Lysogeny in *Streptococcus dysgalactiae* subsp. *dysgalactiae:* lethargy or failure?

From classical infection approaches to whole-genome sequencing

Mariana Mara Nascimento

Thesis to obtain the Master of Science Degree in

Microbiology

Supervisors: Prof. Rogério Paulo de Andrade Tenreiro

Prof. Isabel Maria de Sá Correia Leite de Almeida

Examination Committee

Chairperson: Prof. Jorge Humberto Gomes Leitão

Supervisor: Prof. Rogério Paulo de Andrade Tenreiro

Member of the Committee: Prof. Leonilde de Fátima Morais Moreira

November, 2017

Acknowledgements

This dissertation arises from the collaboration between the Bugworkers group at M&B-BioISI (the Microbiology and Biotechnology unit within the BioSystems & Integrative Sciences Institute) and the Molecular Microbiology group at UCIBIO (Research Unit on Applied Molecular Biosciences). It is the end-product of my one-year stay within the Bugworkers group, but most importantly, it is something that could not have been achieved without the collaboration of a number of individuals to whom I would like to express my sincere gratitude.

Firstly, I want to thank my supervisor, Professor Rogério Tenreiro, for entrusting me with a project that has ultimately shaped my stubbornness into persistence and unquestionably tested my fear of failure. I thank him for all his support, guidance and patience throughout this past year. Most of all, I thank him for presenting me with this challenge and allowing me to make mistakes and not only learn from them, but learn how to surpass them. Of all the lessons I've learned during the past year, this is perhaps the most precious.

I would also like to thank Professor Ilda Sanches, Professor Rosario Mato and Cinthia Barroco, for the resources placed at my disposal and for their interest in this work. My gratitude goes out to everyone at UCIBIO involved in this project as well.

Next, I want to thank Doctor Ricardo Dias for his willingness to place such delicate equipment in my inexperienced hands, but most of all for all the help in acquainting me with the conundrums of genome sequencing. His insight and suggestions regarding the present work are greatly appreciated.

I'd like to express my gratitude to Ana Viana as well, for her help with Atomic Force Microscopy and her availability to guide me through the process and answer my questions.

I'd like to thank Professor Mário Santos for his help during the first and most difficult part of this work as well as Filipa Silva, for her work in the previous Strep project and her current work in assuring the lab is in utmost shape.

I thank my internal supervisor, Professor Isabel Sá Correia for the help in the development of this dissertation.

I'd also like to thank Professor Ana Tenreiro, for her help around the lab, her permanently cheerful demeanor and her much needed good luck tokens.

Honorable mentions go, of course, to all my colleagues and friends inside the Bugworkers group, who have faced the daunting task of putting up with me for the last year and have managed to make it not only bearable but deeply enjoyable. To Cláudia, who has moved on to better things, but not before teaching

me how to plan my work and most importantly, think on my feet. To Pedro, for his availability to answer my questions and discuss ridiculous lunchtime topics. To Ana, André, Beatriz and Catarina for all their support and patience through both sushi outings and times of despair. And to João, for his help with my sequencing attempts and heartfelt protests at the lack of progress bars.

To Inês, Ana Marta for their companionship in these two absolutely crazy years and for always being present and bringing a friendly shoulder, a pair of helping hands, coffee, chocolate, or an assortment of the above. And to Jéssica, who, although further away in this last year, is also a big part of this journey. Without the three of you, I would've certainly quit this masters two months in and missed all the sleepless nights that have brought us here.

To all my friends beyond this masters' bubble, that have ceaselessly heard about my successes and even more about my failures and have encouraged me to go on regardless. To Andreia, Miguel, Ana, Fábio, Víctor, my roommates João, Miguel and Leonor, and many others that have helped me along this journey in their own way.

Lastly, my most profound gratitude goes to my family, for their endless love and support, for their sacrifices and their unwavering belief in me, both as a future scientist and as a person. To them, I owe everything.

Abstract

Streptococci are mostly commensal bacteria found in warm-blooded animals (including humans), but may also cause localized and systemic infections with severe sequelae. Their vast virulence gene repertoire, in part encoded within mobile genetic elements, greatly contributes to their pathogenic success. Concerningly, cases of streptococci regarded as animal pathogens crossing the barrier to become zoonotic agents have been reported. *Streptococcus dysgalactiae* subsp. *dysgalactiae* (SDSD), an animal pathogen involved in bovine mastitis, seems to be undergoing this process, given its recent involvement in human infections. At the root of this phenomenon may be the high rate of bacteriophage-mediated horizontal gene transfer observed between streptococci, particularly involving the emerging zoonotic agents and known human pathogens.

To test this hypothesis, protocols for bacteriophage induction were performed, producing putative phage lysates which were subsequently used in infection assays, where no productive infection was obtained. Phage presence was then assessed through phage DNA extraction and virion visualization through Atomic Force Microscopy with positive results, albeit phage tails could not be detected. To assess prophage genome integrity, whole-genome third-generation sequencing was employed and putative prophages were detected in all tested SDSD strains, as well as bacteriophage resistance systems and phage-associated virulence factors. The number, the varying degrees of integrity, as well as the array of phage-associated sequences and their homology with sequences found in human pathogens and zoonotic agents, support the initial hypothesis that phage elements not only mediate the cross-talk between streptococci but also ultimately shape their pathogenic potential.

Keywords: *Streptococcus*; prophages; horizontal gene transfer; third-generation sequencing; pathogenicit

Resumo

Bactérias do género *Streptococcus* encontram-se presentes em mamíferos (incluindo humanos), e embora sejam maioritariamente comensais podem causar infecções com sequelas graves. O seu vasto repertório de genes de virulência, parcialmente codificado por elementos móveis, contribui para a patogenicidade destes organismos. Diversos exemplos de zoonoses causadas por estreptococos estritamente patogénicos para animais têm sido descritos, nomeadamente envolvendo a subespécie *Streptococcus dysgalactiae* subsp. *dysgalactiae* (SDSD), tipicamente associada a mastites bovinas e recentemente descrita como causa de infecções humanas. Na raiz deste fenómeno pode estar a alta taxa de transferência horizontal de genes mediada por bacteriófagos, observada entre agentes zoonóticos e agentes patogénicos para humanos.

Para testar esta hipótese, foram executados protocolos de indução de bacteriófagos, produzindo lisados fágicos subsequentemente usados em ensaios de infecção. Não foi possível observar infecção produtiva, pelo que a presença de fagos foi avaliada por extração de DNA fágico e observação de viriões através de Microscopia de Força Atómica. Embora esta presença tenha sido confirmada, não foram observadas caudas fágicas. Como tal, procedeu-se à sequenciação do genoma bacteriano para aferir a integridade genómica de possíveis profagos, tendo estes sido detectados em todos os genomas, para além de sistemas de resistência a bacteriófagos e genes de virulência de origem fágica. O número sequências de origem fágica, bem como o seu grau divergente de integridade e de homologia com agentes patogénicos para humanos e agentes zoonóticos parece apoiar a hipótese colocada e indicar que os bacteriófagos são elementos modeladores do potencial patogénico em *Streptococcus*.

Palavras-chave: *Streptococcus*; profagos; transferência horizontal de genes; sequenciação de terceira geração; patogenicidade

Contents

List of Tables

List of Figures

List of Abbreviations

ABB	**A**dapter **B**ead **B**inding buffer
Abi	**Ab**ortive **I**nfection Systems
AFM	**A**tomic **F**orce **M**icroscopy
AMX	**A**dapter **M**ix
BAM	**B**arcoded **A**dapter **M**ix
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
BWA	**B**urrows-**W**heeler **A**ligner
CI	**C**hromosomal **I**slands
CRT	**C**yclic **R**eversible **T**ermination
CS	**C**oding **S**equences
DLA	**D**ouble-**L**ayer **A**gar
ELB	**El**ution **B**uffer
GAS	**G**roup **A** **S**treptococci
HGT	**H**orizontal **G**ene **T**ransfer
HMM	**H**idden **M**arkov **M**odels
ICE	**I**ntegrative and **C**onjugative **E**lements
IGV	**I**ntegrative **G**enomics **V**iewer
KDE	**K**ernel **D**ensity **E**stimate
LAB	**L**actic **A**cid **B**acteria
LLB	**L**ibrary **L**oading **B**uffer
M17YE	**M17** medium supplemented with **Y**east **E**xtract
MAP	**M**inION **A**ccess **P**rogram
MGE	**M**obile **G**enetic **E**lements
ML	**M**ixed **L**ysates
NB	**N**utrient **B**roth
NFW	**N**uclease-**F**ree **W**ater
NGS	**N**ext **G**eneration **S**equencing
ONT	**O**xford **N**anopore **T**echnologies
PacBio	**Pac**ific **Bio**sciences
PSS	**P**rotein **S**ynthesizing **S**ystem

PZT	**P**ie**z**oelec**t**ric Actuator
RAST	**R**apid **A**nnotation using **S**ubsystem **T**echnology
RBF	**R**unning **B**uffer with **F**uel
RM	**R**estriction-**M**odification Systems
RNN	**R**ecurrent **N**eural **N**etworks
SBL	**S**equencing **B**y **L**igation
SBS	**S**equencing **B**y **S**ynthesis
SEM	**S**canning **E**lectron **M**icroscopy
SMRT	**S**ingle-**M**olecule **R**eal-**T**ime
SNA	**S**ingle **N**ucleotide **A**ddition
SPYO	***S****treptococcus* ***pyo****genes*
TEM	**T**ransmission **E**lectron **M**icroscopy
THYE	**T**odd-**H**ewitt medium supplemented with **Y**east **E**xtract
WGS	**W**hole-**G**enome **S**equencing

CHAPTER I. Introduction: the phage-host arms race and its impact on bacterial pathogenicity

1. THE BACTERIOPHAGE

1.1 Phage life cycle

Viruses are broadly viewed as parasitic entities to most organisms, infecting plants, animals and other eukaryotes as well as bacteria and archaea. Bacterial viruses, termed bacteriophages, carry out their life cycle by infecting members of the *Bacteria* domain. Bacteriophages represent the largest share of biological material on Earth, outnumbering bacteria in most (if not all) environments, and their study has allowed the establishment of basic virology concepts (Abedon, 2009; Madigan *et al.*, 2015).

While at first phages were mainly studied as simple model systems, phage research has since then shifted to a more ecological point of view, with focal points ranging from the bacteriophages' role in oceanic matter cycling to their role in bacterial pathogenesis (Brüssow *et al.*, 2004; Labrie *et al.*, 2010). The study of fundamental matters such as an individual phage's survival and, consequently, its potential to reproduce, has revealed their impact on bacterial fitness, bacterial diversity, non-host organisms (for example, eukaryotes that may serve as hosts to bacteria) and even the abiotic environment. Due to their ancient nature, bacteriophages have coexisted with cellular organisms since the earliest life forms came to be, and so their influence in evolutionary dynamics cannot be underestimated (Abedon, 2009; Snyder *et al.*, 2013).

Viruses can exist in both intracellular and extracellular forms. In its extracellular life cycle phase, a viral particle consists of nucleic acid surrounded by a proteinaceous capsid (which may or may not contain other macromolecules). This particle, also termed virion, is metabolically inert and unable to perform biosynthesis; however, it serves as vehicle for the viral genome to move from cell to cell. These genomes can be composed of DNA or RNA in their single-stranded or double-stranded forms, the most common being double-stranded DNA (or dsDNA) viruses; in fact, some phages switch genome composition between DNA and RNA depending on their replication cycle stage (Madigan *et al.*, 2015). Like all viruses, bacteriophages depend on their hosts for almost all functions, since the machinery needed to carry out the viral life cycle is host provided. This has been the base for the debate of whether viruses can be considered living entities or not, despite the fact that they are composed of the same biochemical components as their host counterparts (Brüssow *et al.*, 2004).

Bacteriophage replication takes place inside the host cell and utilizes host machinery to produce all the components that make up new virions. The process starts with the adsorption of a virion to a susceptible

host cell, via cell surface receptors, and the subsequent injection of the viral particle (or, sometimes, only its nucleic acid) into the host. The cell metabolism is then redirected towards the production of viral nucleic acid as well as viral proteins, which will then be assembled into new virions. Viral proteins can be grouped into two comprehensive categories: early proteins, which are necessary for the replication of viral nucleic acid and are thus synthesized first; and late proteins, which include capsid components among other proteins needed only in the moment of assembly (illustrated in **Fig. 1**), allowing them to be synthesized later. After assembly is completed, the new virions are released from the cell (Madigan *et al.*, 2015).

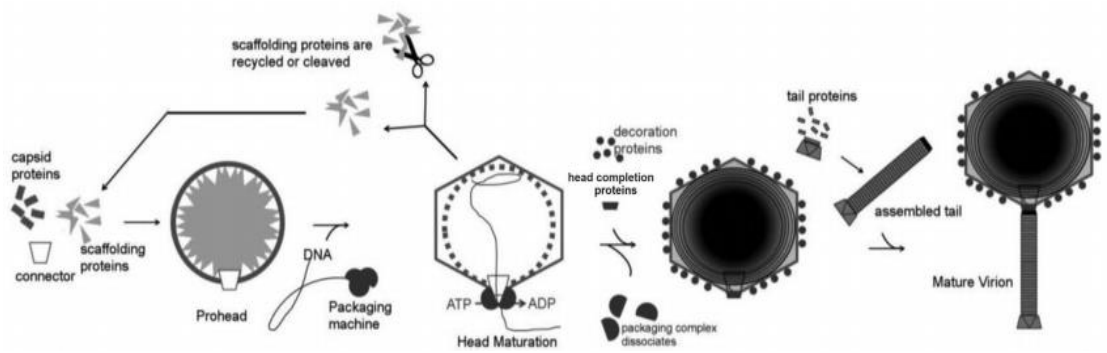


Fig. 1 - **Assembly of a bacteriophage viral particle**. Source: Fokine and Rossmann, 2014

This approach to the bacteriophage life cycle can be rather simplistic, since it does not consider the different pathways it encompasses, as summarized in **Fig. 2.** In the lytic pathway, the host's metabolism is completely overtaken by the virus, resulting in cell lysis shortly after infection; contrastingly, in the lysogenic pathway, the viral genome is replicated along with the host's own genome, allowing it to maintain some degree of control over the metabolic activity and postponing cell lysis. Lysogenic phages, or temperate phages, integrate into the bacterial chromosome or acquire plasmid form to gain control of the host's metabolism, propagating themselves passively as an element of the bacterial chromosome. They retain, however, the ability to revert to a lytic mode of infection under stressful conditions. As long as there is no expression of lytic cycle genes, the host cells, or lysogens, remain unharmed (Fortier and Sekulovic, 2013; Madigan *et al.*, 2015; McShan and Nguyen, 2016).

The mild nature of this infection mode means that a single bacterium can carry multiple prophages, undergoing polylysogeny. Phage gene expression and the possibility of polylysogeny are both controlled by the phage-encoded repressor protein, which prevents gene expression of lytic cycle genes and the insertion of closely related viral entities into the already infected host, providing it with immunity. Disruption of the repressor protein's activity will cancel immunity and induce the prophage, making it enter the lytic pathway and excising it from the bacterial genome. However, the process of viral excision can be hindered by mutations in the viral genome, in which case the prophage becomes a cryptic virus, unable to produce

new virions and infect new hosts. Nevertheless, some phages lacking the necessary components for excision and infection are still able to undergo the process by relying on a helper phage to provide missing functions (Madigan *et al.*, 2015).

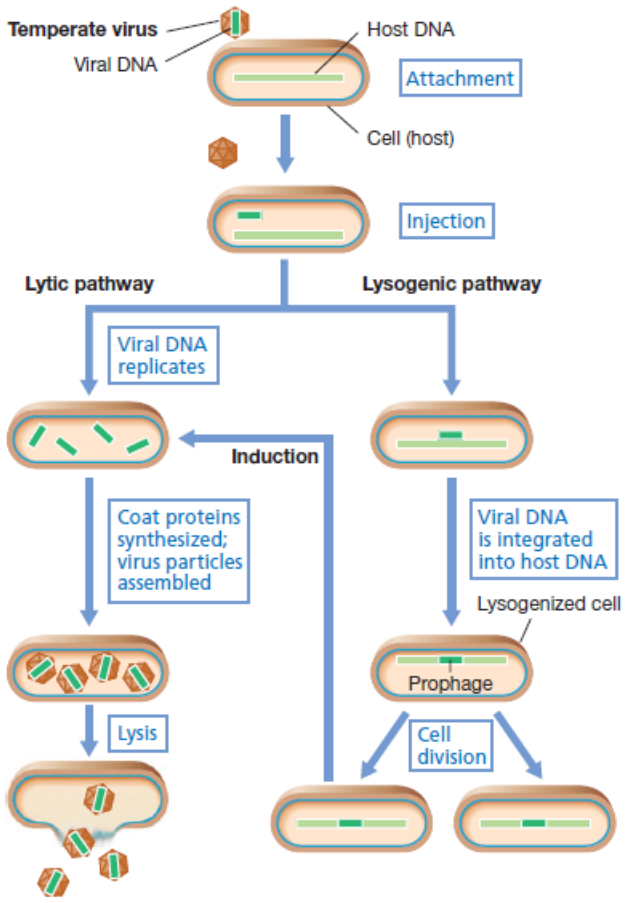


Fig. 2 - **Bacteriophage life cycles**: the lytic and lysogenic pathways. Source: Madigan *et al.* 2012

The lysogenic life cycle is most likely the result of phage adaptation to conditions which are not suitable for rapid virion release. This mechanism allows the phage to delay virion maturation until optimum conditions are reached (Abedon, 2009). However, in extremely unfavorable growth conditions for the host, like starvation for example, the prophage can adopt a third survival strategy – pseudolysogeny. This is a stage of stalled development within the host cell without multiplication of the phage genome (as would happen in lytic development) or its replication in synchrony with the cell cycle (as would happen in lysogenic development); yet, because there is no degradation of the viral genome in a pseudolysogenic state, lytic or lysogenic development can be restarted upon growth condition improvement (Los and Wegrzyn, 2012).

1.2 The *Caudovirales*: phage morphology, genome architecture and evolution

In a similar fashion to their host counterparts, bacteriophages are the object of a formal classification system and are thus grouped into various taxa: orders, families, genus and species, for example. In viral taxonomy, the family taxon is valuable, since members of a given family tend to have a similar virion morphology, genome structure and strategy of replication (Madigan *et al.*, 2015). The most common group are the tailed phages, a category of dsDNA viruses with a protein-only capsid. These phages belong to the *Caudovirales* order which is composed by three families: *Myoviridae, Siphoviridae* and *Podoviridae* (Abedon, 2009).

