Universidade de Lisboa

Faculdade de Farmácia

In Silico Analysis of Mycobacteriophage Ms6

Francisco Olivença Miguel

Mestrado Integrado em Ciências Farmacêuticas

2017

Universidade de Lisboa
Faculdade de Farmácia

In Silico Analysis of Mycobacteriophage Ms6

Francisco Olivença Miguel

Monografia de Mestrado Integrado em Ciências Farmacêuticas
apresentada à Universidade de Lisboa através da Faculdade de Farmácia

Orientador: Professora Doutora Madalena Pimentel

2017

Abstract

Isolated in 1989, Ms6 is a renowned temperate mycobacteriophage that earned a spot as model of mycobacteriophage-induced lysis. Despite the studies on the lytic operon, the remainder of the genome of Ms6 was kept mostly unexplored. A modern next generation sequencing, followed by a complete genome annotation, enlightens several aspects concerning this phage. The genome is composed of linear double stranded DNA, with 54252 bp of length and a GC content of 61,5%. The closest homologues of Ms6 are Dlane and Shauna1. The similarities between Ms6 and these and many other F1 clustered phages allow its allocation to this group although 11 ORFs present a first BLASTp hit that is not a gene product of a F1 phage.

A total of 105 ORFs were identified, of which 43 were given a putative function according to a combination between location in the genome, homology with previously characterized proteins, existence of conserved motifs or structural similarities. The integrase gene divides the genome in two genomic arms, the left and the right. Within the left arm of the genome, it is possible to distinguish genes with structural roles, such as the head and the tail assembly genes and genes required for processes like packaging and lysis. The right arm is less conserved than its left counterpart and comprises ORFs involved either in DNA modification, like exonucleases or methylases, or in phage regulation, for example the WhiB factor or proteins with helix-turn-helix DNA binding motifs. Besides the pin gene and its phage resistance properties, other interesting features include a possible virion associated lysin and a protein that might encourage homologous recombination.

Keywords: Ms6; Bacteriophage; Mycobacteria; Genome Annotation.

Resumo

Isolado em 1989, o Ms6 é um micobacteriófago temperado pertencente à família Siphoviridae e que infecta *Mycobacterium smegmatis*. Apesar dos estudos já efectuados sobre a integração e sobre a lise das micobactérias induzida pelo Ms6, o restante genoma do Ms6 permaneceu maioritariamente por explorar. Através de *Next Generation Sequencing*, seguido de uma completa anotação do genoma, foi possível esclarecer alguns aspetos relativos a este fago.

O genoma é composto por uma dupla cadeia linear de DNA, contendo 54252 bp e um conteúdo em GC de 61,5%. Os fagos mais semelhantes ao Ms6 são o Dlane e o Shauna1. A homologia entre o Ms6 e estes fagos, bem como com muitos outros fagos do subcluster F1, permitem a sua inclusão neste grupo, apesar de 11 ORFs apresentarem um primeiro resultado de BLASTp que não corresponde a proteínas de fagos F1.

Foram identificadas 105 ORFs e foi possível atribuir uma função a 43 das mesmas, por combinação de informação proveniente da sua localização no genoma, homologia com proteínas previamente caracterizadas e presença de motivos conservados ou analogias estruturais. A gene que codifica para a integrase divide o genoma em dois ramos, o esquerdo e o direito. Dentro do ramo esquerdo do genoma, é possível distinguir genes com um papel estrutural, tais como os que codificam proteínas da cápside ou da cauda, e genes necessários a processos como o empacotamento do DNA ou a lise da célula hospedeira. O ramo direito é menos conservado e contém genes envolvidos na modificação do DNA, tais como os que codificam exonucleases ou metilases, ou na regulação do fago, codificando o factor WhiB ou proteínas com domínios de ligação ao DNA do tipo HTH. Além do gene pin, cujo produto confere resistência à superinfeção, outros elementos interessantes incluem uma possível lisina associada ao virião e uma proteína que pode estimular a recombinação homóloga.

Palavras-chave: Ms6; Bacteriófago; Micobactérias; Anotação de genomas.

Acknowledgements

I would like to thank Professor Madalena Pimentel for the unique opportunity she gave me in 2014, allowing me to work at her laboratory and to learn more about microbiology and molecular biology. Since my third year of college I developed two projects concerning phages and my interest in their possibilities grew from there on. Therefore, performing the annotation of Ms6's genome enabled an expansion of my knowledge about phages and their potential applications. Over these three years, including during the elaboration of this thesis, the constant orientation and availability of Professor Madalena were crucial. I'm profoundly grateful for her exceptional teaching skills and for all her help and guidance. The time spent at the laboratory has been a wonderful adventure, with many enjoyable moments alongside Professor Madalena, which I will always remember fondly.

To Dr. Maria João Catalão I would like to thank her support over my senior year and all the precious advices and suggestions. Her kind words were a motivation during the hard times. I would also like to thank her useful explanations at the laboratory.

I consider Professor Madalena and Dr. Maria João Catalão role models and I'm very thankful for crossing my path with both, because they inspire me to follow a scientific career.

To my college friends, namely Catarina Rei, Joana Luz and Marco Nobre, I thank their support in this great adventure of five years and all the good moments we shared. I'm proud of our friendship and hope it continues for many more years. I also want to thank my friends from Caldas da Rainha for understanding my ocasional absences.

I want to thank my family for allowing me to pursue this degree and for all their help during my academic course. It was necessary a lot of scarifices to which I'm forever grateful. Without them, none of this would have been possible. I need to thank my mother in special for her strength and dedication.

Finally, I want to thank Pedro for his attention and patience with me. The hardwork and devotion he inputs in his career, motivate me to give the best of me, always. I also want to thank his consideration for my well-being and all the care over these five years.

Abreviations

%ID	Percentage of identity to query sequence
Aa	Amino acid
Abi	Abortive infection system
ATP	Adenosin triphosphate
attB	Bacterial attachment site
attL	Left host-prophage junction
attP	Phage attachment site
attR	Right host-prophage junction
bp	Base pair
CDD	Conserved domains database
DNA	Deoxyribonucleic acid
ds	Double stranded
EDTA	Ethylenediaminetetraacetic acid
GC content	Guanine-cytosine content
gp	Gene product
HNH	Type of homing endonuclease
HTH	Helix-turn-helix
LPS	Lipopolyssacharide
LysA	Lysin A
LysB	Lysin B
mAGP	Mycolyl-arabinogalactan-peptidoglycan complex
MW	Molecular weight
NAG	N-acetylglucosamine
NAM	N-acetylmuramic acid
NCBI	National Center for Biotechnology Information
NTPase	Nucleotide triphosphate hydrolases
ORF	Open reading frame
PE-PPE	Pro-Glu and Pro-Pro-Glu conserved protein domains
PG	Peptidoglycan
PGBD	Peptidoglycan binding domain
PGRP	Peptidoglycan recognition protein
RNA	Ribonucleic acid

Rpf	Resuscitation-promoting factors
SDS	Sodium dodecyl sulfate
SGNH	Type of esterase domain
Sie	Superinfection exclusion
ss	Single stranded
SSAP	Single-strand annealing protein
TMD	Transmembrane domains
tmRNA	Transfer-messenger ribonucleic acid
tRNA	Transfer ribonucleic acid
UDP-Glc	Uridine diphosphate-α-D-glucose
VAL	Virion-associated lysin
Xis	Excisionase

Index:

Abstract 5
Resumo 6
Acknowledgements 7
Abreviations 8
1 Introduction 13
1.1 Bacteriophages 13
1.1.1 Brief history and relevance of bacteriophage research 13
1.1.2 Phage Morphologies 14
1.1.3 Phage Genomes 14
1.1.4 Phage Lifecycles 15
1.2 Mycobacteriophages 17
1.2.1 General Aspects 17
1.2.2 Mycobacteriophage Genomics 18
1.2.3 Hosts and isolation 19
1.3 Genome Annotation 19
1.3.1 Definition and Data Flow 19
1.3.2 Phage Genome Annotation 21
1.4 Mycobacteriophage Ms6 22
2 Objectives 23
3 Materials and Methods 24
3.1 Phage DNA Extraction and Sequencing 24
3.2 Statistical Gene Prediction by DNA Master 24
3.3 Additional Information and Auto-Annotation Refinement 24
3.4 General Database Search 25
3.4.1 BLASTp and BLASTn 25
3.4.2 HHpred 25
3.5 Prediction of Structural Features 26
3.5.1 CATH 26
3.5.2 TMHMM and TMpred 26
3.6 Specialized Database Search 26
3.6.1 Blastp on phages DB 26

3.6.2 Aragorn and tRNA Scan-SE 26
3.6.3 MOTIF 27
4 Results & Discussion 28
4.1 General Characterization of the Genome of Phage Ms6 28
4.2 Functional ORF Prediction 32
4.2.1 Packaging 32
4.2.2 Head Assembly and Head-to-Tail Connectors 37
4.2.3 Tail Assembly 39
4.2.4 Lysis 40
4.2.5 Integration 43
4.2.6 DNA Modification 43
4.2.7 Regulation 45
4.2.8 Other Interesting Features 47
4.2.8.1 Other potential properties of ORF16 47
4.2.8.2 Structural homology between the gp17 of Ms6 and resuscitation-promoting factor B 47
4.2.8.3 Mycobacteriophage mosaicism is patent in ORFs 18-22 of Ms6 48
4.2.8.4 The Pin protein, codified by ORF37, confers resistance to infection by Ms6 to a mutant M. smegmatis strain 49
4.2.8.5 Multiple tetrapeptide repeats in the central portion of gp48 50
4.2.8.6 ORF60 appears to encode a Sak3-like protein 51
4.2.8.7 A typical phage serine-threonine kinase is associated with ORF102 52
5 Conclusion 53
Bibliography 55
Appendix 63
A1. Guiding Principles of Bacteriophage Genome Annotation 63

Figure Index:

Figure 1 - Diversity in phage morphologies and genomes 15
Figure 2 - Different phage lifecycles 16
Figure 3 - Currently, 10526 actinobacteriophages have been found 17
Figure 4 - Sinteny. 18
Figure 5 - Genome annotation 20
Figure 6 - Map of Ms6 genome 31
Figure 7 - Genome packaging in dsDNA phages 37
Figure 8 - Tail assembly process of Lambda-like phages 40
Figure 9 - Mycobacterial targets of Ms6 lysis proteins 42
Figure 10 - Whole-genome dot plot and BLASTn alignment of phages Ms6, Dlane, Shauna1, Brocalys and Redi 49
Figure 11 - Multiple tetrapeptide repeats of Ms6 51

Table Index:

Table 1 - Coordinates and predicted functions of the 105 ORFs of Ms6 29
Table 2 - Structural analysis, motif search and transmembrane domains for each identified ORF 33

1 Introduction

1.1 Bacteriophages

1.1.1 Brief history and relevance of bacteriophage research

Bacteriophages or phages are bacterial viruses that were first discovered in 1915 by Frederick Twort and described in 1917 by Felix d'Herelle. Due to their antimicrobial properties, early phage-based therapies aroused some enthusiasm. The advent of World War II, associated with the discovery of broad spectrum antibiotics led to the abandonment of phage therapy in a post-war period in the USA and Western Europe, while Russia and Eastern Europe maintained phage applications to some extent (1,2).

During the next decades, bacteriophage studies provided insight on broader molecular biology processes and were determinant to understand that DNA contains genetic information. The perception that phages ubiquitously occupy most environments increased over time and their ability to command many aspects of bacterial or archaeal biology became evident (3). Phages can mediate certain events such as bacterial genome rearrangements, horizontal transference of non-viral genes and, in some cases, conversion of the host to virulence. Currently, it is widely accepted that phages strongly influence their hosts characteristics, distribution and evolution throughout time (4).

With the emergence of multi-drug resistant bacteria and a reduction in antibiotics research and development, the interest in phages as potential antimicrobial tools has been rekindled (5,6). Some of the advantages of phage therapy over chemical antibiotics include the natural bactericidal properties of phages, high host specificity that translates in minimal impact on normal flora, low inherent toxicity and limited potential to induce resistance (7). Other potential phage applications like gene or drug carriers have also emerged (8).

1.1.2 Phage Morphologies

Concerning morphology, phages can be classified as either tailed or non-tailed.

Tailed phages represent the vast majority of phages isolated until today and are represented by the order *Caudovirales*, subdivided into three families: *Myoviridae* (contractile tail), represented by phage T4; *Siphoviridae* (long, noncontractile tail), portrayed by phage λ; *Podoviridae* (short tail), exemplified by phage T7 (9).

Non-tailed phages include polyhedral, filamentous or pleomorphic phages, generally designated as PFP. Polyhedral phages are icosahedral or present a cubic symmetry, filamentous phages have helical symmetry and pleomorphic phages lack symmetry axes.

1.1.3 Phage Genomes

All phages of the *Caudovirales* order contain linear, double stranded DNA. Some families contain genomes with different types of nucleic acid, including circular dsDNA (*Corticoviridae* and *Plasmaviridae*), circular ssDNA (*Inoviridae* and *Microviridae*), ssRNA (*Leviviridae*) and dsRNA phages with segmented chromosomes (*Cystoviridae*). These groups of phages have smaller genomes than tailed phages but appear to evolve by similar mechanisms (10).

Between 90 to 95% of tailed phages DNA consists of protein-coding sequences, revealing an efficient use of their genomes. Genes are typically aligned in co-transcribed groups, often very tightly packed and overlaps of the termination codon of one gene and the initiation codon of the downstream gene are common. In some phages, all the genes are transcribed in one direction, while in others transcription occurs in both forward and reverse directions (10).

Another interesting characteristic of phage genomes is their organization. An analysis of sequence relationships suggests phage genomes present a mosaic architecture that results from horizontal exchanges of segments between members of phage populations. Genome mosaicism refers to a specific assembly of individual modules, each composed of one or more genes, with the location of the modules varying in genomes not closely related. Two alternative models explain this mosaicism: one advocates genetic modules are rearranged by homologous recombination at short conserved boundary

sequences; the other portrays random recombining between viral and nonviral DNA molecules as the driving force of mosaicism, producing mostly genomic trash that is eliminated and some viable genomes that maintain all the necessary functions and adequate size for packaging, originating new virion particles (11). In Figure 1 are depicted several diffrent phage morphologies and their respective nucleic acid types.

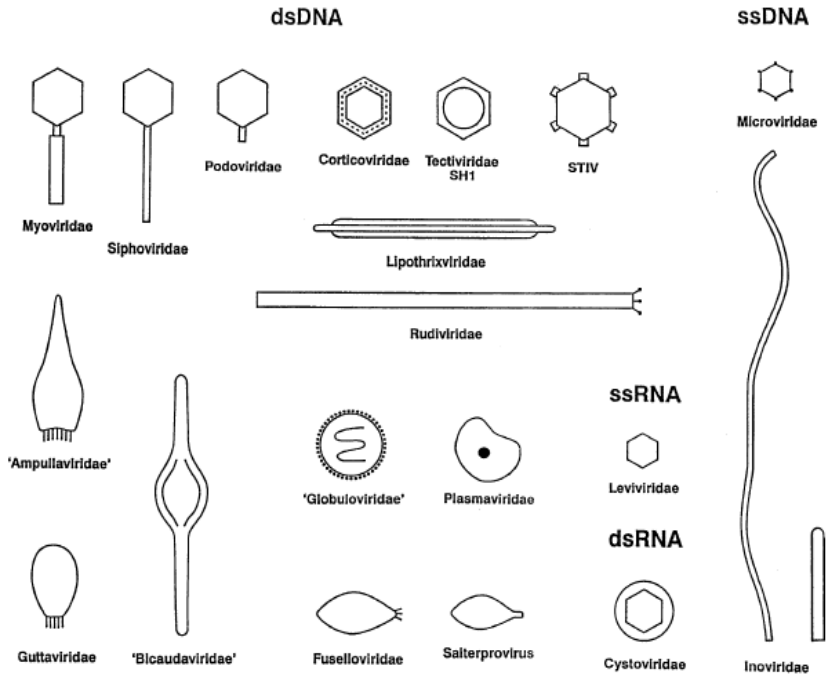


Figure 1 - Diversity in phage morphologies and genomes (Ackermann, 2007).

1.1.4 Phage Lifecycles

Phages are obligate parasites, naturally exploiting bacteria in order to survive and multiply. Concerning their relationship with their hosts, phages can be distinguished as virulent or temperate and may have different exploitation approaches, known as the lytic or the lysogenic lifecycles (Figure 2).

Virulent phages, like *E. coli* phage T4, only pursue the lytic cycle, where the phage replicates its genetic material, expresses structural genes, assembles virions and ultimately releases its progeny through lysis, killing their host.

Temperate phages, like λ phage, can undergo the lytic cycle but can also follow the lysogenic cycle, in which the phage DNA is integrated into the host genome establishing a prophage. Lytic gene expression is repressed and the prophage DNA replicates as part of the bacterial chromosome over the subsequent cell divisions. The

prophage state is very stable until certain triggers, such as DNA-damaging agents, stimulate the switching into the lytic cycle (12).

Sometimes the phage is not able to pursue neither the lytic or the lysogenic cycles and its genome remains as a non-replicating and non-integrated preprophage, only inherited by one of the daughter cells. This is referred as pseudolysogeny and is more frequent in nutrient scarcity. When nutritional conditions are recovered, the phage enters the lytic or lysogenic cycles (13).

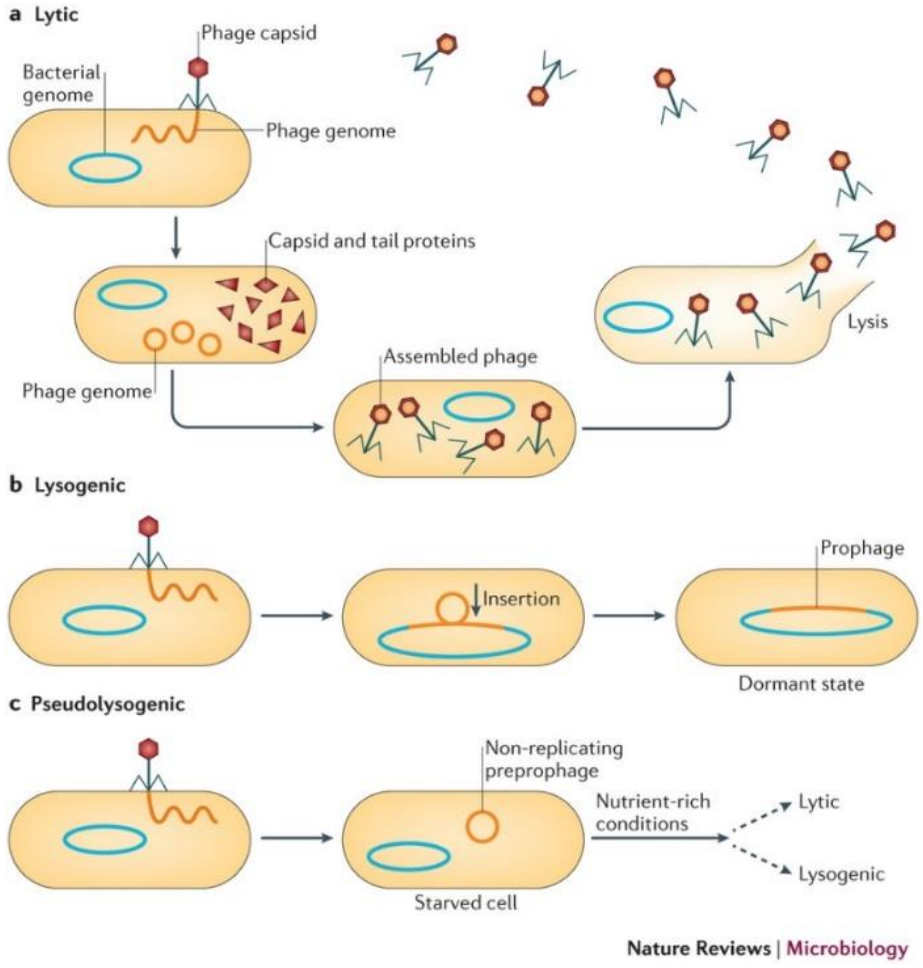


Figure 2 - Different phage lifecycles. a) Lytic lifecycle culminates with the lysis of the host and release of progeny virions; b) Lysogenic lifecycle can only be pursued by temperate phages, by suppression of lytic functions; c) Some phages follow the pseudolysogenic lifecycle, assuming the form of a non-replicating preprophage (Feiner *et al.*, 2015).

