	8	64.30
					3154534-3184955	30.4	N	13	64.87
***Mycobacterium gilvum* PYR-GCK** [f]	N.D.	NC_009338	0	N.A.					

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Mycobacterium gilvum* Spyr1** [f]	Creosote-contaminated soil, Greece	NC_014814	2	1.09%	2607245-2644937	37.6	Q	14	62.06
					4767385-4789945	22.5	N	9	67.40
***Mycobacterium hassiacum* DSM 44199** [d g]	Urine, Germany	NZ_AMRA00000000	4	N.A.	1309867-1331196	21.3	N	7	68.21
					2225906-2244099	18.1	N	7	69.78
					2949565-3006543	56.9	Y	83	63.51
					4283332-4297497	14.1	N	7	69.13
***Mycobacterium hassiacum* DSM 44199** [d g]	Urine, Germany	NZ_ARBU00000000	2	N.A.	3237548-3292409	54.8	Q	82	62.99
					3312556-3321918	9.3	N	8	70.31
***Mycobacterium indicus* pranii MTCC 9506 (DSMZ 45239)**	Animal, South India	NC_018612	2	0.63%	3134984-3158732	23.7	Q	15	64.13
					4731466-4743160	11.6	N	10	65.42
***Mycobacterium intracellulare* ATCC 13950** [g]	Clinical isolates from a Korean pulmonary patient, male, 64 years of age, Human, Korea	NC_016946	0	N.A.					

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Mycobacterium intracellulare* ATCC 13950 [d g]**	Human lymph node	NZ_ABIN00000000	0	N.A.					
***Mycobacterium intracellulare* MOTT-02**	Clinical isolates from a Korean pulmonary patient, male, 64 years of age, Human, Korea	NC_016947	1	0.61%	792142-824918	32.7	N	17	65.17
***Mycobacterium intracellulare* MOTT-64**	Clinical isolates from a Korean pulmonary patient, male, 64 years of age, Human, Korea	NC_016948	2	0.67%	811269-829827	18.5	N	18	65.29
					1064478-1082928	18.4	N	9	66.85
***Mycobacterium kansasii* ATCC 12478 [f g]**	From a fatal case	NC_022663	3	0.62%	1384178-1395201	11	N	9	67.04
					2874153-2884766	10.6	N	14	65.42
					3758746-3776831	18	Q	18	64.33

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Mycobacterium kansasii* ATCC 12478 [d f g]**	From a fatal case	NZ_ACBV00000000	4	N.A.	4411548-4426164	14.6	N	23	64.10
					5358572-5370007	11.4	N	14	64.82
					6048284-6056911	8.6	N	11	62.87
					6376434-6384682	8.2	N	10	63.69
***Mycobacterium leprae* Br4923**	Brazil	NC_011896	0	N.A.					
***Mycobacterium leprae* TN**	N.D.	NC_002677	0	N.A.					
***Mycobacterium liflandii* 128FXT [f]**	N.D.	NC_020133	2	0.30%	4386145-4395998	9.8	N	7	64.23
					4979523-4988579	9	N	10	65.25
***Mycobacterium marinum* M (ATCC BAA-535) [f]**	Tissue samples of human lesions, United States	NC_010612	2	0.46%	1764630-1771903	7.2	N	10	64.70
					4825093-4848031	22.9	N	26	63.15
***Mycobacterium massiliense* 1S-151-0930 [d]**	Bronchial alveolar lavage, Human	NZ_AKUI00000000	1	N.A.	2531963-2597196	65.2	Q	87	63.69
***Mycobacterium massiliense* 1S-152-**	Bronchial alveolar	NZ_AKUJ00000000	1	N.A.	3553849-3619082	65.2	Q	87	63.69

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
0914 [d]	lavage, Human								
***Mycobacterium massiliense* 1S-153-0915** [d]	Sputum, Human	NZ_AKUK00000000	1	N.A.	2890488-2955721	65.2	Q	87	63.69
***Mycobacterium massiliense* 1S-154-0310** [d]	Bronchial alveolar lavage, Human	NZ_AKUL00000000	1	N.A.	2902042-2967275	65.2	Q	87	63.69
***Mycobacterium massiliense* 2B-0107** [d]	Sputum, Human	NZ_AKUN00000000	2	N.A.	313492-346778	33.2	Q	58	59.76
					346833-361617	14.7	N	16	60.04
***Mycobacterium massiliense* 2B-0307** [d]	Sputum, Human	NZ_AKUU00000000	1	N.A.	317738-346567	28.8	N	47	60.03
***Mycobacterium massiliense* 2B-0626** [d]	Bronchial alveolar lavage, Human	NZ_AKUM00000000	2	N.A.	973572-1021697	48.1	Q	74	59.85
					2683294-2713268	29.9	N	16	59.65
***Mycobacterium massiliense* 2B-0912-R** [d]	Sputum, Human	NZ_AKUV00000000	2	N.A.	1391796-1425082	33.2	Q	58	59.76
					1425137-1439921	14.7	N	16	60.04
***Mycobacterium massiliense* 2B-0912-S** [d]	Sputum, Human	NZ_AKUW00000000	2	N.A.	972582-1005868	33.2	Q	58	59.76
					2681757-2711731	29.9	N	16	59.65

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Mycobacterium massiliense* 2B-1231** [d]	Bronchial alveolar lavage, Human	NZ_AKUO00000000	1	N.A.	1034844-1068130	33.2	Q	59	59.76
***Mycobacterium massiliense* CCUG 48898** [d g]	Human sputum and bronchial alveolar lavage, Human, France	NZ_AHAR00000000	3	N.A.	1793291-1822891	29.6	N	29	60.35
					3907391-3945321	37.9	Q	49	63.98
					3945337-3957588	12.2	N	17	64.75
***Mycobacterium massiliense* CCUG 48898** [d g]	Human sputum and bronchial alveolar lavage, Human, France	NZ_AKVF00000000	3	N.A.	1600973-1653490	52.5	Q	78	59.72
					3907330-3945260	37.9	Q	51	63.98
					3945276-3957680	12.4	N	17	64.69
***Mycobacterium massiliense* M18** [d]	Lymph node biopsy of a Malaysian patient with suspected tuberculosis cervical	NZ_AJSC00000000	2	N.A.	4798547-4830969	32.4	N	49	64.36
					4824727-4842815	18	N	18	64.47

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
	lymphadenitis, Human, Malaysia								
***Mycobacterium massiliense* str. GO 06**	N.D.	NC_018150	0	N.A.					
***Mycobacterium neoaurum* VKM Ac-1815D**	N.D.	NC_023036	1	2.49%	4072080-4207674	135.5	Y	127	67.22
***Mycobacterium parascrofulaceum* ATCC BAA-614 [d]**	Clinical specimen (human), Canada	NZ_ADNV00000000	3	N.A.	52-11606	11.5	N	10	59.24
					4336682-4371755	35	Q	23	63.22
					4631977-4695036	63	Y	73	66.63
***Mycobacterium phlei* RIVM601174 [d]**	Human, Netherlands	NZ_AJFJ00000000	2	N.A.	103112-135455	32.3	N	22	68.77
					147443-156919	9.4	N	10	66.96
***Mycobacterium* sp. 155 [d]**	N.D.	NZ_AREU00000000	2	N.A.	1260533-1279296	18.7	N	25	66.39
					3097490-3104486	6.9	N	9	64.68
***Mycobacterium* sp. H4Y [d]**	Sputum, Human, South	NZ_AKIG00000000	4	N.A.	1379632-1404004	24.3	N	10	65.37
					2401904-2412571	10.6	N	11	63.45

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
	Korea				2876718-2894347	17.6	N	16	61.72
					3849582-3885710	36.1	Q	19	64.49
***Mycobacterium* sp. JLS**	Creosote-contaminated soil, United States	NC_009077	0	N.A.					
***Mycobacterium* sp. KMS** [f]	N.D.	NC_008705	2	0.81%	3074173-3108448	34.2	N	19	67.84
					4085194-4097492	12.2	N	21	67.44
***Mycobacterium* sp. MCS** [f]	N.D.	NC_008146	3	1.53%	1770462-1810940	40.4	N	21	63.50
					3056285-3090560	34.2	N	19	67.84
					4050658-4063010	12.3	N	20	67.35
***Mycobacterium* sp. MOTT36Y**	Sputa, Human, South Korea	NC_017904	2	0.36%	967361-980458	13	N	9	65.15
					4758520-4765550	7	N	9	65.50
***Mycobacterium rhodesiae* JS60** [d]	Soil	NZ_AGIQ00000000	7	N.A.	1446937-1458089	11.1	N	13	59.52
					2005316-2016262	10.9	N	12	65.84
					3030284-3060900	30.6	Q	21	66.71

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
					5994993-6024881	29.8	N	66	60.49
					6371973-6397968	25.9	N	19	67.07
					6418923-6448661	29.7	Y	21	61.38
					7156405-7191386	34.9	Y	28	63.27
***Mycobacterium rhodesiae* NBB3**	Estuarine sediment, Australia	NC_016604	0	N.A.					
***Mycobacterium smegmatis* JS623** [f]	Smegma, Human	NC_019966	1	0.40%	4206686-4232683	25.9	N	14	64.24
***Mycobacterium smegmatis* str. MC2 155** [g]	N.D.	NC_008596	2	0.28%	1942190-1952852	10.6	N	10	64.44
					3499988-3508558	8.5	N	9	64.72
Mycobacterium smegmatis str. MC2 155 [g]	N.D.	NC_018289	1	0.28%	1941442-1960865	19.4	N	13	65.58
***Mycobacterium smegmatis* MKD8** [d]	Japan	NZ_AOCJ00000000	3	N.A.	1334549-1341067	6.5	N	7	67.11
					2254937-2278040	23.1	Y	23	65.25
					5526039-5539167	13.1	N	9	64.04

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Mycobacterium thermoresistibile* ATCC 19527 [d]**	Soil	NZ_AGVE00000000	2	N.A.	1438114-1460818	22.7	N	9	66.27
					1690775-1760309	69.5	Q	77	67.11
***Mycobacterium tuberculosis* 02_1987 [d]**	N.D.	NZ_ABLM00000000	2	N.A.	762700-770225	7.5	N	12	65.17
					2582785-2610715	27.9	N	14	66.01
***Mycobacterium tuberculosis* 7199-99**	N.D.	NC_020089	3	1.10%	1162662-1170198	7.5	N	10	65.25
					2970659-2983113	12.4	Q	19	66.27
					3881705-3910421	28.7	N	22	65.18
***Mycobacterium tuberculosis* 94_M4241A [d]**	N.D.	NZ_ABLL00000000	1	N.A.	2706498-2715777	9.2	N	13	66.38
***Mycobacterium tuberculosis* '98-R604 INH-RIF-EM' [d]**	N.D.	NZ_ABVM00000000	3	N.A.	1242115-1249638	7.5	N	13	65.26
					2070321-2083497	13.1	N	13	64.38
					2993899-3010376	16.4	N	14	65.78
***Mycobacterium tuberculosis* str. Beijing/NITR203**	N.D.	NC_021054	2	0.80%	1766846-1788161	21.3	N	17	66.14
					2969790-2983602	13.8	Q	19	66.05
Mycobacterium	N.D.	NZ_AAKR00000000	2	N.A.	796046-803577	7.5	N	13	65.16

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***tuberculosis* C** [d]					2559632-2569050	9.4	N	13	66.89
***Mycobacterium tuberculosis* CAS/NITR204**	N.D.	NC_021193	3	1.18%	1154502-1162028	7.5	N	12	65.25
					1771759-1790415	18.6	N	18	66.62
					2958132-2983883	25.7	N	19	66.24
***Mycobacterium tuberculosis* CCDC5079** [g]	N.D.	NC_017523	2	0.39%	1154189-1161716	7.5	N	8	65.24
					2951861-2961273	9.4	N	9	66.94
***Mycobacterium tuberculosis* CCDC5079** [g]	N.D.	NC_021251	2	0.45%	1156530-1164066	7.5	N	13	65.24
					2964597-2977050	12.4	Q	18	66.28
***Mycobacterium tuberculosis* CCDC5180**	Sputum from patient with secondary pulmonary tuberculosis, Human, China	NC_017522	2	0.38%	1154121-1161648	7.5	N	8	65.24
					2958267-2967679	9.4	N	9	66.93
***Mycobacterium tuberculosis* CDC1551**	N.D.	NC_002755	2	1.18%	677260-703462	26.2	N	16	64.95
					3870574-3896495	25.9	N	20	65.17

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Mycobacterium tuberculosis* CDC1551A** [d]	N.D.	NZ_AELF00000000	4	N.A.	805760-813180	7.4	N	12	64.17
					1650077-1670154	20	N	17	64.38
					2592402-2613751	21.3	Q	15	66.07
					3490212-3518761	28.5	N	19	65.12
***Mycobacterium tuberculosis* CPHL_A** [d]	N.D.	NZ_ACHP00000000	2	N.A.	1618176-1625703	7.5	N	13	65.25
					3418304-3427720	9.4	N	14	66.88
***Mycobacterium tuberculosis* CTRI-2**	Russia	NC_017524	2	0.45%	1158389-1165925	7.5	N	10	65.24
					2964965-2977418	12.4	Q	19	66.26
***Mycobacterium tuberculosis* CTRI-4** [d]	Sputum, Human, Russia	NZ_AIIE00000000	2	N.A.	984864-992391	7.5	N	12	65.25
					2772473-2781885	9.4	N	13	66.93
***Mycobacterium tuberculosis* EAI5/NITR206**	N.D.	NC_021194	3	0.94%	1156406-1163939	7.5	N	13	65.22
					1763970-1785203	21.2	N	11	66.13
					2957342-2969796	12.4	Q	17	66.24
***Mycobacterium tuberculosis* EAI5**	N.D.	NC_021740	1	0.17%	1156610-1164137	7.5	N	12	65.25

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
MUM101									
***Mycobacterium tuberculosis* EAS054** [d]	N.D.	NZ_ABOV00000000	2	N.A.	813712-821239	7.5	N	13	65.22
					1415179-1435797	20.6	N	12	66.11
***Mycobacterium tuberculosis* str. Erdman = ATCC 35801**	Human sputum	NC_020559	3	1.42%	1157016-1164543	7.5	N	13	65.25
					2945745-2970758	25	Q	18	66.28
					3866067-3895979	29.9	N	21	65.16
***Mycobacterium tuberculosis* F11**	N.D.	NC_009565	3	0.67%	1161731-1169267	7.5	N	10	65.28
					1989026-1998675	9.6	N	6	67.59
					2983380-2995833	12.4	Q	18	66.26
***Mycobacterium tuberculosis* GM 1503** [d]	N.D.	NZ_ABQG00000000	1	N.A.	2559244-2580461	21.2	N	13	65.68
***Mycobacterium tuberculosis* H37Ra (ATCC 25177)** [d g]	N.D.	NZ_AAYK00000000	1	N.A.	348817-357557	8.7	N	12	66.03
***Mycobacterium tuberculosis* H37Ra**	N.D.	NC_009525	3	0.97%	1159271-1166807	7.5	N	11	65.25
					1768504-1790023	21.5	N	21	66.10

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
(ATCC 25177) [g]					2982031-2995842	13.8	Q	21	66.05
***Mycobacterium tuberculosis* H37RvCO [d]**	N.D.	NZ_AJSF00000000	2	N.A.	1739822-1760951	21.1	N	11	66.07
					2935352-2946046	10.6	Q	16	66.46
***Mycobacterium tuberculosis* H37Rv (ATCC 27294) [g]**	Lung, Human	NC_000962	3	0.97%	1157963-1165499	7.5	N	10	65.25
					1766987-1788505	21.5	N	22	66.10
					2970063-2983874	13.8	Q	21	66.05
***Mycobacterium tuberculosis* H37Rv (ATCC 27294) [g]**	Lung, Human	NC_018143	3	1.33%	675812-702017	26.2	N	19	64.96
					1766993-1788512	21.5	N	22	66.10
					2970073-2980843	10.7	Q	14	66.57
***Mycobacterium tuberculosis* str. Haarlem [g]**	N.D.	NC_022350	3	1.10%	1162925-1170461	7.5	N	11	65.25
					2965323-2977777	12.4	Q	19	66.28
					3870603-3899317	28.7	N	25	65.17
***Mycobacterium tuberculosis* str. Haarlem [d g]**	N.D.	NZ_AASN00000000	2	N.A.	2457761-2479074	21.3	N	14	65.74
					3366165-3387877	21.7	N	22	65.22
Mycobacterium	N.D.	NC_021192	2	0.42%	1160580-1168142	7.5	N	12	64.62

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***tuberculosis* str. Haarlem/NITR202**					1780017-1790814	10.7	N	11	66.48
***Mycobacterium tuberculosis* K85 [d]**	N.D.	NZ_ACHQ00000000	2	N.A.	1628710-1636237	7.5	N	12	65.22
					3263433-3278255	14.8	N	16	70.76
***Mycobacterium tuberculosis* KZN 605**	Human, South Africa	NC_018078	2	0.63%	1431411-1443844	12.4	Q	19	66.28
					2369462-2384656	15.1	N	12	65.07
***Mycobacterium tuberculosis* KZN 1435**	N.D.	NC_012943	3	0.80%	1431299-1443732	12.4	Q	19	66.28
					2369580-2384774	15.1	N	13	65.07
					3246469-3254005	7.5	N	11	65.25
***Mycobacterium tuberculosis* KZN 4207 [g]**	Human, South Africa	NC_016768	3	0.80%	1431243-1443676	12.4	Q	19	66.28
					2366311-2381505	15.1	N	13	65.07
					3243202-3250738	7.5	N	10	65.25
***Mycobacterium tuberculosis* KZN 4207 [d] [g]**	South Africa	NZ_ACVS00000000	2	N.A.	1155088-1162615	7.5	N	12	65.25
					2021092-2037144	16	N	15	65.08
Mycobacterium	South Africa	NZ_ACVU00000000	3	N.A.	1151605-1159132	7.5	N	12	65.25

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***tuberculosis* KZN R506 [d]**					2013900-2029952	16	N	17	65.08
					2949867-2959279	9.4	N	14	66.92
***Mycobacterium tuberculosis* KZN V2475 [d]**	South Africa	NZ_ACVT00000000	3	N.A.	1147742-1155269	7.5	N	12	65.25
					2008496-2024548	16	N	17	65.08
					2943451-2952863	9.4	N	13	66.92
***Mycobacterium tuberculosis* NA-A0008 [d]**	Sputum from a patient infected with tuberculosis, Human, India	NZ_ALYG00000000	4	N.A.	1116822-1131718	14.8	N	15	64.02
					1465364-1480117	14.7	Q	17	65.38
					1956890-1967034	10.1	N	12	66.25
					4111401-4122982	11.5	N	16	66.48
***Mycobacterium tuberculosis* NA-A0009 [d]**	Sputum from a patient infected with tuberculosis, Human, India	NZ_ALYH00000000	3	N.A.	2605534-2614781	9.2	N	11	65.93
					2722301-2727905	5.6	N	7	66.49
					4127334-4137011	9.6	N	11	65.05
***Mycobacterium tuberculosis* NCGM2209 [d]**	N.D.	NZ_BADQ00000000	2	N.A.	3299583-3308800	9.2	N	13	66.62
					4252824-4279687	26.8	Q	42	69.28
Mycobacterium	Sputum,	NZ_AHHX00000000	5	N.A.	3546002-3559700	13.6	N	12	66.33

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***tuberculosis* OSDD071** [d]	Human, India				3967164-3981125	13.9	N	15	63.84
					4012377-4029249	16.8	N	22	65.53
					4185952-4213724	27.7	N	15	65.32
					4261104-4277115	16	N	22	71.26
***Mycobacterium tuberculosis* OSDD504** [d]	Sputum, Human, India	NZ_AHHY00000000	2	N.A.	4103956-4135263	31.3	N	23	66.32
					4207025-4225760	18.7	N	30	71.14
***Mycobacterium tuberculosis* OSDD518** [d]	Sputum, Human, India	NZ_AHHZ00000000	4	N.A.	778923-790183	11.2	N	11	64.29
					2189485-2215521	26	N	13	64.96
					4173009-4186875	13.8	N	20	66.23
					4225711-4237934	12.2	N	14	70.98
***Mycobacterium tuberculosis* PR05** [d]	Cerebrospinal fluid sample from tuberculosis patient, Human, Malaysia	NZ_AOMG00000000	2	N.A.	4220539-4237804	17.2	N	19	67.79
					4237971-4249673	11.7	N	16	67.76
Mycobacterium	N.D.	NC_017026	1	0.52%	2955788-2978471	22.6	Q	19	65.66

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***tuberculosis* RGTB327**									
***Mycobacterium tuberculosis* RGTB423**	N.D.	NC_017528	3	1.17%	1158796-1166333	7.5	N	12	65.23
					1766223-1787559	21.3	N	16	66.10
					2968585-2991092	22.5	Q	8	66.39
***Mycobacterium tuberculosis* SP21 [d]**	Sputum, Human, Russia	NZ_AOUF00000000	2	N.A.	3028440-3047275	18.8	N	19	66.33
					4226328-4247483	21.1	N	36	67.68
***Mycobacterium tuberculosis* SUMu001 [d]**	N.D.	NZ_ADHQ00000000	2	N.A.	1636588-1658107	21.5	N	10	66.10
					2828905-2858037	29.1	Q	16	66.22
***Mycobacterium tuberculosis* SUMu002 [d]**	N.D.	NZ_ADHR00000000	3	N.A.	2565992-2590251	24.2	Q	16	65.99
					3626157-3633628	7.4	N	10	65.24
					4266416-4292823	26.4	N	13	65.32
***Mycobacterium tuberculosis* SUMu003 [d]**	N.D.	NZ_ADHS00000000	2	N.A.	1145200-1152919	7.7	N	12	65.17
					2906967-2930905	23.9	Q	17	65.98
***Mycobacterium tuberculosis* SUMu004 [d]**	N.D.	NZ_ADHT00000000	2	N.A.	1287368-1298073	10.7	Q	14	64.82
					3312974-3336637	23.6	Q	14	65.68

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Mycobacterium tuberculosis* SUMu005** [d]	N.D.	NZ_ADHU00000000	2	N.A.	1407311-1417951	10.6	Q	14	64.81
					3181196-3190753	9.5	N	11	66.53
***Mycobacterium tuberculosis* SUMu006** [d]	N.D.	NZ_ADHV00000000	2	N.A.	1578121-1585439	7.3	N	10	65.10
					3360456-3379477	19	N	16	66.03
***Mycobacterium tuberculosis* SUMu007** [d]	N.D.	NZ_ADHW00000000	4	N.A.	1597704-1605570	7.8	N	10	65.08
					2231826-2241238	9.4	N	12	66.89
					3550134-3563489	13.1	N	16	67.63
					4265456-4291852	26.3	N	16	65.29
***Mycobacterium tuberculosis* SUMu008** [d]	N.D.	NZ_ADHX00000000	3	N.A.	1616154-1623602	7.4	N	10	65.19
					3375186-3395319	20.1	N	22	66.25
					4270109-4296538	26.4	N	17	65.13
***Mycobacterium tuberculosis* SUMu009** [d]	N.D.	NZ_ADHY00000000	3	N.A.	805722-813334	7.6	N	11	65.03
					2453948-2474856	20.9	N	19	66.04
					3357622-3383547	25.9	N	14	65.22
Mycobacterium	N.D.	NZ_ADHZ00000000	3	N.A.	804355-811998	7.6	N	11	65.15

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***tuberculosis* SUMu010** [d]					1403436-1424954	21.5	N	10	66.10
					2603946-2617822	13.8	Q	20	65.94
***Mycobacterium tuberculosis* SUMu011** [d]	N.D.	NZ_ADIA00000000	3	N.A.	805458-812801	7.3	N	10	65.18
					2593480-2613469	19.9	Q	17	65.87
					3347480-3360445	12.9	N	13	62.72
***Mycobacterium tuberculosis* SUMu012** [d]	N.D.	NZ_ADIB00000000	1	N.A.	2888632-2912934	24.3	Q	16	65.93
***Mycobacterium tuberculosis* T17** [d]	N.D.	NZ_ABQH00000000	3	N.A.	1734479-1757146	22.6	Q	15	66.05
					1998486-2019930	21.4	N	19	64.39
					2921915-2937415	15.5	N	15	64.89
***Mycobacterium tuberculosis* T46** [d]	N.D.	NZ_ACHO00000000	3	N.A.	1586190-1597798	11.6	N	15	65.41
					2189548-2211048	21.5	N	11	66.12
					2451905-2481538	29.6	Q	28	64.41
***Mycobacterium tuberculosis* T85** [d]	N.D.	NZ_ABOW00000000	2	N.A.	802334-809861	7.5	N	13	65.24
					2567273-2587102	19.8	N	14	65.71

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Mycobacterium tuberculosis* T92** [d]	N.D.	NZ_ABLN00000000	3	N.A.	1479054-1486581	7.5	N	12	65.22
					2315352-2344204	28.8	N	27	64.57
					3228210-3253026	24.8	N	14	65.55
***Mycobacterium tuberculosis* UM 1072388579** [d]	Sputum, Human, Malaysia	NZ_AMXW00000000	0	N.A.					
***Mycobacterium tuberculosis* W-148** [d]	N.D.	NZ_ACSX00000000	3	N.A.	2006793-2014239	7.4	N	7	63.15
					2341000-2350413	9.4	N	12	66.93
					2850476-2858003	7.5	N	13	65.22
***Mycobacterium tuberculosis* UT205**	N.D.	NC_016934	0	N.A.					
***Mycobacterium tusciae* JS617** [d]	Granular activated carbon, Germany	NZ_AGJJ00000000	7	N.A.	2439862-2460800	20.9	N	14	66.41
					2614214-2627979	13.7	N	8	68.10
					2652769-2661271	8.5	N	10	67.36
					2671348-2691458	20.1	N	21	68.01
					3302641-3323558	20.9	N	19	66.84

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
					5600545-5660573	60	N	59	66.26
					7083980-7115419	31.4	Y	47	63.38
***Mycobacterium ulcerans* Agy99** [f]	Clinical sample, Human, Ghana	NC_008611	2	0.37%	917436-929425	11.9	N	11	67.20
					3755296-3764098	8.8	N	10	64.97
***Mycobacterium vaccae* ATCC 25954** [d]	Human	NZ_ALQA00000000	1	N.A.	5506007-5526152	20.1	N	18	67.87
***Mycobacterium vanbaalenii* PYR-1**	Estuarine sediments polluted with petrol, United States	NC_008726	1		538064-548679	10.6	N	14	66.15
***Mycobacterium xenopi* RIVM700367** [d]	Human, Netherlands	NZ_AJFI00000000	3	N.A.	58832-68963	10.1	Q	10	64.34
					837197-847850	10.6	Y	11	64.23
					2447053-2492648	45.5	Y	50	65.63
***Mycobacterium yongonense* 05-1390** [f]	Pulmonary patient, male, 64 years old, Human, South Korea	NC_021715	3	0.89%	3542832-3554525	11.6	N	10	65.35
					3643049-3660989	17.9	N	15	62.15
					4993761-5013489	19.7	N	12	64.80

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
Nocardia abscessus **NBRC 100374** [d]	Joint abscess of a 56 year old man with a complete endoprosthesis of one of his knees, Germany	NZ_BAFP00000000	2	N.A.	7411413-7432989	21.5	N	35	66.14
					7441249-7466590	25.3	Y	29	66.14
Nocardia aobensis **NBRC 100429** [d f]	Human, Japan	NZ_BAFQ00000000	1	N.A.	5984929-5998119	13.1	Q	12	65.60
Nocardia araoensis **NBRC 100135** [d]	Human, Japan	NZ_BAFR00000000	2	N.A.	4491835-4508037	16.2	N	10	65.64
					6778720-6806191	27.4	Y	16	65.44
Nocardia asiatica **NBRC 100129** [d]	Human, Japan	NZ_BAFS00000000	2	N.A.	6230003-6239024	9	N	15	66.80
				N.A.	7553518-7567749	14.2	N	19	65.74
Nocardia asteroides **NBRC 15531** [d]	N.D.	NZ_BAFO00000000	1	N.A.	6464139-6483706	19.5	N	12	68.21
Nocardia brasiliensis **ATCC 700358**	Human mycetoma, Mexico	NC_018681	0	N.A.					
Nocardia brasiliensis	N.D.	NZ_BAFT00000000	0	N.A.					

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
NBRC 14402 [d]									
Nocardia brevicatena **NBRC 12119 [d]**	Sputa	NZ_BAFU00000000	2	N.A.	184108-194293	10.1	N	14	66.24
					5608530-5639170	30.6	Y	36	65.31
Nocardia carnea **NBRC 14403 [d]**	N.D.	NZ_BAFV00000000	0	N.A.					
Nocardia cerradoensis **NBRC 101014 [d]**	Soil, Brazil	NZ_BAFW00000000	0	N.A.					
Nocardia concava **NBRC 100430 [d]**	Human, Japan	NZ_BAFX00000000	3	N.A.	4237606-4247437	9.8	N	7	67.21
					7713017-7726248	13.2	Q	16	66.54
					7715411-7742576	27.1	N	26	66.73
Nocardia cyriacigeorgica **GUH-2**	N.D.	NC_016887	0	N.A.					
Nocardia cyriacigeorgica **NBRC 100375 [d]**	Bronchial secretions, Germany	NZ_BAFY00000000	0	N.A.					
Nocardia exalbida **NBRC 100660 [d]**	A bronchoalveolar lavage of a 43-year-old	NZ_BAFZ00000000	1	N.A.	6520306-6528930	8.6	N	9	67.08

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
	immunocompromised patient with a lung abscess, China								
***Nocardia farcinica* IFM 10152** [f]	N.D.	NC_006361	4	1.93%	28359-50108	21.7	Y	24	69.93
					1669018-1707866	38.8	Y	54	67.47
					4103113-4143292	40.1	Y	41	65.33
					4136309-4151800	15.4	N	24	64.19
***Nocardia higoensis* NBRC 100133** [d]	Human, Japan	NZ_BAGA00000000	0	N.A.					
***Nocardia jiangxiensis* NBRC 101359** [d]	Rhizosphere soil (about pH 3.5) of goose-grass (*Elusine indica*) growing next to a copper mine, China	NZ_BAGB00000000	2	N.A.	7163124-7176923	13.8	N	30	58.91
					7209365-7238755	29.3	Q	33	59.68
***Nocardia niigatensis* NBRC 100131** [d]	Human, Japan	NZ_BAGC00000000	1	N.A.	236676-267140	30.4	Y	36	66.87

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Nocardia otitidiscaviarum* NBRC 14405** [d]	Ear of guinea pig	NZ_BAGD00000000	1	N.A.	6425537-6434991	9.4	N	17	66.38
***Nocardia paucivorans* NBRC 100373** [d]	Sputum, Germany	NZ_BAGE00000000	1	N.A.	2423284-2437578	14.2	N	16	64.68
***Nocardia pneumoniae* NBRC 100136** [d]	Human, Japan	NZ_BAGF00000000	0	N.A.					
***Nocardia takedensis* NBRC 100417** [d]	Moat sediment, Japan	NZ_BAGG00000000	3	N.A.	5767857-5774953	7.0	N	8	70.10
					5894626-5916891	22.2	Y	28	67.16
					6181247-6203580	22.3	N	6	67.53
***Nocardia tenerifensis* NBRC 101015** [d]	Soil, Canary Islands	NZ_BAGH00000000	1	N.A.	8418932-8432932	14.0	Y	20	63.04
***Nocardia testacea* NBRC 100365** [d]	Human, Japan	NZ_BAGJ00000000	1	N.A.	1435697-1457180	21.4	Q	21	67.25
***Nocardia thailandica* NBRC 100428** [d]	Human, Thailand	NZ_BAGK00000000	4	N.A.	1008436-1069196	60.7	Q	68	70.64
					3593195-3613718	20.5	Q	17	66.98
					3612039-3624898	12.8	N	19	67.93
					3628939-3643106	14.1	N	28	67.04

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Nocardia veterana* NBRC 100344 [d]**	Bronchioscopic lavage, Australia	NZ_BAGM00000000	0	N.A.					
***Nocardia vinacea* NBRC 16497 [d]**	Soil, Japan	NZ_BAGN00000000	0	N.A.					
***Rhodococcus equi* 103S**	*Equus caballus*, Canada	NC_014659	0	N.A.					
***Rhodococcus equi* ATCC 33707 [d]**	Human abscess and/or skin	NZ_ADNW00000000	2	N.A.	1795183-1844172	48.9	Y	57	66.93
					2174691-2219151	44.4	Y	72	65.57
***Rhodococcus erythropolis* CCM2595 (ATCC 11048) [f]**	Soil	NC_022115	1	0.18%	4660425-4671424	11	N	8	59.09
***Rhodococcus erythropolis* PR4 [f]**	Seawater, Pacific Ocean	NC_012490	3	0.88%	2445447-2470489	25	Q	30	61.96
					4351039-4373917	22.8	N	9	61.35
					4936069-4945592	9.5	N	7	58.63
***Rhodococcus erythropolis* SK121 [d]**	N.D.	NZ_ACNO00000000	2	N.A.	3024554-3041304	16.7	N	7	61.01
					6558023-6563559	5.5	N	7	61.13

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Rhodococcus imtechensis* RKJ300** [d]	Soil, India	NZ_AJJH00000000	3	N.A.	280337-303673	23.3	Q	10	65.51
					5386240-5405206	18.9	N	14	64.01
					5775690-5782910	7.2	N	9	66.72
***Rhodococcus jostii* RHA1** [f]	N.D.	NC_008268	1	0.22%	7055617-7072621	17	N	12	65.65
***Rhodococcus opacus* B4** [f]	N.D.	NC_012522	3	0.76%	320326-329656	9.3	N	11	67.59
					4321619-4339482	17.8	N	28	63.64
					4335275-4368502	33.2	Y	31	63.63
***Rhodococcus opacus* M213** [d]	Fuel-oil contaminated soil, United States	NZ_AJYC00000000	1	N.A.	1919282-1961835	42.5	Y	57	65.47
***Rhodococcus opacus* PD630** [d]	N.D.	NZ_AGVD00000000	0	N.A.					
***Rhodococcus pyridinivorans* AK37** [d]	N.D.	NZ_AHBW00000000	2	N.A.	294536-327014	32.4	Y	48	66.44
					4926772-4948625	21.8	Q	21	67.82
***Rhodococcus pyridinivorans* SB3094** [f]	N.D.	NC_023150	0	N.A.					

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Rhodococcus qingshengii* BKS 20-40 [d]**	Soil, India	NZ_AODN00000000	0	N.A.					
***Rhodococcus rhodochrous* BKS6-46 [d]**	Mangrove forest soil sample, India	NZ_AGVW00000000	3	N.A.	124130-146400	22.2	N	27	64.82
					216568-231833	15.2	N	17	69.28
					4675363-4693294	17.9	N	23	66.96
***Rhodococcus ruber* BKS 20-38 [d]**	Soil, India	NZ_AOEX00000000	4	N.A.	393201-401238	8	N	8	70.09
					2225972-2254658	28.6	N	15	72.00
					3534043-3539655	5.6	N	6	66.29
					5184902-5231057	46.1	Q	44	66.59
***Rhodococcus ruber* Chol-4 [d]**	Sewage sludge, Spain	NZ_ANGC00000000	0	N.A.					
***Rhodococcus* sp. AW25M09 [d]**	N.D.	NZ_CAPS00000000	0	N.A.					
***Rhodococcus* sp. DK17 [d]**	Crude oil contaminated soil, Korea	NZ_AJLQ00000000	0	N.A.					
***Rhodococcus* sp. JVH1 [d]**	Canada	NZ_AKKP00000000	1	N.A.	1529730-1569966	40.2	N	46	64.74

Organism	Isolation source [c]	Accession/GI number	No. putative prophage region(s)	Percentage of genome encoding putative prophage region(s)[e]	Coordinates within host genome [a]	Size (kbp)	Intact [b]	No. CDS	G+C mol % of prophage region
***Rhodococcus* sp. P14** [d]	Crude oil contaminated sediments, China	NZ_AJFC00000000	0	N.A.					
***Rhodococcus* sp. R1101** [d]	Cavitary lung lesion, Human, United States	NZ_AJVB00000000	0	N.A.					
***Rhodococcus triatomae* BKS 15-14** [d]	Soil, India	NZ_AODO00000000	1	N.A.	5554117-5592017	37.9	Y	38	64.55
***Rhodococcus wratislaviensis* IFP 2016** [d]	France	NZ_ANIU00000000	2	N.A.	2724077-2749023	24.9	N	11	64.55
					4251757-4264455	12.6	N	10	67.19
***Tsukamurella paurometabola* DSM 20162** [f]	N.D.	NC_014158	1	0.49%	1490234-1511515	21.2	Q	21	65.13

*[a] Coordinates for draft sequences are based on concatenation of all contigs by PHAST; [b] Y indicates Yes (putative complete predicted prophage); N indicates No (putative incomplete predicted prophage); Q indicates Questionable (completeness of putative predicted prophage is questionable); [c] Isolation data obtained from GenBank and ATCC, DSMZ, CCUG, NTCC culture collection websites; [d] Draft genome sequence in multiple contigs; [e] Calculated where bacterial genome sequences appear complete; [f] Corresponding plasmids' screened (*Table 5.3*); [g] Multiple assemblies screened and listed individually; N.D. indicates no data available; N.A indicates not applicable.*

Table 5.3 Prophages detected in plasmid sequences using PHAST

Plasmid(s) screened using PHAST					**PHAST predicted prophage(s)**					
Source organism	**Plasmid**	**Accession number**	**Plasmid size (bp)**	**Genome structure**	**No. prophage regions**	**Intact [a]**	**Coordinates [b]**	**Size (kbp)**	**No. CDS**	**G+C content (% mol)**
***Gordonia bronchialis* DSM 43247**	pGBRO01	NC_013442	81,410	Circular	0					
***Gordonia* sp. KTR9**	pGKT2	NC_018580	182,454	Circular	0					
	pGKT1	NC_018582	89,480	Circular	0					
	pGKT3	NC_018583	172,385	Circular	1	Y	98742-134788		14	61.13
***Gordonia polyisoprenivorans* VH2**	p174	NC_016907	174,494	Circular	0					
***Gordonia westfalica* strain DSM44215T**	pKB1	NC_005307	101,016	Circular	0					
***Mycobacterium abscessus* subsp. bolletii 50594**	Plasmid 1	NC_021278	172,814	Circular	0					
	Plasmid 2	NC_021279	97,240	Circular	0					
***Mycobacterium* abscessus subsp. bolletii CRM-0020**	plasmid unnamed	NZ_ATFQ01000044	56,466	Linear	0					
***Mycobacterium abscessus* subsp. bolletii F1725**	BRA100	NC_017908	56,265	Circular	0					
***Mycobacterium abscessus* ATCC 19977**	Unnamed plasmid	NC_010394	23,319	Circular	0					

Plasmid(s) screened using PHAST					PHAST predicted prophage(s)					
Mycobacterium avium	pVT2	NC_005016	12,868	Circular	0					
Mycobacterium celatum	pCLP	NC_004963	22,688	Linear	0					
***Mycobacterium chubuense* NBB4**	pMYCCH.02	NC_018023	143,623	Circular	0					
	pMYCCH.01	NC_018022	615,278	Circular	1	N	444996-456797	11.8	10	65.30
Mycobacterium fortuitum	pAL5000	NC_001381	4,837	Circular	0					
***Mycobacterium gilvum* PYR-GCK**	pMFLV01	NC_009339	321,253	Linear	1	Q	164351-215973	51.6	23	65.37
	pMFLV02	NC_009340	25,309	Circular	0					
	pMFLV03	NC_009341	16,660	Circular	0					
***Mycobacterium gilvum* Spyr1**	pMSPYR101	NC_014811	211,864	Circular	0					
	pMSPYR102	NC_014812	23,681	Circular	0					
***Mycobacterium kansasii* ATCC 12478**	pMK12478	NC_022654	144,951	Circular	0					
***Mycobacterium* sp. KMS**	pMKMS02	NC_008704	216,763	Circular	1	Y	174491-216428	41.9	28	65.03
	pMKMS0	NC_008703	302,089	Circular	0					

Plasmid(s) screened using PHAST					PHAST predicted prophage(s)					
	1									
***Mycobacterium liflandii* 128FXT**	pMUM002	NC_011355	190,588	Circular	0					
***Mycobacterium marinum* DL240490**	pMUM003	NC_019018	104,530	Linear	0					
***Mycobacterium marinum* M**	pMM23	NC_010604	23,317	Circular	0					
***Mycobacterium* sp. MCS**	Plasmid 1	NC_008147	215,075	Linear	1	Y	149207-183300	34	35	63.86
***Mycobacterium smegmatis* JS623**	pMYCSM01	NC_019957	394,147	Circular	0					
	pMYCSM02	NC_019958	198,589	Circular	0					
	pMYCSM03	NC_019959	164,114	Circular	0					
***Mycobacterium ulcerans* AGY99**	pMUM001	NC_005916	174155	Circular	1	Y	138963-160330	21.3	24	61.36
***Mycobacterium yongonense* 05-1390**	pMyong1	NC_020275	122,976	Circular	0					
	pMyong2	NC_020276	18,089	Circular	0					
Nocardia aobensis	pYS1	NC_013448	4,326	Circular	0					
***Nocardia farcinica* IFM 10152**	pNF1	NC_006362	184,026	Circular	0					

Plasmid(s) screened using PHAST					PHAST predicted prophage(s)
	pNF2	NC_006363	87,093	Circular	0
***Nocardia* sp. 107**	pXT107	NC_010874	4,335	Circular	0
***Nocardia* sp. C-14-1**	pC1	NC_013538	5,841	Circular	0
***Rhodococcus aetherivorans* I24**	pRA1	NC_010882	9,372	Circular	0
***Rhodococcus equi* 103**	p103	NC_002576	80,609	Circular	0
***Rhodococcus equi* ATCC33701**	pREAT701	NC_004854	80,610	Circular	0
Rhodococcus equi	pVAPB1593	NC_011150	79,251	Circular	0
	pVAPA1037	NC_011151	80,610	Circular	0
	pVAPAMBE116	NC_014247	83,100	Circular	0
Rhodococcus erythropolis	pBD2	NC_005073	210,205	Linear	0
	pFAJ2600	NC_003846	5,936	Circular	0
	pRE8424	NC_006258	5,987	Circular	0
***Rhodococcus erythropolis* CCM2595**	pRECF1	NC_022125	90,223	Circular	0
Rhodococcus *erythropolis* PR4	pREL1	NC_007491	271,577	Linear	0
	pREC1	NC_007486	104,014	Circular	0

Plasmid(s) screened using PHAST					PHAST predicted prophage(s)					
	pREC2	NC_007487	3,637	Circular	0					
***Rhodococcus fascians* D188**	pFiD188	NC_021080	198,917	Linear	0					
***Rhodococcus jostii* RHA1**	pRHL1	NC_008269	1,123,075	Linear	2	Q	452661-481845	29.1	25	65.81
						N	570289-600035	29.7	15	63.16
	pRHL2	NC_008270	442,536	Linear	0					
	pRHL3	NC_008271	332,361	Linear	0					
***Rhodococcus opacus* B4**	pKNR01	NC_006969	4,367	Circular	0					
	pKNR02	NC_006970	2,773	Circular	0					
	pROB01	NC_012520	558,192	Linear	0					
	pROB02	NC_012521	244,997	Linear	1	Y	63094-91700	28.6	14	65.14
	pKNR	NC_012523	111,160	Circular	0					
***Rhodococcus pyridinivorans* SB3094**	plasmid	NC_023144	361,397	Circular	0					
	plasmid	NC_023145	2,035	Circular	0					
Rhodococcus rhodochrous	pNC500	NC_008823	7,637	Circular	0					
***Rhodococcus* sp. B264-1**	pB264	NC_004900	4,970	Circular	0					
***Rhodococcus* sp. BCP1**	pBMC1	NZ_CM002178	120,373	Linear	0					

Plasmid(s) screened using PHAST					PHAST predicted prophage(s)
	pBMC2	NZ_CM002179	103,129	Linear	0
***Rhodococcus* sp. NS1**	pNSL1	NC_010850	117,252	Linear	0
***Tsukamurella paurometabola* DSM 20162**	pTpau01	NC_014159	99,806	Circular	0

[a] Y indicates Yes, N indicates No, Q indicates incomplete; [b] Coordinates for draft sequences are based on concatenation of all contigs by PHAST.

Table 5.4 Genome annotations of temperate phages GAL1, GMA1, and TPA4

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
GAL1-*orf1*	51..593	181	-			
GAL1-*orf2*	720..1169	150	hypothetical protein GOALK_093_00290 [*Gordonia alkanivorans* NBRC 16433]	100	1e-101	
GAL1-*orf3*	1300..1620	107	hypothetical protein [*Gordonia alkanivorans*]	100	5e-63	Putative small terminase
GAL1-*orf4*	1617..3242	542	terminase [*Gordonia alkanivorans*]	99	0.0	Large terminase (COG4626)
GAL1-*orf5*	3245..4624	460	hypothetical protein [*Gordonia alkanivorans*]	100	0.0	Portal protein (pfam05133)
GAL1-*orf6*	4710..5651	314	hypothetical protein [*Gordonia alkanivorans*]	100	0.0	
GAL1-*orf7*	5721..6392	224	hypothetical protein [*Gordonia alkanivorans*]	99	1e-144	
GAL1-*orf8*	6396..6800	135	putative K structural protein [*Gordonia alkanivorans*]	100	7e-84	Bacteriophage lambda head decorator protein (pfam02924)
GAL1-*orf9*	6815..7852	346	putative phage structural protein [*Gordonia alkanivorans*]	100	0.0	Phage major capsid protein E (pfam03864)

ORF[a]	Coordinates	Size (aa)	Significant match [b]			% identity [c]	E value [d]	Putative function (conserved motif) [e]
GAL1-*orf10*	7852..8169	106	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	2e-62	
GAL1-*orf11*	8169..8561	131	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	1e-81	
GAL1-*orf12*	8558..8902	115	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	2e-68	
GAL1-*orf13*	8895..9233	113	hypothetical *alkanivorans*]	protein	[*Gordonia*	99	4e-66	
GAL1-*orf14*	9230..9691	154	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	1e-97	
GAL1-*orf15*	9773..10699	309	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	0.0	Putative major tail protein
GAL1-*orf16*	10786..11244	153	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	2e-95	Putative tail assembly protein
GAL1-*orf17*	11226..11669	148	hypothetical *alkanivorans*]	protein	[*Gordonia*	99	9e-84	Putative tail assembly protein translated by conserved programmed translational frameshift
GAL1-*orf18*	11671..16068	1466	hypothetical *alkanivorans*]	protein	[*Gordonia*	99	0.0	Putative tape measure protein (pfam01576; COG5412)

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
GAL1-*orf19*	16068..16979	304	hypothetical protein [*Gordonia alkanivorans*]	100	0.0	
GAL1-*orf20*	16985..18748	588	hypothetical protein [*Gordonia alkanivorans*]	99	0.0	
GAL1-*orf21*	18745..19113	123	hypothetical protein [*Gordonia alkanivorans*]	99	4e-72	
GAL1-*orf22*	19182..20282	367	hypothetical protein [*Gordonia alkanivorans*]	100	0.0	Lysin (pfam01510; pfam08310)
GAL1-*orf23*	20309..20548	80	holin [*Gordonia alkanivorans*]	100	1e-41	Putative Holin
GAL1-*orf24*	20564..20953	130	membrane protein [*Gordonia alkanivorans*]	100	3e-79	
GAL1-*orf25*	20950..21339	130	hypothetical protein [*Gordonia alkanivorans*]	99	2e-74	
GAL1-*orf26*	21339..22043	235	hypothetical protein [*Gordonia alkanivorans*]	100	3e-153	PE-PPE domain (pfam08237)
GAL1-*orf27*	22156..22695	180	hypothetical protein [*Gordonia alkanivorans*]	99	4e-118	HNH endonuclease (pfam13392; pfam07463)
GAL1-*orf28*	22722..24629	636	hypothetical protein [*Gordonia*	99	0.0	

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
			alkanivorans]			
GAL1-*orf29*	24641..26770	710	hypothetical protein [*Gordonia alkanivorans*]	100	0.0	
GAL1-*orf30*	26783..27679	299	hypothetical protein [*Gordonia alkanivorans*]	100	0.0	
GAL1-*orf31*	28103..29317	405	putative recombinase [*Gordonia alkanivorans*]	100	0.0	Integrase (pfam00589; pfam14659)
GAL1-*orf32*	29450..29785	112	hypothetical protein GOALK_093_00780 [*Gordonia alkanivorans* NBRC 16433]	98	1e-35	
GAL1-*orf33*	complement(29687..30364)	226	hypothetical protein [*Gordonia alkanivorans*]	99	1e-147	HTH DNA binding (pfam12844)
GAL1-*orf34*	30600..31448	283	putative phage protein [*Gordonia alkanivorans*]	100	0.0	Rha phage regulatory protein (pfam09669)
GAL1-*orf35*	31502..31741	80	hypothetical protein [*Gordonia alkanivorans*]	99	8e-44	HTH DNA binding (pfam01381)
GAL1-*orf36*	31738..31854	39	-			
GAL1-*orf37*	31854..31991	46	hypothetical protein [*Gordonia alkanivorans*]	100	2e-22	

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
GAL1-*orf38*	32000..32320	107	hypothetical protein [*Gordonia alkanivorans*]	100	8e-65	
GAL1-*orf39*	32334..33173	280	hypothetical protein [*Gordonia alkanivorans*]	99	1e-179	Phage antirepressor protein (pfam02498; pfam03374)
GAL1-*orf40*	33170..33433	88	hypothetical protein [*Gordonia alkanivorans*]	99	7e-50	
GAL1-*orf41*	33430..33756	109	putative Xre family DNA-binding protein [*Gordonia alkanivorans*]	100	7e-646	HTH DNA binding (pfam13560)
GAL1-*orf42*	33753..33974	74	hypothetical protein [*Gordonia alkanivorans*]	99	1e-40	HTH DNA binding (pfam12728)
GAL1-*orf43*	33971..34180	70	hypothetical protein [*Gordonia alkanivorans*]	100	2e-37	
GAL1-*orf44*	34177..34359	61	hypothetical protein [*Gordonia alkanivorans*]	100	2e-33	
GAL1-*orf45*	34356..34607	84	hypothetical protein [*Gordonia alkanivorans*]	100	2e-49	
GAL1-*orf46*	34604..34783	60	hypothetical protein [*Gordonia alkanivorans*]	100	5e-32	

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
GAL1-*orf47*	34770..35021	84	putative protocatechuate 3,4-dioxygenase subunit beta [*Gordonia alkanivorans*]	99	1e-45	Putative protocatechuate 3,4-dioxygenase subunit beta
GAL1-*orf48*	35018..35191	58	-			
GAL1-*orf49*	35188..35370	61	hypothetical protein [*Gordonia alkanivorans*]	100	2e-34	
GAL1-*orf50*	35370..35588	73	putative pyruvate phosphate dikinase [*Gordonia alkanivorans*]	100	5e-42	Putative pyruvate phosphate dikinase
GAL1-*orf51*	35588..36184	199	hypothetical protein [*Gordonia alkanivorans*]	100	2e-128	
GAL1-*orf52*	36181..36321	47	-			
GAL1-*orf53*	36318..37556	413	putative WhiB family regulatory protein [*Gordonia alkanivorans*]	100	0.0	WhiB (pfam02467)
GAL1-*orf54*	37553..37900	116	hypothetical protein [*Gordonia alkanivorans*]	100	1e-71	
GAL1-*orf55*	38002..38268	89	hypothetical protein [*Gordonia alkanivorans*]	98	3e-29	
GAL1-*orf56*	38270..38929	220	putative methyltransferase [*Gordonia*	100	1e-144	DNA methyltransferase (pfam01555;

ORF[a]	Coordinates	Size (aa)	Significant match [b]			% identity [c]	E value [d]	Putative function (conserved motif) [e]
			alkanivorans]					COG0863)
GAL1-*orf57*	38966..39451	162	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	2e-101	
GAL1-*orf58*	39448..39708	87	hypothetical *alkanivorans*]	protein	[*Gordonia*	99	5e-49	
GAL1-*orf59*	39705..39854	50	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	5e-25	
GAL1-*orf60*	39851..40093	81	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	2e-47	
GAL1-*orf61*	40090..40473	128	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	2e-77	
GAL1-*orf62*	40466..40969	168	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	6e-104	
GAL1-*orf63*	40966..41148	61	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	4e-33	
GAL1-*orf64*	41145..41696	184	methyltransferase *alkanivorans*]		[*Gordonia*	100	4e-121	DNA N-6-adenine-methyltransferase (pfam05869)
GAL1-*orf65*	42011..42133	41	-					

ORF[a]	Coordinates	Size (aa)	Significant match [b]			% identity [c]	E value [d]	Putative function (conserved motif) [e]
GAL1-*orf66*	42130..42576	149	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	1e-94	
GAL1-*orf67*	42573..43475	301	hypothetical *alkanivorans*]	protein	[*Gordonia*	99	0.0	
GAL1-*orf68*	43472..43843	124	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	1e-76	
GAL1-*orf69*	43840..44205	122	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	2e-74	Endodeoxyribonuclease (pfam05866)
GAL1-*orf70*	44202..44411	70	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	1e-38	
GAL1-*orf71*	44515..45402	296	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	0.0	Unknown (DUF3310)
GAL1-*orf72*	45399..45611	71	hypothetical *alkanivorans*]	protein	[*Gordonia*	99	2e-40	
GAL1-*orf73*	45602..45817	72	hypothetical *alkanivorans*]	protein	[*Gordonia*	100	3e-40	
GAL1-*orf74*	45814..46038	75	hypothetical *alkanivorans*]	protein	[*Gordonia*	99	2e-43	

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
GAL1-*orf75*	46097..46876	260	hypothetical protein [*Gordonia alkanivorans*]	100	1e-173	
GAL1-*orf76*	47220..47414	65	hypothetical protein [*Gordonia alkanivorans*]	100	5e-34	
GAL1-*orf77*	47436..47828	131	hypothetical protein [*Gordonia alkanivorans*]	100	2e-80	Unknown (pfam11750)
GAL1-*orf78*	47825..48118	98	hypothetical protein [*Gordonia alkanivorans*]	100	7e-59	
GAL1-*orf79*	48135..48575	147	putative gluconate 2-dehydrogenase (acceptor) [*Gordonia alkanivorans*]	100	2e-92	Putative gluconate 2-dehydrogenase
GAL1-*orf80*	48654..49394	247	hypothetical protein [*Gordonia alkanivorans*]	100	5e-164	HNH endonuclease (pfam01844)
GAL1-*orf81*	49391..49576	62	hypothetical protein [*Gordonia alkanivorans*]	98	7e-34	
GAL1-*orf82*	49576..49935	120	hypothetical protein [*Gordonia alkanivorans*]	100	3e-75	Restriction endonuclease (COG1403)
GMA1-*orf1*	78..440	121	hypothetical protein EN35_19995 [*Rhodococcus qingshengii*]	39	4e-10	Putative small terminase subunit

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
GMA1-*orf2*	437..1684	416	terminase large subunit [*Rhodococcus equi*]	60	4e-172	Large terminase subunit (pfam04466)
GMA1-*orf3*	1681..3249	523	hypothetical protein [*Rhodococcus* sp. UNC363MFTsu5.1]	46	2e-136	Phage portal protein (pfam05133)
GMA1-*orf4*	3255..4343	363	phage minor capsid protein 2 [*Streptomyces globisporus*]	40	1e-71	Phage minor capsid protein (pfam06152)
GMA1-*orf5*	4631..5179	183	hypothetical protein [*Rhodococcus* sp. UNC363MFTsu5.1]	47	1e-29	
GMA1-*orf6*	5231..6211	327	phage capsid protein [*Rhodococcus* sp. UNC363MFTsu5.1]	69	3e-150	Putative phage capsid protein
GMA1-*orf7*	6211..6375	55	hypothetical protein [*Streptomyces* sp. NRRL WC-3795]	53	3e-08	
GMA1-*orf8*	6454..6870	139	hypothetical protein EN35_19955 [*Rhodococcus qingshengii*]	55	2e-36	
GMA1-*orf9*	6867..7187	107	hypothetical protein [*Rhodococcus* sp. UNC363MFTsu5.1]	44	3e-17	
GMA1-*orf10*	7189..7524	112	hypothetical protein [*Corynebacterium aurimucosum*]	41	1e-17	

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
GMA1-*orf11*	7533..7808	92	-			
GMA1-*orf12*	7805..8254	150	hypothetical protein QR64_00255 [*Rhodococcus* sp. Chr-9]	42	2e-21	
GMA1-*orf13*	8266..8748	161	hypothetical protein [*Rhodococcus fascians*]	42	4e-35	Putative major tail protein
GMA1-*orf14*	8784..9359	192	hypothetical protein [*Rhodococcus fascians*]	38	8e-23	Putative tail assembly protein
GMA1-*orf15*	9341..9703	121	hypothetical protein [*Rhodococcus* sp. p52]	52	2e-16	Putative tail assembly protein translated by conserved programmed translational frameshift
GMA1-*orf16*	complement(9700..10146)	149	-			
GMA1-*orf17*	10196..15121	1642	hypothetical protein [*Rhodococcus fascians*]	32	4e-104	Tape measure protein (pfam06737)
GMA1-*orf18*	15118..15930	271	hypothetical protein [*Rhodococcus* sp. 29MFTsu3.1]	29	8e-28	
GMA1-*orf19*	15927..17480	518	hypothetical protein [*Rhodococcus* sp. p52]	44	6e-121	

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
GMA1-*orf20*	17561..18616	352	hypothetical protein [*Gordonia malaquae*]	53	4e-115	Lysin (pfam01510; pfam08310 X2)
GMA1-*orf21*	18613..18861	83	holin [*Dietzia alimentaria*]	58	3e-20	Putative holin
GMA1-*orf22*	18858..19307	150	gp15 [*Mycobacterium* phage Dori]	38	3e-16	
GMA1-*orf23*	19297..19617	107	-			
GMA1-*orf24*	19617..20363	249	hypothetical protein [*Gordonia sihwensis*]	48	2e-49	
GMA1-*orf25*	20374..20823	150	hypothetical protein [*Gordonia sihwensis*]	50	9e-37	
GMA1-*orf26*	20820..21035	72	hypothetical protein [*Gordonia sihwensis*]	41	1e-06	
GMA1-*orf27*	21102..21533	144	hypothetical protein [*Rhodococcus equi*]	63	1e-58	
GMA1-*orf28*	21554..22813	420	hypothetical protein [*Gordonia soli*]	41	3e-24	
GMA1-*orf29*	22915..23067	51	hypothetical protein [*Gordonia malaquae*]	91	3e-21	
GMA1-*orf30*	complement(23064..24152)	363	site-specific recombinase XerD [*Mycobacterium smegmatis*]	48	2e-106	Phage integrase (COG0582)

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
GMA1-*orf31*	complement(24478..24915)	146	hypothetical protein [*Gordonia malaquae*]	99	7e-69	
GMA1-*orf32*	complement(24955..25266)	104	hypothetical protein [*Glycomyces arizonensis*]	35	2e-05	
GMA1-*orf33*	complement(25267..25485)	73	hypothetical protein [*Nocardiopsis dassonvillei*]	53	4e-18	
GMA1-*orf34*	complement(25482..26489)	336	hypothetical protein [*Gordonia malaquae*]	64	1e-23	
GMA1-*orf35*	26567..26794	76	gp41 [*Mycobacterium* phage PMC]	67	1e-16	HTH DNA binding (pfam01381)
GMA1-*orf36*	26839..27690	284	putative phage protein [*Gordonia alkanivorans*]	87	2e-78	BRO family N-terminal protein (pfam02498)
GMA1-*orf37*	27722..28042	107	hypothetical protein [*Gordonia malaquae*]	44	2e-09	
GMA1-*orf38*	28039..28158	40	-			
GMA1-*orf39*	28155..28403	83	-			
GMA1-*orf40*	28400..28561	54	-			
GMA1-*orf41*	28629..28913	95	hypothetical protein [*Gordonia malaquae*]	94	6e-55	

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
GMA1-*orf42*	28910..29065	52	hypothetical protein [*Gordonia malaquae*]	88	2e-18	
GMA1-*orf43*	29052..29600	183	hypothetical protein [*Gordonia malaquae*]	73	5e-77	
GMA1-*orf44*	29606..29722	39	hypothetical protein [*Gordonia malaquae*]	93	1e-08	
GMA1-*orf45*	29719..29916	66	-			
GMA1-*orf46*	29913..30710	266	hypothetical protein [*Gordonia malaquae*]	97	0.0	Exonuclease (PRK09709)
GMA1-*orf47*	30722..31246	175	hypothetical protein [*Gordonia malaquae*]	100	2e-124	Endonuclease (pfam07463; pfam13392)
GMA1-*orf48*	31243..32121	293	hypothetical protein [*Gordonia malaquae*]	98	0.0	
GMA1-*orf49*	32118..32558	147	transcription factor WhiB [*Gordonia sputi*]	43	9e-09	Transcription factor for WhiB (pfam02467)
GMA1-*orf50*	32555..32782	76	-			
GMA1-*orf51*	32775..32900	42	-			
GMA1-*orf52*	32897..33145	83	gp86 [*Mycobacterium* phage Gumball]	47	1e-10	

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
GMA1-*orf53*	33138..33686	183	hypothetical protein [*Rhodococcus* sp. P27]	51	2e-16	
GMA1-*orf54*	33679..33984	102	hypothetical protein [*Gordonia malaquae*]	86	6e-37	
GMA1-*orf55*	complement(33988..34329)	114	hypothetical protein [*Gordonia malaquae*]	91	5e-55	
GMA1-*orf56*	complement(34377..34490)	38	hypothetical protein [*Streptomyces turgidiscabies*]	68	5e-05	
GMA1-*orf57*	complement(34544..34753)	70	-			
GMA1-*orf58*	34814..35008	65	-			
GMA1-*orf59*	35005..35238	78	hypothetical protein [*Gordonia sihwensis*]	44	3e-06	
GMA1-*orf60*	35328..35543	72	-			
GMA1-*orf61*	35540..35677	46	-			
GMA1-*orf62*	35817..37163	449	Gp65 [*Rhodococcus ruber*]	74	0.0	Helicase (pfam00271; pfam00176)
GMA1-*orf63*	37160..38131	324	Gp66 [*Rhodococcus ruber*]	80	0.0	Methylase (pfam01555)

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
GMA1-*orf64*	3807938118..38942	288275	hypothetical protein GSI01S_10_02210 [*Gordonia sihwensis* NBRC 108236]hypothetical protein GSI01S_10_02210 [*Gordonia sihwensis* NBRC 108236]	67	3e-121	
GMA1-*orf65*	39076..39228	51	-			
GMA1-*orf66*	39225..39875	217	hypothetical protein [*Gordonia sihwensis*]	44	4e-40	
GMA1-*orf67*	40045..40320	92	hypothetical protein [*Gordonia malaquae*]	71	6e-29	
GMA1-*orf68*	40382..40945	188	gp54 [*Mycobacterium* phage Mutaforma13]	47	2e-36	Endonuclease (pfam07463; pfam13392)
TPA4-*orf1*	75..497	141	hypothetical protein N505_0105320 [*Rhodococcus* sp. BCP1]	62	7e-49	Putative small terminase
TPA4-*orf2*	526..2238	571	terminase [*Rhodococcus* sp. BCP1]	83	0.0	Putative large terminase
TPA4-*orf3*	2235..3725	497	hypothetical protein [*Gordonia sihwensis*]	59	0.0	Portal protein (pfam05133)
TPA4-*orf4*	3732..4874	381	capsid maturation protease [*Mycobacterium* phage ZoeJ]	43	2e-46	Capsid maturation protease (cd13442)

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TPA4-*orf5*	4871..5194	108	-			
TPA4-*orf6*	5313..5945	211	hypothetical protein [*Gordonia sihwensis*]	42	2e-14	
TPA4-*orf7*	5997..6371	125	hypothetical protein [*Gordonia sihwensis*]	74	2e-59	
TPA4-*orf8*	6373..7398	342	hypothetical protein [*Gordonia sihwensis*]	68	5e-156	Major capsid protein (pfam03864)
TPA4-*orf9*	7567..7953	129	hypothetical protein [*Gordonia sihwensis*]	42	1e-20	
TPA4-*orf10*	7947..8351	135	hypothetical protein [*Mycobacterium abscessus*]	41	6e-21	
TPA4-*orf11*	8351..8638	96	hypothetical protein [*Mycobacterium abscessus*]	45	2e-11	
TPA4-*orf12*	8635..9036	134	hypothetical protein [*Mycobacterium abscessus*]	30	2e-04	
TPA4-*orf13*	9127..9984	286	hypothetical protein I544_3152 [*Mycobacterium abscessus* 103]	46	4e-73	
TPA4-*orf14*	complement(10033..10	165	hypothetical protein [*Dietzia alimentaria*]	61	1e-33	

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
	527)					
TPA4-*orf15*	10600..11064	155	hypothetical protein [*Mycobacterium* sp. 141]	46	5e-22	Putative tail assembly protein
TPA4-*orf16*	11046..11465	140	Hypothetical protein BB31_24245 [*Amcolatopsis lurida* NRRL 2430]	33	8e-05	Putative tail assembly protein translated by conserved programmed translational frameshift
TPA4-*orf17*	11473..16632	1720	hypothetical protein [*Mycobacterium* sp. 141]	37	2e-137	Tape measure protein (COG1196)
TPA4-*orf18*	16629..17726	366	putative gp22 [*Mycobacterium abscessus* MAB_110811_2726]	63	4e-164	
TPA4-*orf19*	17726..19462	579	putative gp23 [*Mycobacterium abscessus* MAB_110811_2726]	74	0.0	
TPA4-*orf20*	19528..19722	65	-			
TPA4-*orf21*	19719..20168	150	putative gp24 [*Mycobacterium abscessus* MAB_110811_2726]	50	2e-31	
TPA4-*orf22*	20165..21250	362	tail protein [*Mycobacterium abscessus*]	54	1e-120	Putative tail protein
TPA4-*orf23*	21250..21615	122	hypothetical protein [*Mycobacterium abscessus*]	59	2e-27	

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TPA4-*orf24*	21615..23492	626	putative structural protein [*Mycobacterium* phage WIVsmall]	39	2e-99	Putative structural protein
TPA4-*orf25*	23492..23755	88	-			
TPA4-*orf26*	23759..25357	533	putative structural protein [*Mycobacterium* phage WIVsmall]	28	1e-06	Putative structural protein
TPA4-*orf27*	25382..25834	151	hypothetical protein WIVsmall_58 [*Mycobacterium* phage WIVsmall]	74	1e-74	
TPA4-*orf28*	25836..26891	352	hypothetical protein [*Proteobacteria bacterium* JGI 0000113-L05]	51	5e-16	
TPA4-*orf29*	26891..27490	200	hypothetical protein [*Gordonia sihwensis*]	51	1e-09	
TPA4-*orf30*	27688..28482	265	hypothetical protein [*Nocardia* sp. BMG111209]	36	4e-19	PE-PPE domain (pfam08237)
TPA4-*orf31*	28502..29614	371	antigen 85 complex protein [*Rhodococcus rhodnii*]	54	2e-122	Lysin (pfam01510; pfam08310)
TPA4-*orf32*	29620..29796	59	-			
TPA4-*orf33*	29799..30185	129	hypothetical protein TM4_gp31 [*Mycobacterium* phage TM4]	31	4e-13	

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TPA4-*orf34*	30182..30700	173	hypothetical protein [*Gordonia malaquae*]	33	4e-13	
TPA4-*orf35*	complement(30775..30942)	56	hypothetical protein [*Gordonia soli*]	45	4e-05	
TPA4-*orf36*	31015..31287	91	-			
TPA4-*orf37*	complement(31376..32608)	411	Integrase [*Gordonia* sp. KTR9]	53	1e-134	Integrase (pfam14659; pfam00589)
TPA4-*orf38*	complement(32979..33404)	142	hypothetical protein RR21198_4189 [*Rhodococcus rhodochrous* ATCC 21198]	47	3e-17	
TPA4-*orf39*	complement(33675..33860)	62	-			
TPA4-*orf40*	complement(33956..34099)	48	-			
TPA4-*orf41*	34071..34304	78	hypothetical protein [*Dietzia* sp. UCD-THP]	55	1e-12	
TPA4-*orf42*	34301..34543	81	-			
TPA4-*orf43*	34540..34899	120	hypothetical protein [*Gordonia amarae*]	52	3e-17	
TPA4-*orf44*	34941..35237	99	-			

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TPA4-*orf45*	35234..35554	107	hypothetical protein [*Mycobacterium avium*]	33	6e-04	
TPA4-*orf46*	35551..36090	180	-			
TPA4-*orf47*	36087..36890	268	hypothetical protein PBI_CATDAWG_12 [*Mycobacterium* phage Catdawg]	41	7e-53	CRISPR-associated Cas4-like protein (PHA00619)
TPA4-*orf48*	36926..37252	109	hypothetical protein [*Mycobacterium* sp. UM_RHS]	59	2e-39	
TPA4-*orf49*	37276..38109	278	gp66 [*Mycobacterium* phage Dori]	56	6e-102	Unknown (pfam10065)
TPA4-*orf50*	38139..38771	211	hypothetical protein [*Salinispora pacifica*]	41	8e-24	
TPA4-*orf51*	38771..38914	48	-			
TPA4-*orf52*	38998..39378	127	-			
TPA4-*orf53*	39393..39623	77	hypothetical protein [*Nocardiopsis gilva*]	54	1e-12	
TPA4-*orf54*	39893..40261	123	hypothetical protein TPA2_gp62 [*Tsukamurella* phage TPA2]	57	7e-29	
TPA4-*orf55*	40258..40635	126	-			

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TPA4-*orf56*	40632..41162	177	hypothetical protein [*Mycobacterium* sp. UM_WWY]	39	1e-16	Exonuclease (cd06127)
TPA4-*orf57*	41159..41554	132	hypothetical protein [*Gordonia otitidis*]	67	3e-45	
TPA4-*orf58*	41855..41971	39	-			
TPA4-*orf59*	41974..42615	214	hypothetical protein PBI_DONOVAN_52 [*Mycobacterium* phage Donovan]	40	3e-32	
TPA4-*orf60*	42608..42820	71	-			
TPA4-*orf61*	42817..43089	91	-			
TPA4-*orf62*	43272..43652	127	hypothetical protein [*Mycobacterium abscessus*]	60	9e-33	
TPA4-*orf63*	43645..44205	187	hypothetical protein [*Nocardia nova*]	69	6e-14	
TPA4-*orf64*	44205..44390	62	-			
TPA4-*orf65*	complement(44357..44842)	162	-			
TPA4-*orf66*	44841..44984	48	-			
TPA4-*orf67*	44981..45280	100	-			

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TPA4-*orf68*	45280..45519	80	-			
TPA4-*orf69*	45516..45704	63	-			
TPA4-*orf70*	45701..45991	97	[*Mycobacterium* phage Dori]	49	2e-04	
TPA4-*orf71*	45988..46362	125	sporulation protein [*Caldanaerobacter subterraneus*]	44	1e-04	WhiA C-terminal HTH domain (pfam02650)
TPA4-*orf72*	46359..46703	115	WhiB family transcriptional regulator [*Corynebacterium efficiens*]	55	6e-24	WhiB transcription factor (pfam02467)
TPA4-*orf73*	46700..47062	121	-			
TPA4-*orf74*	47059..48669	537	DNA methyltransferase [*Mycobacterium abscessus*]	65	0.0	DNA Methylase (pfam00145; cd00315)
TPA4-*orf75*	48666..49040	125	hypothetical protein [*Nocardia brasiliensis*]	53	3e-25	
TPA4-*orf76*	49037..51694	886	gp85 [*Mycobacterium* phage DS6A]	38	3e-62	
TPA4-*orf77*	51945..52427	161	recombination endonuclease VII [*Mycobacterium abscessus*]	51	3e-35	Recombination endonuclease (pfam02945)
TPA4-*orf78*	52424..52702	93	-			
TPA4-*orf79*	52699..53697	333	hypothetical protein [*Nocardia farcinica*]	52	2e-18	

ORF[a]	Coordinates	Size (aa)	Significant match [b]	% identity [c]	E value [d]	Putative function (conserved motif) [e]
TPA4-*orf80*	53694..53990	99	-			
TPA4-*orf81*	53987..54289	101	hypothetical protein [*Tsukamurella* sp. 1534]	45	1e-12	
TPA4-*orf82*	complement(54265..55047)	261	-			
TPA4-*orf83*	55293..55478	62	-			
TPA4-*orf84*	55475..55669	65	-			

[a] ORFs were numbered consecutively; [b] The most closely related gene (only if named) and the name of the organism; [c] Per cent identity is based on the best match when a BLAST P analysis is performed; [d] The probability of obtaining a match by chance as determined by BLAST analysis and only values less than 10^{-4} were considered significant; [e] Predicted function is based on amino acid identity, conserved motifs, and gene location within functional modules.

Table 5.5 Palindrome sequences identified in the genomes of temperate phages GAL1, GMA1, and TPA4.