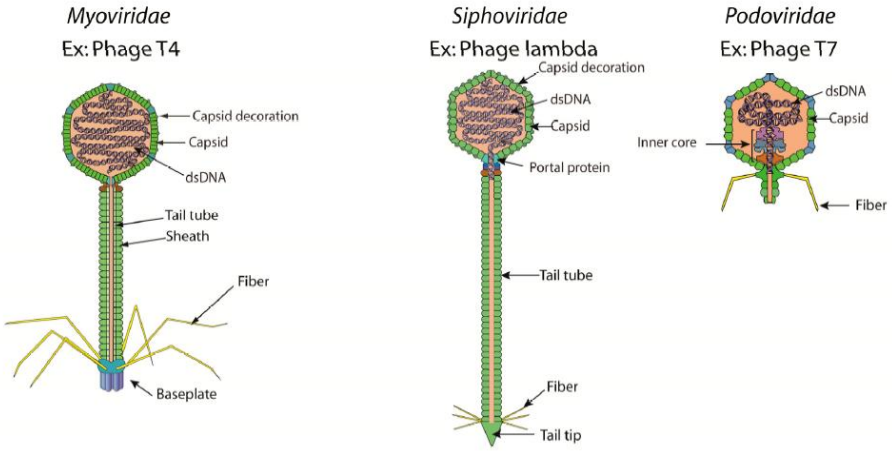


Fig. 3 – **Virion morphotype and structure of *Myoviridae*, *Siphoviridae* and *Podoviridae* members**. Source: Hulo *et al.*, 2017

As can be observed in **Fig. 3**, the *Caudovirales* virions are mainly composed of similar molecular parts, which are assembled in slightly varying pathways. They are composed of an icosahedral protein capsid, which envelops the viral genetic material (and may vary in size according to the size of the genome it contains), the connection between capsid and tail (usually mediated by a protein designated 'portal protein', which works as a channel for genome packaging) and their distinguishing feature, the tail. *Myoviridae* members have long straight contractile tails while *Siphoviridae* phages have long flexible non-contractile tails, and *Podoviridae* have short, stubby, non-contractile tails (Hatfull and Hendrix, 2011; Fokine and Rossmann, 2014). Phage tails have intricate structures and contain many different proteins that help phage during infection, for example, tail fibers which bind to specific sites on the bacterial cell surface (other examples being baseplate proteins and tail tip proteins); because they mediate the phage-bacterium interaction, these structures confer host specificity and influence the phage's host range (Davies *et al.*, 2007; Snyder *et al.*, 2013). Host range can be defined as the breadth of organisms suitable for infection by a given

parasite (in this case, a bacteriophage) and it is constrained by the parasite itself, the host and the environment. Therefore, it is a reflection of the parasite's evolutionary history (Hyman and Abedon, 2010).

Tailed phage genomes can range from 16 kbp up to 500 kbp; however, there is not a uniform distribution of genome sizes across this spectrum, with nearly 50% of all phages in the 30-50 kbp interval. Phages with different virion morphotypes generally have different genome organization and more diverging sequences than phages with the same virion morphotype. This may imply that genome architecture constraints genetic exchange (Hatfull, 2008). Generally, phages with siphoviral morphotypes (the most common among the *Caudovirales*, and thus the most common type of bacteriophages) have synteny among the genes that encode for the virion structure and genes with assembly functions: first are the head/capsid genes (including one or two terminase subunits, the portal protein, a prohead protease, a scaffold protein and the major capsid subunit among others) coupled to the tail genes (including the major tail subunit, the tail tapemeasure protein, minor tail proteins, etc.). Despite this conserved arrangement, phages still contain variable regions with sequences of unknown function; in fact, it is estimated that phages might represent the largest reservoir of unexplored genes. Phages with larger genomes may have less conservation in these regions and more variable regions (Hatfull, 2008; Hatfull and Hendrix, 2011).

Even within regions considered to be more conserved, tailed phage genomes register a staggering amount of recombination events, giving rise to their mosaic structure; the size of modules involved, rates of exchange, and the genome carrying said modules all vary greatly (Hatfull, 2008; Abedon, 2009). The mosaicism of phage genomes is explained through the theory of modular evolution, which proposes "the joint evolution of sets of functionally and genetically interchangeable elements". According to modular theory, the product of phage evolution is "a family of interchangeable genetic elements (modules) each of which carries out a particular biological function"; each viral particle would then be a combination of these modules that is optimized for a given ecological niche. Modules with the same biological function can be exchanged through recombination involving viruses with similar modular construction; although the modules must have the same function, it does not mean that it must be carried out in the same exact manner. Consequently, evolution would act primarily at the modular level, exerting selection according to the following criteria: good execution of function; retention of flanking homology (for proper placement on the genome); and functional compatibility with the maximum number of combinations of other functional units. This means that a module with good function and good compatibility with other modules may be preferred in detriment of a module with excellent function execution, but lower compatibility, ensuring the strive for maximum genetic diversity (Botstein, 1980). Accordingly, it has been observed that disparate phages (or at least parts of them) are often more closely related than their bacterial counterparts, supporting

the hypothesis that most of the genes present on contemporary phages derive from a common ancestral gene pool. The consistent gene order among related phages, suggested by the modular theory, may increase the likelihood of recombinant particles between them being viable (even if the genetic targets do not have close sequence homology). Moreover, the retention of gene order may help preserve patterns of gene regulation, facilitating the coordination of the viral life cycle (Abedon, 2009; Aksyuk *et al.*, 2012).

Another staggering departure from bacterial evolution lies in the fact that virion infectivity is influenced by the amount of DNA packaged within a given capsid – both an insufficient and excessive quantity of genetic material will lead to loss of viability of the phage particle. Consequently, there must be a selection for genome size within the mechanisms of bacteriophage evolution and the processes of DNA gain and loss can be carried out in a way that is independent of gene function. This mechanism counteracts the more familiar selection of genetic sequences for immediate utility, allowing them to be selected for potential future use (Hatfull and Hendrix, 2011).

1.3 Phages as a shaping force in bacterial fitness and bacterial population dynamics

Bacterial evolution differs greatly from that of higher eukaryotes, since sexual life cycles are absent. Consequently, in addition to vertical evolution mechanisms, genetic exchange within a given population is achieved by horizontal gene transfer (HGT), allowing the import of functional genetic units from other individuals belonging to the same or even to different species. While vertical evolutionary mechanisms are regarded as the slower gear of bacterial evolution, horizontal gene transfer represents the faster mode. Thus, while the genetic gains from HGT can be short-lived, they may represent a brief selective advantage which can be crucial in unstable environments and in allowing bacteria to exploit these rather difficult niches (Brüssow *et al.*, 2004). Lysogenic conversion is a phenomenon where a non-defective phage carries genes which are expressed by the lysogen (contrary to most prophage genes) and lead to changes in the lysogen's phenotype. The instance where a phage converts a non-virulent bacterial strain into a virulent one is an example of lysogenic conversion (Fortier and Sekulovic, 2013).

As a result of the action of these evolutionary mechanisms, a bacterial genome can be divided into the core genome sequence (shaped by vertical evolutionary mechanisms) and the variable or accessory genome portion (shaped mainly by horizontal evolutionary mechanisms) (Canchaya *et al.*, 2003). There are several HGT mechanisms, which mediate the transfer of DNA in its various forms: naked DNA, plasmid, conjugative transposon or phage (through transformation, conjugation, transposition and both lysogenization and transduction with bacteriophages, respectively) (Brüssow *et al.*, 2004).

Transduction can happen upon prophage induction and it implies the packaging of DNA fragments into bacteriophage particles and subsequent delivery of this DNA to infected cells (Abedon, 2009; Murray *et al.*, 2009). However, instead of phage DNA, host genome fragments can be accidentally packaged and subsequently incorporated into another host's genome (Brüssow *et al.*, 2004). Additionally, specialized transduction may take place: this term accounts for high-efficiency virus-mediated replication and packaging of a non-viral gene. The difference between generalized transduction and specialized transduction is the degree to which bacterial DNA is incorporated - in generalized transduction it can completely replace the viral genome, causing the loss of phage genetic viability. In specialized transduction, however, the bacterial DNA is merged with viral DNA, maintaining functionality of the phage particle (Abedon, 2009; Hyman and Abedon, 2010; Snyder *et al.*, 2013).

Besides their modes of insertion, it is also noteworthy that, as stated in **section 1.1**, bacteriophages are very prone to recombination events. As with transduction, these events may occur not only between phages, but also involving phages and DNA fragments from plasmids, fragments from the host's chromosome or even foreign DNA. The diverse nature of DNA fragments that can be inserted into a viral particle, along with the various recombination systems available (illegitimate recombination, homologous recombination or even site-specific recombination, for example), contribute to the diversification of bacteriophage genomes and influence their evolutionary dynamics (Brüssow *et al.*, 2004; Fortier and Sekulovic, 2013).

The ability to acquire such diverse genetic patrimony highlights the role of phages in shaping their hosts and the respective bacterial populations, regardless of the development pathway the phage undergoes: lytic phages can shape the host population by eliminating susceptible cells or promoting genetic exchange; lysogenic phages, on the other hand, can alter the host cell phenotype, producing long term effects on the lysogen (McShan and Nguyen, 2016). They also directly affect their host's fitness in several ways: they can serve as anchor points for genome rearrangements, mediate gene disruption, protect the bacterium from lytic infection (by preventing a secondary infection through the synthesis of specific proteins or other mechanisms), lyse competing strains (through prophage induction) and can introduce new fitness factors (Brüssow *et al.*, 2004; Labrie *et al.*, 2010). Phage-lysogen dynamics can be particularly complex, since they depend on the rate at which the phage integrates the bacterial genome, as well as the rate at which it disappears (either through excision or accumulation of mutations and subsequent loss of phage DNA) and the number of recombinational events the viral genome has been exposed to (Brüssow *et al.*, 2004).

Intuitively, it would be expected for prophage integration into the genome to result in the decrease of bacterial fitness, since the viral DNA represents a metabolic burden, but also considering that the

prophage can ultimately cause lysis of the host. To balance these negative effects, the prophage must then provide traits that will increase fitness, or else the lysogen would not be maintained in the population, which would mean disappearance for the phage as well. For example, phage encoded immunity and superinfection exclusion genes (mentioned in **section 1.1**) provide selective advantage to the host, since they protect it from further viral infection; they seem, nevertheless, primarily advantageous to the phage, by preventing competition between foreign viral DNA and the resident prophage (Desiere *et al.*, 2001; Canchaya *et al.*, 2003; Cumby *et al.*, 2012). However, because viral genomes can acquire new functions through recombinational events, they may also contain genes that encode beneficial traits for the host but have no direct use for the phage itself. Because they play no role in the carrying out of the lysogenic phage cycle, and seemed simply a nuisance to the phage, the term "moron genes" was coined to described them. The presence of promoter and terminator elements in moron genes as well as the differences in G+C content from surrounding genetic units set them apart from the prophage genome, further settling their identity as a product of HGT. Besides their presence in bacteriophages, moron gene integration in chromosomal sites has been observed, confirming their existence as selfish genetic entities who explore bacteriophages for mobility purposes (Brüssow *et al.*, 2004; Cumby *et al.*, 2012). Incorporation of moron genes into the viral genome occurs in a similar way to specialized transduction. In fact, moron genes are thought to represent an intermediate state between the two forms of transduction, with generalized transduction being the less biased phenomenon of foreign DNA incorporation into a viral particle and, conversely, specialized transduction being the most biased version (Abedon, 2009; Hyman and Abedon, 2010; Snyder *et al.*, 2013).

Bearing in mind the beneficial traits encoded by phages and the detrimental effects of their presence in a bacterial genome, the most profitable evolutionary outcome would be the selection of lysogens with mutations in prophage DNA that can inactivate the prophage induction process or even a large-scale phage DNA deletion, hampering the lysogenic cycle. However, to attain optimum bacterial fitness, the useful viral genes would be spared from the deletion process, maintaining their position in the genome. For this selective process to occur, a high genomic deletion rate is needed for the removal of deleterious genetic elements, which may help explain the overall constant size of bacterial genomes in spite of the constant integration of parasitic DNA (Desiere *et al.*, 2001; Canchaya *et al.*, 2003).

It becomes clear that prophages are relevant genetic elements, both quantitatively and in their role as HGT vectors, contributing to the host cell's physiology (Canchaya *et al.*, 2003). In fact, the incorporation of prophages into the core bacterial genomes results in much of the diversity observed in closely related bacterial strains (Banks *et al.*, 2004; Cumby *et al.*, 2012). Ultimately, a bacterial population is shaped both by the predatory action of phages and by the presence of phage-encoded genes which may enhance bacterial

survival, help conquer new ecological niches and maintain previously acquired ones (Cumby *et al.*, 2012). Phage-microbe interactions are then a prime example of the Red Queen hypothesis, which posits that environmental interactions lead to continuous variation and selection, leading, in this case, to adaptation of the host and counter-adaptation of the parasite (Stern and Sorek, 2012). These co-evolution cycles involve the emergence of phage-insensitive hosts, which are responsible for preserving bacterial lineages, and the emergence of counter-resistant phages, which threaten new bacterial strains. The back-and-forth mechanism of bacterial resistance to phages and the appearance of new phages is essential in shaping bacterial populations in virtually all known habitats as well as defining phage host range (Hyman and Abedon, 2010; Labrie *et al.*, 2010).

1.4 Co-evolution: the host's bacteriophage resistome

The term "Bacteriophage Resistome" is used to describe the set of defense mechanisms bacteria have developed to prevent bacteriophage infection. These mechanisms can be divided into broad categories: adsorption resistance mechanisms (which work by diminishing the contact between the viral particle and its host, through loss of receptor molecules, for example), restriction mechanisms (which cause the death of phage particles but preserve the host) and abortive infection mechanisms (which result in the death of both the bacteriophage and the host) (Hyman and Abedon, 2010). Of particular relevance to the present work are abortive infection mechanisms and two restriction mechanisms: restriction-modification systems and CRISPR/Cas systems. These widely known restriction mechanisms (which are also useful tools in genetic engineering) have high genetic variability and are also prone to undergo HGT, in order to be spread through bacterial populations (Stern and Sorek, 2012).

These systems might also have costs to the cell since they are not error-free and errors in either system can lead to targeting and destruction of bacterial genetic material. Interestingly, the presence of these systems inside bacteriophages has also been observed, in addition to chromosomal or plasmid presence. This poses two hypotheses: resistance mechanisms allow superinfection exclusion, and are advantageous to the phage for it, or they exist as selfish genetic entities that, resembling moron genes, exploit bacteriophages as means of transportation (Hyman and Abedon, 2010; Stern and Sorek, 2012).

1.4.1 Restriction-modification systems

The widely-known restriction-modification system (RM) is based on its abilities to restrict incoming foreign genetic material and to protect host DNA from restriction (usually through modification of specific bases in the DNA sequence, like methylation for example) as can be seen in **Fig. 4**. To do so, the system recognizes specific phage DNA sequences, 4-8 bp long in average. Because host DNA is modified, unmodified sequences are then assumed to be foreign and thus cleaved. To function, an RM needs a methyltransferase (assuming the modification performed on host DNA is indeed methylation) and a restriction endonuclease (Stern and Sorek, 2012).

If the system fails, intruding phages will be replicated and modified by the cell, becoming resistant to restriction. In response to restriction-modification systems, phages can be equipped with proteins that block restriction, encode their own methyltransferase, stimulate the host's methyltransferase to modify phage DNA or they may avoid containing palindromic sequences in their genomes, since most restriction enzyme recognition sites are of this nature (Hyman and Abedon, 2010; Stern and Sorek, 2012; Vasu and Nagaraja, 2013).

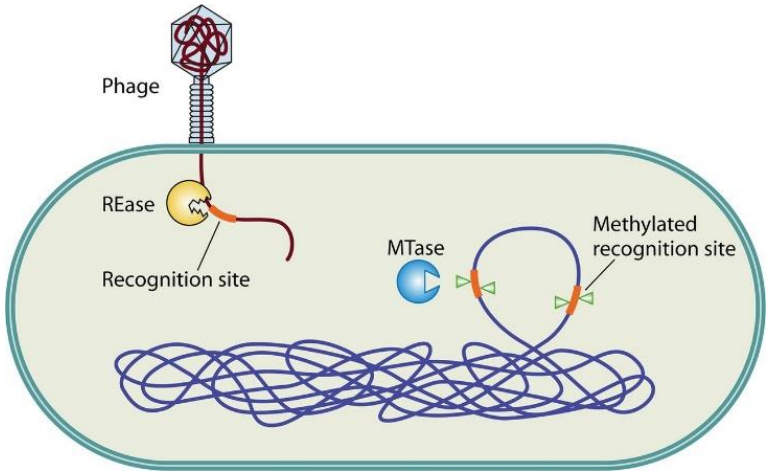


Fig. 4 - **The restriction-modification system.** The MTase (methylase) modifies the host's DNA, making it resistant to the REase's (restriction endonuclease) action. Source: Vasu and Nagaraja, 2013

1.4.2 CRISPR/Cas systems

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) target nucleic acid with specific sequences, providing acquired immunity against phages and plasmids. These loci consist of several noncontiguous direct repeats separated by stretches of spacers (variable sequences acquired from phages or plasmids) often located next to *cas* (CRISPR-associated) genes. Through cleavage of the external sequences and integration into the CRISPR loci, the cell becomes able to recognize the sequence in external elements and avoid subsequent infections with phages containing it (as represented in **Fig. 5**). Even though

spacer sequence acquisition does not seem to have a fitness cost for the host, CRISPR loci cannot expand indefinitely; the optimum parameters and size of the loci, however, are mostly unknown. A CRISPR locus is usually transcribed into a single RNA transcript, which is then cleaved by Cas proteins, generating smaller CRISPR RNA units that target one spacer each. Upon infection, these units pair with foreign nucleic acids, signaling the degradation of foreign sequences (Horvath and Barrangou, 2010; Hyman and Abedon, 2010; Stern and Sorek, 2012).