1.2 Mycobacteriophages

1.2.1 General Aspects

Mycobacteriophages are viruses that infect mycobacteria. In a broader sense, they are considered as actinobacteriophages, phages that infect bacterial hosts within the phylum Actinobacteria. This phylum includes the genus *Mycobacterium*, but also other genera, such as *Arthrobacter*, *Gordonia*, *Propionibacterium* and *Streptomyces*, among others. Mycobacteriophages are, by far, the most numerous actinobacteriophages discovered until today (Figure 3) (14).

All mycobacteriophages identified contain dsDNA and although phages with ssDNA or RNA genomes have not been described yet, is possible they exist. Most mycobacteriophages have siphoviral morphologies, while a smaller number present myoviral morphology. Mycobacteriophages with podoviral characteristics have not been found and this is possibly due to the complex mycobacterial cell wall acting as a barrier and preventing a successful infection by these phages (15).

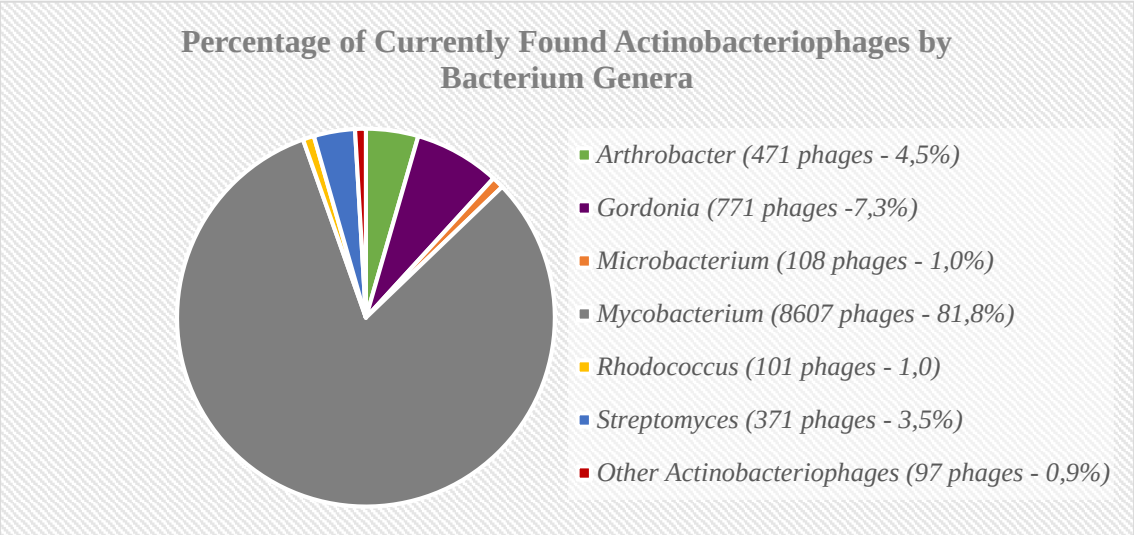


Figure 3 - Currently, 10526 actinobacteriophages have been found. Approximately 82% of these phages are mycobacteriophages, consisting of the larger group of actinobacteriophages, followed by phages infecting the *Gordonia* (7,3%) and the *Arthrobacter* (4,5%) genera (Adapted from The Actinobacteriophages Database, 2017).

1.2.2 Mycobacteriophage Genomics

Although all mycobacteriophages genomes are dsDNA, they vary in length between a little more than 41 kb to close to 165 kb. Average GC% content is close to their common host, *M. smegmatis* mc^2155, but it can also fluctuate between 50,3 and 70% (14).

Mycobacteriophage genome diversity is evident and since some phages are more closely related to each other than to others, an assortment of mycobacteriophages into clusters is possible. Phages of a given cluster must present nucleotide sequence similarity over 50% of genome length. Within some clusters, phages can be subdivided into subclusters based in genome organizations and number of genes that present orthology. A phage with no close relatives is known as a singleton (16). This allocation of phages into clusters and subclusters is not the mirror of precise phylogenetic or evolutionary relationships but rather a convenient way to express phage similarities and distinctions.

The comparison of several mycobacteriophages DNA sequences outlined six general aspects concerning their genomes: a) As in other phages, genes are tightly packed and there is little space for noncoding sequences; b) Genetic mosaicism is present and different segments can have different evolutionary backgrounds; c) Many genes of siphoviral mycobateriophages appear in the same order, namely those involved in virion structure and assembly, a phenomenon named synteny (Figure 4); d) Genes of unknown function exist in large numbers; e) Small open reading frames are numerous, especially in the right parts of genomes; f) The precise role of some genes with attributed putative functions is still hazy (17,18).

Mycobacteriophage genomes have become significant tools in mycobacterial genetics and their genomic characterization has also emphasized a substantial diversity that can be useful to understand aspects related with viral diversity and the evolutionary mechanisms behind it (19).

Figure 4 - The typical order of virion structure and assembly genes in syphoviral mycobacteriophages is a form of synteny (adapted from Jacobs-Sera *et al.*, 2014).

1.2.3 Hosts and isolation

The genus *Mycobacterium* consists of aerobic, rod shaped and non-motile bacilli. Mycobacteria are (G+C) rich (62-70%), with acid-fast alcohol stains and are considered Gram-positive (20). These bacteria show a high resistance to heavy metals, antiseptics and antibiotics, greatly supported by their hydrophobic cell wall, rich in lipids and thicker than most other bacteria, which portrays a distinctive and crucial feature of these bacteria (21). Concerning the growth rate, it is possible to classify mycobacteria in fast or slow-growing. Fast-growing mycobacteria include the laboratory strain *M. smegmatis* mc^2155, while slow-growing mycobacteria include human pathogens *M. tuberculosis* and *M. leprae*, the causative agents of the devastating diseases of tuberculosis and leprosy, respectively (22).

Mycobacteriophages can be isolated from environmental samples like soil or compost, either directly or by enrichment. Other phages are isolated through release from a lysogenic host or are detected in sequenced mycobacterial genomes (17).

In the late 1940s, the first mycobacteriophage was obtained using *M. smegmatis* as a host and a *M. tuberculosis* phage was isolated a few years later (23,24). Nowadays, 9362 mycobacteriophages infecting five mycobacterium species (*M. aurum*, *M. avium*, *M. phlei*, *M. smegmatis* and *M. tuberculosis*) have been identified, from which 1419 are sequenced (14).

1.3 Genome Annotation

1.3.1 Definition and Data Flow

Genome annotation may be considered as a subfield within genome analysis accountable for the analysis and interpretation of raw DNA sequences, necessary to ascertain their relevance in biological events and mechanisms. It is a process with multiple steps that can be divided into three levels (Figure 5, left panel):

a) Nucleotide level – First, genome annotators ask ‘where’ and try to identify known genes, genetic marks and other landmarks in the genetic sequence.

b) Protein-level - After asking 'where', annotators ask 'what' and aim to determine the set of proteins encoded by the DNA sequence of a given organism, to name them and assign them putative functions.

c) Process-level or functional annotation – The final question is 'how' and its goal is to interpret and relate genes, proteins and their functions in the context of biological processes (25).

In (Figure 5, right panel) is depicted a generalized flow chart of genome annotation. Starting with the finished genome sequence, protein-coding genes are predicted by statistical gene prediction methods, like GeneMark or Glimmer. This is an integrated process where general database search (for example, using BLAST to search sequence similarities in databases) and prediction of structural features (such as signal peptides, transmembrane segments or coiled domains) play an important role. Gene identification generates a feedback (FB) that is useful for correction of sequencing errors, especially frameshifts. Then, analysis through specialized databases, like searching for conserved domains in Pfam or CDD, enriches the annotation and adds data to predict gene functions more accurately with a simultaneous, rigorous, context analysis (26).

Genome annotations promote bench work and computational analyses carried at laboratories and are a crucial resource for other genome annotation projects, therefore is paramount they are the most accurate possible and periodically revised and rectified (27).

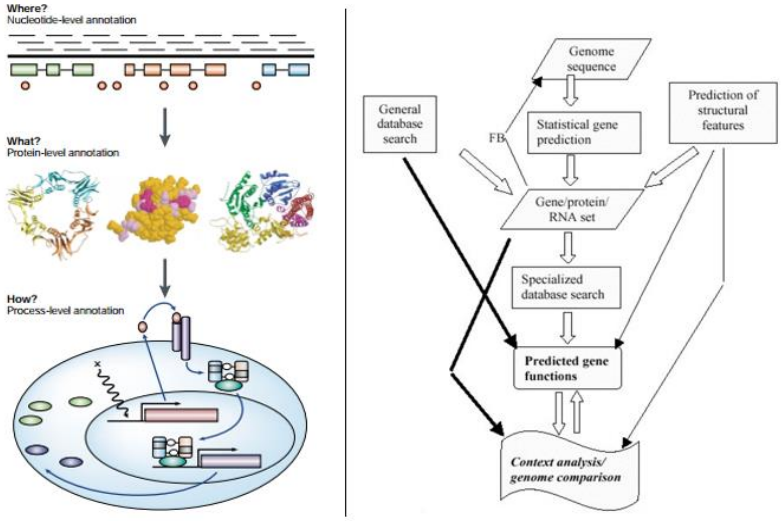


Figure 5 - The levels of genome annotation: nucleotide, protein and functional annotation (left) (Stein, 2001). Flow chart of a general genome annotation process. FB represents the feedback offered by the identification of predicted genes to correct certain sequencing errors (right) (Koonin & Galperin, 2003).

1.3.2 Phage Genome Annotation

Phage genome annotation process gives insight into the possible evolutionary path of a given phage, as well as an idea of species affiliation and if it is admissible for safe use in therapy. The annotation allows the identification of ORFs which may encode mobile genome elements, like transposition modules and introns or dangerous products, such as toxins or specific enzymes (28).

Even though phage genomes have small sizes, they still pose some challenges to a rigorous genome annotation, including gene call. The main issues are the small size of phage genes (~600 bp), especially those in non-structural genomic regions and the abundance of genes with unknown functions (29).

Some bioinformatic tools facilitate phage genome annotation and comparative analysis, like Phamerator, a software that produces a database of gene relationships by arrangement of protein-coding genes into groups of related sequences based in pairwise comparisons (29). These groups are called phamilies and the criteria for one gene to be assigned a specific phamily is to share, at least with one other member, amino acid sequence similarity at an *E* value of 0.001 (or better) or 25% amino acid identity across the length of the sequence (30).

Another useful bioinformatic tool, DNA Master, a sequence editor and analysis program that synthesizes, analyzes and presents data from several DNA analysis programs (GeneMark, Glimmer, Aragorn, BLAST and HHpred) is a key software for phage genome annotation. The automated annotation generated by DNA Master will generally identify over 80% of genes accurately, but since some genes will still need to be manually added, deleted or modified, all gene calls must be reexamined and a set of guiding principles of phage genome annotation ensures the improvement of the draft annotation. These principles were followed during phage Ms6 genome annotation and a transcript of them is provided in the appendices

1.4 Mycobacteriophage Ms6

In 1989, at the current Department of Microbiology of the Faculty of Pharmacy, University of Lisbon, Professors José Moniz-Pereira and Isabel Portugal isolated mycobacteriophage Ms6 from a spontaneously induced culture of *M. smegmatis* HB5688. This temperate phage presents a double stranded DNA genome, with over 50 kb and a siphoviral morphology, with a hexagonally shaped, isometric polyhedral head and long, non-contractile tail (31).

Phage Ms6 has been the subject of several studies, especially concerning the products encoded by its lytic cassette and embodies a model of mycobacteriophage-induced lysis. However, a complete genome annotation was still missing and was essential to assert Ms6 genomic organization and full potential.

2 Objectives

The main goal of this study was to accurately annotate the genome of phage Ms6, a recognized model of mycobacteriophage-induced lysis. Throughout this project I purpose answering the following questions:

- How many ORFs are contained in Ms6's genome?
- How is the genome of Ms6 organized and how does it compare with closely related phages?
- Which ORFs can be assigned putative functions?
- What is the biological relevance of the encoded proteins?

To help answer these questions, it was necessary to:

- Process Ms6's genome sequence with DNA Master to identify gene calls and improve the draft annotation by manual refinement.
- Run the sequences of predicted genes through several databases to obtain data concerning protein functional and/or structural aspects.
- Collect and interpret all available data to suggest how phage encoded proteins operate and contribute to phage biological processes.

We hope to better comprehend the characteristics of Ms6 and to acknowledge its true potential. The research on Ms6 and other bacteriophages leads to progress in the fields of genetics and microbiology by advancing the understanding of phage genomics and infection mechanisms. These findings may ultimately yield useful molecular biology tools or novel therapeutic approaches targeting infectious deceases.

3 Materials and Methods

3.1 Phage DNA Extraction and Sequencing

Genomic DNA extraction was adapted from Sambrook & Russell (2011) (32). A 300 µL sample of CsCl-purified lysate (~1 x 1012 pfu mL-1) was treated with 0.5% SDS, 20 mM EDTA pH 8.0 and 50 mg ml-1 proteinase K, for 1 h at 65°C, allowing virion disruption. Proteins were precipitated by addition of 550 mM KCl followed by a 30 min period of incubation on ice, and removed by centrifugation at 16,900 xg, for 15min at 4°C. The supernatant, containing Ms6 DNA, was transferred to a new tube and an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1) mixture (AppliChem) was added twice, mixed and centrifuged at 16,900 xg, for 5 min at 4°C. This procedure was repeated once with chloroform. Phage DNA was precipitated at -20°C for at least 1h after addition of an equal volume of isopropanol and 10% sodium acetate 3M. DNAwas then centrifuged at 16,900xg, for 45min at 4°C and washed with 70% (v/v) ethanol. After drying, DNA was resuspended in ultra-pure water and quantified at NanoDrop. The complete phage genome sequence was obtained by Next Generation Sequencing, through a Illumina HiSeq2500 system at BaseClear (Leiden, Netherlands).

3.2 Statistical Gene Prediction by DNA Master

The genome sequence of Ms6 was first analyzed with DNA Master (http://cobamide2.bio.pitt.edu/) by performing an automated annotation, which generated consensus gene calls supported by programs that run within DNA Master, like Glimmer (version 3.02) (33) and GeneMark (version 2.0) (34). These programs predicted genes heuristically, determining which codons are more frequently found in the longest ORFs and applying this profile to predict the coding potential of the remaining ORFs.

3.3 Additional Information and Auto-Annotation Refinement

DNA Master was also used to create a six-frame translation file, particularly useful for annotating potential ribosomal frameshifts. A GeneMark-Smeg output, a graphic representation of the coding potential based on a codon usage model for *M.*

smegmatis mc^2155, was obtained using GeneMark.hmm (version 3.25) (http://exon.gatech.edu/GeneMark/gmhmmp.cgi) and is important to find smaller genes that may not be detected by heuristic scans. Starterator (https://seaphages.org/software/#Starterator) highlights common start codons for all the genes in a certain alignment and was used to detect conserved start sites for members of a given phamily.

The collected data allowed the manual refinement of the auto-annotation and supported certain modifications, such as gene deletions, start site alterations or a programed frameshift annotation.

3.4 General Database Search

3.4.1 BLASTp and BLASTn

To predict gene functions, the protein sequences of Ms6 were processed using the BLASTp algorithm. The BLASTp data presented in the results section was obtained within DNA Master but a complementary BLASTp sequence comparison was also conducted through NCBI BLAST server (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Both analyses were performed against all protein sequences in NCBI's non-redundant database.

A whole genome BLASTn analysis was also conducted to expose Ms6's closest homologues.

3.4.2 HHpred

HHpred (https://toolkit.tuebingen.mpg.de/#/tools/hhpred) (35), integrated in the MPI Bioinformatics Toolkit (36), represents another functional assignment tool, with higher sensibility than BLASTp. Each ORF of Ms6 was analyzed with HHpred using three databases simultaneously: PDB_mmCIF70_28_Oct (contains available protein 3D crystal structures), Pfam-A_v31.0 (alignments of genetically mobile domains present in signaling, extracellular and chromatin-associated proteins) and TIGRFAM_v15.0 (collection of curated protein families that provides a tool for identifying functionally related proteins based on sequence homology).

3.5 Prediction of Structural Features

3.5.1 CATH

The existence of typical structural patterns was verified by a CATH analysis (http://www.cathdb.info/) (37). The Classification, Architecture, Topology, Homologous superfamily and percentage of identity with the query sequence of the first match was registered.

3.5.2 TMHMM and TMpred

TMHMM server 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) (38) and TMpred server (http://www.ch.embnet.org/software/TMPRED_form.html) (39) were used to predict possible transmembrane domains. The search was carried with default settings selected.

3.6 Specialized Database Search

3.6.1 Blastp on phages DB

Besides the analysis through DNA Master and NCBI BLAST server, a third BLASTp output was acquired through PhagesDB (http://phagesdb.org/blastp/). While the other sequence comparisons used a general BLAST database, the analysis on PhagesDB relies on a database consisting only on proteins of mycobacteria or mycobacteriophages, quality controlled by the University of Pittsburgh, USA.

3.6.2 Aragorn and tRNA Scan-SE

Although Aragorn (version 1.1) scans for tRNAs within DNA Master, it is necessary to complement its findings. Through web-based Aragorn (version 1.2.36) (http://mbio-serv2.mbioekol.lu.se/ARAGORN/) (40), both Ms6's DNA strands were searched for tRNAs and tmRNAs, against the standard genetic code and considering a circular topology. An analysis with tRNA Scan-SE (http://lowelab.ucsc.edu/tRNAscan-

SE/) (41) was also performed to detect tRNAs, by selecting a mixed sequence source and the legacy (tRNAscan + EufindtRNA) search mode in the search options.

3.6.3 MOTIF

A sequence motif search was performed with the meta site MOTIF (http://www.genome.jp/tools/motif/), from GenomeNet. The search was simultaneously carried out using PROSITE (42), Pfam (43) and NCBI-CDD (44) databases, the last one including COG, SMART and TIGRFAM classification systems. The identified domains, when possible, contributed for structural and functional analysis of Ms6's ORFs.