Repeat number-Phage	Size (bp)	Coordinates	Sequence alignment	Position
P1-GAL1	36	27770-27804	CCGCCCCCACTCTTC-ACGTCGGAGAGTGGGGGCGG	Between *orf30* and *orf31*
		27804-27770	CCGCCCCCACTCTCCGACGT-GAAGAGTGGGGGCGG	
P2-GAL1	27	30767-30793	CTTGGTCGGATTTGAAATCCGACCAAG	Within *orf34*
		30793-30767	CTTGGTCGGATTTCAAATCCGACCAAG	
P3-GAL1	20	26325-26344	CAGCCGTCGATCGACGCCTG	Within *orf29*
		26344-26325	CAGGCGTCGATCGACGGCTG	
P4-GAL1	18	23320-23337	GCAGCGGCCGGCCGCTGC	Within *orf28*
		23337-23320	GCAGCGGCCGGCCGCTGC	
P1-GMA1	39	6388-6426	CAGCCCGGCACCGTGTGGCCCTCCCTCGGTGCCGGGCTG	Between *orf7* and *orf8*
		6426-6388	CAGCCCGGCACCGAGGGAGGGCCACACGGTGCCGGGCTG	
P2-GMA1	18	21453-21470	GGTGCTGACGTCAGCACC	Within *orf27*
		21470-21453	GGTGCTGACGTCAGCACC	
P3-GMA1	16	27417-27432	CTGGTTGATCAACCAG	Within *orf36*

Repeat number-Phage	Size (bp)	Coordinates	Sequence alignment	Position
		27432-27417	CTGGTTGATCAACCAG	
P4-GMA1	16	32932-32947	GCTGGCCGCGGCCAGC	Within *orf52*
		32947-32932	GCTGGCCGCGGCCAGC	
P1-TPA4	86	4475-4556	ACCGGGCCGCGTCGACC--GATCGAAG-GTCCGGAACGTGCTGCAGCACGAGCTCGACACTGCGAACCGGGTCG-CTCGGCTCGGT	Within *orf4*
		4556-4475	ACCGAGCCGAG-CGACCCGGTTCGCAGTGTCGAGCTCGTGCTGCAGCACGTTCCGGAC-CTTCGATC--GGTCGACGCGGCCCGGT	
P2-TPA4	48	9987-10032	CTGGTG--GTGGCCGCGTAGCGGGGAAGCTACGCGGCCACCGCACCAG	Between *orf13* and *orf14*
		10032-9987	CTGGTGCGGTGGCCGCGTAGCTTCCCCGCTACGCGGCCAC--CACCAG	
P3-TPA4	46	38927-38970	GCCGCGGCCAGCAGACCGGCTTCAAAACC--TCTGCTGGCCGCGGC	Between *orf51* and *orf52*
		38970-38927	GCCGCGGCCAGCAGA--GGTTTTGAAGCCGGTCTGCTGGCCGCGGC	
P4-TPA4	36	30268-30301	CGCGGCGGCGAT--CATCGCTGCCATCGCCGCCGCG	Within *orf34*
		30301-30268	CGCGGCGGCGATGGCAGCGATG--ATCGCCGCCGCG	
P5-TPA4	35	31306-31340	AGAAGCCCCCTCCGAGATTCTCGGAGGGGGCTTCT	Between *orf36* and *orf37*

Repeat number-Phage	Size (bp)	Coordinates	Sequence alignment	Position
		31340-31306	AGAAGCCCCCTCCGAGAATCTCGGAGGGGGCTTCT	
P6-TPA4	35	7620-7654	CCTCGAGGAGGCCCGCACCACGGGCCTCCTCGAGG	Within *orf9*
		7654-7620	CCTCGAGGAGGCCCGTGGTGCGGGCCTCCTCGAGG	
P7-TPA4	32	27504-27535	CGTGCCCCGCGGCCAGCGCGCCGCGGGGCACG	Between *orf29* and *orf30*
		27535-27504	CGTGCCCCGCGGCGCGCTGGCCGCGGGGCACG	
P8-TPA4	31	30720-30750	AGCGGCCCCGAGCATCACGCTCGGGGCCGCT	Between *orf34* and *orf35*
		30750-30720	AGCGGCCCCGAGCGTGATGCTCGGGGCCGCT	
P9-TPA4	25	51717-51741	GGTTCCGAGGGCGACCCTCGGAACC	Between *orf76* and *orf77*
		51741-51717	GGTTCCGAGGGTCGCCCTCGGAACC	
P10-TPA4	20	29014-29033	CCGCCGCGACGTCGCGGCGG	Within *orf31*
		29033-29014	CCGCCGCGACGTCGCGGCGG	
P11-TPA4	20	8401-8420	CGACGGCGTCGACGCCCTCG	Within *orf11*

Repeat number-Phage	Size (bp)	Coordinates	Sequence alignment	Position
		8420-8401	CGAGGGCGTCGACGCCGTCG	
P12-TPA4	16	38186-38201	CGGGAAGTACTTCCCG	Within *orf50*
		38201-38186	CGGGAAGTACTTCCCG	
P13-TPA4	16	45597-45612	CGCCGCCCGGGCGGCG	Within *or69*
		45612-45597	CGCCGCCCGGGCGGCG	

Table 5.6 Putative genes in *Gordonia amarae* strain NBRC 15530 genome sequence (accession no. NZ_BAED00000000) that might be associated with phage defence systems

Defence system	Putative function	Contig	Locus	Locus tag	Gene
Restriction Modification (RM) Systems	Type I restriction endonuclease subunit M	GOAMR_20	12642..14267	00160	hsdM
	Type I restriction-modification system specificity subunit		14260..15432	00170	hsdS
	Antirestriction ArdA family protein		complement(135962..136546)	01160	
	HNH endonuclease	GOAMR_06	complement(103654..104838)	01010	
	DNA methyltransferase		complement(104838..106046)	01020	
	Antirestriction protein		complement(107767..108321)	01070	
	Type I restriction-modification system DNA methylase	GOAMR_03	170308..171183	01460	hsdM
	Restriction endonuclease subunit S		171180..172328	01470	hsdS
	Restriction endonuclease subunit R		183755..187090	01600	
	Modification methylase	GOAMR_24	1976..23085	00130	
	HNH endonuclease		23085..24224	00140	
	McrC protein	GOAMR_51	complement(50134..51456)	00510	mcrC
	5-methylcytosine-specific restriction enzyme B		complement(51470..53575)	00520	mcrB

Defence system	Putative function	Contig	Locus	Locus tag	Gene
	ribonuclease H	GOAMR_50	72585..73355	00640	rnhA
	trans-aconitate 2-methyltransferase		72585..73355	00650	tam
	DNA methyltransferase		complement(115171..116358)	01040	
	Endonuclease (pfam03852) [b]		116550..116840	01050	vsr
	Antirestriction protein	GOAMR_25	29863..30420	00340	
	DNA methylase		32177..33307	00390	
	McrBC 5-methylcytosine restriction system component (pfam10117) [a]		complement(33309..34673)	00400	
	exonuclease RecB	GOAMR_04	complement(82276..83154)	00710	
	SAM-dependent methlyltransferase		83251..84081	00720	
Abortive Infection (Abi) Systems	Antitoxin	GOAMR_20	302117..302314	02770	
	Toxin PIN[a]		302314..302721	02780	
	Zeta toxin family protein	GOAMR_43	complement(19300..20175)	00180	
	Abortive infection protein	GOAMR_03	complement(150116..150835)	01270	
Clustered Regularly Interspaced Short Palindromic	CRISPR-associated protein Cas2	GOAMR_28	complement(786..1073)	00010	cas2
	CRISPR-associated protein Cas1		complement(1080..2732)	00020	cas1/4
	CRISPR-associated helicase Cas3		3084..5492	00030	cas3

Defence system	Putative function	Contig	Locus	Locus tag	Gene
Repeats (CRISPR) Systems	CRISPR-associated protein Csx17 (cd09767) [a]		5489..7708	00040	
	CRISPR-associated protein		7737..8660	00050	
	CRISPR-associated protein, GSU0054 family (pfam09609) [a]		8672..10114	00060	

[a] Based on the presence of a motif of this type when checked against GenBank on January 1st 2014 (previously annotated as a hypothetical protein in GenBank entry for Gordonia amarae); [b] Based on the presence of a motif of this type when checked against GenBank on January 1st 2014 (previously annotated as a DNA mismatch repair protein Vsr in GenBank entry for Gordonia amarae).

6. General conclusions and future work

6.1. General conclusions

My PhD project set out to isolate and characterise phages additional to those already isolated, which are infective for the hydrophobic Mycolata group containing the genera *Gordonia, Nocardia*, *Rhodococcus,* and *Tsukamurella.* These organisms are responsible for stabilising the foams formed on aerobic reactors in activated sludge systems. The potential value of phages as biological control agents to control this global operational problem, for which there is no universal control strategy, was also explored.

The major findings were:

- Although a phage lytic for several *Tsukamurella* strains, including *T. spumae* and *T. pseudospumae* had been isolated and characterised (Petrovski *et al.*, 2011a) previously by the La Trobe group, others were sought here to extend the targetable host range. Successful isolation and characterisation of three previously unknown *Tsukamurella* phages TIN2, TIN3, and TIN4 of the *Siphoviridae* viral family, infective for several strains of *Tsukamurella inchonensis* and *T. paurometabola* was achieved. I was able to show that these phages were genetically very similar to each other and to the *Gordonia* phage GTE7, isolated earlier in our lab. Of these, TIN3 and TIN4 genome sequences were identical with the exception of several mutations in the lysis gene module of TIN3 phage. These mutations corresponded to the formation of plaques of an average smaller size than those produced by TIN4 with *T. inchonensis* as host. Furthermore, mass spectroscopy data revealed that phage TIN4 alone encoded a gene associated with virion morphogenesis located in its structural gene module, a highly unusual gene arrangement pattern for *Siphoviridae* phages. Although mass spectroscopy was not conducted on phage TIN3, the same diagnostic motifs were also present in it, suggesting this phage might also contain the same structural gene in its DNA replication module. As these phages appear to be obligatory lytic they might be suitable candidates for potential use in a cocktail of phages for foam biocontrol, except that none of these phages lysed cells of *T.*

spumae and *T. pseudospumae*, both known to stabilise foams in activated sludge plants (de los Reyes III, 2010; Nam *et al.*, 2003). However, these could be added as a cocktail with the phage of Petrovski *et al.* (2011a) included to cover these species

- The phages especially sought in my research included those lytic for *Skermania piniformis*, a very common foaming organism in activated sludge processes in plants around the world (de los Reyes III, 2010), and for which no phage had yet been isolated. I did obtain one such *Siphoviridae* phage, whose host range was limited to some but not all, *S. piniformis* strains against which it was screened. The genome of this phage was novel as it only shared 1% of its genome with Mycobacterium phage Bricole, and a large percentage of its genes could not be functionally annotated. Based on the host range studies I carried out, this phage would not necessarily be suitable as a control weapon against all *S. piniformis* foams, because of its limited effectiveness against only some of the strains of this organism under laboratory conditions

- The other phage target organism I was interested in was *Gordonia amarae*, globally among the most important of the stable foam forming organisms, especially in Australia. Therefore attempts were made to obtain as many novel *Gordonia* phages as possible, from habitats other than activated sludge plants, in the hope that they would contain one or more infective against *G. amarae*. Isolation and characterisation of nine previously unknown *Siphoviridae* phages infective for members of the genus *Gordonia* revealed that none of them targeted any of the large numbers of *G. amarae* isolates from foaming activated sludge plants held at La Trobe University. Some however, did lyse other foaming *Gordonia* strains, among them *G. malaquae*. Many were similar genetically to those isolated in earlier work carried out here. For example, phage GTE8, which was isolated from the sediment in a local stream, was genetically very similar to phages GRU1 and GTE5 for other *Gordonia* species (Petrovski *et al.*, 2012b) obtained previously from activated sludge. From this it seems that Mycolata phages are more widely distributed than we once thought, and in future work samples from environments other than activated sludge should be included in any isolation exercise. Examination of the genome assembly data for phages GMA3, GMA4 and GMA5 all grown on *G. malaquae* strain BEN700 revealed evidence of spontaneous prophage induction of a resident prophage GMA1, occurring despite co-infection with lytic phage infection on this strain.

Furthermore, and not surprisingly, several unusual genes such as those encoding toxins and virulence factors were observed in their genome sequences. Consequently, some of these phages may be unsuitable candidates for foam biocontrol. However, some phage genomes carried genes for integrases, associated with host cell lysogeny and their possession may mean these are unsuitable for foam biocontrol, unless these prophage lytic cycles could be induced *in situ*. This was explored as stated below

- Attempts were made to see how many Mycolata genomes carried prophages, and whether it was possible to induce their lytic cycles. I saw several attractions in being able to achieve this for control of foams, as discussed below. Bioinformatic *in silico* analysis of all the available Mycolata whole genome sequences (wgs) revealed that prophages were common elements there. Thus, putative genes of phage origin were detected in 83% of Mycolata wgs data examined with PHAge Search Tool (PHAST) (Zhou *et al.*, 2011). Of the 259 species of Mycolata examined, 26% contained putative intact prophages potentially inducible with induction agents like mitomycin C. Subsequent induction studies on large numbers of Mycolata strains held at this University, including many *G. amarae* isolates, demonstrated that phages GAL1 from *G. alkanivorans* strain CON70, GMA1 from *G. malaquae* strain BEN700, and TPA4 from *T. paurometabola* strain CON55 were inducible after exposure to mitomycin C. These phages were characterised extensively and while the two induced phages from *G. malaquae* and *T. paurometabola* were novel and different to those seen in the corresponding wgs, that induced from *G. alkanivorans* was almost identical to the prophage detected in the wgs data from strain NBRC 16433

6.2. Discussion & future work

The aims of my PhD were to investigate how phages lytic for the Mycolata could be exploited to provide an environmentally friendly and highly specific method for foam control in activated sludge wastewater treatment plants, and to extend the numbers of Mycolata phages and fully characterise them. What is known about phage-host population dynamics would suggest that complete elimination of either phage or host would not eventuate in activated sludge (Hantula

et al., 1991; Lee *et al.*, 2007; Ogata *et al.*, 1980; Otawa *et al.*, 2007; Shapiro *et al.*, 2010). Achieving total elimination of the Mycolata species responsible would almost certainly minimise the degradation of hydrophobic and xenobiotic substrates in the system, because these populations probably play a major role in their metabolism (Arenskotter *et al.*, 2004; Drzyzga, 2012). Instead a homeostatic balance would develop between their population levels (Hantula *et al.*, 1991; Lee *et al.*, 2007; Ogata *et al.*, 1980; Otawa *et al.*, 2007; Shapiro *et al.*, 2010). Hence any phage control strategy should aim to reduce the host foaming Mycolata population levels below the threshold level required for stable foam formation (Petrovski *et al.*, 2011e). This is not the same for all Mycolata (Petrovski *et al.*, 2011e), despite claims to the contrary (Davenport *et al.*, 2000).

The phage delivery system chosen should be controlled so that only when Mycolata counts exceed the predetermined threshold would it activate and thus risks of over-dosing would be minimised to avoid wastage. Thus, releasing phage into the return activated sludge (RAS) line might be the best strategy for addition as this is a region a low level of turbulence and the high concentrations of biomass there would encourage phage-host adsorption. Proper dosing levels of phage into these systems would depend on regular microscopic examination of the mixed liquor using semi-quantitative FISH with probes targeting the foaming bacteria responsible. The appropriate phage(s) would then be added to the system when the appropriate threshold level of for stable foam formation is approached. When the levels of foaming bacterium fall below this level, phage addition would be ceased.

Before my PhD project began only ten Mycolata phages had been sequenced and characterised, and only two of these appeared to have the capacity to enter a lysogenic lifecycle based the presence of integrase genes in their genomes. All had come from activated sludge, and no other habitats had been used. In my study, a further 16 Mycolata phages were isolated and characterised, with one isolated from a fresh water creek and sediment from a puddle, both with no evidence of any sewage contamination. This suggests that extending the sample sites might prove to be very valuable in obtaining further novel phages.

Special attention was given to seek phages infective for the two major foaming Mycolata in activated sludge plants, *Gordonia amarae* and *Skermania piniformis*. While only successful for some strains of *S. piniformis*, I was unsuccessful in the search for *G. amarae*. The reasons why

such phages were not isolated may be several. Despite many attempts, the lytic effect of phages GRU1, GTE1, GTE2, and GTE4 observed earlier by Thomas *et al.* (2002) showing lysis of *G. amarae* could not be replicated in my studies. This could be explained by the mutation of either phage or host over the last 12 years. There is evidence from wgs data that *G. amarae* strain NBRC 15530^{T}, the only wgs currently available, might possess a CRISPR region, RM systems, and Abi systems, and these may provide it with defence against any phage attack. It may be valuable to obtain wgs data from the many *G. amarae* isolates available from foaming plants to see if any of these possess inherent genetic phage defence systems. No wgs are available yet for *S. piniformis*, but equally, such data may help explain why some strains are not susceptible to phage attack.

There are problems, including those listed below, mostly microbiological, that need solving before phage therapy can be contemplated as a foam control strategy in activated sludge plants. Many of the experiments where these might be tested could be carried out in small laboratory reactors set up to simulate activated sludge plants and laboratory based foam tests:

- Designing appropriate economically viable production for phage titres at an industrial scale, and subsequent downstream harvesting as well as storage protocols for them

- Determining phage host ranges and specificities *in situ*, which may not necessarily be the same as those demonstrated in laboratory systems. There are now protocols for phageFISH fluorescent phage tagging (Allers *et al.*, 2013; Dang & Sullivan, 2014), which coupled with Fluorescence *in situ* hybridisation methods (Carr *et al.*, 2005) using rRNA targeted probes now available for many of the foaming Mycolata, may provide this information

- Clarifying the replication kinetics/efficiencies (including burst sizes) for these phages *in situ*. Real time PCR methods using primers designed against phage genomic sequence information should provide this information

- Determining how long the phages persist within these systems, and how many phage particles need to be added. Again real time PCR methods targeting the added phages and host cells should allow such data to be obtained

- Determine how is it possible to minimise the risks of phage resistance. The literature suggests that phage resistance is a serious consideration in other phage therapy applications (Dy *et al.*, 2014; Hyman & Abedon, 2010; Labrie *et al.*, 2010; Samson *et al.*, 2013; Sorek *et al.*, 2008). Using a cocktail of several phages may solve this important potential problem, especially if they all have different attachment sites on the host cell

- Investigate if it possible to exploit appropriate prophage release from lysogenic hosts as an alternative method of phage biocontrol. Much effort in future work should be directed at developing this attractive approach, and as more extensive wgs data for these organisms becomes available, more clues as to how this could be applied will emerge. This approach would also eliminate the need for large-scale phage production required by a traditional phage therapy approach

- Determine if phage therapy would be appropriate for Mycolata foaming incidents in anaerobic methane digesters, with longer solids retention times and higher concentrations of biomass (Seviour *et al.*, 2010), allowing more ready contact between phages and host cells

7. References

Abedon, S. T., Thomas-Abedon, C., Thomas, A. & Mazure, H. (2011). Bacteriophage prehistory: Is or is not Hankin, 1896, a phage reference? *Bacteriophages* **1**, 174-178.

Ackermann, H. W. (1991). Frequency of morphological phage descriptions. *Arch Virol* **124**, 201-209.

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

Ackermann, H. W. (2001). Frequency of morphological phage descriptions in the year 2000. Brief review. *Arch Virol* **146**, 843-857.

Ackermann, H. W. (2003). Bacteriophage observations and evolution. *Res Microbiol* **154**, 245-251.

Ackermann, H. W. (2007). 5500 Phages examined in the electron microscope. *Arch Virol* **152**, 227-243.

Ackermann, H. W. & Dubow, M. S. (1987). *General Properties of Bacteriophages*, vol. 1. CRC Press.

Ali, Y., Koberg, S., Hessner, S., Sun, X., Rabe, B., Back, A., Neve, H. & Heller, K. J. (2014). Temperate Streptococcus thermophilus phages expressing superinfection exclusion proteins of the Ltp type. *Front Microbiol* **5**, 98.

Allers, E., Moraru, C., Duhaime, M. B., Beneze, E., Solonenko, N., Barrero-Canosa, J., Amann, R. & Sullivan, M. B. (2013). Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. *Environ Microbiol* **15**, 2306-2318.

Almeida, D. R., Miller, D. & Alfonso, E. C. (2010). Tsukamurella: an emerging opportunistic ocular pathogen. *Can J Ophthalmol* **45**, 290-293.

Amann, R. & Ludwig, W. (2000). Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology. *FEMS Microbiol Rev* **24**, 555-565.

Ammelburg, M., Frickey, T. & Lupas, A. N. (2006). Classification of AAA+ proteins. *J Struct Biol* **156**, 2-11.

Anonymous (1969). Milwaukee mystery: unusual operating problem develops. *Water Sewage Works*, 116-123.

Arenskotter, M., Broker, D. & Steinbuchel, A. (2004). Biology of the metabolically diverse genus Gordonia. *Appl Environ Microbiol* **70**, 3195-3204.

Arraj, A., Bohatier, J., Laveran, H. & Traore, O. (2005). Comparison of bacteriophage and enteric virus removal in pilot scale activated sludge plants. *J Appl Microbiol* **98**, 516-524.

Ashbolt, N. J. (2004). Microbial contamination of drinking water and disease outcomes in developing regions. *Toxicology* **198**, 229-238.

Ashelford, K. E., Norris, S. J., Fry, J. C., Bailey, M. J., & Day, M. J. (2000). Seasonal population dynamics and interactions of competing bacteriophages and their host in the rhizosphere. *Appl Environ Microbiol* **66**, 4193-4199.

Bailly-Bechet, M., Vergassola, M. & Rocha, E. (2007). Causes for the intriguing presence of tRNAs in phages. *Genome Res* **17**, 1486-1495.

Bajpai, R., Soni, V., Khandrika, L., Jangir, P. K., Sharma, R. & Agrawal, P. (2012). Genome sequence of a novel actinophage PIS136 isolated from a strain of Saccharomonospora sp. *J Virol* **86**, 9552.

Barr, J. J., Slater, F. R., Fukushima, T. & Bond, P. L. (2010). Evidence for bacteriophage activity causing community and performance changes in a phosphorus-removal activated sludge. *FEMS Microbiol Ecol* **74**, 631-642.

Barr, J. J., Auro, R., Furlan, M., Whiteson, K. L., Erb, M. L., Pogliano, J., Stotland, A., Wolkowicz, R., Cutting, A. S. & other authors (2013). Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc Natl Acad Sci U S A* **110**, 10771-10776.

Beumer, A. & Robinson, J. B. (2005). A broad-host-range, generalized transducing phage (SN-T) acquires 16S rRNA genes from different genera of bacteria. *Appl Environ Microbiol* **71**, 8301-8304.

Bibby, K. (2014). Improved bacteriophage genome data is necessary for integrating viral and bacterial ecology. *Microb Ecol* **67**, 242-244.

Bitton, G. (2005). *Wastewater Microbiology*, 3 edn. New York: John Wiley & Sons Inc.

Blackall, L. L. (1986). *Actinomycete scum problems in actibated sludge plants*. Doctor of Philosophy thesis, University of Queensland, Brisbane.

Blackall, L. L. & Marshall, K. C. (1989). The Mechanism of Stabilization of Actinomycete Foams and the Prevention of Foaming under Laboratory Conditions. *J Ind Microbiol* **4**, 181-187.

Blackall, L. L., Harbers, A. E., Greenfield, P. F. & Hayward, A. C. (1991). Foaming in Activated-Sludge Plants - a Survey in Queensland, Australia and an Evaluation of Some Control Strategies. *Water Res* **25**, 313-317.

Blackall, L. L., Parlett, J. H., Hayward, A. C., Minnikin, D. E., Greenfield, P. F. & Harbers, A. E. (1989). Nocardia-Pinensis Sp-Nov, an Actinomycete Found in Activated-Sludge Foams in Australia. *J Gen Microbiol* **135**, 1547-1558.

Blackbeard, J. R., Ekama, G. A. & Marais, G. V. (1986). A Survey of Filamentous Bulking and Foaming in Activated-Sludge Plants in South-Africa. *Water Pollution Control* **85**, 90-100.

Bohannan, B. J. M. & Lenski, R. E. (2000). Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. *Ecology Letters* **3**, 362-377.

Bose, M. & Barber, R. D. (2006). Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol* **6**, 223-227.

Bouza, E., Perez-Parra, A., Rosal, M., Martin-Rabadan, P., Rodriguez-Creixems, M. & Marin, M. (2009). *Tsukamurella*: a cause of catheter-related bloodstream infections. *Eur J Clin Microbiol Infect Dis* **28**, 203-210.

Bradley, D. E. (1967). Ultrastructure of Bacteriophages and Bacteriocins. *Bacteriol Rev* **31**, 230.

Breitbart, M. & Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13**, 278-284.

Breitbart, M., Miyake, J. H. & Rohwer, F. (2004). Global distribution of nearly identical phage-encoded DNA sequences. *Fems Microbiol Lett* **236**, 249-256.

Breitbart, M., Thompson, L. R., Suttle, C. A. & Sullivan, M. B. (2007). Exploring the Vast Diversity of Marine Viruses. *Oceanography* **20**, 135-139.

Brüsow, H. & Hendrix, R. W. (2002). Phage Genomics: Small Is Beautiful. *Cell* **108**, 13-16.

Brüssow, H. & Desiere, F. (2001). Comparative phage genomics and the evolution of *Siphoviridae*: insights from dairy phages. *Mol Microbiol* **39**, 213-222.

Brüssow, H. & Kutter, E. (2004). Genomics and Evolution of Tailed Phages. In *Bacteriophages Biology and Applications*. Edited by E. Kutter & A. Sulakvelidze. Florida: CRC Press.

Brüssow, H., Canchaya, C. & Hardt, W. D. (2004a). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* **68**, 560-602, table of contents.

Brüssow, H., Canchaya, C. & Hardt, W.-D. (2004b). Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. *Microbiol Mol Biol Rev* **68**, 560-602.

Buckling, A. & Rainey, P. B. (2002). Antagonistic coevolution between a bacterium and a bacteriophage. *Proceedings of the Royal Society B-Biological Sciences* **269**, 931-936.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brüssow, H. (2003). Prophage genomics. *Microbiol Mol Biol Rev* **67**, 238-276.

Carlton, R. M. (1999). Phage therapy: past history and future prospects. *Arch Immunol Ther Exp (Warsz)* **47**, 267-274.

Carr, E. L., Eales, K., Soddell, J. & Seviour, R. J. (2005). Improved permeabilization protocols for fluorescence in situ hybridization (FISH) of mycolic-acid-containing bacteria found in foams. *J Microbiol Meth* **61**, 47-54.

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

Catalano, C. E. (2000). The terminase enzyme from bacteriophage lambda: a DNA-packaging machine. *Cell Mol Life Sci* **57**, 128-148.

Chen, J. & Novick, R. P. (2009). Phage-mediated intergeneric transfer of toxin genes. *Science* **323**, 139-141.

Chibani-Chennoufi, S., Bruttin, A., Dillmann, M. L. & Brussow, H. (2004). Phage-host interaction: an ecological perspective. *J Bacteriol* **186**, 3677-3686.

Chirakadze, I., Perets, A. & Ahmed, R. (2009). Phage Typing. In *Bacteriophages* (Methods in Molecular Biology™), vol. 502, pp. 293-305. Edited by M. J. Clokie & A. Kropinski: Humana Press.

Choi, J., Kotay, S. M. & Goel, R. (2011). Bacteriophage-based biocontrol of biological sludge bulking in wastewater. *Bioeng Bugs* **2**, 214-217.

Chun, J., Blackall, L. L., Kang, S.-O., Hah, Y. C. & Goodfellow, M. (1997). A Proposal To Reclassify *Nocardia pinensis* Blackall et al. as *Skermania pinifomis* gen. nov., comb. nov. *Int J Syst Evol Microbiol* **47**, 127-131.

Chunhong, M., Bhardwaj, K., Sharkady, S. M., Fish, R. I., Driscoll, T., Wower, J., Zwieb, C., Sobral, B. W. S. & Williams, K. P. (2009). Variations on the tmRNA gene. *RNA Biology* **6**, 355-361.

Collins, M. D., Smida, J., Dorsch, M. & Stackebrandt, E. (1988). *Tsukamurella* gen. nov. harboring *Corynebacterium paurometabolum* and *Rhodococcus aurantiacus. Int J Syst Bacteriol* **38**, 385-391.

Crockett, J. K. & Brownell, G. H. (1972). Isolation and characterization of a lysogenic strain of *Nocardia erythropolis*. *J Virol* **10**, 737-745.

Cserzo, M., Wallin, E., Simon, I., Heijne, G. v. & Elofsson, A. (1997). Prediction of transmembrane alpha-helices in procariotic membrane proteins: the Dense Alignment Surface method. *Protein Engineering* **10**, 673-676.

d'Herelle, F. (1917). Sur un microbe invisible antagoniste des bacilles dysentérique. *Academy of Science Paris* **165**, 373-375.

Daims, H. & Wagner, M. (2010). The microbiology of nitrogen removal. In *Microbial Ecology of Activated Sludge*. Edited by R. J. Seviour & P. H. Nielsen. London: IWA Publishing.

Dang, V. T. & Sullivan, M. B. (2014). Emerging methods to study bacteriophage infection at the single-cell level. *Front Microbiol* **5**, 724.

Daniel, A., Bonnen, P. E. & Fischetti, V. A. (2007). First complete genome sequence of two *Staphylococcus epidermidis* bacteriophages. *J Bacteriol* **189**, 2086-2100.

Davenport, R. J., Pickering, R. L., Goodhead, A. K. & Curtis, T. P. (2008). A universal threshold concept for hydrophobic mycolata in activated sludge foaming. *Water Res* **42**, 3446-3454.

Davenport, R. J., Curtis, T. P., Goodfellow, M., Stainsby, F. M. & Bingley, M. (2000). Quantitative use of fluorescent *in situ* hybridization to examine relationships between mycolic acid-containing actinomycetes and foaming in activated sludge plants. *Appl Environ Microbiol* **66**, 1158-1166.

de los Reyes, F. L., 3rd & Raskin, L. (2002). Role of filamentous microorganisms in activated sludge foaming: relationship of mycolata levels to foaming initiation and stability. *Water Res* **36**, 445-459.

de los Reyes III, F. L., (editor) (2010). *Foaming*. London: IWA Publishing.

de los Reyes III, F. L., Ritter, W. & Raskin, L. (1997). Group-Specific Small-Subunit rRNA Hybridization Probes To Characterize Filamentous Foaming in Activated Sludge Systems. *Appl Environ Microbiol* **63**, 1107-1117.

Del Casale, A., Flanagan, P. V., Larkin, M. J., Allen, C. C. & Kulakov, L. A. (2011a). Analysis of transduction in wastewater bacterial populations by targeting the phage-derived 16S rRNA gene sequences. *FEMS Microbiol Ecol* **76**, 100-108.

Del Casale, A., Flanagan, P. V., Larkin, M. J., Allen, C. C. & Kulakov, L. A. (2011b). Extent and variation of phage-borne bacterial 16S rRNA gene sequences in wastewater environments. *Appl Environ Microbiol* **77**, 5529-5532.

Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673-679.

Deng, Z., Kieser, T. & Hopwood, D. A. (1987). Activity of a *Streptomyces* transcriptional terminator in *Escherichia coli*. *Nucleic Acids Res* **15**, 2665-2675.

Dokland, T. (2000). Freedom and restraint: themes in virus capsid assembly. *Structure* **8**, R157-162.

Drzyzga, O. (2012). The strengths and weaknesses of *Gordonia*: a review of an emerging genus with increasing biotechnological potential. *Crit Rev Microbiol* **38**, 300-316.

Dy, R. L., Richter, C., Salmond, G. P. C. & Fineran, P. C. (2014). Remarkable Mechanisms in Microbes to Resist Phage Infections. *Ann Rev Virol* **1**, 307-331.

Eikelboom, D. H. (1991). Scuim-en driflaagvorming opzuiveringsinrichtingen. Delft: TNO Milieu en Energie.

Eikelboom, D. H. (1994). The *Microthrix parvicella* Puzzle. *Water Sci Technol* **29**, 271-279.

El Tahir, Y. & Skurnik, M. (2001). YadA, the multifaceted *Yersinia* adhesin. *Int J Med Microbiol* **291**, 209-218.

Enav, H., Beja, O. & Mandel-Gutfreund, Y. (2012). Cyanophage tRNAs may have a role in cross-infectivity of oceanic *Prochlorococcus* and *Synechococcus* hosts. *ISME J* **6**, 619-628.

Ewert, D. L. & Paynter, M. J. (1980). Enumeration of bacteriophages and host bacteria in sewage and the activated-sludge treatment process. *Appl Environ Microbiol* **39**, 576-583.

Flores, C. O., Meyer, J. R., Valverde, S., Farr, L. & Weitz, J. S. (2011). Statistical structure of host-phage interactions. *Proc Natl Acad Sci U S A* **108**, E288-297.

Flowers, J. J., He, S., Malfatti, S., del Rio, T. G., Tringe, S. G., Hugenholtz, P. & McMahon, K. D. (2013). Comparative genomics of two '*Candidatus* Accumulibacter' clades performing biological phosphorus removal. *ISME J* **7**, 2301-2314.

Fokine, A. & Rossmann, M. G. (2014). Molecular architecture of tailed double-stranded DNA phages. *Bacteriophage* **4**, e28281.

Fokine, A., Chipman, P. R., Leiman, P. G., Mesyanzhinov, V. V., Rao, V. B. & Rossmann, M. G. (2004). Molecular architecture of the prolate head of bacteriophage T4. *Proc Natl Acad Sci USA* **101**, 6003-6008.

Fokine, A., Islam, M. Z., Zhang, Z., Bowman, V. D., Rao, V. B. & Rossmann, M. G. (2011). Structure of the three N-terminal immunoglobulin domains of the highly immunogenic outer capsid protein from a T4-like bacteriophage. *J Virol* **85**, 8141-8148.

Fouts, D. E. (2006). Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* **34**, 5839-5851.

Frazee, R. W., Livingston, D. M., LaPorte, D. C. & Lipscomb, J. D. (1993). Cloning, sequencing, and expression of the *Pseudomonas putida* protocatechuate 3,4-dioxygenase genes. *J Bacteriol* **175**, 6194-6202.

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541-548.

Fujisawa, H. & Morita, M. (1997). Phage DNA packaging. *Genes to Cells* **2**, 537-545.

Fuller, N. J., Wilson, W. H., Joint, I. R. & Mann, N. H. (1998). Occurrence of a sequence in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques. *Appl Environ Microbiol* **64**, 2051-2060.

Furuta, Y. & Kobayashi, I. (2000). Restriction-Modification Systems as Mobile Epigenetic Elements. In *Madame Curie Bioscience Database*. Austin (TX).

Goddard, A. J. & Forster, C. F. (1987). Stable foams in activated sludge plants. *Enzy Microbial Technol* **9**, 164-168.

Goodfellow, M. & Maldonado, L. A. (2006). The Families *Dietziaceae*, *Gordoniaceae*, *Nocardiaceae* and *Tsukamurellaceae*. In *The Prokaryotes*, vol. 3, pp. 843-888. Edited by M. Dworkin, F. Stanley, E. Rosenberg, K. H. Schleifer & E. Stackerbrandt. New York: Springer.

Goodfellow, M., Kumar, V. & Maldonado, L. A. (2012). Genus II. *Gordonia* (Tsukamura 1971) Stackbrandt, Smida and Collins 1988, 345. In *Bergey's Manual of Systematic Bacteriology*, 2 edn, vol. 5, pp. 419-435. Edited by M. Goodfellow, P. Kampfer, H.-J. Busse, M. E. Trujillo, K.-I. Suzuki, W. Ludwig & W. B. Whitman. New York: Springer.

Groth, A. C. & Calos, M. P. (2004). Phage integrases: biology and applications. *J Mol Biol* **335**, 667-678.

Haerter, J. O., Mitarai, N. & Sneppen, K. (2014). Phage and bacteria support mutual diversity in a narrowing staircase of coexistence. *ISME J* **8**, 2317-2326.

Hambly, E. & Suttle, C. A. (2005). The viriosphere, diversity, and genetic exchange within phage communities. *Curr Opin Microbiol* **8**, 444-450.

Han, S., Arvai, A. S., Clancy, S. B. & Tainer, J. A. (2001). Crystal structure and novel recognition motif of rho ADP-ribosylating C3 exoenzyme from *Clostridium botulinum*: structural insights for recognition specificity and catalysis. *J Mol Biol* **305**, 95-107.

Hankin, E. H. (1896). L'action bactericide des eaux de la Jumna et du Gange sur le vibrion du cholera. *Annales de l'Institut Pasteur* **10**, 511.

Hantula, J., Kurki, A., Vuoriranta, P. & Bamford, D. H. (1991). Ecology of bacteriophages infecting activated sludge bacteria. *Appl Environ Microbiol* **57**, 2147-2151.

Hatfull, G. F. (2008). Bacteriophage genomics. *Curr Opin Microbiol* **11**, 447-453.

Hatfull, G. F. (2010). Mycobacteriophages: genes and genomes. *Annu Rev Microbiol* **64**, 331-356.

Hatfull, G. F. & Hendrix, R. W. (2015). The Actinophage Database. Pittsburg: Pittsburgh Bacteriophage Institute.

Hatfull, G. F., Jacobs-Sera, D., Lawrence, J. G., Pope, W. H., Russell, D. A., Ko, C. C., Weber, R. J., Patel, M. C., Germane, K. L. & other authors (2010). Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *J Mol Biol* **397**, 119-143.

Heilmann, S., Sneppen, K. & Krishna, S. (2012). Coexistence of phage and bacteria on the boundary of self-organized refuges. *Proc Natl Acad Sci USA* **109**, 12828-12833.

Hendrix, R. W. (2002). Bacteriophages: evolution of the majority. *Theor Popul Biol* **61**, 471-480.

Hendrix, R. W. (2003). Bacteriophage genomics. *Curr Opin Microbiol* **6**, 506-511.

Hendrix, R. W., Smith, M. C. M., Burns, R. N., Ford, M. E. & Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc Natl Acad Sci U S A* **96**, 2192-2197.

Hennes, K. P. & Suttle, C. A. (1995). Direct counts of viruses in natural waters and laboratory cultures by epifulorescene microscopy. *Limnol Oceanogr* **40**, 1050-1055.

Henthorn, K. S. & Friedman, D. I. (1995). Identification of related genes in phages phi 80 and P22 whose products are inhibitory for phage growth in *Escherichia coli* IHF mutants. *J Bacteriol* **177**, 3185-3190.

Hilton, M. C. & Stotzky, G. (1973). Use of Coliphages as Indicators of Water-Pollution. *Can J Microbiol* **19**, 747-751.

Hiraoka, M. & Tsumura, K. (1984). Suppression of Actinomycete Scum Production - a Case-Study at "Senboku-Wastewater-Treatment-Plant, Japan. *Water Sci Technol* **16**, 83-90.

Hyman, P. & Abedon, S. T. (2010). Bacteriophage host range and bacterial resistance. In *Adv Appl Microbiol*, pp. 217-248. Edited by A. Laskin, G. M. Gadd & S. Sariaslani. London: Academic Press.

Jenkins, D., Daigger, G. T. & Richard, M. G. (1993). *Manual on the causes and control of activated sludge bulking and foaming*, 2 edn. United States: Lewis.

Jenkins, D., Richard, M. G. & Daigger, G. T. (2003). *Manual on the Causes and Control of Activated Sludge Bulking, Foaming, and other Solids Separations Problems*. London: IWA Publishing.

Jensen, E. C., Schrader, H. S., Rieland, B., Thompson, T. L., Lee, K. W., Nickerson, K. W. & Kokjohn, T. A. (1998). Prevalence of Broad-Host-Range Lytic Bacteriophages of *Sphaerotilus natans*, *Escherichia coli*, and *Pseudomonas aeruginosa*. *Appl Environ Microbiol* **64**, 575-580.

Jolis, D. & Marneri, M. (2006). Thermal hydrolysis of secondary scum for control of biological foam. *Water Environ Res* **78**, 835-841.

Jover, L. F., Cortez, M. H. & Weitz, J. S. (2013). Mechanisms of multi-strain coexistence in host-phage systems with nested infection networks. *J Theor Biol* **332**, 65-77.

Juhala, R. J., Ford, M. E., Duda, R. L., Youlton, A., Hatfull, G. F. & Hendrix, R. W. (2000). Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J Mol Biol* **299**, 27-51.

Kamp, D. & Kahmann, R. (1981). The relationship of two invertible segments in bacteriophage Mu and *Salmonella typhimurium* DNA. *Molecular & general genetics : MGG* **184**, 564-566.

Kanda, K., Tan, Y. & Aizawa, K. (1989). A Novel Phage Genome Integrated into a Plasmid in *Bacillus thuringiensis* Strain Af101. *J Gen Microbiol* **135**, 3035-3041.

Khan, M. A., Satoh, H., Katayama, H., Kurisu, F. & Mino, T. (2002a). Bacteriophages isolated from activated sludge processes and their polyvalency. *Water Res* **36**, 3364-3370.

Khan, M. A., Satoh, H., Mino, T., Katayama, H., Kurisu, F. & Matsuo, T. (2002b). Bacteriophage-host interaction in the enhanced biological phosphate removing activated sludge system. *Water Sci Technol* **46**, 39-43.

Klatte, S., Rainey, F. A. & Kroppenstedt, R. M. (1994). Transfer of *Rhodococcus aichiensis* Tsukamura 1982 and *Nocardia amarae* Lechevalier and Lechevalier 1974 to the Genus *Gordona* as *Gordona aichiensis* comb. nov. and *Gordona amarae* comb. nov. *Int J Syst Bacteriol* **44**, 769-773.

Koskella, B. & Brockhurst, M. A. (2014). Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol Rev* **38**, 916-931.

Kotay, S. M., Datta, T., Choi, J. D. & Goel, R. (2011). Biocontrol of biomass bulking caused by *Haliscomenobacter hydrossis* using a newly isolated lytic bacteriophage. *Water Res* **45**, 694-704.

Kragelund, C., Nilsson, B., Eskilsson, K., Bogh, A. M. & Nielsen, P. H. (2010). Full-scale control of Mycolata foam by FEX-120 addition. *Water Sci Technol* **61**, 2443-2450.

Kragelund, C., Remesova, Z., Nielsen, J. L., Thomsen, T. R., Eales, K., Seviour, R. J., Wanner, J. & Nielsen, P. H. (2007). Ecophysiology of mycolic acid-containing *Actinobacteria* (Mycolata) in activated sludge foams. *FEMS Microbiol Ecol* **61**, 174-184.

Kraigher, B., Kosjek, T., Heath, E., Kompare, B. & Mandic-Mulec, I. (2008). Influence of pharmaceutical residues on the structure of activated sludge bacterial communities in wastewater treatment bioreactors. *Water Res* **42**, 4578-4588.

Kruger, D. H. & Bickle, T. A. (1983). Bacteriophage Survival - Multiple Mechanisms for Avoiding the Deoxyribonucleic-Acid Restriction Systems of Their Hosts. *Microbiol Rev* **47**, 345-360.

Kunin, V., He, S., Warnecke, F., Peterson, S. B., Garcia Martin, H., Haynes, M., Ivanova, N., Blackall, L. L., Breitbart, M. & other authors (2008). A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res* **18**, 293-297.

Kutter, E., Raya, R. & Carlson, K. (2005). Molecular mechanisms of phage infection. In *Bacteriophages Biology and Applications*. Edited by E. Kutter & A. Sulakvelidze. Florida: CRC Press.

Labrie, S. J., Samson, J. E. & Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nat Rev Microbiol* **8**, 317-327.

Laslett, D. & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* **32**, 11-16.

Lawrence, J. G., Hendrix, R. W. & Casjens, S. (2001). Where are the pseudogenes in bacterial genomes? *Trends Microbiol* **9**, 535-540.

Lawrence, J. G., Hatfull, G. F. & Hendrix, R. W. (2002). Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J Bacteriol* **184**, 4891-4905.

Lechevalier, M. P. & Lechevalier, H. A. (1974). *Nocardia amarae* sp. nov., an Actinomycete Common in Foaming Activated Sludge. *Nat J Syst Bacteriol* **24**, 278-288.

Lee, S. H., Satoh, H., Katayama, H. & Mino, T. (2004). Isolation, physiological characterization of bacteriophages from enhanced biological phosphorus removal activated sludge and their putative role. *J Microbiol Biotechnol* **14**, 730-736.

Lee, S. H., Onuki, M., Satoh, H. & Mino, T. (2006a). Isolation, characterization of bacteriophages specific to *Microlunatus phosphovorus* and their application for rapid host detection. *Lett Appl Microbiol* **42**, 259-264.

Lee, S. H., Otawa, K., Onuki, M., Satoh, H. & Mino, T. (2006b). Dynamics behavior of phage-host system related to *Microlunatus phosphovorus* in activated sludge with host inoculation. *J Microbiol Biotechnol* **16**, 1518-1522.

Lee, S. H., Otawa, K., Onuki, M., Satoh, H. & Mino, T. (2007). Population dynamics of phage-host system of *Microlunatus phosphovorus* indigenous in activated sludge. *J Microbiol Biotechnol* **17**, 1704-1707.

Lemire, S., Figueroa-Bossi, N. & Bossi, L. (2011). Bacteriophage crosstalk: coordination of prophage induction by trans-acting antirepressors. *PLoS Genet* **7**, e1002149.

Lesnik, E. A., Sampath, R., Levene, H. B., Henderson, T. J., McNeil, J. A. & Ecker, D. J. (2001). Prediction of rho-independent transcriptional terminators in *Escherichia coli*. *Nucleic Acids Res* **29**, 3583-3594.

Lima-Mendez, G., Toussaint, A. & Leplae, R. (2011). A modular view of the bacteriophage genomic space: identification of host and lifestyle marker modules. *Res Microbiol* **162**, 737-746.

Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. (2008). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* **24**, 863-865.

Lindqvist, N., Tuhkanen, T. & Kronberg, L. (2005). Occurrence of acidic pharmaceuticals in raw and treated sewages and in receiving waters. *Water Res* **39**, 2219-2228.

Lindrea, K. C. & Seviour, R. J. (2002). Activated Sludge - The Process. In *Encyclopedia of Environmental Microbiology*, pp. 74-80. Edited by G. Bitton. New York: John Wiley & Sons Inc.

Lipps, G., Weinzierl, A. O., von Scheven, G., Buchen, C. & Cramer, P. (2004). Structure of a bifunctional DNA primase-polymerase. *Nat Struct Mol Biol* **11**, 157-162.

Little, J. W. & Michalowski, C. B. (2010). Stability and instability in the lysogenic state of phage lambda. *J Bacteriol* **192**, 6064-6076.

Loc-Carrillo, C. & Abedon, S. T. (2011). Pros and cons of phage therapy. *Bacteriophages* **1**, 111-114.

Loenen, W. A. & Raleigh, E. A. (2014). The other face of restriction: modification-dependent enzymes. *Nucleic Acids Res* **42**, 56-69.

Loessner, M. J. (2005). Bacteriophage endolysins--current state of research and applications. *Curr Opin Microbiol* **8**, 480-487.

Lu, Z., Altermann, E., Breidt, F. & Kozyavkin, S. (2010). Sequence analysis of *Leuconostoc mesenteroides* bacteriophage Phi1-A4 isolated from an industrial vegetable fermentation. *Appl Environ Microbiol* **76**, 1955-1966.

Madoni, P., Davoli, D. & Gibin, G. (2000). Survey of filamentous microorganisms from bulking and foaming activated-sludge plants in Italy. *Water Res* **34**, 1767-1772.

Markine-Goriaynoff, N., Gillet, L., Van Etten, J. L., Korres, H., Verma, N. & Vanderplasschen, A. (2004). Glycosyltransferases encoded by viruses. *J Gen Virol* **85**, 2741-2754.

McMahon, K. D., He, S. & Oehmen, A. (2010). The microbiology of phosphorus removal. In *Microbial Ecology of Activated Sludge*. Edited by R. J. Seviour. London: IWA Publishing.

McNair, K., Bailey, B. A. & Edwards, R. A. (2012). PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* **28**, 614-618.

Mesquita, M. M. F. & Emelko, M. B. (2012). Bacteriophages as Surrogates for the Fate and Transport of Pathogens in Source Water and in Drinking Water Treatment Processes. In *Bacteriophages*. Edited by D. I. Kurtböke. Croatia: InTech.

Miao, E. A. & Miller, S. I. (1999). Bacteriophages in the evolution of pathogen-host interactions. *Proc Natl Acad Sci U S A* **96**, 9452-9454.

Mielczarek, A. T., Kragelund, C., Eriksen, P. S. & Nielsen, P. H. (2012). Population dynamics of filamentous bacteria in Danish wastewater treatment plants with nutrient removal. *Water Res* **46**, 3781-3795.

Miller, E. S., Kutter, E., Mosig, G., Arosaka, F., Kunisawa, T. & Rüger, W. (2003). Bacteriophage T4 Genome. *Microbiol Mol Biol Rev* **67**, 86-156

Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. (2012). Hypervariable loci in the human gut virome. *Proc Natl Acad Sci USA* **109**, 3962-3966.

Morgan, G., Hatfull, G. F., Casjens, S. & Hendrix, R. W. (2000). Bacteriophage Mu Genome Sequence: Analysis and Comparison with Mu-like Prophages in *Haemophilus*, *Neisseria* and *Deinococcus*. *J Mol Biol* **317**, 337-359.

Mori, T., Itokazu, K., Ishikura, Y., Mishina, F., Sakai, Y. & Koga, M. (1992). Evaluation of Control Strategies for Actinomycete Scum in Full-Scale Treatment Plants. *Water Sci Technol* **25**, 231-237.

Mott, M. L. & Berger, J. M. (2007). DNA replication initiation: mechanisms and regulation in bacteria. *Nat Rev Microbiol* **5**, 343-354.

Munk, A. C., Lapidus, A., Lucas, S., Nolan, M., Tice, H., Cheng, J. F., Del Rio, T. G., Goodwin, L., Pitluck, S. & other authors (2011). Complete genome sequence of *Tsukamurella paurometabola* type strain (no. 33). *Stand Genomic Sci* **4**, 342-351.

Nam, S. W., Kim, W., Chun, J. & Goodfellow, M. (2004). *Tsukamurella pseudospumae* sp. nov., a novel actinomycete isolated from activated sludge foam. *Int J Syst Evol Microbiol* **54**, 1209-1212.

Nam, S. W., Chun, J., Kim, S., Kim, W., Zakrzewska-Czerwinska, J. & Goodfellow, M. (2003). *Tsukamurella spumae* sp. nov., a novel actinomycete associated with foaming in activated sludge plants. *Syst Appl Microbiol* **26**, 367-375.

NCBI (2015). GenBank: NIH.

Nelson, D. (2004). Phage Taxonomy: We Agree To Disagree. *J Bacteriol* **186**, 7029-7031.

Ogata, S., Miyamoto, H. & Hayashida, S. (1980). An investigation of the influence of bacteriophages on the bacterial flora and purification powers of activated sludge. *J Gen Appl Microbiol* **26**, 97-108.

Orhon, D. & Artan, N. (1994). *Modelling of activated sludge systems*. Lancaster: Technomic Publishing Company Inc.

Otawa, K., Lee, S. H., Yamazoe, A., Onuki, M., Satoh, H. & Mino, T. (2007). Abundance, diversity, and dynamics of viruses on microorganisms in activated sludge processes. *Microb Ecol* **53**, 143-152.

Parracho, H. M., Burrowes, B. H., Enright, M. C., McConville, M. L. & Harper, D. R. (2012). The role of regulated clinical trials in the development of bacteriophage therapeutics. *J Mol Genet Med* **6**, 279-286.

Parsley, L. C., Consuegra, E. J., Thomas, S. J., Bhavsar, J., Land, A. M., Bhuiyan, N. N., Mazher, M. A., Waters, R. J., Wommack, K. E. & other authors (2010). Census of the Viral Metagenome within an Activated Sludge Microbial Assemblage. *Appl Environ Microbiol* **76**, 2673-2677.

Paul, J. H. (2008). Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J* **2**, 579-589.

Payne, K., Sun, Q., Sacchettini, J. & Hatfull, G. F. (2009). Mycobacteriophage Lysin B is a novel mycolylarabinogalactan esterase. *Mol Microbiol* **73**, 367-381.

Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J. & other authors (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**, 171-182.

Petrovski, S., Seviour, R. J. & Tillett, D. (2011a). Genome sequence and characterization of the *Tsukamurella* bacteriophage TPA2. *Appl Environ Microbiol* **77**, 1389-1398.

Petrovski, S., Seviour, R. J. & Tillett, D. (2011b). Prevention of *Gordonia* and *Nocardia* Stabilized Foam Formation by Using Bacteriophage GTE7. *Appl Environ Microbiol* **77**, 7864-7867

Petrovski, S., Seviour, R. J. & Tillett, D. (2011c). Characterization of the Genome of the Polyvalent Lytic Bacteriophage GTE2, Which Has Potential for Biocontrol of *Gordonia*, *Rhodococcus*, and *Nocardia* Stabilized Foams in Activated Sludge Plants. *Appl Environ Microbiol* **77**, 3923-3929.

Petrovski, S., Tillett, D. & Seviour, R. J. (2012a). Isolation and complete genome sequence of a bacteriophage lysing *Tetraspheara jensinsii*, a filamentous bacteria responsbile for bulking and foaming in activated sludge. *Virus Genes* **2**, 380-388.

Petrovski, S., Tillett, D. & Seviour, R. J. (2012b). Genome Sequences and Characterization of the Related *Gordonia* Phages GTE5 and GRU1 and Their Use as Potential Biocontrol Agents. *Appl Environ Microbiol* **78**, 42-47.

Petrovski, S., Seviour, R. J. & Tillett, D. (2013a). Genome sequence and characterization of a Rhodococcus equi phage REQ1. *Virus Genes* **46**, 558-590.

Petrovski, S., Seviour, R. J. & Tillett, D. (2013b). Characterization and whole genome sequences of the Rhodococcus bacteriophages RGL3 and RER2. *Arch Virol* **158**, 601-609.

Petrovski, S., Seviour, R. J. & Tillett, D. (2014). Genome sequence of the Nocardia bacteriophage NBR1. *Arch Virol* **159**, 167-173.

Petrovski, S., Dyson, Z. A., Seviour, R. J. & Tillett, D. (2012c). Small but sufficient: the *Rhodococcus* phage RRH1 has the smallest known *Siphoviridae* genome at 14.2 kilobases. *J Virol* **86**, 358-363.

Petrovski, S., Dyson, Z. A., Quill, E. S., McIlroy, S. J., Tillett, D. & Seviour, R. J. (2011d). An examination of the mechanisms for stable foam formation in activated sludge systems. *Water Res* **45**, 2146-2154.

Petty, N. K., Evans, T. J., Fineran, P. C. & Salmond, G. P. (2007). Biotechnological exploitation of bacteriophage research. *Trends Biotechnol* **25**, 7-15.

Pitt, P. & Jenkins, D. (1990). Causes and Control of *Nocardia* in Activated Sludge. *Research Journal of the Water Pollution Control Federation* **62**, 143-150.

Pope, W. H., Anders, K. R., Baird, M., Bowman, C. A., Boyle, M. M., Broussard, G. W., Chow, T., Clase, K. L., Cooper, S. & other authors (2014). Cluster M mycobacteriophages Bongo, PegLeg, and Rey with unusually large repertoires of tRNA isotypes. *J Virol* **88**, 2461-2480.

Pride, D. T., Wassenaar, T. M., Ghose, C. & Blaser, M. J. (2006). Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics* **7**, 8.

Proux, C., van Sinderen, D., Suarez, J., Garcia, P., Ladero, V., Fitzgerald, G. F., Desiere, F. & Brussow, H. (2002). The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like *Siphoviridae* in lactic acid bacteria. *J Bacteriol* **184**, 6026-6036.

Ptashne, M. (2004). *A genetic switch: phage lambda revisited*, 3 edn. New York: Cold Spring Harbor Laboratory Press.

Pujol, R., Duchene, P., Schetrite, S. & Canler, J. P. (1991). Biological Foams in Activated-Sludge Plants - Characterization and Situation. *Water Res* **25**, 1399-1404.

Rao, V. B. & Feiss, M. (2008). The bacteriophage DNA packaging motor. *Annu Rev Genet* **42**, 647-681.

Rao, V. B. & Black, L. W. (2010). Structure and assembly of bacteriophage T4 head. *J Virol* **7**, 356.

Ratcliff, S. W., Luh, J., Ganesan, A. T., Behrens, B., Thompson, R., Montenegro, M. A., Morelli, G. & Trautner, T. A. (1979). The genome of *Bacillus subtilis* phage SPP1: the arrangement of restriction endonuclease generated fragments. *Mol Gen Genet : MGG* **168**, 165-172.

Rohwer, F. & Edwards, R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* **184**, 4529-4535.

Rohwer, F. & Thurber, R. V. (2009). Viruses manipulate the marine environment. *Nature* **459**, 207-212.

Rohwer, F., Prangishvili, D. & Lindell, D. (2009). Roles of viruses in the environment. *Environ Microbiol* **11**, 2771-2774.

Rossetti, S., Tomei, M. C., Nielsen, P. H. & Tandoi, V. (2005). *"Microthrix parvicella"*, a filamentous bacterium causing bulking and foaming in activated sludge systems: a review of current knowledge. *FEMS Microbiol Rev* **29**, 49-64.

Sakamoto, K., Agari, Y., Agari, K., Yokoyama, S., Kuramitsu, S. & Shinkai, A. (2009). X-ray crystal structure of a CRISPR-associated RAMP superfamily protein, Cmr5, from *Thermus thermophilus* HB8. *Proteins* **75**, 528-532.

Salifu, S. P., Valero-Rello, A., Campbell, S. A., Inglis, N. F., Scortti, M., Foley, S. & Vazquez-Boland, J. A. (2013). Genome and proteome analysis of phage E3 infecting the soil-borne actinomycete *Rhodococcus equi*. *Environ Microbiol Rep* **5**, 170-178.

Samson, J. E., Magadan, A. H., Sabri, M. & Moineau, S. (2013). Revenge of the phages: defeating bacterial defences. *Nat Rev Microbiol* **11**, 675-687.

Sandmeier, H. (1994). Acquisition and rearrangement of sequence motifs in the evoiution of bacteriophage tail fibres. *Mol Microbiol* **12**, 343-350.

Savini, V., Fazii, P., Favaro, M., Astolfi, D., Polilli, E., Pompilio, A., Vannucci, M., D'Amario, C., Di Bonaventura, G. & other authors (2012). Tuberculosis-like pneumonias by the aerobic actinomycetes *Rhodococcus*, *Tsukamurella* and *Gordonia*. *Microbes Infect* **14**, 401-410.

Schattner, P., Brooks, A. N. & Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**, W686-689.

Schilling, K. & Zessner, M. (2011). Foam in the aquatic environment. *Water Res* **45**, 4355-4366.

Schrag, S. J. & Mittler, J. E. (1996). Host-parasite coexistence: The role of spatial refuges in stabilizing bacteria-phage interactions. *American Naturalist* **148**, 348-377.

Schwartz, M. A., Tabet, S. R., Collier, A. C., Wallis, C. K., Carlson, L. C., Nguyen, T. T., Kattar, M. M. & Coyle, M. B. (2002). Central venous catheter-related bacteremia due to *Tsukamurella* species in the immunocompromised host: a case series and review of the literature. *Clin Infect Dis* **35**, e72-77.

Schwartzbrod, J., Maux, M. & Chesnot, T. (2002). Parasitic Protozoa: Fate in Wastewater. In *Encyclopedia of Environmental Microbiology*, vol. 4. Edited by G. Bitton. New York: John Wiley & Sons, Inc.

Seong, C. N., Kim, Y. S., Baik, K. S., Choi, S. K., Kim, M. B., Kim, S. B. & Goodfellow, M. (2003). *Tsukamurella sunchonensis* sp. nov., a Bacterium Associated with Foam in Activated Sludge. *J Microbiol Biotechnol* **41**, 83-88.

Seviour, E. M., Williams, C. J., Seviour, R. J., Soddell, J. A. & Lindrea, K. C. (1990). A survey of filamentous bacterial populations from foaming activated sludge plants in eastern states of Australia. *Water Res* **24**, 493-498.

Seviour, E. M., Williams, C., Degrey, B., Soddell, J. A., Seviour, R. J. & Lindrea, K. C. (1994). Studies on Filamentous Bacteria from Australian Activated-Sludge Plants. *Water Res* **28**, 2335-2342.

Seviour, R. J., (editor) (2010a). *An overview of the microbes in activated sludge*. London: IWA Publishing.

Seviour, R. J. (2010b). Factors affecting the bulking and foaming filamentous bacteria in activated sludge. In *Microbial Ecology of Activated Sludge*. Edited by R. J. Seviour & P. H. Nielsen. London: IWA Publishing.

Seviour, R. J. & Blackall, L. L. (1999). Introduction to the microorganisms found in activated sludge processes. In *The Microbiology of Activated Sludge*. Edited by R. J. Seviour & L. L. Blackall. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Seviour, R. J. & Nielsen, P. H. (2010). Methods for the examination and characterization of the activated sludge community. In *Microbial Ecology of Activated Sludge*. Edited by R. J. Seviour & P. H. Nielsen. London: IWA Publishing.

Seviour, R. J., Mino, T. & Onuki, M. (2003). The microbiology of biological phosphorus removal in activated sludge systems. *FEMS Microbiol Rev* **27**, 99-127.

Seviour, R. J., Lindrea, K. C. & Oehmen, A. (2010). The activated sludge process. In *Microbial Ecology of Activated Sludge*. Edited by R. J. Seviour & P. H. Nielsen. London: IWA Publishing.

Seviour, R. J., Kragelund, C., Kong, Y., Eales, K., Nielsen, J. L. & Nielsen, P. H. (2008). Ecophysiology of the Actinobacteria in activated sludge systems. *Antonie Van Leeuwenhoek* **94**, 21-33.

Shapiro, O. H., Kushmaro, A. & Brenner, A. (2010). Bacteriophage predation regulates microbial abundance and diversity in a full-scale bioreactor treating industrial wastewater. *ISME J* **4**, 327-336.

Sharples, G. J., Corbett, L. M. & McGlynn, P. (1999). DNA structure specificity of Rap endonuclease. *Nucleic Acids Res* **27**, 4121-4127.

Short, C. M. & Suttle, C. A. (2005). Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl Environ Microbiol* **71**, 480-486.

Skennerton, C. T., Angly, F. E., Breitbart, M., Bragg, L., He, S., McMahon, K. D., Hugenholtz, P. & Tyson, G. W. (2011). Phage encoded H-NS: a potential achilles heel in the bacterial defence system. *PLoS One* **6**, e20095.

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

Soddell, J. A. (1999). Foaming. In *The Microbiology of Activated Sludge*. Edited by R. J. Seviour. Dordrecht: Kluwer Academic Publishers.

Soddell, J. A. (2002). Activated sludge-foaming. In *Encyclopedia of Environmental Microbiology*, pp. 4-8. Edited by G. Bitton. New York: John Wiley and Sons, Inc.

Soddell, J. A. & Seviour, R. J. (1990). Microbiology of Foaming in Activated-Sludge Plants. *J Appl Bacteriol* **69**, 145-176.

Soddell, J. A. & Seviour, R. J. (1995). Relationship between Temperature and Growth of Organisms Causing *Nocardia* Foams in Activated-Sludge Plants. *Water Res* **29**, 1555-1558.

Soddell, J. A. & Seviour, R. J. (1998). Numerical taxonomy of *Skermania piniformis* and related isolates from activated sludge. *J Appl Microbiol* **84**, 272-284.

Sorek, R., Kunin, V. & Hugenholtz, P. (2008). CRISPR - a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**, 181-186.

Stern, A. & Sorek, R. (2011). The phage-host arms race: shaping the evolution of microbes. *Bioessays* **33**, 43-51.

Steven, A. C., Greenstone, H. L., Booy, F. P., Black, L. W. & Ross, P. D. (1992). Conformational changes of a viral capsid protein: Thermodynamic rationale for proteolytic regulation of bacteriophage T4 capsid expansion, co-operativity, and super-stabilization by soc binding. *J Mol Biol* **228**, 870-884.

Stratton, H. M., Brooks, P. R., Griffiths, P. C. & Seviour, R. J. (2002). Cell Surface hydrophobicity and mycolic acid composition of *Rhodococcus* strains isolated from activated sludge foams. *J Ind Microbiol Biotechnol* **28**, 264-267.

Stratton, H. M., Seviour, R. J., Soddell, J. A., Blackall, L. L. & Muir, D. (1996). The opportunistic pathogen *Nocardia farcinica* is a foam-producing bacterium in activated sludge plants. *Lett Appl Microbiol* **22**, 342-346.

Sulakvelidze, A., Alavidze, Z. & Morris, J. G., Jr. (2001). Bacteriophage therapy. *Antimicrob Agents Chemother* **45**, 649-659.

Summer, E. J., Liu, M., Gill, J. J., Grant, M., Chan-Cortes, T. N., Ferguson, L., Janes, C., Lange, K., Bertoli, M. & other authors (2011). Genomic and functional analyses of *Rhodococcus equi* phages ReqiPepy6, ReqiPoco6, ReqiPine5, and ReqiDocB7. *Appl Environ Microbiol* **77**, 669-683.

Sunairi, M., Watanabe, T., Oda, H., Murooka, H. & Nakajima, M. (1993). Characterization of the genome of the *Rhodococcus rhodochrous* bacteriophage NJL. *Appl Environ Microbiol* **59**, 97-100.

Sutherland, I. W., Hughes, K. A., Skillman, L. C. & Tait, K. (2004). The interaction of phage and biofilms. *FEMS Microbiol Lett* **232**, 1-6.

Suttle, C. A. (2005). Viruses in the sea. *Nature* **437**, 356-361.

Tamaki, H., Zhang, R., Angly, F. E., Nakamura, S., Hong, P. Y., Yasunaga, T., Kamagata, Y. & Liu, W. T. (2012). Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. *Environ Microbiol* **14**, 441-452.

Ternes, T. A. (1998). Occurrence of drugs in German sewage treatment plants and rivers. *Water Res* **32**, 3245-3260.

Theron, J. & Cloete, T. E. (2002). Emerging waterborne infections: contributing factors, agents, and detection tools. *Crit Rev Microbiol* **28**, 1-26.

Thiel, K. (2004). Old dogma, new tricks—21st Century phage therapy. *Nat Biotech* **22**, 31-36.

Thingstad, T. F. & Lignell, R. (1997). Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquatic Microbial Ecology* **13**, 19-27.

Thomas, J. A., Soddell, J. A. & Kurtböke, D. I. (2002). Fighting foam with phages? *Water Sci Technol* **46**, 511-518.

Tree, J. A., Adams, M. R. & Lees, D. N. (2003). Chlorination of indicator bacteria and viruses in primary sewage effluent. *Appl Environ Microbiol* **69**, 2038-2043.

Twort, F. W. (1915). An investigation on the nature of ultra-microscopic viruses. *Lancet* **2**, 1241-1243.

Van Valen, L. (1973). A new evolutionary law. *Evol theory* **1**, 1-30.

ViroCyt White Paper: An Overview of Virus Quantification Techniques. United States: ViroCyt.

Wang, I. N., Smith, D. L. & Young, R. (2000). Holins: the protein clocks of bacteriophage infections. *Annu Rev Microbiol* **54**, 799-825.

Wang, X., Kim, Y., Ma, Q., Hong, S. H., Pokusaeva, K., Sturino, J. M. & Wood, T. K. (2010). Cryptic prophages help bacteria cope with adverse environments. *Nat Commun* **1**, 147.

Wanner, J. (1994). *Activated sludge bulking and foaming control*. Lancaster: Technomic Publishing Company Inc.

Watanabe, H., Kitamura, T., Ochi, S. & Ozaki, M. (1997). Inactivation of pathogenic bacteria under mesophilic and thermophilic conditions. *Water Sci Technol* **36**, 25-32.

Weigele, P. R., Pope, W. H., Pedulla, M. L., Houtz, J. M., Smith, A. L., Conway, J. F., King, J., Hatfull, G. F., Lawrence, J. G. & other authors (2007). Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. *Environ Microbiol* **9**, 1675-1695.

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiol Rev* **28**, 127-181.

Weinbauer, M. G. & Rassoulzadegan, F. (2004). Are viruses driving microbial diversification and diversity? *Environ Microbiol* **6**, 1-11.

Weitz, J. S. & Dushoff, J. (2007). Alternative stable states in host–phage dynamics. *Theor Ecol* **1**, 13-19.

Weld, R. J., Butts, C. & Heinemann, J. A. (2004). Models of phage growth and their applicability to phage therapy. *J Theor Biol* **227**, 1-11.

Westra, E. R., Swarts, D. C., Staals, R. H. J., Jore, M. M., Brouns, S. J. J. & Oost, J. v. d. (2012). The CRISPRs, They Are A-Changin': How Prokaryotes Generate Adaptive Immunity. *Annu Rev Genet* **46**, 311-339.

White, W. R. & Godfree, A. F. (1985). Pollution of freshwater and estuaries. *Soc Appl Bacteriol Symp Ser* **14**, 67S-79S.

Whitman, W. B., Coleman, D. C. & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* **95**, 6578-6583.

Withey, S., Cartmell, E., Avery, L. M. & Stephenson, T. (2005). Bacteriophages—potential for application in wastewater treatment processes. *Sci Total Environ* **339**, 1–18.

Wommack, K. E. & Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* **64**, 69-114.

Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S. & Nasko, D. J. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand Genomic Sci* **6**, 427-439.

Worley-Morse, T. O. & Gunsch, C. K. (2015). Modeling phage induced bacterial disinfection rates and the resulting design implications. *Water Res* **68**, 627-636.

Wu, Q. & Liu, W. T. (2009). Determination of virus abundance, diversity and distribution in a municipal wastewater treatment plant. *Water Res* **43**, 1101-1109.

Wyn-Jones, A. P. & Sellwood, J. (2001). Enteric viruses in the aquatic environment. *J Appl Microbiol* **91**, 945-962.

Xu, J., Hendrix, R. W. & Duda, R. L. (2004). Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. *Mol Cell* **16**, 11-21.

Yassin, A. F., Rainey, F. A., Brzezinka, H., Burghardt, J., Lee, H. J. & Schaal, K. P. (1995). *Tsukamurella inchonensis* sp. nov. *Int J Syst Bacteriol* **45**, 522-527.

Young, I., Wang, I. & Roof, W. D. (2000). Phages will out: strategies of host cell lysis. *Trends Microbiol* **8**, 120-128.

Zhang, Y., Hunt, H. K. & Hu, Z. (2013). Application of bacteriophages to selectively remove *Pseudomonas aeruginosa* in water and wastewater filtration systems. *Water Res* **47**, 4507-4518.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Res* **39**, W347-352.