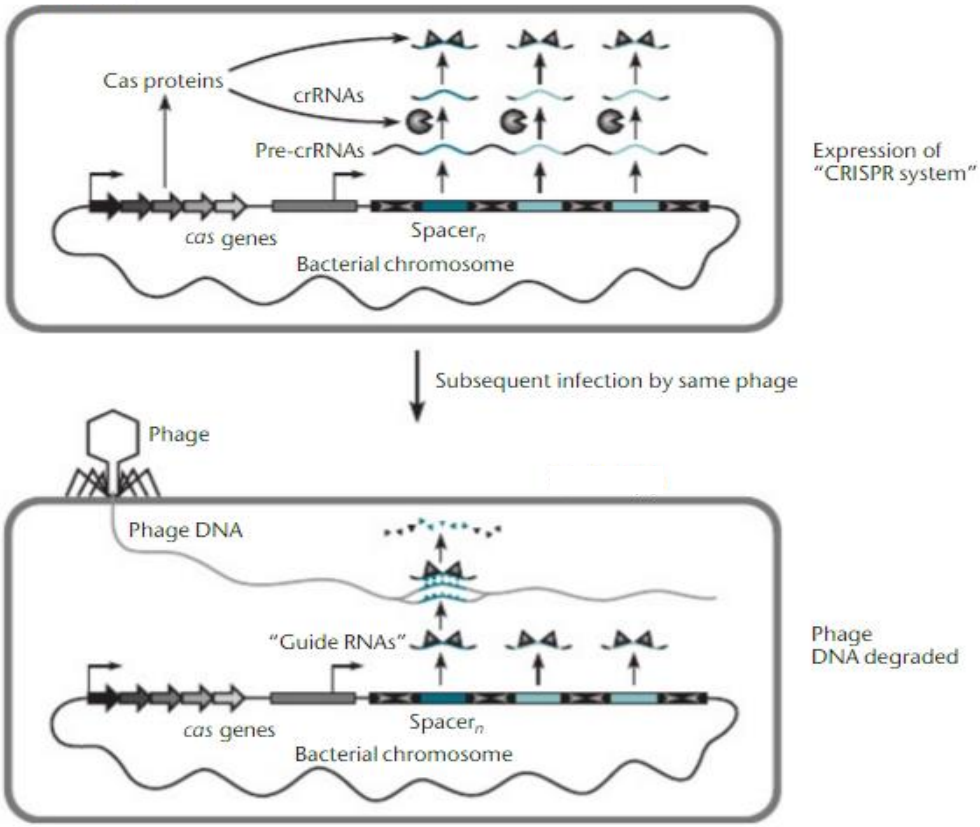


Fig. 5 - **The CRISPR immunity system**. Adapted from: Snyder *et al.*, 2013

Because the viral sequence becomes present in the bacterial genome after the first infection, it allows the microbe to build up inheritable DNA-encoded immunity. However, phages have acquired mutation based strategies to evade CRISPR/Cas systems, by, for example, losing their spacer sequences or encoding products that target Cas proteins (Horvath and Barrangou, 2010; Stern and Sorek, 2012).

1.4.3 Abortive infection systems

Abortive Infection Systems (Abi) is a term used to describe host mechanisms that arrest phage development at its different stages, for example: phage transcription, genome replication, genome packaging, etc. Abi mediated resistance ultimately causes the death of the cell; it is then advantageous as a

selfless defense mechanism, since the host dies, but the surrounding population is benefitted. For their toxicity, Abi systems are tightly regulated (Stern and Sorek, 2012).

Although some of these systems work similarly to toxin-antitoxin systems, Abi systems are vastly diverse and their modes of action are still not completely understood. In Gram-positive bacteria, at least 23 distinct mechanisms have been described, usually mediating the Abi phenotype through a single gene, but requiring two to four in some cases. The systems AbiA, AbiF, AbiK, AbiP and AbiT act at phage DNA replication level; AbiB, AbiG and AbiU act by interfering with RNA transcription; AbiC limits the production of the major capsid protein; AbiE, AbiI and AbiQ affect phage DNA packaging; AbiD1 hampers the action of a phage-encoded RuvC-like endonuclease (responsible for resolving Holliday junctions) and AbiZ causes premature lysis of already infected cells (Iwasaki *et al.*, 1991; Labrie *et al.*, 2010).

1.5 Phage influence across different infection tiers: changing the pathogen-host interaction paradigm

Bacterial adaptation to mammalian hosts poses more of a challenge than adapting to abiotic ecological niches or even simpler life forms, given their extensive defense mechanisms which evolve along with microbes and adapt to them, providing another example of The Red Queen hypothesis (discussed in **section 1.3**). One of the most striking examples of bacteriophage influence in their surrounding environment is their ability to modulate bacterial pathogenicity (Brüssow *et al.*, 2004).

The interaction between a bacterial pathogen and its mammalian host (henceforth referred to as pathogen-host interaction) comprises several steps, including search for an entry site, targeting of a suitable locale for multiplication within the host and becoming persistent in the original host or reaching the next host. The overall success of a pathogen depends on its ability to survive and multiply in a given environment and to propagate itself through several hosts (Wagner and Waldor, 2002; Brüssow *et al.*, 2004).

Some of the features which contribute towards pathogenic success are virulence factors, which play an especially important role in the evasion of host defense mechanisms, engaging, subverting or destroying mammalian host cells. For a given feature to be considered as a virulence factor, it must benefit the cell by either: enhancing the pathogen's fitness in its regular niche within the host, allowing it to outnumber existing competitors; facilitating adaptation to environmental changes in this niche; or mediating the conquest of new niches. Moron genes appear to be strong candidates to serve as virulence factors, since they do provide bacteria with beneficial traits; however, the expression of moron genes must be synced with the metabolism of the bacterium for these benefits to be useful. Additionally, if the product of moron gene expression depends on other bacterial factors (through interaction, for example), further synchronization with these

factors is also required. This is vital for the integration of new factors into a bacterial virulence network, which is usually quite intricate and finely tuned to ensure the pathogen's success (Wagner and Waldor, 2002; Brüssow *et al.*, 2004). Fittingly, because moron genes can also exist as selfish genetic elements, they are usually organized as discrete autonomous elements within prophages, which minimizes interference with possible adjacent prophage structural genes, allowing optimal expression of virulence factors during the lysogenic cycle, in which most prophage structural genes are repressed (Fortier and Sekulovic, 2013).

The long lasting co-evolution between prophages and their bacterial hosts has allowed seamless integration of some prophages in the host's regulatory network, facilitating the phage-bacterium crosstalk and benefitting both parts by enhancing bacterial fitness and altering virulence attributes (Banks *et al.*, 2003; Fortier and Sekulovic, 2013). Although the benefits of phage presence in a bacterial genome have been explored throughout this section mainly from the point of view of acquisition of new genetic material, the transition to a pathogenic phenotype from a commensal one can also be achieved by loss of genes (for example, genes involved in toning down certain virulent traits). Luckily, bacteriophages can mediate both processes: they can be equipped with moron genes that work as virulence factors for the bacterial host and they can also cause single-gene loss when integrating disruptively into the host's genome (by interrupting coding sequences or being placed in intergenic regions essential for coordinated gene transcription) (Wagner and Waldor, 2002; Brüssow *et al.*, 2004).

It becomes clear that the role of bacteriophages as modulators of bacterial pathogenicity is relevant enough to justify the change of the traditional host-pathogen interaction paradigm and introduce phage presence as the third factor. Consequently, in the case of pathogens susceptible to viral influence, the two different tiers of infection should be equated, and thus host-pathogen-phage interactions should be considered instead (Brüssow *et al.*, 2004; Labrie *et al.*, 2010).

The evolutionary dynamics of pathogenic bacteria are one of the many examples that highlight the dual outcomes of phage presence within a bacterial host. This presents a challenge to the traditional view of phages as simply parasitic elements, since both intervenients reap benefits from the established relationship, which would classify them as symbionts (Cumby *et al.*, 2012).

2. BACTERIAL HOSTS: AN OVERVIEW OF THE *STREPTOCOCCUS* GENUS

2.1 General features and pathogenic potential

Streptococci are gram-positive, low G+C content bacteria, first described by Rosenbach. Streptococcal cells are quite small (less than 2 μm in diameter), spherical or ovoid in shape, nonmotile and

unable to form endospores. The genus currently comprises over 90 different species, according to LPSN[1], its type-species being *Streptococcus pyogenes*., and is placed within the *Bacteria* domain, the *Firmicutes* phylum, the *Bacilli* class, the *Lactobacillales* order and finally, the *Streptococcaceae* family (Whiley and Hardie, 2009).

Bacteria belonging to this genus are chemo-organotrophic, presenting a fermentative metabolism during which lactic acid is formed as a result of carbohydrate fermentation (hence the inclusion of this genus in the Lactic Acid Bacteria (LAB) group); besides lactic acid, minor amounts of acetic and formic acids, ethanol and CO_2 may also be produced. The nutritional requirements for these bacteria are both complex and variable. Most *Streptococcus* species are facultatively anaerobic and some require the presence of additional CO_2 for growth. The optimum growth temperature is usually around 37°C but can vary slightly between species (Whiley and Hardie, 2009; Gera and McIver, 2013).

A relevant property of streptococci is their ability to rupture erythrocytes and release their contents into the surrounding environment – hemolysis. There are three distinguishable types of hemolysis: β-hemolysis or complete hemolysis; α-hemolysis or incomplete hemolysis (characterized by a greenish halo around the colonies); and γ-hemolysis which is the absence of hemolysis. This trait, along with biochemical and physiologic properties, can be used to identify and differentiate streptococcal species (Facklam, 2002; Murray *et al.*, 2009). However, the pathogenic features of these bacteria urged the need to create additional classification systems that might help in diagnosis. The most popular of these systems is the Lancefield grouping, which is based on the serotyping of the cell-wall carbohydrate present in *Streptococcus* cells. The groups are designated by letters, according to which cell-wall associated group antigen they possess (A, B, C, E, F, G, etc.) (Whiley and Hardie, 2009).

These bacteria are often associated with warm-blooded animals, including humans. Most species establish relationships of a commensal nature with the respective hosts, inhabiting their mucosal surfaces in the oral cavity, upper respiratory tract and gastrointestinal tract among others; however, given the adequate conditions, streptococci can cause both localized and systemic infections (Whiley and Hardie, 2009).

Human streptococcal diseases can range from infections of the upper respiratory tract, skin and soft tissue to septicemia, meningitis, pneumonia and even bacterial endocarditis (Mims *et al.*, 1998). Additionally, these bacteria are also responsible for several infections in other animals, as well as diseases transmittable from animals to humans (commonly known as zoonoses). The rapid growth of the human population, along

[1] LPSN is available for consultation at: http://www.bacterio.net/

with its high demand for food and animal products and close contact with companion animals may foster the evolution of zoonotic streptococci, thus renewing the importance of understanding both animal and human pathogens as well as the relationships between them (Fulde and Valentin-Weigand, 2012).

The breadth of organisms infected by these bacteria, as well as the multitude of symptoms such infections can have, require a wealthy virulence factor repertoire and means for quick adaptation. This places streptococci as interesting subjects for the exploring of HGT phenomena and phage influence in bacterial fitness.

2.2 Intra-genus phylogenetic relationships

Because of its diversity regarding 16S rRNA sequences, this genus is organized in the following "species groups": "Pyogenic", "Bovis", "Mutans", "Mitis", "Anginosus", "Sanguinis" and "Salivarius" (Whiley and Hardie, 2009). Afterwards, the "Downei" group was created to accommodate *S. downei* and *S. criceti*. This additional grouping was first included in the phylogeny in **Fig. 6.** However, some relationships between groups are poorly resolved, possibly reflecting the effect of frequent HGT during the early diversification of these clusters and attesting to the plasticity of streptococcal genomes. The observed horizontal transference may have a role in adaptation and is more frequent within groups than between them (Richards *et al.*, 2014).

Of special interest for this work is the "Pyogenic" group, which comprises *S. pyogenes*, *S. dysgalactiae, S. agalactiae, S. equi,* among others. The group encompasses varying stages on the spectrum of pathogenicity, including human pathogens, animal pathogens and zoonotic agents (Bentley *et al.*, 1991; Whiley and Hardie, 2009).

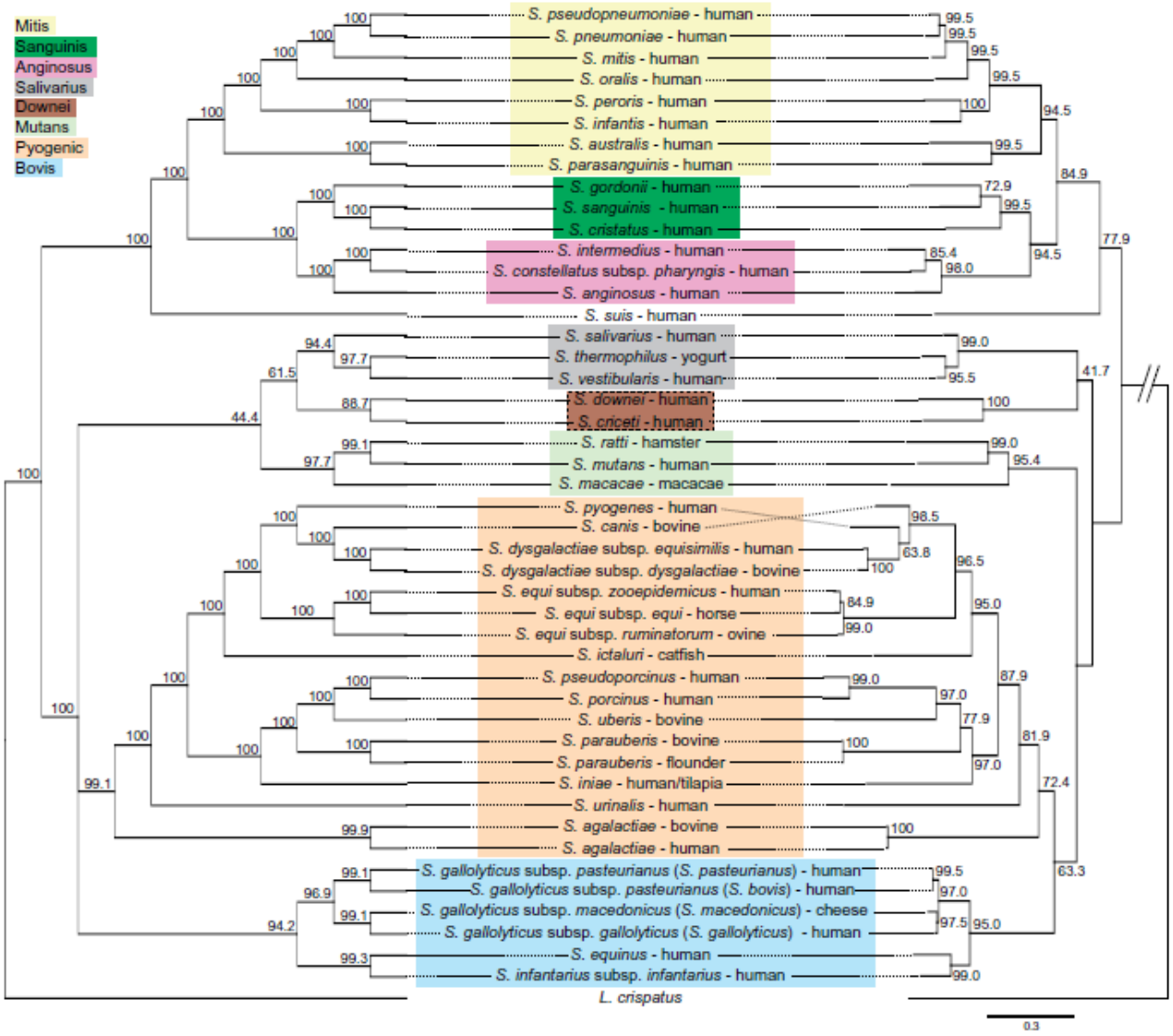


Fig. 6 - **Phylogenies for the genus *Streptococcus* based on a core set of 136 genes**. The left represents the consensus of the phylogenetic signal from each gene (numbers represent the percentage of genes that support each grouping); the right represents the maximum likelihood phylogeny derived from a concatenation of the genes (numbers represent bootstrap support for the grouping). Source: Richards *et al.*, 2014

2.3 *Streptococcus pyogenes*

Streptococcus pyogenes (SPYO) can colonize the throat and skin of its host and can cause several suppurative infections and non-suppurative sequelae, being considered the most pathogenic species within the genus. Strains of *S. pyogenes* can be referred to as GAS (Group A Streptococci), for possessing the "A" antigen in Lancefield grouping, although *S. pyogenes* is not the only species to possess this antigen (Facklam, 2002).

This species represents the most common cause of bacterial pharyngitis, impetigo, and scarlet fever as well as erysipelas and other spreading infections (cellulitis, bacteremia, etc.); furthermore, SPYO can also

be responsible for streptococcal toxic-shock syndrome and necrotizing fasciitis, a condition that is often fatal (Facklam, 2002; Whiley and Hardie, 2009). Thus, infections caused by this organism can range from severe invasive diseases and superficial symptomatic ones to asymptomatic commensal colonization. As in many other species, the virulence of *S. pyogenes* depends on the bacterium's capability of adhering to host cells and invading their target cells (in this case, epithelial cells), of producing toxins and enzymes relevant to the infectious process and of evading the host's immune system (avoiding phagocytosis and opsonization, for example). Streptococci usually carry out infection outside host cells; nevertheless, there may be establishment of an intracellular population able to promote programmed death of the infected cells. The internalization into epithelial cells is believed to be pertinent to the maintenance of persistent infections and invasion of deep tissues (Murray *et al.*, 2009).

A remarkable increase in the incidence of necrotizing fasciitis and streptococcal shock syndrome caused by *S. pyogenes* has been observed worldwide, which has renewed the importance of investigating these bacteria. These changes in virulence may be linked to the streptococcal genome's plasticity, since it includes numerous mobile genetic elements which are largely responsible for genetic differences between SPYO strains, producing an open pan-genome (Banks *et al.*, 2004; Vojtek *et al.*, 2008; Richards *et al.*, 2014; Maruyama *et al.*, 2016).

This species also harbors multiple virulence factor encoding phages in its genome, which may differ from strain to strain, allowing substantial permutation of virulence factor combinations; these permutations may be responsible for the distinct diseases caused by different strains of *S. pyogenes* and for the temporal and geographical variability of clinical isolates (Brüssow *et al.*, 2004; Davies *et al.*, 2007). Numerous streptococcal virulence factors, such as: adhesion factors, lipases, DNases, streptokinases, hyaluronidases, and even the streptococcal pyrogenic exotoxins (a family of superantigens) are encoded by genes located in prophages (Kuhl *et al.*, 2012). The virulence of SPYO is also highly associated with other mobile genetic elements (MGE) such as chromosomal islands (CI) and phage-like chromosomal islands (SpyCI), which confer a mutator phenotype to the host, further increasing intra-species diversity (Nguyen and McShan, 2014).