4 Results & Discussion

4.1 General Characterization of the Genome of Phage Ms6

The genome of phage Ms6 consists of 54252 base pairs of double stranded DNA, with a GC content of 61,5%. Genome annotation allowed the identification of 105 potential ORFs in phage Ms6. 77 of these ORFs (73%) present an ATG initiation codon, while 21 (20%) start with a GTG codon and only 7 genes (7%) use the TTG initiation codon. No tRNA or tmRNA genes were detected, either with DNAMaster or with Aragorn and tRNA Scan-SE.

Each of the 105 ORFs was processed by BLASTp within the DNAMaster software and complementary analysis were performed on NCBI's BLAST server, on PhagesDB platform and on HHpred. Table 1 summarizes information concerning the identified ORFs and Figure 6 is a schematic representation of Ms6 genome.

An overview of the genome enables a distinction between left and right genomic arms, delimited by the integrase gene (ORF36), which is roughly located at the genome center. The left arm genes (ORFs 1-35) encode proteins that play a role in phage structure, assembly and lysis, while the right arm genes (ORFs 37-105) appear to be involved in DNA regulation and modification functions. The mean length of all the proteins is 167 amino acids, but proteins encoded by the genes on the left and right arms have an average of 256 and 119 amino acids, respectively. Of the identified genes, all are transcribed rightwards except eight (ORFs 33-35, 37, 38 and 65-67). Although most of the genome is composed of protein-coding regions, some small non-coding regions can be found upstream ORF1 (128 bp), upstream and downstream ORF11 (126 bp and 136 bp, respectively), between ORFs 35 and 36 (485 bp), 37 and 38 (232 bp), 61 and 62 (159 bp), 93 and 94 (109 bp), and downstream of ORF105 towards the right terminus (128 bp).

A whole genome BLASTn analysis revealed Ms6's two closest homologues. Mycobacteriophage DLane is the closest relative to Ms6, with a total score of 7.60e+04, an *E* value of 0.0 and 98% identity over 81% of the entire genome length. The second closest homologue, mycobacteriophage Shauna1, presents a total score of 7.07e+04, an *E* value of 0.0 and 98% identity for a query coverage of 78%. Homology tends to be highest in the left arm (virion structure and assembly genes) and generally is faint and

Table 1 - Coordinates and predicted functions of the 105 ORFs of Ms6. The position, direction, size in amino acid (aa), molecular weight (MW) in kiloDaltons (kDa) and start codon of ach ORF are listed. The details of the best BLASTp hit within DNA Master are described in the right part of the table. In the last column, the phage cluster or subcluster of each of the gene products is also indicated.

Feature no	Coordinates		Direction	Start Codon	Size (aa)	MW (kDa)	Predicted Function	Best Match (DNA Master BLASTp)	Length	% ID	*E* value	Accession Number	Cluster/ Subcluster
	Start	Stop											
1	129	594	F	atg	151	17.1	Terminase, small subunit	gp1 MP Tweety	151	100	0.0	YP_001469234	F1
2	607	2244	F	atg	545	61.3	Terminase, large subunit	gp2 MP Fruitloop	545	100	0.0	YP_002241687	F1
3	2275	3645	F	ttg	456	51.2	Portal protein	gp3 MP PopTart	456	99	0.0	YP_009214363	F1
4	3632	4387	F	atg	251	27.4	Capsid maturation protease	gp4 MP Fruitloop	251	100	0.0	YP_002241689	F1
5	4470	5063	F	atg	197	22.3	Scaffolding protein	gp5 MP Tweety	216	100	0.0	YP_001469238	F1
6	5082	5903	F	atg	273	29.1	Major capsid protein	gp6 MP Fruitloop	273	100	0.0	YP_002241691	F1
7	5903	6469	F	gtg	188	19.7	Head-to-tail connector	gp7 MP Wee	188	100	0.0	YP_004123829	F1
8	6466	6798	F	gtg	110	11.6	Head-to-tail connector	gp8 MP Ardmore	110	100	0.0	YP_003495150	F1
9	6801	7130	F	gtg	109	11.8	Head-to-tail connector	gp9 MP Fruitloop	109	100	0.0	YP_002241694	F1
10	7117	7524	F	atg	135	14.6	Head-to-tail connector	gp10 MP Wee	135	100	0.0	YP_004123832	F1
11	7651	8460	F	atg	269	29.9	Major tail subunit	gp12 MP Quico	269	99	0.0	YP_009194536	F1
12	8597	9148	F	atg	183	20.4	Tail assembly chaperone	gp12 MP Fruitloop	183	100	0.0	YP_002241698	F1
13	8597	9533	F	atg	311	35.7	Tail assembly chaperone (+1 programmed frameshift)	gp13 MP Shauna1	311	100	0.0	AEJ92993	F1
14	9552	13040	F	atg	1162	119.2	Tape measure protein	gp14 MP Hamulus	1159	98	0.0	YP_008409078	F1
15	13041	14750	F	atg	569	63.6	Minor tail protein	gp15 MP Fruitloop	569	100	0.0	YP_002241700	F1
16	14836	16545	F	ttg	569	64.1	Minor tail protein	gp16 MP Tweety	569	100	0.0	YP_001469249	F1
17	16600	17430	F	atg	276	28.5	Minor tail protein	gp18 MP Drago	282	100	0.0	YP_009016096	F1
18	17427	19739	F	atg	770	82.3	Minor tail protein	gp20 MP Redi	768	90	0.0	YP_009017084	N
19	19750	20904	F	atg	384	38.1	Minor tail protein	gp20 MP Velveteen	385	96	0.0	YP_008409557	F1
20	20901	21140	F	atg	79	8.7		gp22 MP Redi	79	99	0.0	YP_009017086	N
21	21140	21520	F	atg	126	12.8		gp22 MP Velveteen	127	96	0.0	YP_008409559	F1
22	21520	21696	F	atg	58	6.3		gp23 MP Velveteen	58	93	4.0e-19	YP_008409560	F1
23	21767	22000	F	atg	77	8.3	Chaperone-like protein	gp29 MP Ovechkin	77	96	1.4e-31	YP_009211193	F1
24	21997	23151	F	gtg	384	43.4	LysA	gp29 MP Fruitloop	384	100	0.0	YP_002241714	F1
25	23151	24149	F	atg	332	37.2	LysB	gp32 MP Wee	332	100	0.0	YP_004123854	F1
26	24159	24392	F	atg	77	7.9	Holin	gp32 MP Tweety	77	100	2.9e-26	YP_001469265	F1
27	24389	24763	F	atg	124	14.1	Holin	gp35 MP Boomer	124	99	0.0	YP_002014251	F1
28	24788	25021	F	atg	77	8.6		gp36 MP WillSterrel	77	100	0.0	AOQ28491	F1
29	25008	25829	F	atg	273	31.4	DnaQ exonuclease	gp35 MP XFactor	273	100	0.0	YP_009208776	F1
30	25831	25926	F	gtg	31	3.6		gp46 MP Koguma	31	100	6.8e-14	ATN90010	C1
31	25913	26134	F	gtg	73	7.9		gp36 MP Saal	73	99	7.8e-44	YP_009007503	F1
32	26127	26219	F	gtg	30	3.3		gp36 MP Fruitloop	30	97	2.0e-10	YP_002241721	F1
33	26270	26779	R	atg	169	19.9	Transcriptional regulation	gp37 MP Fruitloop	190	98	0.0	YP_002241722	F1
34	26776	26970	R	atg	64	7.3		gp41 MP Bobi	64	98	4.3e-16	YP_008408998	F1
35	26967	27164	R	atg	65	7.4		gp40 MP DLane	80	100	3.3e-39	AEK08584	F1
36	27650	28768	F	atg	372	41.6	Integrase	gp51 MP Mozy	372	100	0.0	AEK09665	F1
37	28834	29322	R	ttg	162	17.0	Pin	gp42 MP Dlane	162	99	0.0	AEK08586	F1
38	29555	31154	R	ttg	199	22.3	Transcriptional regulation	gp44 MP SG4	199	100	0.0	YP_009013246	F1
39	30165	30524	F	ttg	119	13.3	Transcriptional regulation	gp45 MP Fruitloop	119	98	0.0	YP_002241730	F1
40	30581	30868	F	atg	95	10.6		gp50 MP Bobi	95	100	0.0	YP_008409007	F1
41	30865	30987	F	atg	40	4.6		gp51 MP Bobi	40	100	1.1e-17	YP_008409008	F1
42	30984	31250	F	atg	88	10.2	Transcriptional regulation	gp45 MP Velveteen	86	99	1.8e-42	YP_008409582	F1
43	31247	31531	F	atg	94	10.6		gp48 MP Bipolar	76	48	7.0e-10	YP_009200675	F1
44	31524	31688	F	atg	54	6.0		gp49 MP Fruitloop	49	98	5.0e-22	YP_002241734	F1
45	31685	32062	F	gtg	125	14.0		hypothetical protein *Mycobacterium abscessus*	131	69	1.3e-43	WP_074378334	Not applicable
46	32076	32312	F	atg	78	8.8		hypothetical protein *Acinetobacter baumannii*	75	56	1.7e-22	WP_062937353	Not applicable
47	32309	32989	F	gtg	226	24.5		gp51 MP Fruitloop	223	84	0.0	YP_002241736	F1
48	33024	33764	F	atg	246	25.4		gp52 MP XFactor	246	90	0.0	YP_009208793	F1
49	33845	34126	F	atg	93	10.8		gp56 MP Emma	93	100	0.0	ASZ72935	F1

Table 1 (Continued)

Feature no	Coordinates		Direction	Start Codon	Size (aa)	MW (kDa)	Predicted Function	Best Match (DNA Master BLASTp)	Length	% ID	*E* value	Accession Number	Cluster/ Subcluster
	Start	Stop											
50	34123	34311	F	atg	62	6.6		gp61 MP Bobi	62	100	1.7e-35	YP_008409018	F1
51	34315	34650	F	atg	111	11.9		gp54 MP Fruitloop	111	100	0.0	YP_002241739	F1
52	34650	35051	F	atg	133	15.6	Transcriptional Factor WhiB	gp59 MP Hamulus	133	100	0.0	YP_008409123	F1
53	35048	35548	F	atg	166	18.8		gp56 MP Fruitloop	166	100	0.0	YP_002241741	F1
54	35545	35910	F	atg	121	13.7	Transcriptional regulation	gp54 MP PopTart	121	98	0.0	YP_009214414	F1
55	35910	36065	F	atg	51	6.3		gp64 MP SuperGrey	52	100	1.3e-30	APC43595	F1
56	36062	36292	F	atg	76	8.3		gp61 MP Tweety	76	100	2.5e-29	YP_001469294	F1
57	36289	36435	F	atg	48	5.3		gp62 MP Tweety	48	100	2.0e-25	YP_001469295	F1
58	36428	36706	F	atg	92	10.2		gp63 MP Ovechkin	92	99	0.0	YP_009211227	F1
59	36703	37185	F	atg	160	18.1	HNH endonuclease	gp22 MP WIVsmall	160	98	0.0	YP_008059923	F1
60	37182	37760	F	atg	192	20.6		gp63 MP Seagreen	192	100	0.0	YP_009199746	F1
61	37757	38119	F	atg	120	13.8	HNH endonuclease	gp73 MP Quico	120	98	0.0	YP_009194597	F1
62	38279	38434	F	gtg	51	5.7		gp73 MP Wee	51	100	1.3e-30	YP_004123895	F1
63	38427	38777	F	ttg	116	12.8		gp74 MP Wee	116	99	0.0	YP_004123896	F1
64	38774	40279	F	gtg	501	52.4	DNA methylase	gp68 MP Hamulus	486	84	0.0	YP_008409132	F1
65	40248	40409	R	atg	53	6.0	Transcriptional regulation	gp78 MP Quico	53	100	4.7e-29	YP_009194602	F1
66	40402	40779	R	atg	125	13.4		gp57 MP Babsiella	125	99	0.0	YP_009013074	I1
67	40786	41118	R	atg	110	12.3		gp58 MP Babsiella	110	100	0.0	YP_009013075	I1
68	41151	41273	F	atg	40	4.9		gp74 MP Ovechkin	40	98	4.2e-19	YP_009211238	F1
69	41320	42597	F	atg	425	45.8	DNA methylase	gp70 MP DLane	419	87	0.0	AEK08614	F1
70	42594	42728	F	atg	44	5.4		gp79 MP Wee	44	100	3.4e-25	YP_004123901	F1
71	42732	43583	F	gtg	283	32.2		gp76 MP Squirty	283	100	0.0	YP_009124628	F3
72	43580	43843	F	gtg	87	9.4		gp85 MP Quico	87	94	0.0	YP_009194609	F1
73	43840	44031	F	atg	63	7.7		gp62 MP Ardmore	63	98	7.7e-39	YP_003495203	F1
74	44028	44228	F	atg	66	7.2		gp78 MP Hades	66	99	6.0e-41	YP_009125257	F1
75	44212	44604	F	atg	130	14.0		gp82 MP Minerva	113	91	1.2e-16	YP_009124035	J
76	44601	45065	F	gtg	154	16.4		gp73 MP Llij	142	83	0.0	YP_655069	F1
77	45058	45243	F	atg	61	7.3		gp74 MP Llij	61	100	1.9e-22	YP_655070	F1
78	45240	45515	F	atg	91	10.0		gp41 MP NelitzaMV	91	100	0.0	YP_009197709	E
79	45512	45724	F	gtg	70	8.0		gp80 MP SG4	80	100	0.0	YP_009013282	F1
80	45721	46005	F	gtg	94	10.7		gp76 MP Bubbles123	94	100	0.0	APU93073	F1
81	45998	46339	F	atg	113	12.5		gp60 MP Kumao	113	99	0.0	ATN94023	Singleton
82	46336	46749	F	atg	137	15.9		gp85 MP Fruitloop	137	86	0.0	YP_002241770	F1
83	46746	46961	F	atg	71	7.9		gp86 MP ShiLan	71	100	1.8e-28	AEJ93269	F1
84	46954	47142	F	gtg	62	7.2		gp81 MP Inventum	62	100	1.4e-37	YP_009125362	F1
85	47139	47315	F	gtg	58	6.6		gp86 MP Hamulus	58	98	1.1e-34	YP_008409150	F1
86	47312	47503	F	gtg	63	7.4		gp90 MP Bobi	63	100	1.8e-39	YP_008409047	F1
87	47500	47679	F	atg	59	6.8		gp91 MP Bobi	66	98	6.3e-35	YP_008409048	F1
88	47679	47816	F	atg	45	5.0		gp74 MP Ardmore	51	100	2.4e-23	YP_003495215	F1
89	47813	47986	F	atg	57	6.7		gp75 MP Ardmore	57	100	1.0e-33	YP_003495216	F1
90	47983	48102	F	ttg	39	4.3		gp86 MP Saal	39	97	5.8e-20	YP_009007553	F1
91	48099	48428	F	atg	109	12.3		gp82 MP Hamulus	109	100	0.0	YP_008409156	F1
92	48425	48688	F	atg	87	10.0		gp77 MP Ardmore	87	100	6.5e-43	YP_003495218	F1
93	48685	49326	F	atg	213	24.6		gp94 MP Tweety	213	100	0.0	YP_001469327	F1
94	49436	49660	F	atg	74	7.9	Transcriptional regulation	gp95 MP Tweety	74	100	2.2e-34	YP_001469328	F1
95	49657	49824	F	gtg	55	6.1		gp80 MP Ardmore	55	100	1.3e-31	YP_003495221	F1
96	49831	50040	F	atg	69	7.4		gp94 MP Dante	69	100	6.1e-42	YP_009212736	F1
97	50040	50237	F	atg	65	7.0		gp92 MP Llij	65	100	2.5e-40	YP_655088	F1
98	50234	50647	F	atg	137	14.9		gp99 MP Seagreen	137	99	0.0	YP_009199782	F1
99	50644	50886	F	gtg	80	8.7		gp100 MP Tweety	80	100	8.1e-39	YP_001469333	F1
100	50879	51040	F	atg	53	6.3		gp101 MP Job42	53	100	4.7e-31	YP_008126691	F1
101	51057	52493	F	atg	478	54.8	Glycosyltransferase	gp101 MP DLane	478	99	0.0	AEK08645	F1
102	5496	52969	F	atg	157	18.2	Ser/Thr kinase	gp102 MP DLane	157	99	0.0	AEK08646	F1
103	52984	53178	F	atg	64	7.2		gp102 MP Dante	90	100	4.6e-41	YP_009212744	F1
104	53175	53798	F	atg	207	23.5	Glycosyltransferase	gp103 MP Gumbie	207	96	0.0	YP_009018979	F1
105	53798	54124	F	atg	108	12.0	HNH endonuclease	gp101 MP Cabrinians	108	99	0.0	YP_009189823	F1

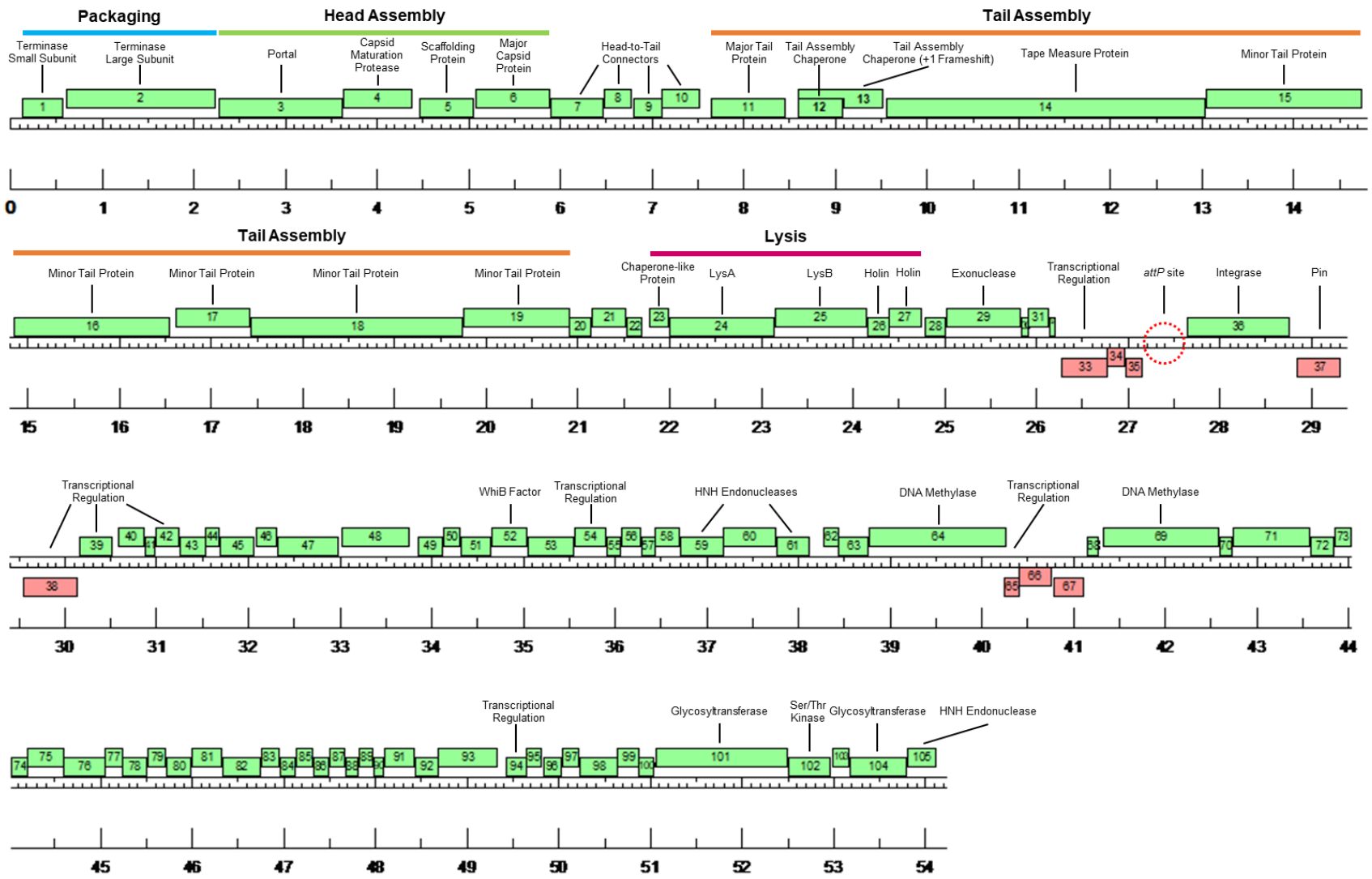


Figure 6 - Map of Ms6 genome. The 105 identified ORFs are represented by green or red colored boxes, depending if they have a forward or reverse direction, respectively. The predicted functions are pointed out, as well as the predicted *attP* site (red circle).

discrete in the right part of the genome (DNA replication and phage regulation genes) (45).