8. Appendices

8.1. Appendix 1 – List of strains used in this study

Table 8.1 List of strains used in this study adapted from Petrovski *et al.* (2011a)

Organism	LAB ID(s) of strains
Dietzia maris	CON27^T
Gordonia aichiensis	CON22^T
Gordonia alkanivorans	CON72/BEN606
Gordonia amarae	CON9, CON44^T, BEN371, BEN374, BEN381, BEN386, BEN389
Gordonia desulfuricans	CON69^T
Gordonia hydrophobica	CON65^T, CON66
Gordonia malaquae	BEN700, CON59, CON60, CON67, A554^T, A448
Gordonia polyisoprenivorans	CON71
Gordonia rubropertincta	CON38^T
Gordonia sputi	CON31, CON48, CON49^T
Gordonia terrae	CON34^T, BEN601, BEN602, BEN603, BEN604, GOR9, G232, G238, G255, G290
Millisia brevis	J81, J82^T
Mycobacterium chlorophenolicus	CON24^T
Mycobacterium smegmatis	CON5
Mycobacterium fortuitum	CON21
Nocardia asteroides	CON12, CON23^T, BEN600
Nocardia brasiliensis	CON42^T
Nocardia brevicatena	CON43
Nocardia carnea	CON30^T
Nocardia nova	CON47^T
Nocardia otitidiscaviarum	CON14, CON15, CON25^T
Nocardia transvalensis	CON40^T
Rhodococcus coprophilus	CON18, CON41^T
Rhodococcus equi	CON10, CON28^T
Rhodococcus erythropolis	CON19, CON29^T, BEN703
Rhodococcus fascians	CON36^T
Rhodococcus globerulus	CON35^T
Rhodococcus rhodnii	CON46^T
Rhodococcus rhodochrous	CON3, CON11S, CON39^T
Rhodococcus ruber	CON33^T
***Rhodococcus* spp.**	J27, J71, J72
Rhodococcus tritomae	RHO1
Streptomyces griseus	CON5
Skermania piniformis	NM40^T, NM41, NM101, NM109, NM168, J8, J20, J50, J54
Tsukamurella inchonensis	BEN701, BEN702, BEN704, CON50^T
Tsukamurella paurometabola	CON37^T, CON51, CON52, CON53, CON54, CON55, CON61
Tsukamurella pseudospumae	TPS1
Tsukamurella spumae	CON58, CON62^T, TSP1
Tsukamurella tyrosinosolvens	CON57^T

^{T} *indicates type strain.*

8.2. Appendix 2 – Repeat sequences identified in this study

Table 8.1 Sequence repeats identified in the genome sequences of phages TIN2, TIN3, and TIN4

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TIN2-D1	68	68564-68622	GGTACTAGCGTACCATGACC--------ACGGCAAGTTT-GTCAACTTGCCGGATCAAAACGCGCTAG
		68128-68194	GGTACTAGCGTACCATGACCGCCTACCTACCGCAACTTTCGTTAGGTAGGC-GATCAAAACACGCTAG
TIN2-D2	52	18548-18599	GCTCGGCGAGTGGCTCGGCAAGGCGTGGACCTGGCTTCAGGAGAATGCGCCT
		18416-18467	GCTTGGCGAGTGGCTTGGCAAGGCTTGGACTTGGCTTCAGGAGAACGCGCCT
TIN2-D3	48	24448-24488	GAGCTCGCA--CGTCATCATGGGCGG---CGGTGGAGCTCG--AGTGG
		20097-20144	GAGATCGCAGACGTCATCATGGGCGGTCCAGGTGGTGCTCGCAAGTGG
TIN2-D4	37	46917-46953	TGCTCCTTCTCGCGCTGAGCTTCCTCGTAACGCTTCT
		14989-15022	TGCTCCTTCTCCCGCTG-GC-CCCTCTTAA-GCTTCT
TIN2-D5	27	31024-31050	GGCGGTGCGCGTGGCTCCGGTGGCGCA
		29408-29434	GGTGGTTCGGGTGGCTCCGGTGGCGCA
TIN2-D6	27	42772-42798	CTCGCCGCCGATGTAGACGTCGAGGCC
		12991-13017	CTCGCGGCCGACGACGACGTCGAGGCC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TIN2-D7	26	7377-7402	CTCCAGGAGGGCGACGTTCTCGTCGC
		1358-1383	CTCAACGGGGGCGACGTTCTCGACGC
TIN2-D8	26	46241-46266	CGTAGGCGCTACGGCGACCGGTGCGC
		26750-26775	CGTAGGAACTACGGCGACCGAGGCGC
TIN2-D9	26	44026-44051	TTTCCTCGAGCGAGCGCAGCTCGGGG
		43002-43027	TTTCCTCGAGCGAACGAAGCACCGGG
TIN2-D10	23	16087-16109	ACGTCACCAACGGTATGGCTGAG
		6481-6503	ACGTCACCAACGGTAACGCTGAG
TIN2-D11	23	27956-27978	CGTCAAGCAGGGACTCAACAACG
		23284-23306	CGTCAAGCCGGGACTCAACAGCG
TIN2-D12	21	45568-45588	AGCCGCCGAGAACGCAGTACT
		21995-22015	AGCCGCCGAGAACGCAGTACT
TIN2-D13	21	49318-49338	CCCGGCCTTGTCCACCAGAAC
		22400-22420	CCCGGCCTTGTGCAGCAGAAC
TIN2-D14	20	71859-71878	TCTCGAGGTAATCGGCGAGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		46552-46571	TCTCGATGTAATCGGCGAGC
TIN2-D15	17	56252-56268	TCTTGTCGTCGTCGGAC
		40406-40422	TCTTGTCGTCGTCGGAC
TIN2-D16	16	47673-47688	CGGCGTGCAGTACTTC
		12672-12687	CGGCGTGCAGTACTTC
TIN2-D17	16	72512-72527	TCGGCACGCTTGCGGT
		58694-58709	TCGGCACGCTTGCGGT
TIN2-I1	36	57698-57733	CATCACTCGAGCTCAACCCCCTCCTGCATGATGCCG
		24490-24457	CACCACTCGAGCTCCACCGCCGCC--CATGATGACG
TIN2-I2	35	68279-68311	GGACCG--TCGCCGAACCGATCGAGGGAGGTGTTC
		4225-4191	GGACCGGCTCGCCGAACCGCTCGAAGTACTTGTTC
TIN2-I3	31	58579-58609	CTTCTTGGCGTACTCGATCTCCGACTTGAGA
		11114-11084	CTGCTTGGCGAACTCGATCTCCTGCTCGAGA
TIN2-I4	29	66356-66384	CCTTGATGTACTCGCCCGCCGGGAGCGCG
		12613-12585	CCTCGATGCGCTCGCCCGCCGCGAGGGCG
TIN2-I5	26	53005-53030	CTGGAGCTTCTCGAGCTGTTTGTGGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		16850-16825	CTGGAGATTCTCGAGCTGCTCGTTGA
TIN2-I6	24	66570-66593	GGCGTCGGCGGCGGCCTGCTTGGC
		19027-19005	GGCG-CGGCGGCAGCCTGCTTGGC
TIN2-I7	24	45081-45104	CGCTTGTCCTTGTTGATGTTGCCG
		8348-8325	CGCTTGATCTTGTTGATGTCGCCG
TIN2-I8	23	40382-40404	CCACCTCGTTGTAGGCGGTCTCG
		9586-9564	CCACCTCGTTCGAGGCGGTCTCG
TIN2-I9	23	60989-61011	TCCCGGCGAGGTCGTCGATCTCG
		54183-54161	TCCCGGCGAGGGCTTCGATCTCG
TIN2-I10	22	51277-51298	CGATCTTGTCGCGCGCCTTCAC
		34932-34911	CGATCTTGTCGCCAGCCTTCAC
TIN2-I11	22	75721-75742	CAGGCTGCACGGAAATGCCGTG
		40793-40772	CAGGCTGCACGGAAAGGCCTTG
TIN2-I12	22	69402-69423	CGAAACTAAAGACTTATTTTAG
		69371-69350	CGAAACTAAAGACTTTTTCTAG
TIN2-I13	20	75082-75101	TGGTCAGCAGGTGTCCACTG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		10422-10403	TGGTCAGCAGGTCTCCACTG
TIN2-I14	19	56989-57007	GAGGGTCGTCCTCCTCGGC
		48180-48162	GACGGTCGTCCTCCTCGGC
TIN2-I15	16	47637-47652	GGCGTCGGGGTCGGTC
		43366-43351	GGCGTCGGGGTCGGTC
TIN2-I16	16	76925-76940	CAAGCCGCGCGAGATG
		45425-45410	CAAGCCGCGCGAGATG
TIN2-I17	16	65539-65554	CGCTGTCGTGACGCCC
		63933-63918	CGCTGTCGTGACGCCC
TIN3-D1	246	68520-68760	TAAGTATATATGTAGTTATACATA--CTAAGTATATGTATGTACTTATGTATACTATGTA-- TATATACTTAGTTATAAGTACTAACGATATGTGTATGTATATCTTTAGTATGAGTAT- CTAAGTATATTGTTATAGTAAGTAATGATGATGAGTAATGATATATCTTTATATTAAGTAATGATATGTGTTAA TGATATATACTTAGTTATGAGAGCTAATGGTATGCGGAAATAAATACTTAGT
		68414-68626	TAAGTATATGGGTA--TATAGATAGGCTAAGTATAT-- AGGTACTTATATAGGGTAGCTAAGTATATACTTAGGTATGAGTACTAATGATATATAGAT---------- AGTACT-GTATACTAAGTATATA--TGTAGTTATACAT-------ACTAAGTATATGTATGTAC-TTATGTA- TACTATGT-------ATATATACTTAGTTATAAGTACTAACGATATGTGTATGTATATCTTTAGT
TIN3-D2	45	67780-67824	TTTGCAACCCTTTCGGTGTTGTCGTTGTGTGGTACTAGTCTATCA
		67448-67491	TTTTCAACCCTTTCGGTGTTGT-GTTCGTTGGTACTAGTCTAGCA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TIN3-D3	34	74079-74111	TCGA-CCTGAACCTCGACGCGCCGCCTACTGGTT
		9725-9758	TCGATCCTGAGACTCGCGCCGCCGCCTACTGGTT
TIN3-D4	30	68712-68741	TATATACTTAGTTATGAGAGCTAATGGTAT
		68472-68501	TATATACTTAGGTATGAGTACTAATGATAT
TIN3-D5	27	42229-42255	CTCGCCGCCGATGTAGACGTCGAGACC
		13082-13108	CTCGCCGCCGACGACGACGTCGAGGCC
TIN3-D6	26	69808-69833	GTTTTACAGTTATTTATGGATTTTAT
		69516-69541	GTTTTACAGTTATTTATGGTTTATAT
TIN3-D7	24	58560-58583	GTCGAAGGCCAACCGAGCCGCGTC
		56000-56023	GTCGAGGGCCGCCCGAGCCGCGTC
TIN3-D8	23	72767-72789	TGAACGACTGAGACTTGCCAACA
		44208-44230	TGAACGACTGAGACTGGCCATCA
TIN3-D9	21	52505-52525	CTCGGCGAGCGCATCGAGCGC
		13130-13150	CTCGACGAGCGCATCGAGCGC
TIN3-D10	21	56052-56072	TCGGCCTCGAGCTCCACGCGG
		54475-54495	TCGGCCTCGAGCGCCACGCGG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TIN3-D11	20	56834-56853	TGTCGTTGTCGAGGTCGATG
		40174-40193	TGTAGTTGTCGAGGTCGATG
TIN3-D12	20	54617-54636	GCTTGAGCTCCACGCGCTTG
		52531-52550	GCTTGAGCTCCGCGCGCTTG
TIN3-D13	19	55052-55070	CCGTTGGTGTCGTTGATCG
		53544-53562	CCGTTGGTGTCGTTGATCG
TIN3-D14	19	58562-58580	CGAAGGCCAACCGAGCCGC
		57881-57899	CGAAGGCCAACCGAGCCGC
TIN3-D15	19	54844-54862	CATGAACTTCACGAGGTCG
		42130-42148	CATGAACTTCACGATGTCG
TIN3-D16	19	70459-70477	GCCAGTCGAGCGCTTCTTC
		51990-52008	GCCACTCGAGCGCTTCTTC
TIN3-D17	16	75994-76009	GAACACGATCAGCTTG
		44635-44650	GAACACGATCAGCTTG
TIN3-D18	16	64158-64173	TGCGGGACTTGTGCAG
		47452-47467	TGCGGGACTTGTGCAG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TIN3-D19	16	65425-65440	CTCGGTGAAGGCGTCG
		63063-63078	CTCGGTGAAGGCGTCG
TIN3-I1	39	68814-68852	TCTAAGAAAAGTCTTTAGTGGCACTAGCTAAAGATATAC
		68395-68357	TCTAAAAGAAGTCTTTAGTTTCAGTACCTACACATCTAC
TIN3-I2	31	59016-59046	CTCGGATCGAGCGTAGGACTGAAGTGCGGGA
		13263-13233	CTCGAGTCGAGCGTAGATCAGAAGTGCGGGA
TIN3-I3	31	43660-43687	CTCGGCGA---CCGGCGACAGGCCGGTCGTC
		8014-7984	CTCGGCGATGGCCGGGGCCAGGCCGGTCGTC
TIN3-I4	29	46955-46982	GGTCTT-GGCGACCTTCTTGAACTTCGGC
		2233-2205	GGTCTTCGGAGAACTCCTTGAACTTCGGC
TIN3-I5	29	65074-65102	CTCGAAGTCCGCGACCTCGTCGTCGGCGG
		13114-13086	CTTGAGGGCCTCGACGTCGTCGTCGGCGG
TIN3-I6	27	46659-46685	GCGGCGGGCGCGGCAGCGGGCTGCTTC
		19114-19088	GCTGCGGGCGCGGCGGCTGCCTGCTTC
TIN3-I7	26	71620-71645	CGAAAGCATGCTGGCCATGAGGTCGT
		4402-4377	CGACAGGATGGTCGCCATGAGGTCGT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TIN3-I8	25	65532-65556	CCTGGGGGGCCTCGACGTCGTTGTC
		13115-13091	CCTTGAGGGCCTCGACGTCGTCGTC
TIN3-I9	21	23140-23159	GCCGATCCCGAACA-ACCCTG
		19997-19977	GCCGATCCCGAACATACCCTG
TIN3-I10	21	72626-72646	GCGAGAGCTGCTTGATCAGAT
		33865-33846	GCGA-AGCTGCTTGATCAGAT
TIN3-I11	19	68950-68968	CGTACTGAATGTACTCTTC
		5267-5249	CGTACTGAATGAACTCTTC
TIN3-I12	19	58235-58253	CCTCGTCTGGGTCAACGAT
		10799-10781	CCTCGTCCGGGTCAACGAT
TIN3-I13	19	42051-42069	TCTTCTCGGGGTTGTACTC
		34569-34551	TCTTCTTGGGGTTGTACTC
TIN3-I14	16	64773-64788	CGTCGCCGTACTCGTC
		24419-24404	CGTCGCCGTACTCGTC
TIN3-I15	16	61542-61557	AGGCGGCGCGCCGCAC
		48263-48248	AGGCGGCGCGCCGCAC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TIN3-I16	16	66856-66871	GTGCGCCGCGATCTTC
		61210-61195	GTGCGCCGCGATCTTC
TIN4-D1	246	68467-68707	TAAGTATATATGTAGTTATACATA--CTAAGTATATGTATGTACTTATGTATACTATGTA-- TATATACTTAGTTATAAGTACTAACGATATGTGTATGTATATCTTTAGTATGAGTAT- CTAAGTATATTGTTATAGTAAGTAATGATGATGAGTAATGATATATCTTTATATTAAGTAATGATATGTGTTAA TGATATATACTTAGTTATGAGAGCTAATGGTATGCGGAAATAAATACTTAGT
		68361-68573	TAAGTATATGGGTA--TATAGATAGGCTAAGTATAT-- AGGTACTTATATAGGGTAGCTAAGTATATACTTAGGTATGAGTACTAATGATATATAGAT---------- AGTACT-GTATACTAAGTATATA--TGTAGTTATACAT-------ACTAAGTATATGTATGTAC-TTATGTA- TACTATGT-------ATATATACTTAGTTATAAGTACTAACGATATGTGTATGTATATCTTTAGT
TIN4-D2	45	67727-67771	TTTGCAACCCTTTCGGTGTTGTCGTTGTGTGGTACTAGTCTATCA
		67395-67438	TTTTCAACCCTTTCGGTGTTGT-GTTCGTTGGTACTAGTCTAGCA
TIN4-D3	34	74026-74058	TCGA-CCTGAACCTCGACGCGCCGCCTACTGGTT
		9673-9706	TCGATCCTGAGACTCGCGCCGCCGCCTACTGGTT
TIN4-D4	30	68659-68688	TATATACTTAGTTATGAGAGCTAATGGTAT
		68419-68448	TATATACTTAGGTATGAGTACTAATGATAT
TIN4-D5	27	42176-42202	CTCGCCGCCGATGTAGACGTCGAGACC
		13030-13056	CTCGCCGCCGACGACGACGTCGAGGCC
TIN4-D6	26	69755-69780	GTTTTACAGTTATTTATGGATTTTAT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		69463-69488	GTTTTACAGTTATTTATGGTTTATAT
TIN4-D7	24	58507-58530	GTCGAAGGCCAACCGAGCCGCGTC
		55947-55970	GTCGAGGGCCGCCCGAGCCGCGTC
TIN4-D8	23	72714-72736	TGAACGACTGAGACTTGCCAACA
		44155-44177	TGAACGACTGAGACTGGCCATCA
TIN4-D9	21	52452-52472	CTCGGCGAGCGCATCGAGCGC
		13078-13098	CTCGACGAGCGCATCGAGCGC
TIN4-D10	21	55999-56019	TCGGCCTCGAGCTCCACGCGG
		54422-54442	TCGGCCTCGAGCGCCACGCGG
TIN4-D11	20	56781-56800	TGTCGTTGTCGAGGTCGATG
		40121-40140	TGTAGTTGTCGAGGTCGATG
TIN4-D12	20	54564-54583	GCTTGAGCTCCACGCGCTTG
		52478-52497	GCTTGAGCTCCGCGCGCTTG
TIN4-D13	19	54999-55017	CCGTTGGTGTCGTTGATCG
		53491-53509	CCGTTGGTGTCGTTGATCG
TIN4-D14	19	58509-58527	CGAAGGCCAACCGAGCCGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		57828-57846	CGAAGGCCAACCGAGCCGC
TIN4-D15	19	54791-54809	CATGAACTTCACGAGGTCG
		42077-42095	CATGAACTTCACGATGTCG
TIN4-D16	19	70406-70424	GCCAGTCGAGCGCTTCTTC
		51937-51955	GCCACTCGAGCGCTTCTTC
TIN4-D17	16	75941-75956	GAACACGATCAGCTTG
		44582-44597	GAACACGATCAGCTTG
TIN4-D18	16	64105-64120	TGCGGGACTTGTGCAG
		47399-47414	TGCGGGACTTGTGCAG
TIN4-D19	16	65372-65387	CTCGGTGAAGGCGTCG
		63010-63025	CTCGGTGAAGGCGTCG
TIN4-I1	39	68761-68799	TCTAAGAAAAGTCTTTAGTGGCACTAGCTAAAGATATAC
		68342-68304	TCTAAAAGAAGTCTTTAGTTTCAGTACCTACACATCTAC
TIN4-I2	31	58963-58993	CTCGGATCGAGCGTAGGACTGAAGTGCGGGA
		13211-13181	CTCGAGTCGAGCGTAGATCAGAAGTGCGGGA
TIN4-I3	31	43607-43634	CTCGGCGA---CCGGCGACAGGCCGGTCGTC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		7962-7932	CTCGGCGATGGCCGGGGCCAGGCCGGTCGTC
TIN4-I4	29	46902-46929	GGTCTT-GGCGACCTTCTTGAACTTCGGC
		2181-2153	GGTCTTCGGAGAACTCCTTGAACTTCGGC
TIN4-I5	29	65021-65049	CTCGAAGTCCGCGACCTCGTCGTCGGCGG
		13062-13034	CTTGAGGGCCTCGACGTCGTCGTCGGCGG
TIN4-I6	27	46606-46632	GCGGCGGGCGCGGCAGCGGGCTGCTTC
		19062-19036	GCTGCGGGCGCGGCGGCTGCCTGCTTC
TIN4-I7	26	71567-71592	CGAAAGCATGCTGGCCATGAGGTCGT
		4350-4325	CGACAGGATGGTCGCCATGAGGTCGT
TIN4-I8	25	65479-65503	CCTGGGGGGCCTCGACGTCGTTGTC
		13063-13039	CCTTGAGGGCCTCGACGTCGTCGTC
TIN4-I9	21	23088-23107	GCCGATCCCGAACA-ACCCTG
		19945-19925	GCCGATCCCGAACATACCCTG
TIN4-I10	21	72573-72593	GCGAGAGCTGCTTGATCAGAT
		33812-33793	GCGA-AGCTGCTTGATCAGAT
TIN4-I11	19	68897-68915	CGTACTGAATGTACTCTTC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		5215-5197	CGTACTGAATGAACTCTTC
TIN4-I12	19	58182-58200	CCTCGTCTGGGTCAACGAT
		10747-10729	CCTCGTCCGGGTCAACGAT
TIN4-I13	19	41998-42016	TCTTCTCGGGGTTGTACTC
		34516-34498	TCTTCTTGGGGTTGTACTC
TIN4-I14	16	64720-64735	CGTCGCCGTACTCGTC
		24367-24352	CGTCGCCGTACTCGTC
TIN4-I15	16	61489-61504	AGGCGGCGCGCCGCAC
		48210-48195	AGGCGGCGCGCCGCAC
TIN4-I16	16	66803-66818	GTGCGCCGCGATCTTC
		61157-61142	GTGCGCCGCGATCTTC

I indicates inverted repeat; D indicates direct repeat.

Table 8.2 Direct repeats identified in the genome sequence of phage SPI1

Repeat number	Size (bp)	Coordinates	Sequence alignment
1	347	27772-28105	GACCGACAACATGGTCCGGCTCCGCTGA----- TCTCGTACCGTCGCAGCGGATAGCGGCGGACCGTCCCATCACCTGGGCGGGGTCGGC- GAGCGTCGTACCCGCCCAGAAGATCGACGTCGCTCGGT- CGACCACATGGACAGGTACCGCCGAGCTGAACGACAAGCCCGG--CGTCGA-- TCCCTCCGGTACGACGACGTGGGCCGGTGCCGTTGACCTGACC-- CGTGTCCGGTCGGTCGATCCCTCCGGTACGACGACGTGGGCCGGTGCCGTTGACCTGACCCGTGTCCGGTCGGCGGAC GCGACCGGCTCAACCACGTGGACCGAGACCGCTGCCCTCAC
		27712-28045	GAGCGACAGCGTGGTCCGGCTCCGC-GAGCGTCGCGCGGACCTACGCCGCTGA-CGCGACG--- CGACCGACAACATGGTCCGGCTCCGCTGATC-TCGTACCGTCGCAGCGGATAG- CGGCGGACCGTCCCATCACCTGGGC-GGGGTCGGCGAGC-GTCGTACCCGCCCAGAAGATCGACGTCGCT-CGGT- CGACCACATGGACAGGTACCGCCGAGCTGAACGACAAGCCCGG-- CGTCGATCCCTCCGGTACGACGACGTGGGCCGGTGCCGTTGACCTGACCCGTGTCCGGTCGGTCGATCCCTCCGGTAC GACGACGTGGGCCGGTGCCGTTGACCTGAC
2	287	28059-28342	GCGGACGCGACCGGCTCAACCACGTGGACCGAGACCGCTGCCCTCACACCATCGCAGCGGATAGCGGTGAATCGCCCG ACCACGTGGACCGGATCGGCGGATCTCCTACTCCGTCAGTACTACGACATCGATTCGTCCGGCACGGTCACATGGTCC GGCGCGGTGGACCTCGCGCAGGGGGT--ATCGGGATACCGACCCACCACGTGGTCCGGAACCGCCGACCTGACCC- CGTCCGGTCGACTCAACGCGGCGGGCACGACCACGTGGGCGCGGACCGTCGACCT
		27759-28042	GCTGACGCGACGCGACCGACAACATGGTCCGGCTCCGCTGATCTCGTACCGTCGCAGCGGATAGCGGCGGACCGTCCC ATCACCTGGGCGGGGTCGGCGAGCGTCGTACCCGCCCAGAAGATCGACGTCGCTCGGTCGACCACATGGACAGGTACC GCCGAGCTGAACGACAAGCCCGGCGTCGATC-- CCTCCGGTACGACGACGTGGGCCGGTGCCGTTGACCTGACCCGTGTCCGGTCG- GTCGATCCCTCCGGTACGACGACGTGGGCCGGTGCCGTTGACCT
3	154	17177-17330	GGGGGTGATGGTGGCGTTCATGTCGTTCTCCTCTGTCTCGGTGGTAGAACCATTACACCACGGTCACCGTGCTAGTTG GCAACCCCACACCGAAAAATAGTTTCCGACCATGTAGTTGAAGTTTTCAACTACCTGCTCTGACCTGGGGAAACGG

Repeat number	Size (bp)	Coordinates	Sequence alignment
		16482-16634	GGTGGTGGTGGTGGCGTTCATGTCGTTCTCCTCTGTTTCGGTGGTAGAACCATTACACCACGGTCACCGTGCTAGTTG GCAACCCCACACCGAAAAATAGTTTCCGACCATGTAGTTGAAG-TTTCAACTACCTACTATGACCAGCAGAAACGG
4	146	32672-32817	TTCGTGGCGGTGCTGCACGATCGCGCGGGTCGCCCGGACCCGGACTGGACGGGGAAAGCCACCACACCGCGACCCGCG CCCGCTACCAGGGCCGGGGTGTGCGCGCTGACCGGCGCGTACGGGGATGTCGTCGACCTCCAGCATGG
		31918-32063	TTCGTGCGCGTACTGCACGATCTCGCCGGCTGCCCCGAGCCAGATTGGACCGGCAAAGCGCTCACCGCTCGACGCGCA CCGACGACCCGATCCGGAGTGTGCGCACTGACTGGGGAGCGCACGGAAGTCGTCGACCTTCAGCACGG
5	72	16572-16643	CGAAAAATAGTTTCCGACCATGTAGTTGAAGTTTCAACTACCTACTATGACCAGCAGAAACGGCAAACCCCC
		16162-16231	CGAAAAATAGTT-CCGACCATGTAGTTGAATTTTCAACTACCCCCGTTGACCTGGGTAAATGG-AAACCGCC
6	68	25013-25080	GGCGGCCGGGCCAGCGGTATCGGGTTCATGCCGAAGCAGACCCTGTTGCCCGAGCGAGTGCTCTCACC
		23735-23802	GGCGGCGTGGCGGTCGGCAAGGGCTTCATGCCGAAGGACGTCATAGCGCCGGAGCGCGTGCTGTCCCC
7	64	17267-17330	CGAAAAATAGTTTCCGACCATGTAGTTGAAGTTTTCAACTACCTGCTCTGACCTGGGGAAACGG
		16162-16223	CGAAAAATAGTT-CCGACCATGTAGTTGAA-TTTTCAACTACCCCCGTTGACCTGGGTAAATGG
8	62	51003-51062	ATGGAGGGGGAAGAGCACGGTGACCGTGCTATGGTGGAGGGATGA--GACCACCGACACGAG
		4268-4328	ATGGGTGGCGAGTAGCACGGTGACCGTGCTACTAT-GAGGGAGTACCAACCACCGACCCGAG
9	61	45558-45614	GTAGGTCATGGTCATC-TCCTCGGTAGGTGGT---CCCTGGACTATAGCACGGTCACCGTG
		16041-16098	GTACGTCATGGGGATCATCCTCTCTCGGTGGTGAACCC---ACTATAGCACGGTCACCGTG
10	56	54273-54328	TAGCACGGTGACCGTGCTATAGTGGAGGGACGCTCACCAACCACCGACCACAGGAG
		4280-4331	TAGCACGGTGACCGTGCTACTAT-GAGGGA---GTACCAACCACCGACCCGAGGAG

Repeat number	Size (bp)	Coordinates	Sequence alignment
11	50	51010-51059	GGGAAGAGCACGGTGACCGTGCTATGGTGGAGGGATGAGACCACCGACAC
		58-102	GGGAAATGCACGGTGACCGTGCTATAGTGGAGG-----GACCACCGACAC
12	46	32056-32099	CAGCACGGTCACCGTGCTATGGTGGTA--CCACCACCGACAAAGAG
		16537-16581	CACCACGGTCACCGTGCTA-GTTGGCAACCCCACACCGAAAAATAG
13	46	32056-32099	CAGCACGGTCACCGTGCTATGGTGGTA--CCACCACCGACAAAGAG
		17232-17276	CACCACGGTCACCGTGCTA-GTTGGCAACCCCACACCGAAAAATAG
14	42	17209-17250	CTGTCTCGGTGGTAGAACCATTACACCACGGTCACCGTGCTA
		16060-16101	CTCTCTCGGTGGTGAACCCACTATAGCACGGTCACCGTGCTA
15	41	38343-38383	TCTCGGTGGCATCGCCAACACTAGCACGGACACCGTGCTAG
		17212-17251	TCTCGGTGGTAGAACCATTAC-ACCACGGTCACCGTGCTAG
16	40	20060-20096	CGGGTCGATGTCGA-CTCCAGCC--CCGCCGCTCTCGCGG
		19976-20015	CGGGTCGATGTCGAGATCGACCCCACCGCCGCGATCGCGG
17	39	52199-52237	CGCGGAGACCGTCGGCGAGATCACGGTGGCGCGACAGCG
		5896-5930	CGCGGAGACCGTCCGAAAGATC---GT-GCGCGACATCG
18	38	17019-17053	GCGCGCAGGGTGTCG-ATCAG--GGCGGACGCGGCGGT
		2745-2782	GCGCGGATGACGTCGGATCAGTGGGCGGACGCGGCGGT

Repeat number	Size (bp)	Coordinates	Sequence alignment
19	37	16519-16555	TCGGTGGTAGAACCATTACACCACGGTCACCGTGCTA
		16065-16101	TCGGTGGTGAACCCACTATAGCACGGTCACCGTGCTA
20	37	54261-54295	GTCCCTCCACTATAGCACGGTGACC--GTGCTATAGT
		50382-50418	GTCCGTCGAATCTAGCACGGTGACCGTGTGCTATAGT
21	33	13085-13116	GCGAGACGCTGGGCT-GCTCGGGGAGCCCGACC
		8321-8352	GCGAGACGCTGCGCTATCCCGGGGA-CCCGACC
22	32	42767-42798	GAACGCGGTCACGGATGATCGACTATGTCGTC
		42020-42051	GAACGCGGTCACGGATGATCGACGACGTCGTC
23	30	54267-54296	CCACTATAGCACGGTGACCGTGCTATAGTG
		16077-16106	CCACTATAGCACGGTCACCGTGCTATAGTG
24	30	47687-47716	CGCTCCCGTTGTGCCGCGCCGCGTCGGTCA
		18476-18505	CGCTCGCGCAGTGCCGCGCCGCCACGCTCA
25	29	54275-54303	GCACGGTGACCGTGCTATAGTGGAGGGAC
		65-93	GCACGGTGACCGTGCTATAGTGGAGGGAC
26	29	54274-54302	AGCACGGTGACCGTGCTATAGTGGAGGGA
		51016-51044	AGCACGGTGACCGTGCTATGGTGGAGGGA

Repeat number	Size (bp)	Coordinates	Sequence alignment
27	29	45224-45252	CGCTCGGAAGCCGGTGCCCGAGCGCGGTG
		13926-13952	CGCTCGGAAGCCG--CCGCGCGCGCGGTG
28	28	54261-54288	GTCCCTCCACTATAGCACGGTGACCGTG
		45587-45614	GTCCCTGGACTATAGCACGGTCACCGTG
29	28	41490-41517	CGCTCAGCGGTAACGCGGCGGTGCGCGT
		1655-1682	CGCTCAGCGGTACCGCGCCGCTGAGCGT
30	27	19437-19461	CGACGAGATCGCCGCCATGC--GAGCG
		10257-10283	CGACGAGATCGCCGCGTTGCTGGAGCG
31	27	26989-27013	GCGGGTCGATGTCG-CT-CAGCTCCGC
		20059-20085	GCGGGTCGATGTCGACTCCAGCCCCGC
32	26	53887-53912	CGGCGGACCCCGGGGGCGCGGCGGCG
		24321-24346	CGGCGGATCTCGGGGGCGCGGCGGCG
33	26	7451-7476	GCGCTCGCGCTCGCGACGTCGAGCCC
		4758-4783	GCGATCGCGCTCGCGTCGTTGACCCC
34	26	9449-9474	TCGAGCTGGTCGCCGCGCTGGCGGTC
		7437-7462	TCGAGGGGGTCGCCGCGCTCGCGCTC

Repeat number	Size (bp)	Coordinates	Sequence alignment
35	25	49190-49214	CTGCTGCAACTCCGGCTCGCTCACC
		6835-6859	CTGGTACGACTCCAGCTCGCTCACC
36	24	51016-51039	AGCACGGTGACCGTGCTATGGTGG
		32057-32080	AGCACGGTCACCGTGCTATGGTGG
37	23	32057-32079	AGCACGGTCACCGTGCTATGGTG
		16084-16106	AGCACGGTCACCGTGCTATAGTG
38	23	37817-37839	CGCCGCGAACGCGGGGGGCGACG
		628-650	CGCCGCGAACGCGCGCGGCGACG
39	23	39371-39393	TCCTCGGGGTATCCGCCGCGATC
		26057-26079	TCCCCGTGGTATCCGCCGCGATC
40	23	50396-50418	GCACGGTGACCGTGTGCTATAGT
		65-85	GCACGGTGACC--GTGCTATAGT
41	23	11088-11110	GCTGATCGAGTGGTTCATGGCGG
		5447-5468	GCTGATCGAGTGG-TCATGGTGG
42	23	41752-41773	CGGT-GACCGATCCCCGGACCAG
		30809-30831	CGGTCGACCGATCCCCGGATCAG

Repeat number	Size (bp)	Coordinates	Sequence alignment
43	23	12668-12690	CCGCGCTGCGCACGTCGCTGGGG
		414-436	CCGCGCTGCGCACCTATCTGGGG
44	23	33801-33823	CGCTGTCGTCGATCTCCCACGGG
		15226-15248	CGCTGTGCGCGATCTCCCACGGG
45	23	38668-38690	GTCCGGCACCGCGCTCGACATCA
		23128-23150	GTCCGGGCGCGCGCTCGACATCA
46	22	16085-16106	GCACGGTCACCGTGCTATAGTG
		65-86	GCACGGTGACCGTGCTATAGTG
47	22	28518-28539	TCGACCCCCGCCGAGTCCGAGA
		2644-2664	TCGACCCC-GCCGAGTCCGGGA
48	22	25617-25638	GTCGTTCTCCGCGCTGCTGGTC
		23917-23937	GTCGATCTCCGCGCTGC-GGTC
49	22	26202-26223	CGACCCGCTCACGGTGTCGCTC
		25353-25373	CGACCCGCTCACCGTGT-GCTC
50	21	52291-52311	CCGGTGTCCGGTCGCCGGACG
		28045-28065	CCCGTGTCCGGTCGGCGGACG

Repeat number	Size (bp)	Coordinates	Sequence alignment
51	20	9630-9649	GGTGGCGGCGGCGGTCATCG
		3098-3117	GGTGGCGGCGGCGGTCGTCG
52	20	49391-49410	CGCGCTGAGCCGGCCGGTCG
		2426-2445	CGCGCTCGGCCGGCCGGTCG
53	20	51901-51920	ACGACGAGCAGATCGAGACC
		5539-5558	ACGACGCGGAGATCGAGACC
54	20	55480-55499	CGCGCTGCGCGAACTCATCG
		6922-6941	CGCGCTGCGCGAACACCTCG
55	20	22198-22217	CGCGGACGCGCTCACCGAGC
		7908-7927	CGCGGACGCGCTCGACGAGC
56	20	27762-27781	GACGCGACGCGACCGACAAC
		9571-9590	GACGCGACGCGGGCGACAAC
57	20	10171-10190	ACCGACGAGGAACTGGCCGA
		10324-10343	ACCGACGAGGACCCGGCCGA
58	20	43494-43513	GGCGGATCTCGGGGACACGG
		24322-24341	GGCGGATCTCGGGGGCGCGG

Repeat number	Size (bp)	Coordinates	Sequence alignment
59	20	53001-53020	TGATTACGCCGACCGGGTGT
		33937-33956	TGATTACGCCGCCCGCGTGT
60	20	52770-52789	GCTGATCGTCGCCGCGATCC
		39354-39373	GCCGATCGTCGCCGTGATCC
61	20	54478-54497	CGACCGGCTGGCCGCGCACG
		47546-47565	CGACCAGCTGGCCGCGCTCG
62	20	54348-54367	TCGAGCGTATGTCCGCGCTG
		49378-49397	TCGCGCGTTTGTCCGCGCTG
63	19	40644-40662	GTCCGCGCGGACGACGTCG
		2741-2759	GTCCGCGCGGATGACGTCG
64	19	7931-7949	TCCGGTCCGTGCCCGCGCT
		5196-5214	TCCGGTCCGTGCCGGCGCT
65	19	51910-51928	AGATCGAGACCGACGCGCT
		17428-17446	AGACCGAGACCGACGCGCT
66	19	42033-42051	GATGATCGACGACGTCGTC
		23482-23500	GATGATCGACGCCGTCGTC

Repeat number	Size (bp)	Coordinates	Sequence alignment
67	18	34654-34671	CACCGTCGGTGACGCGCT
		7869-7886	CACCGTCGGTGACGCGCT
68	18	35889-35906	GAGCCGACCGGGTGGGTG
		1190-1207	GAGCCGACCGGGTCGGTG
69	18	43221-43238	GGACCGGACCCCGGTCCG
		2615-2632	GGACCGGACCCGGGTCCG
70	18	7446-7463	TCGCCGCGCTCGCGCTCG
		3726-3743	TCGCCGCGCTCACGCTCG
71	18	52767-52784	CGAGCTGATCGTCGCCGC
		6751-6768	CGAGCTGATCGTCGGCGC
72	18	26701-26718	CCGATACCGTCGACCCCG
		7517-7534	CCGATACCGTCTACCCCG
73	18	10661-10678	GTGGGGACGCTCCGGCGG
		9242-9259	GTGGGCACGCTCCGGCGG
74	18	25846-25863	GGTCACGATCACCGGCAA
		11754-11771	GGTCACGATCACCGGTAA

Repeat number	Size (bp)	Coordinates	Sequence alignment
75	18	45749-45766	CCACCGCCGCCGCGATCG
		20418-20435	CCGCCGCCGCCGCGATCG
76	18	45755-45772	CCGCCGCGATCGCCACCT
		26069-26086	CCGCCGCGATCCCCACCT
77	18	28692-28709	ACCCGGACCCCGATATCG
		26692-26709	ACCCGGACCCCGATACCG
78	18	35945-35962	CCGGCTACGGTCCACACG
		29747-29764	CCGGCTACGGTCCTCACG
79	18	45720-45737	GCGCAAGACCTCGAGACG
		31428-31445	GCGCAAGACCTCGCGACG
80	18	53089-53106	CGGGTCGAGGTCCGCGAG
		43606-43623	CGCGTCGAGGTCCGCGAG
81	17	4282-4298	GCACGGTGACCGTGCTA
		65-81	GCACGGTGACCGTGCTA
82	17	22181-22197	ACCCTCGGGCAGGCGCT
		21866-21882	ACCCTCGGGCAGGCGCT

Repeat number	Size (bp)	Coordinates	Sequence alignment
83	16	11988-12003	GACGTCGACCCTGACG
		833-848	GACGTCGACCCTGACG
84	16	21635-21650	GGCGGGCTGTTCAGCG
		2562-2577	GGCGGGCTGTTCAGCG
85	16	50394-50409	TAGCACGGTGACCGTG
		4280-4295	TAGCACGGTGACCGTG
86	16	21662-21677	GCGGCGGACATCGCGC
		14998-15013	GCGGCGGACATCGCGC
87	16	19160-19175	CACCACGGTCACCGTG
		16537-16552	CACCACGGTCACCGTG
88	16	19160-19175	CACCACGGTCACCGTG
		17232-17247	CACCACGGTCACCGTG
89	16	36300-36315	TCGACCCGGACCCCGA
		26689-26704	TCGACCCGGACCCCGA
90	16	34645-34660	CGGGTCGACCACCGTC
		31508-31523	CGGGTCGACCACCGTC

Repeat number	Size (bp)	Coordinates	Sequence alignment
91	16	48070-48085	TCCGCGTCCGGCGCCG
		31569-31584	TCCGCGTCCGGCGCCG
92	16	46928-46943	CCGCGTCGAGGTCCGC
		43605-43620	CCGCGTCGAGGTCCGC
93	15	2682-2696	GTCATCGCCGCCGCG
		620-634	GTCATCGCCGCCGCG
94	15	52525-52539	CGTGAGCGATGCCGA
		3619-3633	CGTGAGCGATGCCGA
95	15	34894-34908	GCGGCGGCCGCGCGC
		3923-3937	GCGGCGGCCGCGCGC
96	15	11068-11082	CAGGCGCAGGCGTGG
		5744-5758	CAGGCGCAGGCGTGG
97	15	39353-39367	TGCCGATCGTCGCCG
		6522-6536	TGCCGATCGTCGCCG
98	15	20419-20433	CGCCGCCGCCGCGAT
		8961-8975	CGCCGCCGCCGCGAT

Repeat number	Size (bp)	Coordinates	Sequence alignment
99	15	21369-21383	GCGAGGAGGGGATGC
		11456-11470	GCGAGGAGGGGATGC
100	15	24097-24111	GGAAGCCGCCGCGCG
		13931-13945	GGAAGCCGCCGCGCG
101	15	42564-42578	GGAAACGGTGCCGGA
		17323-17337	GGAAACGGTGCCGGA
102	15	29510-29524	TGGGTGCTCAGCACC
		19149-19163	TGGGTGCTCAGCACC
103	15	21808-21822	CGCCGCGATCGCGGA
		20002-20016	CGCCGCGATCGCGGA
104	15	27332-27346	GACGGTGTCCGGAAC
		20240-20254	GACGGTGTCCGGAAC
105	15	44683-44697	TAGCGCGGTCACCAC
		26941-26955	TAGCGCGGTCACCAC
106	15	45600-45614	AGCACGGTCACCGTG
		32057-32071	AGCACGGTCACCGTG

Repeat number	Size (bp)	Coordinates	Sequence alignment
107	15	42416-42430	ATCGACCGCACCATC
		32544-32558	ATCGACCGCACCATC
108	15	49342-49356	GCGCCGCGACGATCG
		48268-48282	GCGCCGCGACGATCG
109	15	51016-51030	AGCACGGTGACCGTG
		50395-50409	AGCACGGTGACCGTG
110	15	53564-53578	TGACCCGCGTCGACC
		52970-52984	TGACCCGCGTCGACC

Table 8.3 Inverted repeats identified in the genome sequence of phage SPI1

Repeat number	Size (bp)	Coordinates	Sequence alignment
1	62	48934-48993	CCAGCGTCGCGGTC--CCCTGGCCGTACATGCCCGCGAGCACAGCGACCTTCGCGATCTTGC
		4468-4410	CCAGCATCACGGTCAGCCCCGACA--ACATGCCCGCGAGGACAAC-ATCCGCGCGATCGTGC
2	48	37282-37327	GCTCCGCAT--CCGGGTACCTCGGGCGGCGCCGGGTCATCGCGTGAGG
		18090-18045	GCTCAGCATGTCAGCGGACCTCGGGCGGCGCCTGG--AGCGCGTGCGG
3	43	43975-44015	GTCGTGCGC-GTGCCGGAGGATCGCGCGGACCACC-TCGGGGG
		23581-23540	GTCGAGCGCCGCTCCGG-GGATCGCGCGGACCGCCGGCGGGGG
4	40	17209-17248	CTGTCTCGGTGGTAGAACCATTACACCACGGTCACCGTGC
		104-65	CTGTGTCGGTGGTCCCTCCACTATAGCACGGTCACCGTGC
5	40	16514-16553	CTGTTTCGGTGGTAGAACCATTACACCACGGTCACCGTGC
		104-65	CTGTGTCGGTGGTCCCTCCACTATAGCACGGTCACCGTGC
6	39	17911-17949	AGTGCCGACTCCCCCGCCGCGTCCGCGCCGTGACGCGAC
		2794-2756	AGTGCCAACTCCACCGCCGCGTCCGCCCACTGATCCGAC
7	38	50394-50431	TAGCACGGTGACCGTGTGCTATAGTCGGTGTCACCTCC
		16101-16067	TAGCACGGTGACC--GTGCTATAGTGGGT-TCACCACC
8	38	28643-28680	CGAGACCACCCAGATCACCGCGCGCGATCCGGGGGCTT
		13967-13933	CGCGACCAGCCAAAACACCGCGCGCG---CGGCGGCTT

Repeat number	Size (bp)	Coordinates	Sequence alignment
9	35	45580-45614	GTAGGTGGTCCCTGGACTATAGCACGGTCACCGTG
		100-66	GTCGGTGGTCCCTCCACTATAGCACGGTCACCGTG
10	35	16065-16099	TCGGTGGTGAACCCACTATAGCACGGTCACCGTGC
		99-65	TCGGTGGTCCCTCCACTATAGCACGGTCACCGTGC
11	33	8734-8766	GTCACCGCGCCGAGCGTCGACCTCGGCGATGTC
		5904-5873	GTCTCCGCGT-GCGGGTCGTCCTCGGCGATGTC
12	31	28480-28510	GTGAGTCCGGGGAGGTCAGCACGGCGTCGGT
		25364-25336	GTGAG--CGGGTCGCTCAGCACGGCGTCGGT
13	30	54268-54297	CACTATAGCACGGTGACCGTGCTATAGTGG
		16106-16077	CACTATAGCACGGTGACCGTGCTATAGTGG
14	30	18628-18657	GCCTACCGTGACGATCGAGGGACAGGTGCC
		9330-9301	GCCTACCGTGACGAGCGCGAAGCAGGTACC
15	29	54261-54289	GTCCCTCCACTATAGCACGGTGACCGTGC
		93-65	GTCCCTCCACTATAGCACGGTCACCGTGC
16	29	29256-29284	GCGTTCGCGGTGCACACCGCGCGGGCGTC
		26940-26912	GCGTTCGCGGTGCCGGCCGCGCCGGGGTC
17	28	54276-54303	CACGGTGACCGTGCTATAGTGGAGGGAC

Repeat number	Size (bp)	Coordinates	Sequence alignment
		45614-45587	CACGGTGACCGTGCTATAGTCCAGGGAC
18	27	51018-51044	CACGGTGACCGTGCTATGGTGGAGGGA
		45614-45588	CACGGTGACCGTGCTATAGTCCAGGGA
19	27	29173-29199	TCCGCGACGAGCTGGTCGTGCCCGCCG
		28332-28306	TCCGCGCCCACGTGGTCGTGCCCGCCG
20	27	49458-49481	GCGACCA---CGTCCGCGCCGACCCCG
		8963-8937	GCGACCAGTGCGTCCGCGCCGACCTCG
21	26	44702-44727	CCGGGTCGGTGACCGCGACGATGTCG
		11683-11658	CCGGGATGGTGACCGCGAGGCTGTCG
22	26	52813-52838	CTGCCCGCGATCGCGAAGAGACTGGA
		49978-49953	CTGGCCGAGATCGCGAAGATCCTGGA
23	25	54273-54297	TAGCACGGTGACCGTGCTATAGTGG
		16555-16531	TAGCACGGTGACCGTGGTGTAATGG
24	25	54273-54297	TAGCACGGTGACCGTGCTATAGTGG
		17250-17226	TAGCACGGTGACCGTGGTGTAATGG
25	25	23677-23700	ATCG-GCGGCGGTCGATCTCGGGTC
		6521-6497	ATCGCGCGGCGGTCGATGTCGGATC

Repeat number	Size (bp)	Coordinates	Sequence alignment
26	25	54831-54855	CGCTGGAGGTGTGCGACCTCCGCGA
		15522-15498	CGCTGGCGGTCATCGACCTCCGCGA
27	25	35406-35430	CGGCGTGCCGACCGCCGAGTCGACC
		24937-24913	CGCCGCGCCGACCGCCGCGCCGACC
28	24	51016-51039	AGCACGGTGACCGTGCTATGGTGG
		16100-16077	AGCACGGTGACCGTGCTATAGTGG
29	24	54267-54290	CCACTATAGCACGGTGACCGTGCT
		32080-32057	CCACCATAGCACGGTGACCGTGCT
30	24	45599-45622	TAGCACGGTCACCGTGTCAGTCGC
		4298-4275	TAGCACGGTCACCGTGCTACTCGC
31	24	27812-27835	GCAGCGGATAGCGGCGGACCGTCC
		7203-7180	GCGGCGGTCAGCGGCGGACCGTCC
32	24	41462-41485	CACCCACTCGGGCGCGATCGTCAG
		14715-14692	CACCCACTCGGTCCCGAACGTCAG
33	23	17128-17150	CGGACAGGTCGCGCTCGTCGAGC
		7939-7917	CGGACCGGACGCGCTCGTCGAGC
34	23	45592-45614	TGGACTATAGCACGGTCACCGTG

Repeat number	Size (bp)	Coordinates	Sequence alignment
		16108-16086	TGCACTATAGCACGGTGACCGTG
35	23	17676-17698	TCGATCGCGTCCGTCACCATCGA
		1807-1785	TCGATCGCGTCCGCCATCAGCGA
36	23	6500-6522	CCGACATCGACCGCCGCGCGATT
		2905-2883	CCGCCCTCGACCGCCGCGCGCTT
37	23	19372-19394	CGTAGCGCACCGACCGCGTAGCC
		3842-3820	CGTAGCGCACCGCGCGGGTAGCC
38	22	19163-19184	CACGGTCACCGTGGATGTCCCG
		78-57	CACGGTCACCGTGCATTTCCCG
39	22	27584-27605	GTCGACGTACGTGGCCCCGGGG
		9821-9800	GTAGACGTAGGTGGCCCCGGGG
40	20	41788-41807	GATCGGCGACGATCCCGGCC
		533-514	GATCGGCGACGCACCCGGCC
41	20	12230-12249	GACACCGGGGACACGTCGCG
		5403-5384	GACACCACGGACACGTCGCG
42	20	18675-18694	GCGGTGCGCGGCGGACCGTC
		7200-7181	GCGGTCAGCGGCGGACCGTC

Repeat number	Size (bp)	Coordinates	Sequence alignment
43	20	27206-27225	GTGTGCGCGTGATCGAGCGG
		7317-7298	GTGTGCGCGTGCTCGATCGG
44	20	39143-39162	GGCCGCACCCTCACCGACGC
		16750-16731	GGGCGCACCCTCACCAACGC
45	20	27509-27528	GTCGTCGGGACCGATCTCGG
		18462-18443	GTCGTCGATACCGATCTCGG
46	20	34134-34153	GCAAGCTGACCGTGCTGAAG
		32072-32053	GCACGGTGACCGTGCTGAAG
47	19	16083-16101	TAGCACGGTCACCGTGCTA
		4298-4280	TAGCACGGTCACCGTGCTA
48	19	23858-23876	GCGGGTGTCGGCGCGGCGG
		6402-6384	GCGGGTGTCGGAGCGGCGG
49	19	40821-40839	CGCGAGCGCCGCCGACGCG
		24481-24463	CGCGAGCGCCGCCGTCGCG
50	18	32057-32074	AGCACGGTCACCGTGCTA
		4297-4280	AGCACGGTCACCGTGCTA
51	18	3439-3456	CTGAGCGCCGACCTCGGC

Repeat number	Size (bp)	Coordinates	Sequence alignment
		385-368	CTGGGCGCCGACCTCGGC
52	18	12654-12671	CGCGACGACCGGCGCCGC
		3119-3102	CGCGACGACCGCCGCCGC
53	18	25657-25674	GTGCGGTACGGGATCACG
		9862-9845	GTGCGGTACGGGGTCACG
54	18	32618-32635	CCGATCGGTCGACGCCGG
		16330-16313	CCGATCGGTCGATGCCGG
55	18	25924-25941	CCGGTCAGCTCGGCGGTC
		18705-18688	CCGGTCAGCTCGACGGTC
56	18	49827-49844	CCGGCCGCGATCGCGGCG
		21825-21808	CCGTCCGCGATCGCGGCG
57	18	53929-53946	ACCCGGACGCGGTGGTGG
		40314-40297	ACCGGGACGCGGTGGTGG
58	18	55574-55591	CTCGCGGACCTCCACGCG
		43623-43606	CTCGCGGACCTCGACGCG
59	17	50393-50409	CTAGCACGGTGACCGTG
		16556-16540	CTAGCACGGTGACCGTG

Repeat number	Size (bp)	Coordinates	Sequence alignment
60	17	50393-50409	CTAGCACGGTGACCGTG
		17251-17235	CTAGCACGGTGACCGTG
61	17	51016-51032	AGCACGGTGACCGTGCT
		32073-32057	AGCACGGTGACCGTGCT
62	17	53414-53430	GATCGGCGCGGTGTCGA
		38542-38526	GATCGGCGCGGTGTCGA
63	16	32057-32072	AGCACGGTCACCGTGC
		80-65	AGCACGGTCACCGTGC
64	16	17235-17250	CACGGTCACCGTGCTA
		4295-4280	CACGGTCACCGTGCTA
65	16	16540-16555	CACGGTCACCGTGCTA
		4295-4280	CACGGTCACCGTGCTA
66	16	54424-54439	GTACCGCGCGGTACGA
		10481-10466	GTACCGCGCGGTACGA
67	16	36762-36777	TGAGCGGGTCGTCTCC
		26212-26197	TGAGCGGGTCGTCTCC
68	16	50394-50409	TAGCACGGTGACCGTG

Repeat number	Size (bp)	Coordinates	Sequence alignment
		32074-32059	TAGCACGGTGACCGTG
69	16	53973-53988	GACGACGACGACGACG
		46751-46736	GACGACGACGACGACG
70	16	53973-53988	GACGACGACGACGACG
		46754-46739	GACGACGACGACGACG
71	16	53973-53988	GACGACGACGACGACG
		46757-46742	GACGACGACGACGACG
72	15	50214-50228	TCGGCCGGCATCTCG
		10073-10059	TCGGCCGGCATCTCG
73	15	51632-51646	ACCTCCCCCGCCGCG
		15846-15832	ACCTCCCCCGCCGCG
74	15	51016-51030	AGCACGGTGACCGTG
		16554-16540	AGCACGGTGACCGTG
75	15	37939-37953	CGGCCGCCGCGTCGG
		16803-16789	CGGCCGCCGCGTCGG
76	15	51016-51030	AGCACGGTGACCGTG
		17249-17235	AGCACGGTGACCGTG

Repeat number	Size (bp)	Coordinates	Sequence alignment
77	15	34584-34598	GCCGATCCGCCACCG
		18185-18171	GCCGATCCGCCACCG
78	15	47159-47173	CGTCCTCGCTGTACC
		26498-26484	CGTCCTCGCTGTACC
79	15	44088-44102	CGAGATCCGGGATCG
		37378-37364	CGAGATCCGGGATCG
80	15	51164-51178	CTGCGCGATATCGCC
		45146-45132	CTGCGCGATATCGCC