When contacting the mammalian host, pathogenic *S. pyogenes* cells alter their gene expression pattern and, by lysogenization of bystander cells, alter the genomes of commensal *S. pyogenes* strains into potentially virulent ones. However, the relationship between the presence of certain prophages and the bacterial host's virulence is not always linear, since it is possible that the prophage encoded fitness factors increase colonization and persistence capacity but not virulence directly. In this scenario, the bacteria become more successful colonizers and as such, have a better chance of causing infection, but their mechanisms for doing so are not necessarily more efficient (Boyd and Brüssow, 2002; Brüssow *et al.*, 2004;

Vojtek *et al.*, 2008). Beyond the phage-bacteria communication, some data suggest that there is crosstalk between bacteriophages and the mammalian host, since the co-culture of streptococci with mammalian cells can lead to the production of bacteriophage particles (Boyd and Brüssow, 2002).

2.4 *Streptococcus dysgalactiae*

The overall gene content of *S. dysgalactiae* is very similar to *S. pyogenes*; in fact, virulence factors like those of *S. pyogenes* have been detected (Facklam, 2002; Suzuki *et al.*, 2011). The species was divided into two subspecies by Vandamme *et al.* (1996), and several techniques, such as pulsed-field gel electrophoresis, DNA-DNA reassociation experiments, multilocus enzyme electrophoresis, phenotypic experiments, and phylogenetic analysis of several gene sequences have since supported the division (Suzuki *et al.*, 2011; Jensen and Kilian, 2012). A study involving strains from both subspecies found that only 12-16% of their gene content is unique and these differences are related to the assortment of virulence loci present in each one, which is, in turn, connected to the presence of mobile elements (Suzuki *et al.*, 2011).

2.4.1 *S. dysgalactiae* subsp. *equisimilis*

The taxon *S. dysgalactiae* subsp. *equisimilis* (SDSE) was proposed for *S. dysgalactiae* isolates of human origin. The strains are usually β-hemolytic and belong to the A, C, G and L Lancefield groups (although groups C and G are the most frequent) (Vandamme *et al.*, 1996; Facklam, 2002), It was initially regarded as a human commensal organism possibly present in the skin, oropharynx, gastrointestinal and genitourinary tracts (Takahashi *et al.*, 2011). However, this subspecies has become an increasingly important human pathogen responsible for a range of human diseases, such as: acute pharyngitis, bacteremia, cellulitis, endocarditis, endophthalmitis, gas gangrene, meningitis, necrotizing fasciitis, peritonitis, pneumonia, salpingitis, sepsis, septic arthritis, skin infections and toxic shock-like syndrome (Vieira *et al.*, 1998; Suzuki *et al.*, 2011; Genteluci *et al.*, 2015). This infection spectrum partially overlaps with that of *S. pyogenes*, raising the possibility that the disease burden attributed to SDSE has been underestimated (Davies *et al.*, 2007; Jensen and Kilian, 2012).

Acquisition of genetic material through HGT and accumulation of point mutations may have conferred these bacteria the ability to colonize a new ecological niche (Brandt and Spellerberg, 2009). As in the case of *S. pyogenes*, bacteriophages seem to generate diversity within the taxon, being responsible for the differences between pathogenic and commensal isolates. Interestingly, some SDSE phages appear to be related to GAS phages. Chromosomal islands have also been reported to exist in SDSE (Davies *et al.*, 2007).

2.4.2 *S. dysgalactiae* subsp. *dysgalactiae*

The taxon *S. dysgalactiae* subsp. *dysgalactiae* (SDSD) was retained for strains of animal origin, belonging to the C and L Lancefield groups and presenting all types of hemolysis (Rato *et al.*, 2010; Takahashi *et al.*, 2011). This subspecies can be distinguished from *S. dysgalactiae* subsp. *equisimilis* by proteolysis of human fibrin, by a human plasminogen-streptokinase test (SDSD isolates will respond negatively to both tests) and by whole-organism protein electrophoretic patterns (Vandamme *et al.*, 1996; Vieira *et al.*, 1998).

This subspecies is associated with bovine mastitis (along with *S. uberis*, *S. agalactiae*), a highly prevalent disease with major relevance for the dairy industry, as well as toxic shock-like syndrome in cattle among other diseases (Rato *et al.*, 2011, 2013). It has previously been isolated from infected mammary glands, teat injuries and is transmitted primarily during milking. Furthermore, it has been detected in extramammary reservoirs such as cattle tonsils, mouth and vagina (Calvinho *et al.*, 1998). SDSD's ability to cause bovine mastitis is particularly relevant, given the sizeable dairy industry and regular human consumption of products containing dairy (Halasa *et al.*, 2007).

The detection of SDSD infections in farmed fishes has increased recently, and although this subspecies is generally disregarded as a human pathogen, it has been described as the cause of zoonotic infections upon contact with infected fish (Koh *et al.*, 2009; Suzuki *et al.*, 2011; Park *et al.*, 2012; Abdelsalam *et al.*, 2013). Instances of SDSD prosthetic joint infection after total knee arthroplasty and infective endocarditis have also been reported (Park *et al.*, 2012; Jordal *et al.*, 2015). This suggests that *S. dysgalactiae* subsp. *dysgalactiae* may be an emerging zoonotic pathogen (Rato *et al.*, 2011).

2.5 Horizontal gene transfer between streptococcal species

Although genetic transfer in streptococci can be mediated by all mechanisms discussed in **section 1.3**, transduction might be particularly relevant since bacteriophages have been detected in considerable proportion, especially among GAS. Because phages can encode virulence factors, they contribute to the organism's pathogenicity and thus play a role in adaptation of the microbe to different hosts and different environmental pressures; moreover, the contribution of phage presence has been recognized in the generation of streptococcal strains with increased pathogenic potential (Whiley and Hardie, 2009).

All currently known LAB prophages show conservation in their overall gene order, which is as follows: left attachment site (*attL*) – lysogeny – DNA replication – transcriptional regulation – DNA packaging – head – joining – tail – tail fiber – lysis modules – right attachment site (*attR*) (Canchaya *et al.,* 2003). This structural conservation is, as exploited in **section 1.2**, advantageous to phages. Most of the lysogenic phages infecting

the *Streptococcus* genus belong to the *Siphoviridae* family, although infections by *Podoviridae* and *Myoviridae* have also been described (Canchaya *et al.*, 2003).

The case of *S. pyogenes* and *S. dysgalactiae*

Although 16S rRNA analysis suggests *S. agalactiae* to be the closest relative to SDSE, genome wide and gene level comparison places *S. pyogenes* closest both at nucleotide and amino acid sequence level (Shimomura *et al.*, 2011; Maruyama *et al.*, 2016). The mosaic structures present in some *S. pyogenes* and *S. dysgalactiae* subsp. *equisimilis* genes also suggest the recent and ongoing occurrence of interspecies HGT events (Davies *et al.*, 2007). These are particularly important in regards to the observed virulence overlap between these bacteria and are likely to be mediated by MGE such as integrative and conjugative elements (ICE) and prophages, which have been described for both species (Haenni *et al.*, 2010; Jensen and Kilian, 2012).

Putative prophage regions were detected in strains from *S. dysgalactiae* subsp. *dysgalactiae* and *S. dysgalactiae* subsp. *equisimilis* and found to be homologous to prophages from *S. pyogenes*, sharing, in some cases, the same integration sites (Shimomura *et al.*, 2011; Suzuki *et al.*, 2011). In a separate study, SDSD strains were found to carry bacteriophage virulence-associated genes highly similar to those of SPYO, suggesting that bacteriophages may also play a role in the genetic plasticity and virulence of bovine mastitis SDSD isolates (Rato *et al.*, 2010, 2011). Phylogenetic studies seem to support these claims, pointing towards a strong net directionality of gene movement from SPYO donors to SDSE recipients, although HGT phenomena in the reverse direction have also been observed. Directionality of phage movement between species depends on several factors, such as surface characteristics and the bacteriophage resistome of the intervenients (Davies *et al.*, 2007; Vojtek *et al.*, 2008).

Ongoing acquisition of phages between a recognized pathogen and a largely commensal bacterium may not only have drastic effects on the overall population structure of the genus but also result in rapid changes to the pathogenic potential of SDSE and SDSD (Davies *et al.*, 2007). In fact, studies showed SDSD cells to have high adherence and internalization to human cells, suggesting their ability to infect a human host. This capability seems to be species-specific and independent of the strain-virulence gene content (Roma-Rodrigues *et al.*, 2016).

The subspecies of *S. dysgalactiae* do not represent the only example of complex evolutionary interplay with *S. pyogenes*; since phenomena of functional loss, pathogenic specialization and genetic exchange between *S. equi* subsp. *equi; S. equi* subsp. *zooepidemicus* and *S. pyogenes* have been reported

(Holden *et al.*, 2009; Pelkonen *et al.*, 2013). While *S. equi* subsp. *equi* is host-restricted to horses, *S. equi* subsp. *zooepidemicus* is a known zoonotic pathogen.

3. DISSERTATION PURPOSE AND OUTLINE

This dissertation arises from the collaboration between the Bugworkers group at M&B-BioISI (the Microbiology and Biotechnology unit at the BioSystems & Integrative Sciences Institute) and the Molecular Microbiology group at UCIBIO (Research Unit on Applied Molecular Biosciences). It is a part of the Strep-hosp project (reference: PTDC/CVTEPI/6685/2014), which seeks to unveil host specificity and host-pathogen interactions in *Streptococcus*. The project focuses on isolates of *S. dysgalactiae* subsp. *dysgalactiae* from bovine mastitis and isolates of *S. dysgalactiae* subsp. *equisimilis* from both non-invasive and invasive infections and aims to clarify if the presence of *S. pyogenes* virulence genes in these subspecies (particularly the genes encoded by MGE) contribute to the increased bacterial virulence potential. Ultimately, it aims to understand if these animal associated species are in the process of redefining their host-specificity and if they should be considered as infection agents in humans.

The Strep-hosp project is divided into four tasks: (i) molecular characterization of *Streptococcus* isolates; (ii) detection and characterization of mobile genetic elements in *Streptococcus* isolates; (iii) study of *in vitro* and *in vivo* host-pathogen interactions; (iv) transcriptome and proteome analysis of hosts and pathogens. The present dissertation's purpose was to carry out task (ii), using strains of *Streptococcus dysgalactiae* subsp. *dysgalactiae*, detecting the presence of temperate bacteriophages and ICEs and assessing their genomic structure to determine if they are common to *Streptococcus dysgalactiae* subsp. *equisimilis* and *S. pyogenes*, thus establishing their involvement in horizontal gene transfer.

The existence of a previous Strep project involving this subject as well as its results (project reference: PTDC/CVT-EPI/4651/2012) were used to guide the current work.

CHAPTER II. A quest for lysogeny: screening streptococci for functional lysogenic bacteriophages

1. METHODOLOGICAL INTRODUCTION

Bacteriophage isolation

The most straight forward way to study temperate phages would be to isolate them from their hosts, forcing their excision from the host's genome and their assembly into a virion. Production of phage particles can work as an escape mechanism from adverse conditions. Accordingly, prophage induction can be triggered by DNA damage, changes in pH and temperature, oxidative stress, among other cell stress factors (Cumby *et al.*, 2012). Some of the usual methods utilized in achieving prophage induction include exposure to UV or addition of mitomycin C, an antibiotic first isolated from *Streptomyces caespitosus* that acts as a DNA crosslinker (Levine, 1961; Iyer and Szybalski, 1963; Verweij and Pinedo, 1990). This crosslinking action is lethal, meaning that a single crosslink per genome is sufficient to cause cell death (Tomasz, 1995).

Phages obtained through induction experiments may then be used to re-infect bacterial hosts, augmenting phage concentration for further studies, or to infect other hosts, determining host-range. Infection assays can be carried out in both liquid and solid media, although plaque assays (in solid media), which allow the formation of phage plaques, are considered the gold standard technique. A phage plaque is a zone of lysis disrupting a bacterial lawn on a solid media plate, characteristic of the viral infection process, that corresponds to the replication of a single viral particle. Confluent lysis may also occur, when the area of clearing occupies the entire plate (Abedon and Yin, 2009; Madigan *et al.*, 2015).

One of the most common methods for enumerating and identifying phages is the Double-Layer Agar (DLA) method, introduced by Adams in 1959 (cited in Mullan, 2002) in which a layer of chosen medium with 0,4-0,7% agar is plated on top of the same medium with 1,5% agar. A small volume of phage suspension and host cells are mixed in molten medium and then poured into the basal agar layer; alternatively, the phage suspension can be spotted on top of the host cell and molten agar mix (Mullan, 2002).

Abiotic factor influence in bacteriophage isolation

The advantage of using plaque assay based strategies for detection of bacteriophages is that, when positive, they indicate the phages are not only present but capable of productive infection. Getting phages to form plaques, however, can be a time consuming and arduous task, since there are multiple factors that influence the process of viral infection (Mullan, 2002).

There are multiple condition changes that can be tested: utilizing cells in both logarithmic and stationary growth phase, experimenting with different temperatures, lowering the agar percentage used in growth media, supplementing the media with Ca^{2+} or Mg^{2+}, replacing agar with agarose, choosing growth media free of virus inhibitors and agents that chelate co-factors needed in infection (often present in growth media buffers), using activators of the host's SOS system (such as antibiotics). Gelatin can also be added to the buffer used as diluent for phage solutions to prevent phage surface inactivation – the protein saturates the gas-liquid interface and prevents viral access to the surface. These alterations aim to ease the diffusivity of the phages in solid media, their contact with the host and the overall process of infection, counteracting their frailties. However, numerous other factors, such as the extent of phage-bacterium attachment, the phage's latent period, burst size as well as host density can affect the production of plaques (Mullan, 2002; Abedon and Yin, 2009).

2. MATERIALS & METHODS

2.1 Bacterial strains

For the first tasks of the present work, four strains from *S. pyogenes* and five strains from *S. dysgalactiae* subsp. *dysgalactiae* were used to produce phage lysates. The four SPYO strains (encoded as GAP8, GAP58, GAP88 and GAP826) originate from clinical samples collected from human hosts; three out of the four SDSD strains (VSD5, VSD9 and VSD13) are also of clinical/subclinical origin, and were collected from bovine hosts; one SDSD strain (encoded as GCS-Si) is of clinical origin and was collected from a human host in Singapore, who developed cellulitis upon contact with infected fish (Koh *et al.*, 2009). Moreover, two strains of clinical origin in bovine hosts (VSD17 and VSD19) were used as host cells for infection assays. Strains were selected based on their virulence gene repertoire. For further details, including their performance in infection assays during the first Strep project, can be found on **Appendix A**.

2.2 Growth conditions and culture media

Bacteria were recovered from cryopreserved cultures maintained in THYE - Todd-Hewitt (BD) supplemented with 1% yeast extract (Oxoid) - with 20% (v/v) glycerol at -80°C. To potentiate growth and verify their hemolysis features, 10 µL of the preserved cultures were streaked onto COS (Columbia Agar with Sheep Blood Plus) from Oxoid; inoculated plates were incubated overnight at 37°C. In latter experiments, bacteria for liquid pre-inocula were taken either from the COS plates or from the cryopreserved cultures and added to one of the following culture media: THYE and M17YE (M17 (BD) supplemented with 1% yeast extract). For standard solid plate growth, each medium was supplemented with 1.5% bacteriological agar

(BIOKAR Diagnostics). Although THYE is a standard medium for the growth of streptococci, it contains some of the harmful components for phage infection, namely sodium carbonate and disodium phosphate; in turn, M17YE does not contain either substance (containing disodium-β-glycerophosphate instead, which does not harm the process) and so the two media were used in induction assays.

2.3 Bacteriophage induction assays

For phage induction assays, liquid bacterial cultures from all 8 strains were grown overnight at 37°C in THYE and M17YE. Experimental conditions tested are summarized in **Fig. 7**.

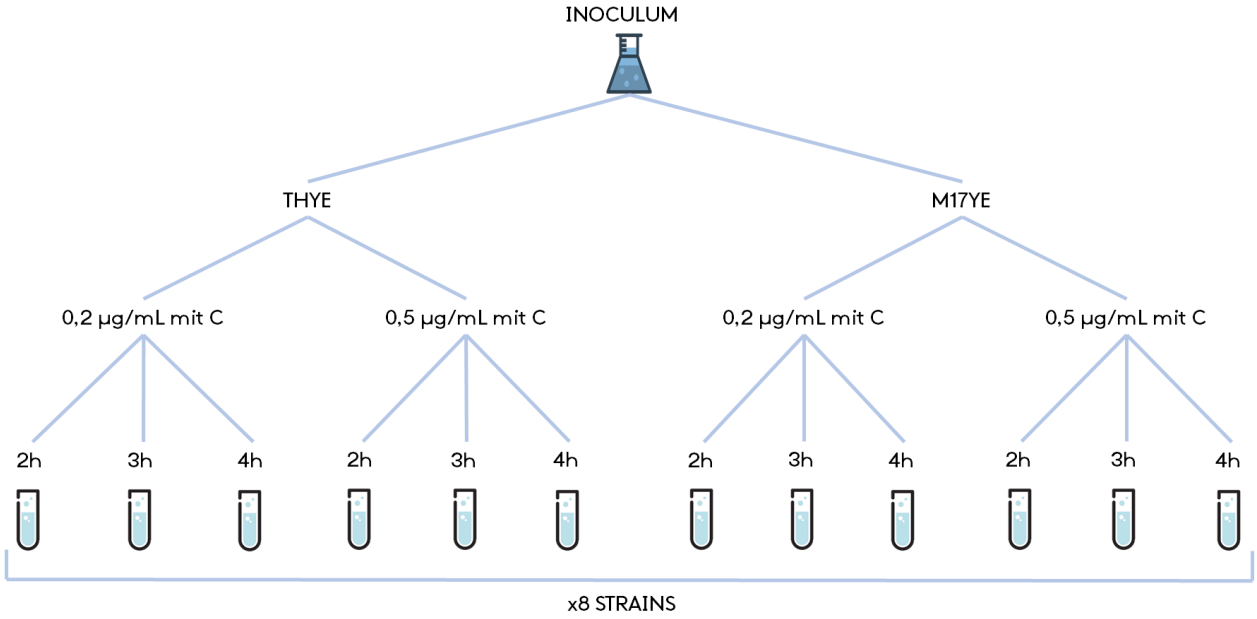


Fig. 7 - **Experimental conditions tested during phage induction assays.** Bacterial cultures from the 8 strains in 2 different culture media (THYE and M17YE) were induced in the early exponential growth phase (OD_{600} ≈ 0.2-0.25) with addition of either 0.2 µg/ml or 0.5 µg/ml of mitomycin C (Sigma-Aldrich) and subsequently incubated, with samples taken at the 2h, 3h and 4h time points.