Ms6's ORFs also match genes found in other mycobacteriophages, such as Fruitloop, Tweety, Ardmore, Hamulus and Ovechkin. Like DLane and Shauna1, all these phages belong to cluster F, specifically subcluster F1, and since Ms6 shares a high degree of homology with most phages of this group, one can assume it belongs to subcluster F1, as was purposed before by Dr. Hatfull (16).

Table 2 gathers additional information obtained through several programs. Structural comparisons with known proteins were computer-predicted with CATH. Motifs in amino acid sequence were searched with MOTIF and the presence of transmembrane helices was evaluated with TMHMM and TMpred.

4.2 Functional ORF Prediction

Of the 105 potential ORFs identified, 43 were assigned putative functions according to homology with characterized genes. Taking into the account typical genome organization of closely related bacteriophages, Ms6's ORFs were arranged into seven distinct segments. Only relevant ORFs are discussed further within each section.

4.2.1 Packaging

In most DNA phages, genome replication results in the accumulation of concatemers, which are cut at precise sites, yielding numerous copies of mature, single virion DNA. DNA cutting is synchronized with DNA packaging, which is an ATP-driven translocation of the processed DNA into the capsid precursor (46). The packaging enzyme, terminase, plays a key role in this process. Terminase is composed of two subunits: the small subunit is a DNA binding protein, while the large subunit possesses nuclease and ATPase activities (47). Phage DNA packaging is represented in Figure 7 (48).

In Ms6, the small and large subunits of terminase are predicted to be encoded by ORF1 and ORF2, respectively. Most mycobacteriophages's homologues of ORF2 are annotated as genes encoding the large subunit of terminase, which strengthens the

Table 2 - Structural analysis, motif search and transmembrane domains for each identified ORF. The classification, architecture, topology and homologous family of the first CATH match is documented on the left part of the table, as well as the respective CATH code and percentage of identity with the query sequence. On the right are listed the results of motif and transmembrane domain search[a].

ORF	CATH Code	%ID	Classification	Architecture	Topology	Homologous Superfamily	Motif search trough MOTIF scanning PROSITE, Pfam and NCBI-CDD databases	Transmembrane helices search with TMHMM and Tmpred
1	2.40.50.140	36.8	Mainly Beta	Beta Barrel	OB fold (Dihydrolipoamide Acetyltransferase, E2P)	Nucleic acid-binding proteins	NCBI-CDD: (317440) Cation channel sperm-associated protein subunit delta	
2	3.40.50.300	6.4	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold	P-loop containing nucleotide triphosphate hydrolases	NCBI-CDD: (226974) Phage terminase-like protein, large subunit	
3	3.40.630.30	12.0	Alpha Beta	3-Layer (aba) Sandwich	Aminopeptidase		NCBI-CDD: (310019) Phage portal protein, SPP1 Gp6-like; Pfam: (Phage_prot_Gp6) Phage portal protein, SPP1 Gp6-like	
4	1.10.20.10	39.3	Mainly Alpha	Orthogonal Bundle	Histone, subunit A	Histone, subunit A	NCBI-CDD: (185513) Cathepsin L protease	
5	1.10.238.10	20.2	Mainly Alpha	Orthogonal Bundle	Recoverin; domain 1	EF-hand	NCBI-CDD: (236945) Sensory histidine kinase UhpB; Pfam: (GP70) Gene 70 protein	
6	2.130.10.10	38.3	Mainly Beta	7 Propellor	Methylamine Dehydrogenase; Chain H	YVTN repeat-like/Quinoprotein amine dehydrogenase	NCBI-CDD: (314507) P22 coat protein - gene protein 5; Pfam: (P22_CoatProtein) P22 coat protein - gene protein 5	
7	3.30.450.20	36.5	Alpha Beta	2-Layer Sandwich	Beta-Lactamase		NCBI-CDD: (275156) Choice-of-anchor C domain; Pfam: (DUF3992) Protein of uknown function	TMpred: One transmembrane helix, N-terminus outside
8	3.20.20.80	91.0	Alpha Beta	Alpha-Beta Barrel	TIM Barrel	Glycosidases	NCBI-CDD: (226067) Putative alpha-1,2-mannosidase; Pfam: (TGT) Queuine tRNA-ribosyltransferase	TMpred: One transmembrane helix, N-terminus outside
9	3.40.50.1110	49.1	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold		NCBI-CDD: (309835) Bacteriophage HK97-gp10, putative tail-component; Pfam: (HK97-gp10_like) Bacteriophage HK97-gp10, putative tail-component	
10	3.60.20.10	22.8	Alpha Beta	4-Layer Sandwich	Glutamine Phosphoribosylpyro-phosphate, subunit 1, domain 1	Glutamine Phosphoribosylpyro-phosphate, subunit 1, domain 1	NCBI-CDD: (182507) Putative outer membrane lipoprotein	TMpred: One transmembrane helix, N-terminus outside
11	3.10.450.50	32.2	Alpha Beta	Roll	Nuclear Transport Factor 2; Chain: A		NCBI-CDD: (240656) Putative D-isomer specific 2-hydroxyacid dehydrogenases, NAD-binding and catalytic domains	TMpred: One transmembrane helix, N-terminus inside
12	3.40.50.1760	21.7	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold		NCBI-CDD: (313222) Predicted integral membrane protein	
13	2.140.10.30	17.6	Mainly Beta	8 Propellor	Methanol Dehydrogenase; Chain A		NCBI-CDD: (162276) Fumarylacetoacetase	
14	1.10.132.20	4.4	Mainly Alpha	Orthogonal Bundle	Topoisomerase I; Chain A, domain 4		NCBI-CDD: (226450) Mu-like prophage protein; Pfam: (TMP_2) Prophage tail length tape measure protein	TMpred: Eight transmembrane helices, N-terminus inside
15	3.40.50.300	16.3	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold	P-loop containing nucleotide triphosphate hydrolases	NCBI-CDD: (310369) Phage tail protein; Pfam: (Sipho_tail) Phage tail protein	
16	3.40.50.1260	14.0	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold		NCBI-CDD: (315249) Patatin phospholipase	TMpred: Two transmembrane helices, N-terminus inside
17	2.20.230.10	15.2	Mainly Beta	Single Sheet	Resuscitation-promoting factor rpfb fold	Resuscitation-promoting factor rpfb	NCBI-CDD: (284802) Merozoite surface protein 1 (MSP1) C-terminus	
18	3.30.420.40	5.3	Alpha Beta	2-Layer Sandwich	Nucleotidyltransferase; domain 5		NCBI-CDD: (227608) Phage-related tail protein	TMpred: Six transmembrane helices, N-terminus outside
19	3.40.800.10	21.6	Alpha Beta	3-Layer (aba) Sandwich	Arginase; Chain A		NCBI-CDD: (180777) Single-stranded DNA-binding protein	TMpred: Two transmembrane helices, N-terminus outside
20	3.40.50.300	61.3	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold	P-loop containing nucleotide triphosphate hydrolases	NCBI-CDD: (238864) Acyloxyacyl-hydrolase like subfamily of the SGNH-hydrolase family	
21	3.40.190.10	26.0	Alpha Beta	3-Layer (aba) Sandwich	D-Maltodextrin-Binding Protein; domain 2	Periplasmic binding protein-like II	NCBI-CDD: (180937) Aminotransferase	TMpred: Two transmembrane helices, N-terminus inside
22	No matches to CATH domains						NCBI-CDD: (234697) Cobalamin synthase; Pfam: (HCV_NS4a) Hepatitis C virus non-structural protein NS4a	TMHMM and TMpred: One transmembrane helix, N-terminus outside
23	3.40.50.1950	48.7	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold		NCBI-CDD: (222949) S-S bond formation pathway protein	
24	3.40.80.10	25.5	Alpha Beta	3-Layer (aba) Sandwich	Lysozyme-like	Lysozyme	NCBI-CDD: (214760) Ami_2 domain; (133475) Peptidoglycan recognition proteins (PGRPs) Pfam: (Amidase_2) N-acetylmuramoyl-L-alanine amidase	TMpred: One transmembrane helix, N-terminus inside
25	3.40.50.1820	52.3	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold		Pfam: (PE-PPE) PE-PPE domain; (Cutinase) Cutinase	TMpred: One transmembrane helix, N-terminus inside
26	3.20.20.70	38.5	Alpha Beta	Alpha-Beta Barrel	TIM Barrel	Aldolase class I	NCBI-CDD: (293550) Putative lactococcus lactis phage r1t holin; Pfam: (Phage_r1t_holin) Putative lactococcus lactis phage r1t holin	TMpred: One transmembrane helix, N-terminus outside
27	3.90.20.10	24.0	Alpha Beta	Alpha-Beta Complex	Hemagglutinin Ectodomain; Chain B		NCBI-CDD: (313949) protein of unknown function; Pfam: (DUF2746) Protein of unknown function	TMHMM and TMpred: One transmembrane helix, N-terminus outside

Table 2 (Continued)

ORF	CATH Code	%ID	Classification	Architecture	Topology	Homologous Superfamily	Motif search trough MOTIF scanning PROSITE, Pfam and NCBI-CDD databases	Transmembrane helices search with TMHMM and Tmpred
28	3.40.50.1100	35.9	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold		NCBI-CDD: (133021) Prokaryotic UGPase	
29	2.60.40.10	18.3	Mainly Beta	Sandwich	Immunoglobulin-like	Immunoglobulins	NCBI-CDD: (176646) DEDDy 3' -5' exonuclease domain of family-B DNA polymerases; Pfam: (Lar_restr_allev) Restriction alleviation protein Lar	
30	No matches to CATH domains						NCBI-CDD: (306751) Lipoxygenase	
31	2.30.30.40	39.2	Mainly Beta	Roll	SH3 type barrels	SH3 Domains		
32	No matches to CATH domains							TMHMM and TMpred: One transmembrane helix, N-terminus outside
33	1.10.357.10	23.5	Mainly Alpha	Orthogonal Bundle	Tetracycline Repressor; domain 2	Tetracycline Repressor; domain 2	NCBI-CDD: (225320) Predicted transcriptional regulator; Pfam: (HTH_23) Homeodomain-like domain	
34	No matches to CATH domains						NCBI-CDD: (276832) Class VII myosin, motor domain	
35	3.40.1280.10	84.9	Alpha Beta	3-Layer (aba) Sandwich	Alpha/beta knot		NCBI-CDD: (132260) 2-hydroxymuconic semialdehyde dehydrogenase	
36	1.10.443.10	43.2	Mainly Alpha	Orthogonal Bundle	hpI Integrase; Chain A	Integrase catalytic core	NCBI-CDD: (271189) C-terminal catalytic domain of integrases from bacterial phages and conjugate transposons; Pfam: (Phage_integrase) Phage integrase family	
37	3.20.20.80	11.0	Alpha Beta	Alpha-Beta Barrel	TIM Barrel	Glycosidases	Pfam: (YjbE) Exopolysaccharide production protein YjbE	TMHMM and TMpred: One transmembrane helix, N-terminus outside
38	No matches to CATH domains						NCBI-CDD: (316117) Helix-turn-helix domain; Pfam: (HTH_31) Helix-turn-helix domain	TMpred: One transmembrane helix, N-terminus inside
39	1.10.260.40	59.2	Mainly Alpha	Orthogonal Bundle	434 Repressor (Amino-terminal Domain)	lambda repressor-like DNA-binding domains	PROSITE profile: (HTH_CROC1) Cro/C1-type HTH domain profile; NCBI-CDD: (238045) Helix-turn-helix XRE-family like proteins; Pfam: (HTH_31) Helix-turn-helix domain	
40	3.30.950.10	26.0	Alpha Beta	2-Layer Sandwich	Methyltransferase, Cobalt-precorrin-4 Transmethylase; Domain 2	Methyltransferase, Cobalt-precorrin-4 Transmethylase; Domain 2	NCBI-CDD: (318377) Domain of unknown function	
41	No matches to CATH domains						Pfam: (DAP) Death-associated protein	
42	2.60.40.10	37.1	Mainly Beta	Sandwich	Immunoglobulin-like	Immunoglobulins	NCBI-CDD: (314606) Histone lysine methyltransferase SET associated; Pfam: (PyocinActivator) Pyocin activator protein PrtN	
43	2.115.10.20	28.4	Mainly Beta	5 Propellor	Tachylectin-2; Chain A	Glycosyl hydrolase domain; family 43	NCBI-CDD: (288228) Protein of unknown function; Pfam: (DUF3138) Protein of unknown function	
44	3.90.79.10	69.1	Alpha Beta	Alpha-Beta Complex	Nucleoside Triphosphate Pyrophosphohydrolase	Nucleoside Triphosphate Pyrophosphohydrolase	NCBI-CDD: (237560) Rod shape-determining protein MreC; Pfam: (DUF4763) Domain of unknown function	
45	1.10.357.10	35.7	Mainly Alpha	Orthogonal Bundle	Tetracycline Repressor; domain 2	Tetracycline Repressor; domain 2	Pfam: (CENP-P) CENP-A-nucleosome distal (CAD) centromere subunit, CENP-P	
46	3.30.500.10	41.8	Alpha Beta	2-Layer Sandwich	Murine Class I Major Histocompatibility Complex, H2-DB;	Murine Class I Major Histocompatibility Complex, H2-DB;	NCBI-CDD: (129037) dDENN; Pfam: (SDP_N) Sex determination protein N terminal	
47	3.20.20.120	20.3	Alpha Beta	Alpha-Beta Barrel	TIM Barrel	Enolase superfamily	NCBI-CDD: (237813) Lipoprotein signal peptidase	TMpred: One transmembrane helix, N-terminus inside
48	3.40.50.200	22.3	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold		NCBI-CDD: (237537) Conjugal transfer protein TrbL	TMpred: Three transmembrane helices, N-terminus outside
49	3.20.20.120	23.4	Alpha Beta	Alpha-Beta Barrel	TIM Barrel	Enolase superfamily		
50	1.20.1090.10	71.4	Mainly Alpha	Up-down Bundle	Dehydroquinate synthase-like, apha domain	Dehydroquinate synthase-like - alpha domain		TMHMM and TMpred: Two transmembrane helices, N-terminus inside
51	3.40.640.10	20.5	Alpha Beta	3-Layer (aba) Sandwich	Aspartate Aminotransferase; domain 2	Type I PLP-dependent aspartate aminotransferase-like (Major domain)	NCBI-CDD: (310133) Protein of unknown function; Pfam: (DUF732) Protein of unknown function	TMpred: One transmembrane helix, N-terminus outside
52	3.40.50.1780	53.7	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold		PROSITE profile: (4FE4S_WBL) 4Fe-4S WhiB-like (Wbl)-type iron-sulfur binding domain profile; NCBI-CDD: (308210) Transcription factor WhiB; Pfam: (Whib) Transcription factor WhiB	
53	3.40.50.1580	29.3	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold		NCBI-CDD: (177363) Hypothetical protein	

Table 2 (Continued)