Table 8.4 Repeat sequences detected in nine *Gordonia* phage genomes

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA2-I1	29	94100-94128	CTTCACGGATGGCAGCGATGCCAGTCTTC
		26333-26305	CTTCCTGGATGTCAGAGATGCCAGTCTTC
GMA2-I2	23	65331-65353	AGCGCAAGATCGCTCTGAGAGGC
		43697-43675	AGCGCTAGATCGCTCTGAGACGC
GMA2-I3	23	61804-61826	TGAGCGCTGCAGTAGGCAGCATC
		50397-50376	TGAG-GCTACAGTAGGCAGCATC
GMA2-I4	22	66679-66700	AACCATCAGCTCAGTGACGGTG
		7193-7172	AACGATCGGCTCAGTGACGGTG
GMA2-I5	21	77503-77523	AGCGCGATCATTCGGTCTTCG
		10181-10161	AGCGCGATCATTCCGTCTTCG
GMA2-I6	20	90351-90370	GAATGTCTGCACAGCCAGCA
		84922-84903	GAATGTCTGCACAGCCATCA
GMA2-I7	19	21963-21981	CCAGGTGCTGAACTCGTCC
		1410-1392	CCAGGTGATGAACTCGTCC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA2-I8	19	70949-70967	ACATGATCGTGTCAACTAC
		4007-3989	ACATGAACGTGTCAACTAC
GMA2-I9	16	102972-102987	CGTTTGGCAGCTTCTT
		28481-28466	CGTTTGGCAGCTTCTT
GMA2-I10	16	46675-46690	ACGCGGATGCCGAAAT
		44178-44163	ACGCGGATGCCGAAAT
GMA2-D1	89	30422-30510	CTCAAGACTCCGCCAATTGACCGCAGCAAGCCACTTGAGCCTGAGGTTGTAGACCCAGACATCAAGGAGCCTAAGACTGACGATAAGAC
		29963-30051	CTCAAGACTCCGCCAATTGACCGCAGCAAGCCACTTGAGCCTGAGGTTGTAGACCCAGACATCAAGGAGCCTAAGACTGACGATAAGAC
GMA2-D2	80	37181-37260	GGAATTCAAGGGCCACCCGGTGAGCAGGGTGATGGTGGTCCGGCTGGACCACCCAACTCCCTTGACATCGGAACTGTAAC
		36629-36708	GGAATCCAGGGGCCTCCTGGAGAACAGGGGGATGAGGGTCCATCAGGTCCACCAAATAGCCTGTCTGTTGGAACTGTAAC
GMA2-D3	63	27173-27235	GTCATTGGCGACATCGTCATGTGGCTTTGGAACACTATTATCAGACCTGCATGGGACGGCATC
		27038-27100	GTCGTTGGCGATGTCGTCATGTGGCTTTGGAATTCCGTGATGAAGCCAGCTTGGGAAGGCATC
GMA2-D4	60	59494-59553	AGGATTTGAACCTCGGACCTTCGCCTTATCAGGGCGATGCTCTAACCAACTGAGCTAAAG
		58417-58474	AGGATTTGAACCTAGGACCTACGGATTAAGAGTCCGCAGCTCTA--CCGCTGAGCTATAG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA2-D5	55	13294-13348	TGTTCGCACTGCAGCTGGAGCTAAGAAGTATGGAGTGCCGATTGGATCTCCAATC
		8379-8433	TGTTCGTACTGCAGCTGGAGCTAAGAAGTATGGAGTGCCGATTGGATCTCCGATC
GMA2-D6	37	36950-36986	AAAGGCGACAAAGGCGACCAGGGCAATGTCGGACCAG
		36797-36833	AAAGGCGACAAAGGCGACCAGGGAAATGCTGGTCCAG
GMA2-D7	36	89007-89042	AGACTATAGTATCACAGTCTATCCCGAATGTAAACT
		50573-50608	AGATTACTGTATCATAAACTATCTGGAATGTAAACT
GMA2-D8	35	88988-89022	GGTGTTCGTTTCGTTGATAAGACTATAGTATCACA
		47883-47915	GGTATTCGTTCCCGT--TACGACTATAGTATCACA
GMA2-D9	31	88997-89027	TTCGTTGATAAGACTATAGTATCACAGTCTA
		79337-79367	TTCGTTGTTGTGTCTACAGTATCACAGTCTA
GMA2-D10	31	97474-97504	TTCGTTCCCGTTAGTAATACTATATCACACT
		47887-47917	TTCGTTCCCGTTACGACTATAGTATCACACT
GMA2-D11	26	58732-58757	CGTACCGCATACGGGAATCGAACCCG
		58075-58100	CGTACCGGAGACGGGAATCGAACCCG
GMA2-D12	26	81998-82023	GCGCTCACGGCTGCACCGCCTCGGCG
		81777-81802	GCGCTCATGACTGCACCGCCTCGGCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA2-D13	25	85019-85043	GTCGAAGCATCCAGAACTTCTTTGG
		4747-4771	GTCGAAGTTTCCAGAACTTCTTTGG
GMA2-D14	25	89253-89277	AACTTCTGCCATGTATGTAGGCACC
		89039-89063	AACTTCTGCCACGTATGTAGGCGCC
GMA2-D15	23	59640-59662	AACGGGATTTGAACCCGTGACCT
		59001-59023	AACGGGATTTGAACCCGTGATCT
GMA2-D16	21	56315-56335	TCCCAGCGAGAACGCCACCAA
		7553-7573	TCCCAGCGAGATCGTCACCAA
GMA2-D17	21	36953-36973	GGCGACAAAGGCGACCAGGGC
		36791-36811	GGCGACAAAGGCGACAAAGGC
GMA2-D18	21	73530-73550	GTACGTCCACTGGACTTTCTT
		65254-65274	GTCCGTACACTGGACTTTCTT
GMA2-D19	19	101384-101402	GTCTAGACCGAGCTCTTCG
		50424-50442	GTCTAGAGCGAGCTCTTCG
GMA2-D20	19	96844-96862	TCGCGAGATCAAGAGCGTC
		54617-54635	TCGCAAGATCAAGAGCGTC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA2-D21	19	60401-60419	CGGGAATCGAACCCGGGTC
		58086-58104	CGGGAATCGAACCCGCGTC
GMA2-D22	19	92459-92477	CGGGCACTTTCTCAGTCAT
		80985-81003	CGGTCACTTTCTCAGTCAT
GMA3-I1	35	77375-77409	CTATTTTCAATTTCCTCTAGCTGGCATAGCTAGGC
		77275-77241	CTATTTTCAATTTCCACTGGCTGGCACAGGTGGGC
GMA3-I2	29	56127-56155	CCGACGAAACTTTGAGCCTTGGATTCAGA
		15148-15120	CCAACGGAACGCTGAGCCTTGGATGCAGA
GMA3-I3	27	61835-61861	CGCTTGCCGCATTCCGAGCAAAACCAA
		8902-8876	CGCTTGCCGCATTCGGAGCAAGGCCAA
GMA3-I4	17	35821-35837	AACGGAATACCATCTGA
		6246-6230	AACGGAATACCATCTGA
GMA3-I5	17	72668-72684	AAACCGCAGGTCAGAGC
		66060-66044	AAACCGCAGGTCAGAGC
GMA3-I6	16	54618-54633	AGAATCGGATGCAGTT
		25123-25108	AGAATCGGATGCAGTT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA3-I7	16	37540-37555	AATTCACCAGTGATGG
		32009-31994	AATTCACCAGTGATGG
GMA3-I8	16	67051-67066	GCGCCACTCAGCGACC
		66457-66442	GCGCCACTCAGCGACC
GMA3-D1	102	66167-66266	TGCCAAATTTCAAAAG-TCTCTGACTTGAAAATAATCTGCCAAATTTCAAAAG- TCTCTGACTTGAAAATAATCTGCCAAATTCTGCCAAATTTTGACTTGA
		66061-66158	TGCCAAATTCCGAACGGTCT-TGACTTGAAA-TAATCTGCCAAATTCCGAACGATCT-TGACTTGAAA- TAATCTGCCAAATCCCGAACGATCTTGACTTGA
GMA3-D2	73	72303-72373	TCAAATATACCCACCAAAAAATCAAATCATATTGTGACTGGCATATTTGTGTAAACGTCGAG--ACACAAGAT
		72231-72303	TCAAATATACCCACCAAAAAATCAAATCATATTGTGACTGGCATATTTATGTAAACCTTGTGTAACACAATAT
GMA3-D3	39	32260-32298	GGCCCAAAGGGCGACAAGGGCGATAAGGGCGATCCTGGA
		31519-31557	GGCCCGAAGGGCGACAAAGGAGACAAGGGCGATACCGGA
GMA3-D4	36	63522-63556	CATCGATCTTGTTGACGATATCCGCCG-CCTCTTCG
		29918-29952	CATCGATCTTGGTGACGAGAT-AGCCGACTTCTTCG
GMA3-D5	35	64516-64550	ATATCCTTGGCGATTCGCATGATGCTATCGAGTAT
		2095-2128	ATATCTTTGGCGATTCGATTGATCCTA-CGAGCAT
GMA3-D6	33	32266-32298	AAGGGCGACAAGGGCGATAAGGGCGATCCTGGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		32113-32145	AAGGGGGATAAGGGTGACCAGGGCGATCCTGGA
GMA3-D7	32	16543-16574	CGCCGATGCAAAGTCGTGCAGCGCAGAATGCG
		10673-10704	CGTCGACGCTAATTCGTGCAGCGCAGGACGCG
GMA3-D8	30	52532-52561	CGGTGCCGGTGGTGCGGCGGGTGCCGCCGG
		52445-52474	CGGTGCCGCAGGAGCGGCGGGTGCTGCCGG
GMA3-D9	28	70585-70612	TTTTCTCTCCCGTGTTCGTGTGGCTGGT
		69094-69121	TTTTCTCTCCAGTGATTGTGTGGCTGGT
GMA3-D10	26	29842-29867	GATGGATCGAGACGCTTGTCGACAAT
		13265-13288	GATGG-TCGA-ACGCTTGTCGACAAT
GMA3-D11	24	32076-32099	AAAGGGCGCAAAGGGTGATCAGGG
		31704-31727	AAAGGGTGACAAGGGTGATCAGGG
GMA3-D12	23	56641-56663	TCAGTCATTTTTCTCTCCTGTGA
		56217-56239	TCAGTCATTTTTCTCTCCTGTGA
GMA3-D13	21	69644-69664	CATGTCATGTACTCCTGTGTG
		68083-68103	CATGTCATTTACTCCTGTGTG
GMA3-D14	21	34848-34868	AGCCAACGAAGTATATGGACA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		25612-25632	AGCCACCGATGTATATGGACA
GMA3-D15	19	31857-31875	CAAGGGTGATCAGGGAAAT
		31713-31731	CAAGGGTGATCAGGGAAAT
GMA3-D16	19	56640-56658	GTCAGTCATTTTTCTCTCC
		40829-40847	GTCAGTCATTTTTTTCTCC
GMA3-D17	19	65083-65101	ATTGCGGCGGCGCGGTACA
		41894-41912	ATTTCGGCGGCGCGGTACA
GMA3-D18	19	48447-48465	TAATTACGATTCTGGTTTC
		41979-41997	TAACTACGATTCTGGTTTC
GMA4-I1	38	36221-36258	GACCCGCAACACCCGCGATGCCGCCAGTGCGGCACGCC
		28933-28898	GAGCCGCATCACG-GCGA-GCGGCCATTGCGGCACGCC
GMA4-I2	32	19528-19559	CCGCCGACACCGACCCGCACACCGGGCACCGT
		10963-10933	CCGCCGACACCGACAGCTTCACC-GGCACCGT
GMA4-I3	30	37813-37840	CCGCCG--GTTCGTCCACGCCCACGGCCAC
		37338-37309	CCGCCGTCGTTCTTCCACGCCCACAGCGAC
GMA4-I4	29	15992-16019	CCGCGTCGA-CCCGCCGGTCGGTGCGTCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		6135-6107	CCGCGACGGTCCCGCCGATCGGTGCGTCG
GMA4-I5	27	29132-29157	TGACCA-GTTCCTCACCGTCGACGACG
		18417-18391	TGACCACGATCGACACCGTCGACGACG
GMA4-I6	26	31579-31604	CCGTCAGGTCGGGGCGCGGGTTCGAC
		24349-24324	CCGGCAGGTCGGGGCGGGTGCTCGAC
GMA4-I7	25	32240-32264	CGTTCCTGCTCGGAGACCTGCAGAA
		2162-2138	CGTTCCTCCTCGCGGACCTGCAGAA
GMA4-I8	25	38567-38591	CGCCTGTACCGCGACCCGTTCGAGT
		19754-19730	CGCCTGTACCGCCACGCGTTCGTGT
GMA4-I9	24	33319-33342	TCATCGACGTCGTCCTCGCCGAGG
		31851-31828	TCATAGGCGTCGTCCTCGCCGAGG
GMA4-I10	24	23437-23460	CCTCGACCGAGGTCACGATCGTGT
		4053-4030	CCTCGACAGCGGCCACGATCGTGT
GMA4-I11	24	22266-22289	GTCCGCCTGGATCGTCAGGATGCG
		12597-12574	GTTCGCCAGGATCGTCAGGGTGCG
GMA4-I12	23	23639-23661	ACGCCGTCGCGCTCGATGCGGTC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		1645-1623	ACGCCGTCAACCTCGATGCGGTC
GMA4-I13	22	25243-25264	CGAGGATCTCCGACGCGTCCCC
		7566-7545	CGAGCATCTCCGACGCGTCACC
GMA4-I14	22	29583-29604	CATCGAGGACGCCCTCGGCGAC
		26684-26663	CATCGAGGACGCGCTCGCCGAC
GMA4-I15	21	26571-26590	GACGGCGGCGGCGTG-GGCGC
		8421-8401	GACGGCGGCGGCGTGCGGCGC
GMA4-I16	21	36326-36346	ACGACGATGCCCGCCGGAGAC
		21379-21359	ACGACGATGCCCGACGGCGAC
GMA4-I17	21	43089-43109	GGGCACCGGTGCCGCCGACGC
		30302-30282	GGGCACCGGTGCCGACCACGC
GMA4-I18	20	29699-29718	CCGAACTCGTCGAAGAGTTC
		10294-10275	CCGAACTCGTCGACGAGTTC
GMA4-I19	20	22172-22191	GCTCGTCGAGACGATCGGTG
		10584-10565	GCTCGCCGAGACGATCGGTG
GMA4-I20	20	26665-26684	CGGCGAGCGCGTCCTCGATG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		17491-17472	CGGCGAGGGCGTCCTCGATG
GMA4-I21	20	43446-43465	GATCACCGTCGGCACCACCA
		23682-23663	GATCACCGTCGGCACCGCCA
GMA4-I22	20	17579-17598	CCGTCATCGTCGATGCGGTC
		1642-1623	CCGTCAACCTCGATGCGGTC
GMA4-I23	20	21822-21841	CGGCGGCGCGGGCAACGTCC
		20844-20825	CGTCGGCGCGGGCAGCGTCC
GMA4-I24	19	25157-25175	ATCCATAGTTATCCACAGG
		25104-25086	ATCCAAAGTTATCCACAGG
GMA4-I25	18	8419-8436	GTCGAGTTCGGCAACCGA
		1172-1155	GTCGAGTTCGGCGACCGA
GMA4-I26	18	16785-16802	GCCGCGACGTGCTCGACG
		7347-7330	GCCGCGACGTGCTCGTCG
GMA4-I27	18	19706-19723	CGGTCGGCACCGAGCTCG
		9172-9155	CGGTCGGCACCGAACTCG
GMA4-I28	18	37485-37502	CCGCCCTCACCACGCGCA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		18387-18370	CCGCCCACACCACGCGCA
GMA4-I29	17	27181-27197	CGGCACGTCCGACAGTC
		14769-14753	CGGCACGTCCGACAGTC
GMA4-I30	16	38239-38254	CGACGGGGACGCCGAG
		29638-29623	CGACGGGGACGCCGAG
GMA4-I31	15	26218-26232	TGAGTTCAGCGACGG
		17473-17459	TGAGTTCAGCGACGG
GMA4-D1	264	19327-19584	CGCACCGTCCAGGTCGCCGCCGGCAGTGCACTCGTCTGCGGTGTGTCCACCGTCGAAA--CGGCGGCCCGACA-- GCTGCAGCTCGCAGCGAACACCGGCAGC- CAGGTCCGCCTCGATCTGGTGGTCCTGCGACTGGTGTGGGCGGGCCTCGGTGCGTCGACGGCTGTC- CTCGACATCAAGCAGGGCACGCCGGGCGCGGTGAACCCGCCGACACCGACCCGCACACCGGGCACCGTCTACGAGGC TCCGCTCGCCGTCGTG
		18249-18500	CGTACCGTGTCGGTCGCCGCCGGTACCGCACAGGTATGCGGCGTG---ACCGT- GAAGTCCGACGCAGCGACGTCGCTGACGTTCGCCGCAAACTCGGGTGGCACA---- CGACTCGATGTCGTCGTGCTGCGCGTGGTGTGGGCGGG---CGCGTCGTCGACGG- TGTCGATCGTGGTCAAGCAGGGAACGTCCGGGTCGAGCACCGTCCCGACGCTCACCCGATCGGCAGGCGCGATGTAC GAGATGCCGCTCGCGGTCGTG
GMA4-D2	49	44526-44574	CGTCACCTCCACCGACCTCGTCACCCTCGGAGTCGGTGACGCCCGATGG
		7469-7509	CGCCACAACCACCGACCTCGTC-------GA-TCGGTGGCGCCCGATGG
GMA4-D3	35	43232-43266	CCATCGACGGTGTCGCCGTCGTCGCAGGCGACCGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		36780-36814	CCATCAACGAGGTCGCCGGCCCCACAGGCGACCGC
GMA4-D4	35	19526-19560	ACCCGCCGACACCGACCCGCACACCGGGCACCGTC
		17550-17583	ACTCGCCGACTCCGGCGT-CGCACCGGGCACCGTC
GMA4-D5	31	42196-42223	CCGCGCTCGTCGACGGACC---CCACCTCCG
		18196-18226	CCGCACTCGTCGACGGCCCGAACGACCTCCG
GMA4-D6	27	37008-37033	ACGC-CGCCACCATCAACGACCCCGAC
		35749-35775	ACGCACGCCACCATCACCGGCCCCAAC
GMA4-D7	26	30366-30391	CGGAGTTCCCGGTCGTCGTCCGCGAG
		6579-6604	CGGAGTTCCCGATCGTCGACTTCGAG
GMA4-D8	26	35889-35914	CGTCGTCCGCACCGAACTCCGCATCG
		17370-17395	CGTCGTCCGCACCGGCGTCCCCATCG
GMA4-D9	26	26794-26819	GCGCGTCGACGAGCATGTCGATCGTG
		18388-18413	GCGCGTCGTCGACGGTGTCGATCGTG
GMA4-D10	24	18611-18634	CGCCGACGGCGCAACCGGCGCCGA
		17669-17692	CGCCGACGGCGCACACGGCGACGA
GMA4-D11	24	25940-25963	CTCGCCGTCGTACGCGTCGTGCCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		19573-19596	CTCGCCGTCGTGCGCGTAGCGCCG
GMA4-D12	23	13347-13369	CCTCGGCCCGATGATGTCGATGG
		869-891	CCTCGGCCCGATGCTGTCGAAGG
GMA4-D13	23	29583-29605	CATCGAGGACGCCCTCGGCGACG
		17472-17494	CATCGAGGACGCCCTCGCCGCCG
GMA4-D14	23	34027-34049	CGGGGAGGCCGACCGTGGCTGAC
		33568-33590	CGGGGAGGCCGATCGTGACTGAC
GMA4-D15	23	41042-41064	CGACCTCGCCAAGACCGAGGACG
		11769-11791	CGAACTCGCCAAGACCAAGAACG
GMA4-D16	21	42245-42265	CGTCGACACCATCGCCGACCG
		37078-37098	CGTCGACACCATCGCCGACCG
GMA4-D17	21	33039-33059	CTGCACCGCCTCGTCCCCGAG
		32707-32727	CTGCACCGCCTTGTCCCCGAG
GMA4-D18	21	19488-19508	TGTCCTCGACATCAAGCAGGG
		12366-12386	TGTCCTCGACATGACGCAGGG
GMA4-D19	21	42553-42573	CGACAACCCACTCGGCCGCCA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		36572-36592	CGACAAAACACTCGGCCGCCA
GMA4-D20	20	38780-38799	GACGGCGCCGCCCGCGCCCT
		1101-1120	GACGGCGCCGCGCGTGCCCT
GMA4-D21	20	20912-20931	GGCGTCGAAGGCGTGGATCG
		3779-3798	GGCGTCGAAGGTGCGGATCG
GMA4-D22	20	9773-9792	GTCTGCGAGACGCCGACCGC
		5979-5998	GTCTCCGCGACGCCGACCGC
GMA4-D23	20	21491-21510	CGCGACGTCCAGCAGGACGA
		9219-9238	CGCGACGTCCAGAAGGAGGA
GMA4-D24	19	29700-29718	CGAACTCGTCGAAGAGTTC
		10274-10292	CGAACTCGTCGACGAGTTC
GMA4-D25	19	13019-13037	AGGCACTGACGCCGATCCT
		12695-12713	AGGCTCTGACGCCGATCCT
GMA4-D26	19	42153-42171	GCGACACTGCTCCTCACGT
		28283-28301	GCGACGCTGCTCCTCACGT
GMA4-D27	18	3654-3671	TCGACGCAGCTGGCAACC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		2120-2137	TCGACGCAGCTCGCAACC
GMA4-D28	18	25776-25793	GTCGAACGCGGTCGGCCG
		8080-8097	GTGGAACGCGGTCGGCCG
GMA4-D29	17	29584-29600	ATCGAGGACGCCCTCGG
		7788-7804	ATCGAGGACGCCCTCGG
GMA4-D30	16	31405-31420	ACCGTGGACCGCCGAG
		2196-2211	ACCGTGGACCGCCGAG
GMA4-D31	16	17473-17488	ATCGAGGACGCCCTCG
		7788-7803	ATCGAGGACGCCCTCG
GMA4-D32	16	36908-36923	CGCGCCGTCATCGACG
		33312-33327	CGCGCCGTCATCGACG
GMA4-D33	15	26151-26165	AGGTCACCGGACTGC
		7212-7226	AGGTCACCGGACTGC
GMA4-D34	15	29092-29106	GCGGTGTCGGCATGA
		28838-28852	GCGGTGTCGGCATGA
GMA4-D35	15	33857-33871	CATCCGCGCACTCGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		29007-29021	CATCCGCGCACTCGA
GMA4-D36	15	33076-33090	CGCTGCGTGCCGCGC
		29681-29695	CGCTGCGTGCCGCGC
GMA4-D37	15	38773-38787	GATCGTCGACGGCGC
		30139-30153	GATCGTCGACGGCGC
GMA4-D38	15	41035-41049	ACCTCATCGACCTCG
		30411-30425	ACCTCATCGACCTCG
GMA4-D39	15	33331-33345	TCCTCGCCGAGGCGG
		32858-32872	TCCTCGCCGAGGCGG
GMA4-D40	15	34021-34035	ACTCGACGGGGAGGC
		33866-33880	ACTCGACGGGGAGGC
GMA5-I1	45	5827-5871	GCAGATTCTGTGGCGTGACCAGGTCCCGGCGGGAACGCTCATCGC
		2308-2264	GCAGATGATGCCGCGTGCCCGTTTGACGGCCGGAACGCTCATCGC
GMA5-I2	41	8067-8107	CGCCGCAGCGGCCCCGGCCGCCGCCGGGGCCGCTCAGGCGG
		5092-5052	CGACGCGGCGGCCGACTCTGCGGCCGGGGCCGCTGCGGCGG
GMA5-I3	32	8068-8099	GCCGCAGCGGCCCCGGCCGCCGCCGGGGCCGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		5064-5033	GCCGCTGCGGCGGCGGGTGCCGCCGGGGCGGC
GMA5-I4	22	13867-13888	CCAGCGGGTCTACGACCTCGGC
		261-240	CCATCGGGCCGACGACCTCGGC
GMA5-I5	20	14957-14976	CTCGTCGTCCTCGCCGTCGA
		7922-7903	CTCGTCGTCCTCGTCGGCGA
GMA5-I6	19	8161-8179	CTCGGTGCCGTCCTCGCCG
		2578-2560	CTCGGTGCCGTTCTCGTCG
GMA5-I7	17	4161-4177	TCGCCTCGGCGGTCGGC
		2805-2789	TCGCCTCGGCGGGCGGC
GMA5-I8	17	4244-4260	TGCGCCGCCGTCGCCGG
		3157-3141	TGCGCCGCCGTCGTCGG
GMA5-I9	17	11509-11525	CCTCGAACTCGAACGGC
		5477-5461	CCTCGAACTCGTACGGC
GMA5-I10	17	9761-9777	GCGACGGCGGCCTCGGT
		8882-8866	GCGACGGCGGCCACGGT
GMA5-I11	16	11808-11823	CGACCTCGAACGGCAT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		6601-6586	CGACCTCGAACGGCAT
GMA5-I12	15	7345-7359	CGCACCGGCCGCCTG
		5608-5594	CGCACCGGCCGCCTG
GMA5-I13	14	17300-17313	TGCCCTCGGTCGTC
		2993-2980	TGCCCTCGGTCGTC
GMA5-D1	425	9274-9690	TGGAACGGCATCAAAGCGGCCGTGATGCTCGTCATCGACGGCATCCGCCTCTACATCGAGCTATGGGCGACGATCAT CACCGCGATCTGGAACGGCATCAAAGCCGCCGCCGTCGCTGTGTGGAACGGCATACAGATC- GCAGTCCAGGTCGTCGTGACCGTCATCCAGACCATCATCACGACACTCGGC- TCGATCATCACCGCCACATGGAACGGCGTCAAAGCGGTCGCCGAGGCTGTATGGAACGGTATCCAGTCGGTCGTCGA CACCGTCGCAGGCGTCATCCGGTCCGCCATCACGACCGCCG--TCGA---- CACCGTCATATCGATCTTCAACCGGGTCAAGGGCGTCGCCGAGACGGTATGGGGCGGCATCCAAGGATTCATCGACA ACGTCGGCAGCGCCGTCCAGTCCGTCATC
		9154-9570	TGGAACGGCATCAAAGCCGCCGCAATGTTCGTGCTCAAGCTCATCGTCGCCTACATCACCGTGTGGAAGACGATCAT CCTCGCCGTCTGGAACGCCATCAAAGCCGCCGCCGTCGCCGTGTGGAACGGCAT- CAAAGCGGCCGTGATGCTCGTCATCGACGGCATCCGCCTCTACATCGAG- CTATGGGCGACGATCATCACCGCGATCTGGAACGGCATCAAAGCCGCCGCCGTCGCTGTGTGGAACGGCATACAGAT CGCAGTCCAGGTCGTCGTGACCGTCATCCAGACCATCATCACGACACTCGGCTCGATCATCACCGCCACATGG---- -- AACGGCGTCAAAGCGGTCGCCGAGGCTGTATGGAACGGTATCCAGTCGGTCGTCGACACCGTCGCAGGCGTCATCCG GTCCGCCATC
GMA5-D2	46	8774-8819	TCGCATCGCTGGTCGCCTCGCTCGTCTCCGGCCTCGGCCCCGCCCT
		8405-8450	TCGCATCCAAGGGCGCAGAGTTCGTCACCCGCCTCGGCCCCGGCCT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA5-D3	41	9842-9882	CCGGCCTCGGCCGCCTCGGGTCCGGCCGAGGAGCGATCACC
		8273-8313	CCGGGTTCGGCCGCCTCGGCGGTGTCCTCGGCGCGATCACC
GMA5-D4	38	8228-8261	CCGCGTTCAAAGCGGCGAT----GTCGTCGAGCTTCGC
		1688-1725	CCGCGTTCACGGCCGCGTTCGACGCCGTCGAGCTTCGC
GMA5-D5	29	8059-8087	GCACTCGCCGCCGCAGCGGCCCCGGCCGC
		5045-5073	GCACCCGCCGCCGCAGCGGCCCCGGCCGC
GMA5-D6	28	7890-7917	CCGCCGTCCGCCCTCGCCGACGAGGACG
		1528-1555	CCGGCGGCCACCGTCGCCGACGAGGCCG
GMA5-D7	27	9357-9383	GATCTGGAACGGCATCAAAGCCGCCGC
		9150-9176	GATCTGGAACGGCATCAAAGCCGCCGC
GMA5-D8	27	10737-10763	TTCGGCGACGACGCCGCGACGTACGTC
		6296-6322	TTCGGCGACCTCGGCGCGACGTACGTC
GMA5-D9	27	11818-11844	CGGCATCATCGGCGACGCGATCCCCGC
		8604-8630	CGGCATCATCGGCGCAGCGATCACCGC
GMA5-D10	25	9239-9263	TCTGGAACGCCATCAAAGCCGCCGC
		9152-9176	TCTGGAACGGCATCAAAGCCGCCGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA5-D11	24	12128-12151	TCGGCCCCGGCCACTACCCGAACA
		8438-8461	TCGGCCCCGGCCTCGACACGATCA
GMA5-D12	24	16533-16556	TCGAAGTCGCCGCCGTCGAAGTGT
		9251-9274	TCAAAGCCGCCGCCGTCGCCGTGT
GMA5-D13	24	16533-16556	TCGAAGTCGCCGCCGTCGAAGTGT
		9371-9394	TCAAAGCCGCCGCCGTCGCTGTGT
GMA5-D14	23	9995-10017	CCCGATGACCGTCAACCCCGCCC
		6663-6684	CCCGATGACCGTC-ACGCCGCCC
GMA5-D15	22	17513-17534	GTCACGTTGTGCGTGTGCTCGC
		2455-2476	GTCACTGTGGGCGTGTGCTCGC
GMA5-D16	21	8663-8683	CCCTCGGTCCCGTCCTCGCCG
		8159-8179	CCCTCGGTGCCGTCCTCGCCG
GMA5-D17	20	4522-4541	CCGCCCCGGCGGCACCCGCC
		5034-5053	CCGCCCCGGCGGCACCCGCC
GMA5-D18	20	12837-12856	CACCGACTACGGCAACACGT
		3207-3226	CACCGTCTACGGCAACACGT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA5-D19	20	11942-11961	CTCAACGGCGGCAAGGGCGC
		3601-3620	CTCAACGGCGGCACGGCCGC
GMA5-D20	17	7662-7678	CGATGTGGGCCGCCGCC
		2778-2794	CGAAGTGGGCCGCCGCC
GMA5-D21	17	9572-9588	CGACCGCCGTCGACACC
		4341-4357	CGACGGCCGTCGACACC
GMA5-D22	15	13708-13722	TCGCCAACAGCGCCG
		5739-5753	TCGCCAACAGCGCCG
GMA5-D23	15	7744-7758	CACCGCACCGGCAGG
		6025-6039	CACCGCACCGGCAGG
GMA5-D24	14	16722-16735	ACGACCTCGTCGAC
		1067-1080	ACGACCTCGTCGAC
GMA5-D25	14	9258-9271	CGCCGCCGTCGCCG
		4246-4259	CGCCGCCGTCGCCG
GMA5-D26	14	8780-8793	CGCTGGTCGCCTCG
		5400-5413	CGCTGGTCGCCTCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA5-D27	14	13545-13558	TCCTCGACGTGCTC
		7586-7599	TCCTCGACGTGCTC
GMA5-D28	14	17048-17061	TACGGCACCGTCCA
		16577-16590	TACGGCACCGTCCA
GMA6-I1	23	21011-21033	CCGCCGGACATGAACCTGGCAGA
		14457-14437	CCGCCGGACATGAAC--GGCAGA
GMA6-I2	22	48594-48615	AAATGCGTAAGTTCATCGGAGC
		1859-1838	AAGTGCGTAAGTTCATCGAAGC
GMA6-I3	16	6792-6807	AGACCGTTCAGAAAAC
		5349-5334	AGACCGTTCAGAAAAC
GMA6-D1	85	31393-31477	GCGGGTGAAGAGGGTTATCTGCTTTACGAATCCCTGAACACGTTGTACTTCGCCACCCCGCAGTGGCTCTTCGACAA GCAGCCGA
		30019-30103	GCCGGGGAAGAAGGCTATATCGCGTACGAGTGCCTGAACACCCTGTACTTCGCGTCCCCTAAGTGGCTCTTCGAGAA CCGCCCGA
GMA6-D2	59	27470-27528	TGGAACATCATCAAGATGGTTGCCTCGGTAGTCTTCAACGCCATCGCCGCAGTCATCCG
		27350-27408	TGGAACACCATCAAGACTGTCTTTATGGCGGTGTGGAACGCGATCCTCGCAGTCATCCG
GMA6-D3	57	10091-10147	GTTCGGGACGCTGACTACTGGGGTGCGCCCGTGGGCACTCCGATCGTGGCTGGGATG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		8952-9008	GTTCGGGACGCTGAGTATTGGGGTATGCCGGTCGGTACTCCGATCACGCCTGGCATG
GMA6-D4	54	50274-50322	TCGGCATCG-AGCC--TTCGCA-GTTGTTCGGCAAGTTCGTGCCCATGC-CGAA
		24473-24526	TCGGCATCGCAGCTCGTCCGCAAGTTGTTCGGCAACGACGGGCACATGCTCGAA
GMA6-D5	41	27470-27510	TGGAACATCATCAAGATGGTTGCCTCGGTAGTCTTCAACGC
		26924-26964	TGGGACACCATCAAGGCGATCGCGATGGGAGTCTTCAACGC
GMA6-D6	35	51064-51098	GGAATGGACCGACGAGGACGACGACTCCGACGACG
		21945-21978	GGAATCGACCGACGAGGACGAC-ACCCCGTCGCCG
GMA6-D7	31	53930-53960	ATCGGCGCACAGATCAAAGAGCGCCAGGACG
		18388-18418	ATCGGCGCACAGCTCTACGATCGTCAGGTCG
GMA6-D8	27	67062-67088	AGAAGCCCCCAAGAAGAAGGTAGGCGC
		54076-54102	AGAAGCGCTGAAGAAGAAGGTCGGCGC
GMA6-D9	26	61643-61668	CTCGGCAAGGAGTACGGGGATGACCT
		6749-6774	CTCAGCAAGTCGTTCGGGGATGACCT
GMA6-D10	24	51136-51159	CGATGACGACGACGAAGACGACGA
		57590-57613	CGATGACGACGAGGACGAGGACGA
GMA6-D11	24	51137-51160	GATGACGACGACGAAGACGACGAG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		58116-58139	GACGAAGACGCCGAAGACGACGAG
GMA6-D12	22	55185-55206	GAGCACTACCGCGACGACCTCC
		49811-49832	GAGCACACCCGCGACGACCTCC
GMA6-D13	21	24267-24287	GCTCGATCATCGGCAAGGCAC
		11469-11489	GCTCGATCATCGGCAGCGCAC
GMA6-D14	21	34531-34551	GGACGATCAACGTATCGATCG
		28497-28516	GGAC-ATCAACGTATCGATCG
GMA6-D15	19	60365-60383	GGTTGGGACATCGCGGAAG
		1201-1219	GGTTGGTACATCGCGGAAG
GMA6-D16	19	41140-41158	GCGTCGACATCGTGGTCTC
		20527-20545	GCGTCGACATCGTTGTCTC
GMA6-D17	19	63891-63909	CCGAAGCGCCGTTCACGAG
		46714-46732	CCGAAGCGCCGGTCACGAG
GMA6-D18	18	69839-69856	CAGAGCGCAATGGGCGAG
		17326-17343	CAGAGCGCAATGGGCGAG
GMA6-D19	16	54208-54223	GACTACGCCGCGTTCA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		50759-50774	GACTACGCCGCGTTCA
GMA6-D20	16	61821-61836	AGCCGGAAGACGAAGA
		58108-58123	AGCCGGAAGACGAAGA
GMA7-I1	21	71407-71427	CGACGTCCTCCTCGGCCTGCT
		12052-12032	CGCCGTCCTTCTCGGCCTGCT
GMA7-I2	21	53040-53060	CTGCCCCTGAAGTTTGAAGTA
		17729-17709	CTGCCCCTGAAGCTTGGAGTA
GMA7-I3	16	68898-68913	CGTTGAGCTTCTCGTA
		7102-7087	CGTTGAGCTTCTCGTA
GMA7-I4	16	58715-58730	CCGAGAAGATCTCGCG
		7270-7255	CCGAGAAGATCTCGCG
GMA7-I5	16	22596-22611	GACAAGCAGGCCGACG
		16877-16862	GACAAGCAGGCCGACG
GMA7-D1	242	19701-19940	GGAAACATTCTCGGATGGCTCGGCAATCTCGGAGGCAAGCTGCTTGAGTGGATGGGGGCTGCGTGGCAATGGCTCGT AGACAATGGCCCAACCATGCTTGCGAAGCTTATTGTGTGGCT-- TGCGTCTCTGCCCGCCAAGTTTATCGGCTGGCTCGGCGATATTGGCGGCAAGCTTCTCGAATGGCTCAGGGCGGGTT GGGATTACCTCAAGGACAACTGGCCTATCATTCTCGCCAAGTTT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		19569-19808	GGAAAGATTCTCGGATGGCTCGGCGACCTCGGCGGCAAGCTGCTCGAGTGGATGGGCGCTGCGTGGAACTGGCTTGT CGAGAATGGGCCGACGATGCTCCTGAATCTCATGACATGGCTCATGGGTATC-- GCTGGAAACATTCTCGGATGGCTCGGCAATCTCGGAGGCAAGCTGCTTGAGTGGATGGGGGCTGCGTGGCAATGGCT CGTAGACAATGGCCCAACCATGCTTGCGAAGCTT
GMA7-D2	50	63890-63936	ACGGTGCCGGTGCTGGGGTCCGAGGCCTCC---ACGGCCCCGTCCACGAC
		47815-47863	ACGGTGCCGGTGCCGGGG-CGGATGCCCACGGGTCGGAACCGTCCACGAC
GMA7-D3	48	19836-19883	AAGTTTATCGGCTGGCTCGGCGATATTGGCGGCAAGCTTCTCGAATGG
		19572-19619	AAGATTCTCGGATGGCTCGGCGACCTCGGCGGCAAGCTGCTCGAGTGG
GMA7-D4	30	66361-66390	CGTGCCCTACGTAGGACTCGAACCTACGCC
		66016-66045	CGTGCCCTACGTAGGACTCGAACCTACGCC
GMA7-D5	28	34428-34455	CCGCCCAGACTGTAGTTGAGGCTTTCGC
		16009-16034	CCGCCCAGACTGTAGT--CGGCTCTCGC
GMA7-D6	24	67186-67209	CGTTGTAGTACTAGTCTAACACGT
		66941-66964	CGTTCTAGTACTAGTCTATCACGT
GMA7-D7	22	51618-51639	CCAGCTTTCACTGTCCCACGAG
		35317-35338	CCAGCTTTCACTTTCCCTCGAG
GMA7-D8	21	29185-29205	CGACGCGATCATCAAGGGTAT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		22607-22627	CGACGCGATCATCAAGGGCAT
GMA7-D9	21	72861-72881	CCGGCTCGGGAAAGAATGCGT
		37682-37702	CCGGCCCTGGAAAGAATGCGT
GMA7-D10	21	62876-62896	TGATTCTTCCCCTTTCGGTGG
		62356-62376	TGATTGATCCCCTTTCGGTGG
GMA7-D11	21	67189-67209	TGTAGTACTAGTCTAACACGT
		67005-67025	TGTAGTACTAGTCTAGCATGT
GMA7-D12	19	64731-64749	GTCCCGTCGTTGTCGATCT
		44094-44112	GTCTCGTCGTTGTCGATCT
GMA7-D13	17	13690-13706	CGAGGGCGGCAAGGCTG
		6352-6368	CGAGGGCGGCAAGGCTG
GMA7-D14	16	26359-26374	ACCTGGACGGCGACCT
		5573-5588	ACCTGGACGGCGACCT
GRU3-I1	33	14343-14373	GCCTCGGTCTCGTAGTTGA--AGATCGTTCGTC
		2805-2773	GCCTCGGTCTCGTCGATGACGAGGTCGTTGGTC
GRU3-I2	30	8990-9019	TCGTCCTCGCCGTGATCGCCGTCGTCGCCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		1186-1157	TCATCGTCGCCGGGGTCGACGTCGTCGCCG
GRU3-I3	29	7668-7696	CGACGGCACCGAACACAACGACGTCGTCG
		1446-1418	CGACGGCGGCGACGATCACGACGTCGTCG
GRU3-I4	27	15140-15164	GGCGA--AGGTTGACCGCGTTGCCGAC
		2722-2696	GGCGAGCAGGTTGACCGCGTGGCGGAC
GRU3-I5	20	12572-12591	CGGCGTACCCGACGCGCTCG
		3367-3348	CGGCGTACCCGCCGCGCACG
GRU3-I6	19	9721-9739	GCCGGTCTCGGTGTCGTGG
		3160-3142	GCCGGTCTCGGCGTCGAGG
GRU3-I7	19	11767-11785	CGGCACCGGCCGCCGCAAC
		8203-8185	CGACACCGGCCGCCGCTAC
GRU3-I8	19	17410-17428	GGGCGGTCGTCCGGCGGCG
		9795-9777	GGACGGGCGTCCGGCGGCG
GRU3-I9	18	15348-15365	CATGTCGGGTTCCTCTCG
		97-80	CATGTCGGGTTCCTTTCG
GRU3-I10	18	14787-14804	GATGACGACCGCCGAGGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		4228-4211	GATGCCGACCGCCGAGGC
GRU3-I11	17	17410-17426	GGGCGGTCGTCCGGCGG
		1317-1301	GGGCGGTCGTCGGGCGG
GRU3-I12	16	7134-7149	TCGAGTTCGGCGGCGA
		1597-1582	TCGAGTTCGGCGGCGA
GRU3-I13	14	8799-8812	CGACGTCGTCGCCG
		1170-1157	CGACGTCGTCGCCG
GRU3-I14	14	6104-6117	GCGATGACCGCGCC
		2541-2528	GCGATGACCGCGCC
GRU3-I15	14	6336-6349	CGGGCGGCGTCGAG
		3218-3205	CGGGCGGCGTCGAG
GRU3-I16	14	12379-12392	ACAGCCAGGGCGGC
		9227-9214	ACAGCCAGGGCGGC
GRU3-D1	158	9223-9380	GCTGTCTGGAATGCGATAAAGGCCGTCGTTCAGTTCGTGATCGATGCGCTGCTCGCCTACATCCAAGTGTGGTCGAT GACCATTACCGCGATCTGGAACGCGATCAAGTTCGTAGCGCTCGCGGTCTGGACCGGAATTCAGATCGCCGTGCAGG TCGT
		9103-9260	GCGGTGTGGAACGCGATCAAGACCGTCGCCGCCGTCGTGATCGCCGCCGTCACCGCCTACGTCAAGGCATGGCAGGC GGTCATTGTCGCCATCTGGAACGCGATAAAGACTGCCGCCCTGGCTGTCTGGAATGCGATAAAGGCCGTCGTTCAGT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
			TCGT
GRU3-D2	78	14151-14228	CGTCGCGCCCTTCACGCCGCTGTGACCAGCGAAAGCAACTCATCAGCGGGTTCGGGGTTCGAGTCCCTGATGGCGCAC
		14074-14151	CGTCGCGCCCTTCACGCCGCTGTGACCAGCGAAAGCAACTCATCAGCGGGTTCGGGGTTCGAGTCCCTGATGGCGCAC
GRU3-D3	45	5011-5055	TGTCGACGCGGTCACCGCCGACGACCTGTCGACCGCCCCGGTAGA
		1622-1666	TGTCGACGCGGTCGGCGACGCTGACTTGTCGGCGGCCACGGTCGA
GRU3-D4	30	1412-1441	CGACCGCGACGACGTCGTGATCGTCGCCGC
		412-441	CGCCCGCGACGACGCCGCGAGCGTCGACGC
GRU3-D5	30	9309-9338	CGCGATCTGGAACGCGATCAAGTTCGTAGC
		9102-9131	CGCGGTGTGGAACGCGATCAAGACCGTCGC
GRU3-D6	29	9123-9151	GACCGTCGCCGCCGTCGTGATCGCCGCCG
		1413-1440	GACCG-CGACGACGTCGTGATCGTCGCCG
GRU3-D7	29	9125-9153	CCGTCGCCGCCGTCGTGATCGCCGCCGTC
		8831-8859	CCGTCATCGCCGACCTCGTCGCCGCCGTC
GRU3-D8	28	9123-9150	GACCGTCGCCGCCGTCGTGATCGCCGCC
		1430-1457	GATCGTCGCCGCCGTCGAGAACGCGGCC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GRU3-D9	27	9377-9403	TCGTCGTCAACATCATCACCGCGATCA
		8672-8698	TCCTCGGCCAGATCATCACCGCGATCA
GRU3-D10	26	9012-9036	CGTCGCCGCGATC-GTGCTGCTGGCG
		3969-3994	CGTCGCCGCGATCGGTGCCGCTGTCG
GRU3-D11	24	9135-9158	CGTCGTGATCGCCGCCGTCACCGC
		8997-9020	CGCCGTGATCGCCGTCGTCGCCGC
GRU3-D12	24	9196-9219	TGGAACGCGATAAAGACTGCCGCC
		9109-9132	TGGAACGCGATCAAGACCGTCGCC
GRU3-D13	23	11758-11780	GGCGGCCAACGGCACCGGCCGCC
		7392-7414	GGCCGCCAAGCGCACCGGCCGCC
GRU3-D14	22	13567-13588	ATCGCCGCGCACCCTGGCTATC
		1194-1214	ATCGCCGCGCACCC-GGCCATC
GRU3-D15	22	8381-8400	AGATG--CAAGGCGTGGGCCGC
		2811-2832	AGATGGTCAAGGCGTGGGCCGC
GRU3-D16	22	8930-8951	TCGCCGCCGCCGCAACCGCGCT
		8126-8147	TCGCCGCCGCCGCGATCCCGCT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GRU3-D17	22	12868-12889	GGGTGTTCAACGGCAAGACCGC
		8411-8432	GGGTGTTCAACGGCCTGGCCGC
GRU3-D18	21	8846-8866	TCGTCGCCGCCGTCGTCAACG
		1432-1452	TCGTCGCCGCCGTCGAGAACG
GRU3-D19	20	9006-9025	CGCCGTCGTCGCCGCGATCG
		8799-8818	CGACGTCGTCGCCGCGATCG
GRU3-D20	20	8799-8818	CGACGTCGTCGCCGCGATCG
		3963-3982	CGACACCGTCGCCGCGATCG
GRU3-D21	19	11349-11367	GCCGACCTCGTCACGATCG
		951-969	GCCGACCTCGTCGTGATCG
GRU3-D22	19	10778-10796	CGCCGTCACGACCGCCGAC
		4384-4402	CGCCGCGACGACCGCCGAC
GRU3-D23	19	13555-13573	GTCTTCGACGTGATCGCCG
		8992-9010	GTCCTCGCCGTGATCGCCG
GRU3-D24	19	15716-15734	TCGCCCGCGCCGCTGAACA
		11365-11383	TCGCCCGCGCCGCCGAGCA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GRU3-D25	18	8796-8813	CGCCGACGTCGTCGCCGC
		517-534	CGCCGAGGTCGTCGCCGC
GRU3-D26	18	8932-8949	GCCGCCGCCGCAACCGCG
		3704-3721	GCCGCCGCCGCATCCGCG
GRU3-D27	18	9836-9853	TCACCGTCGACAACCGCG
		6036-6053	TCACCGTCGACAACCCCG
GRU3-D28	17	15560-15576	GCCGGACGCCACCGCCG
		7210-7226	GCCGGACGCCACCGCCG
GRU3-D29	17	5761-5777	GTCCTCGGGCACGCTCC
		2683-2699	GTCCTCGGGCACGGTCC
GRU3-D30	17	12241-12257	TCGGCCCCATCGCCAAC
		8603-8619	TCGGCCCCATCGTCAAC
GRU3-D31	15	8120-8134	CGTCGATCGCCGCCG
		1576-1590	CGTCGATCGCCGCCG
GRU3-D32	15	11140-11154	ACGCCGCCGCCCTCG
		2934-2948	ACGCCGCCGCCCTCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GRU3-D33	15	6151-6165	CGCCGGGTGGCTCGG
		5332-5346	CGCCGGGTGGCTCGG
GRU3-D34	15	9126-9140	CGTCGCCGCCGTCGT
		8847-8861	CGTCGCCGCCGTCGT
GRU3-D35	14	15564-15577	GACGCCACCGCCGC
		209-222	GACGCCACCGCCGC
GRU3-D36	14	11990-12003	GGCTCACTCGCACT
		1128-1141	GGCTCACTCGCACT
GRU3-D37	14	14907-14920	CTTCGACCGCGACG
		1409-1422	CTTCGACCGCGACG
GRU3-D38	14	6814-6827	CCCACCGGTCGCCC
		1880-1893	CCCACCGGTCGCCC
GRU3-D39	14	17356-17369	AGACCGGCGGCCGA
		3941-3954	AGACCGGCGGCCGA
GRU3-D40	14	10046-10059	CGACGACCTGTCGA
		5029-5042	CGACGACCTGTCGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GRU3-D41	14	6519-6532	CCAACCCGGCCGAC
		5793-5806	CCAACCCGGCCGAC
GRU3-D42	14	9140-9153	TGATCGCCGCCGTC
		7921-7934	TGATCGCCGCCGTC
GTE6-I1	56	39696-39751	ACCGCCGGACGGCGAGCAAGTCGAGCAGCTGATCGCGCTCGCGCAGGATCTCGTCG
		7842-7795	ACCGC-GGTCGCCG-GCAGGTCGA-CACC-GATCGCGC----GCAGGATCTCGACG
GTE6-I2	55	10343-10394	GCTCGCCGAACTGCAGGGCCTCGACCTCGAAC-AGCTGCGCCG--GCGCGAGTTC
		9690-9639	GCTCGCCGAACTGCA---CCTCGATCTCATCCGAGTCACGCCGGCGCGCGAGTTC
GTE6-I3	51	48160-48208	CGCGGCCGCGATCGCCGAGCACG-CAGCC-GGCCGCGACCAGCGCAGCAAC
		37923-37875	CGCGGCCGCGATCG--TAGGACGCCTGCCGGGCGGCGATCAGCTTTGCAAC
GTE6-I4	42	53601-53640	CCGGCCGG--CGTCGCCGCGGCCGGCAGCATTGCCTACACCG
		9099-9058	CCTGCCGGATCTTCGCCGCGGCCAGCTGCCCGGCCGACACCG
GTE6-I5	38	38196-38233	CGCCGAGGCTCGCGCCGCCGGCGAACTGCTGTACCTGC
		23824-23790	CGCCGAGGATCGGGCCG---GCGAACTGCACCACCTGC
GTE6-I6	37	54932-54968	CGGCGGTGGAAACACGAGCAGGCGCGCAGCGGTCCCG
		31987-31951	CGGCGGCGTGATCATGAGGATGCGCGCAGCGGTGCCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GTE6-I7	35	52080-52114	CGCCCGGATCTGAACGACGACGACGAGACGGGCAA
		43835-43804	CGCCCGGATCT---CGTCGACGACGCGGCGCGCAA
GTE6-I8	35	50812-50846	GCGCGCGCTGATCGCCGAGATCGAGGTCGACGACG
		33984-33950	GCGCAGGTTGATCGCCGAGATCCGGCCCGACGCCG
GTE6-I9	33	54279-54311	TCGACCCGCTCGTCGTCGACAAGCACCTCGACC
		12264-12232	TCGACCGGCTCGTCGTCGAACTGCTCGTCGACC
GTE6-I10	31	42277-42307	CTCGCCGGGCACCTCGAAAACGGCGGCCGGC
		35880-35850	CTCGCCGTGCAGCCAGAACCCGGCGGCCGGC
GTE6-I11	28	3940-3967	TCGACCCGGCCGAAGAGAGCGAGTAGCC
		1331-1304	TCGACCCGGCCGTCGCGCGCGAGTAGCC
GTE6-I12	28	36470-36497	CCGGCATCGCCGGCACGCTCGCGCTCGT
		20125-20098	CCGTCATCGCCGGCAGGCTCGCTATCGT
GTE6-I13	28	33924-33951	CGCCACGATCGCGGCCACCGGCTCGCCG
		8453-8428	CGCCCCGATCGCGGCCA--GGGTCGCCG
GTE6-I14	28	50817-50844	CGCTGATCGCCGAGATCGAGGTCGACGA
		30635-30609	CGGTGATCGCCGAGATC-ACGTCGCCGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GTE6-I15	28	52742-52769	TCGTCGGCGCGATCGACGACAACCCGCC
		9406-9379	TCGTCGGCGCGGTCGGCGTCATCCGGCC
GTE6-I16	28	37149-37176	GCTGCCCGGCGTGTTCGGAGCGGCTGGC
		26700-26673	GCTGCCCGGCGTGGTCGTGGTGGCCGGC
GTE6-I17	27	23533-23559	TTCTTCCAGTCGACGGCGGCCGCGGTC
		11688-11662	TTCTGCGCGTCGTCGGCGGCCGCGGTC
GTE6-I18	26	22386-22411	TTCGGCGGTGTCGGTGACGATGACCG
		2363-2338	TTCGGCGACGTTGGTGACGATGACCG
GTE6-I19	26	49588-49613	GCCTGAACGGCGAGTTCGCCGCGCTC
		288-263	GCCTGGTCGGCGAGTTCGGCGCGTTC
GTE6-I20	26	21895-21920	GCTGATCGGCCGGCAACTCGTCGGCG
		9870-9845	GCGGATCGGCCGGCACCACGTCGCCG
GTE6-I21	25	4751-4772	CGGCACGCCGGCCG---AGGCGATC
		3159-3135	CGGCACGCCGGCCGGATAGGCGATC
GTE6-I22	25	50866-50890	CACGCTCGGCGCGACCGGCCGCCGG
		19591-19567	CACGCTCGGCGCATCCTGCGGCCGG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GTE6-I23	24	36165-36188	CGTCGCCGACCTCGCCGAGCGCGC
		10724-10702	CGTCGCCCAC-TCGCCGAGCGCGC
GTE6-I24	24	55094-55117	ACCACGACGACGAGCAGATCGTCG
		9856-9833	ACCACGTCGCCGGGCAGATCGTCG
GTE6-I25	24	40693-40716	CGAGGCACCCGGCGTCGACGACGA
		26226-26203	CGAGGTACCCGGCGTCGTCGTCGA
GTE6-I26	24	30865-30885	CGGTGGT---GCGATCGCCGACGT
		3813-3790	CGGTGGTCACGCGATCGCCGACGT
GTE6-I27	24	23604-23627	CGGTACCGTCGCGCCGGCGATCGC
		22696-22675	CGGTAGCG--GCGCCGGCGATCGC
GTE6-I28	23	34643-34665	TGTTTCAGCAGCAGGTTCGGCAG
		683-661	TGTTTCAGCAGCAGTTTCTGCAG
GTE6-I29	23	22748-22770	GGTCGACGCTGGTCGCCGGCGCG
		16689-16667	GGTCGAGGCTGGTCGCCGTCGCG
GTE6-I30	23	35104-35126	GAGGTCGACGACGAGCACACGGT
		4085-4063	GAGGTCGACGACGCGCGCAGGGT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GTE6-I31	23	13910-13932	GCAGTCGGCGAGATCCGGCCCGA
		9398-9376	GCGGTCGGCGTCATCCGGCCCGA
GTE6-I32	23	52447-52469	CTCGTCGACGACGCCGGCCGCAA
		43826-43804	CTCGTCGACGACGCGGCGCGCAA
GTE6-I33	22	31267-31288	CACCGGCTCGACGGGCTGGTCG
		12271-12250	CACCGGCTCGACCGGCTCGTCG
GTE6-I34	22	49598-49619	CGAGTTCGCCGCGCTCGTCGCC
		13793-13772	CGAGTTCGCCGCCCTCGGCGCC
GTE6-I35	22	54005-54026	CCGACGCGCGGCGTCGAGGTGC
		43980-43960	CCGA-GCGCGGCGTCGTGGTGC
GTE6-I36	21	33953-33973	CGTCGGGCCGGATCTCGGCGA
		13934-13914	CGTCGGGCCGGATCTCGCCGA
GTE6-I37	21	52209-52229	GGCCGCCCGAGCGCATCGACC
		1105-1085	GGCCACGCGAGCGCATCGACC
GTE6-I38	21	9892-9912	GCTCGCAGGATCTCGACGCCG
		7812-7792	GCGCGCAGGATCTCGACGGCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GTE6-I39	21	37751-37771	CGCACTCGGCGACCTGCTCGA
		9432-9412	CGCGCTCGGCGACCTGCACGA
GTE6-I40	21	43241-43261	CTGCCGAGCACCCGTCGGCCG
		10896-10876	CTACCGAGCACCCGCCGGCCG
GTE6-I41	21	53608-53628	GCGTCGCCGCGGCCGGCAGCA
		12314-12294	GCGGCGCCTCGGCCGGCAGCA
GTE6-I42	21	31980-32000	CGCCGCCGGCGACCGAGCCGG
		18696-18676	CGCCGCCGGCGAGCGTGCCGG
GTE6-I43	21	39008-39028	CGCATCGACCGCGACGAGGAT
		24138-24118	CGCACCGACGGCGACGAGGAT
GTE6-I44	21	48482-48502	CGTACTCGTCGCCGTCGACGA
		33457-33437	CGGACACGTCGCCGTCGACGA
GTE6-I45	20	48187-48206	CGGCCGCGACCAGCGCAGCA
		3085-3066	CGGCCGCGGCCAGCGCAGCA
GTE6-I46	20	16879-16898	AGGCGCCGGCGGCCGGGCAG
		4588-4569	AGACGCCGGCGGCCGGGCAG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GTE6-I47	20	20958-20977	ACCCGCAGCAGCGCACGAAC
		7056-7037	ACCGGCAGCAGCGCACGAAC
GTE6-I48	20	26610-26629	GACCTCGCAGGGCACGGTGC
		241-223	GACCT-GCAGGGCACGGTGC
GTE6-I49	20	50827-50846	CGAGATCGAGGTCGACGACG
		500-481	CGCGATCGAGGTCGTCGACG
GTE6-I50	20	45151-45170	TCGCCGCCCGACTCGTCGAC
		1661-1642	TCGCCGCCCGACGGGTCGAC
GTE6-I51	20	44774-44793	CGGCGGCCGCGTCCTGCGCA
		3088-3069	CGGCGGCCGCGGCCAGCGCA
GTE6-I52	20	30112-30131	GCGGCGAGGCTGTGCGCCGG
		4852-4833	GCGGCGAGGCTGCGGGCCGG
GTE6-I53	20	28355-28374	TCGACGAGTGCAACGACGAC
		6899-6880	TCGACGAGTGCACCGACCAC
GTE6-I54	20	31465-31484	CGCGCAGCTGCAGAACGCCG
		21573-21554	CGCGCAGCTGCAGTTCGCCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GTE6-I55	20	47451-47470	GCTCGACCCGAACGTGCCGC
		27209-27190	GCTCGACCCGAGCGTGACGC
GTE6-I56	19	26481-26499	GATCGCGGCGAAGATCTTC
		24471-24453	GATCGCGGCGAAGATCTTC
GTE6-I57	18	38745-38762	TCGGCGTCGATCACTACG
		812-795	TCGGCGTCGATCACTTCG
GTE6-I58	18	39562-39579	ATTCCAGTCGCGCAGCGT
		1530-1513	ATTCCAGTCGCGCAGGGT
GTE6-I59	18	33455-33472	CCGGGTCGGCCGGCATCG
		9501-9484	CCGGCTCGGCCGGCATCG
GTE6-I60	18	43612-43629	CCGACGTGCAGATCGCCG
		11286-11269	CCGACGGGCAGATCGCCG
GTE6-I61	18	54446-54463	CTCGCCGAGGTCGACGAC
		13227-13210	CTCGCCGAGGTGGACGAC
GTE6-I62	18	51371-51388	AGTGCGACCAGTGCGGCG
		17068-17051	AGTGCGACCAGCGCGGCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GTE6-I63	18	32562-32579	CGCCGACACGGTGGCCGC
		19498-19481	CGCCGCCACGGTGGCCGC
GTE6-I64	18	39735-39752	CGCGCAGGATCTCGTCGA
		30842-30825	CGCGCAGGATCGCGTCGA
GTE6-I65	18	55101-55118	CGACGAGCAGATCGTCGA
		50739-50722	CGCCGAGCAGATCGTCGA
GTE6-I66	18	55089-55106	CGATGACCACGACGACGA
		52579-52562	CGGTGACCACGACGACGA
GTE6-I67	17	14475-14491	TCGAACTGCTCGTCGAC
		12249-12233	TCGAACTGCTCGTCGAC
GTE6-I68	17	53593-53609	CCGGCGCACCGGCCGGC
		25837-25821	CCGGCGCACCGGCCGGC
GTE6-I69	16	22111-22126	TTCGACCGCATGGCCG
		8087-8072	TTCGACCGCATGGCCG
GTE6-I70	16	34966-34981	GGCAGCATCGGCACCG
		19815-19800	GGCAGCATCGGCACCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GTE6-I71	16	34383-34398	ACCCACCGATCGAGGG
		21676-21661	ACCCACCGATCGAGGG
GTE6-I72	15	55366-55380	GACGATCGCCTCGCG
		1269-1255	GACGATCGCCTCGCG
GTE6-I73	15	49958-49972	ATCGGGCACAGCGAG
		1922-1908	ATCGGGCACAGCGAG
GTE6-I74	15	4571-4585	GCCCGGCCGCCGGCG
		2955-2941	GCCCGGCCGCCGGCG
GTE6-I75	15	54195-54209	TGAACGCGCCGTTCG
		8712-8698	TGAACGCGCCGTTCG
GTE6-I76	15	33741-33755	CTTCAACCGGTCCGA
		8756-8742	CTTCAACCGGTCCGA
GTE6-I77	15	23549-23563	CGGCCGCGGTCGGCA
		10814-10800	CGGCCGCGGTCGGCA
GTE6-I78	15	15215-15229	CTCGATCTGCCCGAG
		12751-12737	CTCGATCTGCCCGAG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GTE6-I79	15	33386-33400	CGGATCTCGCCGACT
		13926-13912	CGGATCTCGCCGACT
GTE6-I80	15	32443-32457	GCCGCCGGCGCCGCA
		18091-18077	GCCGCCGGCGCCGCA
GTE6-I81	15	24337-24351	GCCACGTTCACGGCG
		23369-23355	GCCACGTTCACGGCG
GTE6-I82	15	41200-41214	CGGCGCCGACGGCAC
		24685-24671	CGGCGCCGACGGCAC
GTE6-I83	15	44737-44751	CGATCGCCTGCCCGA
		29185-29171	CGATCGCCTGCCCGA
GTE6-I84	15	52549-52563	GCCGCCGGGCACGTC
		29846-29832	GCCGCCGGGCACGTC
GTE6-I85	15	43357-43371	GCGAACAGATCGCCG
		30786-30772	GCGAACAGATCGCCG
GTE6-I86	15	56753-56767	CCGACGATCGAGTCG
		39192-39178	CCGACGATCGAGTCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GTE6-I87	15	54669-54683	CGTCGACCTCGGCGA
		54461-54447	CGTCGACCTCGGCGA
GTE6-D1	110	19997-20093	CGATCCGGCCGACGGCGAGTTCGACACCTCGAC-CGC---ACCG---ACGACCCCG----CCGG-- CGCAGGCGCCGAGCGCCGGCAGCGACGACGCCGGCGAGGCCGAC
		2799-2905	CGATC-- GCCGACGGCGACCACGACACCGCGGCACGCGCTACCGCTCGCGACGCCGAGACCCGGCTCGCAGCCGCTCA- CGCCGGCATCCTCGCGATCGTCGAGGCCGAC
GTE6-D2	101	24673-24771	GCCGTCGGCGCCGCGGTTATGTGGTTCTGGAACACCATTATCGCGCCGGCGTTCGCCGCGATCGGCGCGATCATCTC GGC--GTGGTGGACCGGCGTGCAG
		24541-24639	GCCGTCGGCGCCGTGTTTACCTGGCTGTGGCAGACGATCATCGTGCCGGCGTTCACCGCGATACGGGCCGT-- TTTCGACCTGTGGTGGGCCGGCGTGCAG
GTE6-D3	92	13584-13675	ACCGGCTCGAAGAGGACCAGGTACCGGACACCGGCGGCGAGGGCGGCAAGCTCGTCGAGTTCGAGAACGGCGTCGCG AAGTACGACGACGGC
		13326-13417	ACCCGATGGAAGAGGACGAGTCGCCCGACAAGGGTGCGCTCGGCGGCAAGCTGATCAGCTACGGCGACGGCCGCGCC GACTACGACGACGGC
GTE6-D4	85	5677-5761	GCGGTGAAGAAAGCCGCCGAGTACATGGATTACTCGTCGTTCTCGCAGCAGGGCCTCGTCGATCAACTCGTGTTCGA GGGATTCA
		5536-5620	GCCGTGAGCAAGGCCAAGGATTACCTCGAATACTCGGCGTTCTCCCGTTCGGGCCTGATCAAGCAACTCGCGTTCGA GGGATTCA
GTE6-D5	75	13772-13846	GGCGCCGAGGGCGGCGAACTCGTGTCGTTCGGTGACGGCGTGGCCGTGTATGACGACGGCACCGAGACCGACGGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		13616-13690	GGCGGCGAGGGCGGCAAGCTCGTCGAGTTCGAGAACGGCGTCGCGAAGTACGACGACGGCACCGAAACGAACGGC
GTE6-D6	60	44292-44351	CGCCCGCGCCAACACGCCGCCGTGGACGATCACGGCGTGGCTCGTCGTCGACCCGTCCGG
		1598-1655	CGACCGCGCCGGCAC-CCGACGAGTTCCGTCGCGTCGTGA-TCGGCGTCGACCCGTCGGG
GTE6-D7	55	55032-55085	CCGTGCACC-CCGAGGACGACTACACGATGACAGGAGTACCCACCGAATGACCGA
		49725-49779	CCGTGCATCACCGAGCACGGCACCACCACCACGGAAGGATCGACCGAATGACCGA
GTE6-D8	54	17792-17840	CTCGTGTCGATCGACGTGA--GCG---CCGAGTACGAGGACGGCACCGAGAACG
		13790-13841	CTCGTGTCGTTCG--GTGACGGCGTGGCCGTGTATGACGACGGCACCGAGACCG
GTE6-D9	54	47895-47948	CGCCGAACGTCACCCTGCACGGCGCCGACTGGGAGCCGATCAACCCGCTCGACC
		29827-29876	CGCCGGACGTG-CCCGGC--GGCGGCGACCTGCCGCCGATCAACCTGC-CGACC
GTE6-D10	52	10011-10062	GCCGAGGGCCGGCTGAACGAGGCGACGCAGGCTGCCGAGGTCGACGACGACG
		20057-20100	GCCGAGCGCCGGCAG--CGA--CGACGCCGGC----GAGGCCGACGACGACG
GTE6-D11	52	46606-46656	CCGGCAAGGGCGCCGGCGCCGGCACCGATAAGACCGGCGTC-GCCGCCGGCA
		3374-3425	CCGGCGTGAGCGCGGTCGCCGGCACCGTGGCGATCATCGTCGGCCGCCGGCA
GTE6-D12	50	47664-47712	CGACGCCGTGCACGAGGTGATCGACCG-GTTCGGGCAGCAGATCGCCGAC
		9905-9953	CGACGCCGTGCAGGCCGAGTTCGACCGTGTGCAGG-ATGAGTTCGCCGAC
GTE6-D13	49	51782-51830	CATCGACCGCGACGACGAGCACCCGGACCTCGACCCCGACGTGCCGGTC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		35100-35142	CATCGAGGTCGACGACGAGCACACGGTGCT-GA-----ACGTGCCGGTC
GTE6-D14	48	35888-35935	AGGCGTACGCCGCGCAGAAGTGGGAGCAGGGCGCGCTCGGCTGGCCGA
		35714-35761	AGGCGTACGCCGCGCGCAACTGGGAGGCCGGCCCGCTCGGGTTCCCGA
GTE6-D15	47	45972-46016	AGCCGAGGCGGATCTCGCCGCAGT--CGATCAGGCGCTCGCCGAGCT
		10310-10354	AGCCGAGGCCGAAG--GCATCAGTTACGAGCAGGCGCTCGCCGAACT
GTE6-D16	46	44103-44147	CGACGCACGCGGCAACCTGACCGTGC-CGACGCCGGCCGAACTCGA
		34536-34579	CGACGCAGACGGCAACCTGACGTTGCGCAACCTCG--CGAACTCGA
GTE6-D17	45	15282-15325	CGATCTACGCGGGCGGCATGGC-CGGCCGTGTGATGACCGAGGCG
		1826-1869	CGAACTACGGCGGCGGCATGGCTCGGCAG-CTGATCACGCAGGCG
GTE6-D18	44	40882-40922	CGACGC---ACCAGCCGACCCCGCAGACCTCACCGCCGACGAGG
		7926-7967	CGACGCTCGACAAGCCGACCCCGCAGA--TCATCCCCGGAGAGG
GTE6-D19	41	50808-50848	GCCGGCGCGCGCTGATCGCCGAGATCGAGGTCGACGACGAC
		9689-9728	GCCGGAACACGCCGA-CGACGTGCTCGAAGTCGACGACGAC
GTE6-D20	40	50765-50804	GCCGACGGCGTGCAGTTCCGGTTCGACATGCCGGTGACCG
		542-575	GCCGACGGCGTGGA--TC----TCGACCTGCCGGTGACCG
GTE6-D21	38	53591-53625	CACCGGCGCACCGGCCGG---CGTCGCCGCGGCCGGCA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		5376-5413	CACCCGCGCAATCGCCGGACTCGTCGCCGCGGCCGGCA
GTE6-D22	38	54699-54736	CGTGTTCGTCGTCGACGGCAACACCCCGCCGAAGGCGC
		39217-39254	CGTGAACCTCGTCGACGGCGAGACCGCGGCCAAGGCGC
GTE6-D23	38	48464-48501	ACTCGGCTACCCCGCCGGCGTACTCGTCGCCGTCGACG
		5394-5431	ACTCGTCGCCGCGGCCGGCATCGTCGTCGCCGTCGCCG
GTE6-D24	37	30472-30505	GCAGAACGTCGACCCCGACG---CGCTCGGCGCCGAG
		26696-26732	GCAGCACGACGACCCCGACGACACCCTCGACGCCGAG
GTE6-D25	36	9328-9363	TCGACGAGCAGGGCGGTAAGGGGCCGGTCGAGCCGG
		12234-12269	TCGACGAGCAGTTCGACGACGAGCCGGTCGAGCCGG
GTE6-D26	36	43690-43722	CGACCGTCGACGA---CAGTCGCGAGAACGCCGCAC
		477-512	CGACCGTCGACGACCTCGATCGCGAGATCGCCGAAC
GTE6-D27	36	54166-54201	GTGGGCCGCCGGCTACCTCGTGGCCGTGATGAACGC
		52545-52580	GTGGGCCGCCGGGCACGTCGTCGTCGTGGTCACCGC
GTE6-D28	36	22768-22801	GCGGGTGCCGGCATCGC--CGCGATCGGTGCGCTCG
		3662-3696	GCCGGTGCCGGCATCGCGGGGCGA-CGTTGCCCTCG
GTE6-D29	36	19571-19603	CCGCAGGATG-CGCCGA--GCGTGGTCGCGCTCGAC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		18329-18364	CCGCCGGCTGACACCGACGGCGCGGTCGCGCTCGAC
GTE6-D30	35	36462-36496	GCAGGCCGCCGGCATCGCCGGCACGCTCGCGCTCG
		31807-31841	GCTGGCCGCCGGCAGCGCGGTCACGCTCGTGTTCG
GTE6-D31	35	18902-18936	TCGGTGGCTGCGCCGCGTCGACGGGCTCGCGTGGC
		5161-5193	TCGGCGACGGCGCC-CG-CGACGGGCTCGGGTGGC
GTE6-D32	35	52033-52067	CTGCCGGCGACGGCGTGTTCCCCCTCGGCTCGATC
		23438-23469	CTGCCGGCGACGGCAATTT---CCTCGGCTCGCTC
GTE6-D33	35	11698-11732	GATCGCGCTCACTGCCCGCGACGTGAAGGCGCGCG
		4640-4674	GACCGGGCTGACCGCGAGCAACGTGAAGGCGCTCG
GTE6-D34	35	20832-20866	CGACCGACACCGGCGAGGATCGCGAGGTCGAGCGC
		14649-14683	CGACCGACACCGGTGACCCTGCCAAGGTCGAGGGC
GTE6-D35	35	53694-53728	ACCCCGCTCGGCGTCGTCATCTACAACACCGGCAA
		49913-49947	ACCCCGCTCGGGTTCGACGTGCAGATCACCGGCAA
GTE6-D36	34	48297-48330	AGTAGCATCACGCCGTACCCACCAAGTACCGACC
		46167-46197	AGTAGCATCACCGTGTACCCACCA---ACCGACC
GTE6-D37	34	33945-33976	CTCGCCGGCGTCGG-GCCG-GATCTCGGCGATCA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		30599-30632	CTCGACGGCATCGGCGACGTGATCTCGGCGATCA
GTE6-D38	34	52731-52764	CGAGTTCGAGATCGTCGGCGCGATCGACGACAAC
		13639-13672	CGAGTTCGAGAACGGCGTCGCGAAGTACGACGAC
GTE6-D39	34	46298-46331	CGACGGACTCAGCGACGCCGACCACGCCGCACTG
		38893-38923	CGACGGAC--AGGTAC-CCGACCACGCCGAACTG
GTE6-D40	33	2648-2680	CGACGACCTCGCTGACGAACTCGCCGAACGCGC
		484-516	CGACGACCTCGATCGCGAGATCGCCGAACTCGC
GTE6-D41	32	46709-46740	CGCGCTCGTGCTCGCCTCGATCGGCGCGGTCA
		24714-24745	CGCGCCGGCGTTCGCCGCGATCGGCGCGATCA
GTE6-D42	30	32916-32945	CGGCGGCAAGCTCGTCACGTCGCAGGTGCC
		12103-12132	CGGCGGCAAGCTCGCCGCGTCGCTGCTGCC
GTE6-D43	30	33084-33113	GACCGGCTGGCAGCAGGTCGCCTATCCGGC
		3121-3149	GAGCGGCTAG-AGCAGATCGCCTATCCGGC
GTE6-D44	30	52677-52705	CCGACCCG-ACGTGGACGTGAAGCGCGACA
		7985-8014	CCGAACCGCACGTCGACGAGAAGCGCGACA
GTE6-D45	30	52511-52540	CGAACGCGCAGGCGATCGCGTATTGCGTGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		48598-48627	CGCACGCGAAGGCGATCGCGAACGGCGTGC
GTE6-D46	30	39142-39171	GCGGTACGAGCCGGCGCTCACTGCCCGAGA
		11689-11718	GCGGGTCGAGATCGCGCTCACTGCCCGCGA
GTE6-D47	30	51856-51885	CGGTCGAGCCGGTACTCACGCACGGCAAGA
		12258-12287	CGGTCGAGCCGGTGCACACCGGCGGCATGA
GTE6-D48	29	47935-47960	CAACCCGCTCGAC---CCGGTGCCGGCAT
		3647-3675	CAACCCGCTCGACGAGCCGGTGCCGGCAT
GTE6-D49	29	47398-47426	ACCCCGAGGTCGACACCGACGAGAAGGGC
		20658-20686	ACCCCGACGTAGACACCGACGACATGGGC
GTE6-D50	29	33426-33453	CGACGGC-ACGATCGTCGACGGCGACGTG
		33290-33318	CGACGGCGAAGATCGTCGACGCCAACGTG
GTE6-D51	29	29418-29446	GCCGGGCCGCCCGTACTGGCCGGGCGCAC
		2058-2086	GCCGGGCAGCACGTGGTCGCCGGGCGCAC
GTE6-D52	28	12343-12370	CGCCGAGCTACACCTCACGCTCGCGTTC
		10448-10475	CGACGAGCTACACCGCCGGCTCGCGTTC
GTE6-D53	28	37577-37603	CCACC-AGAGAGGCACGACCATGACCGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		36038-36065	CCACCTAGAAGGGCACGACCATGACCGA
GTE6-D54	28	53799-53825	CAGCCGTG-CGACGCCGCCGGCACGATC
		4536-4563	CAGCCGCGACGAGGCCGCCGGCACGTTC
GTE6-D55	28	11858-11885	CTACTCGCTGCCGGCGTCGAGTCGGCCG
		8155-8181	CTGCTCGCTGCCGGCGACGACT-GGCCG
GTE6-D56	28	30490-30515	CGCGCT--CGGCGCCGAGGCGAACCTCG
		17146-17173	CGCGCTGCCGGCGCCGAGGCGCATCTCG
GTE6-D57	28	24717-24744	GCCGGCGTTCGCCGCGATCGGCGCGATC
		22774-22800	GCCGGCAT-CGCCGCGATCGGTGCGCTC
GTE6-D58	28	14387-14414	GAGGTCGACGCGCTGCTGAACGAGGCGA
		10008-10035	GAGGCCGAGGGCCGGCTGAACGAGGCGA
GTE6-D59	28	49298-49325	CGACACCCTCGAATCGAAGTACGACGAC
		26714-26741	CGACACCCTCGACGCCGAGCACGACGAC
GTE6-D60	27	32385-32411	GACGGCGACCGTGCCGGCGCTCGTCGA
		1593-1619	GACGGCGACCGCGCCGGCACCCGACGA
GTE6-D61	27	9479-9505	GCAGTCGATGCCGGCCGAGCCGGTCGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		3174-3200	GCCGTCGAAGCCGGCCGAGCAGTTCGA
GTE6-D62	27	9985-10011	CCGAGGCGCTTGCACGGCTCGACGAGG
		4762-4788	CCGAGGCGATCGCGTGGCTCGACGAGG
GTE6-D63	27	47691-47714	GTTCGGGCAGCAG---ATCGCCGACCG
		11797-11823	GTCCGGGCAGCAGCTGATCGCCGACCG
GTE6-D64	27	52444-52466	CACCTCGTCGACGAC----GCCGGCCG
		29961-29987	CACCTCGTCGACGACCAGGGCCGGCCG
GTE6-D65	27	30259-30283	GGCTCGTCGACTCGGA-GCG-TGCCCG
		12531-12557	GGCTCGTCGACTCGCATGCGCTGCCCG
GTE6-D66	27	55827-55853	CCGACCCGCGAGATCGCCGAGACCGAG
		43612-43637	CCGACGTGC-AGATCGCCGAGGCCGAG
GTE6-D67	26	18912-18937	CGCCGCGTCGACGGGCTCGCGTGGCC
		1080-1105	CGGCGGGTCGATGCGCTCGCGTGGCC
GTE6-D68	26	54440-54465	CCGAAGCTCGCCGAGGTCGACGACGC
		7007-7032	CCGAAGCTCGCCGCGCTGGCCGACGC
GTE6-D69	26	22671-22696	CACCGCGATCGCCGGCGCCGCTACCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		12703-12728	CATCGCGATCGCCGGCGACGTGACCG
GTE6-D70	26	52235-52260	CTGCTCGACGACGACCTCGCCGAGGA
		40310-40335	CTGCTCGAATACGGACTCGCCGAGGA
GTE6-D71	26	25815-25837	CACCGGGCCGGCCGGT---GCGCCGG
		16862-16887	CACCGGGCCGGCCGATCAGGCGCCGG
GTE6-D72	25	45215-45239	CCGCAGGCGATCGCCGGCACCGCAC
		16983-17007	CCGCAGACGAACACCGGCACCGCAC
GTE6-D73	25	35489-35513	TCGCGACCGCGCCGGCACCGATCAA
		19272-19296	TCGTGACCGCGCCGGCGACGATCAA
GTE6-D74	25	32592-32616	CGCGACGACCACGGCGAAGGGCGGC
		27588-27612	CGCGACGACCAGGGCGACGGCCGGC
GTE6-D75	25	18916-18940	GCGTCGACGGGCTCGCGTGGCCGAC
		3227-3249	GCGTCGACGGGCT--CGTGCCCGAC
GTE6-D76	25	37761-37785	GACCTGCTCGACAAGCTCAACACCG
		10440-10462	GACCTGCTCGACGAGCT--ACACCG
GTE6-D77	25	48602-48626	CGCGAAGGCGATCGCGAACGGCGTG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		529-553	CGCGAAGGCGAAGGCCGACGGCGTG
GTE6-D78	25	40116-40140	TCATGACGTGCGCACAGCTGGGCCT
		2299-2323	TCAAGCCGGGCGCACAGCTGCGCCT
GTE6-D79	25	39783-39807	CGCGCTACTCGCCGCCGACACCGAG
		11854-11878	CGCGCTACTCGCTGCCGGCGTCGAG
GTE6-D80	25	17142-17166	GCTGCGCGCTGCCGGCGCCGAGGCG
		11857-11881	GCTACTCGCTGCCGGCGTCGAGTCG
GTE6-D81	25	51903-51927	CGCCGGGCTCAACTCGACCAGCTGC
		44132-44155	CGCCGGCCG-AACTCGACCAGCTGC
GTE6-D82	24	47177-47200	CGCCGACGGGCACGCCGACCTCGC
		39346-39369	CGCCGACGGGCACGCCGTCATCGC
GTE6-D83	24	45768-45791	CCTCGACGGGCTGATCGACGCCGC
		39819-39842	CCTCGACGGGCTGATCGAGGACGC
GTE6-D84	24	54113-54135	GGCCGC-GACTACTCGGGCGGGCT
		51229-51252	GGCCGCCGACTACTCGGGCGGGCT
GTE6-D85	24	17817-17840	AGTACGAGGACGGCACCGAGAACG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		13662-13684	AGTACGACGACGGCACCGA-AACG
GTE6-D86	24	30408-30431	CACCCGGCGGCGGCACGCCGATCG
		16820-16842	CACCCGGCT-CGGCACGCCGATCG
GTE6-D87	24	34518-34541	CGTCGAGACCGCCGTGAACGACGC
		11872-11895	CGTCGAGTCGGCCGTGAACGATGC
GTE6-D88	24	44802-44825	ACCGGGATCGCCGCCGGCTACCGG
		34204-34227	ACCGGTATGGGCGCCGGCTACCGG
GTE6-D89	24	50070-50093	CGGCCGACGACGAGGCGCCGGTCG
		40752-40775	CGGCCGACGACGAACCGCCGGCCG
GTE6-D90	24	11600-11620	CTCGTCGCCGC---CGGCATCGTC
		5395-5418	CTCGTCGCCGCGGCCGGCATCGTC
GTE6-D91	23	31762-31784	CAAGCAGCCGTTCGAGTTCTACG
		19402-19424	CAAGCCGCCGTTCGAGTTCTGCG
GTE6-D92	23	23772-23794	GCTCGGGCAGCTGATCGCGCAGG
		1846-1867	GCTCGG-CAGCTGATCACGCAGG
GTE6-D93	23	1320-1342	CGGCCGGGTCGAAGAGGGCGGCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		14607-14627	CGGCCGG--CGAAGAGGGCGGCG
GTE6-D94	23	33456-33478	CGGGTCGGCCGGCATCGCCGCGA
		22769-22789	CGGGT--GCCGGCATCGCCGCGA
GTE6-D95	23	40090-40112	CGAGGCGGCGACCGTGCTCGGCG
		32382-32403	CGAGACGGCGACCGTGC-CGGCG
GTE6-D96	23	44247-44269	GTGCCTGCTCGACCCCGAACCGT
		3608-3630	GTGCCTGCTCGACGGCGAGCCGT
GTE6-D97	23	12094-12116	GAACAAAGACGGCGGCAAGCTCG
		7162-7184	GAACAAAGACGGCGAGAAGGTCG
GTE6-D98	23	22872-22894	CGAGAACGCCGAGAAGTTCAACG
		7558-7580	CGAGAACGCCGACGAGTTCAGCG
GTE6-D99	23	33762-33784	CGGGCTCGGCGGCGACTGGTCGA
		31078-31100	CGGGCTCGGCGCCGACGGGACGA
GTE6-D100	23	46389-46411	GTCCGCCGGCGACGCGTCCAAGC
		34104-34126	GTCCGCCGGCGACGTGTTCGAGC
GTE6-D101	23	49043-49065	CGACCGAGGCAAGCGGGCCGTCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		45948-45970	CGACCAACGCAAGCGGGCCGACG
GTE6-D102	23	55182-55204	CGGCGAGATCTACCCGTACAGCG
		47983-48005	CGGCGCGATCCGCCCGTACAGCG
GTE6-D103	22	44355-44376	CGTCGCCGCACACCGCGGCCAC
		25766-25787	CGTCGTCGCACACCGCGGCCAC
GTE6-D104	22	46298-46319	CGACGGACTCAGCGACGCCGAC
		5610-5631	CGAGGGATTCAGCGACGCCGAC
GTE6-D105	22	53441-53462	GATCGACGGCCGCCGGCTCGGC
		38728-38749	GATCGTCGCCCGCCGGCTCGGC
GTE6-D106	22	49253-49274	CGACGACGGGCTCGTGACCCGA
		3228-3248	CGTCGACGGGCTCGTG-CCCGA
GTE6-D107	21	27252-27272	CAGTGGGCCTATCAGGAGGGC
		7634-7654	CAGTGGGCCTATCAGCAGGGC
GTE6-D108	21	7821-7841	TCGACCTGCCGGCGACCGCGG
		558-578	TCGACCTGCCGGTGACCGAGG
GTE6-D109	21	32516-32536	CGGCCGCCGGCAGTGCCGACG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		3414-3434	CGGCCGCCGGCAGTACCGCCG
GTE6-D110	21	15891-15911	CGACCAGCGTGAAGGCGCTCG
		4654-4674	CGAGCAACGTGAAGGCGCTCG
GTE6-D111	21	8841-8861	GACCGTATCGGCCGTGTGCTG
		7721-7741	GAACGTACCGGCCGTGTGCTG
GTE6-D112	21	50955-50975	AGGCCGCGCTCGACGTGTACG
		10672-10692	AGGCCGGGCTCGACGTGTTCG
GTE6-D113	21	47844-47864	AGATCCCCGAGGATCTGTTCG
		15042-15062	AGTTCCCCGAGGATCTGGTCG
GTE6-D114	21	50399-50418	CCCCGAAGAGGTCGA-GCGCA
		15457-15477	CCCCGAAGAGGTCGAGGCGCA
GTE6-D115	21	51147-51167	TCGGCGCGATCGCGAACGCCG
		24014-24034	TCGTCGCGCTCGCGAACGCCG
GTE6-D116	21	37574-37594	ACCCCACCAGAGAGGCACGAC
		29686-29706	ACCCCACCAGAGAGGACCGAC
GTE6-D117	21	41203-41223	CGCCGACGGCACGTTCCTCGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		33423-33443	CGCCGACGGCACGATCGTCGA
GTE6-D118	21	42270-42290	GCTCGCGCTCGCCGGGCACCT
		34769-34789	GCTCGGGCTCGCCGGGCAGCT
GTE6-D119	20	30694-30713	GTTCGGCAACCTGCTCGACG
		6763-6782	GTTCGGCAACCTGCCCGACG
GTE6-D120	20	21748-21767	CCGGCGCTGACGACATGGCT
		11430-11449	CCGGCACTGACGACATGGCT
GTE6-D121	20	19366-19385	GGCCGGTGTCCCTGGCCGTG
		14899-14918	GGCCGGTGTCCGTGGCCGTG
GTE6-D122	20	33110-33129	CGGCGGCACCGGTGTCGAGC
		15847-15866	CGGCGGCACCGGTGTCGGGC
GTE6-D123	20	36473-36492	GCATCGCCGGCACGCTCGCG
		18736-18755	GCATCGCCGGCACGATCGCG
GTE6-D124	20	35949-35968	CGCAGCACCCGAGTTCGGCA
		30317-30336	CGCACCACCCGAGTTCGGCA
GTE6-D125	20	39219-39238	TGAACCTCGTCGACGGCGAG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		1031-1050	TGAACCTCGTCGCCGACGAG
GTE6-D126	20	56961-56980	ACGCCGGCGACCAAACCGAC
		2407-2426	ACGCCGGCGACCACGCCGAC
GTE6-D127	20	52242-52261	ACGACGACCTCGCCGAGGAA
		2647-2666	ACGACGACCTCGCTGACGAA
GTE6-D128	20	15568-15587	GAACTTCACCGGCACGTTCG
		3842-3861	GAATTTCACCGGCACGATCG
GTE6-D129	20	17199-17218	GCCGGCGTTCGAGGTGCCCG
		4091-4110	GCCGCCGTTCGAGGTGCACG
GTE6-D130	20	42489-42508	GCGCCGGCAACGGGCTCGGG
		5170-5189	GCGCCCGCGACGGGCTCGGG
GTE6-D131	20	34865-34884	ACCTCGTGGCCGGCGACGTG
		6690-6709	ACCACGAGGCCGGCGACGTG
GTE6-D132	20	17143-17162	CTGCGCGCTGCCGGCGCCGA
		8155-8174	CTGCTCGCTGCCGGCGACGA
GTE6-D133	20	25514-25533	GGCTCGCGTCGAAGGATCTC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		8917-8936	GGCTCGCGTCGCCGGATCTC
GTE6-D134	20	11266-11285	CGACGGCGATCTGCCCGTCG
		9830-9849	CGACGACGATCTGCCCGGCG
GTE6-D135	20	50720-50739	CGTCGACGATCTGCTCGGCG
		9830-9849	CGACGACGATCTGCCCGGCG
GTE6-D136	20	30953-30972	GCCGACGCCGCGGCCGCGCA
		11327-11346	GCCGTGGCCGCGGCCGCGCA
GTE6-D137	20	50046-50065	CGGTCGCGTTCCTCGGCGAG
		12360-12379	CGCTCGCGTTCCTCGGTGAG
GTE6-D138	20	36467-36486	CCGCCGGCATCGCCGGCACG
		23267-23286	CCGACGGCATCGCCGGCGCG
GTE6-D139	20	49026-49045	TCGCTGACCTCGATCATCGA
		23476-23495	TCACTGACCTCGATCAACGA
GTE6-D140	20	38697-38716	TCGGGCGCGAGGTCGAGTCG
		26546-26565	TCGGGCGCGAGGACGAGGCG
GTE6-D141	20	37811-37830	CGACGTGAAGATCAAAAGCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		28827-28846	CGACGTGAAGAACAAGAGCG
GTE6-D142	20	32837-32856	CGCACGCGACGACCAAGGCG
		32588-32607	CGGACGCGACGACCACGGCG
GTE6-D143	20	48857-48876	CACCATCGTCGACGGCGCCG
		33432-33451	CACGATCGTCGACGGCGACG
GTE6-D144	20	46615-46634	GCGCCGGCGCCGGCACCGAT
		35491-35510	GCGACCGCGCCGGCACCGAT
GTE6-D145	20	52239-52258	TCGACGACGACCTCGCCGAG
		36164-36183	TCGTCGCCGACCTCGCCGAG
GTE6-D146	20	49883-49902	GCCGCCACACTGCCCGGCGT
		37141-37160	GCCGCCAAGCTGCCCGGCGT
GTE6-D147	20	46783-46802	ACATGAACCTCGTCGCCGGC
		39216-39235	ACGTGAACCTCGTCGACGGC
GTE6-D148	19	43754-43772	CGCCGGCATCCTCGGGATC
		2875-2893	CGCCGGCATCCTCGCGATC
GTE6-D149	19	53516-53534	GCAGCTGGCCGCCGAGGAT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		3435-3453	GCAGCTGGCCGCCGAGCAT
GTE6-D150	19	39673-39691	TGGATGACCGACCCGAACG
		3493-3511	TGGATGAACGACCCGAACG
GTE6-D151	19	37941-37959	CTCGACGACGGCAGCCTGC
		4206-4224	CTCGACGACGGCAGCGTGC
GTE6-D152	19	50812-50830	GCGCGCGCTGATCGCCGAG
		4670-4688	GCTCGCGCTGATCGCCGAG
GTE6-D153	19	12570-12588	AGATCCTGCGCGCGCTCGG
		7800-7818	AGATCCTGCGCGCGATCGG
GTE6-D154	19	9934-9952	TGCAGGATGAGTTCGCCGA
		9316-9334	TGCAGGATGAGTTCGACGA
GTE6-D155	19	11885-11903	GTGAACGATGCGACCGGGC
		10572-10590	GTGAACGATGCGACCGCGC
GTE6-D156	19	33843-33861	CGTCGCGACCACCGACGGC
		11254-11272	CGTCCCGACCACCGACGGC
GTE6-D157	19	46643-46661	CGTCGCCGCCGGCATGGTC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		11602-11620	CGTCGCCGCCGGCATCGTC
GTE6-D158	19	24171-24189	CTCGGCGGTGCGCACCGCG
		22593-22611	CTCGGCGGTGCGCGCCGCG
GTE6-D159	19	16619-16637	GACGACCACCACGCCGGCG
		26655-26673	GACGACGACCACGCCGGCG
GTE6-D160	19	30875-30893	ATCGCCGACGTGATCTCGG
		30608-30626	ATCGGCGACGTGATCTCGG
GTE6-D161	19	41197-41215	GCTCGGCGCCGACGGCACG
		31081-31099	GCTCGGCGCCGACGGGACG
GTE6-D162	19	35491-35509	GCGACCGCGCCGGCACCGA
		31713-31731	GCGACCCCGCCGGCACCGA
GTE6-D163	19	42269-42287	TGCTCGCGCTCGCCGGGCA
		32996-33014	TGCTCGCGCTCACCGGGCA
GTE6-D164	19	51789-51807	CGCGACGACGAGCACCCGG
		34239-34257	CGCGACGACGAGCGCCCGG
GTE6-D165	19	40696-40714	GGCACCCGGCGTCGACGAC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		39106-39124	GGCACTCGGCGTCGACGAC
GTE6-D166	19	47782-47800	CGTCGACCCGAAGTGCGAC
		46679-46697	CGTCGACCCGAAGGGCGAC
GTE6-D167	19	53399-53417	CGACGACGCCGAGGTCGAC
		47393-47411	CGACGACCCCGAGGTCGAC
GTE6-D168	19	50342-50360	ATCGCCCGTCGCCGACGTG
		48015-48033	ATCGCCCGTCGCCGCCGTG
GTE6-D169	18	21572-21589	CGGGCTCGACGTGTTCGA
		10676-10693	CGGGCTCGACGTGTTCGA
GTE6-D170	18	34363-34380	TCGCTGGCTAGGGGCCGG
		14197-14214	TCGCTGGCTAGGGGCCGG
GTE6-D171	18	34365-34382	GCTGGCTAGGGGCCGGGC
		17626-17643	GCTGGCTAGGGGCCGGGC
GTE6-D172	18	8781-8798	GACATACGGCCGGAACTG
		637-654	GACCTACGGCCGGAACTG
GTE6-D173	18	39793-39810	GCCGCCGACACCGAGGGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		1459-1476	GCCGACGACACCGAGGGC
GTE6-D174	18	43162-43179	GACCAACCGACATGACCG
		4463-4480	GACCAACCGACATGATCG
GTE6-D175	18	44661-44678	CTCGCCGGCTACAGCCTC
		4710-4727	CTCGTCGGCTACAGCCTC
GTE6-D176	18	51438-51455	GGCCGAGCACGGCGCCGC
		5772-5789	GGCCGAGCACGGAGCCGC
GTE6-D177	18	12840-12857	CGATCGAGGCCGCCGGCG
		8578-8595	CGATCGAGGCCGCCGTCG
GTE6-D178	18	19019-19036	GGCGGGGCAGCTGGCCGC
		9065-9082	GGCCGGGCAGCTGGCCGC
GTE6-D179	18	10046-10063	CGAGGTCGACGACGACGG
		9713-9730	CGAAGTCGACGACGACGG
GTE6-D180	18	41294-41311	GCGATCGACGACGACGAG
		9756-9773	GCGAGCGACGACGACGAG
GTE6-D181	18	29514-29531	GTACGGCATCGACACCGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		17484-17501	GTTCGGCATCGACACCGA
GTE6-D182	18	22818-22835	GATCAAGACCGCGTTCGC
		18025-18042	GATCGAGACCGCGTTCGC
GTE6-D183	18	41293-41310	CGCGATCGACGACGACGA
		18851-18868	CGCGATCGACGGCGACGA
GTE6-D184	18	55276-55293	GCCGGCTACCGCGTGATG
		19084-19101	GCCGGCTACCGCGTGGTG
GTE6-D185	18	39716-39733	TCGAGCAGCTGATCGCGC
		23774-23791	TCGGGCAGCTGATCGCGC
GTE6-D186	18	56065-56082	CGCGGCCGGGCTGCAGAA
		24399-24416	CGCTGCCGGGCTGCAGAA
GTE6-D187	18	55850-55867	CGAGCACCGCGTCGGGCA
		25142-25159	CGAGCGCCGCGTCGGGCA
GTE6-D188	18	46526-46543	GGGCACCCTGCTCGACGG
		28863-28880	GGGCACGCTGCTCGACGG
GTE6-D189	18	48152-48169	GCCGACACCGCGGCCGCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		30953-30970	GCCGACGCCGCGGCCGCG
GTE6-D190	18	44495-44512	CCGAGCGCGCGGCGTGCT
		37220-37237	CCGAGCGCGCGGCGTCCT
GTE6-D191	18	53359-53376	AACGACTCGCGAAGGTAC
		38162-38179	AACGACTCGCGACGGTAC
GTE6-D192	18	51782-51799	CATCGACCGCGACGACGA
		39010-39027	CATCGACCGCGACGAGGA
GTE6-D193	18	44054-44071	CGGCCGTCGACGCGCTGC
		40650-40667	CGGCCGTCGACGCGCAGC
GTE6-D194	18	56104-56121	GCAGCTGCCCGACAACCT
		45436-45453	GCAGCTGCCCGACGACCT
GTE6-D195	17	35491-35507	GCGACCGCGCCGGCACC
		1597-1613	GCGACCGCGCCGGCACC
GTE6-D196	17	31515-31531	CGCTGTGGCAGCCGGGC
		2048-2064	CGCTGTGGCAGCCGGGC
GTE6-D197	17	38438-38454	CCCACCAGCGAAGGACA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		5924-5940	CCCACCAGCGAAGGACA
GTE6-D198	17	49448-49464	GCTGCTCGCCGCGCTCG
		23922-23938	GCTGCTCGCCGCGCTCG
GTE6-D199	17	50831-50847	ATCGAGGTCGACGACGA
		35101-35117	ATCGAGGTCGACGACGA
GTE6-D200	16	11466-11481	GCTCGTCGCTGTCGGG
		1686-1701	GCTCGTCGCTGTCGGG
GTE6-D201	16	22609-22624	GCGAAGTCGATCGCCG
		2792-2807	GCGAAGTCGATCGCCG
GTE6-D202	16	16881-16896	GCGCCGGCGGCCGGGC
		2940-2955	GCGCCGGCGGCCGGGC
GTE6-D203	16	54449-54464	GCCGAGGTCGACGACG
		10044-10059	GCCGAGGTCGACGACG
GTE6-D204	16	50833-50848	CGAGGTCGACGACGAC
		10046-10061	CGAGGTCGACGACGAC
GTE6-D205	16	11815-11830	CGCCGACCGCCGGCGG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		10265-10280	CGCCGACCGCCGGCGG
GTE6-D206	16	41018-41033	GAGCAGATCAAGGGGC
		12029-12044	GAGCAGATCAAGGGGC
GTE6-D207	16	17626-17641	GCTGGCTAGGGGCCGG
		14199-14214	GCTGGCTAGGGGCCGG
GTE6-D208	16	42111-42126	ACGATCCGGCCGACGG
		19996-20011	ACGATCCGGCCGACGG
GTE6-D209	16	36123-36138	CGTGGCCGTCGCCGGC
		22695-22710	CGTGGCCGTCGCCGGC
GTE6-D210	16	30568-30583	CGGCGGGTTCGCGTCG
		25686-25701	CGGCGGGTTCGCGTCG
GTE6-D211	16	44867-44882	CCCGAACGACGGCATC
		34602-34617	CCCGAACGACGGCATC
GTE6-D212	16	56494-56509	CGACGGCAAGCCGATC
		48437-48452	CGACGGCAAGCCGATC
GTE6-D213	15	39115-39129	CGTCGACGACCTCGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		481-495	CGTCGACGACCTCGA
GTE6-D214	15	19021-19035	CGGGGCAGCTGGCCG
		740-754	CGGGGCAGCTGGCCG
GTE6-D215	15	46786-46800	TGAACCTCGTCGCCG
		1031-1045	TGAACCTCGTCGCCG
GTE6-D216	15	56466-56480	TCAAGCCGGGCGCAC
		2299-2313	TCAAGCCGGGCGCAC
GTE6-D217	15	25301-25315	CGCACCTCGCCGGCG
		2757-2771	CGCACCTCGCCGGCG
GTE6-D218	15	5084-5098	CAAGGTGCGCCGCGA
		3039-3053	CAAGGTGCGCCGCGA
GTE6-D219	15	19024-19038	GGCAGCTGGCCGCCG
		3434-3448	GGCAGCTGGCCGCCG
GTE6-D220	15	33429-33443	CGGCACGATCGTCGA
		3851-3865	CGGCACGATCGTCGA
GTE6-D221	15	43745-43759	CACCGAGGCCGCCGG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		3914-3928	CACCGAGGCCGCCGG
GTE6-D222	15	52612-52626	GCGTGGCTCGACGAG
		4773-4787	GCGTGGCTCGACGAG
GTE6-D223	15	40171-40185	GCTGCCGTTCTGGGA
		8334-8348	GCTGCCGTTCTGGGA
GTE6-D224	15	22126-22140	GGCGCGTTCGAGGAC
		8769-8783	GGCGCGTTCGAGGAC
GTE6-D225	15	9563-9577	CGAGGAACGGCTCGC
		8909-8923	CGAGGAACGGCTCGC
GTE6-D226	15	10048-10062	AGGTCGACGACGACG
		9823-9837	AGGTCGACGACGACG
GTE6-D227	15	35103-35117	CGAGGTCGACGACGA
		10046-10060	CGAGGTCGACGACGA
GTE6-D228	15	19563-19577	CGAACCGGCCGCAGG
		10243-10257	CGAACCGGCCGCAGG
GTE6-D229	15	45436-45450	GCAGCTGCCCGACGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		10958-10972	GCAGCTGCCCGACGA
GTE6-D230	15	55432-55446	GCCGAGGTCGAGGCG
		11642-11656	GCCGAGGTCGAGGCG
GTE6-D231	15	25864-25878	GCCGACGACGCGCAG
		11672-11686	GCCGACGACGCGCAG
GTE6-D232	15	18082-18096	CGCCGGCGGCAAGCG
		12850-12864	CGCCGGCGGCAAGCG
GTE6-D233	15	32917-32931	GGCGGCAAGCTCGTC
		13625-13639	GGCGGCAAGCTCGTC
GTE6-D234	15	44059-44073	GTCGACGCGCTGCTG
		14390-14404	GTCGACGCGCTGCTG
GTE6-D235	15	41133-41147	CCGAGGCCGGCGAGG
		14691-14705	CCGAGGCCGGCGAGG
GTE6-D236	15	39277-39291	CGACGGGCAGGTGCT
		15079-15093	CGACGGGCAGGTGCT
GTE6-D237	15	40366-40380	CGGCGCCCGAGGTGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		16939-16953	CGGCGCCCGAGGTGC
GTE6-D238	15	37909-37923	ACGATCGCGGCCGCG
		18747-18761	ACGATCGCGGCCGCG
GTE6-D239	15	53399-53413	CGACGACGCCGAGGT
		19624-19638	CGACGACGCCGAGGT
GTE6-D240	15	50855-50869	CTCACCCTGTACACG
		20924-20938	CTCACCCTGTACACG
GTE6-D241	15	26988-27002	CGGCATGGACCCGAA
		22314-22328	CGGCATGGACCCGAA
GTE6-D242	15	33854-33868	CCGACGGCACGTACT
		26264-26278	CCGACGGCACGTACT
GTE6-D243	15	35982-35996	GCAGGACTTCGCCGG
		26799-26813	GCAGGACTTCGCCGG
GTE6-D244	15	44833-44847	CGCTCGGGCTCGGCA
		29195-29209	CGCTCGGGCTCGGCA
GTE6-D245	15	30818-30832	AACCTGCTCGACGCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		30701-30715	AACCTGCTCGACGCG
GTE6-D246	15	36676-36690	CGTGACCGGCGACGG
		32220-32234	CGTGACCGGCGACGG
GTE6-D247	15	44943-44957	CTCGACGACTGGGCG
		34288-34302	CTCGACGACTGGGCG
GTE6-D248	15	46428-46442	CTGAAAGCCGGCACC
		35599-35613	CTGAAAGCCGGCACC
GTE6-D249	15	53806-53820	GCGACGCCGCCGGCA
		39629-39643	GCGACGCCGCCGGCA
GTE6-D250	15	55628-55642	CGACGACTGGGCCGA
		45924-45938	CGACGACTGGGCCGA
GTE6-D251	15	51051-51065	TCGCCGCGCTCGTCG
		49603-49617	TCGCCGCGCTCGTCG
GTE6-D252	15	55921-55935	CAGCGAGGACTCGAC
		49966-49980	CAGCGAGGACTCGAC
GTE8-I1	45	61772-61816	CGGTGATTGTTCGCGATGACGACGAGATCATCGACGACGACGACG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		43836-43795	CGGTGATCGT---CGACGACGACGAGATCGTGGTGGACGACGACG
GTE8-I2	43	47400-47442	CTTCTTGGCCTTCGGGACCTTGCGGTATCCCTCGGCGGGGCCG
		24416-24374	CTTCTTGCCCTTCCGGACCTTGCGGCACTTGTCGACGCGGCCG
GTE8-I3	37	54497-54533	CTCCCGGACGCCGAACGTCACGCCGTCGAGAGCGCAG
		12896-12860	CTCCAGGACGCCGAACGGATCGACGTCGGGATCGGAG
GTE8-I4	37	39192-39228	TGTCTCCGAGCAGGCCGACGTCGCCGTCGTCGATCGC
		28000-27965	TGTCTCCGAGCTGGCCGACCTCTTCGTGGACG-TCGC
GTE8-I5	36	57386-57421	CGGCCGTCGTCGCACCGACACCGGCGCCGCAACCGC
		262-227	CGGACGCCGCCGGGCTGCCACCGCCGCCGCAACCGC
GTE8-I6	34	52495-52528	TCGGTGTAGGTGACCTGGTCGCTGTCGGCGTCCG
		11360-11327	TCGGCGTCGGTGTCGTCGTCCTTGTCGGCGTCCG
GTE8-I7	33	47695-47726	GGCCGTCAATGAGCATCT-TGGTCAGACGCTCG
		38729-38697	GGCCGTCAATGAGCACCTCCAGGCAGACGGTCG
GTE8-I8	31	52198-52228	ATCGCCTCCATCTCCTCGCTCGGGAGATGGT
		28592-28562	ATCGCCTCCATCTCGGCGCTCGTGATCTGGT
GTE8-I9	31	16481-16511	GACCGGCGCCGTCGTCACTGCGACGACCGCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		8143-8113	GACCGGCGCCGTCGCCCCTTCGTTGATCGCG
GTE8-I10	31	22744-22774	AGTTCGCGCTCGGCTTCGGCAAGAGCGCGCC
		11486-11457	AGTTCGCGCTCGGCTTTGTC-GGCGCGCGCC
GTE8-I11	29	58642-58670	CGCGGCCGACGCTCTCGCGCTGCTCGACG
		47891-47863	CGCGGCCGACGACCTCGTGTTCCTCGACG
GTE8-I12	28	38027-38054	CCCACTGCTTCAGCCGCTCGTCGGTGAT
		6815-6788	CCCACTGGTCCCGCTGCTCGTCGGTGAT
GTE8-I13	27	40585-40611	CAGGTCGGGGTACTTCGCGACGATCTC
		26994-26968	CAGGTCGAGGATCTTCGCGACGATCTC
GTE8-I14	26	52147-52172	TCGGGGACCGTCTTGCCGACGGTGGG
		11336-11311	TCGGCGTCCGTCTTGCCGCCGTTGGG
GTE8-I15	24	39201-39224	GCAGGCCGACGTCGCCGTCGTCGA
		10976-10953	GCTGGCCGACGTCGCGGCCGTCGA
GTE8-I16	24	35248-35271	CGACGTCGCCGAGCGCGTCGAGGA
		27697-27674	CGAGGTCCCCGAGCGCGTCGACGA
GTE8-I17	23	20467-20489	GCTCGCTGATCGGCGCCATCTTC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		18708-18686	GCTCGTTGATCGGCGCCATCTTC
GTE8-I18	23	30580-30602	GAACGCCGACCGCATCGCCATCG
		10771-10749	GAACGCCGACCGGTTCGCCATCG
GTE8-I19	23	14014-14035	TGTTCCTG-GGTCGGGCCTCTCG
		13917-13895	TGGTCCTGTGGTCGGGCCTCTCG
GTE8-I20	22	46050-46071	GCGACGATGTCGAAGGTGCCGT
		3703-3682	GCGCCGACGTCGAAGGTGCCGT
GTE8-I21	22	63600-63621	AAGATCGGTGAGCTGCGCAAGG
		37571-37550	AAGATCGCCGAGCTGCGCAAGG
GTE8-I22	22	63123-63143	GCGG-TCGCCGACGAGGTCGCG
		45551-45530	GCGGATCGCCGACGAGGTCGCG
GTE8-I23	21	44915-44935	TCCTCGACGCGCTCGGCGACG
		35271-35251	TCCTCGACGCGCTCGGCGACG
GTE8-I24	21	47368-47388	CTTCCAGATGGTGTGTGTCAC
		21377-21357	CTTCCAGATGGTGTGCGTCAC
GTE8-I25	21	42828-42847	TCGCGCTCGATCTC-CTGTCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		7701-7681	TCGCGCTCGATCTCGCTGTCG
GTE8-I26	21	22521-22541	GTCACCCAGAACACCCCGTCG
		14245-14225	GTCACCGAGAACACCCCGGCG
GTE8-I27	21	56230-56250	ACTCGCCACCGGACGACACCA
		36306-36286	ACCCGCCGCCGGACGACACCA
GTE8-I28	21	55279-55299	CTCGACGCCGGGTGGCGCAGC
		36355-36335	CTCGACGCCGGTTGGCACAGC
GTE8-I29	20	46950-46969	ACCGCGGCCGGACGGCCATC
		40772-40753	ACCGCGGCCGGACGGCCATC
GTE8-I30	19	29126-29144	GCAGCGGTCCCGGCCGACG
		18526-18508	GCCGCGGTCCCGGCCGACG
GTE8-I31	19	45893-45911	CCGCTCGTTCTTGTGGATC
		32529-32511	CCGCACGTTCTTGTGGATC
GTE8-I32	17	40587-40603	GGTCGGGGTACTTCGCG
		7856-7840	GGTCGGGGTACTTCGCG
GTE8-I33	17	45679-45695	CGTCGACCTTCAGCTTC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		18024-18008	CGTCGACCTTCAGCTTC
GTE8-I34	16	49028-49043	GTCGAGGATCTTCTTG
		21623-21608	GTCGAGGATCTTCTTG
GTE8-I35	16	51047-51062	AGGTCGCCGAGCGACT
		37352-37337	AGGTCGCCGAGCGACT
GTE8-I36	16	64179-64194	AAGACCTGACCCCGGT
		46098-46083	AAGACCTGACCCCGGT
GTE8-D1	141	29568-29707	GCGAGCTTCGAGGGCACCGGCTCGTTCTCGGCCGTCGCAGCGGCGCACCTGCTGGCGTCGATGGAGGGGCTCGGGTC CTTCGCCGCCAGTC-AGCTCGCGCACCTCTTCGCGAGCTTCACTGGGACTGGGGCCTTCTCGGC
		29352-29491	GCGAGCTTCGAGGGGACCGGCGACTTCCCGGCCTCCCTCGCTGCTCACCTGAACGCGACAGTCGAGGCGACCGGCTC CTTC-CCGGCAGCCGCGGTGGCCCACCTGCTCGCGACCCTCACTGCCACCGGCGCATTCTCGGC
GTE8-D2	98	19686-19777	GCCGGGGTCAAGGAGGCCGTCCAGGCGCTCGCGCCGATCATCCAGATCG-TCGGCAG--CGTCCTGC--- TCACCGTGCTCGGACCGGCGCTGACCGA
		19020-19117	GCCGGGGTCAACCAGCTCGACCAGGCGATGGCGAAGCTCTCGCCGAACGCGCAGCAGTTCGTCCGGCAGATCCACGC GCTCGGCCCGGCCTGGACCGA
GTE8-D3	44	49508-49551	GCCGCCGAGTTCCTTCAGCCGCTCCTTCATGACCTCGCGCTCGG
		38226-38269	GCCGCCGAGCCGGATCAGCCGCTCCATCATGAACACGCCCGCGG
GTE8-D4	43	29665-29707	CGCACCTCTTCGCGAGCTTCACTGGGACTGGGGCCTTCTCGGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		29341-29383	CGCAGCTCCTCGCGAGCTTCGAGGGGACCGGCGACTTCCCGGC
GTE8-D5	42	8499-8538	CGGACGTCGC--AGGACCACGACACCGACGGTGACAAGGTCG
		11327-11368	CGGACGCCGACAAGGACGACGACACCGACGCCGACAAGGACG
GTE8-D6	42	4707-4748	CTCGTCGACAACCTCTCGATGAAGCTCCAGGCGCTCCTGGGC
		1534-1575	CTCGCCGACGACGACCCGCTGATGTTCCAGGCGCTCCTGCGC
GTE8-D7	42	28331-28371	ACGCCACC-TCGGCCAGCTTCGCCGCGCTCATCGGCGCCTCG
		18340-18380	ACGCCGCCGTCGGCGGGCT-CACCGCGCTCATCGGTGCCGCG
GTE8-D8	40	58638-58675	CCACCGCGGCCGACGCTCTCGCGCTGCTCGAC--GCTGCT
		32728-32767	CCACCGCGGTCGGCGCTCTCGCGCTGATCGTCTGGCTGCT
GTE8-D9	36	64542-64574	CGGCGATGGA-GA--TGATCGCGAAGATCGAAGCCG
		62422-62457	CGGCGATGGATGACCTGCTCGCGCAGATCGAAGCCG
GTE8-D10	35	28974-29008	GCGCAACAGCGCTGACGGCACCTTCGTCGCTGGCG
		3671-3702	GCGCAACGGCG---ACGGCACCTTCGACGTCGGCG
GTE8-D11	33	61197-61229	AAACGTACTCTGACCTGCTGCTTAACGAGGCGT
		61132-61163	AAACATACTCTGACCTGC-GATTATCCAGGGGT
GTE8-D12	30	48894-48923	TCGTTGAACAGGCCCATCTTGCCGGTGCGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		42489-42518	TCGTTGAACAGGCTCATCTTCGCGGTGCGC
GTE8-D13	29	52306-52334	TCCGTATGGGCGGCACTATGGGCGGCACT
		52234-52262	TCGGTATGGGCGGCACTATGGGCGGCACT
GTE8-D14	29	17989-18017	CGAAGCTCGACGCCACCCGGAAGCTGAAG
		5439-5467	CGAAGAACGCCGCCGCTCGGAAGCTGAAG
GTE8-D15	29	40704-40732	CGAAGACCTCGGAGAAGTCGGCCAGCTCG
		27966-27994	CGACGTCCACGAAGAGGTCGGCCAGCTCG
GTE8-D16	27	39203-39229	AGGCCGACGTCGCCGTCGTCGATCGCG
		18016-18042	AGGTCGACGTCGCCGTCGAGGCTCGCG
GTE8-D17	27	27896-27922	TCATCCAGAACGTCATCGACACGATCC
		27098-27124	TCATCCAGAACACCATCGGCACCATCC
GTE8-D18	27	30016-30039	CGTCGACGATCACCG---CGTCGGTCG
		43822-43848	CGTCGACGATCACCGGCTCCTCGGTCG
GTE8-D19	26	41300-41325	GCGCCATCGAGCGCGGCGTCGCCCTC
		34292-34315	GCGCCA--GAGCGCGGCGTCGCCGTC
GTE8-D20	26	63707-63732	GGTCATGAAGGACGCCTACCTGCGCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		32373-32398	GGGCAAGAAGCTCGCCTACCTGCGCG
GTE8-D21	25	51831-51855	GCCGCCGTCGCCGTCGACGATCGCG
		39205-39229	GCCGACGTCGCCGTCGTCGATCGCG
GTE8-D22	24	31505-31528	AAGGTCGCCGACGACTTCCTCGGC
		20805-20828	AAGGCCGTCGACAACTTCCTCGGC
GTE8-D23	24	57260-57283	CGTCGAGCGGGAGCGTCGAGGGTC
		54834-54857	CGTCGAGCGGGACCGTTGCGGGTC
GTE8-D24	23	33046-33068	CCTGGACTACCTCACCGGCGATC
		17247-17269	CCTCGACTACTTCACCGGCGATC
GTE8-D25	23	58201-58223	CCGTCGGCCGCTGCGCATCGCGA
		18055-18076	CCGTCG-CCGCCGCGCATCGCGA
GTE8-D26	22	13368-13389	CACGCGATCGCGATGCCGGTCC
		4278-4299	CACGCGATCGCGATGTCGATCC
GTE8-D27	22	23668-23689	CCGGATCAAGGTGCGCAACTTC
		13115-13136	CCGGATCAAGGTGCGCGACATC
GTE8-D28	21	27568-27587	CGGC-TCCAGGTCGAGGTCGG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		104-124	CGGCATCCAGGTCGAGGTCGG
GTE8-D29	21	54652-54672	GACCGCGCCGGCCGGATCGCC
		153-173	GACCGGGCCCGCCGGATCGCC
GTE8-D30	21	13209-13229	GCGGTGGCGGGGTCGAGCGAG
		5001-5021	GCCGTGGCGGGGTCGGGCGAG
GTE8-D31	21	61622-61642	GGATCACGATCCCGGACGCCG
		7548-7568	GGCTCACGATCCCGGACGGCG
GTE8-D32	21	22446-22466	TTCGGCCGCGAGGACGAGGCG
		17970-17990	TTCGGCCGCGAGGCCCAGGCG
GTE8-D33	21	44914-44934	CTCCTCGACGCGCTCGGCGAC
		27673-27693	CTCGTCGACGCGCTCGGGGAC
GTE8-D34	21	32416-32436	TGGAACGAGCTGGTCTACGAC
		28126-28146	TGGAACGAGCTGTTCTACAAC
GTE8-D35	21	62916-62936	ATCGCGAAGATCTCGGCGAAG
		40277-40297	ATCCCGAAGATCTCGTCGAAG
GTE8-D36	20	41515-41534	GGTGGGGATGCCGTCGGCCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		14942-14961	GGCGGGGATGCCGTCGGCCG
GTE8-D37	20	57281-57300	GTCCCCGGCGCCGTCGGTGG
		28513-28532	GTCCCCGGCGCCATCGGTGG
GTE8-D38	19	54292-54310	CGACGACACCGGCGCCGAC
		11343-11361	CGACGACACCGACGCCGAC
GTE8-D39	19	50660-50678	CGTCGGCACCCAGAACACC
		22517-22535	CGTCGTCACCCAGAACACC
GTE8-D40	19	50926-50944	CTGGCGCAGCTCGGCGATC
		37551-37569	CTTGCGCAGCTCGGCGATC
GTE8-D41	18	12495-12512	TACGAGGCCGACACCGGC
		3435-3452	TACGAGGCCGACACCGGC
GTE8-D42	18	67239-67256	CGACCGAGACCATCCTGC
		21907-21924	CGACCGAGACCATCCTGC
GTE8-D43	17	62595-62611	AACCTGCTCACCGCGAC
		3906-3922	AACCTGCTCACCGCGAC
GTE8-D44	17	60395-60411	GCTGGTCATGGCCCTCC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		20192-20208	GCTGGTCATGGCCCTCC
GTE8-D45	17	65256-65272	GGATGCGCGACATGGTC
		21007-21023	GGATGCGCGACATGGTC
GTE8-D46	17	36142-36158	CGTCGTAGTCGTCGACG
		34837-34853	CGTCGTAGTCGTCGACG
GTE8-D47	16	29756-29771	CGTCGGCGCCTTCACC
		24983-24998	CGTCGGCGCCTTCACC
GTE8-D48	16	47727-47742	CGGTTGCCCGCGGTGA
		34998-35013	CGGTTGCCCGCGGTGA