Overnight cultures were diluted 1:100 in the fresh corresponding culture medium (for a total volume of 20 mL per culture) and allowed to grow until OD_{600} ≈ 0.2-0.25, to ensure induction occurred in the early exponential growth phase. Mitomycin C (Sigma-Aldrich) was then added to each culture to reach a final concentration of either 0.2 µg/mL or 0.5 µg/mL. Cultures were then incubated at 37°C for 4 hours, with samples being collected at the 2h, 3h and 4h time points[2]. Samples were then centrifuged at 1500 × *g* and 4°C for 15 minutes (using an Eppendorf 5810 R centrifuge). The supernatant was collected and filtered using

[2] Concentration of mitomycin C and exposure times were chosen in contrast with the procedure from the first Strep project, in which 1 µg/mL of mitomycin C was used and strains were exposed to the stress agents for 24 hours.

0.45 µm pore membrane filters (Sarstedt) and the resulting filtrate was diluted 1:1 in SM buffer 2x (0.06% gelatin, 20mM NaCl, 16 mM $MgSO_4$, 100mM Tris-HCl) and stored at 4°C.

2.4 Infection assays

Previously obtained phage lysates were subsequently tested through different protocols: spot assays, incorporation assays and cross assays. Spot and incorporation experiments were performed using isolates VSD17 and VSD19 as the hosts for phage infection, based on their virulence gene repertoire (VSD17 contains no phage-associated virulence genes while VSD19 does – thus, it would be expected for VSD17 to be a much more permissive host than VSD19, since the latter may already contain prophages in its genome).

For infection assays, only M17YE medium was used (since it seemed like the most promising) and it was supplemented with 5mM of $CaCl_2$ in both liquid and solid form. In total, four different approaches were performed, divided in two categories: experiments in molten medium and experiments in liquid medium.

2.4.1 Experiments in molten medium

2.4.1.1 Spot assays

Lysates were diluted up to 10^{-4} in SM buffer. Cultures of the VSD17 and VSD19 isolates (host strains) were incubated overnight at 37°C in M17YE. Overnight cultures were then diluted 1:100 in fresh M17YE (for a total volume of 50 mL per culture) and incubated at 37°C until $OD_{600} \approx 0.8$.

In this variation of the Double-Layer Agar method (discussed in **section 1.1** of the present chapter) Plates of M17YE (with 1.5% agar and supplemented with $CaCl_2$) were previously prepared, as well as 5 mL aliquots of molten M17YE (with 0.5% agar and supplemented with $CaCl_2$) which were kept stabilized in a 45°C water bath. 200 µL of the host culture were then mixed with the 5 mL aliquot of molten media, which was then plated upon the correspondent bottom layer 1.5% agar medium and left to dry. In each plate, half of the original lysates for a given strain, along with their respective dilutions, were spotted (each spot corresponding to 10 µL of lysate) in a chess pattern, to avoid contact between spots and left to dry. For each host strain, two control plates were made: a plate containing only the host culture and a plate in which the phage lysate was substituted for a solution of mitomycin C in SM buffer at the highest concentration used in the induction assays (0.5 µg/mL)[3], as depicted in **Fig. 8**.

[3] The mitomycin C control was made to verify that putative lysis plaques were not caused by the residual mitomycin C still present in the lysates and their dilutions.

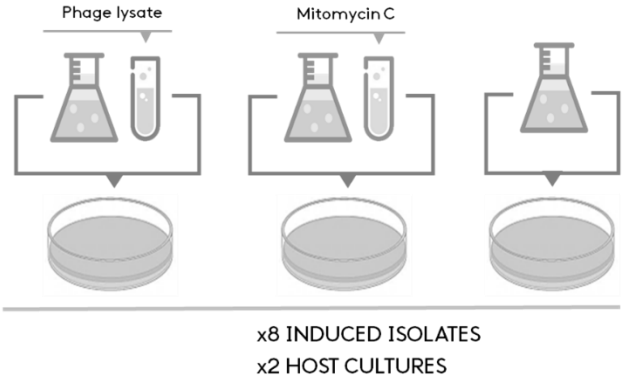


Fig. 8 - **Spot assay experimental scheme**. Plates produced per strain and per host culture.

2.4.1.2 Incorporation assays

Incorporation assays were performed using single-layer agar plates with molten media and only the original phage lysates were tested. Similarly to spot assays, cultures of the VSD17 and VSD19 isolates (host strains) were incubated overnight at 37°C in M17YE. Overnight cultures were then diluted 1:100 in fresh M17YE (for a total volume of 50 mL per culture) and incubated at 37°C until $OD_{600} \approx 0.8$. Meanwhile, 5 mL aliquots of M17YE molten media (with 0.5% agar and supplemented with $CaCl_2$) were kept stabilized in a 45°C water bath. 200 µL of the overnight host culture were then mixed with 10 µL of an original lysate and the 5 mL of culture medium and poured onto a small Petri dish. Controls were the same as those used for the spot assay.

2.4.1.3 Crossed assays

Crossed assays are an "all vs. all" experimental scheme – testing all produced lysates against all possible hosts. As such, instead of using strains VSD17 and VSD19, the 8 strains used in phage induction assays (GAP8, GAP58, GAP88, GAP826, VSD5, VSD9, VSD13 and GCS-Si) were used as hosts. Cultures of the host strains were incubated overnight at 37°C in M17YE. Overnight cultures were then diluted 1:100 in fresh M17YE (for a total volume of 50 mL per culture) and incubated at 37°C until $OD_{600} \approx 0.8$. To reduce the number of plates produced, lysates from the same strain (from the conditions depicted in **Fig. 7**) were mixed in equal parts (mixing a total of 12 lysates, with a volume of 70 µL each). The resulting 8 mixed lysates (ML) were tested against all 8 hosts in an incorporation assay, as described in **section 2.4.1.2** of the present chapter, in which 200 µL of host culture were mixed with 60 µL of the ML and 5 mL of growth medium. The same controls from previous infection experiments were applied.

2.4.2 Experiments in liquid medium

Infection experiments were also carried out in liquid medium, using strains VSD17 and VSD19 as hosts. Mixed lysates (total volume of 1 mL per mixed lysate) for each induced strain were prepared. Liquid bacterial cultures (with a volume of 100 mL) from the two host strains were grown overnight at 37°C in M17YE broth. The overnight cultures were then diluted 1:100 in fresh M17YE broth (total volume of 50 mL per culture) and incubated until OD_{600} ≈ 0.2-0.25. At this point, each culture was infected with a ML and then checked hourly to assess bacterial lysis.

2.5 Phage elution and purification

In case of putative phage plaque formation, isolated plaques were extracted from the plate and placed in 200 µL of SM buffer 1x (0.03% gelatin, 10mM NaCl, 8 mM $MgSO_4$ and 50 mM Tris-HCl). As for plates with possible confluent lysis, the entire plate was flooded with 2mL of SM buffer 1x and left to elute for 4 hours; the liquid was then collected, filtered through a 0.45 µm pore membrane filter (Sarstedt) and stored at 4°C. Resulting phage elutes were then tested in spot assays and incorporation assays.

3. RESULTS & DISCUSSION

Results from the previous Strep project suggested productive infection was possible within this selection of *Streptococcus* strains. However, reproducibility of such results was a problem during the first project, and to assess whether these inconsistencies were due to abiotic factors influencing the infection process, the first phase of this work consisted in testing different induction and infection conditions.

Induction assays occurred as expected, with cultures responding appropriately to the introduction of mitomycin C through OD_{600} reduction. The obtained lysates were then used in different infection assays. Putative phage plaques were detected in all three types of molten media infection assays (as depicted on **Fig. 9**); strains VSD5, VSD9 and GCS-Si had seemingly positive results in more than one type of infection experiment and strain VSD17 was the only host in spot and incorporation assays to register possibly positive results. Plaques were consequently eluted and purified. After purification, putative phage elutes were re-tested through spot, incorporation and cross assays; yet, productive phage infection was never achieved. The seemingly negative results across all attempted approaches suggested that no bacteriophages with plaquing ability were present. However, absence of plaque-forming ability is not necessarily equivalent to absence of a productive infection and broth-based host range determination might help determine whether productive infection is really occurring (Hyman and Abedon, 2010). Assays in liquid growth media were also

performed, but negative results persisted and the putative phages exhibited inability to clear liquid cultures. If present, isolated phages were unable to conduct a productive infection.

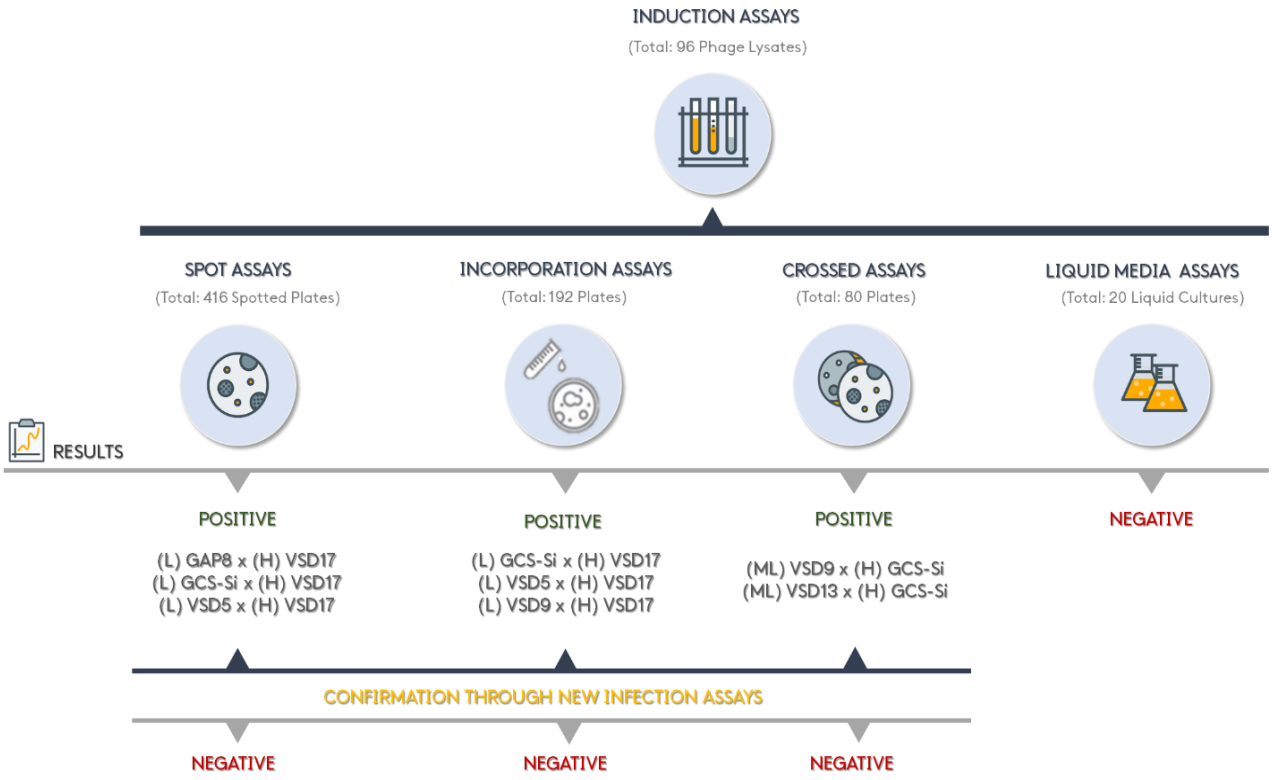


Fig. 9 - **Bacteriophage induction and infection results.** The different types of performed assays are represented along with their yields. For positive results, **(L)** and **(ML)** indicate the strain that produced the lysate or mixed lysate, respectively; **(H)** represents the host strain in which positive results were detected.

The areas of apparent lysis detected in molten media experiments were not due to the residual effect of mitomycin C (given that the control plates did not show lysis), but could be due to the action of other substances, such as bacteriocins. Bacteriocins are small heat-stable peptides common in Gram-positive bacteria; producers of these peptides are often more efficient in host colonization, since they enable the producer to eliminate competitor strains, which may or may not belong to different species (Lux *et al.*, 2007). Streptococci (including *S. pyogenes* and *S. dysgalactiae* subsp. *dysgalactiae*) are prolific producers of these ribosomally synthesized antibiotics (Wescombe *et al.*, 2009), whose presence in a solid medium culture can be confused with phage plaques since bacteriocins also originate clear zones in a plate, similar to phage plaques (Heng *et al.*, 2006; Wescombe *et al.*, 2009).

Spotting assays using putative eluates and their respective dilutions should help differentiating the two: bacteriocins are proteins and their action is concentration-dependent, as such, the lysis area should be maintained in more concentrated lysate solutions and disappear in higher dilutions; viral infection, however,

should still occur in the same manner even when bacteriophages are diluted since they can replicate within the host and multiply. Confirmation using this type of experiment was not conclusive, since no lysis areas were detected at this stage. Because both bacteriocins and phage capsids are proteins, procedures to exclude bacteriocins but keep phage capsids intact are not straight-forward, and thus the nature of the lysis areas detected could not be determined with certainty.

Not all possible conditions were exploited during this series of experiments, leaving certain modifications that remain to be tried. Among these is the replacement of agar with agarose, a purified substance which does not contain agaropectin, a compound with sulphate and carboxyl groups that can inhibit viruses, or other host and virus growth inhibitors (Mullan, 2002; Abedon and Yin, 2009). Additionally, other inducers might be tested, such as hydrogen peroxide, which has proved useful regarding strains of *S. pyogenes*, and fluoroquinolones, which have been successfully used with other pyogenic streptococci (Banks *et al.*, 2003; Ingrey *et al.*, 2003; Brüssow *et al.*, 2004). The latter might prove interesting, since fluoroquinolones are routinely administered to bovines suffering from mastitis, providing similar conditions to those which bacteria are subjected to *in vivo* (Kroemer *et al.*, 2012).

There are also other substances that can be added during plaque assays to enhance phage performance, such as antibiotics (which can also be added in combination with glycerol). Some antibiotics activate the SOS bacterial system, causing cells to divide poorly, increase in size and increase the protein synthesizing system (PSS) activity as well, possibly increasing phage production in turn; as for glycerol, it may increase phage diffusion in the medium, enhancing phage plaque size. The same logic applies to sodium azide and glycine. Thus, any substance or condition that directly or indirectly stimulates an increase of PSS should increase phage production and subsequently plaque size (Santos *et al.*, 2009).

4. CONCLUSIONS

By the end of this chapter, the most likely conclusion would be that there seem to be no infective bacteriophages present in any of the tested strains. Even though not all possible experimental conditions were exhausted during the first phase of this work, this approach proved to be extremely time-consuming and led to the conclusion that obtaining productive lysogenic particles from this collection did not seem to be possible. Yet, it provided no information on whether bacteriophages were or not present in the obtained lysates, for lack of infection productivity does not equal absence of phages. In light of these results, the most suitable approach seems to be the confirmation of both the presence and integrity of phage particles themselves, rather than assessing their functionality through classic infection experiments.

CHAPTER III. The virion: determining genomic and physical integrity of phage particles

1. METHODOLOGICAL INTRODUCTION

Due to the complexity and limitations of infection assays, other methods for detection of bacteriophages, independently of their lysogenic productivity, should also be employed. Methods such as microscopy and extraction of viral DNA can be combined to have a better assessment of the state of phage particles not only in terms of their genome, but also their physical integrity. For example, mishaps during DNA packaging inside the viral capsid originate a seemingly functional virion structure-wise, but render it non-productive due to absence of a complete genome; conversely, while phage particles may appear to have a full genome, abnormalities in physical structure can condemn the infection process. Consequently, determination of integrity at the genome level as well as physical integrity should be paired.

Phage detection methods are constantly evolving: besides more traditional viral DNA extraction techniques, it can also be achieved through powerful microscopy (such as TEM – Transmission Electron Microscopy, or AFM – Atomic Force Microscopy). New microscopy, PCR or genomic-based methodologies, as well as improvements to well established protocols are still proposed regularly (Mullan, 2002; Anderson *et al.*, 2011).

Microscopy-based phage detection

Microscopy-wise, visualization of phage particles using TEM is considered the gold standard technique. However, bacteriophage preparation methods for TEM viewing involve adsorbing previously purified samples to a carbon-coated copper grid, allowing them to dry and performing negative contrast with either methylamine tungstate or uranyl acetate. The purification of the samples often implies the use of ultra-centrifugation based methods (such as CsCl density-gradient centrifugations), which may physically disrupt frail bacteriophages (Beniac *et al.*, 2014). Consequently, using a type of microscopy that does not require such intricate preparation methods, which are in turn more likely to produce image artifacts, may prove beneficial.

Atomic Force Microscopy

Atomic force microscopy belongs to the broad family of scanning probe microscopes which use a proximal probe to investigate properties of surfaces with subnanometre resolution. At first, what was considered the main improvement of AFM was its much higher imaging resolution in comparison to optical

microscopy, but the possibilities of spectroscopic analysis, surface modification and molecular manipulation opened an entire new realm of possibilities for AFM use (Alessandrini and Facci, 2005).

As for biological applications, the most appealing advantage of this type of microscopy over TEM and SEM (Scanning Electron Microscopy) is the fact that it allows measurements of native biological samples in physiological-like conditions, simplifying the sample preparation process and avoiding preparation-related image artifacts. Biological samples studied through AFM range from phospholipids, proteins, DNA, RNA, to subcellular structures, living cells and tissues (Alessandrini and Facci, 2005).