ORF	CATH Code	%ID	Classification	Architecture	Topology	Homologous Superfamily	Motif search trough MOTIF scanning PROSITE, Pfam and NCBI-CDD databases	Transmembrane helices search with TMHMM and Tmpred
54	1.10.357.10	25.4	Mainly Alpha	Orthogonal Bundle	Tetracycline Repressor; domain 2	Tetracycline Repressor; domain 2	PROSITE Pattern: (HTH_LUXR_1) LuxR-type HTH domain signature; NCBI-CDD: (259851) Helix-turn-helix domain of Hin and related proteins; Pfam: (HTH_23) Homeodomain-like domain	
55	3.40.50.1360	82.7	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold		NCBI-CDD: (129593) Glucosamine-6-phosphate isomerase	
56	No matches to CATH domains							
57	No matches to CATH domains						Pfam: (Pex14_N) Peroxisomal membrane anchor protein conserved region	
58	3.10.100.10	32.3	Alpha Beta	Roll	Mannose-Binding Protein A; Chain A	Mannose-Binding Protein A, subunit A	PROSITE Pattern: (HEMOPEXIN) Hemopexin domain signature; NCBI-CDD: (164916) Hypothetical protein	
59	2.30.30.20	19.9	Mainly Beta	Roll	SH3 type barrels	Aspartate carbamoyltransferase, Regulatory-chain, C-terminal domain	NCBI-CDD: (315955) HNH endonuclease; Pfam: (HNH_3) HNH endonuclease	
60	3.30.1490.20	17.1	Alpha Beta	2-Layer Sandwich	DNA Ligase; domain 1	ATP-grasp fold, A domain	NCBI-CDD: (107186) hypothetical protein; Pfam: (Rad52_Rad22) Rad52/22 family double-strand break repair protein	
61	1.10.1130.10	37.2	Mainly Alpha	Orthogonal Bundle	Flavocytochrome C3; Chain A, domain 2	Flavocytochrome C3; Chain A	NCBI-CDD: (238038) HNH nucleases; Pfam: (HNH) HNH endonuclease	
62	No matches to CATH domains						NCBI-CDD: (225307) Uncharacterized FAD-dependent dehydrogenases [General function prediction only].	
63	3.40.50.10730	28.2	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold	Urocanase like domains		
64	3.40.50.150	34.9	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold	Vaccinia Virus protein VP39	PROSITE Pattern: (CYTOCHROME_B5_1) Cytochrome b5 family, heme-binding domain signature; PROSITE Profile: (SAM_MT_C5) C-5 cytosine-specific DNA methylase (Dnmt) domain profile; NCBI-CDD: (223348) Site-specific DNA methylase; Pfam: (DNA_methylase) C-5 cytosine-specific DNA methylase	
65	3.20.20.80	48.2	Alpha Beta	Alpha-Beta Barrel	TIM Barrel	Glycosidases	NCBI-CDD: (279710) Ribbon-helix-helix protein, copG family; Pfam: (RHH_1) Ribbon-helix-helix protein, copG family	
66	2.40.128.20	71.4	Mainly Beta	Beta Barrel	Lipocalin		NCBI-CDD: (313308) Uncharacterized small protein; Pfam: (DUF2292) Uncharacterized small protein	
67	3.90.550.10	27.8	Alpha Beta	Alpha-Beta Complex	Spore Coat Polysaccharide Biosynthesis Protein SpsA; Chain A	Spore Coat Polysaccharide Biosynthesis Protein SpsA; Chain A	NCBI-CDD: (225373) NTP pyrophosphohydrolases containing a Zn-finger, probably nucleic-acid-binding	TMpred: One transmembrane helix, N-terminus outside
68	No matches to CATH domains							
69	3.40.50.150	26.1	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold	Vaccinia Virus protein VP39	PROSITE Pattern: (N6_MTASE) N-6 Adenine-specific DNA methylases signature; NCBI-CDD: (307613) DNA methylase; Pfam: (N6_N4_Mtase) DNA methylase	
70	No matches to CATH domains						NCBI-CDD: (185174) Subtilase cytotoxin subunit B-like protein	
71	1.25.40.20	15.9	Mainly Alpha	Alpha Horseshoe	Serine Threonine Protein Phosphatase 5, Tetratricopeptide repeat		NCBI-CDD: (311458) SecA DEAD-like domain	
72	3.40.50.300	47.7	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold	P-loop containing nucleotide triphosphate hydrolases	NCBI-CDD: (131771) Phenylphosphate carboxylase, beta subunit	
73	2.60.40.10	39.1	Mainly Beta	Sandwich	Immunoglobulin-like	Immunoglobulins	NCBI-CDD: (275128) Putative methyltransferase	
74	3.90.180.10	61.2	Alpha Beta	Alpha-Beta Complex	Quinone Oxidoreductase; Chain A, Domain 1	Medium-chain alcohol dehydrogenases, catalytic domain	NCBI-CDD: (132245) 6-hydroxycyclohex-1-ene-1-carbonyl-CoA dehydrogenase	
75	2.115.10.20	19.1	Mainly Beta	5 Propellor	Tachylectin-2; Chain A	Glycosyl hydrolase domain; family 43	PROSITE profile: (PROKAR_LIPOPROTEIN) Prokaryotic membrane lipoprotein lipid attachment site profile; NCBI-CDD: (223004) UL37 tegument protein	TMpred: One transmembrane helix, N-terminus outside
76	3.80.10.10	22.6	Alpha Beta	Alpha-Beta Horseshoe	Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)	Ribonuclease Inhibitor	NCBI-CDD: (307807) Protein of unknown function DUF 45	
77	1.10.630.10	59.7	Mainly Alpha	Orthogonal Bundle	Cytochrome p450	Cytochrome p450	NCBI-CDD: (177654) Protochlorophyllide reductase	

Table 2 (Continued)

ORF	CATH Code	%ID	Classification	Architecture	Topology	Homologous Superfamily	Motif search trough MOTIF scanning PROSITE, Pfam and NCBI-CDD databases	Transmembrane helices search with TMHMM and Tmpred
78	3.50.90.10	39.1	Alpha Beta	3-Layer (aba) Sandwich	YerB-like fold	YerB-like	NCBI-CDD: (226909) ABC-type uncharacterized transport system, periplasmic component	
79	2.60.40.10	39.4	Mainly Beta	Sandwich	Immunoglobulin-like	Immunoglobulins	NCBI-CDD: (214439) Cytochrome c oxidase subunit III	
80	No matches to CATH domains						NCBI-CDD: (183649) Putative monovalent cation/H+ antiporter subunit A	
81	No matches to CATH domains						NCBI-CDD: (223634) Phosphoserine phosphatase	
82	3.20.20.70	23.2	Alpha Beta	Alpha-Beta Barrel	TIM Barrel	Aldolase class I		
83	No matches to CATH domains						NCBI-CDD: (240647) Putative D-isomer specific 2-hydroxyacid dehydrogenases	
84	No matches to CATH domains						NCBI-CDD: (200575) Catalytic NodB homology domain of Colletotrichum lindemuthianum chitin deacetylase and similar proteins	
85	No matches to CATH domains						NCBI-CDD: (183451) Translation initiation factor IF-2 subunit beta; Pfam: (Ribosomal_S27e) Ribosomal protein S27	
86	2.60.40.10	73.4	Mainly Beta	Sandwich	Immunoglobulin-like	Immunoglobulins	Pfam: (DUF3619) Protein of unknown function	
87	3.30.70.150	78.3	Alpha Beta	2-Layer Sandwich	Alpha-Beta Plaits		NCBI-CDD: (165164) Hypothetical protein	
88	No matches to CATH domains							
89	3.20.20.70	48.3	Alpha Beta	Alpha-Beta Barrel	TIM Barrel	Aldolase class I	NCBI-CDD: (237030) Alpha-ketoglutarate decarboxylase	
90	No matches to CATH domains						NCBI-CDD: (223430) Recombinational DNA repair protein (RedF pathway)	
91	1.50.10.10	44.6	Mainly Alpha	Alpha/alpha barrel	Glycosyltransferase		NCBI-CDD: (274106) DNA-directed RNA polymerase subunit A'; Pfam: (RNA_pol_Rpb1_1) RNA polymerase Rpb1, domain 1	
92	1.10.490.10	39.8	Mainly Alpha	Orthogonal Bundle	Globin-like	Globins	NCBI-CDD: (270335) C-lobe of transferrin	
93	3.30.360.10	27.1	Alpha Beta	2-Layer Sandwich	Dihydrodipicolinate Reductase; domain 2	Dihydrodipicolinate Reductase; domain 2	NCBI-CDD: (223625) Zn-finger domain associated with topoisomerase type I; Pfam: (zf-C4_Topoisom) Topoisomerase DNA binding C4 zinc finger	
94	1.10.1660.10	62.7	Mainly Alpha	Orthogonal Bundle	Multidrug-efflux Transporter Regulator; Chain: A; Domain 2		NCBI-CDD: (315411) Helix-turn-helix domain; Pfam: (HTH_17) Helix-turn-helix domain	TMHMM and TMpred: One transmembrane helix, N-terminus outside
95	No matches to CATH domains						NCBI-CDD: (211425) Middle domain of the origin recognition complex, subunit 6	
96	3.90.80.10	81.4	Alpha Beta	Alpha-Beta Complex	Inorganic Pyrophosphatase	Inorganic Pyrophosphatase	NCBI-CDD: (270166) Engulfment and cell motility protein 1 pleckstrin homology domain	
97	4.10.70.10	33.3	Few Secondary Structures	Irregular	Echistatin	Echistatin	NCBI-CDD: (185216) RatA-like protein; Pfam: (Tetradecapep) Myoactive tetradecapeptides family	
98	3.40.50.300	34.1	Alpha Beta	3-Layer (aba) Sandwich	Rossmann fold	P-loop containing nucleotide triphosphate hydrolases	NCBI-CDD: (274420) Methane monooxygenase/ammonia monooxigenase, subunit C	
99	2.130.10.10	25.9	Mainly Beta	7 Propellor	Methylamine Dehydrogenase; Chain H	YVTN repeat-like/Quinoprotein amine dehydrogenase	NCBI-CDD: (184198) Cobalt transporter ATP-binding subunit	TMpred: One transmembrane helix, N-terminus outside
100	3.30.2080.10	68.5	Alpha Beta	2-Layer Sandwich	GH92 mannosidase fold	GH92 mannosidase domain	NCBI-CDD: (226031) Phage protein D	
101	3.20.20.80	23.9	Alpha Beta	Alpha-Beta Barrel	TIM Barrel	Glycosidases	NCBI-CDD: (132997) Glycosyltransferase family A; Pfam: (Glycos_transf_2) Glycosyl transferase family 2	
102	1.10.510.10	26.0	Mainly Alpha	Orthogonal Bundle	Transferase (Phosphotransferase); domain 1	Transferase (Phosphotransferase) domain 1	NCBI-CDD: (279908) Phosphotransferase enzyme family; Pfam: (APH) Phosphotransferase enzyme family	TMpred: One transmembrane helix, N-terminus inside
103	2.30.42.10	43.1	Mainly Beta	Roll	Pdz3 Domain		NCBI-CDD: (275338) Heme b synthase	
104	3.10.180.10	10.6	Alpha Beta	Roll	2,3-Dihydroxybiphenyl 1,2-Dioxygenase; domain 1	2,3-Dihydroxybiphenyl 1,2-Dioxygenase, domain 1	NCBI-CDD: (307736) Glycosyltransferase family 25; Pfam: (Glyco_transf_25) Glycosyltransferase family 25	TMpred: One transmembrane helix, N-terminus outside
105	3.40.605.10	17.4	Alpha Beta	3-Layer (aba) Sandwich	Aldehyde Dehydrogenase; Chain A, domain 1	Aldehyde Dehydrogenase; Chain A, domain 1	NCBI-CDD: (280088) HNH endonuclease; Pfam: (HNH) HNH endonuclease	

a) For NCBI-CDD results, the numbers inside parenthesis represent the PSSM-id (Position-Specific Scoring Matrix).

prediction. In addition, a motif search detected a phage terminase-like large subunit motif and CATH analysis showed similarity with the homologous family of P-loop containing nucleotide triphosphate hydrolases (P-loop NTPase). This P-loop NTPase fold suggests that the product of Ms6 ORF2 can hydrolyze the beta-gamma phosphate bond of bound nucleoside triphosphate, like ATP, matching the ATP-binding properties described for the large subunit. BLASTp matches for Ms6 ORF1 were more diverse, however the first match was the gene product of mycobacteriophage Tweety *gp1*, which is annotated as coding for the small subunit of terminase. Taking together these results, and location adjacent to the large terminase, Ms6 ORF1 was annotated as encoding the small subunit of terminase.

The gene products of ORF1 and ORF2 also exhibited homology with sequences of hypothetical proteins from several members of the genus *Mycobacterium* (especially with *M. abscessus*), as well as with *Nocardia* or *Gordonia* species, among other bacteria.

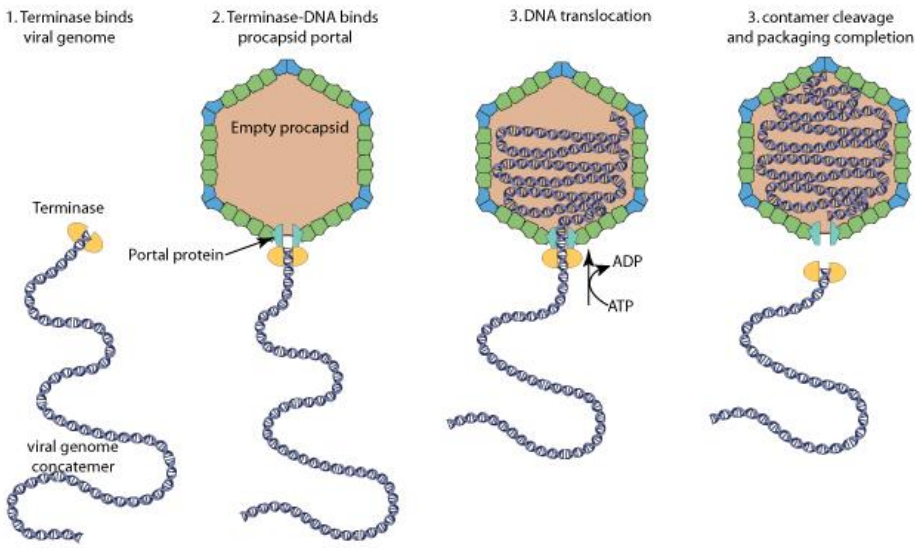


Figure 7 - Schematic representation of genome packaging in dsDNA phages (Swiss Institute of Bioinformtics, 2014).

4.2.2 Head Assembly and Head-to-Tail Connectors

The head assembly structural cluster of tailed dsDNA phages includes genes that encode several important proteins: the major capsid protein, whose hundreds of copies arrange in a regular pattern, resulting in an icosahedral capsid; the portal protein, which forms a ring complex that acts both as a gate for DNA entering or exiting the capsid and

as a binding site for gate-sealing proteins after DNA translocation; the scaffolding protein, responsible for overseeing a proper head assembly process; the protease, encoded by most phages, regulates capsid maturation by scaffold digestion or auto-inactivation (49). A scaffold protein gene is not detected in some mycobacteriophages, but its role may be played by a domain of the capsid subunit (50).

In Ms6 these functions are assigned to ORFs 3-6, which comprise the head assembly structural cluster. A BLASTp analysis for ORF3 showed the highest identity (99% identical) with the portal protein of mycobacteriophage PopTart gp3. An homology with a Pfam domain (pfam05133) of phage portal proteins (*Bacillus subtilis* phage SPP1 gp6-like) was identified for ORF3 and many other alignments support the notion that ORF3 is translated into Ms6's portal protein. ORF4 is predicted to code for the capsid maturation protease since the encoded protein shares a high identity with amino acid sequences designated as mycobacteriophage capsid maturation proteases, alongside a cathepsin L protease conserved domain (PTZ00203), thus reinforcing a proteolytic function for the product of ORF4. The major capsid protein of Ms6 appears to be encoded by ORF6, since it is 99% identical to 100% of phage Fruitloop major capsid protein sequence (gp6). ORF6 also bears a domain of a *Salmonella typhimurium* phage P22 coat-like protein (pfam11651), that extends for about three quarters of the amino acid sequence and which indicates a structural role for this gene product. Unlike the described ORFs, ORF5 does not present such evident structural motifs; however, ORF5 is 100% identical to the whole sequences of the scaffolding proteins of phages Tweety and SuperGrey, and is placed between the protease and the capsid genes, making it the prime candidate for the scaffolding protein in Ms6.

ORFs 7-10 are located immediately downstream the gene predicted to encode the major capsid protein. These genes encode four proteins, all with a similar size, molecular weight and a common characteristic: a high identity to several BLASTp entries annotated as head-to-tail connectors. ORF9 additionally owns a bacteriophage HK97 gp10-like conserved domain (pfam04883), representative of a putative *Escherichia coli* phage HK97 tail component. These ORFs fit into the set of four to eight genes that most mycobacteriophages present between the head assembly and the tail assembly clusters, which are believed to intervene in the head-tail attachment process (17).

4.2.3 Tail Assembly

Siphoviridae phages have a long non-contractile tail made of several copies of a major tail protein. In these phages, tail assembly occurs as a gradual process, with each player being added in a strict, sequential manner, thereby assuring proteins are attached into growing complexes and not lost on unproductive structures (51). In λ's tail assembly pathway (Figure 8), the tape measure protein establishes the length of the tail and interacts with two tail assembly chaperones. A tail initiator then arises from the interaction between this structure and the baseplate. The larger chaperone appears to assemble copies of the major tail protein, which in turn displace the chaperones and generate a tube. The terminator proteins cease the extension once the tail achieves an adequate length and the tape measure protein undergoes a maturation process, yielding a mature tail (52,53).

The major tail protein in Ms6 is probably encoded by ORF11, since this is highly identical to many other mycobacteriophage proteins with this function associated. BLASTp analysis of ORFs 12 and 13 suggest these are translated as a +1 programmed translational frameshift, an event that occurs due to a slippery ribosomal translation, producing two different proteins from a single nucleotide sequence (54). The location of ORFs 12 and 13 in the genome, their overlapping region with a slippery sequence and the absence of an evident ribosome binding site for ORF13 are typical aspects of these frameshifts (55). This is a widely preserved feature in tail genes of dsDNA phages and both the production and the ratio of the two proteins are important for efficient tail assembly (52). ORF14 is 3489 bp long, easily recognizable as the largest ORF in the genome and, therefore, the most likely to encode the tape measure protein. The presence of a tape measure domain (TIGR02675) and the homology with several members of this group of tail proteins (Phatniss gp15, SiSi gp14 and XFactor gp14) consolidate this assumption. The size of the tape measure protein is proporcional to the tail length, with each amino acid generally accounting for 0.15 nm in tail length (45). In the Ms6 case, the translation product of ORF14 is a 1162 amino acid protein, generating a predicted 174.3 nm tail, displayed as an α-helical structure, which is close to the real length of 188 nm (Gigante, personal communication).

According with the syntenically organization of most mycobacteriophage genomes, phages encode 5-10 minor tail proteins (17), which in Ms6 might be ORFs 15-19. All these ORFs possess substantial homology with several phage minor tail proteins and both ORFs 15 and 18 show phage tail-related motifs (pfam05709 and COG52839,

respectively). Interestingly, ORF17 presents a structural homology to *M. tuberculosis* H37Rv resuscitation-promoting factor (Rpf) B. This aspect is discussed later in this study.

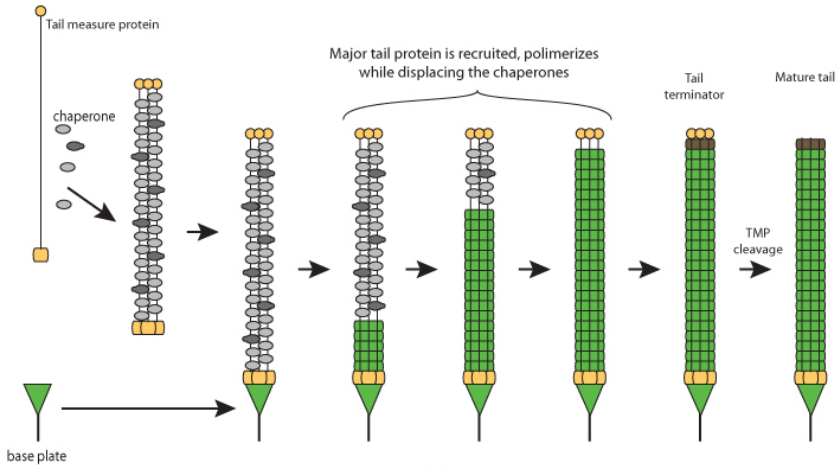


Figure 8 - Representation of the sequential tail assembly process of Lambda-like phages (Swiss Institute of Bioinformatics, 2014).

4.2.4 Lysis

The proteins encoded by the lytic cassette of Ms6 have been well characterized in several studies (56–61) and Ms6 can be regarded as a mycobacteriophage-induced lysis model, employing the holin-endolysin strategy enhanced by some extra features (62). The products of ORFs 23-27 correspond, respectively, to gp1-5 described in the mentioned studies.