I indicates inverted repeat, D indicates direct repeat.

Table 8.5 Repeat sequences identified in the genomes of phages GAL1, GMA1, and TPA4.

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GAL1-D1	70	41915-41984	CGACCAGTTGAAGGCGATCGGTAACGGGGTGTGCCCGCCGCAAGCGTTCCGGGCGTTGCAAATCCTGGAC
		31892-31961	CGATCAGCTGAAGGCCGTCGGCAACGGTGTGTGCCCGCCTCAGGCGTACCGGGCACTGGAACTGCTGCAC
GAL1-D2	41	17225-17265	GGCGACGACCTGGACGACTTCTGTTTGTGGGCGCGGGAGAT
		17003-17043	GGCGACGACATTGACGAGTTCTGCCAGTGGGCGCGGGACAT
GAL1-D3	31	28065-28095	CTGGGCTAAACGTAAGTACAAACGTAAGTGA
		27866-27896	CTGCGATAGACGTAAGTCTAAACGTAAGTGA
GAL1-D4	31	41061-41091	CACCGTCCGCAACGCCCTCTGGTCCAGCGGC
		35426-35456	CACCCTACGCATCGCCCTCTGGTGCGGAGGC
GAL1-D5	28	38946-38973	CACCGGGGCGGTGTACTGACATGAGCGA
		38250-38277	CATCGGGGCGGTGTACTGACATGGGCGA
GAL1-D6	28	26958-26985	GGTCGTTGCTCGCCACCACCGACGGCAC
		25629-25656	GGACGTTGTACGGCTCCACCGACGGCAC
GAL1-D7	26	39440-39465	GGTGTCCGGTGACCGCCGATCCCGCC
		38918-38943	GGTGTCCGGTGACCGCCGATCCCGCC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GAL1-D8	25	6568-6592	AGGTCACCGCGACCGGCCTGTTCGG
		6242-6265	AGGGCACCT-GACCGGCCTGTTCGG
GAL1-D9	25	26561-26584	GCCCGCAATCTCAACCCCG-GTGGT
		25562-25586	GCCCGCAATCTCATCCACGAGTGGT
GAL1-D10	21	35271-35291	CATCATCGACGGAGAGACGTG
		32441-32461	CATCATCGACGGCGAGCCGTG
GAL1-D11	21	37771-37791	GGCGTACGCGAAGGGCTGGTT
		37280-37300	GGCGAAGGCGAAGGGCTGGTT
GAL1-D12	21	47370-47390	ACCGACCACATCACGATCGCG
		44205-44225	ACCGACCACATCAGCATCGCG
GAL1-D13	20	19889-19908	CGGCCTGTTCGAGGCGTACG
		6581-6600	CGGCCTGTTCGGGCCGTACG
GAL1-D14	20	30736-30755	GCCACAACCTGGACGACTTC
		17226-17245	GCGACGACCTGGACGACTTC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GAL1-D15	20	46932-46951	TCCCCGTCGTCATCCGCGAC
		24049-24068	TCCCCGTCGTCATCGTCGAC
GAL1-D16	20	43974-43993	CATCCGGCCGCTGACAGCCC
		26659-26678	CATCCGGCCGCTCACCGCCC
GAL1-D17	20	47822-47841	GCTATGACCACCACCGACCT
		39848-39867	GCTATGACCACCCCCAACCT
GAL1-D18	19	37488-37506	CGGCATGGTCGGCGGTGTC
		12354-12372	CGGCATCGTCGGCGGTGTC
GAL1-D19	19	42335-42353	GCTGGATCGGGTTCGACAC
		38786-38804	GCTGGATCGGGTTCCACAC
GAL1-D20	18	13705-13722	CCGGTCCTCGGGCAGCTG
		13393-13410	CCGTTCCTCGGGCAGCTG
GAL1-D21	18	15844-15861	GGCGAGTACGTGGTCAAC
		14512-14529	GGCGAGTACGTCGTCAAC
GAL1-D22	17	40090-40106	ATGACCACCCCCAACCT
		39851-39867	ATGACCACCCCCAACCT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GAL1-D23	16	9601-9616	GGAAACCCAGCAGCAG
		7357-7372	GGAAACCCAGCAGCAG
GAL1-D24	16	23976-23991	GCGCAGCGTGACGCCA
		12922-12937	GCGCAGCGTGACGCCA
GAL1-D25	16	32437-32452	CCGTCATCATCGACGG
		16894-16909	CCGTCATCATCGACGG
GAL1-D26	16	47860-47875	CGTGTCTGCCGCCTCG
		34060-34075	CGTGTCTGCCGCCTCG
GAL1-D27	16	43540-43555	CCGCATCGACGACCCC
		37325-37340	CCGCATCGACGACCCC
GAL1-D28	15	47445-47459	GCTGAGGCGATCGCC
		2358-2372	GCTGAGGCGATCGCC
GAL1-D29	15	33341-33355	CTCGAGATCGGCCCG
		10058-10072	CTCGAGATCGGCCCG
GAL1-D30	15	36652-36666	AGCACGGCGTCCTCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		20152-20166	AGCACGGCGTCCTCG
GAL1-I1	44	12142-12184	GGTCTGGATCGGCG-CATCAACATCAACGTCAACACCAGCGGCC
		5004-4961	GGTCGAGATCGGCCACACCATCGTCGACCTCAACACCAGCGGCC
GAL1-I2	43	30221-30260	GATGACGTTCTGCGCG---TCAGGGCGGTTCCTGCCGAGCTTC
		10950-10908	GATGACGTACGGCTCGAGTTCGACGCGGTTCTTGCCGAGCTTC
GAL1-I3	28	35789-35816	CCCATCCCCGGGCCGACGAGCTTCACCC
		13349-13323	CCCATCCCCGGGCCGGCG-GCGGCACCC
GAL1-I4	26	32493-32518	GGATACCGCGACGCGAGCAACGCAGC
		21565-21540	GGATACCGCGACGCGTGGACAGCAGC
GAL1-I5	23	37176-37198	GCGGGCCGGGTGGAGAACATCAC
		35040-35018	GCGTGGCGGGTGGAGAACGTCAC
GAL1-I6	20	40924-40942	GGTCG-CGCACCAACTCCTC
		20658-20639	GGTCGGCGCACCAACTCCTC
GAL1-I7	19	23906-23924	GATGCGCCCGTACTCGCCC
		7096-7078	GATGCGCTCGTACTCGCCC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GAL1-I8	18	18606-18623	TGTCGTTGCCGTTGGCGA
		7397-7380	TGCCGTTGCCGTTGGCGA
GAL1-I9	18	43961-43978	CGAAACAGGCCCGCATCC
		39435-39418	CGAAGCAGGCCCGCATCC
GAL1-I10	16	25447-25462	CACCCACGTCGACGGG
		17558-17543	CACCCACGTCGACGGG
GAL1-I11	16	41609-41624	CAAGGGTGACCGTCCG
		33577-33562	CAAGGGTGACCGTCCG
GAL1-I12	16	84-99	GGCACTCAGTGCCGGT
		49940-49925	GGCACTCAGTGCCGGT
GAL1-I13	15	7492-7506	GCGGCCGGCGGATCG
		680-666	GCGGCCGGCGGATCG
GAL1-I14	15	12653-12667	CAGCTGCAGCCGCAG
		5860-5846	CAGCTGCAGCCGCAG
GAL1-I15	15	8831-8845	AGGGTGACCGAGGTG
		6093-6079	AGGGTGACCGAGGTG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GAL1-I16	15	9271-9285	CGACATCGGTGATCT
		8717-8703	CGACATCGGTGATCT
GAL1-I17	15	49695-49709	CACATCCACATCAGC
		12726-12712	CACATCCACATCAGC
GAL1-I18	15	21868-21882	TCGACAGCATCGCCG
		14509-14495	TCGACAGCATCGCCG
GAL1-I19	15	45708-45722	ACCCCGGCGGCTACG
		29106-29092	ACCCCGGCGGCTACG
TPA4-D1	193	15953-16143	CGAAGCAGATCCGCGACGCCGAGCAGAAGGTCGCCGACAAGGAGGCCGCCGCGCGGCGCGCGCACCCAGCTCGCGGAGG CGCGGAACAACCCGAAGGCGAAGAAGTCGATGATC-- CAGGCCGCCGAGGACCGGCTCACCGTCGCCGAGCGCGAGGCAGCCGACGCCAAGACCGACCTCGAGACGCTGAAG
		15719-15909	CGAAGAAGCTCCGCGACGCGCAGCAGAAGGTCACCGACAAGGAGGCCGCCGCGGCACTCGCGCAGACCAAGCTCAACGAGA CCCTGAACAACCCGAAGGCCAAGGAGTCG-- GCTCGGCAGGCCGCCCGCGACCGCCTCACCATCGCGCAGCGCGAGGCCGCCGACGCGAAGACCGACCTTGAGACGCTCAAG
TPA4-D2	85	24402-24482	CGGCGGTGATGCCTGGGGCAGCAGCGGGGGTTCCGCCCGCCGTG---GT- CGCGGGGTCGGTGCTGCAGCCGCCGCCGATGACGG
		24087-24170	CGGCGCGGATGCCCGCAGCCGCCGAGGTGTATCCGCCGACGGTGCGGGTGCTCGTGG- CGGAGGTCCTGCCGCCGCCGATGACGG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D3	66	5650-5715	TGCAGAAGCAGCTCGACGACGCCGTCGCCGCGCAGAAGAAGTCCGACGAGGCCGCCGCGACGGCGA
		5557-5610	TGCAGAAGCAGGTCGACG-CGC--TCACC---------AAGGCAGCCGAGGACGCCGCGAAGGCGA
TPA4-D4	63	15637-15695	GGCAGCGGCTCGGGCTCGGGCTCG-GGCAGTGG--CGGCGGTAGCGGGTCCGG-TTCGTCGGC
		15541-15603	GGCAGCAGCTCGGGCTCGGGGTCGTCGAAGTCGTACGGCGGCGGCGGGTTCGGCAGCGTCGGC
TPA4-D5	51	53573-53623	AGCGGATCGCCGGCACCGTCACCGCCGTCGCGCTCAACGCCGACGGCAGCC
		12800-12849	AGCGAATCACCGCGGCCACCACCGCCGTCGCCGCGAACGCCG-CGGCGGCC
TPA4-D6	49	44595-44642	GGACCGCACGAG-GCGCAGGTCGACGCCCTCGTCAACAAGTTCAGCGGC
		15192-15237	GGACTGCTCGGGCGCGCAGG-CGGCGA--TCGTCAACAAGATCACCGGC
TPA4-D7	47	42031-42074	ATCGGCGACCGGATGCCGCGCATC-ACCGGC--GCCATGCCGGCGGC
		16426-16472	ATCGGCGACTTCATGAAGCCGATCGACCGTCCGGCCATGCCGGCGGC
TPA4-D8	47	43938-43983	CGAGGTGCCCGATCAGGTGGAGCT-GCTCCGCCTGGCGCAGGCCGAG
		20066-20111	CGAGGTGCCCGACCCGGACGAGCTCGACCTGCCTGAC-CCGGCCGAG
TPA4-D9	45	13067-13111	TCGGTCAGCGTGTGCCGGTCATCGCGTCGATGCAGCAGGCGTTCG
		1709-1753	TCGCTCCGCGCGAGAAGGTCGACGCGACGGTGCGGCAGGCGTTCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D10	44	37853-37896	GCGAGGAGGTGTCCGCGACGGCGGGCCGGACCTCGCAGATCGAG
		13508-13548	GCGACGAGGG--CCGCGAAGGCAC-CCGGAACGCGCAGATCGAG
TPA4-D11	43	38973-39012	CCACCCCACCACC---AGGAGACCCGCGATGGCCGACTACGAG
		36901-36940	CCACCCCACCACCGAAAGGA-ACC--CCATGTCCGACAACGAG
TPA4-D12	42	40327-40368	GCCCTGCTCCGCGGCGAACGCGACGAAGCCCGCCGGCAGCTC
		29231-29271	GCCGTGCCGCGCGCCGACCG-GTCGATGCCCGCCGGCACCTC
TPA4-D13	41	43785-43825	GCACAAGGCCGCGCTCGTGCAGGCGATCGACGGGAAGGTCA
		9507-9547	GCACTACGCCTACCTCGTCGAGACCTTCGACGGGAAGGTCA
TPA4-D14	40	43663-43699	CTGACCGAGA---TCCGCGCCGACCTCGACGACCTGCACG
		40791-40830	CTGCCCGAGAAGTTCCGCGCCGACCACGACGAGCGGTACG
TPA4-D15	39	48273-48310	CATCGCCGAGCTCCGCG-GCGGCGGCTCGACGGCGCGCC
		35757-35794	CATCACCGAGCTCCGCGAGCAGCGCGACGA-GGCCCGCC
TPA4-D16	38	54100-54137	CACCGTCCGCCTCGTCGACTTCGCCGGCCGCCTCCTCC
		53382-53419	CACCGTCCGCCTCATCGCCCCCGACGGCCGCCCCATCC
TPA4-D17	38	52704-52741	CACCGACCAGACCGACCGCGCCATCGCCGCGATCGACG
		1965-2002	CACCGAGCAGGCCGAGCTCACCGCCCGCGCGATCGACG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D18	38	45281-45318	TGACCGCCAGCGTCACCAGCCTCATCGCGCACCGCGCC
		3745-3782	TGATCGACGGCCTCGCCCACCTCATCGCGCAGCTCGCC
TPA4-D19	38	47672-47709	CGCAGCACGGCGTCGTGCGCGCGATGCAGGCGTTCAAG
		29355-29392	CGAAGGACGCCGCCCCGGGCGCGGTGCAGGCGTTCGAG
TPA4-D20	37	46738-46771	CCGCGCCCGCCGCGTCC---TCCTCGGCGAGACCACC
		15000-15036	CCGCGCCCGTTCCGTCGCGATCCTCGGCGAGACCGCC
TPA4-D21	37	41653-41689	GCCATCGCCGCCGCGCAGGCGCACGGGATCCTCGCGC
		30287-30320	GCCATCGCCGCCGCGATCGCTCTCG---TCCTCGCGC
TPA4-D22	37	54589-54623	ATCATCCAGGCCGCCAAGTACCTCCGCACC--CGCCG
		16063-16099	ATGATCCAGGCCGCCGAGGACCGGCTCACCGTCGCCG
TPA4-D23	36	19434-19465	TCGTCGGCGGCGCCGC-GGTGGCC---GCCTGACCC
		16382-16417	TCGTCGGCGGCGCGGCGGGCGGCCTCGGCCTGACCC
TPA4-D24	36	21380-21415	CCGTCGTCGAGTGGCGCGACATGGGCGAGCTCGACC
		20059-20094	CCGCCGGCGAGGTGCCCGACCCGGACGAGCTCGACC
TPA4-D25	35	16810-16843	GGAACCG-CGGCGTCGTCCTCGACGAGGGACTCGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		2466-2500	GGCACCGTCGTCGACGTCCTCGACGAGCGTCTCGA
TPA4-D26	35	8588-8622	CGCGCCGCCGCGCACGAGGCGAAGACCCACGCGCT
		22827-22861	CGCGCCGCCGCCGTCGAGGCGCGGATCCGCGCGCT
TPA4-D27	35	41640-41674	GGCGCCTGACCCGGCCATCGCCGCCGCGCAGGCGC
		39297-39331	GGCGCTGGACGCGGAGATCGCCGCCGCGGCGGCGC
TPA4-D28	35	51511-51545	CGCCGAGGGGCCGGCCGGTGCACGGGAGCTGCGGA
		49933-49967	CGGCGAGGCGCTCGCCGCCGAACGGGAGCTGCAGA
TPA4-D29	33	22478-22509	CCTCGCCGACGGCACCGTCCGCTTCGAC-CGGA
		17420-17452	CCGCGCCGACGGCATGGTCCTCGTCGACACGGA
TPA4-D30	33	14904-14936	CCGCCGCGCCGACATCTACGCGGGCCGCGGCGC
		11538-11570	CCGCCGCGCCGTCAACTCGGCAGAGCGCGGCGC
TPA4-D31	32	50766-50797	TCGACGCCCTCGCCGCCCTGTCCGCCCGCTCG
		21874-21905	TCGACGCCCTCGCCAGCCTGTTCGGCGGCGCG
TPA4-D32	32	42155-42186	TGTTCGCGCTCGCCGACGAACTCGGTGTCAAG
		36724-36755	TGATCGTGCTCGCCGACGTTCTCGGCGTCGAG
TPA4-D33	32	2932-2963	TCACCCGCGAGATCGCCTACCTGACCGACGCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		52336-52367	TCGCCCGCGCCATCGCCTACCTCGCCGACCCG
TPA4-D34	31	34598-34626	CCGCGCCCGCCGCGGCGACGC--CCGGCCTC
		6666-6696	CCGCGCGCGGGGCGGCGACGCCGCCGGCCTC
TPA4-D35	31	45367-45397	GCCGGCACGGACACCGTCGGCGACGTCGTGC
		4132-4157	GCCGGCACG-----CGTCGGCGACGGCGTGC
TPA4-D36	31	50747-50777	CTCGTCAGCATCAACGAGGTCGACGCCCTCG
		8390-8420	CTCCTCAACAGCGACGGCGTCGACGCCCTCG
TPA4-D37	31	48345-48375	CGGCAACCACCACGCCCTCGTCATGCGGAAC
		47514-47544	CGGCTACCGGCACCGCCTCGTCATGCTGAAC
TPA4-D38	30	51887-51915	CAGAGG-TTCAGAGGGCGACCCTCGGAACC
		51712-51741	CAGAGGGTTCCGAGGGCGACCCTCGGAACC
TPA4-D39	30	53964-53993	ATGCACGGAACCGAGGAGACCCGATGACCG
		53672-53700	ATGC-CGAACCCGAGGAGACCCGATGACCG
TPA4-D40	30	37039-37068	CGCCGAGTGCGTCGACGCGGCCCTCGCCAC
		7059-7088	CGCCGACGGTGTGAACGCGGCCCTCGCCAC
TPA4-D41	29	54870-54898	CCGCCTGTTCACCGCAGCCGAACTCGTCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		45733-45761	CCGCCTGTTCACCCCGGACGAACTCGTCG
TPA4-D42	29	37560-37588	CGTGATCTACGACGAGCTCGGCGCCGACG
		14355-14383	CGCGATCACCGACGAGCTCGGCCCGGACG
TPA4-D43	29	44614-44642	TCGACGCCCTCGTCAACAAGTTCAGCGGC
		21874-21902	TCGACGCCCTCGCCAGCCTGTTCGGCGGC
TPA4-D44	29	48083-48111	GCCGCGACGGCAAGCAGGCGCAGGCCGCG
		29016-29044	GCCGCGACGTCGCGGCGGCGCAGGCCGCG
TPA4-D45	29	35954-35982	CCGAGAAGGACGCCGCCTACCGCACGCGG
		29352-29378	CCGCGAAGGACGCCGCC--CCGGGCGCGG
TPA4-D46	29	43702-43728	CCGAA--CGACGACACCGCCGACCAGGTC
		36124-36152	CCGAAGCCGACCGCATCGCCGACCAGGTC
TPA4-D47	28	24628-24655	GCCCGGTCTGCCGTACACGCTGCCGTTC
		24238-24265	GCACGGGCTGCCGTACACGCTGCCGTTC
TPA4-D48	28	7774-7801	TCCGACGAGATCCCCGACCCCGCGGTCA
		1432-1458	TCCGACGAGATCCTCGA-CCCGCGGACA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D49	28	4350-4377	ATCGCGAACGCGATGGACAAGGCCGAGG
		958-985	ATCGCGAACGCGATGTTCAGTGGCGAGG
TPA4-D50	28	35004-35031	GTCGTCGCCGACGACCGCAGGCGCGTCG
		5006-5033	GTGGTCGCCGACGACAGCTGGGACGTCG
TPA4-D51	28	43704-43731	GAACGACGACACCGCCGACCAGGTCTTC
		23659-23686	GAACGACGACAAGGGCAACCAGGTCATC
TPA4-D52	27	54549-54574	CACGACCGCCCGGAAGC-TCCCGCACC
		39619-39645	CATGACCGCCCGGAAGCGGCCCGCGCC
TPA4-D53	26	55013-55038	GCGGCGACGTCCTCGACGCCCTCGCC
		21862-21887	GCCGCGGCGTGCTCGACGCCCTCGCC
TPA4-D54	26	47314-47339	CGCGCCCTGCGCGGCGACGCCCAGCC
		34599-34624	CGCGCCCGCCGCGGCGACGCCCGGCC
TPA4-D55	26	22403-22428	GCAGCAGCTCACCGGCGAGTGGACCG
		7839-7864	GCAGCAGCTCAACGGCGACATGACCG
TPA4-D56	26	27891-27916	CGACGGCGCGTACGCCGCCGCGCGCG
		8578-8603	CGACGGGGCGCGCGCCGCCGCGCACG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D57	26	35422-35447	GACGTGGCGTGAGGCGCCGCGCACGA
		8579-8604	GACGGGGCGCGCGCCGCCGCGCACGA
TPA4-D58	26	42954-42979	CTTCACCGGCGCCCGCACGCCCGCCG
		34309-34334	CTTCACCGGCGCCCACCTGACCGCCG
TPA4-D59	26	45530-45555	CATCGCCGCAGCCGTCGCCGCGGTGG
		34560-34585	CACCGTCGCGGCCGTCGCCGCAGTGG
TPA4-D60	26	40280-40305	GGATCGCGCAGCTCCGCGCCACCGCC
		40008-40033	GGATCGCGCAGCTCTGGACCACCACC
TPA4-D61	25	38560-38583	CCTCGCAGGCGGCGAACGC-GCTCC
		7752-7776	CCTCGCAGGCGCCGAACGCGGCTCC
TPA4-D62	25	34591-34613	TGCAGC--GCCGCGCCCGCCGCGGC
		38623-38647	TGCAGCAGGCCCCGCCCGCCGCGGC
TPA4-D63	25	9438-9462	CAACGTCACCCCGGCGACCGCGGCA
		1675-1699	CAGCGGCCCCACGGCGACCGCGGCA
TPA4-D64	25	35216-35240	CGCAGAGTCGGTGGCGGCATGAGCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		23589-23612	CGCAGTGGCG-TGGCGGCATGAGCG
TPA4-D65	25	30158-30182	GCGGCACCGCCGCGGCGGCACCGAG
		28110-28134	GCGGCACCGCCTCGGCTGGGCCGAG
TPA4-D66	25	51502-51526	CGCCGTGCTCGCCGAGGGGCCGGCC
		30002-30026	CGCCGTGCTCGCCGCCGGCCAGGCC
TPA4-D67	25	44115-44139	CGCCGAGGCGCGCCGGCAGCTCGCC
		40347-40371	CGACGAAGCCCGCCGGCAGCTCACC
TPA4-D68	25	53578-53602	ATCGCCGGCACCGTCACCGCCGTCG
		45364-45388	ATCGCCGGCACGGACACCGTCGGCG
TPA4-D69	24	42774-42797	GGCAGATCACCCTCACCGAGACCA
		40361-40384	GGCAGCTCACCATCACCGAGACCA
TPA4-D70	24	13848-13871	CAAGACCGCCCTCGACGCGCTGAA
		13575-13598	CAAGACCGCCCTCGACAACCTGAA
TPA4-D71	24	21283-21306	CCGATCCTCCGCAGCGAGTGGACC
		16626-16649	CCGATGAGCCGCAGCGAGTGGACC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D72	24	54856-54879	GACGCCCTCGTCGACCGCCTGTTC
		17382-17405	GACGGCCTCGTCGACCGCATGGTC
TPA4-D73	24	18479-18502	GAGGACACCGGCGTCGGCCTGATC
		17933-17956	GACGACACCGGCGTCGGCACGATC
TPA4-D74	24	37043-37066	GAGTGCGTCGACGCGGCCCTCGCC
		36034-36057	GAGTGCGTCGACGCCTACCTCGCC
TPA4-D75	24	52709-52732	ACCAGACCGACCGCGCCATCGCCG
		49041-49064	ACCAGACCGACAGCCCCACCGCCG
TPA4-D76	23	14860-14882	GTCCCGATGCTCGCGGGCGGCCT
		13252-13274	GTCGCGATGGTCGCGGGCGGCCT
TPA4-D77	23	14717-14739	CGAAGGCCGCGGTGGACGCGATC
		193-215	CGAAGGCCGCGGTCCGCGCGATC
TPA4-D78	23	20039-20061	GCTCCCGGAGCCGATCACCGCCG
		2429-2451	GCTCCCGGAGCTGATCCGCGCCG
TPA4-D79	23	22990-23012	CCGTGTGGCAGCAGTTCGGCGGC
		22396-22418	CCGTGTGGCAGCAGCTCACCGGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D80	23	38460-38482	CGGCTGACCATCGCCGTCGTGGC
		28796-28818	CGGCTGACCATCGGCATCGAGGC
TPA4-D81	23	45933-45955	CCACCGACCCCGCGGTGCTCGAC
		42086-42108	CCACCGACCCCGAGATCCTCGAC
TPA4-D82	23	55318-55340	TCGCCGACCTCGGCGACTGCGGC
		43128-43150	TCCCCGACATCTGCGACTGCGGC
TPA4-D83	23	52509-52531	TCGTCGCCCAGCGCACCCGCCTC
		49878-49900	TCGTCGCCCAGGCCACCCGCATC
TPA4-D84	22	6068-6089	CACCGAGACGCCGAAGTCGGCG
		4937-4958	CACCTCGACGCCGAAGTCGGCG
TPA4-D85	22	23554-23575	CCGCGTCCTGAAGGAGCTCGCG
		12066-12087	CCGCGGGCTGAAGGAGCTCGCG
TPA4-D86	22	49325-49346	GGCACCGGCAAGTCCGGCGACG
		16150-16171	GGCACCACCAAGTCCGGCGACG
TPA4-D87	22	49201-49222	CGCCCTCGCCGACGGCGACGTC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		22475-22496	CGCCCTCGCCGACGGCACCGTC
TPA4-D88	22	39205-39226	GGCGTCGTCGCCCGCCACCCCG
		28610-28631	GGCGTCGTCGCCCACCACACCG
TPA4-D89	22	44240-44261	CGGCGCGGTCCGCGTCGCCGTC
		37123-37144	CGGCGCGGTCCGCCTCGCCATC
TPA4-D90	21	12412-12432	ATCGACGGCCTCGGCCACCTC
		3747-3767	ATCGACGGCCTCGCCCACCTC
TPA4-D91	21	28895-28915	CGCGGCGTCGCGGCGATCCTC
		21271-21291	CGCGGCGTCGCGCCGATCCTC
TPA4-D92	21	55016-55036	GCGACGTCCTCGACGCCCTCG
		8400-8420	GCGACGGCGTCGACGCCCTCG
TPA4-D93	21	23132-23152	CGTCGGGGCGCGCGCCGACGC
		8578-8598	CGACGGGGCGCGCGCCGCCGC
TPA4-D94	21	54491-54511	TCGCCTCCGACGTCGCCACCG
		13046-13066	TCGGCTCCGCCGTCGCCACCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D95	21	22947-22967	GGCGGCGGCGCCGGCACCCTC
		14926-14946	GGCCGCGGCGCCGGCACCGTC
TPA4-D96	21	51153-51173	CCGCCGCCGGTGACGAGATCG
		23260-23280	CCGCCGCCGGTGACGTGTTCG
TPA4-D97	21	54423-54443	CGGGCCCCGCGGCCGGCGCGC
		27504-27524	CGTGCCCCGCGGCCAGCGCGC
TPA4-D98	21	53046-53066	CGACCACGACGACGCCCGCGA
		50263-50283	CGACCACGACGACGACCACGA
TPA4-D99	20	37577-37596	TCGGCGCCGACGCGACGACC
		6325-6344	TCGCCGCCGACGCGACGACC
TPA4-D100	20	36204-36223	GAGCTCGGCCCGTCCGAGGT
		21406-21425	GAGCTCGACCCGTCCGAGGT
TPA4-D101	20	29484-29503	GCGGCCCGCTCGGCTACCCG
		29310-29329	GCGGCCCGCTCGGCTTCCCG
TPA4-D102	20	50026-50044	GGTCGA-GGCCGCGGTCCGC
		190-209	GGTCGAAGGCCGCGGTCCGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D103	20	34359-34378	TCGTCCTCGTCCGCGCGTAC
		3211-3230	TCGACCTTGTCCGCGCGTAC
TPA4-D104	20	8691-8710	CGCCGAGCTCGGCGTCCCCG
		3782-3801	CGCCCAGCTCGGCGTCCGCG
TPA4-D105	20	16940-16959	CCAGCTCGGCGTCCACATCG
		3785-3804	CCAGCTCGGCGTCCGCGTCG
TPA4-D106	20	20113-20132	TCGACGACCTCATCCCGTGG
		6899-6918	TCGACGACCTCATCACCTGG
TPA4-D107	20	29700-29719	GGCGCGCATCGCCGCGCACG
		8584-8603	GGCGCGCGCCGCCGCGCACG
TPA4-D108	20	41211-41230	GCACCGCGCTCGGCGAGACC
		11197-11216	GCACCGCGCTCCGCGAGGCC
TPA4-D109	20	29380-29399	GCAGGCGTTCGAGGGCGGCG
		13101-13120	GCAGGCGTTCGTGCGCGGCG
TPA4-D110	20	46297-46316	GACACCGTCGCCGGGCAGCT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		13651-13670	GACACCGTCGCCGCGAAGCT
TPA4-D111	20	52212-52231	AAGGCCCGCGCGCACACCGG
		14176-14195	AAGGCCCGCGCGCAGATCGG
TPA4-D112	20	53580-53599	CGCCGGCACCGTCACCGCCG
		14934-14953	CGCCGGCACCGTCTTCGCCG
TPA4-D113	20	27912-27931	GCGCGGCCGGTTCGTCGGCC
		15677-15696	GCGGGTCCGGTTCGTCGGCC
TPA4-D114	20	55013-55032	GCGGCGACGTCCTCGACGCC
		22768-22787	GCGACGACGTCCTCGGCGCC
TPA4-D115	20	36966-36985	AGTTCACCGACTACGCGGTC
		26524-26543	AGATCGCCGACTACGCGGTC
TPA4-D116	20	34174-34193	TCGGCCTCGGCAAAAGCCAG
		27926-27945	TCGGCCTCGGCTACAGCCAG
TPA4-D117	20	29401-29420	GCTGTACCGCCGCGACGGCG
		29008-29027	GCTGTTCCGCCGCGACGTCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D118	20	43674-43693	CCGCGCCGACCTCGACGACC
		43565-43584	CCGCCGCGACCTCGACGACC
TPA4-D119	20	54515-54534	TCGCCGGCACCCTCAACGCC
		53579-53598	TCGCCGGCACCGTCACCGCC
TPA4-D120	19	12486-12504	GGGCAGCAACGGCCTGCTC
		4700-4718	GGGCAGCAACGTCCTGCTC
TPA4-D121	19	44613-44631	GTCGACGCCCTCGTCAACA
		8408-8426	GTCGACGCCCTCGTGAACA
TPA4-D122	19	42873-42891	GAAGATCGGCAACGTCGTC
		14388-14406	GAAGACCGGCAACGTCGTC
TPA4-D123	19	37315-37333	GAGCCGATCACCGGCGCCG
		20046-20064	GAGCCGATCACCGCCGCCG
TPA4-D124	19	45318-45336	CCTCACCCAAGCCGACAAG
		20416-20434	CCTCACCGAAGCCGACAAG
TPA4-D125	19	30162-30180	CACCGCCGCGGCGGCACCG
		29291-29309	CAGCGCCGCGGCGGCACCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D126	19	32243-32261	GGCCTCGCCGATGCGGATG
		31928-31946	GGCCTCGCCGATGCGGGTG
TPA4-D127	19	43571-43589	CGACCTCGACGACCCCGCG
		35811-35829	CGACCTCGACGACCCTGCG
TPA4-D128	19	50085-50103	TCGCGGAGGAGCTCGGCGA
		42590-42608	TCGCGGAGGAGCACGGCGA
TPA4-D129	19	45051-45069	ACGTCAACGCGATCGGCGA
		44293-44311	ACGGCAACGCGATCGGCGA
TPA4-D130	19	53852-53870	CGAGCAGCGCGCCCCGGGC
		53013-53031	CGAGCAGCGCGCCCGGGGC
TPA4-D131	18	29407-29424	CCGCCGCGACGGCGACCC
		5701-5718	CCGCCGCGACGGCGACCC
TPA4-D132	18	17234-17251	GCGGATGCTCACCCGCGA
		2882-2899	GCGGATGCTCAACCGCGA
TPA4-D133	18	23340-23357	GGCGCGGTGATCCTCGAC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		4865-4882	GGCGCAGTGATCCTCGAC
TPA4-D134	18	15874-15891	GCCGCCGACGCGAAGACC
		6327-6344	GCCGCCGACGCGACGACC
TPA4-D135	18	23026-23043	GCATCGCCCGGTACACCG
		8804-8821	GCATCGCCCGGCACACCG
TPA4-D136	18	28217-28234	GCCTCGCCGACCCGATCG
		9584-9601	GCCTCGGCGACCCGATCG
TPA4-D137	18	49670-49687	GGCGCCCCGGTCACCGTC
		10355-10372	GGCGCCGCGGTCACCGTC
TPA4-D138	18	19849-19866	GCGGCCCGGGTGGCGAGC
		15350-15367	GCGGCCCGGGTGGCGGGC
TPA4-D139	18	37581-37598	CGCCGACGCGACGACCGA
		15876-15893	CGCCGACGCGAAGACCGA
TPA4-D140	18	55013-55030	GCGGCGACGTCCTCGACG
		16816-16833	GCGGCGTCGTCCTCGACG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D141	18	27573-27590	TCGATCACCGCGCACGCC
		20540-20557	TCGATCACCGCGTACGCC
TPA4-D142	18	40750-40767	TCACCGCGTACGTCGAGC
		20544-20561	TCACCGCGTACGCCGAGC
TPA4-D143	18	42102-42119	CCTCGACGGCATCGACGA
		22610-22627	CCTCGACGGCATCGTCGA
TPA4-D144	18	50440-50457	CCCGTGGGCCGTCCTCGG
		25065-25082	CCGGTGGGCCGTCCTCGG
TPA4-D145	18	53769-53786	CCGATCGCCCGCGGCGAC
		34741-34758	CCGATCACCCGCGGCGAC
TPA4-D146	18	50931-50948	CCCGCGCCGGTGCGCTCC
		39911-39928	CCCGTGCCGGTGCGCTCC
TPA4-D147	18	52403-52420	CCCCACCCACCAGGAGAA
		41950-41967	CCCCAACCACCAGGAGAA
TPA4-D148	18	54918-54935	CGGCCAGCGCATCACCGC
		45521-45538	CGGCCAGCGCATCGCCGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D149	18	52762-52779	ATCGACGACATCGTCGAC
		49367-49384	ATCGACGACATCGCCGAC
TPA4-D150	17	20988-21004	GCCTCGGCGACCCGATC
		9584-9600	GCCTCGGCGACCCGATC
TPA4-D151	17	50160-50176	TCGGCCGGCCCGCCGCG
		21452-21468	TCGGCCGGCCCGCCGCG
TPA4-D152	16	40284-40299	CGCGCAGCTCCGCGCC
		2318-2333	CGCGCAGCTCCGCGCC
TPA4-D153	16	45010-45025	GACCGACCTCGAGGAG
		7614-7629	GACCGACCTCGAGGAG
TPA4-D154	16	38735-38750	GCAGGCCGCGCTGCTC
		11796-11811	GCAGGCCGCGCTGCTC
TPA4-D155	16	25110-25125	CCGCGTCGGCGGCGAC
		12165-12180	CCGCGTCGGCGGCGAC
TPA4-D156	16	23171-23186	GCGCGTGCGGATCGGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		17117-17132	GCGCGTGCGGATCGGC
TPA4-D157	16	32032-32047	CGCGGGCGCGCAGCTG
		22304-22319	CGCGGGCGCGCAGCTG
TPA4-D158	16	36772-36787	CCGCCGCCGTCGAGGC
		22831-22846	CCGCCGCCGTCGAGGC
TPA4-D159	16	37270-37285	AACCCCATGTCCGACA
		36920-36935	AACCCCATGTCCGACA
TPA4-D160	16	54915-54930	CCTCGGCCAGCGCATC
		47835-47850	CCTCGGCCAGCGCATC
TPA4-D161	16	55025-55040	TCGACGCCCTCGCCGC
		50766-50781	TCGACGCCCTCGCCGC
TPA4-D162	15	11561-11575	AGCGCGGCGCGAAGG
		776-790	AGCGCGGCGCGAAGG
TPA4-D163	15	23900-23914	GCGGCGGCCGGGATG
		3246-3260	GCGGCGGCCGGGATG
TPA4-D164	15	35311-35325	CCGTCTCGAGAAGGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		6750-6764	CCGTCTCGAGAAGGC
TPA4-D165	15	26050-26064	GCCAGGGCGACGACC
		10196-10210	GCCAGGGCGACGACC
TPA4-D166	15	18887-18901	CCGACGATCCTGAAC
		12637-12651	CCGACGATCCTGAAC
TPA4-D167	15	42856-42870	GGTGACCGCATCGCC
		13432-13446	GGTGACCGCATCGCC
TPA4-D168	15	48225-48239	CGTCGCCGACGGCGC
		14588-14602	CGTCGCCGACGGCGC
TPA4-D169	15	43180-43194	CGAGATCCCCGCCGA
		17618-17632	CGAGATCCCCGCCGA
TPA4-D170	15	53478-53492	CGAGCTCGACCTGCC
		20084-20098	CGAGCTCGACCTGCC
TPA4-D171	15	36120-36134	CTCACCGAAGCCGAC
		20417-20431	CTCACCGAAGCCGAC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-D172	15	26736-26750	GGGACGGCGCGCACG
		22129-22143	GGGACGGCGCGCACG
TPA4-D173	15	50113-50127	CTACAGCGGCGGCGG
		22940-22954	CTACAGCGGCGGCGG
TPA4-D174	15	49142-49156	GGCGCCGGGAAGGTC
		26034-26048	GGCGCCGGGAAGGTC
TPA4-D175	15	52693-52707	CCGATCATGAGCACC
		40626-40640	CCGATCATGAGCACC
TPA4-D176	15	50763-50777	AGGTCGACGCCCTCG
		44611-44625	AGGTCGACGCCCTCG
TPA4-D177	15	54447-54461	GCGCGGCGCCACCGA
		45438-45452	GCGCGGCGCCACCGA
TPA4-D178	15	50770-50784	CGCCCTCGCCGCCCT
		46323-46337	CGCCCTCGCCGCCCT
TPA4-D179	15	52561-52575	CCCGCCGCGCCTGCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		50168-50182	CCCGCCGCGCCTGCG
TPA4-I1	60	37008-37061	GCCGCGCGCTCGC---CGAGG--CGTCCCGC-GCCCTCGCCGAGTGCGTCGACGCGGCCC
		5524-5465	GCCGCGCGCTCGCTCTCGAGTGCCTTCTTGCCGCCCTCGCCGAGCGGCTCATCGCCGCCC
TPA4-I2	52	42939-42989	CGGCGACTACGGCATCTTC-ACCGGCGCCCGCACGCCCGCCGGCACCTACCC
		10361-10313	CGGCGCCCACGGCATCTTCGACCGCCGCGC---CGGCCGCCCTCACCTACCC
TPA4-I3	51	15634-15684	CGCGGCAGCGGCTCGGGCTCGGGCTCGGGCAGTGGCGGCGGTAGCGGGTCC
		38675-38630	CGCGGCAGCGGCCGGGGCCTGGGC---GGGAGCCGCGGCGG--GCGGGGCC
TPA4-I4	41	24444-24484	TGGTCGCGGGGTCGGTGCTGCAGCCGCCGCCGATGACGGTG
		9208-9168	TGGTGGCGAGGTCGTACCGCCAGATGCCACCGATGACGGTG
TPA4-I5	40	29921-29960	CGCCCTCGCCACCGCGGTCCTCGGACCCGTGATCTCCGCG
		14268-14229	CGCCTCCGCCACCGCGGCGAGCTGCCGCGTGGTCTCCGCG
TPA4-I6	36	51889-51924	GAGGTTCAGAGGGCGACCCTCGGAACCCTCGGAACC
		51743-51708	GAGGTTCCGAGGGTCGCCCTCGGAACCCTCTGAACC
TPA4-I7	36	24150-24185	TCCTGCCGCCGCCGATGACGGCGGCGGCGGACGTGC
		19573-19539	TCTTGCCGCCCTCGATGACGGC-GTGGCGTACGTGC
TPA4-I8	36	41093-41126	CGCGCTGGACGACGCCGCGGCGGCG--CTCCGCATC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		39339-39304	CGCCCTGCGCGCCGCCGCGGCGGCGATCTCCGCGTC
TPA4-I9	35	44992-45026	GACCGTCCCCGACCAGGCGACCGACCTCGAGGAGC
		2862-2828	GACCGTGCTCGTCGCGGCGCACGACCTCGAGGCGC
TPA4-I10	34	25057-25088	CGGCGGCGCCGGT--GGGCCGTCCTCGGCGCTCG
		12865-12833	CGGCGGCGCCGGTGCGGGCCG-CCGCGGCGTTCG
TPA4-I11	33	42886-42918	GTCGTCGGCGACGACTATGACCCCGGCGGCGTC
		6714-6682	GTCGACGCCGCCGAGCATGAGGCCGGCGGCGTC
TPA4-I12	32	53103-53134	CCCCGTCGCGGAGGATCGAGAGCGCCTCGCCG
		1042-1011	CCCCGTCGCGGAGGATCGTGCGCGTCTCGCCG
TPA4-I13	32	25491-25522	CGCCGAACCAGGACACCTGGCGGTACCGGTCG
		1780-1752	CGCCGAACCA---CACCACGCGGTACCGGTCG
TPA4-I14	32	53001-53032	CGCGCACGAGCGCGAGCAGCGCGCCCGGGGCG
		29396-29365	CGCCCTCGAACGCCTGCACCGCGCCCGGGGCG
TPA4-I15	31	41862-41892	ACGCGCTCGAGCTCCTCGCCGCGCTGGCCGC
		27542-27512	ACGCGCGCGTGCCCCGCGGCGCGCTGGCCGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-I16	31	34722-34752	CCTCACCGTCATCCAGGGGCCGATCACCCGC
		32591-32561	CCCCATCGGCACCCACGGGACGATCACCCGC
TPA4-I17	30	25877-25905	CGCCGCCGACGCGAACGACGTCGC-GAACG
		12178-12149	CGCCGCCGACGCGGAGGATCTCGCGGAACG
TPA4-I18	30	52015-52044	CCACCGCCCGCCCCCTCGCGAAGCGCGCCG
		39304-39275	CCAGCGCCCGCCCCCACGGCACGCGGGCCG
TPA4-I19	28	38393-38420	CTGGATGACCGGCGCGCTCGGCGACGCG
		24382-24355	CTGCACGACCGGCGCGCGCAGCGCCGCG
TPA4-I20	28	30268-30295	CGCGGCGGCGATCATCGCTGCCATCGCC
		39324-39297	CGCGGCGGCGATCTCCGCGTCCAGCGCC
TPA4-I21	27	19762-19788	ACCTCGTCGAACCCGACGCCGACGACC
		5052-5026	ACCTCGGCGTACACGACGCCGACGTCC
TPA4-I22	27	54742-54768	GCCTGCCCCATCTGCAAGACCGACCTG
		23703-23678	GCCTGCCCCATCTGCACGA-TGACCTG
TPA4-I23	27	35964-35990	CGCCGCCTACCGCACGCGGGCCGAGGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		39296-39271	CGCCCCC-ACGGCACGCGGGCCGCGGA
TPA4-I24	27	35530-35556	CCGGTACACGGGCGGCGCGAAGTGACC
		39677-39652	CCGGGACGCGGGCGGCGCGGA-TGACC
TPA4-I25	26	54695-54720	GCGCCGTCCGCATCGTCGACCGCCGC
		10359-10334	GCGCCCACGGCATCTTCGACCGCCGC
TPA4-I26	26	52480-52505	GACGAATCACGAGCGCGAGCATCTGC
		24601-24577	GACGA-TCAGAAGCGCGAGCATCTGC
TPA4-I27	25	43565-43587	CCGCCG--CGACCTCGACGACCCCG
		15581-15557	CCGCCGTACGACTTCGACGACCCCG
TPA4-I28	25	44098-44122	GGGCGCTGGTCCGGGCACGCCGAGG
		33934-33912	GGGCGC--GTGCGGGCACGCCGAGG
TPA4-I29	25	48395-48418	ACGGCGCCGAGCACGTC-ACGCCGG
		22789-22765	ACGGCGCCGAGGACGTCGTCGCCGG
TPA4-I30	25	54679-54703	CTCCACGCCATCACCCGCGCCGTCC
		32230-32206	CTCGACGCGTTCATCCGCGCCGTCC
TPA4-I31	25	42894-42918	CGACGACTATGACCCCGGCGGCGTC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		34152-34128	CGTCGACGGTGACACCGGCGGCGTC
TPA4-I32	24	31604-31627	GCCATCGGTGACGACGACGGTGCG
		6534-6511	GCCGTCGGTGCCGACGACGGTGCG
TPA4-I33	24	50098-50121	CGGCGACCACCAGCACTACAGCGG
		9999-9977	CGGCCACCACCAGCACTAC-GCGG
TPA4-I34	24	53568-53591	GCGCGAGCGGATCGCCGGCACCGT
		31853-31830	GCGCGTGACGATCGCCGGCACCGT
TPA4-I35	24	47617-47640	GAAGGCGACCGGGAGCCGGACCTC
		41527-41504	GATGGCGACCGGGATCCGGGCCTC
TPA4-I36	23	29822-29844	CGCGGACCTCCGCGCCGCGCTCT
		11581-11559	CGCGGACCTTCGCGCCGCGCTCT
TPA4-I37	23	50137-50159	CCTCGAGGCGCTGCGCGAGCTCA
		5646-5624	CCTCGAGGCGCTTCTCGAGTTCA
TPA4-I38	23	53421-53443	GCTCGCCCGCCACGTCCGCGCGC
		11782-11760	GCGCGCCCGCCACGCCCGCGAGC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-I39	23	42069-42091	GGCGGCGTGGTGGCTGACCACCG
		27079-27057	GGCGGCGCGGATGCTGACCACCG
TPA4-I40	23	55008-55030	CCGCTGCGGCGACGTCCTCGACG
		34585-34563	CCACTGCGGCGACGGCCGCGACG
TPA4-I41	22	45481-45502	CCGCCCACGACAGCGGCACCGC
		24791-24770	CCGCCCCGGACAGCGGCACCGC
TPA4-I42	22	31916-31937	GCGGATCGCGAGGGCCTCGCCG
		30038-30017	GCCGATCGCGAGGGCCTGGCCG
TPA4-I43	22	19727-19748	GCCGCGCGAGGTTCTGCTCACC
		7968-7949	GCCGCGCGAGGTTCT--TCACC
TPA4-I44	22	47755-47776	CGCTGCCCGCGCGTCGAGTGCC
		46636-46616	CGCTCCCC-CGCGTCGAGTGCC
TPA4-I45	21	48276-48296	CGCCGAGCTCCGCGGCGGCGG
		2331-2311	CGCGGAGCTGCGCGGCGGCGG
TPA4-I46	21	33300-33320	CGTCGGCCACGACTTCGGCGT