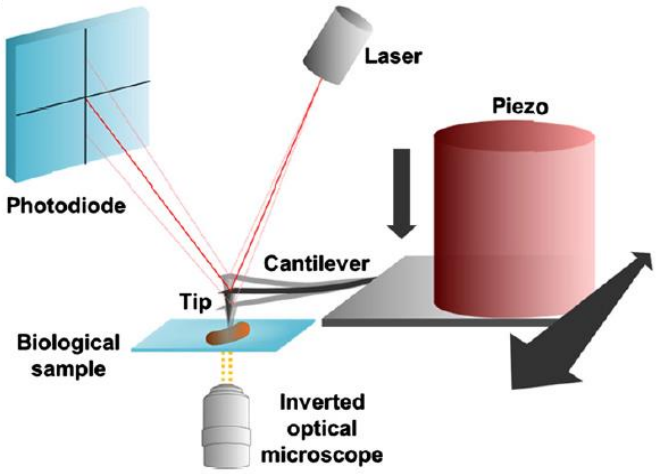


Fig. 10 - **Functioning scheme of AFM system.** Coupling atomic force microscopes to inverted optical microscopes is optional, but frequent. Source: Pillet *et al.*, 2014

Unlike lens-based technologies, scanning probe microscopes rely on the measure of a parameter between a sharp tip and a surface – AFM relies on measurements of force. The setup consists of a micro-machined cantilever probe and a sharp tip mounted to a Piezoelectric (PZT) actuator[4] and a position-sensitive photodetector (the photodiode referenced in **Fig. 10**) receiving a laser beam reflected off the end-point of the beam, providing cantilever deflection feedback. The principle of AFM is to scan the tip over the sample surface, at sub-Ångström accuracy, with feedback mechanisms that enable the PZT scanners to maintain the tip at a constant force or constant height above the sample surface. As the scanning occurs, the tip moves up and down according to the contour of the surface, and the laser beam deflected from the cantilever provides measurements of the difference in light intensities between upper and lower photo detectors. It is then the feedback from the photodiode difference signal that, through software control in an associated computed, enables the tip to maintain constant force, upholding the principle of AFM.

[4] A piezoelectric actuator converts an electrical signal into a precisely controlled physical displacement. If this displacement is prevented, a blocking force will develop, which is then utilized in AFM (Muralt, 2000).

The amount of feedback signal measured at each point allows to form a 3D reconstruction of the sample topography, which is usually displayed as an image (Jalili and Laxminarayana, 2004; Alessandrini and Facci, 2005; Pillet *et al.*, 2014). AFM has three main operational modes: contact mode, non-contact mode and tapping mode, based on how the tip interacts with the sample, as depicted in **Fig. 11** (Jalili and Laxminarayana, 2004).

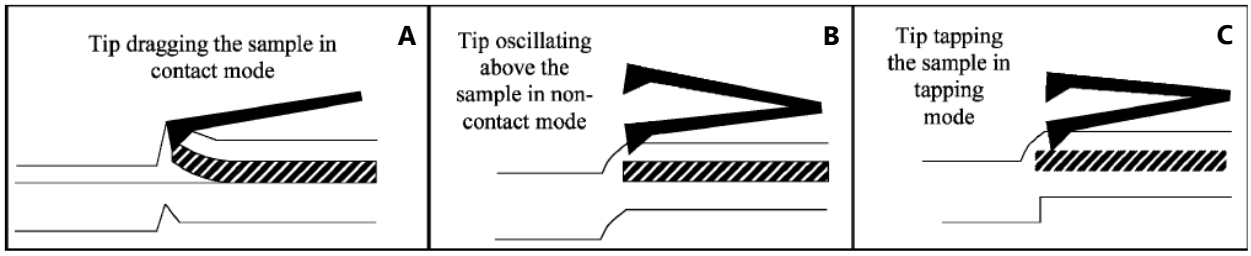


Fig. 11 - **Main AFM operational modes.** (A) contact mode or repulsive mode, in which the tip is in close contact with the sample and measures mainly repulsive van der Waals forces; (B) non-contact mode, in which the cantilever tip hovers 50-150 Å and detects attractive van der Waals forces between the tip and the sample; and (C) tapping mode, in which the cantilever is oscillated near its neutral resonant frequency and is then deviated according to the sample topography, lightly tapping the sample – the values are then compared with the set reference value and the "error signal" is used to represent the topography. Source: Jalili and Laxminarayana, 2004

2. MATERIALS & METHODS

2.1 Bacterial strains

For the second phase of the present work, infection assay results from the Strep project were revisited and strains with diverse viral infection profiles, as well as virulence gene repertoires, were selected. Because the main focus of the project is to investigate HGT from *S. pyogenes* to *S. dysgalactiae* subsp. *dysgalactiae*, the presence of phage particles extracted from SDSD isolates takes priority. Consequently, 4 SDSD strains were carried over from the first phase of this work - VSD13, VSD17, VSD19 and GCS-Si – and a new strain was added - VSD4. This new strain is similar to other "VSD" encoded ones, in that it is also of clinical/subclinical origin, and was collected from a bovine host. Further details on these strains can be found on **Appendix A**.

A strain of *Escherichia coli* (*E. coli* K12 MG1655) was used, along with the T7 bacteriophage, as a positive control for these experiments.

2.2 Growth conditions and culture media

SDSD strains were recovered from cryopreserved cultures maintained in THYE - Todd-Hewitt (BD) supplemented with 1% yeast extract (Oxoid) - with 20% (v/v) Glycerol at -80°C. For subsequent experiments,

strains were incubated overnight at 37°C in M17YE broth or M17YE agar (supplemented with 1.5% Bacteriological Agar (BIOKAR Diagnostics)).

The *E. coli* strain was recovered from cryopreserved cultures maintained in NB (Nutrient Broth (BIOKAR Diagnostics)) with 20% (v/v) glycerol at -80°C. For subsequent experiments, the strain was incubated overnight at 37°C in NB or NA (Nutrient Agar – NB supplemented with 1.5% bacteriological agar (BIOKAR Diagnostics)).

2.3 Modified phage induction assay

For phage induction assays performed in this stage, liquid bacterial cultures from all 5 SDSD strains were incubated overnight at 37°C in M17YE broth. Overnight cultures were diluted 1:100 in fresh M17YE (for a total volume of 400 mL per culture) and allowed to grow until $OD_{600} \approx 0.2$-0.25, to ensure induction occurred in the early exponential growth phase. Mitomycin C (Sigma-Aldrich) was then added to each culture to reach a final concentration of 0.5 µg/mL and cultures were then incubated overnight at 37°C to allow lysis[5]. Crude lysates obtained from this procedure were then used for phage DNA extraction and AFM sample preparations.

For the *E. coli* strain, a similar procedure was followed with the adequate culture medium, but instead of mitomycin C, 100 µL of a highly concentrated T7 phage solution were added.

2.4 Bacteriophage DNA extraction

For DNA extraction, 200 mL of the crude lysate obtained in **section 2.3** were treated with DNase I (Sigma-Aldrich) and RNase A (Sigma-Aldrich) with final concentrations of 5 µg/mL and 2 µg/mL, respectively, and incubated for 2 h at 37°C. Then, NaCl (Duchefa Biochemie) was added to a final concentration of 1M and lysates were agitated and incubated in ice for 1 h. Cell residues were deposited through centrifugation: 15000 × *g* and 4°C for 45 minutes (using a Beckman J2-21 centrifuge equipped with the Beckman JLA-16.250 rotor) – and the supernatants were transferred to new tubes. Phages were then concentrated by precipitation with 10% (w/v) PEG8000 (Sigma-Aldrich) overnight at 4°C. After centrifugation (15000 × *g* and 4°C for 25 minutes), the resulting pellet was resuspended in 5 mL in SM buffer (0.03% gelatin, 10mM NaCl, 8 mM $MgSO_4$ and 50 mM Tris-HCl). PEG was extracted by adding an equal volume of a 1:1 phenol/chloroform mixture (Sigma-Aldrich) and centrifuging at 4020 × *g* and 4°C for 15 minutes (using an Eppendorf 5810 R centrifuge). The aqueous phase was transferred to a new tube and to it were added: SDS to a final

[5] Considering the results from Chapter II, mitomycin C concentration and time of exposure were adjusted to intermediate levels between these results and those obtained in the first Strep project.

concentration of 0.5%, EDTA pH 8.0 to a final concentration of 0.02 mol/L and proteinase K (Invitrogen) to a final concentration of 0.05 mg/mL. Lysates were then incubated at 37°C for 1 h. Phenol extraction was performed by adding 1 vol. of a 1:1 phenol/chloroform mixture, centrifuging at 4020 × *g* and 4°C for 15 minutes, then adding 1 vol. of a 24:1 chloroform/isoamyl alcohol mixture (Carlo Erba Reagents) and centrifuging again at 4020 × *g* and 4°C for 15 minutes. Subsequently, phage DNA was mixed with 1 vol. of isopropanol and left to precipitate overnight at 4°C. In the following day, the samples were centrifuged at 3000 × *g* and 4°C for 10 minutes, washed with 70% (v/v) ethanol and resuspended in 50 µL of TE buffer (10mM Tris, 1 mM EDTA; pH 8.0).

2.5 DNA agarose gel electrophoresis

Phage DNA (30 µL of each sample) was then submitted to electrophoresis in a 0.8% (w/v) agarose (Invitrogen) gel, with 0.5X TBE buffer (40 mM Tris; 45 mM Boric acid; 1 mM EDTA; pH 8.3) and a constant voltage of 4 V/cm for 1 h. The gel was stained with Ethidium bromide and revealed in an Alliance 4.7 UV transilluminator (UVItec) and the image retrieved using the Alliance software. The molecular weight marked used was a "1kb DNA Ladder" (Invitrogen) and purified λ phage DNA (Invitrogen) was also used as a reference.

2.6 DNA quantification

DNA quantification was performed using the Qubit 2.0 Fluorometer (Invitrogen) with the dsDNA High-Sensitivity Kit - suitable for samples expected to have between 0.2-100 ng of DNA - and according to the manufacturer's instructions. The volume of sample dispensed for Qubit quantification was 1 µL.

2.7 AFM sample preparation

2.7.1 Precipitation of phage particles

AFM sample preparation consists of executing the phage DNA extraction protocol up until the resuspension in SM buffer: 200 mL of the crude lysates obtained in **section 2.3** were treated with DNase I (Sigma-Aldrich) and RNase A (Sigma-Aldrich) with final concentrations of 5 µg/mL and 2 µg/mL, respectively, and incubated for 2 h at 37°C. Then, NaCl (Duchefa Biochemie) was added to a final concentration of 1M and lysates were agitated and incubated in ice for 1 h. Cell residues were deposited through centrifugation - 15000 × *g* and 4°C for 45 minutes (using a Beckman J2-21 centrifuge equipped with the Beckman JLA-16.250 rotor) – and the supernatants were transferred to new tubes. Phages were then concentrated by precipitation with 10% (w/v) PEG8000 (Sigma-Aldrich) overnight at 4°C. After centrifugation (15000 × *g* and

4°C for 25 minutes), the resulting pellet was resuspended in 500 µL in SM buffer (0.03% gelatin, 10mM NaCl, 8 mM $MgSO_4$ and 50 mM Tris-HCl). To facilitate the process, AFM sample preparation and phage DNA extraction were done in parallel.

2.7.2 Preparation for AFM visualization

Atomic Force Microscopy was carried out in a Multimode 8 HR produced by Bruker, using Peak Force Tapping mode. All measurements were performed by placing a drop (ca. 50 µL) of each sample onto freshly cleaved mica for 20 min, rinsing with ultrapure water and drying with pure N_2. The images were acquired in ambient conditions (ca. 21°C), using etched silicon tips with a spring constant of ca. 0.4 N/m (SCANASYST-AIR, Bruker), at a scan rate of about 1.3 Hz.

3. RESULTS & DISCUSSION

3.1 Genomic integrity

After provoking bacterial lysis, Phage DNA extraction was performed and its product submitted to electrophoresis, as seen in **Fig. 12**. The first steps of the DNA extraction protocol exclude bacterial DNA and RNA (through DNase and RNase treatment) without affecting viral DNA, since it should still be inside the protein phage capsid. After elimination of bacterial residues and phage precipitation with PEG8000, the capsids are destroyed and phage DNA extraction follows.

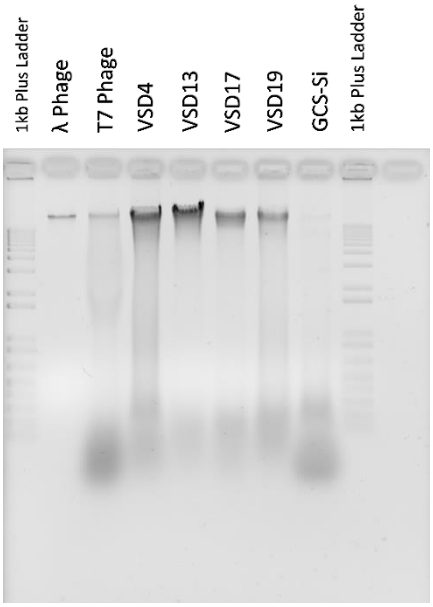


Fig. 12 - **Electrophoresis of phage DNA samples.** Purified λ phage DNA was used as a control for size and expected fragment aspect; the T7 phage DNA was extracted using the same procedure applied to the VSD strains and served as a control for the success of the protocol.

The presence of gel bands similar to the λ phage DNA and the T7 phage DNA, suggests that induction experiments were successful and the SDSD strains do contain prophage sequences integrated in their genomes. Furthermore, it suggests that they are capable of excision from the bacterial genome and successful encapsidation. SDSD phage DNA appears to be less defined than that of the T7 phage. This may be due to the difference in earlier protocols: to obtain *E. coli* lysates containing the T7 phage, an otherwise phage-free bacterial culture was infected with a concentrated T7 stock, meaning only fragments corresponding to the T7 genome can be recovered; as for SDSD strains, their phage repertoire is unknown, and the presence of several integrated prophage sequences somewhat close in size can explain the initial dragging observed in these gel bands. As for smears observed in the lower section of the gel, they might be due to insufficient RNase treatment in the first steps of DNA extraction.

Phage DNA samples were also quantified using the Qubit Fluorometer, and suggested that even though the GCS-Si lysate appeared negative in **Fig.12**, some DNA may be present, as assessed in **Table 1**. To assess whether the smears present in the gel were indeed due to RNA presence, T7 and VSD17 samples (with larger and smaller smears, respectively) were used to perform a Qubit RNA Quantitation assay. Both samples contained RNA: the T7 phage sample registered a concentration of 36,8 ng/μL and the VSD17 sample registered a concentration of 18,24 ng/μL.

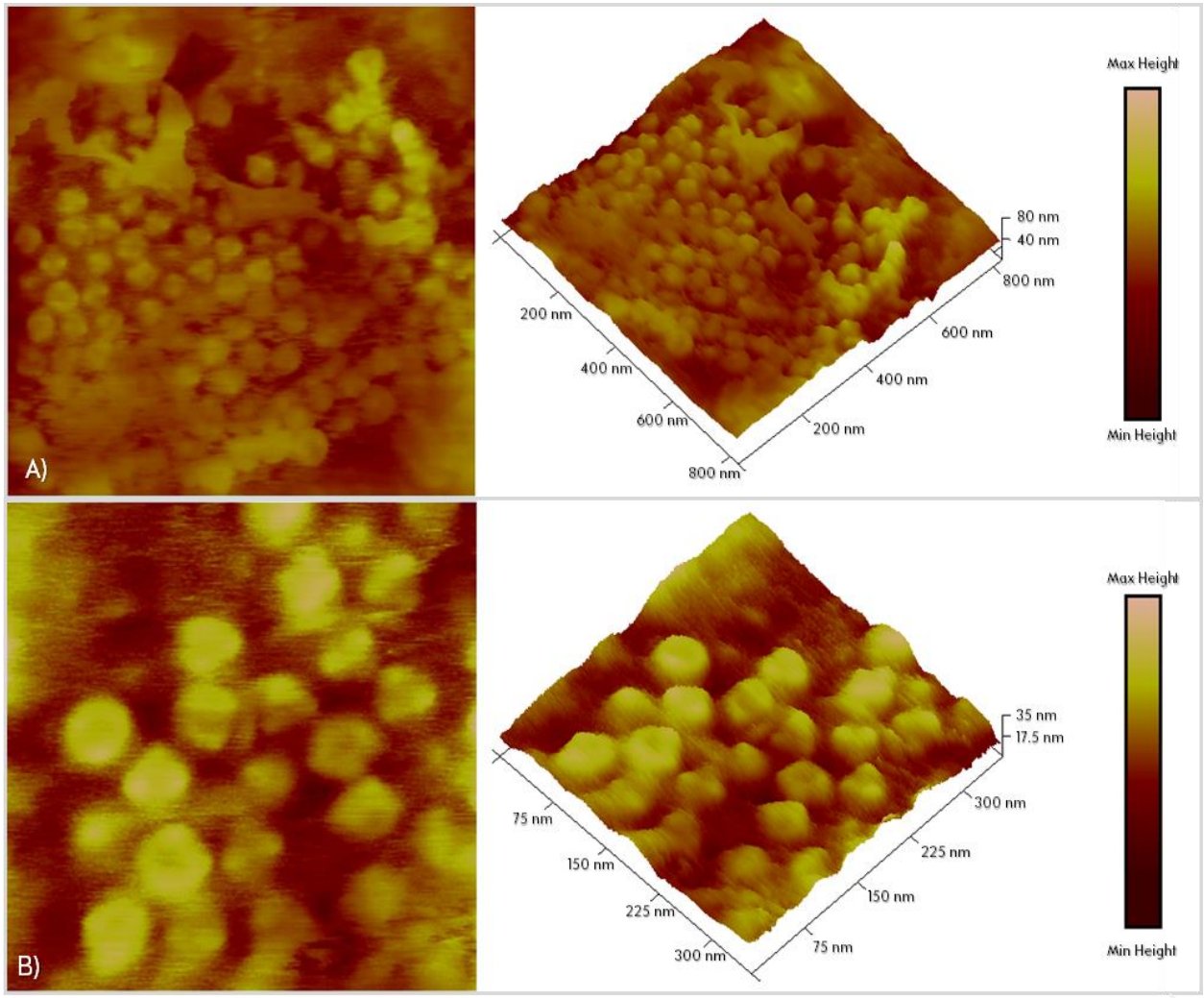
Table 1 - Phage DNA quantitation results.

Phage DNA Sample	Quantitation
T7 phage	33,2 ng/μL
VSD4	150,8 ng/μL
VSD13	71,6 ng/μL
VSD17	26,2 ng/μL
VSD19	31,6 ng/μL
GCS-Si	21 ng/μL

Although these results confirm phage presence in SDSD strains, with phage genome fragments appearing within the expected size of 50 kb (average sizes of *Siphoviridae* members), the phages' physical integrity remains unknown. While encapsidation and capsid functionality are required for recovery of phage genome fragments through this method, it provides no information on other vital components for viral infection, such as the virion tail for example.