ORF24 encodes the endolysin of Ms6, lysinA (LysA), an enzyme previously proposed by Catalão *et al*. (60) to target the peptidoglycan structure, hydrolyzing the amide bond between the N-acetylmuramic acid (NAM) and the oligopeptide L-alanine residues. A second potential translation site was identified by the group and this was detected in the endolysin nucleotide sequences of other bacteriophages, like *Lactococcus* phage ϕvML3 or *Streptococcus* phage C1. The authors also demonstrated that *lysA* gene translation produces two polypeptides, one with 384 amino acids ($Lysin_{384}$) and other with only 241 ($Lysin_{241}$), both necessary for an efficient host lysis. We find two noteworthy conserved domains in ORF24: a peptidoglycan recognition protein (PGRP) conserved domain (cd06583) and the Ami_2 domain (smart00644). The first recognizes bacteria cell wall PG, while the second is specifically related with the N-acetylmuramoyl-L-alanine amidase activity, which cuts the bond between N-acetylmuramic acid (NAM)

and the peptide chain. An amidase activity was already demonstrated for the Ms6 LysA (63). Endolysins may target other PG bonds and are also classified in *N*-acetyl-β-d-muramidases (generically also termed lysozymes) and lytic transglycosylases, *N*-acetyl-β-d-glucosaminidases and endopeptidases according to the specific bond they attack (64). The generic term "lysozymes" have been ambiguously used in the literature and public databases to designate enzymes that hydrolyze the PG. This explains why the CATH analysis suggests a structural similarity with the lysozyme homologous superfamily, while the Ami_2 domain (E value 6e-15) clearly correlates with the N-acetylmuramoyl-L-alanine amidase activity.

Located immediately upstream *lysA* is ORF23. The bioinformatic analysis revealed that all homologues of this ORF in other mycobacteriophages are hypothetical proteins. However, prior studies showed this ORF encodes a chaperone-like protein necessary for LysA delivery to the peptidoglycan (58), contrasting with the lambda model, in which active endolysin only reaches the peptidoglycan after a membrane destabilization is initiated by a canonical holin (65). In fact, gp23 has characteristics related to type III secretion systems chaperones and it homoligomerizes to interact with LysA; deletion of the associated gene hindered normal host lysis, stressing the advantage conferred by this feature (58). Some significant findings about this protein denoted by Catalão *et al.* (61) include a N-terminal $Lysin_{384}$ binding domain, a C-terminal homooligomerization domain and two motifs: a C-terminal GXXXG and N-terminal AXXXAXXXA. Whilst the first motif might be involved in the molecular stabilization within gp23, the second may secure intermolecular interactions between LysA and the chaperone-like protein (61).

ORF 25 (previous gp3) encodes LysB, a protein that was shown to have lipolytic activity (56) and that has an important role in lysis. Ms6 LysB was shown to cleave the ester bond that links the mycolic acids of the outer membrane to the arabinogalactan in the cell wall, releasing thus the lipid-rich outer membrane of mycobacteria at the end of a phage lytic cycle (57). ORF25 bears a PE-PPE (Pro-Glu and Pro-Pro-Glu) motif (PF08237) between amino acids 134-224 associated with cell surface features in mycobacteria (66) and a cutinase between amino acids 120-208. Similarly to cutinases, lipases and esterases, the pentapetide Gly-Tyr-Ser168-Gln-Gly is found, matching the typical Gly-X-Ser-X-Gly motif present in lipolytic enzymes (67). Recently, although unnoticed by conserved domains analysis, HHpred search tool detected a LysB N-

terminal homology with the peptidoglycan binding domain (PGBD) of *Pseudomonas aeruginosa* phage phiKZ endolysin (Probability 98.82, *E* value 1.5e-08).

ORF26 encodes a small sized protein (77 aa), with a low molecular weight (7.9 kDa), a hydrophobic nature and the presence of two potential TMDs. A motif for *Lactococcus lactis* phage r1t holin (pfam16945) is also present. Even though the TMD1 between residues 17 and 34 is composed mainly of weakly hydrophobic or polar uncharged residues, and therefore not detected by all TMD search tools, the gene product of ORF26 is recognized as a member of class II holins (59). TMD1 has caracteristics of a signal-arrest-relase domain and it was suggested to function as a pinholin, creating small holes that allow ion passage and subsequent membrane despolarization, opposed to the large holes formed by canonical holins (59,68). The product of the adjacent ORF27 contains a single TMD helix and it was demonstrated to act together with Gp26 as a holin, interacting and cooperating for a precise timing of lysis, as elimination of genes *26* or *27* led to, respectively, accelerated and delayed lysis phenotypes (59).

In Figure 9 is displayed a representation of the mycobacterial cell envelope, with the target layer of each phage lysis protein indicated with an arrow.

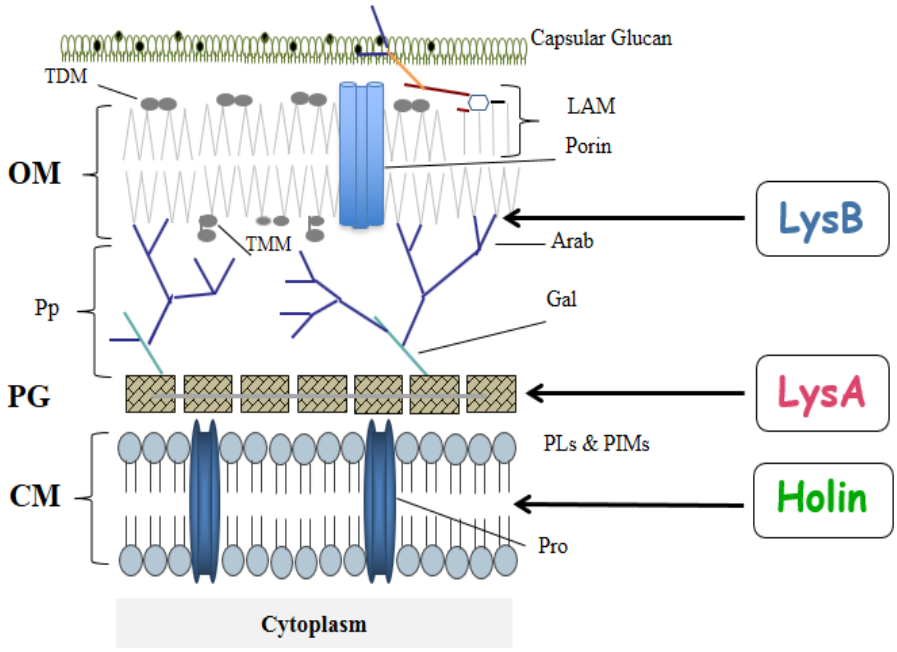


Figure 9 - Mycobacterial targets of Ms6 lysis proteins. Arab, arabinan; CM, cytoplasmic membrane; Gal, galactan; LAM, lipoarabinomannan; OM, outer membrane; Pro, protein; PG, peptidoglycan; PIMs, phosphatidylinositolmannosides; PLs, phospholipids; Pp, periplasm; TDM, trehalose dimycolate; TMM, trehalose monomycolate. (Pimentel, 2014).

4.2.5 Integration

Temperate phages integrate into the bacterial genome through site-specific recombination. This event which occurs between a phage attachment site (*attP*) and a bacterial attachment site (*attB*) (69), requires a phage-encoded recombinase which in Ms6 is encoded by ORF36.

The site-specific integration system of Ms6 was previously studied by Freitas-Vieira *et al.* (70), enabling the identification of both the integrase gene and the *attP* site. An integrative plasmid containing the integrase gene and the *attP* site was constructed and a nonsense mutation within the integrase gene entirely extinguished the integration capacity of the plasmid. The authors also showed that the C-terminus of the integrase gene was similar to other conserved C-terminal regions typical of the phage integrase family. Currently, a BLASTp analysis of the product of ORF36 identified several homologous proteins producing significant alignments with mycobacteriophage Mozy gp51 and mycobacteriophage Fruitloop gp40 showing the highest identity (99%). Additionally, the C-terminal catalytic domain of integrases from bacterial phages and conjugate transposons (cd01189) was detected on MOTIF.

The common core within the attP site is 26 bp in length and lies 98 bp upstream the 5' end of the integrase gene and in the chromosomal *attB* site it overlaps a $tRNA^{Ala}$ gene, indicating that the phage genome is inserted at the 3' end of this gene (70). The common core sequence is centrally located in the genome (bp 27552..27577), since it is situated at a distance from the left end that corresponds to 50.8% of the entire genome length.

4.2.6 DNA Modification

We detected several ORFs encoding putative DNA modification enzymes, most of which residing in the right genomic arm.

ORF29 displays a cd05160 domain (*E* value 0.025) between amino acid residues 83 and 201. This conserved domain reflects a connection with the DnaQ-like 3'-5' exonuclease superfamily, enzymes that promote the 3'-5' oriented excision of nucleoside monophosphates at nucleic acids termini. Between residues 221 and 259, we also find the Lar_restr_allev Pfam motif (PF14354) of protein Lar (*E* value 0.013), a Rac prophage product that was described to alleviate restriction and enhance modification by *E. coli* K-

12 restriction and modification system (71). Nonetheless, the high homology of ORF29 with several annotated mycobacteriophage DnaQ-like exonucleases, with over 90% identity for 100% query covers and *E* values of 0.0, supports an exonuclease function for this ORF.

Homing consists in the endonuclease-catalyzed lateral transfer of an intervening sequence to a homologous allele missing that sequence (72). Mycobacteriophages frequently have one or more putative HNH homing endonucleases and it is possible that these mobile elements contribute to genomic mosaicism (16). ORF59, ORF61 and ORF105 of Ms6 possess motifs of HNHc endonucleases and the closest homologues to each of these ORFs are all assigned with this function: ORF59 is 98% identical to gp22 of mycobacteriophage WIVsmall (*E* value 0.0); ORF61 has 98% identity with mycobacteriophage Quico's gp73 (*E* value 0.0); ORF105 is 99% identical to gp101 of mycobacteriophage Cabrinians (*E* value 0.0). These results sustain a putative HNH endonuclease role for these ORFs.

BLASTp analysis shows that ORF64 is 84% homologue to a methyltransferase of phage Hamulus (gp68) (*E* value 0.0). A methyltransferase function is consistent with the presence of a site-specific DNA methylase conserved domain (COG0270) (*E* value 1e-40) between amino acid 1 and 182. Besides ORF64, ORF69 is also predicted to encode a DNA methyltransferase due to significant BLASTp alignments with proteins annotated with this function. A DNA methylase motif (pfam01555) with an *E* value of 9e-19 was also detected. Additionally, PROSITE results suggest that ORF64 is a C-5 cytosine-specific DNA methylase, while ORF69 conveys a N-6 adenine-specific signature. The proteins specified by ORF64 and ORF69 likely protect newly replicated DNA from host restriction endonucleases, as it has been described for ORF18 of *Lactococcus lactis* phage Tuc2009, a methyltransferase (73).

ORF101 and ORF104 appear to encode proteins with a glycosyltransferase activity, potentially establishing glycosidic bonds. Between amino acids 25-108, ORF101 bears the PF00535 glycosyltransferase family 2 motif (*E* value 8.2e-5), a group of enzymes that catalyzes the transfer of sugar from uridine diphosphate-α-D-glucose (UDP-Glc) and other molecules, to a range of substrates. Interestingly, CATH predicted a structural similarity with the homologous superfamily of glycosidases between residues 361-475, a class of enzymes that hydrolyze glycosidic linkages. Albeit this apparent paradox, ORF101 is highly similar to many other phage glycosyltransferases. ORF104

contains a PF01755 motif of the glycosyltransferase family 25 (*E* value 6.7e-8). This family of enzymes assist lipopolysaccharide biosynthesis by promoting the transfer of various sugars onto the growing chain. Some of this ORF's closest homologues are also designated as phage glycosyltransferases. In general, phage-encoded glycosyltransferases are possibly involved in bacteria cell wall glycosylation and/or phage DNA modification, with subsequent protection from host restriction enzymes. This DNA modification strategy has been reported for the β-glycosyltransferase of phage T4, which transfers glucose UDP-Glc to 5-hydroxymethylcytosine residues of phage DNA (74).

4.2.7 Regulation

The genome bioinformatic analysis shows that several ORFs possibly encode transcriptional regulators. Transcriptional regulation is partly controlled by repressors. In phage lambda, CI and Cro repressors regulate the lysis-lysogeny decision as a bistable genetic switch: Cro stimulates lysis by progressively silencing early viral functions, while CI maintains lysogeny through a total shut down of lytic function expression (75).

In Ms6, Garcia *et al*. (63) hypothesized that ORF33, which is transcribed leftwards, could function as a possible repressor gene, given the homology between the encoded protein and the repressors of phages Bxb1 (gp69) and L5 (gp71), possibly two of the most well studied repressors in mycobacteriophages. Analysis by CATH identified a structural similarity with the tetracycline repressor domain 2. The tetracycline repressor forms a homodimer and interacts with DNA by two helix-turn-helix (HTH) motifs (76), however a search with MOTIF predicted a HTH domain (cd00569) with a low e-value (0.058). The first BLASTp hit for ORF33 is gp37 of mycobacteriophage Fruitloop. Fruitloop's gp37 may not be directly implicated in immunity regulation, since homologues are not present in all cluster F phages and the number of stoperator sites (genetic sites that prevent transcription elongation and potentiate the repressor's activity) throughout the genome is not as significant as in phage Bxb1 (16). Thus, it is possible that ORF33 may be involved in regulation, but not necessarily the immunity function.

ORFs 38 and 39 of Ms6 also display BLASTp hits annotated as mycobacteriophage immunity repressors or Cro proteins, respectively. Both ORFs contain HTH motifs, with ORF39 having more motifs of this nature and with more relevant *E* values than ORF38. Furthermore, CATH determined a structural similarity for

ORF39 with lambda repressor-like DNA-binding domains (%ID 59,2%, *E* value 2.3e-4), accentuating a possible connection between the product of this ORF and a repressor-like activity. However, without experimental data, we cannot assure with certainty if any of the described ORFs (ORF33, ORF38 and ORF39) have a role as immunity repressors.

One of the BLASTp matches for ORF42 is a DNA-binding protein of *M. abcessus* (identity 38%, *E* value 3e-6). HHpred also detected a HTH domain (probability 98.19%, *E* value 1.6e-7) in ORF42, which was not detected by MOTIF but that correlates with a putative transcription regulation role.

A WhiB-like transcriptional factor is predicted to be encoded by ORF52, as high homology (identity > 97%, *E* values 0.0) with other annotated phage WhiB factors was detected. This is strengthened by the identification of a WhiB Pfam motif, PF02467 (*E* value 3.9e-21), through MOTIF. In *M. tuberculosis*, the WhiB-like genes are involved in an array of processes, including cell division, stress sensing, pathogenesis and antibiotic resistance (77). Although WhiB family proteins are common in many mycobacteriophages, it is still unspecified if they regulate host or phage expression (17). The only exception is for the WhiB factor of phage TM4, reported to exert a dominant negative effect on the host's *whiB2* gene and to promote superinfection exclusion (78).

ORF54 is predicted to have structural similarities with the homologous superfamily of tetracycline repressor (domain 2) and the encoded protein contains HTH domains of LuxR-type (PS00622) and Hin-like (cd00569) proteins. All these proteins homodimerize and have DNA-binding properties (76,79,80), strengthening the idea that the product of ORF54 possibly regulates DNA transcription as a homodimer.

ORF65 encodes a protein with a distant relation to a PF01402 conserved domain of the CopG family ribbon-helix-helix (*E* value 0.045). CopG is a transcriptional repressor, whose dimers display topological similarities with the Arc repressor of phage P22 (81). A homology with this Arc repressor was also detected by HHpred (probability 98.96%, *E* value 4.5e-11). Unlike the HTH motif, the ribbon-helix-helix region of CopG is involved in oligomerization rather than DNA recognition (81), but an association between ORF65 and transcriptional regulation is still plausible.

The closest homologue of ORF94 is gp95 of mycobacteriophage Tweety, a protein with a HTH DNA binding domain. Besides pfam12728 HTH domain (*E* value 4e-5), motif search also exposed conserved domains of the transcriptional regulator families

SoxR (COG0789) (*E* value 9e-5) and MerR (TIGR02043) (*E* value 4e-4). These proteins are part of a large family of transcription factors that regulate oxidative stress (e.g. SoxR) and metal detoxification (e.g. MerR) as dual regulators, meaning they can both repress and activate gene transcription (82). Structurally, MerR transcriptional regulators resemble phage lambda excisionase (Xis) (83), but the distant location of ORF94 from the integrase-encoding gene does not support a Xis-like function for this ORF.

4.2.8 Other Interesting Features

4.2.8.1 Other potential properties of ORF16

As stated before, ORF16 appears to encode a minor tail protein. Interestingly, HHpred identified a homology for ORF16 with the protein family Siphovirus ReqiPepy6 Gp37-like (Probability 99.72%, *E* value 4.0e-17). This family includes many phage proteins from Siphoviruses, whose function is unknown but related to PF06605, prophage tail proteins that probably act as endopeptidases. This homology and the putative structural role of this feature, led us to suspect that the encoded protein might function as a virion-associated lysin (VAL). VALs are enzymes synthesized by phages, generally associated to an element of the virion, such as the tail, that allow phages to overcome the bacterial PG (84). Much like endolysins, VALs act as glycosidases, amidases or endopeptidases, but they only create a hole large enough to permit phage DNA injection into the cytoplasm, which is distinctive from the extensive PG degradation performed by endolysins during lysis (84). Although a peptidoglycan hydrolase domain was not detected, due to these similarities, one may suspect that the product of ORF16 may also have a function in PG hydrolysis, facilitating Ms6's infection, but experimental support is required to fully assert this function.

4.2.8.2 Structural homology between the gp17 of Ms6 and resuscitation-promoting factor B

M. tuberculosis encodes five resuscitation-promoting factors (RpfA-E) involved in this pathogen's transition from dormant to active state (85). The muralytic properties of Rpfs are responsible for the biological activity of these proteins (86). One Rpf-like

protein motif was identified in the tape measure protein of mycobacteriophage Barnyard and two other predicted peptidoglycan-hydrolyzing motifs are detectable in the tape measure proteins of other phages (87). In Ms6, structural homology with the homologous superfamily of Rpf B (lytic transglycosylase) of *M. tuberculosis* H37Rv was encountered by CATH, not for the more expected tape measure gene, but for ORF17, which was predicted to code for a minor tail protein. Nonetheless, it is reasonable a Rpf-like activity in a putative minor tail protein may exert the same influence as one associated with tape measure proteins. In a broader sense, since most cells in their natural environment are in a latent state, encoding a Rpf-like protein would grant phages an advantage by overcoming potential growth phase-dependent structural alterations in host PG layer (11,88).

4.2.8.3 Mycobacteriophage mosaicism is patent in ORFs 18-22 of Ms6

Although Ms6 is highly similar to other F1 phages, 11 ORFs have a first BLAST hit that is unrelated with the members of this subcluster. Two of these correspond to genes from phages of subcluster I1, other two to phages of cluster N, one gene is analogous to a singleton's gene and four are related with cluster/subcluster C1, E, F3 or J. No function has been attributed to these genes in Ms6, except for ORF 18, which is homologous to gene 20, coding for a predicted minor tail protein of mycobacteriophage Redi, a member of the cluster N. ORFs 45 and 46 resemble genes from *M. abscessus* and *Acinetobacter baumannii*, respectively, for which no fuction is assigned. This is not unexpected, since phages can acquire genes from their hosts, especially temperate phages due to their integration into the bacterial genome.

The whole genome BLASTn analysis of Ms6 reveals the existence of a region that is not as similar with Dlane, Shauna1 or Fruitloop's sequences as the rest of the genome. This segment is located between bp 17661 and bp 21713, has 4053 bp of length and roughly corresponds to ORFs 18-22 of Ms6. This fragment is not found in the phages mentioned above, but was detected, albeit not complete, in other phages from cluster F1, like Brocalys and Velveteen, which have a lower overall similarity with the genome of Ms6. Interestingly, this region is 95% conserved in phages Redi and Phancyphin from cluster N, reflecting genome mosaicism. A DotPlotter representation (a) and a BLASTn comparison (b) of phages Ms6, Dlane, Shauna1, Brocalys and Redi is provided in Figure 10 and strongly emphasizes the described mosaicism.