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		4964-4944	CGTCGGCGCCGACTTCGGCGT
TPA4-I47	21	44884-44904	TATTCGTCGACGCCGACTTCG
		4968-4948	TACTCGTCGGCGCCGACTTCG
TPA4-I48	21	40491-40511	CATCGACGCCGCCGGGGACGT
		7500-7480	CATCGACGCCGCGAGGGACGT
TPA4-I49	21	51329-51349	GCGATGGCCGTGTCGAAGCTG
		15614-15594	GCGATGGCCGTGCCGACGCTG
TPA4-I50	21	45376-45396	GACACCGTCGGCGACGTCGTG
		19191-19171	GACACCTTCGGCGACGTCTTG
TPA4-I51	21	37916-37936	TCAGCGCGCCGCCGTTCGAGG
		25787-25767	TCAGCGTGCCGCCGTTCGGGG
TPA4-I52	21	49890-49910	CCACCCGCATCGCCGCGGCCC
		31947-31927	CCACCCGCATCGGCGAGGCCC
TPA4-I53	21	52746-52766	CCTCACCGACGACCAGATCGA
		32069-32049	CCTCACCGACGACGAGCTCGA
TPA4-I54	21	45946-45966	GGTGCTCGACCTCGACGTCCA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		42421-42401	GGAGCTCGACCTCGAGGTCCA
TPA4-I55	21	48073-48093	CCGGTCGAGGGCCGCGACGGC
		46344-46324	CCGGTCGAGGGCGGCGAGGGC
TPA4-I56	20	37784-37803	GGGCGAGCTCGTCCGGGTCG
		20094-20075	GGTCGAGCTCGTCCGGGTCG
TPA4-I57	20	22258-22276	GGCTCGCCGGCCGC-GCGTC
		2656-2637	GGCTCGCCGGCCGCGGCGTC
TPA4-I58	20	31745-31764	GGCCTCGGCGGCGAAGTCGG
		10902-10883	GGACTCGACGGCGAAGTCGG
TPA4-I59	20	15832-15850	CGGC-AGGCCGCCCGCGACC
		13279-13260	CGGCGAGGCCGCCCGCGACC
TPA4-I60	20	43849-43868	GACGCCCTGCGCGCCGAGCG
		14520-14501	GACGGCCTGCGCGCCGGGCG
TPA4-I61	20	41559-41578	CGCCGAGCACGTCGCCGAGG
		17900-17881	CGCCGAGCACGGTGCCGAGG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-I62	20	52170-52189	GAACTCCTCGCCTTCCAAGA
		19596-19577	GAACTCCTCGCCATCCAGGA
TPA4-I63	20	45836-45855	GTCTGGCACCTCGCCAGCGG
		20078-20059	GTCGGGCACCTCGCCGGCGG
TPA4-I64	20	30710-30729	CGCACCGGACAGCGGCCCCG
		24790-24771	CGCCCCGGACAGCGGCACCG
TPA4-I65	20	29062-29081	GCCGGGGCCGGTGCGCAACG
		28200-28181	GCCGGCGCCGGTGCGCAGCG
TPA4-I66	20	42231-42250	CGCGGAGATGCGCGCCGACC
		29715-29696	CGCGGCGATGCGCGCCGGCC
TPA4-I67	20	45076-45095	CGAGGACGACGCCGGCCACG
		33310-33291	CGTGGCCGACGCCGGCCACG
TPA4-I68	20	48392-48411	GCGACGGCGCCGAGCACGTC
		37589-37570	GCGTCGGCGCCGAGCTCGTC
TPA4-I69	20	34267-34286	CCGCGGACCTGCGCGCCGCC
		39343-39324	CCGCCGCCCTGCGCGCCGCC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-I70	19	38403-38421	GGCGCGCTCGGCGACGCGG
		3673-3655	GGCGCGCTCGGCGCCGCGG
TPA4-I71	19	50413-50431	CCGCGTCGCGCTGCTCGCG
		35788-35770	CCTCGTCGCGCTGCTCGCG
TPA4-I72	18	8776-8793	GTCCTCGGCGGCGGCAAG
		1512-1495	GTCCTCGGCGGCGGCCAG
TPA4-I73	18	24554-24571	GCGGCGACCGCGGCGCCG
		19458-19441	GCGGCCACCGCGGCGCCG
TPA4-I74	18	32080-32097	CGCCGCGGTCGACGCGTC
		24568-24551	CGCCGCGGTCGCCGCGTC
TPA4-I75	18	51118-51135	GCACGCCGGCCGCGGCGA
		25277-25260	GCACGCCGGCCGCGCCGA
TPA4-I76	18	44335-44352	ACTGCGACGACGGCCGCG
		34583-34566	ACTGCGGCGACGGCCGCG
TPA4-I77	18	42101-42118	TCCTCGACGGCATCGACG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		36796-36779	TCCTCGACGGCCTCGACG
TPA4-I78	18	47312-47329	AGCGCGCCCTGCGCGGCG
		40068-40051	AGCGCGACCTGCGCGGCG
TPA4-I79	18	51399-51416	CCCGGGAGCGCGGCGAGC
		51285-51268	CCCGCGAGCGCGGCGAGC
TPA4-I80	17	21372-21388	GGAGCTCACCGTCGTCG
		7458-7442	GGAGCTCACCGTCGTCG
TPA4-I81	17	28933-28949	CGCGAGCCACGCGATCG
		12762-12746	CGCGAGCCACGCGATCG
TPA4-I82	17	50472-50488	GCACCCTCGCCGCGATC
		25944-25928	GCACCCTCGCCGCGATC
TPA4-I83	16	25111-25126	CGCGTCGGCGGCGACC
		6338-6323	CGCGTCGGCGGCGACC
TPA4-I84	16	47886-47901	CGCGCGGATCCGCGCC
		22859-22844	CGCGCGGATCCGCGCC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-I85	16	41077-41092	ACGAGCACCCGCACCG
		24141-24126	ACGAGCACCCGCACCG
TPA4-I86	16	50162-50177	GGCCGGCCCGCCGCGC
		30332-30317	GGCCGGCCCGCCGCGC
TPA4-I87	15	53441-53455	CGCACCAGCTGCAGC
		599-585	CGCACCAGCTGCAGC
TPA4-I88	15	32388-32402	TCGCGGGTGATCTCG
		2941-2927	TCGCGGGTGATCTCG
TPA4-I89	15	35883-35897	CCCGGCCGCCGCGGC
		3257-3243	CCCGGCCGCCGCGGC
TPA4-I90	15	5384-5398	GCCGGCGCTCGGCGC
		3357-3343	GCCGGCGCTCGGCGC
TPA4-I91	15	22164-22178	GGCGTCCTCGGCTGC
		5600-5586	GGCGTCCTCGGCTGC
TPA4-I92	15	19984-19998	GCCAGATGCCGGGGA
		6288-6274	GCCAGATGCCGGGGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
TPA4-I93	15	12166-12180	CGCGTCGGCGGCGAC
		6338-6324	CGCGTCGGCGGCGAC
TPA4-I94	15	49676-49690	CCGGTCACCGTCGCG
		9875-9861	CCGGTCACCGTCGCG
TPA4-I95	15	24554-24568	GCGGCGACCGCGGCG
		11732-11718	GCGGCGACCGCGGCG
TPA4-I96	15	15834-15848	GCAGGCCGCCCGCGA
		14884-14870	GCAGGCCGCCCGCGA
TPA4-I97	15	37522-37536	TCGACGGTGTGGGCC
		20543-20529	TCGACGGTGTGGGCC
TPA4-I98	15	30318-30332	CGCGGCGGGCCGGCC
		21468-21454	CGCGGCGGGCCGGCC
TPA4-I99	15	39846-39860	CGACCGCCGAGCCCC
		23783-23769	CGACCGCCGAGCCCC
TPA4-I100	15	48757-48771	GCACCCGCACCGTCG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		24137-24123	GCACCCGCACCGTCG
TPA4-I101	15	42886-42900	GTCGTCGGCGACGAC
		35018-35004	GTCGTCGGCGACGAC
TPA4-I102	15	43440-43454	CGCGACCTGCGCGGC
		40066-40052	CGCGACCTGCGCGGC
TPA4-I103	15	41425-41439	CGACCTCGGCGCCGA
		40169-40155	CGACCTCGGCGCCGA
GMA1-D1	120	121-240	CCGCCGCCGTCATCGCCGACGGCCGCACCCGCCGCGCCGCCCTCGAACAACGCCTCATCGACGAAGCCAACAAGGCGTTGG ACACCATGTGGGCGCCGCACGAGATAGGCGCATTCGGTG
		1-120	CCGCCGCCGTCATCGCCGACGGCCGCACCCGCCGCGCCGCCCTCGAACAACGCCTCATCGACGAAGCCAACAAGGCGTTGG ACACCATGTGGGCGCCGCACGAGATAGGCGCATTCGGTG
GMA1-D2	54	4744-4791	CGACACCGACAAGTCG-----ACGACAGCCGACACCGCGGCCGACAAGGGCGA
		2631-2684	CGACACCGACAAGACGATCTTCACGCCAGTCGGCGCAGCGTCAGATAAGGGCGA
GMA1-D3	39	20641-20679	ACGCCGGTCCGACTATGGCCGCTGATCGACGCGGGCATG
		9133-9167	ACGCCGGGCGGAC----GGCGCTGATCGACGTCGGCATG
GMA1-D4	35	27747-27778	CGAAGTTGACGGAGATCGGCG---AGGAAACGTTG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		17158-17192	CGAAGCTGAAGGAGATCGGCGGCCAAGACACGTTG
GMA1-D5	32	13862-13893	CTGTTCCAGATCGCGAAGGGCACGTGGACGTC
		13724-13755	CTGATCCAGAACCCGGACGGCACGTGGACGTC
GMA1-D6	29	14344-14372	GGCCCGCCACGACAAGGCCGAAGCCGAGA
		4813-4841	GGCCCGCAAGCACGAGGCCGAAGCCAAGA
GMA1-D7	25	31228-31252	TGAAAGGAAAGTCGCATGACCAGCA
		28614-28638	TGAAAGGAAATCAGCATGACCAGCA
GMA1-D8	25	35101-35125	GGCTGCACCGACTGGACGATCCGCA
		20856-20880	GGCCGCACGGTCTTGACGATCCGCA
GMA1-D9	25	32218-32241	GCCCGGTCCG-TGTCGCGTGCCGTG
		30330-30354	GCCTGGTCTGGTGTCGCGTGCCGTG
GMA1-D10	24	6949-6972	TCGCCGCACCAACCACCGAAGTCG
		2819-2842	TCGCCGCAACCGCCACCGAAGTCG
GMA1-D11	23	17829-17851	GGTCGCGCGCGTCGTGGACGCGC
		7819-7841	GGTCGCGCGCGTGCTGGACGCGC
GMA1-D12	23	35475-35497	CCGACCATCCGCGCCGCCGACGG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		33952-33974	CCGACCATCCGCGCGTTCGACGG
GMA1-D13	20	25205-25224	TCGAGGTCGCTGTCGCCGTC
		21891-21910	TCGAGGTAGCTGTCGCCGTC
GMA1-D14	20	29617-29636	TGTCGCCGCCGCCCTCATCC
		11233-11252	TGTCGCCGCCGGTCTCATCC
GMA1-D15	20	38752-38771	GCAAGCTCCGTGACGAGCAG
		15051-15070	GCAAGCTCCGTGACACGCAG
GMA1-D16	20	38140-38159	GGTTCAAGGTCGACGACGGT
		17257-17276	GGTTCAAGGTCGGCGACCGT
GMA1-D17	19	36349-36367	CGGCGACGCCCGCACCGAA
		9468-9486	CGGCGACGCCCGCACCGAA
GMA1-D18	19	33587-33605	CGACCAGGCGATCGCCGCG
		5404-5422	CGACCAGGCGAACGCCGCG
GMA1-D19	19	12915-12933	TCGTCATCAACCCTGACGA
		12744-12762	TCGTCATCAACCGTGACGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA1-D20	18	24932-24949	GACCAGGAACTACACGCG
		16382-16399	GACCAGGAACTACACGCG
GMA1-D21	18	41160-41177	CGCGACGCCTCGGCCTCG
		2603-2620	CGCGAGGCCTCGGCCTCG
GMA1-D22	18	14064-14081	CCGGCGGCGACCCCGGCA
		13794-13811	CCGGCGGCGACCCCGACA
GMA1-D23	18	37225-37242	GCTCGGCGACTCGTGCGA
		36965-36982	GCTCGGCGACTCGTACGA
GMA1-D24	17	5758-5774	CGTCATCGACGAGTTCC
		5290-5306	CGTCATCGACGAGTTCC
GMA1-D25	17	32755-32771	CGCCGAGGCGGGTGAGG
		32534-32550	CGCCGAGGCGGGTGAGG
GMA1-D26	16	9173-9188	CACCCTCGCTGAGGCG
		5242-5257	CACCCTCGCTGAGGCG
GMA1-D27	15	33173-33187	CCGCGCCGCCCTCGA

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		32-46	CCGCGCCGCCCTCGA
GMA1-D28	15	33173-33187	CCGCGCCGCCCTCGA
		152-166	CCGCGCCGCCCTCGA
GMA1-D29	15	33799-33813	GACGCCGTCGACGCG
		3223-3237	GACGCCGTCGACGCG
GMA1-D30	15	19873-19887	ACACCGTCACCGACG
		6335-6349	ACACCGTCACCGACG
GMA1-D31	15	10592-10606	TCCACCATCGGCATC
		10010-10024	TCCACCATCGGCATC
GMA1-D32	15	34043-34057	GTCGCCGTCGTGGAC
		17750-17764	GTCGCCGTCGTGGAC
GMA1-D33	15	34254-34268	GAGGAGTCGTGAAGT
		32547-32561	GAGGAGTCGTGAAGT
GMA1-D34	15	35154-35168	CGCCGAGCGCGCCGA
		33569-33583	CGCCGAGCGCGCCGA
GMA1-I1	47	22486-22531	CACGATGTCGGGGTCCGGTGGCGCATTC-CGCACGATGATCGTCGCC

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
		13816-13771	CACGATGTCGGGGT-CGCCGCCGGATTCGCGCTTGATGAGCGCCGCC
GMA1-I2	38	22006-22043	GGATGCGACGGTGTCACCGCCGGAGATGGTCGCGTTCG
		1459-1422	GGATGCGACGGTGCCGATGCCGTCGATGACAGCGTTCG
GMA1-I3	33	30408-30440	ACGTCGTCACCTGCGGCGTTCGCGAAGAGTGTC
		7024-6992	ACTTCCTCACCTGCGGCGTTGCGGACGAGTTTC
GMA1-I4	23	10499-10521	CAGGCCACCGCGGCCCTTCGTGC
		4401-4379	CAAGCCGCCGAGGCCCTTCGTGC
GMA1-I5	20	26623-26642	CGGTCACAGCCTCACCTCGC
		20752-20733	CGGTCACAGCCTCAGCACGC
GMA1-I6	17	38627-38643	GCGCGAGACTTCCCGCC
		4038-4022	GCGCGAGACTTCCCGCC
GMA1-I7	16	28081-28096	GAAGGTTGGGCGCCCG
		23291-23276	GAAGGTTGGGCGCCCG
GMA1-I8	15	24463-24477	TCCGGTCGTCGTCGG
		2248-2234	TCCGGTCGTCGTCGG

Phage-Repeat number	Size (bp)	Coordinates	Sequence alignment
GMA1-I9	15	26806-26820	GCCGCCGCCTGCGGG
		5119-5105	GCCGCCGCCTGCGGG
GMA1-I10	15	12764-12778	CTCGATGCGGCGTCG
		10794-10780	CTCGATGCGGCGTCG
GMA1-I11	15	25463-25477	GGATGCCGCGGCCGT
		21584-21570	GGATGCCGCGGCCGT

I indicates inverted repeat, D indicates direct repeat.

8.3. Appendix 3 – Candidate's publications included in this thesis

Dyson, Z. A., Tucci, J., Seviour, R. J. & Petrovski, S. (2015). Three of a kind: Genetically similar *Tsukamurella* phages TIN2, TIN3, and TIN4. *Appl and Enviro Microbiol* **81**, 6767-72.

Dyson, Z. A., Tucci, J., Seviour, R. J. & Petrovski, S. (2015). Lysis to kill: evaluation of the lytic abilities, and genomics of nine bacteriophages infective for *Gordonia* spp. and their potential use in activated sludge foam bio-control. *PLoS ONE* **10**, e0134512.

Dyson, Z. A., Tucci, J., Seviour, R. J. & Petrovski, S. (2015, *In press*). Isolation and characterisation of bacteriophage SPI1, which infects the activated sludge foaming-bacterium *Skermania piniformis*. *Arch Virol.*

AEM
Journals.ASM.org

Three of a Kind: Genetically Similar *Tsukamurella* Phages TIN2, TIN3, and TIN4

Zoe A. Dyson,[a] Joseph Tucci,[a] Robert J. Seviour,[b] Steve Petrovski[b]

La Trobe Institute of Molecular Sciences, Bendigo, Victoria, Australia[a]; Department of Physiology, Anatomy and Microbiology, La Trobe University, Bundoora, Victoria, Australia[b]

Three *Tsukamurella* phages, TIN2, TIN3, and TIN4, were isolated from activated sludge treatment plants located in Victoria, Australia, using conventional enrichment techniques. Illumina and 454 whole-genome sequencing of these *Siphoviridae* viruses revealed that they had similar genome sequences, ranging in size between 76,268 bp and 76,964 bp. All three phages shared 74% nucleotide sequence identity to the previously described *Gordonia* phage GTE7. Genome sequencing suggested that phage TIN3 had suffered a mutation in one of its lysis genes compared to the sequence of phage TIN4, to which it is genetically very similar. Mass spectroscopy data showed the unusual presence of a virion structural gene in the DNA replication module of phage TIN4, disrupting the characteristic modular genome architecture of *Siphoviridae* phages. All three phages appeared highly virulent on strains of *Tsukamurella inchonensis* and *Tsukamurella paurometabola*.

The genus *Tsukamurella* contains Gram-positive organisms in the *Corynebacteriales* (1) and currently includes 11 species (2). Members of this genus are strictly aerobic and weakly/variably acid-fast, nonmotile, non-spore-forming organisms whose cell envelopes contain long-chain unsaturated mycolic acids (3–8). Together with other mycolic acid-producing genera, including *Mycobacterium*, *Millisia*, *Nocardia*, *Rhodococcus*, *Williamsia*, *Rhodococcus*, and *Gordonia*, they are often referred to collectively as the mycolata, a group of organisms whose mycolic acids render their cells highly hydrophobic (9, 10).

The high hydrophobicity and the biosurfactants produced by the mycolata implicate them as stabilizing agents of foams in activated sludge plants (9, 11). Several *Tsukamurella* species including *T. spumae* (12), *T. sunchonensis* (13), and *T. pseudospumae* (12) have been isolated from these foams. Stable foams complicate sludge management, have a negative impact on plant esthetics, and increase maintenance costs (9).

Several opportunistic *Tsukamurella* pathogens are known to exist, thus posing a potential health hazard to plant operators from aerosol foam dispersal (9). Thomas et al. (14) proposed that a phage therapy biocontrol approach might be an attractive option for treating this global operational problem and successfully isolated 17 phages infective for foaming mycolata from Australian wastewater treatment plants. More mycolata phages have been isolated and described since then (15–21). Among them is the TPA2 phage, infective for members of the genus *Tsukamurella* (15).

As of March 2015, genomes of 228 phages infective for members of the closely related genus *Mycobacterium* have been sequenced. To better understand their genomics, Hatfull et al. (22) devised a clustering system that claimed to reveal insights into their evolutionary interrelationships. In this system, TPA2 phage would be placed into their cluster B as its genome organization is similar to that of the *Mycobacterium* phage Rosebush (15). Despite this similarity between TPA2 and *Mycobacterium* phages, little else is known about the extent of genetic diversity of phages infective for *Tsukamurella*.

In this study, we report the isolation and characterization of three lytic *Tsukamurella* phages, TIN2, TIN3, and TIN4. These phages were characterized based on their morphologies, host ranges, complete genome sequences, and structural protein analysis using mass spectroscopy (MS) and compared to other phages lytic for members of this genus.

MATERIALS AND METHODS

Isolation and preliminary characterization of phages TIN2, TIN3, and TIN4. The bacterial strains used in this study are listed in Table 1, and methods for their growth are those described previously (15). Phage TIN2 was isolated and subsequently purified from an activated sludge sample collected from Bendigo, Victoria, Australia, as detailed elsewhere (15).

Phages TIN2, TIN3, and TIN4 were isolated on *T. inchonensis* strain DSMZ 44067 lawn plates from samples from treatment plants in Victoria, Australia. TIN3 and TIN4 phages were obtained from the same sample. Subsequent enrichment experiments then allowed each to be recognized based on its plaque size during phage purification since plaques formed by TIN4 phage were slightly larger (~0.50 mm) than those formed by TIN3 phage (~0.25 mm), as discussed later.

Host range determinations and preparation of grids for transmission electron microscopy (TEM) visualizations of virions were performed as described previously (15). The carbon- and Formvar-coated grids (Electron Microscopy Sciences, Australia) were stained with 2% uranyl acetate and examined using a Tenaci Fei T30 transmission electron microscope at an accelerating voltage of 300 kv.

Genome sequencing of bacteriophages TIN2, TIN3, and TIN4. Genomic DNA was extracted from phage TIN2 and sequenced using a Roche GS FLX genome sequencer and titanium chemistry (15). Genomic

Received 7 April 2015 Accepted 15 July 2015
Accepted manuscript posted online 17 July 2015
Citation Dyson ZA, Tucci J, Seviour RJ, Petrovski S. 2015. Three of a kind: genetically similar *Tsukamurella* phages TIN2, TIN3, and TIN4. Appl Environ Microbiol 81:6767–6772. doi:10.1128/AEM.01145-15.
Editor: M. J. Pettinari
Address correspondence to Steve Petrovski, steve.petrovski@latrobe.edu.au.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/AEM.01145-15.

doi:10.1128/AEM.01145-15

TABLE 1 Strains used in this study

Tsukamurella organism	Strain no.		Phage-forming lytic plaque(s)	Reference
	La Trobe University culture collection[a]	Synonym(s)[b]		
T. inchonensis	CON50^T	DSMZ 44067	TIN2, TIN3, TIN4	8
	CON52	NCTC10741	TIN2, TIN3, TIN4	40
T. paurometabola	CON37^T	Tpau37, ATCC 25938	TPA2	7
	CON51	DSMZ 20162	TPA2	7
	CON53	IMRU1283	TPA2	41
	CON54	M337, IMRU1505	TPA2, TIN2, TIN3, TIN4	6
	CON55	M343, IMRU1312	TPA2, TIN2, TIN3, TIN4	41
	CON61	M334, IMRU1520, DSMZ44119	TPA2, TIN2, TIN3, TIN4	7
T. pseudospumae	TPS^T	N1176, DSMZ44118	TPA2	12
T. spumae	CON62^T	N1171, DSMZ44113	TPA2	6
T. tyrosinosolvens	CON57^T	DSMZ44234	TPA2	42

[a] The superscript "T" indicates a type strain. The cultures have been deposited in or obtained from public culture collections.
[b] Synonym(s) from other culture collections. DSMZ, Deutsche Sammlung von Mikroorganismen und Zellkulturen (German Collection of Microorganisms and Cell Cultures); ATCC, American Type Culture Collection; NCTC, National Collection of Type Cultures (Central Public Health Laboratory, England); IMRU, Institute of Microbiology, Rutgers University.

DNA was extracted from phages TIN3 and TIN4 in the same manner, and DNA sequencing libraries were prepared using an Illumina Nextera XT sample preparation kit as per the manufacturer's instruction. The prepared DNA libraries were sequenced on an Illumina MiSeq as a 150-bp paired-end run. Sequence reads were assembled for all three phages using the CLC Workbench version (version 6.5.1).

Genome annotation. The genome open reading frames (*orf* genes) within the *de novo*-assembled sequences were detected initially using Glimmer (version 3.02), where *orf* genes as small as 90 bp were revealed (23). All predicted start codons were inspected for the presence of putative ribosomal binding sites and corrected manually if required. Sequence similarity searches were carried out against the GenBank database. The presence of tRNA and transfer-messenger RNA (tmRNA) was also determined using ARAGORN (24) and tRNAScan-SE (25). Transmembrane domains were predicted with the DAS Transmembrane Prediction Server (26).

Mass spectroscopy of phage proteins. To identify phage structural proteins, high-titer purified phage particles (PFU/ml of $>10^{10}$) were precipitated using $(NH_4)_2SO_4$, followed by a second precipitation with $ZnCl_2$. Pellets were resuspended in 8 M urea to a final volume of 100 μl prior to analyses at the Mass Spectroscopy and Proteomics Facility at the La Trobe Institute of Molecular Sciences (Melbourne, Australia). Peptides reconstituted in 0.1% formic acid and 2% acetonitrile (buffer A) were loaded onto a trap column (C_{18} PepMap, 300-μm inner diameter [i.d.], 2-cm trapping column [Thermo-Fisher Scientific, Scoresby, Australia]) at 5 μl/min for 6 min and washed for 6 min before switching the precolumn in line with the analytical column (Vydac MS C_{18}, 75-μm i.d. by 25 cm; Grace Davison). The separation of peptides was performed at 300 nl/min using a linear acetonitrile (ACN) gradient of buffer A and buffer B (0.1% formic acid, 80% ACN) from 5% buffer B to 40% over 60 min. Data were collected on a hybrid quadrupole/time of flight MS (MicroTOF-Q; Bruker, Germany) with a nano-electrospray ion source in the data-dependent acquisition mode and an MS scan range of *m/z* 150 to 2,500. Nitrogen was used as the collision gas. The ionization tip voltage and interface temperature were set at 4,200 V and 205°C, respectively. Collision-induced dissociation (CID) tandem MS (MS/MS) spectra were collected for the four most intense ions. Dynamic exclusion parameters were set as follows: repeat count, 2; duration, 60 s. The data were collected and analyzed using data analysis software (Bruker Daltonics, Bremen, Germany).

Nucleotide sequence accession numbers. The nucleotide sequences for phages TIN2, TIN3, and TIN4 have been deposited in the GenBank under accession numbers KR011062, KR011063, and KR011064, respectively.

RESULTS AND DISCUSSION

Isolation and characterization of TIN2, TIN3, and TIN4 phages. Samples of activated sludge were screened for the presence of virulent phages by enrichment and plaque plating on lawn plates of *Tsukamurella inchonensis* (DSMZ 44067). Three phages were detected and named TIN2, TIN3, and TIN4. TIN2 and TIN3 produced plaques with an average size of ~0.25 mm, and TIN4 plaques were slightly larger, with an average size of ~0.5 mm. All three purified phages were then screened against the 73 mycolata isolates, which came mainly from activated sludge plants, held in the La Trobe University culture collection (11). They included species of *Gordonia*, *Nocardia*, *Rhodococcus*, *Mycobacterium*, *Skermania*, and *Tsukamurella*. As well as lysing *T. inchonensis* strain DSMZ 44067, all three phages lysed another *T. inchonensis* strain as well as three strains of *Tsukamurella paurometabola* (Table 1).

The identical host ranges of phage TIN2, TIN3, and TIN4 suggested that they might be similar genetically. Polyvalency has been reported for *Tsukamurella* phage TPA2, which lysed strains of *T. paurometabola*, *T. pulmonis*, *T. tyrosinosolvens*, *T. pseudospumae*, and *T. spumae* (Table 1) (15).

When the three phages were screened against other *T. paurometabola* strains, including DSMZ 20162 and IMRU1283, no plaque formation occurred with any of them. A similar level of stringent strain specificity has been observed in other actinophages (15–21, 27). Possible explanations include an absence of a compatible receptor site in the resistant host cell, absence of appropriate molecular machinery to support infection, or possession of phage resistance systems like clustered regularly interspersed repeat (CRISPR) systems, restriction modification (RM) systems, and abortive infection (Abi) systems (28).

TIN2, TIN3, and TIN4 phages are all members of the family *Siphoviridae*. Phages TIN2, TIN3, and TIN4 all display an isometric capsid and a long noncontractile tail, characteristic of phages belonging to the *Siphoviridae* (Fig. 1). All three phages demonstrated very similar morphologies, with each having type B1 isometric capsids (29) ranging from ~57 to ~69 nm in size and tails ranging from ~450 to ~471 nm in length.

Phages TIN2, TIN3, TIN4 are all genetically very similar. The genome sequence of TIN2 phage was obtained with Roche/454

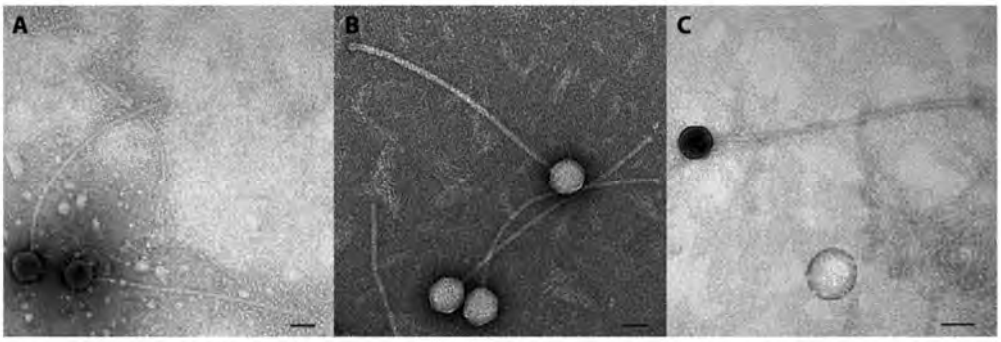

FIG 1 Virion morphologies of phages TIN2 (A), TIN3 (B), and TIN4 (C). Scale bar, 50 nm.

pyrosequencing that generated 47,091 reads, with an average of 224-fold coverage following assembly. The genome sequences of phages TIN3 and TIN4 were generated with Illumina sequencing that gave 1,432,104 and 712,336 reads, respectively. When assembled, the genomes of phages TIN3 and TIN4 had averages of 1,987-fold and 900-fold coverage, respectively.

The TIN2, TIN3, and TIN4 phages possessed double-stranded DNA genomes of 76,964 bp, 76,269 bp, and 76,268 bp, respectively. Phage TIN2, TIN3, and TIN4 genomes had G+C contents of 58.9 mol%, 59.3 mol%, and 59.3 mol%, respectively, which are lower than those of their corresponding hosts, which range between 67 and 78 mol% (30). All three phage genomes contained a single $tRNA^{Asn}$ that might aid in the acquisition of rare codons (31, 32), but no tmRNAs were detected.

An alignment of these genome sequences demonstrated that they are very similar to each other and also to the sequence of phage GTE7, as shown in Fig. 2 and Table S1 in the supplemental material (18). Whole-genome alignments showed that phage TIN2 shared high levels of nucleotide sequence identity (89% identity) with both TIN3 and TIN4. However, among the three, phages TIN3 and TIN4 are most similar, with the TIN3 genome differing from that of TIN4 by a single short variable region of 20 bp located in the lysis module in TIN3. In all three, genes are oriented in both forward and reverse orientations, with 107 to 109 putative *orf* genes in total and 34 arranged in a forward orientation in each. Only 21 to 23% of these putative genes in each could be annotated functionally, and a high number of ORFans (orphan genes, without known homologues in other organisms) of unknown function were present in all three genomes (see Table S1).

Genomes of all three phages are largely modular, being organized with genes of similar functions clustered together in DNA packaging, head and tail morphogenesis, cell lysis, and DNA replication/maintenance modules. In the absence of an obvious origin of replication in any of these phages, all *orf* genes were numbered consecutively, starting with the small terminase gene, and in the same transcriptional direction as this gene (*orf1*).

Sequence repeats occur in all three *Tsukamurella* phage genomes. Several repeat structures were identified in these sequences, as summarized in Tables S2 and S3 in the supplemental material. All three contained 11 or 12 palindromes ranging from 16 to 54 bp in size. The majority of these were located intergenically, where they might act as *rho*-independent transcriptional terminators (43). All three phage genomes contained 16 or 17 inverted repeat structures ranging in size from 16 to 68 bp. These repeat structures may be associated with replication origins (44) and transposable elements, but neither were seen here.

A total of 17 to 19 direct repeat structures were also identified in each of the genomes, ranging in size from 16 to 246 bp. Similar repeat structures have been observed in *Tsukamurella* phage TPA2 (15) as well as in other mycolata phages (21), where similar hypotheses have been proposed for their possible functions.

Genome annotation of phages TIN2, TIN3, and TIN4. Annotations of the TIN2, TIN3, and TIN4 genomes are summarized in

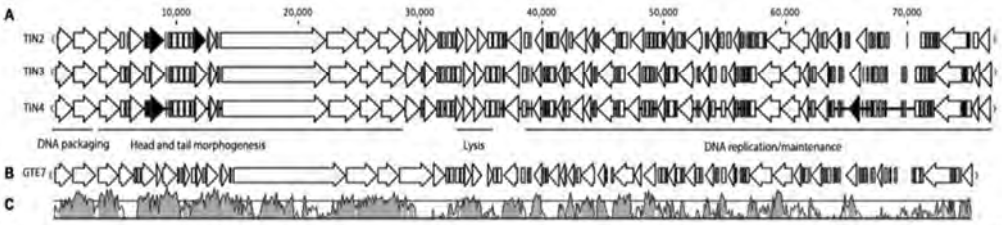

FIG 2 Genome map of phages TIN2, TIN3, and TIN4. (A) Phage genome modular organization of TIN2, TIN3, and TIN4. Shaded genes indicate those identified as encoding structural proteins by mass spectroscopy data. (B) Phage GTE7 genome organization. (C) Pairwise alignment of phages TIN4 and GTE7.

Table S1 in the supplemental material and clearly demonstrate how similar they are at the amino acid level.

The packaging modules of TIN2, TIN3, and TIN4 phages contain genes *orf1* and *orf2*, which encode the large and small terminase subunits, respectively. Typically, these two genes function together as a complex, with the small terminase subunit determining the specificity of DNA packaging (35) while the large terminase subunit mediates cleavage of the phage DNA packaged into the prohead (36).

The phage structural module was located immediately downstream of the DNA packaging module in all three genomes. All contained within this a conserved translational slippage mechanism commonly seen in *Siphoviridae* phages (37), located in the two genes immediately upstream of the putative tape measure protein encoded by *orf18*. These genes are thought to function in tail assembly (37). The same slippage mechanism was also seen in phage GTE7 (18).

The phage lysis module is located after the structural genes in all three genomes and typically consists of one or more phage lysin genes located adjacent to a phage holin gene. Together, these lysins and holin are responsible for the release of phage progeny at the end of the replication cycle (38). *orf29* in all three phages appears to encode one of the phage lysis enzymes as it shares between 74 and 77% identity to a lysozyme-encoding gene from phage GTE7. *orf30* in all three phages may encode holins as their transcribed and translated proteins all share amino acid sequence similarities to the Gp54 protein of *Tsukamurella* phage TPA2, identified previously as a putative holin (15). However, Gp54 does not satisfy the listed criteria (39) for such a protein, which should be less than 150 amino acid residues in size and contain two or more transmembrane regions. As neither occurs here, its function remains unclear.

A second putative lysin is encoded by *orf31* in all three phages and shares between 74 and 75% of its protein amino acid sequence identity with a putative peptidase gene of phage GTE7. It also contains the peptidase motif pfam01551. *orf32* in all three might also encode a holin protein as its translated protein is 152 amino acids in size, which is close to the 150 amino acids suggested to be its likely size (39), and contains between four and five transmembrane regions.

Immediately downstream of the phage lysis module in all three phages are the genes responsible for DNA replication/maintenance. More differences between the genome of phage TIN2 and the genomes of phages TIN3 and TIN4 were noticed here. Several motifs are present in the putative genes in this module in phage TIN2 that were absent from phages TIN3 and TIN4. For example, in the genome sequences of phage TIN2, *orf38* shares 49% identity to *gp034* of phage GTE7 and encodes a motif for a prolipoprotein diacylglyceryl transferase (PRK131008). This motif was absent from the equivalent gene (*orf37*) in phages TIN3 and TIN4, which shared only 57% identity to the product of *gp034* of phage GTE7. Orf40 in all three phages shares 86 to 87% amino acid sequence identity to the DNA methylase in phage GTE7 and possesses the pfam01555 motif indicative of a DNA methylase. Thus, *orf40* might aid phage resistance to any host restriction modification systems for cleaving foreign DNA (28).

No genes like those encoding integrases, excisionases, and repressor proteins associated with lysogeny (33) were identified in these genomes, suggesting that all are highly virulent phages. This would be an attractive attribute in any potential application for phage control of its *Tsukamurella* host.

Structural proteomics of TIN2, TIN3, and TIN4 virions. Structural genes appear to be located from *orf7* to *orf18* in all three phages. Mass spectroscopy studies confirmed this function for the *orf7*, *orf8*, and *orf15* genes in phage TIN2 and the *orf7*, *orf8*, and *orf89* genes in TIN4 (see Table S4 in the supplemental material). Because phages TIN3 and TIN4 were genetically very similar, TIN3 phage proteins were not analyzed. The function of Orf89 as a structural protein gene in phage TIN4 is unusual as it is located in the DNA replication/maintenance module, thus disrupting the otherwise expected modular structure of these genomes. This arrangement is likely to hold for TIN3 phage as well, where the same motifs (see Table S4) are present in Orf89. What functional attribute this gene plays in virion morphogenesis remains to be determined.

TIN3 contains mutations in the lysis module. The phage lysis module in TIN3 phage appears to extend from *orf29* to *orf32* and contains putative holin-encoding genes (*orf30* and *orf32*) and two putative lysin genes (*orf29* and *orf31*). Phage genomes of TIN3 and TIN4 differ only in their sequences within one gene (*orf29*) by 20 bp in TIN3, where it appears to result in a missense mutation. Subsequently, a truncated phage lysis/lysozyme would be expected to be synthesized compared to that formed in phage TIN4. Whether this mutation in phage TIN3 contributes to reduced phage replication kinetics or rates of cell lysis remains to be determined.

Phages TIN3 and TIN4 had been separated, as described earlier, on the basis of plaque size and morphology after infection of the same *T. paurometabola* strain from the same wastewater sample. Thus, it could be argued that this 20-bp mutation may be the basis for this observed difference in plaque size. It is possible that these two phages represent variants of the same phage, with phage TIN3 having undergone several genetic mutations. The second putative lysin gene is encoded by *orf31* in all three phages, so if Orf29 is rendered dysfunctional by mutation, Orf31 alone might facilitate host cell lysis.

Evolutionary relationships of phages TIN2, TIN3, and TIN4. The high degree of genome sequence similarity of the three *Tsukamurella* phages suggests a common evolutionary ancestry, most strikingly displayed with phages TIN3 and TIN4, where the only difference is in the putative lysin gene *orf29* in TIN3. Several putative indels were seen after their genome sequences were compared. For example, their putative tape measure protein genes (*orf18*) differed in size by only 9 bp, consistent with either a putative insertion or deletion event having occurred. Several putative substitution events were also noticed in the putative holin-encoding *orf32* gene. This gene in all three phages has the same length and has a similar sequence but produces different best-match results after a BLAST search using the GenBank database. Here, the closest relative of *orf32* from TIN2 is found in *Mycobacterium* phage Jolie1 while the homologues in TIN3 and TIN4 are most similar to the homologue in phage 39HC.

The only sequenced lytic *Tsukamurella* phage genome published to date is that of phage TPA2, which contains several chimeric genes. However, none of these were seen in the genomes of the TIN2, TIN3, and TIN4 phages (15). Yet possible evidence for recombination events resulting in new modular arrangements of genes, like those seen in phage TPA2 (15), was noticed here. A good example is in the genome sequences of phages TIN3 and

TIN4, where the translated amino acid sequence of genes *orf54* to *orf57* in both show high sequence similarities to genes *gp051* to *gp054* in phage GTE7. Following these genes in phages TIN3 and TIN4 are three novel genes (*orf58* to *orf60*) and a GTE7 phage gene encoding a *gp056* protein homologue (Orf61). Furthermore, several genes in the phage GTE7 genome are absent from the genomes of phages TIN2, TIN3, and TIN4. For example, the third lysin gene of phage GTE7, *gp040*, has no homologue in any of the three phages characterized here. This finding might indicate that some of these genes are accessory elements and superfluous to core phage functions, or perhaps their roles are carried out by other genes in these TIN phages.

The only protein detected in the three *Tsukamurella* phages described here that is similar in its amino acid sequence to any of the proteins present in the previously characterized *Tsukamurella* phage TPA2 is the *orf30* protein that is a homologue of Gp054, the holin gene in phage TPA2. The high degree of nucleotide sequence similarity between these three *Tsukamurella* phages and the phage GTE7 infective for certain members of the genera *Gordonia* and *Nocardia* (18) makes their host range differences surprising.

The three *Tsukamurella* phages described here could not be assigned to any cluster in the scheme of Hatfull et al. (22) and, instead, group with a similarly unrelated singleton phage. As several phages of this type exist that are likely to be related evolutionarily, perhaps this *Mycobacterium* phage grouping scheme should be extended to include a new cluster for phages ReqiDocB7, GTE7, TIN2, TIN3, and TIN4 (18, 34).

Conclusions. The isolation of three genetically very similar *Tsukamurella* phages has provided insights into their genomics and evolution. These include identifying mutations within the lysis module of phage TIN3 alone. Mass spectroscopy data suggest the presence of a structural protein gene within the DNA replication module of phage TIN4, disrupting its expected modular genome architecture. Furthermore, TIN2, TIN3, and TIN4 phages are genetically quite different from the *Tsukamurella* phage TPA2 (15). The three phages described here were highly virulent on all the *T. inchonensis* strains from our culture collection and on some *T. paurometabola* strains. In contrast, TPA2 did not form plaques on *T. inchonensis* but was highly virulent on the remainder of the *Tsukamurella* strains listed in Table 1 (15). This work demonstrates the genetic diversity among *Tsukamurella* phages isolated from activated sludge.

ACKNOWLEDGMENTS

We thank Eric Hanssen (Bio21 Institute) for assistance with transmission electron microscopy, Pierre Faou (La Trobe Institute of Molecular Sciences) for assistance with mass spectroscopy, Stephen Doyle (Department of Animal, Plant and Soil Sciences, La Trobe University) for assistance in Illumina sequencing, and Daniel Tillett (Nucleics, Australia) for useful discussions as well as technical advice.

Z.A.D. was the recipient of an Australian Postgraduate Award Ph.D. scholarship.

REFERENCES

1. Schwartz MA, Tabet SR, Collier AC, Wallis CK, Carlson LC, Nguyen TT, Kattar MM, Coyle MB. 2002. Central venous catheter-related bacteremia due to *Tsukamurella* species in the immunocompromised host: a case series and review of the literature. Clin Infect Dis 35:e72–e77. http://dx.doi.org/10.1086/342561.
2. Munk AC, Lapidus A, Lucas S, Nolan M, Tice H, Cheng J-F, Del Rio TG, Goodwin L, Pitluck S, Liolios K, Huntemann M, Ivanova N, Mavromatis K, Mikhailova N, Pati A, Chen A, Palaniappan K, Tapia R, Han C, Land M, Hauser L, Chang Y-J, Jeffries CD, Brettin T, Yasawong M, Brambilla E-M, Rohde M, Sikorski J, Göker M, Detter JC, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk H-P. 2011. Complete genome sequence of *Tsukamurella paurometabola* type strain (no. 33^T). Stand Genomic Sci 4:342–351. http://dx.doi.org/10.4056/sigs.1894556.
3. Bouza E, Perez-Parra A, Rosal M, Martin-Rabadan P, Rodriguez-Creixems M, Marin M. 2009. *Tsukamurella*: a cause of catheter-related bloodstream infections. Eur J Clin Microbiol Infect Dis 28:203–210. http://dx.doi.org/10.1007/s10096-008-0607-2.
4. Almeida DRP, Miller D, Alfonso EC. 2010. *Tsukamurella*: an emerging opportunistic ocular pathogen. Can J Ophthalmol 45:290–293. http://dx.doi.org/10.3129/i09-252.
5. Esteban J, Calvo R, Molleja A, Soriano F. 1998. Isolation of *Tsukamurella*-like organisms from human samples: contamination, colonization, or infection? Clin Microbiol Newsl 20:6–8. http://dx.doi.org/10.1016/S0196-4399(01)80037-4.
6. Nam SW, Chun J, Kim S, Kim W, Zakrzewska-Czerwinska J, Goodfellow M. 2003. *Tsukamurella spumae* sp. nov., a novel actinomycete associated with foaming in activated sludge plants. Syst Appl Microbiol 26:367–375. http://dx.doi.org/10.1078/072320203322497392.
7. Collins MD, Smida J, Dorsch M, Stackebrandt E. 1988. *Tsukamurella* gen. nov. harboring *Corynebacterium paurometabolum* and *Rhodococcus aurantiacus*. Int J Syst Evol Microbiol 38:385–391.
8. Yassin AF, Rainey FA, Brzezinka H, Burghardt J, Lee HJ, Schaal KP. 1995. *Tsukamurella inchonensis* sp. nov. Int J Syst Bacteriol 45:522–527. http://dx.doi.org/10.1099/00207713-45-3-522.
9. de los Reyes FL, III. 2010. Foaming, p 215–258. *In* Seviour RJ, Nielsen PH (ed), Microbial ecology of activated sludge. IWA Publishing, London, United Kingdom.
10. Savini V, Fazii P, Favaro M, Astolfi D, Polilli E, Pompilio A, Vannucci M, D'Amario C, Di Bonaventura G, Fontana C, D'Antonio D. 2012. Tuberculosis-like pneumonias by the aerobic actinomycetes *Rhodococcus*, *Tsukamurella* and *Gordonia*. Microbes Infect 14:401–410. http://dx.doi.org/10.1016/j.micinf.2011.11.014.
11. Petrovski S, Dyson ZA, Quill ES, McIlroy SJ, Tillett D, Seviour RJ. 2011. An examination of the mechanisms for stable foam formation in activated sludge systems. Water Res 45:2146–2154. http://dx.doi.org/10.1016/j.watres.2010.12.026.
12. Nam SW, Kim W, Chun J, Goodfellow M. 2004. *Tsukamurella pseudospumae* sp. nov., a novel actinomycete isolated from activated sludge foam. Int J Syst Evol Microbiol 54:1209–1212. http://dx.doi.org/10.1099/ijs.0.02939-0.
13. Seong CN, Kim YS, Baik KS, Choi SK, Kim MB, Kim SB, Goodfellow M. 2003. *Tsukamurella sunchonensis* sp. nov., a bacterium associated with foam in activated sludge. J Microbiol Biotechnol 41:83–88.
14. Thomas JA, Soddell JA, Kurtböke DI. 2002. Fighting foam with phages. Water Sci Technol 46:511–553.
15. Petrovski S, Seviour RJ, Tillett D. 2011. Genome sequence and characterization of the *Tsukamurella* bacteriophage TPA2. Appl Environ Microbiol 77:1389–1398. http://dx.doi.org/10.1128/AEM.01938-10.
16. Petrovski S, Dyson ZA, Seviour RJ, Tillett D. 2012. Small but sufficient: the *Rhodococcus* phage RRH1 has the smallest known *Siphoviridae* genome at 14.2 kilobases. J Virol 86:358–363. http://dx.doi.org/10.1128/JVI.05460-11.
17. Petrovski S, Seviour RJ, Tillett D. 2011. Characterization of the genome of the polyvalent lytic bacteriophage GTE2, which has potential for biocontrol of *Gordonia*-, *Rhodococcus*-, and *Nocardia*-stabilized foams in activated sludge plants. Appl Environ Microbiol 77:3923–3929. http://dx.doi.org/10.1128/AEM.00025-11.
18. Petrovski S, Seviour RJ, Tillett D. 2011. Prevention of *Gordonia*- and *Nocardia*-stabilized foam formation by using bacteriophage GTE7. Appl Environ Microbiol 77:7864–7867. http://dx.doi.org/10.1128/AEM.05692-11.
19. Petrovski S, Seviour RJ, Tillett D. 2013. Genome sequence and characterization of a *Rhodococcus equi* phage REQ1. Virus Genes 46:588–590. http://dx.doi.org/10.1007/s11262-013-0887-1.
20. Petrovski S, Seviour RJ, Tillett D. 2014. Genome sequence of the *Nocardia* bacteriophage NBR1. Arch Virol 159:167–173. http://dx.doi.org/10.1007/s00705-013-1799-z.
21. Petrovski S, Tillett D, Seviour RJ. 2012. Genome sequences and characterization of the related *Gordonia* phages GTE5 and GRU1 and their use as

potential biocontrol agents. Appl Environ Microbiol 78:42–47. http://dx.doi.org/10.1128/AEM.05584-11.
22. Hatfull GF, Jacobs-Sera D, Lawrence JG, Pope WH, Russell DA, Ko CC, Weber RJ, Patel MC, Germane KL, Edgar RH, Hoyte NN, Bowman CA, Tantoco AT, Paladin EC, Myers MS, Smith AL, Grace MS, Pham TT, O'Brien MB, Vogelsberger AM, Hryckowian AJ, Wynalek JL, Donis-Keller H, Bogel MW, Peebles CL, Cresawn SG, Hendrix RW. 2010. Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. J Mol Biol 397:119–143. http://dx.doi.org/10.1016/j.jmb.2010.01.011.
23. Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23:673–679. http://dx.doi.org/10.1093/bioinformatics/btm009.
24. Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32:11–16. http://dx.doi.org/10.1093/nar/gkh152.
25. Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33:W686–W689. http://dx.doi.org/10.1093/nar/gki366.
26. Cserzo M, Wallin E, Simon I, von Heijne G, Elofsson A. 1997. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng 10:673–676. http://dx.doi.org/10.1093/protein/10.6.673.
27. Petrovski S, Seviour RJ, Tillett D. 2013. Characterization and whole-genome sequences of the *Rhodococcus* bacteriophages RGL3 and RER2. Arch Virol 158:601–609. http://dx.doi.org/10.1007/s00705-012-1530-5.
28. Labrie SJ, Samson JE, Moineau S. 2010. Bacteriophage resistance mechanisms. Nat Rev Microbiol 8:317–327. http://dx.doi.org/10.1038/nrmicro2315.
29. Ackermann HW. 2003. Bacteriophage observations and evolution. Res Microbiol 154:245–251. http://dx.doi.org/10.1016/S0923-2508(03)00067-6.
30. Goodfellow M, Maldonado LA. 2006. The families *Dietziaceae*, *Gordoniaceae*, *Nocardiaceae* and *Tsukamurellaceae*, p 843–888. *In* Dworkin M, Stanley F, Rosenberg E, Schleifer KH, Stackerbrandt E (ed), The Prokaryotes, vol 3. Springer, New York, NY.
31. Enav H, Béjà O, Mandel-Gutfreund Y. 2012. Cyanophage tRNAs may have a role in cross-infectivity of oceanic *Prochlorococcus* and *Synechococcus* hosts. ISME J 6:619–628. http://dx.doi.org/10.1038/ismej.2011.146.
32. Bailly-Bechet M, Vergassola M, Rocha E. 2007. Causes for the intriguing presence of tRNAs in phages. Genome Res 17:1486–1495. http://dx.doi.org/10.1101/gr.6649807.
33. Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H. 2003. Prophage genomics. Microbiol Mol Biol Rev 67:238–276. http://dx.doi.org/10.1128/MMBR.67.2.238-276.2003.
34. Summer EJ, Liu M, Gill JJ, Grant M, Chan-Cortes TN, Ferguson L, Janes C, Lange K, Bertoli M, Moore C, Orchard RC, Cohen ND, Young R. 2011. Genomic and functional analyses of *Rhodococcus equi* phages ReqiPepy6, ReqiPoco6, ReqiPine5, and ReqiDocB7. Appl Environ Microbiol 77:669–683. http://dx.doi.org/10.1128/AEM.01952-10.
35. Catalano CE. 2000. The terminase enzyme from bacteriophage lambda: a DNA-packaging machine. Cell Mol Life Sci 57:128–148. http://dx.doi.org/10.1007/s000180050503.
36. Fujisawa H, Morita M. 1997. Phage DNA packaging. Genes Cells 2:537–545.
37. Xu J, Hendrix RW, Duda RL. 2004. Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. Mol Cell 16:11–21. http://dx.doi.org/10.1016/j.molcel.2004.09.006.
38. Daniel A, Bonnen PE, Fischetti VA. 2007. First complete genome sequence of two *Staphylococcus epidermidis* bacteriophages. J Bacteriol 189:2086–2100. http://dx.doi.org/10.1128/JB.01637-06.
39. Wang IN, Smith DL, Young R. 2000. Holins: the protein clocks of bacteriophage infections. Annu Rev Microbiol 54:799–825. http://dx.doi.org/10.1146/annurev.micro.54.1.799.
40. Chun J, Goodfellow M. 1995. A phylogenetic analysis of the genus *Nocardia* with 16S rRNA gene sequences. Int J Syst Bacteriol 45:240–245. http://dx.doi.org/10.1099/00207713-45-2-240.
41. Goodfellow M, Zakrzewska-Czerwinska J, Thomas EG, Mordarski M, Ward AC, James AL. 1991. Polyphasic taxonomic study of the genera *Gordonia* and *Tsukamurella* including the description of *Tsukamurella wratislaviensis* sp.nov. Zentralbl Bakteriol 275:162–178. http://dx.doi.org/10.1016/S0934-8840(11)80063-0.
42. Yassin AF, Rainey FA, Burghardt J, Brzezinka H, Schmitt S, Seifert P, Zimmermann O, Mauch H, Gierth D, Lux I, Schaal KP. 1997. *Tsukamurella tyrosinosolvens* sp. nov. Int J Syst Bacteriol 47:607–614. http://dx.doi.org/10.1099/00207713-47-3-607.
43. Lesnik EA, Sampath R, Levene HB, Henderson TJ, McNeil JA, Ecker DJ. 2001. Prediction of rho-independent transcriptional terminators in *Escherichia coli*. Nucleic Acids Res 29:3583–3594. http://dx.doi.org/10.1093/nar/29.17.3583.
44. Mott ML, Berger JM. 2007. DNA replication initiation: mechanisms and regulation in bacteria. Nat Rev Microbiol 5:343–354. http://dx.doi.org/10.1038/nrmicro1640.

PLOS | ONE

RESEARCH ARTICLE

Lysis to Kill: Evaluation of the Lytic Abilities, and Genomics of Nine Bacteriophages Infective for *Gordonia* spp. and Their Potential Use in Activated Sludge Foam Biocontrol

Zoe A. Dyson[1], Joseph Tucci[1], Robert J. Seviour[2], Steve Petrovski[2]*

1 La Trobe Institute of Molecular Sciences, Bundoora, Victoria, Australia, **2** Department of Physiology, Anatomy and Microbiology, La Trobe University, Bundoora, Victoria, Australia

* steve.petrovski@latrobe.edu.au

OPEN ACCESS

Citation: Dyson ZA, Tucci J, Seviour RJ, Petrovski S (2015) Lysis to Kill: Evaluation of the Lytic Abilities, and Genomics of Nine Bacteriophages Infective for *Gordonia* spp. and Their Potential Use in Activated Sludge Foam Biocontrol. PLoS ONE 10(8): e0134512. doi:10.1371/journal.pone.0134512

Editor: Raymond Schuch, ContraFect Corporation, UNITED STATES

Received: April 24, 2015

Accepted: July 9, 2015

Published: August 4, 2015

Data Availability Statement: The nucleotide sequence for all phages have been deposited GenBank under the following accession numbers; GTE6 (KR053200), GTE8 (KR053201), GMA2 (KR063281), GMA3 (KR063279), GMA4 (KR053199), GMA5 (KR053198), GMA6 (KR063280), GMA7 (KR063278), and GRU3 (KR053197).

Funding: This work was supported by La Trobe University.

Abstract

Nine bacteriophages (phages) infective for members of the genus *Gordonia* were isolated from wastewater and other natural water environments using standard enrichment techniques. The majority were broad host range phages targeting more than one *Gordonia* species. When their genomes were sequenced, they all emerged as double stranded DNA *Siphoviridae* phages, ranging from 17,562 to 103,424 bp in size, and containing between 27 and 127 genes, many of which were detailed for the first time. Many of these phage genomes diverged from the expected modular genome architecture of other characterized *Siphoviridae* phages and contained unusual lysis gene arrangements. Whole genome sequencing also revealed that infection with lytic phages does not appear to prevent spontaneous prophage induction in *Gordonia malaquae* lysogen strain BEN700. TEM sample preparation techniques were developed to view both attachment and replication stages of phage infection.

Introduction

Many isolates of members of the actinobacterial genus *Gordonia* have been cultured from wastewater treatment plants [1] where they probably play a key role in degrading the more recalcitrant influent substrates [2, 3]. They include *Gordonia amarae*, an organism with a characteristic right-angled branching morphology, and among the first foam forming bacteria isolated and cultured [1, 4, 5]. Other *Gordonia* species and members of closely related genera share this distinctive morphology, and so in the absence of more precise identification, those with it are commonly referred to as *Gordonia amarae*-like organisms, or GALO [1].

Members of the *Corynebacteriales*, which include *Gordonia*, *Nocardia*, *Rhodococcus*, *Tsukamurella* and *Mycobacterium*, are often referred to collectively as the Mycolata because they

Competing Interests: The authors have declared that no competing interests exist.

alone synthesize long chain hydroxylated mycolic acids, organized as an exocellular outer membrane [6]. Their presence renders these cells highly hydrophobic. In activated sludge, high levels of these Mycolata stabilize foams formed on the surface of aeration tanks and clarifiers [1]. Formation of these stable foams is a global problem that impacts negatively on plant aesthetics, increases maintenance costs, and complicates sludge management [7]. Some of the Mycolata in these foams are opportunistic pathogens, thus posing a potential health hazard to plant operators from their aerosol dispersal [1, 7, 8].

Formation of these stable foams requires air bubbles, surface active agents, and hydrophobic particles, in this case the Mycolata cells [9]. A successful control strategy must be directed at the hydrophobic bacteria because neither air bubbles nor detergents can be eliminated from the activated sludge process [9]. Current foam control strategies are not effective universally, and no single method reliably controls all foams. This probably is a reflection of how little is known about the microbial ecology of these causative bacteria [1]. It was Thomas et al., [10] who first proposed that phage therapy could be exploited using the natural lytic cycles of phages as an attractive and environmentally friendly approach to selectively control their population levels without affecting other desirable bacteria in these systems.

Currently (February 2015) 228 phages targeting members of the genus *Mycobacterium* have had their genomes sequenced, and only four lytic *Gordonia* phage genome sequences are available. These are phages GTE2 [11], GTE7 [12], GRU1 and GTE5 [13]. All *Gordonia* phages isolated so far have distinctive genome sequences [11–13]. Yet with such a small sample size, it is not sensible to comment on the general characteristics of *Gordonia* phages and draw conclusions from these as to their suitability or otherwise for foam bio-control. Therefore, more *Gordonia* lytic phages are needed, including those from habitats other than activated sludge plants.

This study set out to increase the small existing library of *Gordonia* phages, and to characterize them in terms of their host ranges, morphologies, and genomics. Nine phages infective for members of this genus were isolated and their suitability for use in *Gordonia* foam biological control was investigated.

Materials and Methods

No specific permission was required for sample collection from the water locations sampled as all samples were publically available for researchers to collect. All fieldwork conducted in this study did not involve endangered or protected species.

Isolation and preliminary characterization of *Gordonia* phages

Host strains held in the La Trobe University culture collection used in this study, and the techniques for their growth are those detailed by Petrovski et al., [14], together with those listed in Table 1, which were grown in the same manner. All phages were isolated and subsequently purified from water samples collected from a variety of locations using enrichment pools of multiple host strains, as shown in Table 1 and described previously [14]. Phage host range specificity determinations were also carried out as described by [14].

Transmission electron microscopy of virion morphology

Grids for visualization of virions were prepared with the negative stain uranyl acetate [14]. Both carbon and formvar coated grids were used (Electron Microscopy Sciences, Australia), with the exception of phage GTE6 which was examined on grids coated with formvar only. Prepared grids were subsequently examined with a JEOL JEM-100CX, JEOL JEM-2010HC, or a Tenaci Fei T30 Transmission Electron Microscope (Table 2).

Table 1. Isolation and characterization of nine *Gordonia* phage.

Phage	Sample	Strain	Lab ID	Enrichment pool members	Host range
GMA2	Activated sludge, Kyneton, Victoria, Australia	*G. malaqaue*	A448	See GTE8	*G. terrae* (CON34, GOR9, G238), *G. malaquae* (CON59, CON60, A554, A448), *G. hydrophobica* (CON65, CON66)
GMA3	Wastewater, Glenelg, South Australia, Australia	*G. malaquae*	BEN700	See GTE8	*G. terrae* (G238), *G. malaquae* (BEN700)
GMA4	Puddle water and sediment, Reservoir, Victoria, Australia	*G. malaquae*	BEN700	See GTE8	*G. malaquae* (BEN700)
GMA5	Activated sludge, Carrum (Eastern Treatment Plant), Victoria, Australia	*G. malaquae*	BEN700	See GTE8	*G. rubropertincta* (CON38), *G. terrae* (G238, G232), *G. malaquae* (BEN700)
GMA6	Activated sludge, Bendigo, Victoria, Australia	*G. malaquae*	CON67	See GTE8	*G. malaquae* (CON59, CON60, CON67, A554, A448, BEN700), *G. terrae* (G238)
GMA7	Activated sludge, Werribee, Victoria, Australia	*G. malaquae*	CON60	*G. terrae* (GOR9, G232, G238), *G. malaquae* (A554, A448, CON60, BEN700), *T. paurometabola* (CON61)	*G. terrae* (CON34, GOR9, G238), *G. rubropertincta* (CON38), *G. malaquae* (CON59, CON60, A554, A448, BEN700), *G. hydrophobica* (CON65, CON66)
GRU3	Wastewater, Inverell, Queensland, Australia	*G. rubropertincta*	CON38	See GTE8	*G. rubropertincta* (CON38), *G. terrae* (GOR9, G232)
GTE6	Activated sludge, Nambour, Queensland, Australia	*G. terrae*	CON34	*G. terrae* (CON34, BEN601, BEN604), *G. sputi* (CON48, CON49), *G. amarae* (CON44, CON9)	*G. terrae* (CON34, GOR9), *G. malaquae* (CON59, CON60, A554, A448), *G. hydrophobica* (CON65, CON66)
GTE8	Bendigo creek water, Bendigo, Victoria, Australia	*G. terrae*	G232	*G. terrae* (CON34, G238, G290, G255, G232, GOR9), *G. sputi* (CON48, CON49), *G. amarae* (CON44, CON9), *G. hydrophobica* (CON65, CON66), *G. desulfuricans* (CON69), *G. polyisoprenovorans* (CON71), *G. alkanivorans* (CON72), *G. malaquae* (A554, A448, BEN700, CON67), *T. inchonensis* (BEN701), *R. erythropolis* (BEN703) *G. aicheiensis* (CON22)	*N. asteroids* (CON12), *G. terrae* (CON34, GOR9, G232), *G. rubropertincta* (CON38)

doi:10.1371/journal.pone.0134512.t001

Table 2. *Gordonia* phage virion measurements.

Phage name	Capsid diameter (nm)	Tail length (nm)
GMA2[b]	61 ± 4	386 ± 3
GMA4[b]	54 ± 2	244 ± 2
GMA5[b]	37 ± 2	85 ± 9
GMA6[b]	62 ± 2	143 ± 7
GMA7[c]	63 ± 3	474 ± 9
GRU3[b]	43 ± 2	93 ± 10
GTE6[a]	48 ± 8	152 ± 12
GTE8[b]	56 ± 2	239 ± 12

[a] electron micrographs obtained using a JEOL JEM-100CX,
[b] electron micrographs obtained using a Tenaci Fei T30,
[c] electron micrographs obtained using a JEOL JEM-2010HC.

doi:10.1371/journal.pone.0134512.t002

Transmission Electron Microscopy to show phage infection

To visualize phage attachment, a single colony of *Gordonia malaquae* (CON67) was taken from a streak plate incubated at 30°C for 3 days. The cells were added to 20 μL of high titer GMA6 phage lysate ($>10^{10}$ PFU/mL), and left to stand for 10 min to allow attachment before they were adsorbed onto the surface of carbon/formvar coated 200 mesh copper grids (Electron Microscopy Sciences, Australia). Grids were washed twice in sterile double-distilled water (ddH_2O), and then negatively stained with 2% (w/v) uranyl acetate for 2 min. Excess liquid was absorbed onto filter paper and the grid was allowed to air dry. These grids were then examined under a JEOL JEM-2010HC Electron Microscope.

For phage assembly, a 1 mL aliquot of a *Gordonia terrae* (CON34) culture incubated at 30°C for three days in PYCA broth was removed carefully and added to 20 mL of PYCa broth together with 100 μL of high titer phage GTE6 suspension ($>10^{10}$). This mixture was allowed to stand for 10 mins before further incubation at 30°C for three days. A 1.5 mL aliquot was then centrifuged (3,000 x g for 30 min) and the supernatant discarded. Pelleted cells were re-suspended and fixed in 2.5% (v/v) glutaraldehyde in 0.1 M phosphate buffer (pH 6.8–7.3), and incubated at 4°C overnight, then harvested (14,000 x g for 5 min) and washed in the same phosphate buffer 3 times, with 10 min between washes. Cells were post-fixed in 1% osmium tetroxide in 0.1 M phosphate buffer for 90 min and washed three times in sterile ddH_2O. They were then dehydrated through an acetone series of increasing concentrations (30%, 50%, 70%, 90% and 100%) for 10 min each, prior to a final washing with 100% acetone with a molecular sieve (ProSciTech, Australia) (10 min). Dried specimens were infiltrated with Spurr's epoxy resin (ProSciTech, Australia), initially with 50% resin, 50% dehydrated acetone, and incubated overnight at room temperature. The mixture was replaced by 100% Spurr's resin with a further incubation of 1–2 h, before finally being replaced by fresh Spurr's expoxy resin, and polymerised at 65°C overnight. Thin sections (100 nm) were cut with a glass knife on an LKB Microtome and post-stained with uranyl acetate and lead citrate. Sections were placed on 200 mesh copper grids and examined as described above.

Mass spectroscopy

To identify phage structural proteins, purified virions $>10^{13}$ PFU/mL were precipitated with $(NH_4)_2SO_4$ followed with exposure to $ZnCl_2$ to remove any residual polyethylene glycol from the previous step. Pellets were re-suspended in 8 M urea to a final volume of 100 μL prior to transfer to the Mass Spectroscopy and Proteomics facility at the La Trobe University Institute of Molecular Sciences. Here peptides reconstituted in 0.1% formic acid and 2% acetonitrile (buffer A) were loaded onto a trap column (C18 PepMap 300 μm i.d. × 2 cm trapping column, Thermo-Fisher Scientific) at 5 μL/min for 6 min and washed for 6 min before switching the precolumn in line with the analytical column (Vydac MS C18, 75 μm i.d. × 25 cm, Grace Davison). The separation of peptides was performed at 300 nL/min using a linear acetonitrile (ACN) gradient of buffer A and buffer B (0.1% formic acid, 80% ACN), starting from 5% buffer B to 40% over 60 min. Data were collected on an hybrid quadrupole/time-of-flight MS (Micro-TOF-Q, Bruker, Germany) with a nano-electrospray ion source using Data Dependent Acquisition mode and *m/z* 150–2500 as MS scan range. Nitrogen was used as the collision gas. The ionisation tip voltage and interface temperature were set at 4200 V and 205°C respectively. Collision Induced Dissociation (CID) MS/MS spectra were collected for the 3 most intense ions. Dynamic exclusion parameters were set as follows: repeat count 2, duration 60 s. The data were collected and analysed using Data Analysis Software (Bruker Daltonics, Bremen, Germany).

Genome sequencing of *Gordonia* phages

Genomic DNA was extracted from phages GTE6, GMA2, and GMA6 and sequenced using a Roche GS FLX genome sequencer and titanium chemistry, as described in Petrovski, Seviour [14]. Genomic DNA extracted from all other phages in the same manner was prepared with an Illumina Nextera XT sample preparation kit as per manufacturers' instructions. The prepared DNA libraries were sequenced on an Illumina MiSeq as a 150 bp paired end run.

Genome annotation

The genome open reading frames (ORFs) were screened initially using Glimmer (v3.02), where ORFs with a minimum size of 90 bp were detected [15]. All predicted start codons were inspected for the presence of putative ribosomal binding sites and corrected as necessary. Sequence similarity searches were carried out against the GenBank database, as described by Petrovski et al. [11]. The presence of tRNA and tmRNA were also determined using both ARAGORN [16], and tRNAScan-SE [17]. Transmembrane domains were predicted with the DAS Transmembrane Prediction server [18].

Phage DNA when analyzed by gel electrophoresis gave results consistent with circularly permuted DNA genomes. Therefore, for consistency the genomes annotations were conducted starting with the DNA packaging operon.

Nucleotide sequence accession number

The nucleotide sequences for all phages have been deposited GenBank under the following accession numbers; GTE6 (KR053200), GTE8 (KR053201), GMA2 (KR063281), GMA3 (KR063279), GMA4 (KR053199), GMA5 (KR053198), GMA6 (KR063280), GMA7 (KR063278), and GRU3 (KR053197).

Results and Discussion

Phage isolation and host range characterization

All *Gordonia* phages isolated previously were obtained from wastewater, with most coming from activated sludge plants on the east coast of Australia [10–13]. While most phage isolates described here were also from wastewater (Table 1), an additional two phages GMA4 and GTE8, were obtained from puddle sediment (Reservoir, Australia), and creek water (Bendigo, Australia), respectively.

One of these, phage GMA4, lysed a single Mycolata strain, *Gordonia malaquae* strain BEN700. Phage GMA3 lysed the same *G. malaquae* strain (BEN700), but, also *G. terrae* (G238). All the other phage's lysed multiple *Gordonia* strains, with phage GMA7 attacking 11 strains from four different *Gordonia* species i.e. *G. terrae* (CON34, GOR9, G238), *G. rubropertincta* (CON38), *G. malaquae* (CON59, CON60, A554, A448, BEN700), and *G. hydrophobica* (CON65, CON66). As well as phage GTE8 lysing three strains of *Gordonia terrae* (CON34, GOR9, G232) and one of *G. rubropertincta* (CON38), it could also lyse *Nocardia asteroides* (CON12).

Phages able to lyse members of both these genera have been reported before. They include phage GRU1, which targets *Nocardia nova* strain CON47 and *Gordonia terrae* strains CON34, and G232 and also *Gordonia rubropertincta* strain CON38 [13]. This outcome might reflect the close phylogenetic relationship of these host bacteria.

Many of these overlap in their host ranges. For example phages GTE6, GMA2, GMA6, and GMA7 all lysed the same four strains of *G. malaquae* (CON59, CON60, A554, and A448), a property which might make them useful additions to any phage cocktail designed to target

foaming caused by *G. malaquae*, especially if they use different host receptor sites. Phage GMA5 was lytic against two of eight *G. terrae* strains (G238, G232). A similar situation has been reported for other *Gordonia* phages, including GTE2 [11] that lysed only one of five *G. terrae* strains. No phages infective for *G. amarae* were obtained in this study.

Virion morphology

All phages examined by TEM had both the isometric type B1 capsids (~ 37 to ~ 63 nm in diameter) and long, non-contractile tails (~ 85 to ~ 474 nm long) characteristic of members of the family *Siphoviridae*. Phage GMA3 was not examined by TEM, but based on its genome sequence which contained a gene encoding a long tape measure protein and its dsDNA genome, it too is most likely to be a member of the *Siphoviridae* [19]. Further details are provided in Fig 1, and Table 2. With TEM, the morphology of phage GMA6 was not that expected of a *Siphoviridae* member, since its tail appeared to be uncharacteristically thick and rigid (Fig 1). To resolve this concern, phage GMA6 virions were exposed briefly to *G. malaquae* strain CON67, it's isolating host, and then examined by TEM. Images showed clearly (Fig 2a) that its phage tail can be flexible, confirming it as a *Siphoviridae* member. Furthermore, TEM (Fig 2a) shows that virion attachment can involve many phages simultaneously. Whether superinfections where more than one phage genome successfully invades the host cell, was not explored. We could also visualize post replication mature phage progeny within the host cell, thus Fig 2b shows mature GTE6 virions inside the host cells, prior to cell lysis and release of phage progeny.

Structure and organization of *Gordonia* phages genomes

When the assembled genomes for all nine *Gordonia* phages were examined, they were in most cases distinctively different to each other. Genome sizes ranged from 17, 562 to 103, 424 bp, and they contained between 27 to 127 putative *orfs*, arranged mostly in the modular architecture commonly seen in the *Siphoviridae* phages (Table 3, Fig 3). All contained putative genes orientated in both forward and reverse directions, with the one exception being phage GTE6, where all its genes were in a forward orientation. Only between 22 and 50% of the putative genes identified in the nine *Gordonia* phage genomes could be annotated functionally (Table 3, S1 Table). The G+C mol % contents of all phages ranged from 51.3 to 67.8 mol % (Table 3) and for the majority this value was close to that of the corresponding host cells [6].

Most of these phage genomes contained no putative tRNAs, and no tmRNA could be identified in any of them (Table 3). Of those phages where tRNAs were seen, phage GMA7 contained 1 putative tRNA-Asn, GMA4 contained 1 putative tRNA-Trp, and phage GMA2 contained a tRNA cluster of 57, 883 to 60, 154 bp where 16 putative tRNA were located (S2 Table). Such tRNA clusters have been observed previously in *Mycobacterium* phages where they appear to be important in late lytic growth, where they may compensate for degradation and inadequacy of host tRNA [20].

The assembled DNA sequence of all nine phages were compared to one another and to other sequences deposited in public databases. Two phage's isolated in this study, GMA4 and GMA5, were 77% homologous, suggesting a close evolutionary relationship. The genome sequence of GMA7 shared a 97% homology to the DNA sequence in phage GTE7, a polyvalent *Gordonia* phage [12]. Similarly, phage GTE8 shared a lower level of homology (81% and 83% respectively) to two closely related *Gordonia* phages GTE5 and GRU1 described previously [13]. The remainder of the genomes were substantially different to all other phage sequences deposited in GenBank.