3.2 Physical integrity

To assess the physical structure of phage particles, PEG precipitated samples were viewed using Atomic Force Microscopy. Although very advantageous to this particular case, AFM can be somewhat time-consuming and several adjustments to the drying steps of sample preparation must be made, as well as experimenting with diluting the samples in more appropriate buffers and/or perform multiple washing steps. This is crucial because attempting to view samples that are very concentrated and rich in background components (as is the case for these phage lysates) can damage the tip used to engage the surface and quickly increase the costs of this process. To this end, three out of the six lysates were chosen to undergo AFM: the T7 phage lysate (which served as a positive control), and the VSD13 and VSD17 lysates (which had intermediate concentration values expected to be more suited for this technique). Results of 2D image capturing as well as rendering of 3D images are shown in **Fig. 13.**



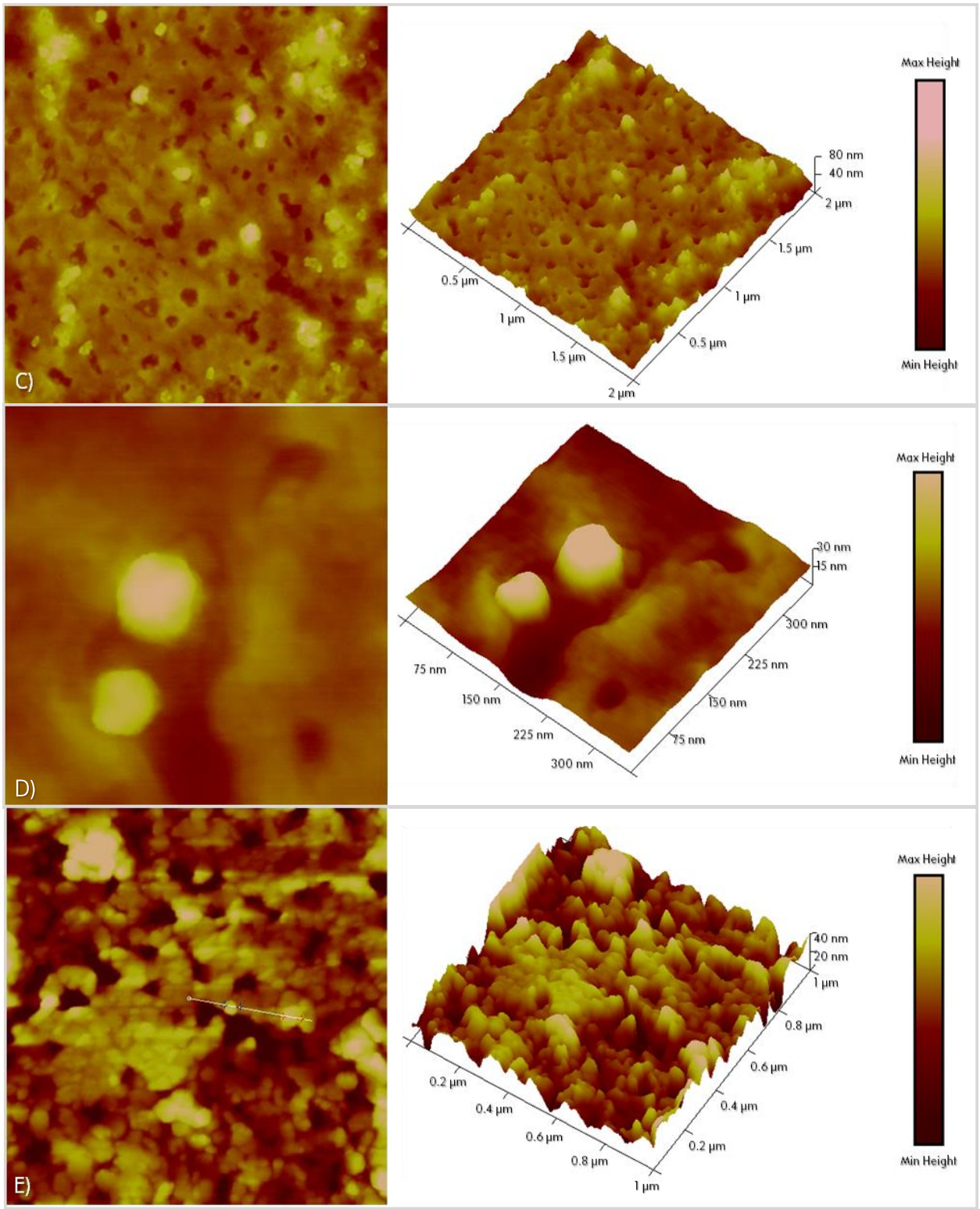


Fig. 13 - **AFM 2D and 3D images.** Sets **A)** and **B)** correspond to the VSD13 sample; sets **C)** and **D)** correspond to the VSD17 sample; set **E)** corresponds to the T7 phage control sample. The indicated section in E) 2D image was used for size comparison between samples.

Globular structures peaking in height were detected across all samples. To assess their size, several sections from each 2D image were inspected and measured using the NanoScope Software (further details on measurements can be found on **Appendix B**). These structures were consistent in size (averaging at about 60 nm in diameter) as well as morphology and their abundance in the samples seemed to reflect that of phage DNA. Because the samples are mounted onto a hydrophilic surface, slight deviations from the canonic icosahedral structure and TEM-obtained dimensions (expectable capsid diameters are around 50 nm, although sizes do vary) are predictable – the adherence of phage capsid proteins to the surface may cause them to appear larger and to lose their shape. The long period of exposure to PEG (a highly hydrophilic compound) the samples were subjected to can also affect capsid shape.

Irregularities in the background are due to the complexity of the sample, which still contains leftover culture medium, PEG8000 and SM buffer. Proteins and other compounds present will adhere to the hydrophilic support and create irregularities in the surface. Although washing steps (applied to sample VSD13, represented on sets **A)** and **B)** of **Fig. 13**) did contribute to eliminate this effect, dilution of samples in a cleaner buffer is advised, to both adjust concentration and get rid of background irregularities, resulting in clearer images.

Phage tails could not be observed in any of the samples submitted to AFM. While for the T7 phage this could just be due to its morphology - T7 is a member of the *Podoviridae* family, characterized by very small non-contractile tails – the same does not apply to SDSD samples, given that streptococci are most commonly infected by *Siphoviridae* phages, with long flexible non-contractile tails. Phage tails have been observed before through AFM, albeit in much more purified samples with no need to undergo PEG precipitation (Ivanovska *et al.*, 2007; Arkhangelsky and Gitis, 2008; Kuznetsov *et al.*, 2013; Szermer-Olearnik *et al.*, 2017). Due to scheduling constraints, samples did not undergo AFM shortly after their preparation, and as such the prolonged exposure to PEG may also have tempered with phage tail integrity.

4. CONCLUSIONS

Although presence of infective bacteriophages could not be assessed from **Chapter II** results, bacteriophages do seem to be present among SDSD strains. Moreover, there appear to be phage genomes able to not only replicate but carry out integration, excision, capsid formation and encapsidation in a successful manner.

Absence of phage tails in AFM images could either be an artifact caused by the lack of sample purification procedures and prolonged exposure to PEG or by a genomic abnormality rendering bacteriophages uncappable of synthesizing or correctly assembling tails. Even assuming that virions are

indeed intact, the answer as to why these viral particles are incapable of conducting successful infection may still lie in a genomic approach, by looking not only at the phages' genomes, but also their bacterial counterparts.

CHAPTER IV. The prophage state: mining bacterial genomes for integrated phage sequences

1. METHODOLOGICAL INTRODUCTION

The key to the unproductive infection continuously verified throughout **Chapter II** does not seem to lie with physical frailty or assay conditions, but may be related to phage defectiveness. Whether it indeed lies within the tail modules or is related to other factors, looking at the phage in its prophage state – still inserted in the host genome – seems to be the most promising option. Although still viewed as a daunting task, genome sequencing has evolved greatly since Sanger sequencing platforms[6] both in terms of its throughput and accessibility, originating the plethora of diverse methodologies now known as NGS (Next Generation Sequencing) (Goodwin *et al.*, 2016).

1.1 Next-generation sequencing platforms

Over the last 15 years genome sequencing technologies have evolved greatly, from the sequencing of short oligonucleotides to millions of bases, enhancing the diversity and number of sequenced genomes and decreasing sequencing cost per megabase. This allows NGS platforms to provide considerable quantities of data in comparison to first-generation sequencing (traditional Sanger sequencing), although not without disadvantages. NGS also competes with alternative technologies, such as DNA microarrays, qPCR, optical mapping (combining long-read technology with low-resolution sequencing) and NanoString (a technology relying on target-probe hybridization with labelled molecules bound in a discrete order) (Goodwin *et al.*, 2016; Heather and Chain, 2016).

Because NGS encompasses such a diverse group of technologies, it can be divided in short-read NGS (or second-generation sequencing) and long-read NGS (or third-generation sequencing), as summarized in **Fig. 14**. These two sections of NGS have dramatically different properties, with second-generation being associated with detection of clonally amplified DNA and third-generation associated with single-molecule detection. Transversely to sequencing generations, methodologies can also be divided based on whether they use optical (Illumina, Pacific Biosciences, Roche 454) or non-optical detection (Ion Torrent and Oxford Nanopore) of signals to perform base-calling. Different technologies often complement

[6] Sanger sequencing is an approach that relies on the mix of dye-labelled deoxynucleotides (dNTPs) and dideoxy-modified dNTPs. This modification halts the incorporation of any other nucleotide and thus, when a PCR reaction is carried out, the incorporation of a dideoxy-dNTP terminates elongation. Resulting strands are then separated on gel and the terminal base is identified by laser excitation and spectral emission analysis (Goodwin *et al.*, 2016).

each other and usage of platforms from different generations in a single experiment is now commonplace (Goodwin *et al.*, 2016; Levy and Myers, 2016). Although additional technologies exist (both current niche technologies as well as already extinct platforms), for the purpose of this introduction only the major NGS technologies were considered.

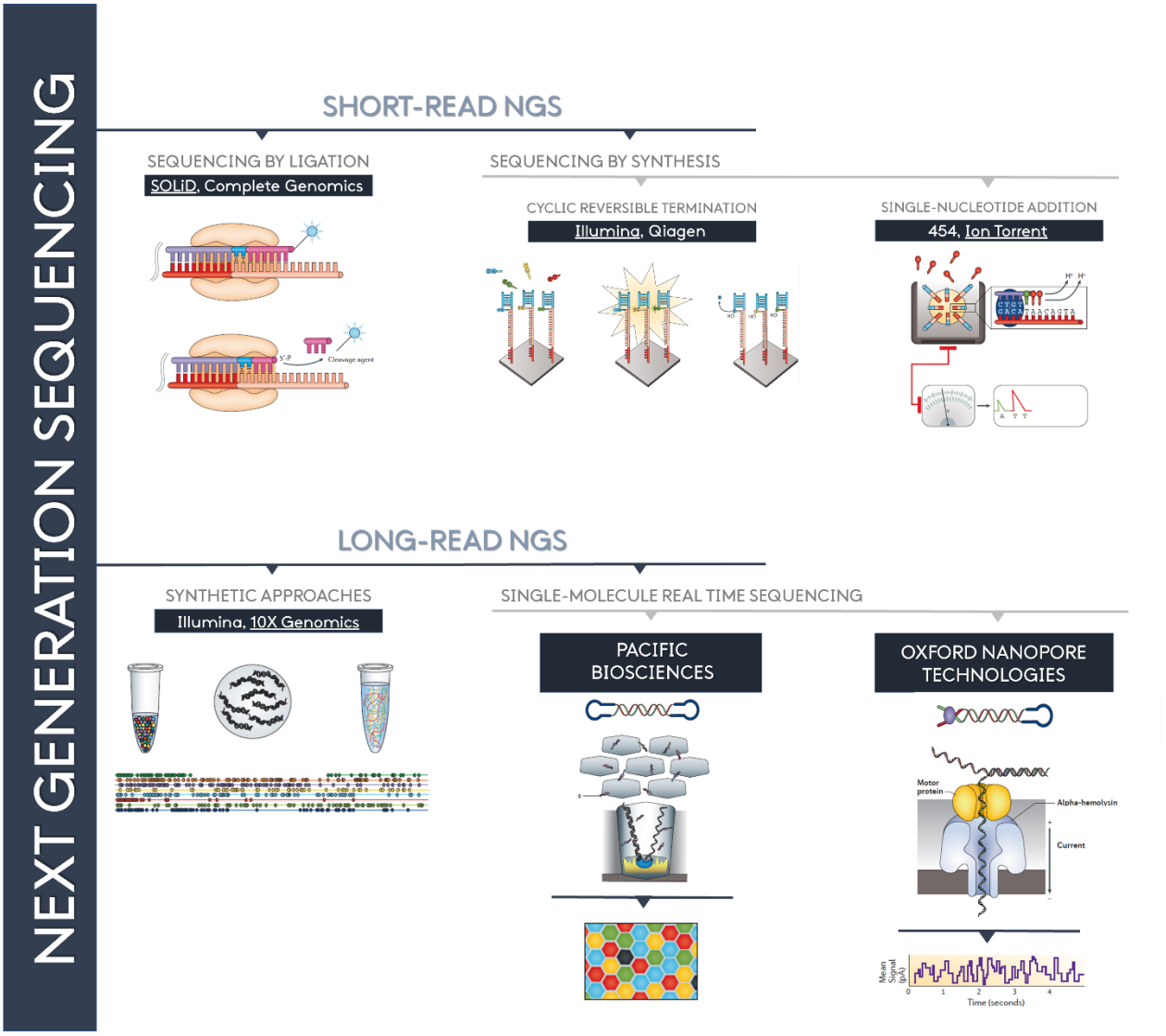


Fig. 14 - Overview of Next Generation Sequencing methods. In methods with more than one associated technology/company, illustrated examples correspond to the underlined option. Adapted from: Goodwin *et al.*, 2016.

1.1.1 Short-read NGS

Short-read NGS, or second-generation sequencing, represents the first wave of progress within NGS, departing from the inference of nucleotide identity through radio- or fluorescence-based labeling of dNTPs and oligonucleotides. Additionally, NGS usually allows visualization in real time, instead of using

electrophoresis-based methods which were standard for first-generation sequencing (Heather and Chain, 2016).

Second-generation approaches can themselves be divided into two categories: sequencing by ligation (SBL) and sequencing by synthesis (SBS). SBL methodologies rely on the binding of a probe (with a couple of known bases followed by degenerate or universal bases) to a fluorophore, hybridization of the probe-fluorophore complex to a DNA fragment and subsequent cleaving of the fluorophore, originating a signal whose emission spectrum allows the determination of the bases complementary to the probe's known nucleotides. After cleavage of the fluorophore, subsequent probes are added until complete hybridization of the fragment – finishing a round of probe extension – and then the fragment is reset, initiating a new round of probe extension with either an (n+1) or (n+2) offset from the previous round. Offset rounds help build coverage and increase confidence in sequencing results (Goodwin *et al.*, 2016; Levy and Myers, 2016).

SBS approaches rely on the action of a polymerase and detection of nucleotide incorporation into an elongating strand, either by fluorophore signaling or changes in ionic concentration. Sequencing by synthesis can be achieved either through Cyclic Reversible Termination (CRT) or Single Nucleotide Addition (SNA). CRT methodologies use terminator molecules that block the ribose 3'-OH group, preventing elongation in a similar way to Sanger sequencing. SNA, on the other hand, does not force termination, but rather operates in an iterative way, adding only one type of nucleotide at a time and marking the incorporation of a single dNTP into an elongating strand – thus, elongation stops simply due to the absence of the following nucleotide. In homopolymer regions, where more than one nucleotide of the same type will be added at once, identification is achieved through detection of proportional signal increases (Goodwin *et al.*, 2016).

Both SBL and SBS require the clonal amplification of DNA, given that having a high number of DNA copies enhances the distinction of the signal from background noise. Generation of these clonal template populations can be achieved through three strategies: bead-based (using emulsion PCR), solid-state (amplification directly on a slide) or DNA nanoball (template enrichment in solution) clonal template generation (Goodwin *et al.*, 2016).

It becomes clear that, although usually regarded as a high-fidelity and short-read group of approaches, second-generation sequencing encompasses a remarkable diversity of methodologies, varying in terms of their chemistry, capabilities and specifications (with some methods reaching 600 bp in read-length or 99,99% accuracy). Still, certain downfalls regarding Sanger sequencing, such as a higher error rate and reads shorter than the 700 bp achievable by Sanger, as well as the difficulty in resolving homopolymer regions, cannot be overlooked. Furthermore, although parallel use of different short-read methodologies is

often employed, the Illumina technology dominates the second-generation market as the most well established option, offering reads up to 300 bases and an average accuracy of 99,50% in platforms with variable throughputs. Furthermore, Illumina provides paired-end sequencing, allowing the sequencing of both ends of each DNA fragment, generating alignable sequence data that directly improves the quality of the dataset (Reuter *et al.*, 2015; Heather and Chain, 2016).

1.1.2 Long-read NGS

One of the main applications of NGS is whole-genome sequencing (WGS). Genomes are quite complex, containing long repetitive elements and structural variations that directly impact the evolution and adaptation of an organism. The length and complexity of these structural features can be a challenge even to paired-end sequencing, making *de novo* genome sequencing one of the greatest shortcomings in second-generation sequencing (in addition to secondary structures and modified or non-canonical bases). Longer-reads, capable of spanning over these problematic regions, may help increase the accuracy of WGS as well as improving transcriptomic research, and thus third-generation sequencing was born. Long-read NGS can be performed using *in silico* approaches or single-molecule real-time sequencing (SMRT sequencing), and unlike short-read NGS, it does not require chemical cycling for each dNTP nor does it depend on the on a clonal amplified DNA population to generate detectable signals (Goodwin *et al.*, 2016).

The *in silico* approaches, or synthetic long-read technology, are not true sequencing systems but rather utilize existing short-read sequencers along with barcoding systems. DNA fragments are distributed in partitions, sheared, barcoded and sequenced using second-generation technology; barcoding facilitates the process of assembly, given that fragments with the same barcode must be derived from the same long fragment, and allows the generation of long reads *in silico* (Goodwin *et al.*, 2016; Levy and Myers, 2016).

SMRT sequencing allows actual generation of reads with thousands of bases per read. It is dominated by two companies with different detection methodologies: Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). PacBio relies on the optical detection of a sequencing-by-synthesis reaction while ONT performs detection through nanopores (Levy and Myers, 2016).

The PacBio technology, which is the most widely proven among long-read methods, is represented in **Fig. 15**, generating reads that average at over 10,000 nt and can exceed 40,000 nt with a per-base error rate close to 15%, mitigated by the generation of consensus sequences. Additionally, and in theory, the errors (mostly indels) are randomly distributed within reads, which allows them to be overcome by a high enough coverage (Goodwin *et al.*, 2016; Levy and Myers, 2016).