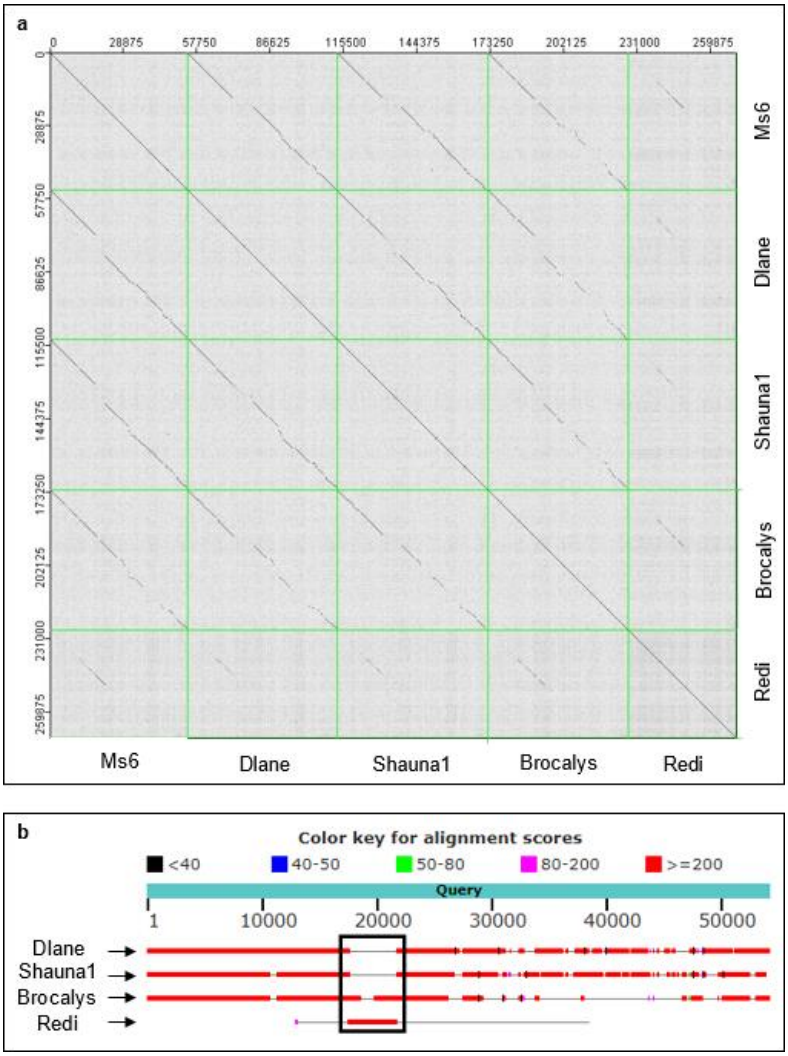


Figure 10 - a) Whole-genome dot plot of phages Ms6, Dlane, Shauna1, Brocalys and Redi; dot plots were constructed in JDotter (89), available at https://virology.uvic.ca, with a sliding window of 50 bp. b) BLASTn alignment of the genome of Ms6 with the phages mentioned; the black square demarcates a clear exemple of genetic mosaicism.

4.2.8.4 The Pin protein, codified by ORF37, confers resistance to infection by Ms6 to a mutant *M. smegmatis* strain

ORF37 encodes a product that was named Pin and explored by Pimentel (90), who demonstrated that a mutant strain expressing Pin (*M. smegmatis* mc^2155 13B) is resistant to infection by Ms6. However, the transfection of phage DNA into this strain resulted in

regular plaque formation, exposing that the resistance mechanism operates in the early steps of infection, possibly at the adsorption or DNA injection levels (90).

4.2.8.5 Multiple tetrapeptide repeats in the central portion of gp48

ORF48 contains a central core of very high mol% GC content (Figure 11, panel a), which corresponds to a roughly 200 bp region. Within this region, 16 repeat units of 12 bp are detected, where the first 6 nucleotides (GCCGCA) are conserved, as well as the ninth and twelfth nucleotide (G and C, respectively). A greater variation is noticed for nucleotides in positions seven (8 Gs, 7 Ts and 1 C), eight (11 Gs and 1 A), ten (10 As and 2 Ts) and eleven (10 Gs and 2 As), which are first and second codon positions. As a consequence, the encoded tetrapeptides always contain alanine as the first and second residues, glycine (8 times), tryptophan (7 times) or glutamine (once) at the third amino acid position and serine (10 times) or tyrosine (twice) as the fourth residue. The amino acid sequence of Ms6 gp48 is depicted in panel b of Figure 11, with the 16 tetrapeptides colored according to the four different possibilities (AAGS, AAWS, AAGY and AAQS).

This pattern of multiple tetrapeptide repeats was already described by Pham *et al*. (45) for mycobacteriophage Tweety's gp54, which contains 48 tetrapeptide repeats with the same amino acid profile found in Ms6 tetrapeptide sequence. The authors refer that the number of tetrapeptide units vary among phages and that these tetrapeptide repeats may also be absent in some phages. In Ms6, sequence similarity of the product of ORF 48 with related phage proteins (gp52 of XFactor, gp53 of Shauna1 and gp57 of Bobi) is much higher in the N- and C- termini, as the result of the variation of the number of tetrapeptide repeats in the central portion of the proteins. An alignment of the central region of Ms6 gp48 with the stated gene products is found in Figure 11, panel c. Like the product of ORF 48 of Ms6, XFactor gp52 has 16 repeats, while Shauna gp53 and Bobi gp57 both have 17 repeats and Tweety gp54 remarkably has 48 repeats. Notably, a similar tetrapeptide repeated sequence was not detected for phage Dlane, Ms6's closest relative.

While structures related to these gene products have been reported for other mycobacteriophages and *Bordetella* phages, hinting that these repeats are possibly widely spread among phage populations, the function of these features and the rationale behind their diversity remains unknown (45).

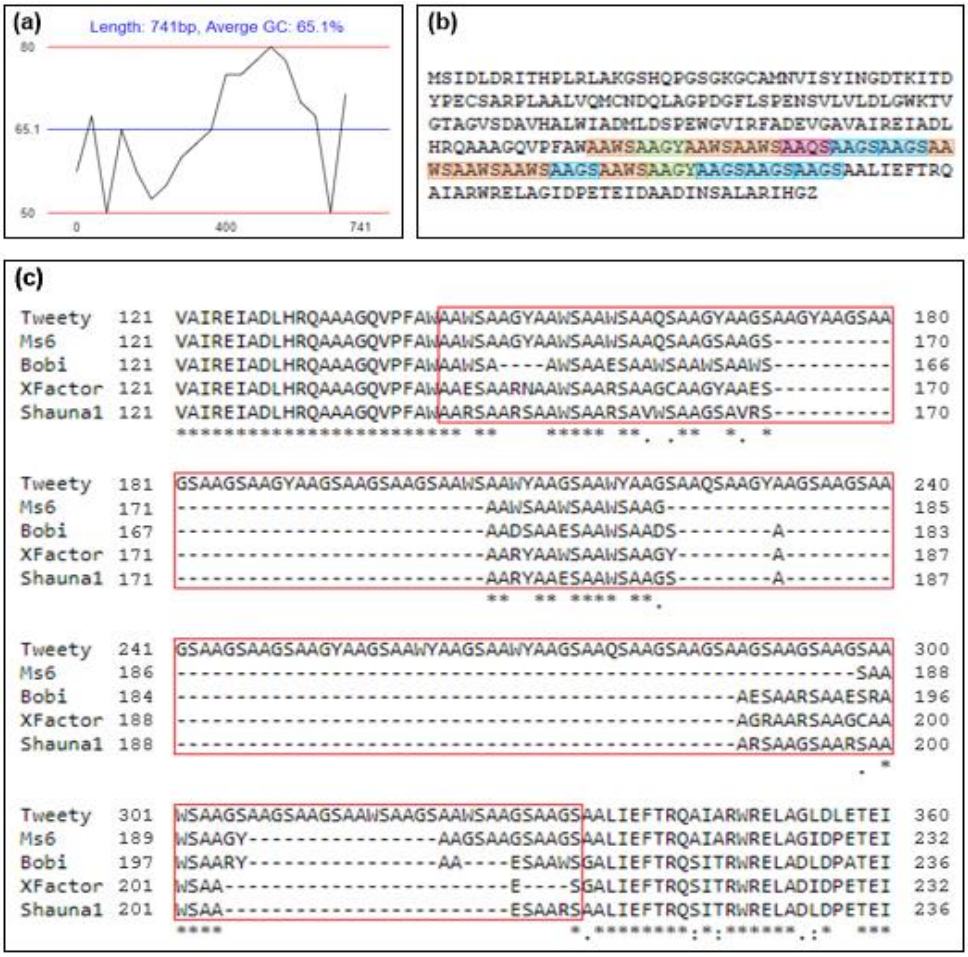


Figure 11 - (a) Plot of mol% G + C for ORF48 of Ms6. The average GC content of Ms6 genome is 61,5%, but for this region is 65,1%. The plot was obtained with DRAW GC, by ENDMEMO, available at http://www.endmemo.com/index.php. (b) Sequence of the product of OF 48 of Ms6 with the multiple tetrapeptide repeats pointed out. Four types of tetrapeptides are present: AAGS (7 times), AAWS (6 times), AAGY (twice) and AAQS (once). (c) Alignment of region of Tweety gp54, Ms6 gp48, Bobi gp57, XFactor gp52 and Shauna gp53. The red squares mark multiple tetrapeptide repeats across the different gene products of these phages. Asterisks show amino acid identities, colons represent conserved substitutions and periods show semi-conserved substitutions. The alignment was performed with Clustal Omega at EMBL-EBI, available at *https://www.ebi.ac.uk/services*.

4.2.8.6 ORF60 appears to encode a Sak3-like protein

Analysis of ORF60 with HHpred search tool detected a homology with the protein Sak3 from *Lactococcus* phage p2 (probability 99.8, *E* value 1.9e-21). Sak3 is a single-

strand annealing protein (SSAP) that stimulates homologous recombination and is also involved in the antiviral AbiK mechanism, since mutations in the Sak3-coding gene can reduce sensitivity to this abortive infection (Abi) strategy (91). One of the best-characterized SSAPs is the eukaryotic Rad52, with a major role in recombination and DNA repair (92). Although the E value is unreliable, Motif search for ORF60 identified a conserved domain of Rad52/22 protein family (pfam04098), further complementing the HHpred finding. In addition, CATH predicted an ATP-grasp fold for ORF60, suggesting the encoded protein interacts with this molecule, which in turn is consistent with the ATPase-dependent activity of Sak3 (91). Based on the structural homology and the presence of relevant motifs, we conceive that the product of ORF60 participates in equivalent biological events to the ones Sak3 intervenes.

4.2.8.7 A typical phage serine-threonine kinase is associated with ORF102

ORF102 is closely related with gene *100* of phage Fruitloop (99% Identity, E value 0.0), that encodes a putative serine-threonine kinase (Ser/Thr kinase) (16). In fact, CATH predicted a structural similarity for ORF102 with the superfamily of transferase (phosphotransferase) domain 1, which includes Ser/Thr kinases. The APH phosphotransferase motif (PF01636) was also found between amino acids 68-144 (*E* value 1.5e-7). APH phosphotransferases are a class of bacterial antibiotic resistance proteins, which confer resistance to various aminoglycosides through antibiotic phosphorylation (93). The Ser/Thr protein kinase (gp0.7) of phage T7 is known to participate in host transcription shutoff to favor virus replication (94). With the available data it is not possible to assign a fuction for this ORF.

5 Conclusion

Mycobacteriophage Ms6 was isolated in 1989, from a culture of *M. smegmatis* HB5688 at the Faculty of Pharmacy, University of Lisbon. Over the years, Ms6 research contributed to a better comprehension of phage biology, namely through studies on the lytic operon, giving this phage a special role as model of mycobacteriophage-mediated lysis. Other genetic elements of Ms6 have also been explored, such as the ones involved in the integration functions, but most of the genome remained uncharted. Therefore, a complete genome annotation was lacking and was essential to fully grasp Ms6's characteristics and potential.

In this study, after sequencing of the phage genome, a series of bioinformatic analysis was performed and the results carefully interpreted, generating a complete and integrated genome annotation of phage Ms6.

The genome of Ms6 is composed of a double stranded DNA molecule, with 54252 bp and a GC content of 61.5%. High homology with several subcluster F1 phages sustains the inclusion of Ms6 in this group, with its closest relative, phage Dlane, being 98% identical over 81% of the entire genome length. In addition, about 10% of Ms6's first BLASTp hits are proteins encoded by phages included in clusters or subclusters other than F1. These findings are consistent with the notion that subcluster F1 is among the most diverse sets of phages.

105 ORFs were identified and 43 (≈41%) putative functions were assigned by a conjunction of several parameters, such as protein homology, presence of motifs and conserved domains, location in genome or contiguity to certain genes. These 43 features were assorted into seven distinct sections: ORFs 1 and 2 are involved in DNA packaging; ORFs 3-6 are responsible for the head assembly, while ORFs 7-10 may act as head-to-tail connectors; the tail assembly cluster is possibly composed of ORFs 11 to 19 and contains a typical +1 transcriptional frameshift associated to the tail assembly chaperones; the lytic functions are attributed to ORFs 23-27; roughly in the middle of the sequence lies ORF36, the *int* gene, which is followed by *pin*, a reverse transcribed gene associated with a superinfection exclusion mechanism; the remaining ORFs are mostly located in the right genomic arm and presumably have a role in either DNA modification or phage regulation.

Some notable findings concern ORF16, which perhaps has a VAL-like activity associated to its apparent structural role as a minor tail protein, or ORF60, that encodes a Sak3-like protein, possibly promoting homologous recombination and control over host abortive infection strategies. Another intriguing feature is the 16 multiple tetrapeptide repeats found within gp48, to which no function is yet linked.

Genome annotation of Ms6 resulted in a better understanding of this phage potential characteristics, but the function of many genes still remains a mystery. In a broader scale, this is also valid for many other phage genomes, as well as for bacteria or other organisms since experimental data is necessary to support bioinformatic previsions. Therefore, genome annotation must be a continuous process, incorporating tomorrow's findings into yesterday's predictions, which in turn will avoid propagation of poorly-supported deductions and contribute to a more precise annotation of the ensuing genomes.

Bibliography

1. Cisek AA, Dąbrowska I, Gregorczyk KP, Wyżewski Z. Phage Therapy in Bacterial Infections Treatment: One Hundred Years After the Discovery of Bacteriophages. Curr Microbiol. 2017;74(2):277–83.

2. Summers WC. Bacteriophage Therapy. Annu Rev Microbiol. 2001;(55):437–51.

3. Clokie MRJ, Millard AD, Letarov A V, Heaphy S. Phages in nature. Bacteriophage. 2011;1(1):31–45.

4. Carter JB, Saunders VA. Viruses and their importance. In: Carter JB, Saunders VA, editors. Virology: Principles and Applications. 1st ed. West Sussex: John Wiley & Sons Ltd; 2007. p. 1–7.

5. O'Flaherty S, Ross RP, Coffey A. Bacteriophage and their lysins for elimination of infectious bacteria: Review article. FEMS Microbiol Rev. 2009;33(4):801–19.

6. Drulis-Kawa Z, Majkowska-Skrobek G, Maciejewska B. Bacteriophages and phage-derived proteins--application approaches. Curr Med Chem. 2015;22(14):1757–73.

7. Loc-Carrillo C, Abedon ST. Pros and cons of phage therapy. Bacteriophage. 2011;1(2):111–4.

8. Domingo-Calap P, Georgel P, Bahram S. Back to the future: bacteriophages as promising therapeutic tools. HLA. 2016;87(3):133–40.

9. Ackermann HW. 5500 Phages examined in the electron microscope. Arch Virol. 2007;152(2):227–43.

10. Hendrix RW. Phage evolution. In: Abedon ST, editor. Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses. 1st ed. Cambridge: Cambridge University Press; 2008. p. 177–94.

11. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, et al. Origins of highly mosaic mycobacteriophage genomes. Cell. 2003;113(2):171–82.

12. Oppenheim AB, Kobiler O, Stavans J, Court DL, Adhya S. Switches in Bacteriophage Lambda Development. Annu Rev Genet. 2005;39(1):409–29.

13. Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I, Herskovits AA. A new

perspective on lysogeny: prophages as active regulatory switches of bacteria. Nat Rev Microbiol. 2015;13(10):641–50.

14. The Actinobacteriophage Database at PhagesDB.org [Internet]. [cited 2017 May 15]. Available from: http://phagesdb.org/

15. Hatfull GF, Cresawn SG, Hendrix RW. Comparative genomics of the mycobacteriophages: Insights into bacteriophage evolution. Res Microbiol. 2008;159(5):332–9.

16. Hatfull GF. The Secret Lives of Mycobacteriophages. 1st ed. Vol. 82, Advances in Virus Research. Elsevier Inc.; 2012. p. 179-288.

17. Hatfull GF. Molecular Genetics of Mycobacteriophages. Microbiol Spectr. 2014;2(2):1–36.

18. Jacobs-sera D, Bowman CA, Pope WH, Russell DA, Cresawn SG, Hatfull GF. Annotation and Bioinformatic Analysis of Bacteriophage Genomes: A User Guide to DNA Master. 2014.

19. Jacobs-sera D, Marinelli L, Bowman C, Broussard G, Guerrero C, Boyle M, et al. On the nature of mycobacteriophage diversity and host preference. Virology. 2012;434(2):187–201.

20. Cook GM, Berney M, Gebhard S, Heinemann M, Cox RA, Danilchanka O, et al. Physiology of mycobacteria. Adv Microb Physiol. 2013;2911(9):81–319.

21. Jarlier V, Hikaido H. Mycobacterial cell wall: Structure and role in natural resistance to antibiotics. FEMS Microbiol Lett. 1994;123(1–2):11–8.

22. Hett EC, Rubin EJ. Bacterial Growth and Cell Division: a Mycobacterial Perspective. Microbiol Mol Biol Rev. 2008;72(1):126–56.

23. Gardner GM, Weiser R. A Bacteriophage for *Mycobacterium smegmatis*. Proc Soc Exp Biol Med. 1947;66(1):205.

24. Froman S, Will DW, Bogen E. Bacteriophage active against virulent *Mycobacterium tuberculosis*. I. Isolation and activity. Am J Public Health. 1954;44(10):1326–33.

25. Stein L. Genome Annotation: From Sequence To Biology. Nat Rev Genet. 2001;2(7):493–503.

26. Koonin EV, Galperin MY. Genome Annotation and Analysis. In: Koonin EV, Galperin MY, editors. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. 1st ed. Boston: Kluwer Academic; 2003.

27. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13(5):329–42.

28. Krylov VN. Bacteriophages of *Pseudomonas aeruginosa*: Long-term prospects for use in phage therapy. In: Maramorosch K, Murphy FA, editors. Advances in Virus Research. First edit. Elsevier Inc.; 2014. p. 227–78.