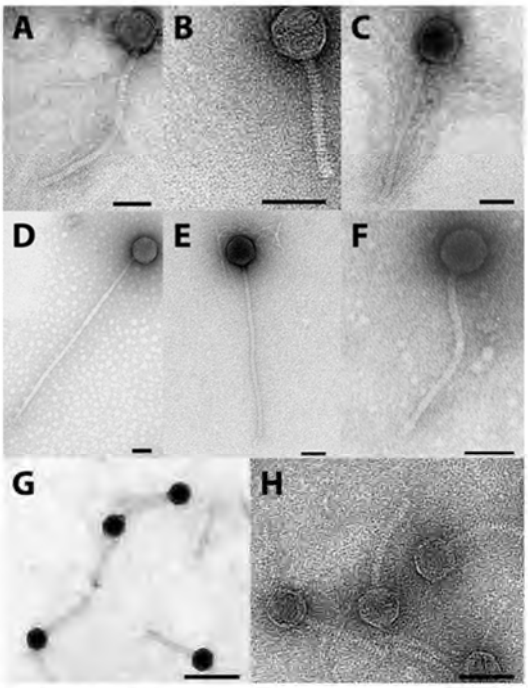

Fig 1. TEM *Gordonia* phage morphologies. (A) Phage GMA4 (B) Phage GRU3 (C) Phage GTE8 (D) Phage GMA7 (E) Phage GMA2 (F) Phage GTE6 (G) Phage GMA6 (H) Phage GMA5. Scale = 50 nm.

doi:10.1371/journal.pone.0134512.g001

Evidence for potential spontaneous prophage induction events

Whole genome sequencing using next generation DNA sequencing technology allows a high level of genome coverage. During the phage isolation procedure, GMA3, GMA4 and GMA5 were isolated in the same host, *G. malaquae* (BEN700). After multiple rounds of purification, the phages were grown to high titre and DNA was isolated for sequencing. Upon sequencing and assembly, it was clear that the phage contigs obtained had >1200-fold coverage and a smaller contaminating contig appears in all three isolations of approximately 41 kb with a lower coverage, 17-fold to 227-fold (Table 4). PCR analysis of the genomic DNA of *G. malaquae* (Ben700) revealed the contaminating contig was present in the host presumably as a prophage and therefore named GMA1.

This observation suggests that this putative lysogenic *G. malaquae* strain could tolerate co-infection with these three phages, as well as the previously described GTE2 phage [11]. Whether these phages interact while co-infecting is unknown.

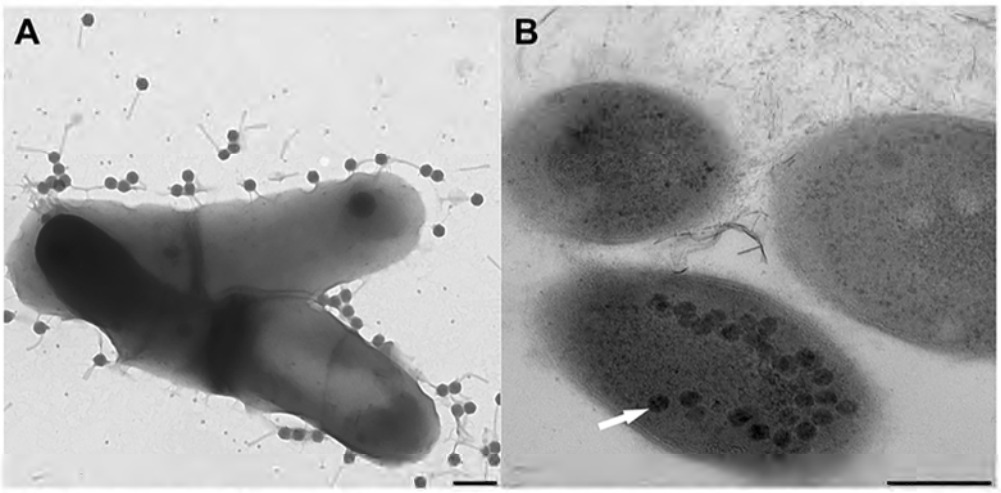

Fig 2. Stages in *Gordonia* phage infection cycles. (A) Attachment stage of phage infection cycle between phage GMA6 and host *Gordonia malaquae* strain CON67. Scale = 200 nm. (B) Replication of phage GTE6 inside *G. terrae* strain CON34 cells prior to cell lysis. Arrows indicate phage replicated inside bacterial cells. Scale = 200 nm.

doi:10.1371/journal.pone.0134512.g002

Sequence repeats in *Gordonia* phage genomes

Repeat structures have been reported previously in genome sequences of several related phages [13–14]. All nine phage genomes contain between 1 to 18 palindromic sequences of between 14 to 98 bp in length (S3 Table). Some of these are located in what appear to be intergenic areas, which might support their roles as putative *rho*-independent transcriptional terminators [21]. They also contained 14 to 252 direct repeats ranging in length between 14 and 425 bp (S4 Table). Also seen in these genomes were 3 to 87 inverted repeats of 56 to 14 bp long (S4 Table). Inverted repeats may indicate replication origins and transposable elements [22], but neither of these could be identified in any of these phages, and so their roles remain unknown.

Gordonia phage DNA packaging modules

In *Siphoviridae* phage genomes, the large terminase subunit protein usually functions in a complex with a small terminase subunit, and together these act to mediate cleavage of the phage DNA at specific sites prior to packaging into the prohead [23–24]. The gene encoding the large terminase subunit was identified in all nine *Gordonia* phages examined here by either amino acid sequence homology to other known terminase genes (GMA4, GMA3, GMA5, GMA7, and GTE8) or the presence of the diagnostic conserved motif. In *Siphoviridae* phages the small terminase gene is typically located upstream, and is transcribed in the same direction as the large terminase [23–24]. In all phages except GMA2 this pattern could be recognized, and in some cases supported by amino acid sequence homology to other known small terminases (S1 Table).

Table 3. Summary of characters of the nine *Gordonia* phage genomes.

Phage name	Average coverage (fold)	Total read count	Genome size (bp)	G+C content (mol %)	No. putative tRNA	No. putative genes	No. putative genes in forwards orientation	No. functionally annotated putative genes	No. novel genes	No. palindromes	No. direct repeats	No. inverted repeats
GMA2 [a c]	1, 212	336, 750	103, 424	53.4	16	126	42	42	62	7	22	10
GMA3 [b e d]	1, 200	677, 981	77, 779	51.3	0	104	32	27	47	16	18	8
GMA4 [b e d]	1, 981	716, 641	45, 537	66.4	1	68	61	22	11	6	40	31
GMA5 [b f d]	6, 793	930, 480	17, 562	66.4	0	28	24	14	4	11	28	13
GMA6 [a c]	247	55, 269	83, 324	58.2	0	115	109	38	68	1	20	3
GMA7 [b c]	1, 603	947, 843	73, 419	56.6	1	101	32	23	5	18	14	5
GRU3 [b e d]	520	89, 131	17, 727	66.5	0	26	23	12	6	3	42	16
GTE6 [a c]	915	141, 321	56, 982	67.8	0	86	86	23	49	3	252	87
GTE8 [b c]	1, 605	777, 336	67, 617	66.0	0	94	67	23	22	5	48	36

[a] sequenced using 454,

[b] sequenced using Illumina,

[c] reads assembled using CLC workbench (v6.5.1),

[d] reads assembled using CLC workbench (v7.5.1),

[e] reads assembled using Spades (v3.1.0),

[f] reads assembled using ABySS (v1.3.7).

doi:10.1371/journal.pone.0134512.t003

Gordonia phage structural protein genes and their proteomics

Phage structural protein genes are located typically adjacent to the DNA packaging module, usually beginning with head morphogenesis genes, followed by tail morphogenesis genes [19]. Some departures from this gene arrangement were seen in these nine phages. For example, in GMA6, only one (*orf8*) of the genes identified between the terminase genes (*orf2* and *orf3*) and the putative portal protein gene (*orf11*), could be assigned a putative function in encoding a nucleoside triphosphate pyrophosphohydrolase. This interpretation was based on the presence of the cd11541 motif. Seemingly involved in DNA maintenance, it was seen between the structural and packaging modules, a location different to the typical modulated genome architecture of *Siphoviridae* phages, where all genes of similar function are clustered together [25]. Furthermore, *orf14*, within the structural gene module of GMA6, appears to encode a HNH endonuclease based on its amino acid sequence homology to the diagnostic pfam01844 motif. Gene arrangements in phage GMA2 suggest that *orf21*, encoding a putative DNA methyltransferase is in the structural gene module.

In all nine phages the tape measure proteins were encoded by their longest gene, which is usual in *Siphoviridae* phages [19] (S1 Table, Fig 3). The only exception was in phage GRU3 where *orf6* encoding a putative phage head protein was slightly larger in size (657 amino acids) than *orf12* encoding its tape measure protein, 622 amino acids long. In most of these (GMA2, GMA3, GMA4, GMA6, GMA7, and GTE8), the two genes preceding that encoding the tape

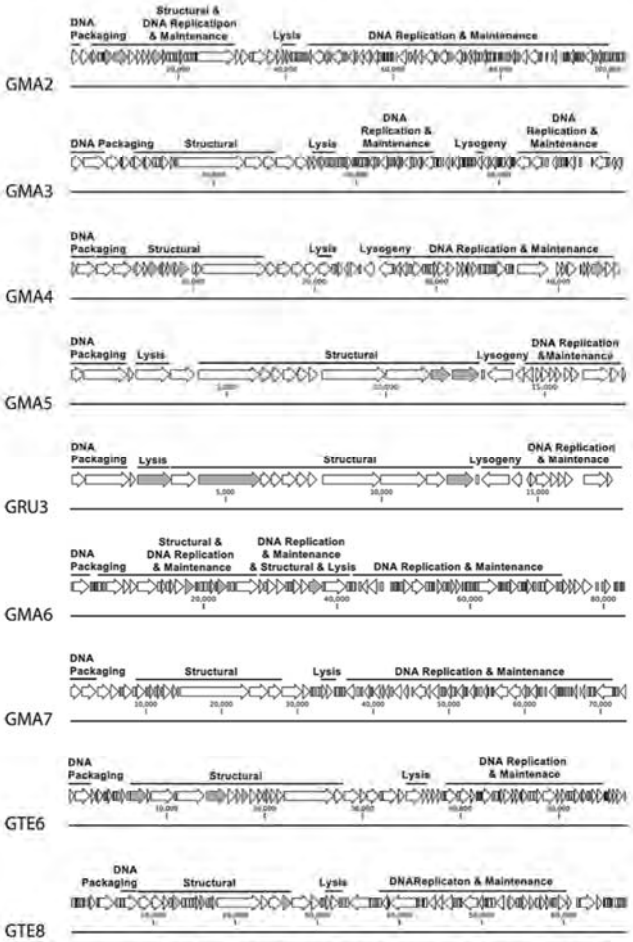

Fig 3. Genome map of nine *Gordonia* phages. Grey indicates structural genes identified with mass spectroscopy data.

doi:10.1371/journal.pone.0134512.g003

Table 4. Coverage of phage GMA1 in the assemblies of phages GMA3, GMA4 and GMA5.

Phage Sequenced	Length of contaminating contig	Average coverage	Total reads
GMA3	41,106	17-fold	5,097
GMA4	40,897	227-fold	70,089
GMA5	41,106	32-fold	10,135

doi:10.1371/journal.pone.0134512.t004

measure protein were identified as encoding putative tail assembly proteins, where the latter appeared to be translated using a conserved programmed frameshift, a common feature of *Siphoviridae* phages [26]. Usually the gene immediately upstream of these is that encoding the putative major tail protein [26].

Mass spectrometry data (S5 Table) seemed to suggest several structural genes are located outside the structural gene module. For example, in phage GMA4 *orf66* was located in the DNA replication module, with a translated protein sequence homologous to a hypothetical protein from *Aeromicrobium marinum*, but also a motif for a phage tail fiber protein (COG5301). Similarly, in GMA6, *orf43* was located within the DNA replication gene modules.

Gordonia phage lysis gene modules are diverse

Lysin genes were identified in all nine *Gordonia* phages, but their locations and numbers varied, and as with many already discussed, they often appeared to disrupt the usual and expected modular genome architecture of *Siphoviridae* phages. Phages GMA5 and GRU3 contain a D-alanyl-D-alanine carboxypeptidase encoding gene (*orf5* in both) showing amino acid sequence homology to a hypothetical protein in *Gordonia soli*, located within what appears to be the phage structural module (S1 Table, Fig 3). A phage lysin motif within the structural gene module was reported for *Rhodococcus* phage RRH1, suggesting this is not an uncommon occurrence [27, 28].

In the genomes of GMA7 and GTE6 their lysin genes were adjacent to their structural protein encoding genes (*orf28* –*orf29* and *orf38*, respectively), and unusually, both had additional lysin genes in their DNA replication gene modules (*orf41* and *orf58*, respectively). The same pattern was reported for phage GTE7, [12], to which GMA7 is genetically similar at a nucleotide level (97% identity, 95% coverage).

Phages GMA2, GMA3, and GMA6 also had unusual lysin gene arrangements, with higher numbers of such genes than the more common lysin A and B arrangement [29]. Phage GMA2 unusually possessed four putative lysin genes (*orf35* –*orf38*), identified from their amino acid homologies and presence of the diagnostic pfam13529 (peptidase), pfam01510 (N-acetylmuramoyl-L-alanine amidase), and cd02619 (peptidase) motifs in Orf35, Orf36, and Orf38, respectively. Phage GMA6 also had four lysin genes (*orf34*, *orf37*, *orf40*, and *orf45*), many of which were separated by genes associated with DNA replication/maintenance and virion morphogenesis. A similar pattern was seen in phage GMA3, which contained three putative lysin genes (*orf22*, *orf24*, and *orf26*) separated by a putative nuclease gene (*orf25*), again associated with DNA replication/maintenance (S1 Table, Fig 3).

Orf45 in GMA6 is a more complex lysin gene than those seen in all other actinophages. It alone encodes an unusually high number of different lysin motifs. These include an N-terminal BacA motif of a bacterial lysin from *Enterococcus faecalis* (cd06418), an N-acetylmuramoyl-L-alanine amidase motif (pfam01510) downstream of this, a peptidase motif (pfam01551) further downstream, and an additional C-terminal motif (pfam13810) of unknown function (S1 Table).

Holins could not be identified in phages GMA4, GMA5, and GRU3 by nucleotide or amino acid sequence homologies, nor by the criteria of Wang et al. [30], which state that expressed products should be less than 150 amino acid residues and contain two or more transmembrane regions. If holins are present in these two phages, it would seem that their genes are novel in their locations and/or translated amino acid sequences.

DNA replication/maintenance genes

DNA replication modules in all other actinophage genomes sequenced so far are arranged in a modular architecture, where genes functioning in DNA replication/maintenance are located

adjacent to lysin genes [11–14, 27, 28, 31, 32]. In GMA4, GMA7, and GMA2 phages, this region contains putative DNA-methylase encoding genes, of these GMA2 appears to possess at least two (*orf21* and *orf51*). If functional, they may play a role in protecting their DNA from host cell restriction attack [33]. Such enzymes have been identified in other *Gordonia* phages including GTE7 phages [12]. Metagenomic studies by Tamaki et al. [34] have suggested that methylase genes are more prevalent in phages within the activated sludge habitat from where most actinophages have come. Glycosyltransferase encoding genes are also seen in many phage genomes [35], including that of GMA2 (*orf4* and *orf22*), and all appear to have similar functions to phage methylases where they help protect phage DNA from digestion with restriction endonucleases from host RM systems [33]. These genes can also have other functions including serotype conversion in temperate phages [35], and so their purpose here remains unclear.

Lysogeny and lysogenic conversion genes

Genomes of phages GMA3, GMA4, GMA5, and GRU3 all contain putative genes that are homologues of phage integrase genes (*orf75*, *orf29*, *orf17*, and *orf17* respectively), based on their product amino acid sequence similarities to those of known phage integrases, and the possession of the integrase specific motif pfam00589. If functional, their presence suggests the capability for a lysogenic lifecycle as well as a lytic one.

The GMA4 genome appears to encode several moron genes that may confer a selective advantage to its host. For example, it possesses a gene associated with phage resistance (*orf34*) [36]. The N-terminal region of Orf34 contains a Rha motif (pfam09669), thought to interfere with further phage infection of bacterial host strains lacking the integration host factor (IHF) [36]. It regulates expression of the *rha* gene, and so may confer resistance to further phage attack of any bacterial host infected by it in a lysogenic cycle [36].

Unexpected features of the *Gordonia* phages

As mentioned, most of the nine *Gordonia* phages sequenced in this study had highly distinctive genomes, with high percentages of ORFans (5 to 59%) (S1 Table), for which no statistically significant identifications could be made against sequences held in GenBank. Yet their genes encode motifs suggestive of their putative function. For example, both GMA2 and GMA6 possess a cd00233 motif in their Orf14 and Orf13 putative proteins, respectively. The *orf14* and *orf13* genes appear to encode a VIP2 family actin-ADP-ribosylating toxin with high specificity against the insect pest corn rootworms, and sharing a statistically significant sequence similarity with enzymatic components of other binary toxins, including the *Clostridium botulinum* C2 toxin, *C. perfringens* iota toxin, *C. piroforme* toxin, *C. piroforme* toxin and *C. difficile* toxin [37].

Furthermore, phage GTE6 genome appeared to contain a gene (*orf12*), encoding a host cell surface-exposed lipoprotein since its expressed amino acid sequence shares homology with the pfram07553 motif. Such motifs are usually involved in superinfection exclusion, acting at the stage of DNA release from the phage head into the host cell. These motifs have been associated with Superinfection exclusion (Sie) systems in temperate phages, where they interfere with co-infections involving other phages [33, 38, 39]. Presence of such a motif in what appears to be an obligatory lytic phage is unexpected. Equally unexpected is that *orf21* in phage GTE6 encodes a putative Eppstein-Barr nuclear antigen (Orf21), showing 35% amino acid sequence similarity to that of *Saccharomonospora* phage PIS 136 in this region [40]. Whether this homology reflects a similar function for the pair, or a distant evolutionary relationship between them, is unknown. The *orf4* of phage GMA7 also appears to encode an unexpected motif (cd12820) normally associated with a putative adhesion virulence factor, forming a matrix on the bacterial outer membrane, which mediates binding to collagen and epithelial cells [41].

Evolutionary ancestry of *Gordonia* phages

From the data presented here, it is clear that phages GTE8, isolated from creek water and GMA7, from activated sludge are genetically very similar to phages GTE5/GRU1, and GTE7, respectively. It is reasonable to propose that these similarities reflect a closely shared ancestral past. Similar comments apply to phage GMA5 and GRU3. Despite not sharing nucleotide sequence identity with phage GTE7 DNA, the expressed amino acid sequences of phage GMA3 expressed genes are highly similar to it. Nine genes of GMA3 were most similar to those from GTE7, while 23 other genes were most similar to those from phage ReqiDocB7, to which GTE7 genome is closely related at an amino acid level [12, 42]. As a similar closeness was not reflected at the nucleotide sequence level, one suggestion might be that more distant evolutionary relationships exist between GTE7, ReqiDocB7, and GMA3. GMA3 contains a gene showing homology to a putative RNA-binding gene from the chimpanzee *Pan troglodytes* (*orf91*), and a centromere protein F-like gene from the banana plant *Musa acuminata* (*orf74*). The expected values for these matches are borderline statistically significant (2e-04 and 6e-04, respectively), so whether these data reflect real distant evolutionary relationships remains unresolved.

Each individual *Gordonia* phage genome sequence was unique, but given the close genetic relationships between the Mycolata host genera, attempts were made to classify these according to the system of Hatfull, Jacobs-Sera [43] designed to show evolutionary relationships the Mycobacteriophages. It was not possible to place these *Gordonia* phages into any of their pre-existing clusters. For example, while GMA7 is highly similar to GTE7 at a nucleotide sequence level, and GMA3 contains genes encoding several putative proteins also similar to those of GTE7, none could be grouped with any Mycobacteriophages. Instead they emerge as singletons since none of the existing clusters embraced them.

Suitability of these phages for use in foam bio-control

Of the nine phages examined in this study, GMA3, GMA5, GMA5 and GRU3 contain putative integrase genes, suggesting that they may undertake a lysogenic lifecycle. If these genes are functional, then these are probably undesirable candidates for standard phage therapy for activated sludge foam control.

Of the remaining five phages GMA2 and GMA6 both appear to contain a putative VIP2 family actin-ADP-ribosylating toxin gene, which target eukaryotic proteins upon infection. Consequently, neither phage would be considered being suitable for bio-control strategies. Their release into the environment may potentially result in the spread of these undesirable genes and an increased virulence of other bacteria hosts.

Other *Gordonia* phages GMA7, GTE6, and GTE8 appear to be obligatory lytic. Of these, GMA7 and GTE8 seem particularly attractive as both have impressive broad host ranges. For instance, GMA7 targets eleven strains of *Gordonia* including those of *G. terrae*, *G. malaquae*, *G. rubropertincta and G. hydrophobica*. Similarly, GTE8 targets several species including *G. terrae* (CON34, GOR9, G232) *and G. rubropertincta* (CON38) and *Nocardia asteroids* (strain CON12) (Table 1). Furthermore, phage GMA7 contains a putative DNA methylase gene (*orf38*) containing a pfam01555 motif. If this gene is functional, then this phage may evade cleavage by host defense RM systems [33, 38] and thus become an even more powerful addition to any phage therapy cocktail.

Conclusions

Nine phages infective for members of the genus *Gordonia* were isolated from wastewater and natural water environments, several of which had broad host ranges. Methods for visualization

of the phage infection cycle using TEM were successful and may prove to be useful in studies of mechanisms of phage infection. Whole genome sequencing of these phages revealed that their genomes were all distinctively different, failing to cluster with those of known Mycobacteriophages, based on both nucleotide and amino acid sequence similarities. Some are less modular in their genomic architecture than those characterized previously, and contain a higher number of lysin genes seen in other Actinophage genomes previously. Of these nine phages, three broad host range phages GMA7, GTE6, and GTE8 appear obligatory lytic and hence potentially suitable candidates for phage therapy cocktails to control activated sludge foaming.

Supporting Information

S1 Table. Genome annotations of nine phages infecting *Gordonia* spp. [a] ORFs were numbered consecutively, [b] The most closely related gene (only if named) and the name of the organism, [c] Percentage identity is based on the best match when a BLAST P analysis is performed, [d] The probability of obtaining a match by chance as determined by BLAST analysis. Only values less than 10^{-4} were considered significant, [e] Predicted function is based on amino acid identity, conserved motifs, and gene location within functional modules.
(DOCX)

S2 Table. Putative tRNA detected in *Gordonia* phage genomes.
(DOCX)

S3 Table. Palindromes in the genome sequence of nine *Gordonia* spp. phages.
(DOCX)

S4 Table. Repeats in the genome sequences of nine *Gordonia* spp. phages. I indicates inverted repeat, D indicates direct repeat.
(DOCX)

S5 Table. Summary of *Gordonia* phage structural genes identified by mass spectroscopy.
(DOCX)

Acknowledgments

We wish to thank Dr. Robert Glaishier (LIMS), Dr. Sabine Wilkens (LIMS), Ms. Glenys Shirley (LIMS), Dr. Eric Hanssen (Bio21 Institute) and Dr. Christopher Adda (LIMS) for assistance with Transmission Electron Microscopy, Dr. Pierre Faou (LIMS) for assistance with mass spectroscopy, and Dr. Daniel Tillett for useful discussions.

Author Contributions

Conceived and designed the experiments: RS ZD SP. Performed the experiments: ZD. Analyzed the data: ZD SP. Contributed reagents/materials/analysis tools: JT. Wrote the paper: RS JT SP ZD.

References

1. de los Reyes FL III. 2010. Foaming. p 215–258. *In* Seviour RJ, Nielsen PH (ed), Microbial Ecology of Activated Sludge. IWA Publishing, London.
2. Drzyzga O. The strengths and weaknesses of Gordonia: a review of an emerging genus with increasing biotechnological potential. Crit Rev Microbiol. 2012; 38(4):300–16. doi: 10.3109/1040841X.2012.668134 PMID: 22551505
3. Arenskötter M, Bröker D, Steinbüchel A. Biology of the Metabolically Diverse Genus Gordonia. App Environ Microbiol. 2004; 70:3195–204.

4. Klatte S, Rainey FA, Kroppenstedt RM. Transfer of *Rhodococcus aichiensis* Tsukamura 1982 and *Nocardia amarae* Lechevalier and Lechevalier 1974 to the Genus *Gordonia* as *Gordona aichiensis* comb. nov. and *Gordona amarae* comb. nov. Int J Syst Bacteriol. 1994; 44(4):769–73. PMID: 7981103
5. Lechevalier MP, Lechevalier HA. *Nocardia amarae* sp. nov., an Actinomycete Common in Foaming Activated Sludge. Nati J Syst Bacteriol. 1974; 24(2):278–88.
6. Goodfellow M, Kumar V, Maldonado LA. Genus II. Gordonia (Tsukamura 1971) Stackbrandt, Smida and Collins 1988, 345^{VP}. In: Goodfellow M, Kampfer P, Busse H-J, Trujillo ME, Suzuki K-I, Ludwig W, et al., editors. Bergey's Manual of Systematic Bacteriology. 5. 2 ed. New York: Springer; 2012. p. 419–35.
7. Soddell JA, Seviour RJ. Microbiology of foaming in activated sludge plants. J App Bacteriol. 1990; 69:145–76.
8. Soddell JA. Foaming. In: Seviour RJ, editor. The microbiology of activated sludge. Dordrecht: Kluwer Academic Publishers; 1999.
9. Petrovski S, Dyson ZA, Quill ES, McIlroy SJ, Tillett D, Seviour RJ. An examination of the mechanisms for stable foam formation in activated sludge systems. Water Res. 2011; 45(5):2146–54. doi: 10.1016/j.watres.2010.12.026 PMID: 21239035
10. Thomas JA, Soddell JA, Kurtböke DÍ. Fighting foam with phages. Water Sci Technol. 2002; 46:511–53. PMID: 12216679
11. Petrovski S, Seviour RJ, Tillett D. Characterization of the Genome of the Polyvalent Lytic Bacteriophage GTE2, Which Has Potential for Biocontrol of *Gordonia*, *Rhodococcus*, and *Nocardia* Stabilized Foams in Activated Sludge Plants. App Environ Microbiol. 2011; 77(12):3923–9.
12. Petrovski S, Seviour RJ, Tillett D. Prevention of *Gordonia* and *Nocardia* Stabilized Foam Formation by Using Bacteriophage GTE7. App Environ Microbiol. 2011; 77(21):7864–7
13. Petrovski S, Tillett D, Seviour RJ. Genome Sequences and Characterization of the Related *Gordonia* Phages GTE5 and GRU1 and Their Use as Potential Biocontrol Agents. App Environ Microbiol. 2012; 78(1):42–7.
14. Petrovski S, Seviour RJ, Tillett D. Genome sequence and characterization of the *Tsukamurella* bacteriophage TPA2. Appl Environ Microbiol. 2011; 77(4):1389–98. doi: 10.1128/AEM.01938-10 PMID: 21183635
15. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007; 23(6):673–9. PMID: 17237039
16. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004; 32(1):11–6. PMID: 14704338
17. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005; 33(Web Server issue):W686–9. PMID: 15980563
18. Cserzo M, Wallin E, Simon I, Heijne Gv, Elofsson A. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the Dense Alignment Surface method. Prot Eng. 1997; 10(6):673–6.
19. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, et al. Origins of Highly Mosaic Mycobacteriophage Genomes. Cell. 2003; 113:171–82. PMID: 12705866
20. Pope WH, Anders KR, Baird M, Bowman CA, Boyle MM, Broussard GW, et al. Cluster M Mycobacteriophages Bongo, PegLeg, and Rey with Unusually Large Repertoires of tRNA Isotypes. J Virol. 2014; 88(5):2461–80. doi: 10.1128/JVI.03363-13 PMID: 24335314
21. Lesnik EA, Sampath R, Levene HB, Henderson TJ, McNeil JA, Ecker DJ. Prediction of rho-independent transcriptional terminators in Escherichia coli. Nucleic Acids Res. 2001; 29(17):3583–94. PMID: 11522828
22. Mott ML, Berger JM. DNA replication initiation: mechanisms and regulation in bacteria. Nat Rev Microbiol. 2007; 5:343–54. PMID: 17435790
23. Catalano CE. The terminase enzyme from bacteriophage lambda: a DNA-packaging machine. Cell Mol Life Sci. 2000; 57:128–48. PMID: 10949585
24. Fujisawa H, Morita M. Phage DNA packaging. Genes Cells. 1997; 2(9):537–45. PMID: 9413995
25. Hatfull GF. Bacteriophage Genomics. Curr opin microbiol. 2008; 11(5):447–53. doi: 10.1016/j.mib.2008.09.004 PMID: 18824125
26. Xu J, Hendrix RW, Duda RL. Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. Mol Cell. 2004; 16(1):11–21. PMID: 15469818
27. Petrovski S, Seviour RJ, Tillett D. Characterization and whole genome sequences of the *Rhodococcus* bacteriophages RGL3 and RER2. Arch Virol. 2013; 158(3):601–9. doi: 10.1007/s00705-012-1530-5 PMID: 23129131

28. Petrovski S, Dyson ZA, Seviour RJ, Tillett D. Small but sufficient: the *Rhodococcus* phage RRH1 has the smallest known Siphoviridae genome at 14.2 kilobases. J Virol. 2012; 86(1):358–63. doi: 10.1128/JVI.05460-11 PMID: 22013058

29. Payne K, Sun Q, Sacchettini J, Hatfull GF. Mycobacteriophage Lysin B is a novel mycolylarabinogalactan esterase. Mol microbiol. 2009; 73(3):367–81. doi: 10.1111/j.1365-2958.2009.06775.x PMID: 19555454

30. Wang IN, Smith DL, Young R. Holins: the protein clocks of bacteriophage infections. Annu Rev Microbiol. 2000; 54:799–825. PMID: 11018145

31. Petrovski S, Seviour RJ, Tillett D. Genome sequence and characterization of a Rhodococcus equi phage REQ1. Virus Genes. 2013; 46(3):588–90. doi: 10.1007/s11262-013-0887-1 PMID: 23381579

32. Petrovski S, Seviour RJ, Tillett D. Genome sequence of the Nocardia bacteriophage NBR1. Arch Virol. 2014; 159(1):167–73. doi: 10.1007/s00705-013-1799-z PMID: 23913189

33. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010; 8(5):317–27. doi: 10.1038/nrmicro2315 PMID: 20348932

34. Tamaki H, Zhang R, Angly FE, Nakamura S, Hong P-Y, Yasunaga T, et al. Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. Environ Microbiol. 2012; 14(2):441–52. doi: 10.1111/j.1462-2920.2011.02630.x PMID: 22040222

35. Markine-Goriaynoff N, Gillet L, Van Etten JL, Korres H, Verma N, Vanderplasschen A. Glycosyltransferases encoded by viruses. J Gen Virol. 2004; 85(10):2741–54.

36. Henthorn KS, Friedman DI. Identification of related genes in phages phi 80 and P22 whose products are inhibitory for phage growth in Escherichia coli IHF mutants. J Bacteriol. 1995; 177(11):3185–90. PMID: 7768817

37. Han S, Arvai AS, Clancy SB, Tainer JA. Crystal structure and novel recognition motif of Rho ADP-ribosylating C3 exoenzyme from Clostridium botulinum: structural insights for recognition specificity and catalysis1. Journal of Mol Biol. 2001; 305(1):95–107.

38. Samson JE, Magadan AH, Sabri M, Moineau S. Revenge of the phages: defeating bacterial defences. Nat Rev Micro. 2013; 11(10):675–87.

39. Ali Y, Koberg S, Heßner S, Sun X, Rabe B, Back A, et al. Temperate Streptococcus thermophilus phages expressing superinfection exclusion proteins of the Ltp type. Front Microbiol. 2014; 5.

40. Bajpai R, Soni V, Khandrika L, Jangir PK, Sharma R, Agrawal P. Genome Sequence of a Novel Actinophage PIS136 Isolated from a Strain of Saccharomonospora sp. J Virol. 2012; 86(17):9552. doi: 10.1128/JVI.01529-12 PMID: 22879621

41. El Tahir Y, Skurnik M. YadA, the multifaceted Yersinia adhesin. International Journal of Medical Microbiology. 2001; 291(3):209–18. PMID: 11554561

42. Summer EJ, Liu M, Gill JJ, Grant M, Chan-Cortes TN, Ferguson L, et al. Genomic and functional analyses of *Rhodococcus equi* phages ReqiPepy6, ReqiPoco6, ReqiPine5, and ReqiDocB7. Appl Environ Microbiol. 2011: 77(2):669–83. doi: 10.1128/AEM.01952-10 PMID: 21097585

43. Hatfull GF, Jacobs-Sera D, Lawrence JG, Pope WH, Russell DA, Ko CC, et al. Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. J Mol Biol. 2010; 397(1):119–43. doi: 10.1016/j.jmb.2010.01.011 PMID: 20064525

Arch Virol
DOI 10.1007/s00705-015-2631-8

BRIEF REPORT

Isolation and characterization of bacteriophage SPI1, which infects the activated-sludge-foaming bacterium *Skermania piniformis*

Z. A. Dyson[1] · J. Tucci[1] · R. J. Seviour[2] · S. Petrovski[2]

Received: 12 June 2015 / Accepted: 26 September 2015

Abstract Foaming in activated sludge plants is a worldwide problem commonly caused by proliferation of bacteria of the order *Corynebacteriales*. These include *Skermania piniformis*, a filamentous bacterium that has been documented to be a major cause of foaming globally, and particularly in Australian treatment plants. Phage SPI1 is the first phage that was isolated and shown to infect this organism. It targets seven of the nine strains of *S. piniformis* held in our culture collection, but none of the other 73 mycolata strains of different genera, mostly isolated from wastewater, against which it was tested. Phage SPI1 is a member of the family *Siphoviridae* and has a circularly permuted dsDNA genome of 55,748 bp with a G+C content of 67.8 mol %. It appears to be obligatorily lytic, with no evidence of genes related to a lysogenic mode of existence.

Keywords Phage · Bacteriophage · *Skermania piniformis* · PTLO · Activated sludge foaming · Biocontrol · Wgs · Phage therapy

Skermania piniformis is a common causative agent of stable foams and scums developing on the surfaces of aerobic reactors in activated sludge plants worldwide, and for reasons still not understood, especially those in Australia [2, 33, 34, 36]. Like other mycolic-acid-synthesising Gram-positive bacteria of the order *Corynebacteriales*, (referred to hereafter as the mycolata), which are associated with foam stabilization, the cells of *S. piniformis* are strongly hydrophobic [4]. Formation of stable foams is a serious operational problem, impacting negatively on treatment-plant aesthetics, increasing maintenance costs, and complicating sludge management [7, 11, 13].

Foaming is a flotation event that requires air bubbles, surface-active agents, and hydrophobic particles – in this case, the bacterial cells. Any control strategy should be directed at this bacterial component, because neither air bubbles nor detergents can be eliminated from the process [20]. Currently, no universal method is available to control foaming in wastewater treatment plants, a situation that probably reflects our present poor understanding of the bacteria involved. While earlier studies [21–28, 37] have resulted in successful isolation of phages that infect many members of the foaming mycolata, no phage has been reported yet that is lytic for *S. piniformis*.

An activated sludge sample containing a large number of pine-tree-like branching filaments typical of *S. piniformis* [4], collected from Albury, NSW, Australia, was screened for *S. piniformis* phages by multiple host enrichment (Table 1) and subsequent plaque plating, as described by Petrovski et al. [21]. Single plaques were obtained on lawn plates of *S. piniformis* strain NM40^T that were ~ 0.5 mm in diameter. This phage was plaque purified and named phage SPI1. It is the first *Skermania* phage to be described.

Phage SPI1 belongs to the family *Siphoviridae*, possessing the characteristic long, non-contractile tail (~ 239 nm) of members of the order *Caudovirales* with a B1 isometric capsid (~ 60 nm) morphotype (Fig. 1). When

Electronic supplementary material The online version of this article (doi:10.1007/s00705-015-2631-8) contains supplementary material, which is available to authorized users.

✉ S. Petrovski
steve.petrovski@latrobe.edu.au

[1] La Trobe Institute of Molecular Sciences, Bendigo, VIC 3550, Australia

[2] Department of Physiology, Anatomy and Microbiology, La Trobe University, Bundoora, VIC 3086, Australia

Published online: 12 October 2015

Table 1 *Skermania piniformis* strains used for the pooled enrichment isolation and host range determination of phage SPI1

Organism	Lab ID	Lysis observed
Skermania piniformis	NM40^{T,a}	+
Skermania piniformis	NM41^{a}	-
Skermania piniformis	NM101	+
Skermania piniformis	NM109	+
Skermania piniformis	NM168	+
Skermania piniformis	J8	+
Skermania piniformis	J20^{a}	-
Skermania piniformis	J50	+
Skermania piniformis	J54	+

Where ^{T} indicates type strain, ^{a} indicates that the strain was used in enrichment isolation of phage SPI1, + indicates lysis, and – indicates no lysis. Phages were isolated and subsequently purified from activated sludge samples as described by Petrovski et al. [21], using the enrichment pool containing these three *S. piniformis* strains. Host range determinations were carried out as described by Petrovski et al. [21]. The bacterial strains used and methods for their growth are described by Petrovski et al. [21]. Additional bacterial strains used in this study were *S. piniformis* (NM40^{T}, NM41, NM101, NM109, NM168, J8, J20, J50, and J54, all code numbers for *S. pinensis* strains held in the La Trobe University culture collection). These relatively slowly growing bacteria were grown on homemade R2A medium consisting of 0.5 g of yeast extract (Oxoid, Adelaide, Australia), 0.5 g of Proteose peptone (Difco, North Ryde, Australia), 0.5 g of Casamino acids (Difco, North Ryde, Australia), 0.5 g of glucose, 0.5 g of soluble starch (Difco, North Ryde, Australia), 0.3 g of K_2HPO_4, 0.005 g of $MgSO_4.7H_2O$, and 0.3 g of sodium pyruvate (BDH, Murarrie, Australia) per liter, either broth or agar (14 g/L) R2A (Oxoid, Adelaide, Australia) at 25 °C. All other chemicals were obtained from Sigma (Sydney, Australia) unless otherwise noted

screened against 73 mycolata strains isolated mainly from activated sludge treatment plants, plaques were produced on seven of the nine *S. piniformis* strains screened, including the type strain (NM40^{T}) and strains NM101, NM109, NM168, J8, J50, and J54, isolated from several different activated sludge plants. A similar level of strain specificity has been noted for other actinophages [21–28] and probably reflects the absence of a suitable receptor site, or action of bacterial phage resistance systems, including clustered regularly interspersed repeat (CRISPR) systems, restriction modification (RM) systems, or abortive infection (Abi) systems in the resistant strains [14]. No lysis was observed when mock infection controls were performed with each strain, revealing that cell lysis was not a consequence of spontaneous release of prophages from the susceptible host strains.

The genome sequence of phage SPI1 was determined using Roche/454 pyrosequencing. This generated 15,051 reads, with an average of 99-fold sequence coverage after assembly. This phage possesses a circularly permuted dsDNA genome 55,748 bp in size. It has a G+C content of 67.8 mol %, which is similar to that of its *S. piniformis* host

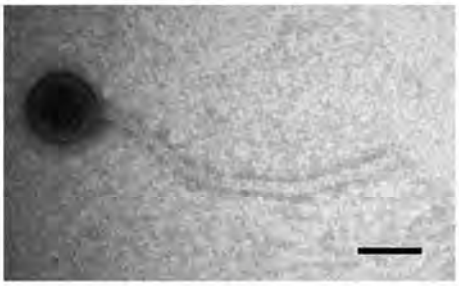

Fig. 1 Electron micrograph of phage SPI1. Scale bars represent 50 nm. Carbon and formvar grids (Electron Microscopy Sciences, Australia) for visualization of virions were stained with uranyl acetate as described by Petrovski et al. [21] prior to examination using a Jeol JEM-2010HC Electron Microscope

(67.5 mol %) [4], suggesting that SPI1 is well adapted to its host. At the DNA level, this genome has a sequence that is quite different from those of all other phage genomes, and therefore, comparisons of its putative open reading frames with others were based on their predicted expressed amino acid sequences. Analysis of this genome sequence revealed 67 putative open reading frames (ORFs). These are numbered consecutively, with 51 oriented in a forward orientation and 16 in a reverse orientation (Fig. 2). While 47 of the ORFs showed statistically significant sequence similarity to those reported previously, only 18 could be functionally annotated (Table 2). The predicted proteins encoded by 18 ORFs (27 %) exhibited no statistically significant similarity to any hypothetical protein.

The SPI1 genome is modularly organized, like those of other actinophages, consisting of DNA packaging, structural, cell lysis, and DNA replication/maintenance modules (Fig. 2). In the absence of a putative origin of replication, the ORFs were ordered from the beginning, with ORF1, which is located immediately upstream of the packaging module discussed later, overlapping ORF2, the putative small terminase gene in the same transcriptional direction in what appears to be an operon-like structure.

A large number of repeat structures were observed in the genome sequence of SPI1, among which were 25 palindromes (Table 3), and some of these were located in intergenic positions. These intergenic palindromes may function as *rho*-independent transcriptional terminators, although not all downstream flanking sequences displayed the T-rich region typical of those reported in *Escherichia coli* [16]. In *Streptomycetes lividans* [9] this T-rich region is not required for transcriptional termination, so what role, if any, these palindromic sequences might play in phage SPI1 remains to be determined. Also present in the SPI1 genome are large numbers of other repeat sequences, with

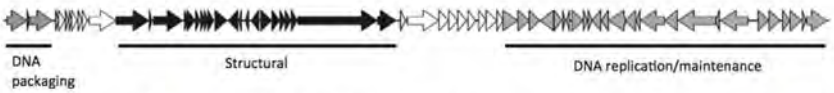

Fig. 2 Genome map of phage SPI1. Genes shaded in black indicate the lysis module. The genome of phage SPI1 was sequenced as described by Petrovski et al. [21], and *de novo* assembly was performed using CLC workbench (v6.5.1). Genome open reading frames (ORFs) within the assembled sequence with a minimum size of 90 bp [8] were detected initially using Glimmer (v3.02). All predicted 'start' codons were inspected manually for the presence of putative ribosomal binding sites and corrected as required. Sequence similarity searches were carried out against sequences held in the GenBank database, as described by Petrovski et al. [22]. The presence of tRNA and tmRNA were determined by using tRNAScan-SE [32] and ARAGORN [15]. Transmembrane domains were predicted using the DAS Transmembrane Prediction Server [5]. The nucleotide sequence for *Skermania piniformis* phage SPI1 has been deposited in GenBank under accession number KR011061

110 direct repeats and 80 inverted repeats ranging from 15 to 347 bp in size (Tables S1 and S2). Some of these repeats are quite large in size and in number compared to those seen in other actinophage genomes [21, 22, 24], suggesting that they probably have functional roles that are still unknown.

The DNA packaging module is comprised of the small and large terminase genes, ORF2 and ORF3 respectively. These genes encode the enzymes involved in packaging of phage DNA [29], and they were identified based on homology and a conserved motif (pfam03237) in ORF3. Typically, the small and large terminases function together as a complex, with the small terminase subunit determining the specificity of DNA packaging [3], while the large terminase subunit mediates cleavage of the phage DNA packaged into the prohead [10].

ORF12–ORF31 appear to represent the structural module of phage SPI1, which is typically located adjacent to the packaging module in phage genomes. ORF12-ORF16 encode proteins that share significant levels of sequence similarity to other proteins expressed in actinobacterial cells (Table 2). Some contain conserved motifs, adding further support to the notion that these genes encode structural proteins, e.g., ORF12 encodes a common motif for an F-like protein found in phage mu, which is commonly found in *Caudovirales* phages and is considered necessary for viral head morphogenesis [30].

The protein encoded by ORF21 is most similar in its amino acid sequence to a hypothetical protein from *Rhodococcus ruber* (70 % identity). However, it also shares amino acid sequence similarity with the main tail structural protein from *Tsukamurella* phage TPA2 (65 % identity) [21], and on this basis, we propose that it may carry out the same function or a similar one in phage SPI1.

The protein encoded by ORF28 was most similar in its sequence to a hypothetical protein from *R. rhodochrous* (38 % identity), but also to the gp30 protein of *Tsukamurella* phage TPA2 (52 % identity), located in its structural module. This location suggests that ORF28 could also be a structural gene in phage SPI1. Both ORF29 and ORF30 are located upstream of the putative tape measure protein gene (ORF31), although no putative ribosomal slippage sequences indicative of a conserved translational frameshift mechanism [39] could be identified. The protein encoded by ORF31 contains 1801 amino acid residues and is the largest protein encoded by the SPI1 genome, making it a strong candidate for the putative tape measure protein and thus typical of phages in the family *Siphoviridae* [19]. This suggestion is supported by the fact that it possesses a motif (TIGR01760) that is diagnostic for a tape measure protein. The remainder of the genes within the structural module have no predicted or known function, but given their location, with the exception of ORF18, could encode other phage structural proteins.

A phage lysis genome module consists typically of one or more lysis genes and a holin gene. Together, these are responsible for host cell lysis and the release of phage progeny at the end of the replication cycle [6]. ORF18, located within the putative structural gene module encodes a protein that shares amino acid sequence homology with a putative holin (Table 1). This finding was surprising, given that holin genes are typically located adjacent to the lysin genes in dsDNA phage genomes [6, 18] and ORF18 was not seen in the vicinity of a recognizable lysin gene. Furthermore, the predicted gene product contained one putative transmembrane region, thus only partially satisfying the listed criteria for a holin protein [38]. These suggest that phage holin proteins should contain two or more transmembrane regions and be less than 150 amino acid residues in size. Therefore, ORF18 is unlikely to encode a holin.

A higher level of amino acid sequence homology supported the annotation of ORF49 and ORF50 as components of the lysis module of phage SPI1, despite their location within the DNA replication module. Such a placement of lysis genes has been seen elsewhere in the genome of *Gordonia* phage GTE7 [23] and *Rhodococcus* phage REQ1 [26]. Therefore, this arrangement may be a more common feature in this group of phage genomes than presently considered. The protein encoded by ORF49 shares 63 % amino acid sequence identity with a hypothetical protein from *Rhodococcus* sp. strain 29MFTsu3.1 and contains a

Table 2 Genome annotation of *Skermania piniformis* phage SPI2

ORF[a]	Coordinates	Size (aa)	Significant match[b]	% Identity[c]	E-value[d]	Putative function (conserved motif)[e]
ORF1	125..1504	460	-			
ORF2	1501..1629	43	-			Putative small terminase
ORF3	1626..3227	534	Hypothetical protein [*Rhodococcus ruber*]	55	0.0	Large terminase (pfam03237)
ORF4	3435..3623	63	Hypothetical protein [*Rhodococcus rhodochrous*]	39	2e-06	
ORF5	3620..4051	144	Hypothetical protein TPA2_gp76 [*Tsukamurella* phage TPA2]	47	3e-09	
ORF6	4097..4282	62	-			
ORF7	4339..4503	55	-			
ORF8	4508..4885	126	-			
ORF9	4947..5195	83	-			
ORF10	5232..5567	112	gp10 [*Mycobacterium* phage Pipefish]	39	2e-14	Unknown (pfam07098)
ORF11	5702..7516	605	Hypothetical protein [*Rhodococcus rhodochrous*]	47	6e-162	
ORF12	7513..9762	750	Hypothetical protein [*Rhodococcus rhodochrous*]	39	4e-145	Putative capsid protein (pfam04233)
ORF13	9762..9938	59	Hypothetical protein TPA2_gp19 [*Tsukamurella* phage TPA2]	42	2e-04	
ORF14	10048..12063	672	Hypothetical protein [*Rhodococcus ruber*]	49	4e-174	
ORF15	12149..12370	74	putative Gp13 [*Nocardia cyriacigeorgica*]	52	3e-11	
ORF16	12322..12894	191	putative Gp13 [*Nocardia cyriacigeorgica*]	47	4e-36	
ORF17	12909..13205	99	-			
ORF18	13257..13538	94	Holin [*Rhodococcus* phage ReqiDocB7]	40	2e-07	
ORF19	13510..13746	79	-			
ORF20	13743..14132	130	-			
ORF21	14233..15042	270	Hypothetical protein [*Rhodococcus ruber*]	70	5e-133	
ORF22	Complement (15110..15778)	223	-			
ORF23	Complement (15852..16001)	50	-			
ORF24	Complement (16278..16502)	75	-			
ORF25	Complement (16676..17197)	174	gp46 [*Mycobacterium* phage Acadian]	42	1e-26	
ORF26	17337..18080	248	Hypothetical protein [*Rhodococcus rhodochrous*]	54	2e-74	
ORF27	18082..18612	177	Hypothetical protein [*Rhodococcus ruber*]	54	6e-51	

Table 2 continued

ORF[a]	Coordinates	Size (aa)	Significant match[b]	% Identity[c]	E-value[d]	Putative function (conserved motif)[e]
ORF28	18626..18979	118	Hypothetical protein [*Rhodococcus rhodochrous*]	38	1e-07	
ORF29	18972..19376	135	Tail assembly chaperone [*Mycobacterium* phage Jolie1]	47	1e-24	Putative tail assembly protein
ORF30	19402..19797	132	Hypothetical protein TPA2_gp32 [*Tsukamurella* phage TPA2]	39	3e-09	
ORF31	19838..25240	1801	Hypothetical protein [*Nocardia otitidiscaviarum*]	38	3e-180	Putative tape measure protein (TIGR01760; COG5412)
ORF32	25249..26508	420	Hypothetical protein [*Rhodococcus ruber*]	54	4e-163	
ORF33	26804..27139	112	Hypothetical protein [Rhodococcus ruber]	77	3e-18	
ORF34	27255..29411	719	Hypothetical protein [*Rhodococcus ruber*]	50	2e-115	
ORF35	29414..30061	216	Hypothetical protein [*Rhodococcus ruber*]	37	8e-38	
ORF36	30071..30691	207	Hypothetical protein [*Rhodococcus ruber*]	50	3e-65	
ORF37	30688..31281	198	Hypothetical protein [*Rhodococcus rhodochrous*]	48	2e-55	
ORF38	31386..31886	167	Hypothetical protein [*Rhodococcus rhodochrous*]	38	4e-27	
ORF39	32109..32612	168	-			
ORF40	32609..33346	246	Putative uncharacterized protein [*Rhodococcus* sp. AW25M09]	45	3e-10	
ORF41	33333..33746	138	Hypothetical protein [*Gordonia paraffinivorans*]	33	1e-04	
ORF42	33743..34762	340	Putative uncharacterized protein [*Rhodococcus* sp. AW25M09]	38	2e-36	CRISPR associated RAMP superfamily protein Csf2 (cd09706)
ORF43	34759..35541	261	Hypothetical protein [*Mycobacterium smegmatis*]	32	3e-13	
ORF44	35523..36242	240	Hypothetical protein [*Rhodococcus imtechensis*]	37	1e-23	Phosphoadenosine phosphosulfate reductase family (pfam01507)
ORF45	Complement (36247..37182)	312	MULTISPECIES: ribonucleotide-diphosphate reductase subunit beta [*Rhodococcus*]	89	0.0	Ribonucleotide reductase (pfam00268)
ORF46	37216..37737	174	Hypothetical protein [*Rhodococcus rhodochrous*]	31	2e-06	
ORF47	37803..37958	52	-			
ORF48	Complement (38040..38258)	73	Hypothetical protein [*Gardnerella vaginalis*]	52	3e-11	Dehydrogenase (PRK08324)

Table 2 continued

ORF[a]	Coordinates	Size (aa)	Significant match[b]	% Identity[c]	E-value[d]	Putative function (conserved motif)[e]
ORF49	38423..39205	261	Hypothetical protein [*Rhodococcus* sp. 29MFTsu3.1]	63	8e-92	Lysin (pfam13539)
ORF50	39202..39486	95	Hypothetical protein [*Micromonospora parva*]	48	2e-05	Putative Holin
ORF51	39515..39790	92	Hypothetical protein PBI_BERNAL13_1 [*Mycobacterium* phage Bernal13]	42	3e-06	
ORF52	Complement (39778..40371)	198	RuvC-like resolvase superfamily protein [*Rhodococcus* phage E3]	40	6e-32	
ORF53	Complement (40401..40883)	161	gp48 [*Mycobacterium* phage Daisy]	48	2e-24	
ORF54	Complement (40880..42031)	384	gp49 [*Mycobacterium* phage Arbiter]	36	2e-54	Nuclease (pfam12705)
ORF55	Complement (42028..42738)	237	Putative uncharacterized protein [*Rhodococcus* sp. AW25M09]	33	7e-08	
ORF56	Complement (42775..43041)	89	-			
ORF57	Complement (43057..44694)	546	Helicase [*Tsukamurella* phage TPA2]	52	1e-160	Helicase (COG1061)
ORF58	Complement (44691..45566)	292	-			
ORF59	Complement (45615..48164)	850	Primase [*Tsukamurella* phage TPA2]	40	2e-171	Primase/polymerase (pfampfam09250; pfam13481)
ORF60	Complement (48158..48418)	87	-			DNA binding (cd00569)
ORF61	Complement (48488..50353)	622	gp58 [*Mycobacterium* phage Acadian]	47	4e-163	DNA polymerase I - 3′-5′ exonuclease (COG0749)
ORF62	51044..51766	241	-			
ORF63	51783..52529	249	-			
ORF64	52705..53217	171	Hypothetical protein [*Nocardia cyriacigeorgica*]	36	3e-14	
ORF65	53214..54002	263	-			
ORF66	54033..54275	81	Type B dihydrofolate reductase DfrB6 [*Salmonella enterica* subsp. enterica serovar Infantis]	55	4e-09	R67 dihydrofolate reductase (pfam06442)
ORF67	54335..55735	467	Hypothetical protein [*Corynebacterium falsenii*]	33	6e-28	Pentapeptide repeats (pfam13599;pfam13599)

[a] ORFs were numbered consecutively, [b] The most closely related gene (only if named) and the name of the organism, [c] Percentage identity is based on the best match when a BLAST P analysis is performed, [d] The probability of obtaining a match by chance as determined by BLAST analysis. Only values less than 10^{-4} were considered significant, [e] Predicted function is based on amino acid identity, conserved motifs, and gene location within functional modules

Table 3 Palindromic sequences present in the SPI1 genome sequence

Palindrome number	Size (bp)	Coordinates	Sequence alignment	Position in genome
P1	63	54251-54313	GTTGGTGAGCGTCCCTCCACTATAGCACGGTGACCGTGCTATAGTGGA GGGACCGCTCACCAAC	Overlapping the 3' end of ORF66
		54313-54251	GTTGGTGAGCGTCCCTCCACTATAGCACGGTCACCGTGCTATAGTGGA GGGACCGCTCACCAAC	
P2	46	12070-12115	GGGGAACCGGGCGGGCGGCACGATCGCTTCGCCCGCCCGGCTCCCC	Between ORF14 and ORF15
		12115-12070	GGGGAGCCGGGCGGGCGAAGCGATCGTGCCGCCCGCCCGGTTCCCC	
P3	40	51772-51811	GTAACGGTCGCATGTGGCTTACGGCACATGCGACCGTTAC	Overlapping the 5' end of ORF63
		51811-51772	GTAACGGTCGCATGTGCCGTAAGCCACATGCGACCGTTAC	
P4	40	15057-15096	AACGGGCACGGTATCCGAAAGGACGGATACCGTGCCCGTT	Between ORF21 and ORF22
		15096-15057	AACGGGCACGGTATCCGTCCTTTCCGGATACCGTGCCCGTT	
P5	38	9359-9396	GCGTACCCGCTCCAGGAGCCAGCCCGGAGCGGGTACGC	Within ORF12
		9396-9359	GCGTACCCGCTCCGGGCTGGCTCCTGGAGCGGGTACGC	
P6	37	50457-50493	GAACGGGCTCCCCCTCTCTCCGCGGGGGAGCCCCGTTC	Between ORF61 and ORF62
		50493-50457	GAACGGGCTCCCCCGCGGAGAGAGGGGGAGCCCCGTTC	
P7	31	3349-3379	CGATACTAGCACGGCCGACCGTGCTAGTATCG	Between ORF3 and ORF4
		3379-3349	CGATACTAGCACGGTCGCCGTGCTAGTATCG	
P8	31	6204-6234	TCTTTCGGGTGTGGGACCCGCACCCGAAAGA	Within ORF11
		6234-6204	TCTTTCGGGTGCGGGTCCCACACCCGAAAGA	
P9	30	9497-9526	TGGTTCGCGTCCGGAGCGACGGACGCGACCA	Within ORF12
		9526-9497	TGGTCGCGTCCGTCGCTCCGGACGCGACCA	
P10	30	8497-8526	GCCGTCATCGGTACACCGACCGATGACGGC	Within ORF12
		8526-8497	GCCGTCATCGGTCGGTGTACCGATGACGGC	
P11	29	16078-16106	CACTATAGCACGGTCACCGTGCTATAGTG	Between ORF23 and ORF24
		16106-16078	CACTATAGCACGGTGACCGTGCTATAGTG	
P12	29	1654-1682	ACGCTCAGCGGTACCGCGCCGCTGAGCGT	Within ORF3
		1682-1654	ACGCTCAGCGGCGCGGTACCGCTGAGCGT	
P13	25	7910-7934	CGGACGCGCTCGACGAGCGCGTCCG	Within ORF12
		7934-7910	CGGACGCGCTCGTCGAGCGCGTCCG	
P14	23	4278-4300	AGTAGCACGGTGACCGTGCTACT	Overlapping the 3' end of ORF6
		4300-4278	AGTAGCACGGTCACCGTGCTACT	
P15	22	46302-46323	CACGCTCGCGCGCGCGAGCGTG	Within ORF59
		46323-46302	CACGCTCGCGCGCGCGAGCGTG	
P16	22	15378-15399	TCTTCCGCGAGCTCGCGGAGGA	Within ORF22
		15399-15378	TCCTCCGCGAGCTCGCGGAAGA	
P17	20	15197-15216	CAGCGTGGCATGCCACGCTG	Within ORF22
		15216-15197	CAGCGTGGCATGCCACGCTG	

Table 3 continued

Palindrome number	Size (bp)	Coordinates	Sequence alignment	Position in genome
P18	20	40924-40943	GGGCCAAGATATCTTGGCCC	Within ORF54
		40943-40924	GGGCCAAGATATCTTGGCCC	
P19	20	48547-48566	GACCGCCCCGCCGGGCGATC	Within ORF61
		48566-48547	GATCGCCCGCCGGGGCGGTC	
P20	18	43017-43034	TATCTATCCGGATAGATA	Within ORF56
		43034-43017	TATCTATCCGGATAGATA	
P21	16	2618-2633	CCGGACCCGGGTCCGG	Within ORF3
		2633-2618	CCGGACCCGGGTCCGG	
P22	16	7154-7169	GGTCCGCCCGGCGGACC	Within ORF11
		7169-7154	GGTCCGCCCGGCGGACC	
P23	16	23992-24007	GGTCATCGCGATGACC	Within ORF31
		24007-23992	GGTCATCGCGATGACC	
P24	16	45388-45403	GTGTCTCCGGAGACAC	Within ORF58
		45403-45388	GTGTCTCCGGAGACAC	
P25	16	46832-46847	ACATCGACGTCGATGT	Within ORF59
		46847-46832	ACATCGACGTCGATGT	

pfam13539 motif diagnostic of a D-alanyl-D-alanine carboxypeptidase phage lysin protein. In this case, ORF50 satisfies the criteria for a holin-encoding gene detailed by Wang et al. [38], as its product possessed two transmembrane regions and was only 95 amino acids in size, supporting its identification as a phage holin.

The DNA replication/maintenance module of phage SPI1 appears to be encoded by genes ORF42–ORF66. Many of these genes are distinctive, while some share homology with other actinobacterial genes. For example, ORF48 encodes a protein sharing 52 % amino acid sequence identity with a hypothetical protein from *Gardenerella vaginalis* and contains the motif for a short-chain dehydrogenase (PRK08324). The amino acid sequence of the protein encoded by ORF52 is 40 % identical to a RuvC-like resolvase protein of *Rhodococcus* phage E3 [31], suggesting that this gene might provide a mechanism for DNA recombination events [35] in phage SPI1. Putative Holliday junction resolvase genes have been found in the genome sequences of several other actinophages, including phages TPA2 [21] and GTE2 [22].

ORF57 appears to encode a helicase, as its product shares 52 % amino acid sequence identity with the helicase of *Tsukamurella* phage TPA2 and contains a motif (COG1061) that is diagnostic for a helicase. In the case of gene ORF59, its predicted amino acid sequence shares 40 % identity with the primase gene from the phage TPA2. Furthermore, its N-terminal region contains a motif (pfam09250) that is diagnostic for a bifunctional DNA primase/polymerase [17], while its C-terminal region contains the motif (pfam13481) for the AAA domain in many presumed DNA repair proteins [1]. While no statistically significant matches could be identified for ORF60, based on either its nucleotide or translated amino acid sequence, its product did contain an HTH DNA binding motif (cd00569). This suggests that it might have a regulatory function in SPI1. The protein encoded by ORF61 shares 47 % amino acid sequence identity with the gp58 protein of *Mycobacterium* phage Acadian, which encodes a DNA polymerase I-3′-5′ exonuclease and polymerase. ORF61 also possesses a motif (COG0749) for a DNA polymerase I 3′-5′ exonuclease and polymerase, suggesting that it might function in this capacity in phage SPI1.

Analysis of the SPI1 genome suggests a number of complex recombination events may have contributed to its evolution. The presence of a Holliday junction resolvase is supportive of the potential for homologous recombination events in its past, as these junction structures form during recombination events [35]. It is also possible that homologous recombination events have influenced the order of the genes in SPI1, since ORF5, which appears to encode a homologue of the Gp76 protein of phage TPA2, is located upstream of ORF30, which itself appears to encode a

homologue of the Gp32 protein of TPA2. Several genes with no similarity to TPA2 genes are interspersed between them [21]. Similar recombination events have been observed in the genome sequence of phage TPA2, which might suggest that despite the conserved modular arrangement of genes within phage genomes, they can recombine to form new modular arrangements [21], an idea that is further supported by the presence of the chimeric gene ORF14 in phage SPI1.

Although the genome sequence of phage SPI1 is distinctive at the nucleotide sequence level, examination of its putative gene product sequences suggests that 34 % of its genes share statistically significant sequence similarities to those from the cluster B phages in the clustering system of Hatfull et al. [12]. This cluster also includes *Tsukamurella* phage TPA2 [21].

The information provided in this paper will allow development and application of real-time PCR assays to quantify this SP1 phage *in situ* and to understand better its population dynamics. It will enable several key questions to be addressed, which will determine the suitability of this phage and other yet to be isolated *S. piniformis* phages for control of foams caused by this organism. These include determining its optimal dosing rate and its persistence in treatment plants. What is known about phage population dynamics in natural habitats would suggest that the complete elimination of either the phage or the host cells would not occur, and indeed, elimination of *S. piniformis* might actually reduce wastewater treatment efficiency because this organism plays a positive role in degradation of hydrophobic substrates. The aim would therefore be to reduce the *Skermania* population levels below the threshold level required for stable foam formation. This is not the same for all mycolata in all plants [20], and hence would need to be determined on an individual-plant basis.

Acknowledgments The authors thank Dr. Alexander Fink (La Trobe University) for assistance with transmission electron microscopy, and Dr. Daniel Tillett for useful discussions. Z.A. Dyson was the recipient of an Australian Postgraduate Award PhD Scholarship. The authors declare no conflict of interest.

References

1. Ammelburg M, Frickey T, Lupas AN (2006) Classification of AAA+ proteins. J Struct Biol 156:2–11
2. Blackall LL, Harbers AE, Greenfield PF, Hayward AC (1991) Foaming in activated sludge plants: a survey in Queensland, Australia and an evaluation of some control strategies. Water Res 25:313–317
3. Catalano CE (2000) The terminase enzyme from bacteriophage lambda: a DNA-packaging machine. Cell Mol Life Sci 57:128–148
4. Chun J, Blackall LL, Kang S-O, Hah YC, Goodfellow M (1997) A proposal to reclassify *Nocardia pinensis* Blackall et al. as *Skermania pinifomis* gen. nov., comb. nov. IJSEM 47:127–131
5. Cserzo M, Wallin E, Simon I, Gv Heijne, Elofsson A (1997) Prediction of transmembrane alpha-helices in procariotic membrane proteins: the Dense Alignment Surface method. Protein Eng 10:673–676
6. Daniel A, Bonnen PE, Fischetti VA (2007) First complete genome sequence of two *Staphylococcus epidermidis* bacteriophages. J Bacteriol 189:2086–2100
7. de los Reyes FL III (2010) Foaming. In: Seviour RJ, Nielsen PH (eds) Microbial ecology of activated sludge. IWA Publishing, London, pp 215–258
8. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23:673–679
9. Deng Z, Kieser T, Hopwood DA (1987) Activity of a *Streptomyces* transcriptional terminator in *Escherichia coli*. Nucleic Acids Res 15:2665–2675
10. Fujisawa H, Morita M (1997) Phage DNA packaging. Genes Cells 2:537–545
11. Goddard AJ, Forster CF (1987) Stable foams in activated sludge plants. Enzyme Microb Technol 9:164–168
12. Hatfull GF, Jacobs-Sera D, Lawrence JG, Pope WH, Russell DA, Ko CC, Weber RJ, Patel MC, Germane KL, Edgar RH, Hoyte NN, Bowman CA, Tantoco AT, Paladin EC, Myers MS, Smith AL, Grace MS, Pham TT, O'Brien MB, Vogelsberger AM, Hryckowian AJ, Wynalek JL, Donis-Keller H, Bogel MW, Peebles CL, Cresawn SG, Hendrix RW (2010) Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. J Mol Biol 397:119–143
13. Jenkins D, Daigger GT, Richard MG (1993) Manual on the causes and control of activated sludge bulking and foaming. Lewis, United States
14. Labrie SJ, Samson JE, Moineau S (2010) Bacteriophage resistance mechanisms. Nat Rev Microbiol 8:317–327
15. Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32:11–16
16. Lesnik EA, Sampath R, Levene HB, Henderson TJ, McNeil JA, Ecker DJ (2001) Prediction of rho-independent transcriptional terminators in *Escherichia coli*. Nucleic Acids Res 29:3583–3594
17. Lipps G, Weinzierl AO, von Scheven G, Buchen C, Cramer P (2004) Structure of a bifunctional DNA primase-polymerase. Nat Struct Mol Biol 11:157–162
18. Lu Z, Altermann E, Breidt F, Kozyavkin S (2010) Sequence analysis of *Leuconostoc mesenteroides* bacteriophage Φ1-A4 isolated from an industrial vegetable fermentation. Appl Environ Microbiol 76:1955–1966
19. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR, Hendrix RW, Hatfull GF (2003) Origins of highly mosaic mycobacteriophage genomes. Cell 113:171–182
20. Petrovski S, Dyson ZA, Quill ES, McIlroy SJ, Tillett D, Seviour RJ (2011) An examination of the mechanisms for stable foam formation in activated sludge systems. Water Res 45:2146–2154
21. Petrovski S, Seviour RJ, Tillett D (2011) Genome sequence and characterization of the *Tsukamurella* bacteriophage TPA2. Appl Environ Microbiol 77:1389–1398
22. Petrovski S, Seviour RJ, Tillett D (2011) Characterization of the genome of the polyvalent lytic bacteriophage GTE2, which has potential for biocontrol of *Gordonia*, *Rhodococcus*, and *Nocardia* stabilized foams in activated sludge plants. Appl Environ Microbiol 77:3923–3929
23. Petrovski S, Seviour RJ, Tillett D (2011) Prevention of *Gordonia* and *Nocardia* stabilized foam formation by using bacteriophage GTE7. Appl Environ Microbiol 77:7864–7867

24. Petrovski S, Dyson ZA, Seviour RJ, Tillett D (2012) Small but sufficient: the *Rhodococcus* phage RRH1 has the smallest known Siphoviridae genome at 14.2 kilobases. J Virol 86:358–363
25. Petrovski S, Tillett D, Seviour RJ (2012) Genome sequences and characterization of the related *Gordonia* phages GTE5 and GRU1 and their use as potential biocontrol agents. Appl Environ Microbiol 78:42–47
26. Petrovski S, Seviour RJ, Tillett D (2013) Characterization and whole genome sequences of the *Rhodococcus* bacteriophages RGL3 and RER2. Arch Virol 158:601–609
27. Petrovski S, Seviour RJ, Tillett D (2013) Genome sequence and characterization of a *Rhodococcus equi* phage REQ1. Virus Genes 46:588–590
28. Petrovski S, Seviour RJ, Tillett D (2014) Genome sequence of the *Nocardia* bacteriophage NBR1. Arch Virol 159:167–173
29. Rao VB, Feiss M (2008) The bacteriophage DNA packaging motor. Ann Rev Gene 42:647–681
30. Ratcliff SW, Luh J, Ganesan AT, Behrens B, Thompson R, Montenegro MA, Morelli G, Trautner TA (1979) The genome of *Bacillus subtilis* phage SPP1. Mol Gen Genet 168:165–172
31. Salifu SP, Valero-Rello A, Campbell SA, Inglis NF, Scortti M, Foley S, Vázquez-Boland JA (2013) Genome and proteome analysis of phage E3 infecting the soil-borne actinomycete *Rhodococcus equi*. Environ MicrobiolRep 5:170–178
32. Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33:W686–W689
33. Seviour EM, Williams C, DeGrey B, Soddell JA, Seviour RJ, Lindrea KC (1994) Studies on filamentous bacteria from Australian activated sludge plants. Water Res 28:2335–2342
34. Seviour EM, Williams CJ, Seviour RJ, Soddell JA, Lindrea KC (1990) A survey of filamentous bacterial populations from foaming activated sludge plants in eastern states of Australia. Water Res 24:493–498
35. Sharples GJ, Corbett LM, McGlynn P (1999) DNA structure specificity of Rap endonuclease. Nucleic Acids Res 27:4121–4127
36. Soddell JA, Seviour RJ (1998) Numerical taxonomy of *Skermania piniformis* and related isolates from activated sludge. J Appl Microbiol 84:272–284
37. Thomas JA, Soddell JA, Kurtböke DÍ (2002) Fighting foam with phages. Water Sci Technol 46:511–553
38. Wang IN, Smith DL, Young R (2000) Holins: the protein clocks of bacteriophage infections. Annu Rev Microbiol 54:799–825
39. Xu J, Hendrix RW, Duda RL (2004) Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. Mol Cell 16:11–21

8.4. Appendix 4 – Candidate's publications not included in this thesis

Petrovski, S., Dyson, Z. A., Seviour, R. J. & Tillett, D. (2012). Small but sufficient: the *Rhodococcus* phage RRH1 has the smallest known *Siphoviridae* genome at 14.2 kilobases. *J Virol* **86**, 358-363.

Petrovski, S., Dyson, Z. A., Quill, E. S., McIlroy, S. J., Tillett, D. & Seviour, R. J. (2011). An examination of the mechanisms for stable foam formation in activated sludge systems. *Water Res* **45**, 2146-2154.

JVI
Journals.ASM.org

Small but Sufficient: the *Rhodococcus* Phage RRH1 Has the Smallest Known *Siphoviridae* Genome at 14.2 Kilobases

Steve Petrovski,* Zoe A. Dyson, Robert J. Seviour,* and Daniel Tillett

La Trobe Institute for Molecular Sciences, La Trobe University, Victoria 3552, Australia

Bacteriophages are considered to be the most abundant biological entities on the planet. The *Siphoviridae* are the most commonly encountered tailed phages and contain double-stranded DNA with an average genome size of ~50 kb. This paper describes the isolation from four different activated sludge plants of the phage RRH1, which is polyvalent, lysing five *Rhodococcus* species. It has a capsid diameter of only ~43 nm. Whole-genome sequencing of RRH1 revealed a novel circularly permuted DNA sequence (14,270 bp) carrying 20 putative open reading frames. The genome has a modular arrangement, as reported for those of most *Siphoviridae* phages, but appears to encode only structural proteins and carry a single lysis gene. All genes are transcribed in the same direction. RRH1 has the smallest genome yet of any described functional *Siphoviridae* phage. We demonstrate that lytic phage can be recovered from transforming naked DNA into its host bacterium, thus making it a potentially useful model for studying gene function in phages.

There are thought to be at least 10^{31} phage particles on the planet (22), making them the most abundant of biological entities. Our understanding of their genetic variety is scant, and we have only just begun to explore this enormous phage diversity (35). Phages are classified on the basis of their morphology into 13 families and one named order (*Caudovirales*) that embraces more than 95% of all described phages (3). The *Caudovirales* are divided into three families differing in their tail morphology and function: the *Myoviridae* (contractile tails), the *Podoviridae* (short noncontractile tails), and the *Siphoviridae* (long noncontractile tails) (2). The most commonly observed phages (~60%) are members of the *Siphoviridae* (1).

Siphoviridae phages contain typically 35 to 70 kb of DNA in either an icosahedral capsid ranging in size from 55 to 60 nm or a prolate capsid (2, 20). Despite phage lambda, a member of this family, being one of the most studied model organisms, what constitutes the core genome of this family is still unknown. Complicating any attempt to determine the minimal functional *Siphoviridae* genome is that the known genomes cover a wide size range (16.9 kb to more than 121 kb), a lack of gene conservation between phages (8), and DNA packing constraints limit the ability to create deletion mutants (10). Identification of new phages with inherently small genomes and gene sets should provide useful models for unraveling the complexities of phage reproduction and ecology (39).

One group of phages that has received little attention is the group of phages that target *Rhodococcus* species, organisms found in a wide range of habitats, especially soil (16). Most species are nonpathogenic, although *R. equi* causes serious infections in foals (28), while other strains are responsible for operational problems in wastewater treatment systems (30). Summers et al. (41) and Petrovski et al. (32) have characterized five phages able to propagate on *Rhodococcus* species. These phages all have novel genome sequences and more than 50% of each genome is composed of new genes of unknown function.

In this study, we describe a novel *Rhodococcus* phage lytic for RRH1. This phage has the smallest known genome of any *Siphoviridae* phage (14.2 kb), with only 20 putative genes. These attributes, together with the ability to regenerate lytic phage in a host cell from transformed naked phage DNA demonstrated here, offer an exciting new model for better understanding phage biology.

MATERIALS AND METHODS

Bacterial strains and media. The bacterial strains used in this study are listed in a study by Petrovski et al. (31). All were grown in peptone yeast extract calcium (PYCa) broth or agar (31). All chemicals unless otherwise noted were obtained from Sigma, Australia.

Phage purification, host range determination, and single-step growth curves. Phage recovery and purification were performed using *R. rhodochrous* (Rrho39), *R. equi* (Requ10), and *R. erythropolis* (Rery29) as described by Petrovski et al. (32). Ten rounds of phage dilution and single plaque isolation were performed to ensure the final phage suspension resulted from a single virion, before undertaking any further characterization studies.

After purification, host range studies were performed using a drop dilution series of RRH1 phage (~10^{10} PFU/ml) spotted onto swabbed lawn plates of 65 bacterial strains (31). Plates were incubated for 2 days at 30°C before inspection for the presence of plaques. Single-step phage growth experiments were performed as described previously (4).

Electron microscopy. RRH1 virions (with and without added T2 phage) were allowed to absorb to Formvar coated 200-mesh copper grids for 5 min. These were washed twice for 1 min in double-distilled water (ddH_2O) and negatively stained with 2% (wt/vol) uranyl acetate for 2 min. Excess liquid was absorbed by filter paper and the grids allowed to air dry before being examined under a JEOL JEM-2010HC and a JEOL JEM-100cx transmission electron microscope (TEM) at an accelerating voltage of 100 kV. Electron micrographs were calibrated using catalase crystals (Electron Microscopy Sciences, Hatfield, PA), and the known *Escherichia*

Received 20 June 2011 Accepted 6 October 2011
Published ahead of print 19 October 2011
Address correspondence to Daniel Tillett, d.tillett@latrobe.edu.au.
* Present address: S. Petrovski, Department of Primary Industries, Biosciences Research Division, Victorian AgriBiosciences Centre, 1 Park Drive, La Trobe University Research and Development Park, Bundoora, Victoria 3083, Australia; R. J. Seviour, Department of Microbiology, La Trobe University, Bundoora, Victoria 3086, Australia.

doi:10.1128/JVI.05460-11

coli Myoviridae phage T2. RRH1 particle sizes were determined by measuring 40 symmetrical phage capsid vertices and 10 tail lengths.

DNA isolation and sequencing. RRH1 phages were purified with NaCl-polyethylene glycol (PEG) 8000, and DNA was isolated by using SDS-proteinase K as described previously (31). Four independent isolates of RRH1 from different geographical locations were sequenced using a Roche GS FLX genome sequencer and Titanium chemistry by Genoseq (UCLA, Los Angeles, CA). The pyrosequencing reads were assembled separately using the gsAssembler (Roche Applied Science, Indianapolis, IN). The resulting single contigs from each isolate had a minimum of 50-times sequence read coverage.

Genome annotation. The genome of RRH1 was annotated using the Integrative Services for Genomic Analysis (http://isga.cgb.indiana.edu) (21) interface to the Egatis (29) software package, followed by manual inspection of all gene predictions.

Putative open reading frames (ORFs) longer than 90 bases were predicted using Glimmer3 (13) with the iterative process described by Delcher et al. (12) to enhance predictive accuracy. All predicted start codons were inspected manually for the presence of putative ribosomal binding sites and adjusted as required.

Sequence similarity searches were performed using BLAST X against a nonredundant database, including data sourced from NCBI, Swiss-Prot, and PDB databases by using a significance value of 1e−04. The BLAST X results were used as input for the BLAST-Extend-Repraze algorithm (http://sourceforge.net/projects/ber/) to identify potential frameshifts or point mutations. Protein domain searches were performed using hmmpfam (http://hmmer.janelia.org/) against the PFAM (5) and TIGRFAM HMM (17) databases to identify protein family or domain matches. Each ORF was also checked manually using the conserved domain database (CDD) (14). Possible transmembrane domains were recognized using DAS (dense alignment surface method) transmembrane prediction (http://www.sbc.su.se/~miklos/DAS/) (11). Identified ORFs were also screened for the presence of lipoprotein motifs with the predicted protein sequence (6). Any possible tRNAs and tmRNA genes were screened by using tRNAscan-SE (24, 37).

Mass spectroscopy. To identify phage structural proteins, purified virions (~10^{10} PFU) were precipitated from concentrated stocks by using $ZnCl_2$ (36) to remove contaminating PEG. The pellet was reduced using 100 mM dithiothreitol (DTT) and heat denatured (100°C for 5 min). Samples were loaded into a 12% SDS-polyacrylamide gel electrophoresis gel prior to staining with Coomassie brilliant blue. The entire lane containing all visible protein bands was excised from the gel to create a protein pool. The excised gel was trypsin digested (40) followed by analysis using electrospray ionization (ESI)-time of flight mass spectrometry (TOF-MS) by the Mass Spectroscopy and Proteomic Facility at La Trobe University.

Electroporation of phage DNA into *Rhodococcus erythropolis*. Electrocompetent *R. erythropolis* (Rery29) cells were prepared as described by Sekizaki et al. (38) and stored at −70°C until required. Aliquots (100 μl) were thawed on ice prior to electroporation using 100 to 500 ng of RRH1 genomic DNA at 2.5 kV/cm. The transformed cells were incubated at 30°C for 2 h in 1 ml PYCa medium, plated onto lawn plates of *R. erythropolis* (Rery29), and incubated for 2 days prior to being examined for plaques.

Nucleotide sequence accession number. The nucleotide sequence for RRH1 has been deposited in GenBank under accession number JN116822.

RESULTS AND DISCUSSION

Isolation and general features of RRH1 phage. Four phage isolates were obtained using three *Rhodococcus* strains as hosts (two lysed *R. rhodochrous* [Rrho39], one lysed *R. equi* [Requ10], and one lysed *R. erythropolis* [Rery29]). These phage isolates were obtained from samples of activated sludge from three wastewater treatment plants in Victoria, Australia (Daylesford, Werribee, and Ballarat), and from one in Queensland, Australia (Nambour). Sequencing revealed that all four isolates had identical genomes (see below), suggesting that this phage RRH1 may be widely present in Australian activated sludge communities.

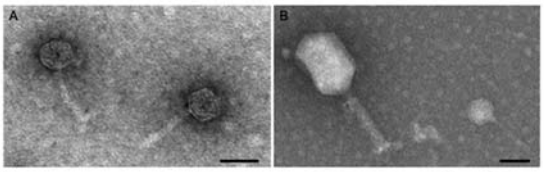

FIG 1 TEM micrograph of RRH1 (A) and RRH1 and T2 (B) (scale bars = 50 nm).

Electron microscopy revealed that RRH1 has a noncontractile tail (~81 ± 1.9 nm) and an isometric capsid (~43 ± 1.2 nm) (Fig. 1A and B). Its morphology is characteristic of *Siphoviridae* phages, with a capsid among the smallest ever reported for a member of this family.

The average burst size was calculated to be 45 ± 5 particles per infective center, with a latency period of ~2 h in PYCa broth at 30°C. RRH1 phage displays a broad host range within the genus *Rhodococcus*. From the bacterial strains tested in our collection, RRH1 can propagate lytically on *R. equi* (Requ28), *R. equi* (Requ10), *R. erythropolis* (Rery29), *R. globerulus* (Rglo35), *R. rhodnii* (Rrho46), and *R. rhodochrous* (Rrho39).

Genome sequencing of RRH1. These four RRH1 phages were each sequenced using Roche/454 pyrosequencing (see Materials and Methods). The data generated over 10,000 reads for each, with a minimum of 50-times sequence coverage, and identical DNA sequences emerged. The assembled sequences of all four phages were consistent with a circularly permuted, double-stranded DNA (dsDNA) genome, a proposition confirmed by PvuII restriction digest profiles (data not shown). Consequently, only that of RRH1 phage isolated on *R. rhodochrous* from the Daylesford treatment plant is detailed here. Its genome consists of 14,270 bp with a G+C content of 68.3 mol% that falls within the range 67 to 73 mol% G+C for *Rhodococcus* species (16). A comparison of the RRH1 genome with those of other *Siphoviridae* family members revealed that this phage has the smallest genome described to date. The most similar phage at the DNA level is a putative prophage sequence found in the complete genome of *Nocardia farcinica* IFM 10152 (23), but their similarity is low (5%). The remainder of the RRH1 genome shares no identity with any other sequenced genome.

The RRH1 genome carries 20 putative genes larger than 90 nucleotides and contains no tRNAs (Table 1). The ORFs are numbered consecutively from *orf1* to *orf20*. All 20 genes are located on the same strand and transcribed in the same direction (Fig. 2). Only 12 ORF products share significant homology with other protein sequences in the GenBank database, and only 6 could be annotated functionally. The RRH1 phage genome appears to be modularly organized (as is typical for the *Siphoviridae*), consisting of gene clusters involved in DNA packaging, cell lysis, and head and tail morphogenesis (Fig. 2). Each is discussed below.

DNA packaging module. The putative large terminase appears to be encoded by *orf2*, as it shares significant similarity with a large terminase found in a prophage of *Corynebacterium diphtheriae*, as well as containing the motif pfam03354 characteristic of large terminases. This gene product is essential for packaging DNA into the phage head during phage replication (34). Typically, the large

TABLE 1 Summary of genes carried by RRH1

ORF[a]	Genome coordinates	Protein function (Pfam)[b]	Match (% identity)[c]	Protein size in kDa (no. of TM domains)[d]	E value[e]
1	219–599	Putative small terminase		13.3 (1)	
2	655–1965	Large terminase (pfam03354)	Terminase protein, *Corynebacterium diphtheriae* (44)	46.5 (0)	$7e^{-85}$
3	1995–2327	Unknown	Hypothetical protein, *Gordonia neofelifaecis* (58)	11.9 (0)	$3e^{-07}$
4	2471–3589	Portal protein (pfam04860)	Hypothetical protein, *Corynebacterium diphtheriae* (52)	40.5 (1)	$7e^{-54}$
5	3586–4320	Lysin (pfam01471)	Hypothetical protein, *Rhodococcus erythropolis* (52)	26.7 (0)	$7e^{-61}$
6	4308–4442	Unknown		4.5 (1)	
7	4439–6424	Prohead protease (pfam04586) and capsid fusion protein	Hypothetical protein, *Gordonia neofelifaecis* (41)	68.4 (4)	$7e^{-115}$
8	6428–6784	Unknown	Hypothetical protein, *Gordonia neofelifaecis* (44)	12.3 (0)	$5e^{-14}$
9	6793–7125	Putative structural protein	Hypothetical protein, *Gordonia neofelifaecis* (44)	11.8 (0)	$6e^{-14}$
10	7140–7571	Unknown	Hypothetical protein, *Gordonia neofelifaecis* (60)	15.0 (0)	$2e^{-41}$
11	7884–8024	Putative structural protein		5.3 (0)	
12	8025–8324	Unknown	Hypothetical protein, *Gordonia neofelifaecis* (36)	10.9 (1)	$3e^{-05}$
13	8442–10040	Putative tape measure protein	Hypothetical protein, *Corynebacterium diphtheriae* (38)	53.3 (6)	$2e^{-75}$
14	10037–11458	Unknown	Hypothetical protein, *Gordonia neofelifaecis* (32)	52.3 (0)	$1e^{-52}$
15	11472–12083	Putative structural protein		21.5 (0)	
16	12087–12329	Unknown		9.0 (1)	
17	12510–13157	Unknown		23.4 (0)	
18	13348–13467	Unknown		4.2 (1)	
19	13445–13912	Putative structural protein		16.0 (1)	
20	13909–14184	HNH endonuclease (pfam01844)	HNH endonuclease, *Corynebacterium diphtheriae* (62)	10.0 (0)	$3e^{-24}$

[a] ORFs were numbered consecutively.
[b] Predicted function is based on amino acid identity, conserved motifs, (ESI-)TOF-MS, and gene location within functional modules.
[c] The most closely related gene (only if named) and the name of the organism. The percent identity was based on the best match when a BlastP analysis was performed.
[d] The predicted protein size and the predicted transmembrane (TM) domains were determined using DAS (11).
[e] The probability of obtaining a match by chance as determined by BLAST analysis. Only values less than 10^{-4} were considered significant.

terminase functions as a complex with a small terminase. The latter is responsible for determining specificity of DNA binding (9), while the large terminase performs DNA cleavage after the packing into the prohead (15). Such terminase genes are typically transcribed together in an operon-like structure, in which the small subunit is located upstream of the large subunit (8). On this basis, we predict that the preceding gene (*orf1*) may encode a novel putative small terminase.

Structural and lysin modules. The portal protein appears to be encoded by *orf4*, based on the presence of the distinctive conserved motif pfam04860. The portal protein gene is typically the first gene in the structural protein module of *Siphoviridae* phages (8), and this arrangement is also the case with RRH1. Phage portal protein genes are typically followed by genes encoding capsid and tail proteins (8). This genomic structure is not maintained in RRH1, in which a putative lysin gene (*orf5*) is found downstream of *orf4*. While the predicted amino acid sequence of Orf5 is most closely related to that of a hypothetical protein in a putative prophage of *Rhodococcus erythropolis*, it also shares identity with the putative lysin proteins of the *Mycobacterium* phage Bethlehem and other related *Mycobacterium* phages (19). In addition, Orf5 contains the pfam01471 motif, composed of three alpha helices and associated with peptidoglycan binding and bacterial cell wall degradation (7). On the basis of these data, we conclude that *orf5* encodes a lysin that is an unusual location within the gene module encoding phage structural proteins. No putative holin could be recognized within this genome, but it is possible the adjacent gene (*orf6*) may encode a novel holin. Holin proteins are typically smaller than 150 amino acids in length, with at least one transmembrane domain and their genes located adjacent to the lysin gene (42). These three criteria are satisfied by *orf6* (Table 1).

The genome region from *orf7* to *orf19* appears to encode the structural proteome of RRH1. SDS-PAGE analysis revealed three major structural proteins (~15 kDa, ~40 kDa, and ~200 kDa) and five minor proteins (~23 kDa, ~26 kDa, ~30 kDa, ~37 kDa, and ~50 kDa) (Fig. 3). The largest protein is likely to result from the covalent self-linking of the capsid proteins, a structural feature observed with other phage capsid proteins (32, 43). The mass spectroscopy data revealed extensive peptide coverage of the C-terminal region of Orf7 but no coverage of the N-terminal re-

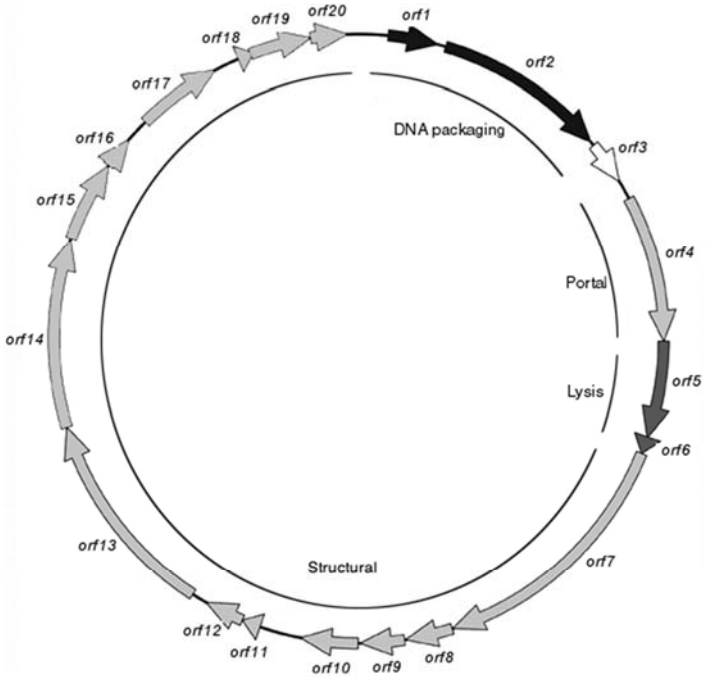

FIG 2 Circular map of the RRH1 genome. The arrows represent the putative genes and the direction they are transcribed. Modules are shaded in similar colors and the inner circle indicates the functional modules.

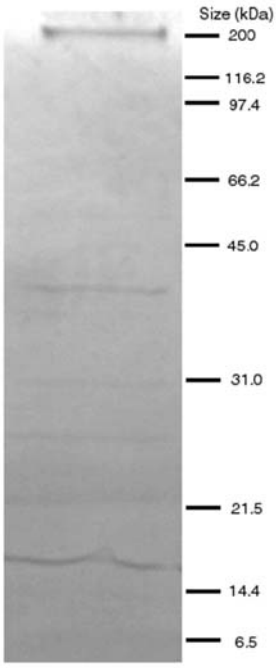

FIG 3 SDS-PAGE electrophoretic analysis of RRH1 proteins.

gion (data not shown). The N-terminal region of Orf7 contains a pfam04586 motif, characteristic of prohead proteases. The predicted size of Orf7 is 68.4 kDa, but from the SDS-PAGE data (Fig. 3), we were unable to recognize a protein of this size. These data suggest that Orf7 is a fusion protein product involving the prohead protease and the major capsid protein, as reported before for the *Lactococcus* phage c2 (25). This hypothesis is supported by the observation that one of the most abundant RRH1 phage proteins (and the likely main capsid protein) is only ~40 kDa (Fig. 3).

SDS-PAGE analysis revealed the presence of a major structural protein of approximately 15 kDa. *In silico* analysis of the 20 putative genes of RRH1 revealed only two putative protein products that match such a size (Orf10 and Orf19). Whole-phage mass spectroscopy revealed 55% peptide coverage of Orf19 but no coverage for Orf10, suggesting that Orf19 is the major structural tail protein. No putative ribosomal slippage sequence in the structural protein genes could be detected, although the sequence has been observed in other *Rhodococcus* phages (32, 41). The genes *orf8* to *orf18* have no predicted function, with the sequences of Orf11 and Orf16 to Orf19 sharing no similarity to any other protein. Mass spectroscopy analysis revealed sequences of both the major structural proteins (Orf7 and Orf19) and identified Orf11, Orf13, and Orf15 as minor structural proteins. The largest putative gene in the RRH1 phage genome is *orf13*. This gene is predicted to encode the tape measure protein, since its closest sequence homologues are those of other phage tape measure proteins.

orf20 encodes a putative endonuclease, based on sequence similarity with a putative HNH endonuclease found in a prophage of

C. diphtheriae and the presence of the conserved motif pfam01844. Surprisingly, the RRH1 phage genome does not appear to carry any DNA replication genes, such as those encoding helicases, primases, or DNA polymerases. The absence of such gene products suggests that this phage relies on its host to provide all the required genome replication machinery.

It is interesting to note that the RRH1 phage displays a broad host range and relatively small burst size (i.e., phage particles produced per cell infected) for such a small phage. It may be that its limited genome size allows it to function only as a generalist of low efficiency. This hypothesis would suggest that most genes present in larger phage genomes are not essential for lytic replication but are instead important for achieving replication efficiency through host specialization and better use of host resources. Phages like RRH1 would be expected to have a competitive advantage over more specialized phages in environments where there is a high diversity of *Rhodococcus* spp., since they would be able to exploit the available host resource more fully but be at a competitive disadvantage in environments with a limited range of strains and where host utilization efficiency is paramount. Such a model may explain why both broad- and narrow-host-range phages for the same host can be isolated.

RRH1 has the smallest known *Siphoviridae* phage genome. The genome size of RRH1 is unusually small (14.2 kb). To our knowledge, no functional *Siphoviridae* phage with a smaller genome has been identified. A search of the NCBI database for complete dsDNA virus genomes using the keywords "bacteriophage" or "phage" identified 537 complete phage genome sequences as of June 2011. The next smallest of the *Siphoviridae* genomes deposited belongs to the *Enterococcus faecalis* phage EFRM31, at 16.95 kb (26, 27).

Can RRH1 be used as a model to study phage gene function? RRH1 is genetically very simple, suggesting that it may be of value in aiding our understanding of the *Siphoviridae*. With the aim of assessing RRH1 phage as a model system, a DNA transformation protocol was developed. Electroporation of naked phage DNA into electrocompetent *R. erythropolis* cells (Rery29) resulted in viable plaques ($\sim 10^3/\mu$g DNA) developing on lawn plates of the host organism. Furthermore, the calculated DNA packing efficiency for RRH1 is at the lower end for a *Siphoviridae* phage (33), suggesting that it may be possible to add additional genes to the RRH1 genome. This outcome raises the feasibility of being able to apply standard *in vitro* molecular biology techniques to study gene function in this phage. Work exploring this attractive possibility is currently in progress.

Conclusion. Phages are considered to represent the most genetically diverse biological entities on the planet (18). How little we still know of phage genomics is emphasized by the RRH1 phage described in this paper. Despite its small genome size and limited number of genes, most of its gene products cannot be assigned a function based on sequence similarities. New and improved model systems are needed to better elucidate how phages reproduce and exploit their hosts. The minimal genome size of phage RRH1 combined with its suitability for gene manipulation and other standard *in vitro* genetic techniques suggests that it may be an excellent candidate for this task.

ACKNOWLEDGMENTS

The research was supported by the Australian Research Council Linkage Grant (LP0774913) in collaboration with Melbourne Water (David Gregory) and South East Water (Graham Short), who are thanked for their financial support. Z. A. Dyson was the recipient of an Australian Postgraduate Award Ph.D. scholarship. S. Petrovski was funded by the ARC linkage grant and La Trobe University.

We thank Carolyn Bell and Judy Oliver for providing bacteriophage T2.

REFERENCES

1. **Ackermann HW.** 1992. Frequency of morphological phage descriptions. Arch. Virol. **124**:201–209.
2. **Ackermann HW.** 1998. Tailed bacteriophages: the order *Caudovirales*. Adv. Virus Res. **51**:135–201.
3. **Ackermann HW.** 2003. Bacteriophage observations and evolution. Res. Microbiol. **154**:245–251.
4. **Adams MH.** 1959. Bacteriophages. Intersciences Publishers, Inc., New York, NY.
5. **Bateman A, et al.** 2004. The Pfam protein families database. Nucleic Acids Res. **32**:D138–D141.
6. **Billion A, Ghai R, Chakraborty T, Hain T.** 2006. Augur–a computational pipeline for whole genome microbial surface protein prediction and classification. Bioinformatics **22**:2819–2820.
7. **Briers Y, et al.** 2007. Muralytic activity and modular structure of the endolysins of Pseudomonas aeruginosa bacteriophages phiKZ and EL. Mol. Microbiol. **65**:1334–1344.
8. **Brüssow H, Desiere F.** 2001. Comparative phage genomics and the evolution of *Siphoviridae*: insights from dairy phages. Mol. Microbiol. **39**: 213–222.
9. **Catalano CE.** 2000. The terminase enzyme from bacteriophage lambda: a DNA-packaging machine. Cell. Mol. Life Sci. **57**:128–148.
10. **Catalano CE, Cue D, Feiss M.** 1995. Virus DNA packaging: the strategy used by phage λ. Mol. Microbiol. **16**:1075–1086.
11. **Cserzö M, Wallin E, Simon I, von Heijne G, Elofsson A.** 1997. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng. **10**:673–676.
12. **Delcher AL, Bratke KA, Powers EC, Salzberg SL.** 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics **23**:673–679.
13. **Delcher AL, Harmon D, Kasif S, White O, Salzberg SL.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27**: 4636–4641.
14. **Finn RD, et al.** 2010. The Pfam protein families database. Nucleic Acids Res. **38**:D211–D222.
15. **Fujisawa H, Morita M.** 1997. Phage DNA packaging. Genes Cells **2**:537–545.
16. **Goodfellow M, Maldonado LA.** 2006. The families *Dietziaceae*, *Gordoniaceae*, *Nocardiaceae* and *Tsukamurellaceae*. *In* Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E, (ed.) The prokaryotes. Archaea. Bacteria: firmicutes, actinomycetes. Springer, New York, NY.
17. **Haft DH, Selengut D, White O.** 2003. The TIGRFAMs database of protein families. Nucleic Acids Res. **31**:371–373.
18. **Hatfull GF.** 2008. Bacteriophage genomics. Curr. Opin. Microbiol. **11**: 447–453.
19. **Hatfull GF, Cresawn SG, Hendrix RW.** 2008. Comparative genomics of the mycobacteriophages: insights into bacteriophage evolution. Res. Microbiol. **159**:332–339.
20. **Hatfull GF, et al.** 2010. Comparative genomic analysis of 60 Mycobacteriophage geneomes: genome clustering, gene acquisition and gene size. J. Mol. Biol. **397**:119–143.
21. **Hemmerich C, Buechlein A, Podicheti R, Revanna KV, Dong Q.** 2010. An Ergatis-based prokaryotic genome annotation web server. Bioinformatics **26**:1122–1124.
22. **Hendrix RW.** 2003. Bacteriophage genomics. Curr. Opin. Microbiol. **6**:506–511.
23. **Ishikawa J, et al.** 2004. The complete genomic sequence of *Nocardia farcinica* IFM 10152. Proc. Natl. Acad. Sci. U. S. A. **101**:14925–14930.
24. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**: 955–964.
25. **Lubbers MW, Waterfield NR, Beresford TP, Le Page RW, Jarvis AW.** 1995. Sequencing and analysis of the prolate-headed lactococcal bacteriophage c2 genome and identification of the structural genes. Appl. Environ. Microbiol. **61**:4348–4356.

26. **Mazaheri Nezhad Fard R, Barton MD, Heuzenroeder MW.** 2010. Novel bacteriophages in Enterococcus spp. Curr. Microbiol. **60**:400–406.
27. **Mazaheri Nezhad Fard R, Barton MD, Arthur JL, Heuzenroeder MW.** 2010. Whole-genome sequencing and gene mapping of a newly isolated lytic enterococcal bacteriophage EFRM31. Arch. Virol. **155**:1887–1891.
28. **Meijer WG, Prescott JF.** 2004. *Rhodococcus equi*. Vet. Res. **35**:383–396.
29. **Orvis J, et al.** 2010. Ergatis: a web interface and scalable software system for bioinformatics workflows. Bioinformatics **26**:1488–1492.
30. **Petrovski S, et al.** 2011. An examination of the mechanisms for stable foam formation in activated sludge systems. Water Res. **45**:2146–2154.
31. **Petrovski S, Seviour RJ, Tillett D.** 2011. Genome sequence and characterization of the Tsukamurella phage TPA2. Appl. Environ. Microbiol. **77**:1389–1398.
32. **Petrovski S, Seviour RJ, Tillett D.** 2011. Characterization of the genome of the polyvalent lytic bacteriophage GTE2, which has potential for biocontrol of *Gordonia*-, *Rhodococcus*- and *Nocardia*-stabilized foams in activated sludge plants. Appl. Environ. Microbiol. **77**:3923–3929.
33. **Purohit PK, et al.** 2005. Forces during bacteriophage DNA packaging and ejection. Biophys. J. **88**:851–866.
34. **Rao VB, Feiss M.** 2008. The bacteriophage DNA packaging motor. Annu. Rev. Genet. **42**:647–681.
35. **Rohwer F.** 2003. Global phage diversity. Cell **18**:113–141.
36. **Santos MA.** 1991. An improved method for small scale preparation of bacteriophage DNA based on phage precipitation by zinc chloride. Nucleic Acids Res. **19**:5442.
37. **Schattner P, Brooks AN, Lowe TM.** 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. **33**:W686–W689.
38. **Sekizaki T, et al.** 1998. Improved electroportation of *Rhodococcus equi*. J. Vet. Med. Sci. **60**:277–279.
39. **Shapiro OH, Kushmaro A.** 2011. Bacteriophage ecology in environmental biotechnology processes. Curr. Opin. Biotechnol. **22**:449–455.
40. **Shevchenko A, et al.** 1996. A strategy for identifying gel-separated proteins in sequence databases by MS alone. Biochem. Soc. Trans. **24**: 893–896.
41. **Summers EJ, et al.** 2011. Genomic and functional analysis of *Rhodococcus equi* phages ReqiPepy6, ReqiPoco6, ReqiPine5 and ReqiDocB7. Appl. Environ. Microbiol. **77**:669–683.
42. **Wang I, Smith DL, Young R.** 2000. Holins: the protein clocks of bacteriophage infections. Annu. Rev. Microbiol. **54**:799–825.
43. **Wikoff WR, et al.** 2000. Topologically linked protein rings in the bacteriophage HK97 capsid. Science **289**:2129–2133.

WATER RESEARCH 45 (2011) 2146–2154

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/watres

An examination of the mechanisms for stable foam formation in activated sludge systems

Steve Petrovski, Zoe A. Dyson, Eben S. Quill, Simon J. McIlroy, Daniel Tillett, Robert J. Seviour*

Biotechnology Research Centre, La Trobe University, PO Box 199, Bendigo, Victoria 3552, Australia

ARTICLE INFO

Article history:
Received 26 August 2010
Received in revised form
21 December 2010
Accepted 23 December 2010
Available online 28 December 2010

Keywords:
Activated sludge
Bacillus subtilis
Foaming
Mycolata
Surface scum
Surfactants

ABSTRACT

Screening pure cultures of 65 mycolic acid producing bacteria (Mycolata) isolated mainly from activated sludge with a laboratory based foaming test revealed that not all foamed under the conditions used. However, for most, the data were generally consistent with the flotation theory as an explanation for foaming. Thus a stable foam required three components, air bubbles, surfactants and hydrophobic cells. With non-hydrophobic cells, an unstable foam was generated, and in the absence of surfactants, cells formed a greasy surface scum. Addition of surfactant converted a scumming population into one forming a stable foam. The ability to generate a foam depended on a threshold cell number, which varied between individual isolates and reduced markedly in the presence of surfactant. Consequently, the concept of a universal threshold applicable to all foaming Mycolata is not supported by these data. The role of surfactants in foaming is poorly understood, but evidence is presented for the first time that surfactin synthesised by *Bacillus subtilis* may be important.

1. Introduction

The generation of stable foam on the surface of aerated reactors is a common feature of activated sludge systems around the world (de los Reyes, 2010; Soddell, 1999; Soddell and Seviour, 1990). In attempts to seek strategies for its control, considerable effort has been directed at better understanding the microbial ecology of foaming (de los Reyes, 2010; Kragelund et al., 2007; Müller et al., 2007; Seviour et al., 2008). Microscopic examination reveals that most foams contain either long unbranched Gram positive filaments of *Candidatus* 'Microthrix parvicella' or short Gram positive branched filaments of mycolic acid producing bacteria (Mycolata) (de los Reyes, 2010; Kragelund et al., 2007; Seviour et al., 2008). The latter may fragment during their life cycles into coccoid unicells, which are often reported in foams (de los Reyes, 2010).

In many of the early microbiological foam surveys, the molecular techniques now available to allow their unequivocal identification were not applied (de los Reyes, 2010; Nielsen et al., 2009; Seviour et al., 2008), and so attempts to relate foaming incidents to specific operational conditions were compromised by the methodology used. It is now clear from these molecular approaches that a single bacterial morphotype may include several phylogenetically unrelated organisms, differing markedly often in their physiology and ecology (Nielsen et al., 2009; Seviour et al., 2008; Soddell, 1999). For example, the *Gordonia amarae* like organisms (GALO) morphotype of right-angled branching Gram positive filaments is shared by members of several other genera in the Mycolata (de los Reyes, 2010; Seviour et al., 2008; Soddell, 1999). On the other hand the branching frequencies and angles of the pine tree like organisms (PTLO) can vary considerably in activated sludge.

* *Corresponding author*. Tel.: +61 3 5444 7868; fax: +61 3 5444 7476.
E-mail address: s.petrovski@latrobe.edu.au (S. Petrovski).
0043-1354/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.watres.2010.12.026

Yet on isolation all these differing morphotypes belong to a single bacterial species *Skermania piniformis* (Soddell and Seviour, 1998). Other phylogenetically unrelated GALO and PTLO probably exist in activated sludge communities (Kragelund et al., 2007), emphasising that our present understanding of the microbial ecology of foams is incomplete.

The current view is that these stable foams are generated by a selective enrichment of hydrophobic bacteria in them by a process of flotation (de los Reyes, 2010; Soddell and Seviour, 1990). Flotation requires three components: gas bubbles surrounded by liquid films, generated by the aeration system; surfactants which reduce the surface tension and thus prevent liquid drainage from the gas bubble walls; and small hydrophobic particles (the bacterial cells), responsible for the long term stabilisation of such foams (Blackall and Marshall, 1989; Soddell and Seviour, 1990).

According to this flotation model, all hydrophobic bacteria, regardless of their morphology have the potential to stabilise foams, and so populations other than those discussed above are probably involved in stable foaming incidents, even if not as major contributors (Lemmer et al., 2005). Equally, in the absence of hydrophobic particles, any foam that might develop on reactors would be expected to be ephemeral, and rapidly collapse. Yet whether this is the case remains unknown.

Fluorescence *in situ* hybridisation (FISH) has been used to elucidate possible relationships between Mycolata population sizes in mixed liquor samples and the onset of foaming incidents (Davenport et al., 2000; de los Reyes and Raskin, 2002). Two foaming threshold values have been proposed for *Gordonia* spp by de los Reyes and Raskin (2002). The *formation threshold* was thought to correspond to the foaming potential of the mixed liquor, while the *stability threshold* was considered to be the population level at which a stable foam developed. Davenport et al. (2000) estimated their foaming thresholds with a FISH probe targeting all the known Mycolata, and expressed it as number of cells ml^{-1} mixed liquor. So it is not possible to compare their values with those of de los Reyes and Raskin (2002), who based theirs on whole filament lengths. Davenport et al. (2008) extended this approach to suggest that their threshold was a universal value applicable to all foaming activated sludge plants.

This proposal was based on several untested assumptions. Thus, all Mycolata cells were assumed to possess the same cell surface hydrophobicity and propensity to foam regardless of plant operating conditions. Yet many foaming plants contain more than a single foam stabilising Mycolata population. Furthermore, whether their foaming threshold value might be affected by the presence of surfactants, which will probably vary in type and concentration within and among individual plants, was not considered. FISH probes are not available for all foaming Mycolata and will only allow metabolically active cells to be quantified (Amann and Ludwig, 2000). Any dead or moribund Mycolata likely to be present (Kragelund et al., 2007), which may retain their hydrophobicity and hence foam stabilising properties, will not be included in any FISH based enumeration.

Therefore this important question of whether such a universal stability threshold value exists for Mycolata is not yet resolved conclusively and needs to be examined further.

Equally only a small selection of the Mycolata isolated from foams have been used in standardised pure culture experiments in attempts to understand what factors might determine their foaming potentials (de los Reyes and Raskin, 2002). So whether this flotation theory applies to all Mycolata is unknown, as is the role if any of surfactants in foam generation and foaming capacity. Where these experiments have been carried out, usually with small numbers of '*Nocardia*' (probably *Gordonia*) and *Rhodococcus* foam isolates, the data are often confusing and contradictory (Ho and Jenkins, 1991; Stratton et al., 1998, 2003).

Consequently our study set out to examine the foaming behaviour under standardised laboratory conditions of pure cultures of most of the known non-pathogenic Mycolata, especially those isolated from foams in attempts to see whether flotation explains their foaming behaviour. It also examined closely whether the concept of a universal threshold value for Mycolata foaming stabilisation has any credibility. Particular attention in this study was paid to the possible role of surfactants in stable foaming incidents, and their influence on Mycolata foaming behaviour.

2. Materials and methods

2.1. Isolates

Sixty five Mycolata strains listed in Table 1 were used. Their identity was confirmed in this study by partial sequencing (>500 bp) of their 16S rRNA genes, which in all cases were 100% identical to those generated previously for them. All were grown either in peptone yeast extract (PYCa; 0.5 g/L peptone, 0.3 g/L yeast extract, 0.1 g/L glucose and 0.1 g/L calcium chloride) broth or in an activated sludge mixed liquor medium at 30 °C for two to seven days. This medium used mixed liquor from the Carrum plant (Victoria, Australia) and was first filter sterilised through a 0.22 μm nitrocellulose filter (Millipore). The filtered effluent was supplemented with several different carbon sources as detailed later.

2.2. Assessment of foaming ability of the Mycolata strains

A foaming apparatus described by Stratton et al. (2002) with a sintered glass disc fitted to its base was connected to a rotameter. A 20 ml aliquot of each Mycolata broth culture (A_{600} adjusted to 1.0) was added to the cylinder and aerated at 100 ml/min for 1 min. Adjusting culture absorbance was achieved by diluting each culture in its own supernatant to ensure that any exocellular surfactants present there are not removed. Foaming abilities were assessed using modified criteria of Blackall and Marshall (1989), and distinctions between foam and scum formation made as described later (Table 2).

Foaming thresholds were determined by assessing foaming abilities of strains over a range of different cell densities as determined by their A_{600}. The lowest A_{600} supporting stable foam formation (≥category 3 in Table 2) and the corresponding cfu/ml on the appropriate medium (see above) were then determined.

Table 1 – Strains used in this study.

Name of organism	Culture collection numbers and other synonyms[a]
Gordonia sp.	
G. aichiensis	Raic22^T, DSMZ 43978
G. alkanivorous	Ben606
G. amarae	Gama44^T, DSMZ 43392
G. amarae	Gama9, UQCC2810
G. amictica	Ben607
G. australis	18F3M
G. defluvii	J4^T, DSMZ 44981
G. desulfuricans	213E^T, NCIMB 40816
G. hydrophobica	N1123^T, DSMZ 44015
G. malaquae	A554^T, ATCC 35215
G. malaquae	A448
G. polysoprenororaus	Ben605
G. rubropertincta	Grub48^T, DSMZ 43197
G. sputi	Gspu49^T, ATCC 29627
G. sputi	Gspu48, ATCC 336609
G. terrae	Gter34^T, DSMZ 43249
G. terrae	Ben601
G. terrae	Ben602
G. terrae	Ben603
G. terrae	Ben604
Rhodococcus sp.	
R. coprohilus	Rcop41^T, DSMZ 43347
R. coprohilus	Rcop18, UQCC 1259
R. equi	Requ10, UQCC 702
R. equi	Requ28^T, UQCC20307, DSMZ 20307
R. erythropolis	Rery19, UQCC 379
R. erythropolis	Rery29^T, DSMZ 43066
R. globerulus	Rglo35^T, DSMZ 43954
R. luteus	IMV 385^T, AUCNM A-594
R. obuensis	ATCC 33610^T
R. rhodnii	Rrho46^T, DSMZ 43336
R. rhodochrous	Rrho3, UQCC 2807
R. rhodochrous	Rrho39^T, DSMZ 43241
R. rhodochrous	Rrho11S, UQCC 2808
R. ruber	Rrub33^T, DSMZ 43338
R. tritomae	DSM 44892^T
Nocardia sp.	
N. asteroides	Nast23^T, DSMZ 43757
N. asteroides	Noast4, UQCC 131
N. brasiliensis	Nbra42^T, DSMZ 43758
N. carnea	Ncar30^T, DSMZ 43397
N. nova	Nnov47^T, ATCC 33726
N. otitdidiscariarum	Noti14, AMMRL 19.11
N. otitdidiscariarum	Noti25^T, DSMZ 43242
N. otitdidiscariarum	Noti1S, AMMRL 19.12
N. transvalensis	Ntra40^T, DSMZ 43405
Tsukamurella sp.	
T. inchonensis	DSMZ 44067^T
T. inchonensis	NCTC 10741
T. paurometabola	Tpau37^T, ATCC25938
T. paurometabola	NCTC107411
T. paurometabola	IMRU1283
T. paurometabola	DSMZ 20162
T. paurometabola	IMRU 1520, M334, DSMZ 44119
T. paurometabola	IMRU 1312, M343
T. paurometabola	IMRU 1505, M337
T. pseudospumae	N1176^T, DSMZ 44118
T. pulmonis	DSM44142
T. spumae	N1171^T, DSMZ 44113, NCIMB 139647
T. spumae	JC85
T. tyrosinosolvens	DSMZ 44234^T
Mycobacterium sp.	
M. chlorophenolicus	Mchl24^T, DSMZ 43826
M. smegmatis	Msme1, UQCC 120
M. fortuitum	Mfor21, UQCC 422
Other	
Dietzia maris	Dmar27^T, DSMZ 43672
Streptomyces griseus	Sgri05
Millisia brevis	J82^T, DSMZ 44463

^{T} Type strain.
All other culture numbers have been derived from the La Trobe University Bendigo culture collection.
a Culture collection numbers have been obtained from the following organisations; DSMZ = German collection of microorganisms and cell cultures. ATCC = American type culture collection. UQCC = Australian collection of microorganisms (now ACM). AMMRL = Australian national reference laboratory in medical mycology. NCIMB = National collection of industrial bacteria.

2.3. *Determination of cell surface hydrophobicity and surface tension*

Cell surface hydrophobicities were determined using the microbial adherence to hydrocarbon (MATH) assay with n-hexadecane (Rosenberg et al., 1980) as the solvent. The percentage cell hydrophobicities were calculated by

Table 2 – Foaming assay scale.

Rating	Description
0	As for pure water; No foam
1	1.0–3.0 cm of foam with fragile ill formed bubbles. Insufficient stability to form films. Immediate collapse on cessation of aeration.
1a[a]	Flotation of clumped bacterial cells to the surface of the air–water interface. Clumped cells remain afloat upon cessation of aeration producing a scum layer.
2	Intermitted films sufficiently stable. Usually generated from a fragile foam structure of limited height. Films unstable on cessation of aeration.
3	Substantial foaming (i.e., bubbles about 10 cm diameter) to 3–8 cm height. Infrequent or regular film formation, with both film and foam semi-stable on cessation of aeration.
4	Initially 8–15 cm of foam (about 1 cm diameter bubbles) with stable films being formed at regular intervals. Body of the foam and films stable for 3–5 min once aeration ceases.
5	Stable foam 5–10 cm in height in 2 min, after which collapse to 3–5 cm height. Foam is stable when aeration ceases. No films.
6	Stable foam 15–30 cm in height with no films. Bubble size about 0.5 cm during aeration and increases to 2.0–3.0 cm diam. in 3–5 min from the time aeration ceases.

a This key has been added to represent the 'scummers'.

determining the A_{600} before and after n-hexadecane addition. Surface tensions of broth culture media were obtained by the Wihelmy plate method using an Analyte surface tension metre Model 2141 (Mc Van Instrument Pty Ltd, Australia).

2.4. 16S rRNA gene sequencing

Amplification of 16S rDNA genes was performed by colony PCR using universal primers 27f (AGAGTTTGATCMTGGCTCAG) and 1525r (AAGGAGGTGWTCCARCC) (Lane, 1991). The reaction mixture (final volume 30 μl) contained the following components: ~100 ng of template DNA, 0.2 mM dNTPs, 1× PCR reaction buffer, 1.5 mM $MgCl_2$, 1 μg of each oligonucleotide (Geneworks), 2.5 U Taq DNA polymerase and 10% (v/v) DMSO. The mixture was subjected to 36 thermal cycles, as follows: 92 °C, 3 min (first cycle only); 92 °C, 1 min; 52 °C, 70 s; 72 °C, 2 min; 72 °C, 5 min (last cycle only). The presence of appropriate reaction product was assessed by agarose gel electrophoresis. PCR products were cleaned up using a Wizard SV gel and PCR clean-up kit (Promega) and sequenced at Australian Genome Research Facility (AGRF, University of Queensland, Australia).

2.5. Fluorescence in situ hybridisation (FISH)

Samples were fixed in 50% [v/v] ethanol overnight at 4 °C and stored at −20 °C. Fluorescence *in situ* hybridisation (FISH) was performed using the protocol of Daims et al. (2005). The LGC354B FISH probe (5′-CGGAAGATTCCCTACTGC-3′), targeting some members of the *Firmicutes* (Meier et al., 1999) was used to screen foam samples for bacteria of interest (see later). Pre-treatment of the biomass with achromopeptidase, lysozyme and mild acid hydrolysis as detailed by Kragelund et al. (2007), together with extended FISH hybridisation times of 16 h, did not make an appreciable difference to the FISH signal fluorescence intensities.

2.6. High performance liquid chromatography (HPLC)

Surfactin concentrations in mixed liquors and foams were determined using a modification of the method described by Gong et al. (2009). Analyses by HPLC used a Shimadzu LC-10Ai fitted with a SCL-10Avp PDA detector and a Synergy Hydro-RP column (150 × 4.6 mm I.D., 4 μm, Phenomenex) with a mobile phase of 3.8 mM trifluoroacetic acid (20%) and acetonitrile (80%) at a flow rate of 1 ml/min. All samples (20 μL) were run after a 20 min column equilibration period. UV spectra of surfactin were collected during analysis with the PDA detector, revealing a λ_{max} of 205 nm, which was used for surfactin quantification. Surfactin standards (purchased from Sigma–Aldrich) were prepared from a stock solution of 0.0016 g surfactin in 10 ml analytical grade methanol and sufficient Milli-Q water added to bring the final solution volume to 100 ml.

2.7. Nucleotide sequence accession number

The nucleotide sequence for the 16S rRNA gene of *B. subtilis* has been deposited in GenBank under accession number HQ190905.

3. Results

3.1. Foaming ability of Mycolata cells and insights into the flotation theory in activated sludge foaming

Sixty five independently isolated Mycolata strains mainly from activated sludge systems behaved differently in their abilities to form stable foams under the conditions chosen here. While the majority of strains produced foams, the amount produced and its stability varied considerably (Table 3). With some strains e.g. *Gordonia malaquae*, *Rhodococcus coprophilus* and *Tsukamurella spumae*, no foam was generated, but instead a greasy scum layer was formed at the air–water interface after aeration, and these are described here as scummers. The production of this scum layer seemed to occur exclusively only with those cultures growing as large aggregates or clumps, and where unicells in the liquid medium were rare (ie broth turbidity was very low). The foaming capacity scoring system of Blackall and Marshall (1989) does not allow for such scum producing organisms, therefore we have amended this table to incorporate the scummers (Table 2).

Surface tension measurements were used as indicators of surfactant production levels of individual cultures. Most broth cultures of strains producing a scum or a foam at the low arbitrary foaming level of 1 had surface tensions >60 nm/M (Table 3). Those with foaming capacities of ≥2 gave generally lower surface tension values, although there were exceptions to this generalisation, as with *Rhodococcus equi* (Requ28). This culture gave a very persistent foam (level 6), but no correspondingly large drop in broth surface tension. Cell surface hydrophobicity determinations on all these Mycolata strains revealed that they were all hydrophobic, although the CSH values, reflecting the percentages of hydrophobic cells in each suspension, varied considerably among them (Table 3). When the cells of *N. otitidiscaviarum* (Noti25), *R. equi* (Requ28) and *G. amarae* (Gama44) were rendered no longer viable after autoclaving and subjected to the foaming test, they were still able to produce foam. Even allowing for the limitations known to exist with using the MATH assay (Rosenberg, 2006; Stratton et al., 2002), these results raised questions as to whether a universal threshold value for stable foaming incidents caused by these Mycolata is valid. Consequently, attempts were made to clarify this.

3.2. Does a universal Mycolata foaming threshold value exist?

Whether the same number of cells is required to obtain stable foams with pure culture of all foaming Mycolata strains was examined. The data presented here clearly show that foaming threshold values varied with individual strain under the controlled conditions used in this study, as the data shown in Table 3 demonstrate. Although it appears that stable foams are only generated at high cell numbers, these vary from 2.5×10^6 cfu ml^{-1} with *G. sputi* (Gspu48) to 1.5×10^9 cfu ml^{-1} with *G. terrae* (Ben601). *G. amarae* the most commonly recorded Mycolata in activated sludge foams (de los Reyes, 2010) requires 1.5×10^8 cfu ml^{-1} to produce a stable foam under these conditions, which is in the higher range when compared to the other Mycolata strains (Table 3).

Table 3 – Foaming capacity, hydrophobicity, surface tension and threshold results of Mycolata cultures tested.

Strain[a]	Foaming capacity[b]	Hydrophobicity (%)[c]	Surface tension (nm/M)	Foaming threshold (cfu/ml)[d]
Water[e]	0	–	75.0	NA
Broth[e]	0	–	66.6	NA
G. aichiensis (Raic22)	6	++++	50–55	2.0×10^8
G. alkanivorous (Ben606)	1	++	60–65	5.5×10^8
G. amarae (Gama44)	3	+++	55–60	1.5×10^8
G. amarae (Gama9)	1a	+++	55–60	NA
G. amictica (Ben607)	1	++++	65–70	4.0×10^8
G. australis (18F3M)	1a	++++	60–65	NA
G. defluvii (J4)	4	++++	65–70	4.1×10^7
G. desulfuricans (213E)	2	+++	60–65	7.3×10^6
G. hydrophobica (N1123)	3	++++	55–60	1.0×10^8
G. malaquae (A554)	1a	++++	60–65	NA
G. malaquae (A448)	1a	++++	60–65	NA
G. polysoprenororaus (Ben605)	3	++	55–60	5.1×10^8
G. rubropertincta (Grub48)	1	+++	65–70	2.6×10^7
G. sputi (Gspu49)	1a	++++	>65	NA
G. sputi (Gspu48)	2	++++	55–60	2.5×10^6
G. terrae (Gter34)	2	+++	60–65	4.2×10^8
G. terrae (Ben601)	2	+++	55–60	1.5×10^9
G. terrae (Ben602)	2	+++	65–70	4.0×10^8
G. terrae (Ben603)	1a	+++	65–70	NA
G. terrae (Ben604)	1	+++	65–70	1.7×10^8
R. coprophilus (Rcop41)	1a	++++	>70	NA
R. coprophilus (Rcop18)	1a	++	65–70	NA
R. equi (Requ10)	2	+++	60–65	3.4×10^8
R. equi (Requ28)	6	++	55–60	1.0×10^7
R. erythropolis (Rery19)	4	+++	60–65	4.0×10^8
R. erythropolis (Rery29)	6	+++	50–55	2.0×10^8
R. globerulus (Rglo35)	5	++	55–60	1.0×10^8
R. luteus (IMV 385)	5	++	50–55	3.9×10^8
R. obuensis (ATCC 33610)	2	+++	60–65	1.0×10^8
R. rhodni (Rrho46)	2	+++	55–60	1.3×10^8
R. rhodochrous (Rrho3)	2	++	55–60	1.8×10^8
R. rhodochorus (Rrho39)	5	++++	55–60	1.0×10^7
R. rhodochorus (Rrho11S)	2	++	55–60	2.0×10^8
R. ruber (Rrub33)	2	++	55–60	2.1×10^8
R. tritomae (DSM44892)	3	++++	60–65	8.0×10^7
N. asteroides (Nast23)	4	+++	55–60	2.0×10^8
N. asteroides (Noast4)	6	++++	60–65	1.0×10^7
N. brasiliensis (Nbra42)	1a	+++	60–65	NA
N. carnea (Ncar30)	3	++	55–60	8.5×10^7
N. nova (Nnov47)	1a	++	60–65	NA
N. otitidiscaviarum (Noti14)	6	++++	55–60	2.0×10^7
N. otitidiscaviarum (Noti25)	5	++++	60–65	1.0×10^8
N. otitidiscaviarum (Noti15)	1a	++++	65–70	NA
N. transvalensis (Ntra40)	1a	++	55–60	NA
Tsukamurella inchonensis (DSMZ 44067)	1a	++++	65–70	NA
T. paurometabola (Tpau37)	3	+++	45–50	5.1×10^7
T. paurometabola (DSM20162)	3	+++	35–40	1.7×10^8
T. paurometabola (IMRU1520)	4	+++	55–60	5.0×10^7
T. paurometabola (IMRU1312)	1a	+++	65–70	NA
T. paurometabola (IMRU1505)	1a	+++	65–70	NA
T. paurometabola (NCTC107411)	1a	+++	65–70	NA
T. pseudospumae (N1176)	1a	++++	65–70	NA
T. pulmonis (DSM44142)	3	++++	55–60	5.0×10^7
T. spumae (N1171)	1a	++++	65–70	NA
T. spumae (JC85)	1a	++++	60–65	NA
T. tyrosinosolvens (DSMZ 44234)	1a	++	55–60	NA
M. chlorophenolicus (Mchl24)	3	++	55–60	5.0×10^7
M. smegmatis (Msme1)	1a	++++	60–65	NA

Table 3 (continued).

Strain[a]	Foaming capacity[b]	Hydrophobicity (%)[c]	Surface tension (nm/M)	Foaming threshold (cfu/ml)[d]
M. fortuitum (Mfor21)	1a	++++	65–70	NA
D. maris (Dmar27)	1	++	55–60	4.0×10^8
S. griseus (Sgri05)	1	+++	55–60	3.1×10^8
M. brevis (J82)	1a	++++	60–65	NA

NA = Not applicable.
a Strains used in this study.
b Foaming capacity determined by the foaming assay described by Blackall and Marshall (1989). The scale is represented in Table 2.
c Cell hydrophobicity. % cell hydrophobicity; − = 0%; + = 1%–30%; ++ = 31%–60%; +++ = 61%–80%; ++++ = 80%–100%.
d The average minimum number of cells required to form a stable foam.
e Controls that contain no bacteria.

Of course this pure culture behaviour may not represent the situation in activated sludge systems, where mixtures of substrates are available to these organisms, and so each of the cultures listed in Table 1 was also grown in activated sludge mixed liquor. Carbon and nitrogen sources (glucose and yeast extract) were added to this clarified mixed liquor to allow the cells to grow and cell numbers to increase to the levels needed for stable foam formation. These cultures grown in mixed liquor gave similar threshold values to those obtained in artificial media (e.g. *G. amarae* (Gama44) = 2.5×10^8 cfu ml^{-1}; *G. aichiensis* (Raic22) = 1.5×10^8 cfu ml^{-1}, *N. asteroides* (Nast23) = 2.0×10^8 cfu ml^{-1} and *T. paurometabola* (Tpau37) = 3.5×10^7 cfu ml^{-1}).

The presence of any biosurfactants, likely to vary in both their composition and concentrations between plants would be expected to affect substantially these individual foaming threshold values, as shown below.

3.3. Investigation of biosurfactant producing organisms in foaming WWTP

Addition of a surfactant (Triton-X 100) to pure cultures of all of the scumming Mycolata seen in this study converted all of them to stable foam formers. For example after the addition of Triton-X 100 to the scum producing bacterial cultures the foaming capacity increased from 1a to 3 for *G. terrae* (Ben603), *N. brasiliensis* (Nbra42) 4, *T. spumae* (N1171) and increased to 4 for *R. coprophilus* (Rcop18) and *M. smegmatis* (Msme1). Thus, it is clear that biosurfactants impact considerably on the foaming capacities of these Mycolata, and the flotation model incorporates an essential role for them. Some will enter the plants in the raw influent, yet comparatively little attention has been given previously to which surface active agents might be produced by the existing activated sludge communities (de los Reyes, 2010).

A sample from a biodiesel producing activated sludge plant in northern Victoria with a very serious foaming problem was sent to this laboratory for microscopic examination. The foam, overflowing from the reactor and responsible for severe loss of biomass appeared greyish in colour but had collapsed completely within three days of collection. It contained no Mycolata, but instead was dominated by Gram positive rod shaped bacteria growing in chains. Streaking this sample onto PYCa agar yielded three different colony types, which were repeatedly subcultured until a pure culture of each was obtained, as assessed by microscopy. Colony PCR was used to generate partial sequences of each of their 16S rRNA genes, identifying tentatively one of the isolates as *Bacillus subtilis*.

When pure cultures of this *B. subtilis* were examined by the laboratory foaming test, this organism produced considerable amounts of foam, far in excess of the foams formed by any of the Mycolata cultures examined earlier. Yet this foam rapidly collapsed. The MATH assay showed that these *B. subtilis* cells were not hydrophobic, but the surface tension of its culture medium decreased sharply after 3 days incubation from 66.6 nm/M to 29.5 nm/M (Table 4), suggesting it was producing surfactant material. HPLC analysis of the culture medium identified this surfactant as surfactin, known to be synthesised by *B. subtilis* strains especially when grown with glucose as carbon source (Besson and Michel, 1992).

3.4. How common is B. subtilis in activated sludge foaming plants?

A survey of foaming activated sludge plants in Australia suggested that *B. subtilis* is commonly found in foams. Thus, when foams from 12 plants in New South Wales, Victoria and South Australia were streaked onto PYCa agar, a variety of colony morphologies were seen. Each was screened for its foaming ability using the foaming assay. All isolates producing foams in excess of 10 cm were identified after 16S rRNA sequencing, and their ability to synthesise surfactants assessed by surface tension changes in their culture medium after incubation (Table 4). Isolates of *B. subtilis* sharing >99% similarity in their 16S rRNA sequences and producing an unstable foam were obtained from six of the 12 plants.

Analysis of these foam samples suggested that each was dominated by a different hydrophobic bacterial population (see Table 5). *In situ* analysis of these was then carried out. It was not possible to design a probe specific for the 16S rRNA sequence of our *B. subtilis* isolates, or an encompassing phylotype of closely related sequences, because of their high levels of sequence similarity shared with other related organisms. Differentiation of individual *Bacillus* species is problematic because of the highly conserved nature of their 16S rRNA genes (Rooney et al., 2009). Subsequent FISH analysis applying the LGC354B FISH probe targeting some of the members of the *Firmicutes*, including the *B. subtilis* strains

Table 4 – Bacterial isolated from foam samples producing surfactant.

Source WWTP	Strain number	Foam height (cm)	Surface tension (mn/M)	Putative identity based on 16S rRNA sequence
Biodiesel	BioD2	17	29.5	*Bacillus subtilis*
Carrum	ETP1	16	32.3	*Bacillus subtilis*
Carrum	ETP2	10	70.9	*Pseudomonas nitroreducens*
Carrum	ETP4	20	46.8	*Bacillus subtilis*
Pakenham	Pak3	16	51.5	*Bacillus subtilis*
Liverpool	Liv1	16	40.1	*Bacillus subtilis*
Liverpool	Liv3	10	77.0	*Pseudomonas alcaligenes*
Somers	Som4	13	72.5	*Klebsiella* sp.
Heatherton	Hea1	15	71.4	*Acinetobacter xiamenensis*
Heatherton	Hea2	16	32.9	*Bacillus subtilis*
Heatherton	Hea3	14	23.3	*Bacillus pumilus*
Heatherton	Hea4	12	55.8	*Rhodococcus erythropolis*
Mt Martha	MtMA3	16	31.1	*Bacillus subtilis*
Boneo	BonC3	16	37.3	*Bacillus subtilis*

isolated from these foam samples, detected small numbers of hybridised rod shaped bacteria in 3 of the foam samples (Mt Martha, Heatherton and Liverpool) indicating that their presence in these foams was low. However, surfactin could be detected in Mt Martha samples as well as those from the Biodiesel plant samples at low concentration (<0.1 mg/l), and laboratory foaming tests with commercial surfactin suggest that this biosurfactant enhances foaming abilities at levels below those detectable by HPLC analyses (<0.1 mg/l) (data not shown).

Table 5 – Summary of the WWTP surveyed and their dominant foam associated filaments.

Plant	State	Dominant filament[a]
Biodiesel	Vic	Gram +ve rods in chains
Carrum ETP	Vic	GALO
Pakenham (aeration tank)	Vic	*N. limicola II* (Gram +ve) 0041/0675
Liverpool	NSW	GALO
Glenfield	NSW	GALO *Haliscomenobacter hydrossis*
Somers (aeration tank)	Vic	*Microthrix parvicella*
Heatherton	Vic	0041/0675 Cocci/rods related to GALO
Mt Martha (aeration tank)	Vic	*Microthrix parvicella* *N. limicola II* (Gram −ve)
Mt Martha (Digester)	Vic	*Microthrix parvicella*
Boneo (Clarifier)	Vic	*Microthrix parvicella* 0041/0675
Boneo (Bioreactor)	Vic	1863 (*Acinetobacter*) *Microthrix parvicella* 0041/0675
Whyalla (SBR1)	SA	GALO 0041/0675
Whyalla (SBR2)	SA	GALO 0041/0675

a GALO = *Gordonia amarae* like organism.

3.5. *Are foaming thresholds affected by the presence of surfactants?*

When PYCa broth was supplemented with Triton-X 100 at a concentration of 0.00075% (v/v), unstable foam could be generated when aerated in the foaming test. The addition of this small quantity of Trition-X 100 to Mycolata cultures had a dramatic affect on the extent and stability of foaming with the laboratory foaming assay system. In particular, the threshold cell number required for stable foam formation decreased by at least 10 fold in all cases. The foaming threshold of the precultured Mycolata strains was also determined in the presence of the *B. subtilis* culture supernatant containing surfactin. Again the foaming threshold was reduced by at least a 10 fold decrease in cell numbers and in some instances up to a 100 fold decrease. Such outcomes confirm that in the presence of surfactant, substantially fewer cells are needed to form a stable foam.

3.6. *Effect of hydrophobic substrates on foaming abilities of Mycolata*

When selected Mycolata were grown in media containing the hydrophobic substrate olive oil instead of glucose these organisms grew as clumps which stuck to the glass surfaces of the culture vessels. This suggests that their cell hydrophobicities increase markedly when in the presence of such a hydrophobic carbon source, giving them an increased access to it as a carbon and energy source. Unfortunately the MATH and foaming assays could not be performed with these cultures because the cells did not disperse within the culture medium, but instead adhered strongly to the glass surface of the culture vessel as a film, thus preventing performing any such analyses.

4. Discussion

The data presented in this study confirm that the flotation theory helps to explain most of what is known about the formation of stable foams by members of the Mycolata on activated sludge reactors. However, not all strains tested form stable foams under the conditions used in this study, suggesting that foaming is not necessarily an inherent feature of these organisms, but a complex end result of the interaction of many environmental conditions (Stratton et al., 2002, 2003). Furthermore, in some strains, foaming ability did not always correspond to a high cell percentage MATH hydrophobicity value or substantial culture surface tension reduction thought to reflect biosurfactant production. Similar inconsistencies have been recorded in previous studies (Stratton et al., 2002), and suggest that these measurements are not always appropriate for such foaming experiments (Eikelboom, 1991). Nevertheless, the results clearly show that stable foam production is an event which requires the presence of three essential components; air bubbles, surfactants and hydrophobic bacterial cells. With insufficient hydrophobic cells, but in the presence of the other two, large amounts of an unstable foam will be generated, and explains why these are commonly seen during the start up of plants (Jenkins et al., 2004), when the population sizes of such organisms would be expected to

be low. In some activated sludge plants the biosolids form thin greasy layers or scums over the surface of the mixed liquor (Lemmer et al., 2005). How their formation relates to foaming incidents has been unclear, and the terms foaming and scumming have often been used to describe the same operational problem (Lemmer et al., 2005). However, these data suggest that unlike foams, scums only form in the presence of sufficient levels of hydrophobic bacteria but insufficient surfactant levels for foaming. Consequently these terms should not be used synonymously.

The data presented here also show that stable foams will form with non-viable Mycolata cells, which from MATH assay results appear to retain their hydrophobicity. This is an important observation, since it questions the validity of using FISH, where detection of cells relies on high ribosome levels, for determining bacterial cell numbers in activated sludge foams (Davenport et al., 2008; de los Reyes and Raskin, 2002). Thus, the current concept of a universal foaming threshold based on this approach should remain speculative, especially since the results presented here clearly show that the stable foaming threshold number varies with individual Mycolata, and falls substantially in the presence of surfactants. Production of surfactants by Mycolata is well documented (de los Reyes, 2010; Lechevalier, 1975), although the level of their synthesis can vary with the carbon source used. Here we show that the non-hydrophobic *B. subtilis* may be an important contributor to stable foam formation in activated sludge plants. Its frequent isolation from foams, and demonstrated ability to generate large amounts of an unstable foam in pure culture, together with production of the powerful surfactant surfactin are consistent with such an involvement. Surfactin is a complex molecule, unlikely to be degraded rapidly in activated sludge, and likely to persist there long enough to encourage foaming. *In situ* analysis with FISH in activated sludge foams indicated that the detectable presence of *B. subtilis* was low. However, as these cells are not hydrophobic by the MATH assay (Table 4), they may occur more commonly in the mixed liquor and end up in foams incidentally, where they may then sporulate under the nutrient deficient conditions likely to exist there (Eales et al., 2005), thus becoming important but indirect participants in stable foaming events. Even in the mixed liquor their abundance need not be high for a possible important contribution to foaming, given the low levels of surfactin demonstrated as needed for foam stabilisation (<0.1 mg/l) and the high level of its production by these strains.

This is the first report where a possible important role for *B. subtilis* and surfactin production has been demonstrated in foaming, and further work directed at understanding better their influence on its formation in activated sludge plants is required. This population is unlikely to be alone in affecting foaming by *in situ* surfactant production, and with the availability of a simple assay for detecting bacterial biosurfactant formation (Burch et al., 2010), a search for other populations, especially in unstable activated sludge foams may be rewarding.

5. Conclusion

This study presents data from 65 foaming Mycolata that the flotation theory can be applied to most of these to explain their role in stable activated sludge foam formation. Not all conformed to this theory under the experimental conditions used in the study, and with some isolates, the bases for their foaming ability remains uncertain, which suggests that foaming is a complex phenomenon but requires both hydrophobic cells and surfactants as required by the flotation theory. The data also demonstrate a clear distinction between the properties of foaming and scumming bacteria, and proposes that these terms should not be used synonymously. When surfactants are absent but hydrophobic Mycolata are present, a scum is produced. On the other hand with no hydrophobic particles but a surfactant present, an unstable foam is generated. The concept of a universal threshold for foaming applicable to all plants is not supported by these data, where the numbers of cells required depends on the bacterial strain and the presence of surfactant. The study shows for the first time that *B. subtilis*, shown here to be commonly culturable from foams may be an important participant, by its production of the powerful surfactant surfactin.

Acknowledgements

We wish to thank Dr Michael Angove for his advice on the analysis of the HPLC data. The research was supported by the Australian Research Council Linkage Grant (LP0774913) in collaboration with Melbourne Water (David Gregory) and South East Water (Graham Short) who are thanked for their financial support. Z.A. Dyson was the recipient of an Australian Postgraduate Award PhD Scholarships. S. Petrovski was funded by the ARC linkage grant and La Trobe University.

REFERENCES

Amann, R., Ludwig, W., 2000. Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology. FEMS Microbiology Reviews 24 (5), 555–565.

Besson, F., Michel, G., 1992. Biosynthesis of iturin and surfactin by *Bacillus subtilis*. Biotechnology Letters 14 (11), 1013–1018.

Blackall, L.L., Marshall, K.C., 1989. The mechanism of stabilization of actinomycete foams and the prevention of foaming under laboratory conditions. Journal of Industrial Microbiology 4 (3), 181–188.

Burch, A.Y., Shimada, B.K., Browne, P.J., Lindow, S.E., 2010. Novel high-throughput detection method to assess bacterial surfactant production. Applied and Environmental Microbiology 76 (16), 5363–5372.

Daims, H., Stoecker, K., Wagner, M., 2005. Fluorescence *in situ* hybridization for the detection of prokaryotes. In: Osborn, A. M., Smith, C.J. (Eds.), Molecular Microbial Ecology. Taylor & Francis Group, New York, pp. 213–240.

Davenport, R.J., Curtis, T.P., Goodfellow, M., Stainsby, F.M., Bingley, M., 2000. Quantitative use of fluorescent *in situ* hybridization to examine relationships between mycolic acid containing actinomycetes and foaming in activated sludge plants. Applied and Environmental Microbiology 66 (3), 1158–1166.

Davenport, R.J., Pickering, R.L., Goodhead, A.K., Curtis, T.P., 2008. A universal threshold concept for hydrophobic mycolata in activated sludge foaming. Water Research 42 (13), 3446–3454.

de los Reyes, F.L., Raskin, L., 2002. Role of filamentous microorganisms in activated sludge foaming: relationship of Mycolata levels to foaming initiation and stability. Water Research 36 (2), 445–459.
de los Reyes, F.L., 2010. Foaming. In: Seviour, R., Nielsen, P.H. (Eds.), Microbial Ecology of Activated Sludge. IWA Publishing, London, UK, pp. 215–258.
Eales, K.L., Nielsen, J.L., Kragelund, C., Seviour, R., Nielsen, P.H., 2005. The *in situ* physiology of pine tree like organisms (PTLO) in activated sludge foams. Acta Hydrochimica et Hydrobiologica 33, 203–209.
Eikelboom, D.H., 1991. Scuim-en driflaagvorming op zuiveringsinrichtingen. Report. TNO Milieu en Energie, Delft, Netherlands.
Gong, G., Zheng, Z., Chen, H., Yuan, C., Wang, P., Yao, L., Yu, Z., 2009. Enhanced production of surfactin by *Bacillus subtilis* E8 mutant obtained by ion beam implantation. Food Technology and Biotechnology 47 (1), 27–31.
Ho, C., Jenkins, D., 1991. The effect of surfactants on nocardia foaming in activated sludge. Water Science and Technoogy 23 (4–6), 879–887.
Jenkins, D., Richard, M.G., Daigger, G.T., 2004. Manual on the Causes and Control of Activated Sludge Bulking, Foaming and Other Solids Separation Problems. IWA Publishing, London.
Kragelund, C., Remesova, Z., Nielsen, J.L., Thomsen, T.R., Eales, K., Seviour, R., Wanner, J., Nielsen, P.H., 2007. Ecophysiology of mycolic acid-containing *Actinobacteria* (Mycolata) in activated sludge foams. FEMS Micriobiology Ecology 61 (1), 174–184.
Lane, D.J., 1991. 16S/23S rRNA sequencing. In: Goodfellow, M., Stackebrandt, E. (Eds.), Nucleic Acid Techniques in Bacterial Systematics. John Wiley & Sons, Chichester, UK, pp. 115–175.
Lechevalier, H., 1975. Actinomycetes of sewage treatment plants. In: U.S. Dept. Of Commerce National Technical Information Services Report.
Lemmer, H., Lind, G., Muller, E., Schade, M., 2005. Non-famous scum bacteria: biological characterization and troubleshooting. Acta Hydrochemica et Hydrobiologica 33 (3), 197–202.
Meier, H., Amann, R., Ludwig, W., Schleifer, K.H., 1999. Specific oligonucleotide probes for in situ detection of a major group of gram-positive bacteria with low DNA G + C content. Systematic and Applied Microbiology 22 (2), 186–196.
Müller, E., Schade, M., Lemmer, H., 2007. Filamentous scum bacteria in activated sludge plants: detection and identification quality by conventional activated sludge microscopy versus flouresence *in situ* hybridization. Water Environment Research 79 (11), 2274–2286.
Nielsen, P.H., Daims, H., Lemmer, H., 2009. FISH Handbook for Biological Wastewater Treatment. IWA Publishing, London.
Rooney, A.P., Price, N.P.J., Ehrhardt, C., Swezey, J.L., Bannan, J.D., 2009. Phylogeny and molecular taxonomy of the *Bacillus subtilis* species complex and description of *Bacillus subtilis* subsp. *inaquosorum* subsp. nov. International Journal of Systematic and Evolutionary Microbiology 59, 2429–2436.
Rosenberg, M., 2006. Microbial adhesion to hydrocarbons:twenty-five years of doing MATH. FEMS Microbiology Letters 262, 129–164.
Rosenberg, M., Cutnick, D., Rosemberg, E., 1980. Adherence of bacteria to hydrocarbons. A simple method for measuring cell surface hydrophobicity. FEMS Microbiology Letters 9, 29–33.
Seviour, R.J., Kragelund, C., Kong, Y., Eales, K.L., Nielsen, J.L., Nielsen, P.H., 2008. Ecophysiology of the *Actinobacteria* in activated sludge systems. Antonie van Leeuwenhoek 94 (1), 21–33.
Soddell, J.A., 1999. Foaming. In: Seviour, R.J., Blackall, L.L. (Eds.), Microbiology of Activated Sludge. Kluwer, Dordrecht, pp. 161–202.
Soddell, J.A., Seviour, R.J., 1990. Microbiology of foaming in activated sludge plants – a review. Journal of Applied Bacteriology 69 (2), 145–176.
Soddell, J.A., Seviour, J.R., 1998. Numerical taxonomy of *Skermania piniformis* and related isolates from activated sludge. Journal of Applied Microbiology 84 (2), 272–284.
Stratton, H., Seviour, B., Brooks, P., 1998. Activated sludge foaming: what causes hydrophobicity and can it be manipulated to control foaming? Water Science and Technology 37 (4–5), 503–509.
Stratton, H.M., Brooks, P.R., Griffiths, P.C., Seviour, R.J., 2002. Cell surface hydrophobicity and mycolic acid composition of *Rhodococcus* strains isolated from activated sludge foams. Journal of Industrial Microbiology and Biotechnology 28 (5), 264–267.
Stratton, H.M., Brooks, P.R., Carr, E.L., Seviour, R.J., 2003. Effects of culture conditions on the mycolic acid composition of isolates of *Rhodoccous* spp. From activated sludge foams. Systematic and Applied Microbiology 26 (2), 165–171.