29. Cresawn SG, Bogel M, Day N, Jacobs-Sera D, Hendrix RW, Hatfull GF. Phamerator: a bioinformatic tool for comparative bacteriophage genomics. BMC Bioinformatics. 2011;12(1):395.

30. Hatfull GF, Pedulla ML, Jacobs-Sera D, Cichon PM, Foley A, Ford ME, et al. Exploring the mycobacteriophage metaproteome: Phage genomics as an educational platform. PLoS Genet. 2006;2(6):0835–47.

31. Portugal I, Anes E, Moniz-Pereira J. Temperate mycobacteriophage from *M. smegmatis*. Acta Leprol. 1989;7(1):243–4.

32. Sambrook J, Russell D. Molecular Cloning: A Laboratory Manual. 3rd ed. Sambrook J, Russell D, editors. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2001.

33. Delcher A, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 1999;27(23):4636–41.

34. Lukashin A V., Borodovsky M. GeneMark.hmm: New solutions for gene finding. Nucleic Acids Res. 1998;26(4):1107–15.

35. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005;33(SUPPL. 2):244–8.

36. Alva V, Nam S-Z, Söding J, Lupas AN. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. Nucleic Acids Res. 2016;44(W1):W410–5.

37. Orengo C, Michie A, Jones S, Jones D, Swindells M, Thornton J. CATH – a

hierarchic classification of protein domain structures. Structure. 1997;5(8):1093–109.

38. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. Edited by F. Cohen. J Mol Biol. 2001;305(3):567–80.

39. Hofmann K, Stoffel W. TMbase - A database of membrane spanning proteins segments. Biol Chem. 1993;374(166).

40. 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.

41. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. 2016;44(W1):W54–7.

42. Sigrist CJA, De Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucleic Acids Res. 2013;41(D1):344–7.

43. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: The protein families database. Nucleic Acids Res. 2014;42(D1):222–30.

44. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, et al. CDD: Conserved domains and protein three-dimensional structure. Nucleic Acids Res. 2013;41(D1):348–52.

45. Pham TT, Jacobs-sera D, Pedulla ML, Hendrix RW, Hatfull GF. Comparative genomic analysis of mycobacteriophage Tweety : evolutionary insights and construction of compatible site-specific integration vectors for mycobacteria. Microbiology. 2007;153(Pt 8):2711–23.

46. Catalano CE, Cue D. MicroReview Virus DNA packaging : the strategy used by phage lambda. Mol Microbiol. 1995;16(6):1075–86.

47. Fujisawa H, Morita M. Phage DNA packaging. Genes Cells. 1997;2(9):537–45.

48. ViralZone, Swiss Institute of Bioinformatics. Viral genome packaging [Internet]. 2014 [cited 2017 May 12]. Available from: https://viralzone.expasy.org/3944

49. Huet A, Duda RL, Hendrix RW, Boulanger P, Conway JF. Correct Assembly of the Bacteriophage T5 Procapsid Requires Both the Maturation Protease and the Portal Complex. J Mol Biol. 2016;428(1):165–81.

50. Hatfull GF. Mycobacteriophages : Genes and Genomes. Annu Rev Microbiol. 2010;64:331–56.

51. Aksyuk AA, Rossmann MG. Bacteriophage Assembly. Viruses. 2011;3(3):172–203.

52. Xu J, Hendrix RW, Duda RL. Chaperone – Protein Interactions That Mediate Assembly of the Bacteriophage Lambda Tail to the Correct Length. J Mol Biol. 2013;426(5):1004–18.

53. ViralZone, Swiss Institute of Bioinformatics. Viral tail assembly [Internet]. 2014 [cited 2017 Aug 20]. Available from: http://viralzone.expasy.org/3955?outline=all_by_species

54. Farabaugh PJ. Programmed Translational Frameshifting. Microbiol Rev. 1996;60(1):103–34.

55. Henry M, O'Sullivan O, Sleator RD, Coffey A, Ross RP, McAuliffe O, et al. In silico analysis of Ardmore, a novel mycobacteriophage isolated from soil. Gene. 2010;453(1–2):9–23.

56. Gil F, Catalão MJ, Moniz-Pereira J, Leandro P, McNeil M, Pimentel M. The lytic cassette of mycobacteriophage Ms6 encodes an enzyme with lipolytic activity. Microbiology. 2008;154(5):1364–71.

57. Gil F, Grzegorzewicz AE, Catalão MJ, Vital J, McNeil MR, Pimentel M. Mycobacteriophage Ms6 LysB specifically targets the outer membrane of *Mycobacterium smegmatis*. Microbiology. 2010;156(5):1497–504.

58. Catalão MJ, Gil F, Moniz-Pereira J, Pimentel M. The mycobacteriophage Ms6 encodes a chaperone-like protein involved in the endolysin delivery to the peptidoglycan. Mol Microbiol. 2010;77(3):672–86.

59. Catalão MJ, Gil F, Moniz-Pereira J, Pimentel M. Functional analysis of the Holin-Like proteins of mycobacteriophage Ms6. J Bacteriol. 2011;193(11):2793–803.

60. Catalão MJ, Milho C, Gil F, Moniz-Pereira J, Pimentel M. A second endolysin gene is fully embedded in-frame with the lysA gene of mycobacteriophage Ms6. PLoS One. 2011;6(6):e20515.

61. Catalão MJ, Gil F, Moniz-Pereira J, Pimentel M. The endolysin-binding domain

encompasses the N-terminal region of the mycobacteriophage Ms6 Gp1 chaperone. J Bacteriol. 2011;193(18):5002–6.

62. Pimentel M. Genetics of phage lysis. Microbiol Spectr. 2014;2(1):1–13.

63. Garcia M, Pimentel M, Moniz-Pereira J. Expression of mycobacteriophage Ms6 lysis genes is driven by two σ70-like promoters and is dependent on a transcription termination signal present in the leader RNA. J Bacteriol. 2002;184(11):3034–43.

64. Loessner MJ. Bacteriophage endolysins - Current state of research and applications. Curr Opin Microbiol. 2005;8(4):480–7.

65. Wang I-N, Smith DL, Young R. Holins: The Protein Clocks of Bacteriophage Infections. Annu Rev Microbiol. 2000;54(1):799–825.

66. Adindla S, Guruprasad L. Sequence analysis corresponding to the PPE and PE proteins in *Mycobacterium tuberculosis* and other genomes. J Biosci. 2003;28(2):169–79.

67. Pio TF, Macedo GA. Chapter 4 Cutinases:. Properties and Industrial Applications. 1st ed. Vol. 66, Advances in Applied Microbiology. Elesvier Inc.; 2009. p. 77-95.

68. Park T, Struck DK, Dankenbring CA, Young R. The pinholin of lambdoid phage 21: Control of lysis by membrane depolarization. J Bacteriol. 2007;189(24):9135–9.

69. Peña CE, Lee MH, Pedulla ML, Hatfull GF. Characterization of the mycobacteriophage L5 attachment site, attP. J Mol Biol. 1997;266(1):76–92.

70. Freitas-vieira A, Anes E, Moniz-Pereira J. The site-specific recombination locus of mycobacteriophage Ms6 determines DNA integration at the $tRNA^{Ala}$ gene of *Mycobacterium* spp. Microbiology. 1998;144(Pt 12):3397–406.

71. King G, Murray NE. Restriction alleviation and modification enhancement by the Rac prophage of *Escherichia coli* K-12. Mol Microbiol. 1995;16(4):769–77.

72. Chevalier BS, Stoddard BL. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. Nucleic Acids Res. 2001;29(18):3757–74.

73. McGrath S, Seegers JFML, Fitzgerald GF, Van Sinderen D. Molecular characterization of a phage-encoded resistance system in *Lactococcus lactis*. Appl

Environ Microbiol. 1999;65(5):1891–9.

74. Ünligil UM, Rini JM. Glycosyltransferase structure and mechanism. Curr Opin Struct Biol. 2000;10(5):510–7.

75. Takeda Y, Folkmanis A, Echols H. Cro Regulatory by Bacteriophage λ. J Biol Chem. 1977;252(17):6177–83.

76. Kisker C, Hinrichs W, Tovar K, Hillen W, Saenger W. The complex formed between Tet repressor and tetracycline-Mg^{2+} reveals mechanism of antibiotic resistance. J Mol Biol. 1995;247(2):260–80.

77. Alam MS, Garg SK, Agrawal P. Studies on structural and functional divergence among seven WhiB proteins of *Mycobacterium tuberculosis* H37Rv. FEBS J. 2009;276(1):76–93.

78. Rybniker J, Nowag A, Van Gumpel E, Nissen N, Robinson N, Plum G, et al. Insights into the function of the WhiB-like protein of mycobacteriophage TM4 - A transcriptional inhibitor of WhiB2. Mol Microbiol. 2010;77(3):642–57.

79. Li Z, Nair SK. Quorum sensing: How bacteria can coordinate activity and synchronize their response to external signals? Protein Sci. 2012;21(10):1403–17.

80. Lee SY, Lee HJ, Lee H, Kim S, Cho EH, Lim HM. In vivo assay of protein-protein interactions in Hin-mediated DNA inversion. J Bacteriol. 1998;180(22):5954–60.

81. Xavier Gomis-Rüth F, Solà M, Acebo P, Párraga A, Guasch A, Eritja R, et al. The structure of plasmid-encoded transcriptional repressor CopG unliganded and bound to its operator. EMBO J. 1998;17(24):7404–15.

82. Permina EA, Kazakov AE, Kalinina O V, Gelfand MS. Comparative genomics of regulation of heavy metal resistance in Eubacteria. BMC Microbiol. 2006;6:49.

83. Sam MD, Cascio D, Johnson RC, Clubb RT. Crystal structure of the excisionase-DNA complex from bacteriophage lambda. J Mol Biol. 2004;338(2):229–40.

84. Latka A, Maciejewska B, Majkowska-Skrobek G, Briers Y, Drulis-Kawa Z. Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers during the infection process. Appl Microbiol Biotechnol. 2017;101(8):3103–19.

85. Kana BD, Gordhan BG, Downing KJ, Sung N, Vostroktunova G, Machowski EE,

et al. The resuscitation-promoting factors of *Mycobacterium tuberculosis* are required for virulence and resuscitation from dormancy but are collectively dispensable for growth in vitro. Mol Microbiol. 2008;67(3):672–84.

86. Mukamolova G V, Murzin AG, Salina EG, Demina GR, Kell DB, Kaprelyants AS, et al. Muralytic activity of *Micrococcus luteus* Rpf and its relationship to physiological activity in promoting bacterial growth and resuscitation. Mol Microbiol. 2006;59(1):84–98.

87. Piuri M, Hatfull GF. A peptidoglycan hydrolase motif within the mycobacteriophage TM4 tape measure protein promotes efficient infection of stationary phase cells. Mol Microbiol. 2006;62(6):1569–85.

88. Kana BD, Mizrahi V. Resuscitation-promoting factors as lytic enzymes for bacterial growth and signaling. FEMS Immunol Med Microbiol. 2010;58(1):39–50.

89. Brodie R, Roper RL, Upton C. JDotter: A Java interface to multiple dotplots generated by dotter. Bioinformatics. 2004;20(2):279–81.

90. Pimentel M. Biologia Molecular da Resistência à Superinfecção por Bacteriófagos: Estudo do Gene pin do Micobacteriófago Ms6. University of Lisbon; 1999.

91. Scaltriti E, Launay H, Genois MM, Bron P, Rivetti C, Grolli S, et al. Lactococcal phage p2 ORF35-Sak3 is an ATPase involved in DNA recombination and AbiK mechanism. Mol Microbiol. 2011;80(1):102–16.

92. Singleton MR, Wentzell LM, Liu Y, West SC, Wigley DB. Structure of the single-strand annealing domain of human RAD52 protein. Proc Natl Acad Sci U S A. 2002;99(21):13492–7.

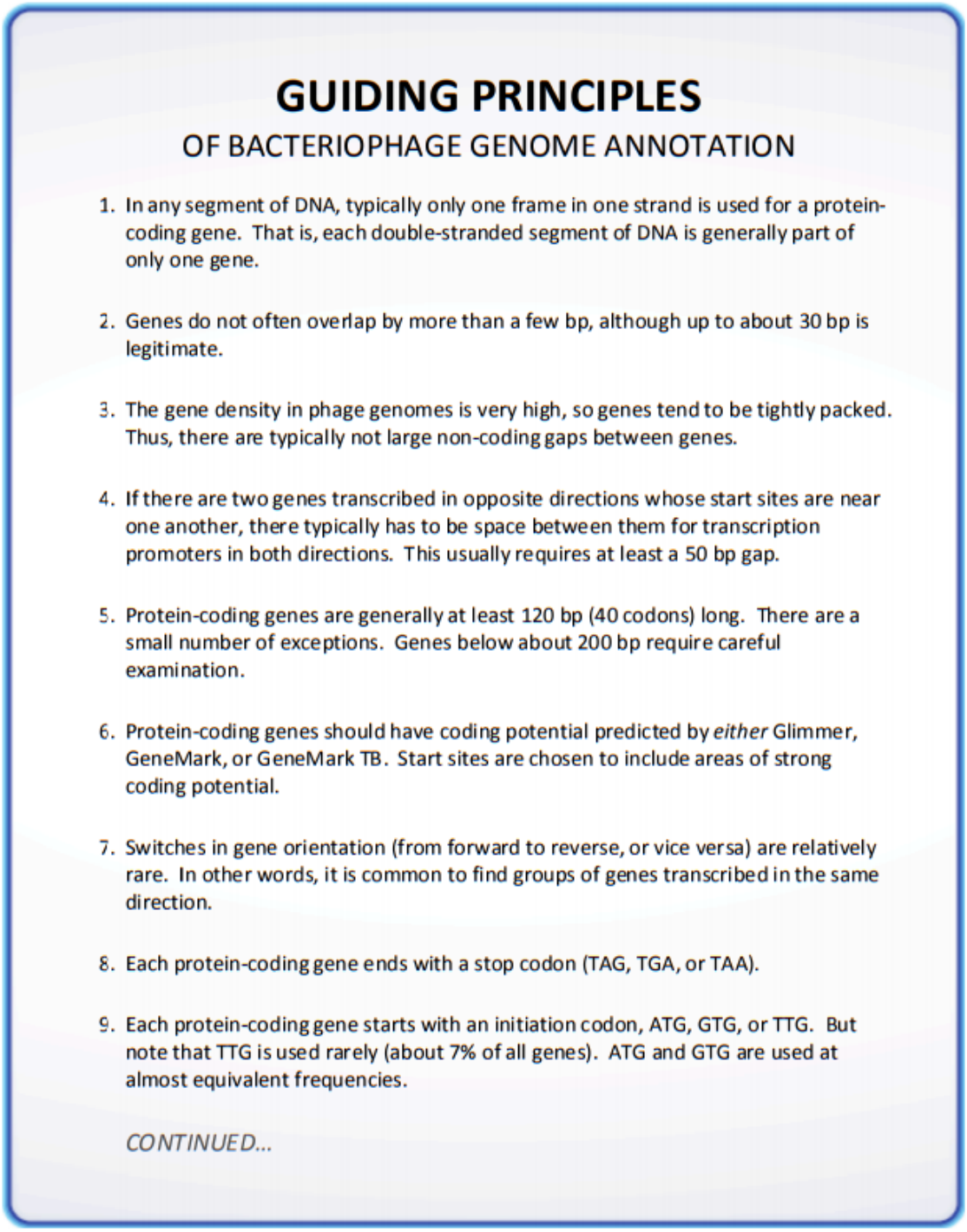
93. Trower MK, Clark KG. PCR cloning of a streptomycin phosphotransferase (aphE) gene from *Streptomyces griseus* ATCC 12475. Nucleic Acids Res. 1990;18(15):4615.

94. Severinova E, Severinov K. Localization of the *Escherichia coli* RNA Phosphorylated by Bacteriophage T7 Polymerase β′ Subunit Residue Kinase Gp0.7. J Bacteriol. 2006;188(10):3470–6.

Appendix

A1. Guiding Principles of Bacteriophage Genome Annotation

The following images contain the guiding principles of bacteriophage genome annotation, followed during Ms6 analysis. These principles are found on "Annotation and Bioinformatic Analysis of Bacteriophage Genomes: A User Guide to DNA Master" by Jacobs-Sera *et al.*, 2014.



GUIDING PRINCIPLES

OF BACTERIOPHAGE GENOME ANNOTATION

1. In any segment of DNA, typically only one frame in one strand is used for a protein-coding gene. That is, each double-stranded segment of DNA is generally part of only one gene.
2. Genes do not often overlap by more than a few bp, although up to about 30 bp is legitimate.
3. The gene density in phage genomes is very high, so genes tend to be tightly packed. Thus, there are typically not large non-coding gaps between genes.
4. If there are two genes transcribed in opposite directions whose start sites are near one another, there typically has to be space between them for transcription promoters in both directions. This usually requires at least a 50 bp gap.
5. Protein-coding genes are generally at least 120 bp (40 codons) long. There are a small number of exceptions. Genes below about 200 bp require careful examination.
6. Protein-coding genes should have coding potential predicted by *either* Glimmer, GeneMark, or GeneMark TB. Start sites are chosen to include areas of strong coding potential.
7. Switches in gene orientation (from forward to reverse, or vice versa) are relatively rare. In other words, it is common to find groups of genes transcribed in the same direction.
8. Each protein-coding gene ends with a stop codon (TAG, TGA, or TAA).
9. Each protein-coding gene starts with an initiation codon, ATG, GTG, or TTG. But note that TTG is used rarely (about 7% of all genes). ATG and GTG are used at almost equivalent frequencies.

CONTINUED...

GUIDING PRINCIPLES

OF BACTERIOPHAGE GENOME ANNOTATION

...CONTINUED

10. An important task is choosing between different possible translation initiation (i.e., start) codons. The correct start site can often be distinguished by association with a credible ribosome-binding site (RBS; Shine-Dalgarno (SD) sequence). Identifying the correct start site, however, is not always easy and is predicated on the following sub-principles:

    a. The preferred start site usually has one of the higher SD scores of all the potential start codons, but not necessarily the highest.

    b. Manual inspection can be helpful to distinguish between possible start sites. The consensus is as follows: **AAGGAGG – 3-12 bp – start codon**.

    c. The relationship to the closest upstream gene is important. Usually, there is neither a large gap nor a large overlap (i.e., more than about 4 bp). A short overlap of 1-4 bp—where the start codon overlaps the stop codon of the upstream gene—is very common.

    d. The position of the start site is often conserved among homologues of genes. Therefore, the start site of a gene in your phage is likely to be in the same position as those in related genes in other genomes. But be aware that one or more previously annotated and published genes could be suboptimal, and you may have the opportunity to help change it to a more optimal one.

    e. Your final start-site selection will likely represent a compromise of these sub-principles. For example:

        i. A start codon that overlaps the stop codon of a previous gene trumps a somewhat lower score.

        ii. A higher SD score or canonical RBS trumps a more extended gene overlap.

        iii. If choosing between several starts with similar SD scores, it is usually best to choose the one that gives the longest open reading frame.

11. tRNA genes are not called precisely in the program embedded in DNA Master, and require extra attention. (Please refer to **Section 9.5**.)