This strategy involves generating a capped template termed SMRT-bell: this is achieved by ligating single-stranded hairpin adapters onto both ends of a digested molecule of either DNA or cDNA. Because there are hairpins at either end, making the template circular, the original DNA molecule can be sequenced several times by using a strand displacing polymerase; this way, native (and potentially modified) DNA can be directly sequenced. It is this circularization that allows the increase of accuracy up to 99.90%. DNA synthesis is carried out in microfabricated nanostructures called zero-mode waveguides: zeptoliter-sized chambers with a single polymerase immobilized at the bottom. These structures are meant to reduce background noise in optical detection by making the zone of detection extremely small, ensuring only the polymerase is illuminated by light diffusion. With no forced deterrence of sequencing needed, polymerization occurs continually and fluorescent signals can be read in real-time. Time of residence of phospholinked nucleotides in an active site depends on the rate of catalysis; thus, recorded fluorescent pulses tend to be on the millisecond scale, allowing only the bound nucleotide to occupy the zero-mode waveguide detection zone and making the signal more reliable. The polymerase then cleaves the fluorophore and allows it to diffuse away from the detection area, clearing the signal before the next dNTP is incorporated (Metzker, 2010; Reuter *et al.*, 2015).

This method allows for the discrimination between methylated and unmethylated versions of the same base, as well as between methylated cytosine and methylated adenine, based on the polymerase's timings during elongation – modified sites force the polymerase to pause for longer, increasing interpulse duration and indicating the presence of a modified base (Flusberg *et al.*, 2010; Goodwin *et al.*, 2016).

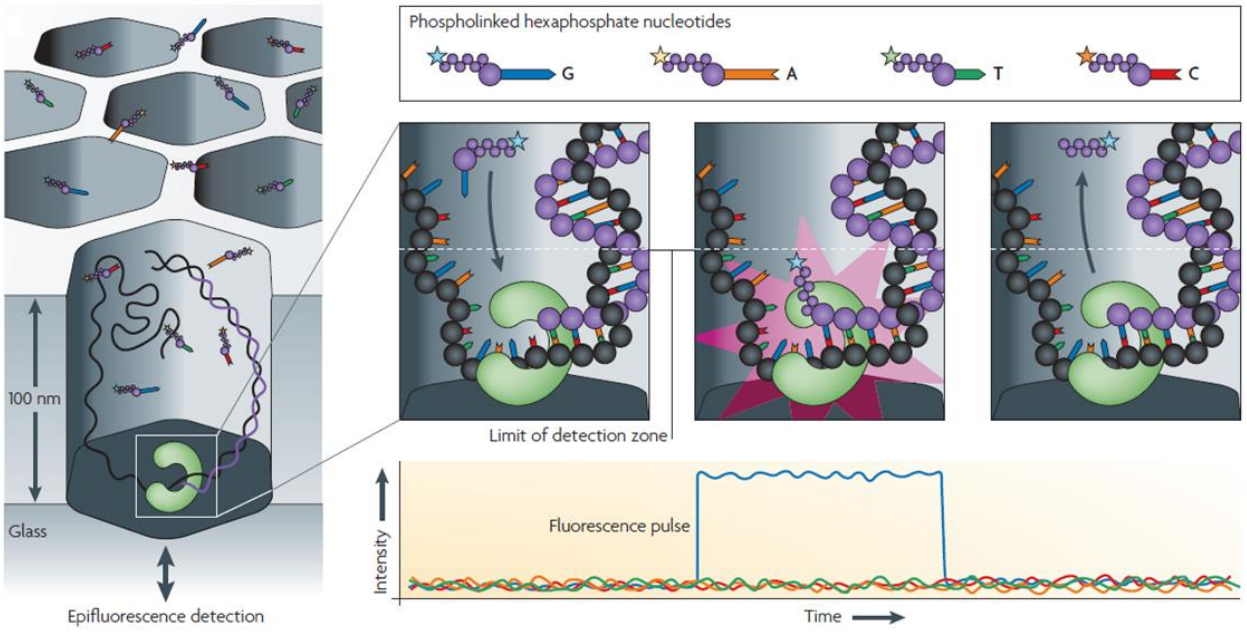


Fig. 15 - The PacBio SMRT sequencing methodology. On the left, the zero-mode waveguide is depicted while the right represents the incorporation and signaling of fluorescent dNTPs. Source: Metzker, 2010.

Oxford Nanopore Technologies represents perhaps the most disruptive technology within the NGS group, using sequencers based on nanopore biosensors. These biosensors can be divided into solid-state pores and biological pores: solid-state pores are fabricated from diverse materials using semiconductor production processes, allowing them to work in several experimental conditions; biological nanopores consist of transmembrane protein channels, usually genetically engineered and embedded in a matrix. The current ONT biosensor is based on mutants of the Curlin sigma S-dependent growth nanopore (CsgG)[7] (Magi *et al.*, 2017).

Nanopore sequencers directly detect the composition of a native ssDNA molecule, making them exempt from the usage of secondary signals (such as light, color or pH) customary to other sequencing technologies. Nanopore sequencers work by passing DNA molecules through a protein pore where current is applied and translocation speed is controlled by coupling an enzyme motor to the nanopore; this allows for the lowering of speed to a point that permits single-nucleotide resolution. As nucleotides pass through the pore, the current is affected with current charges being traced temporally to create squiggle space graphs – these graphs represent shifts in voltage which are characteristic of the DNA sequence that passed through the pore at that given time. This process is illustrated in **Fig. 16.** Library preparation for nanopore sequencing involves the fragmentation of DNA and ligation of adapters to both ends of the molecule (leader and hairpin adapters) pre-loaded with motor proteins. The leader adapter guides the dsDNA fragments towards the pores and the respective motor protein mediates the unzipping of dsDNA and the passage of the template strand through the pore. When the strand is finished, the hairpin motor protein then moves the complement strand through the same pore. Therefore, even though the prepared library consists of dsDNA, molecules are sequenced in single strands thanks to the action of motor proteins (Reuter *et al.*, 2015; Magi *et al.*, 2017).

Because signals aren't interpreted base to base, but rather as k-mers (oligomers of length "k" that comprise the DNA molecule), there are more than 4 possible signals to interpret. In fact, because nanopore sequencing also allows the detection of modified bases, there are over 1,000 possible signals. Each sequencing flow cell (a cartridge onto which the DNA library is inserted) has 2048 individual protein

[7] The CsgG is a secretion channel involved in curli formation. Curli are functional amyloid fibers present in the extracellular matrix of biofilms formed by some bacteria, including *α-Proteobacteria* and *γ-Proteobacteria* (Goyal *et al.*, 2014).

nanopores arranged in 512 channels, allowing it to process up to 512 DNA molecules at once (Reuter *et al.*, 2015; Goodwin *et al.*, 2016).

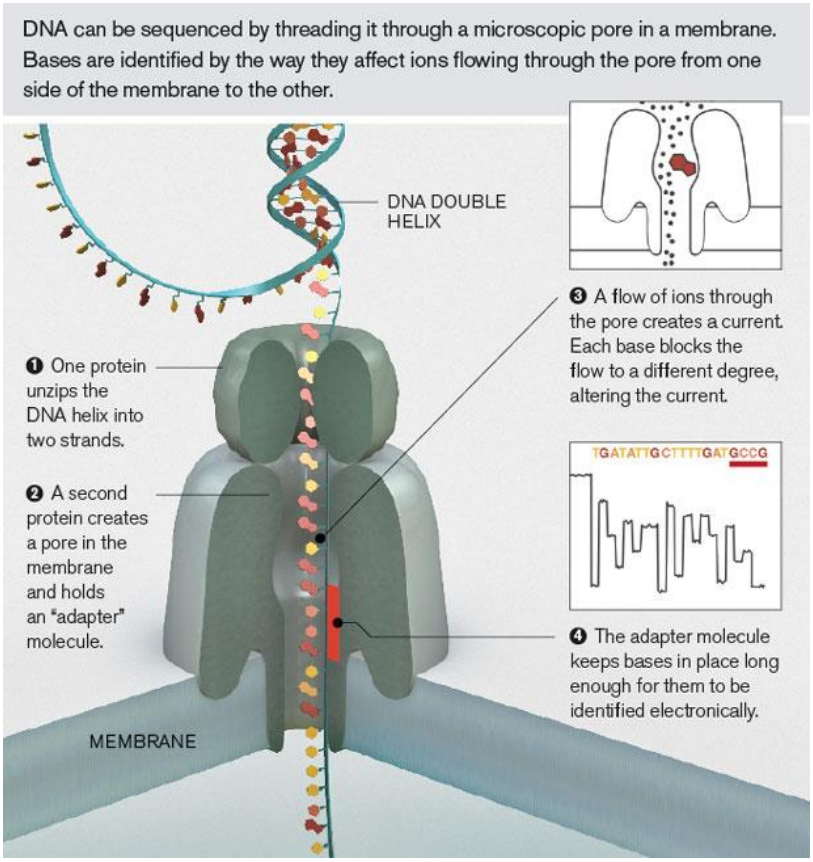


Fig. 16 - The nanopore sequencing process. Source: n.d. author, MIT Technology Reviews

Similarly to PacBio data, the main concern error-wise are indels, but the creation of consensus sequences (through addition of the aforementioned hairpin adapter) helps circumvent this issue. The presence of the hairpin adapter connects the two strands and, when successful, creates the so-called "2D" reads, a more accurate consensus option. When the process isn't effective, only the template strand – or "1D" read – is provided (Goodwin *et al.*, 2016; Levy and Myers, 2016; Magi *et al.*, 2017).

Beyond the benchtop options, GridION and PromethION, ONT's technology is also available in handheld form, through a device called MinION (shown in **Fig. 17**) – the first handheld sequencer and also the lowest-cost option, requiring only an active USB port to operate on a laptop. The device, 10 cm in length and weighing 90 g, generates DNA sequences with average length of 2-10 kb although superior sizes are achievable, given that nanopore sequencing imposes few restraints in fragment size. ONT has released several chemistry versions for the MinION (R6.0, R7.0, R7.3, R9, R9.4 and R9.5), as well as software updates, increasing the MinION's capabilities in terms of throughput, speed and error rate. For R9.4 chemistry, the 1D reads score about 90% accuracy, while 2D reads reach about 95% accuracy. Due to the recent nature of

R9.5 chemistry (released in May 2017), no official data for its performance was released yet. R9.5 introduces the $1D^2$ reads, which improve accuracy in a similar way to 2D reads but retain the simplicity in library preparation characteristic of 1D read generation (Magi *et al.*, 2017).

The portability and relatively simple library preparation make MinION very suitable for sequencing in remote locations; in fact, MinION was the device used when first sequencing DNA in microgravity conditions in the International Space Station, according to ONT's website. ONT also seems to be focusing on the portability aspects of its technology, seeking further miniaturization through the SmidgION, a nanopore sequencing device powered by a mobile phone[8].

1.2 Analysis of MinION-generated sequencing data

The unique nature of this technology also means it challenges most available bioinformatics methods which were designed to work with second-generation data. However, MinION was launched through an independent beta-testing program aimed towards a developer community – the MinION Access Program (MAP) – allowing the development and adaptation of computational approaches towards MinION generated data. As a result of the MAP, researchers had a chance to evaluate the performance of the MinION in terms of base throughput and read quality and nanopore-oriented algorithms for base-calling, read mapping, *de novo* assembly, variant discovery as well as overall data handling are now available (Magi *et al.*, 2017). Base-calling as well as assembly and polishing steps are crucial to the nanopore data analysis workflow, directly influencing the informational content that can be retrieved from sequencing data.

1.2.1 Base-calling

As previously discussed, DNA translocation through a nanopore causes current drops; this generates signals documented in squiggle space graphs, where the signals are represented by shifts in the mean current according to DNA base passage through the pore. These current signals are then decoded into bases, or base-called. To that effect, current measurements must be segmented to determine the delimitations of current shifts, a process that is often hampered by the non-uniform nature of the DNA translocation process. Because segmentation steps themselves introduce "noise" that complicates the analysis of current data, base-calling requires machine learning approaches to accurately determine sequences, using either Hidden Markov Models (HMM) or Recurrent Neural Networks (RNN) to achieve it. Currently there are several base-calling options: HMM based Nanocall and RNN based Albacore, DeepNano,

[8] More information can be found at https://nanoporetech.com/products/smidgion

Nanonet, basecRAWller, as well as ONT proprietary base-callers (such as Metrichor, now integrated into ONT's EPI2ME platform). RNN based algorithms seem to be the primary choice in dealing with nanopore sequencing data (Magi *et al.*, 2017; Stoiber and Brown, 2017).

For 1D reads, template and complement strands are usually base-called in a straightforward way. As for 2D base-calling, information from separate event sequences – corresponding to the template and complement strands – is combined, and the DNA sequence with maximum likelihood is produced (Boža *et al.*, 2016).

1.2.2 *De novo* genome assembly and polishing

When performing whole-genome sequencing, the primary objective tends to be the computational reconstruction of the genome utilizing the reads obtained from the sequencing run. Ultra-long read generating technology proves especially useful when attempting *de novo* assembly, since the reads are long enough to span over repetitive regions and other cumbersome genomic elements (Magi *et al.*, 2017).

As expected, most available assembly tools are designed to deal with second-generation sequencing data and consequently are not the most suitable for error-prone nanopore data, given that the assembly process also usually implies rounds of read correction. At first, hybrid assembly methods for long-read NGS appeared, utilizing complementary second-generation data to correct long reads. This allows the exploitation of well-established second-generation algorithms to deal with a relatively new type of data, combining advantages of both technologies. In due time, non-hybrid methods came about, using only nanopore data for iterative self-correction, and eventually attained a comparable performance to that of hybrid methodologies. Non-hybrid methods can be hierarchical or direct: hierarchical methods perform multiple rounds of read overlapping and correction as a means to improve ultra-long reads before the actual assembly process; direct methods, on the other hand, skip these prior correction steps and perform assembly directly (Koren *et al.*, 2016; Magi *et al.*, 2017).

One of the most commonly used assemblers, Canu, consists of a three-staged pipeline with correction, trimming and assembly steps that can be performed independently or in series, making it capable of operating in both a hierarchical and direct fashion. Canu supports both PacBio and Oxford Nanopore data, and was found to outperform several other non-hybrid methodologies in a study carried out by Deschamps *et al.* in 2016. Additional polishing can be carried out using tools such as Pilon or Nanopolish (Koren *et al.*, 2016; Magi *et al.*, 2017).

Pilon is an all-in-one automated tool design to improve draft assemblies by correcting indels, gaps and read alignment discontinuities. However, its peak performance is when supplied with paired-end data from Illumina libraries (Walker *et al.*, 2014).

Nanopolish, on the other hand, works by taking the draft assembly (generated by Canu, for example) and progressively modifying it through small localized changes that improve average identity and contig length of non-hybrid assemblies, using HMM on MinION-generated electric current signals. Thus, it requires no other sequencing library to perform correction (Loman *et al.*, 2015; Magi *et al.*, 2017).

Although these tools possess other possible applications and are valuable on other types of biological material, those features are not part of the present introduction.

1.3 Whole-genome sequencing and prophage detection

As previously discussed in **Chapter I**, viral evolution greatly differs from evolution of other lifeforms, since it relies heavily on genome structure rather than sequence homology to convey evolutionary relationships between organisms. Furthermore, there is also a certain degree of protein conservation among the *Siphoviridae*, regarding protein such as the integrase, the portal protein, the terminase and the tail tape measure protein (Canchaya *et al.*, 2003).

Although mosaic in nature, phage genomes undergo modular evolution, and the relative position of these modules is vital for survival (Abedon, 2009; Aksyuk *et al.*, 2012). Thus, when deciding on an experimental approach to address phage genomes, recovering information on their structure should be a priority. As seen in this chapter, the properties of third-generation sequencing make them especially suitable as a more straight-forward approach to *de novo* genome assembly of small bacterial and viral genomes, because they allow the resolution of structural conundrums (Lavezzo *et al.*, 2016).

Phage prediction tools

As sequencing technologies evolved, their potential in unveiling the prophage state of the lysogenic life cycle was eventually noticed, and dedicated tools for the detection of these prophage sequences within bacterial genomes were created, such as Prophinder and PhiSpy.

Prophinder is a prophage detection algorithm that uses similarity searches coupled with statistical detection of phage-gene enriched regions. To this effect, Prophinder is combined with the ACLAME database (standing for A CLAssification of Mobile genetic Elements), which consists of prophage predictions in sequenced prokaryotic genomes (Lima-Mendez *et al.*, 2008). Because it works through homology with known phages, the exclusive usage of tools such as Prophinder may hamper the discovery of unknown

phage regions. Other tools, such as PhiSpy, were created to counteract this issue. PhiSpy is a weighted phage detection algorithm that works based on prophage characteristics: protein length, transcription strand directionality, customized AT and GC skew, the abundance of unique phage words, phage insertion points, in addition to similarity of phage proteins. Consequently, PhiSpy is able to detect previously unknown sequences (Akhter *et al.*, 2012).

2. MATERIALS & METHODS

2.1 Bacterial strains

The same SDSD strains from Chapter III were used. For further details, consult **Appendix A.**

2.2 Growth conditions and culture media

SDSD strains were recovered from cryopreserved cultures maintained in THYE - Todd-Hewitt (BD) supplemented with 1% yeast extract (Oxoid) - with 20% (v/v) glycerol at -80°C. For subsequent experiments, strains were incubated overnight at 37°C in M17YE broth, generally in a total volume of 250 mL.

2.3 Genomic DNA extraction

DNA extraction was performed using the Wizard ® Genomic DNA Purification Kit (Promega), with modifications to the protocol for isolation of genomic DNA from gram-positive bacteria: OD_{600} of overnight cultures was measured to gauge the volume needed to place the kit's yield between 6 and 13 µg of genomic DNA[9]. The corresponding volume was then centrifuged at 4000 × *g* and 20°C for 15 minutes (using an Eppendorf 5810 R centrifuge), after which the supernatant was discarded and the pellet washed twice with 1 mL of ultrapure water. The washed pellet was then resuspended in 480 µL of 50 mM EDTA. An enzymatic lysis cocktail was prepared, consisting of lysozyme, in a final concentration of 10 mg/mL (Sigma-Aldrich), and mutanolysin, in a final concentration of 0.08 mg/mL (Sigma-Aldrich); the final volume of the lysis cocktail must be 120 µL, as not to affect downstream steps. This lysis cocktail was then added to the cells + EDTA mixture, mixed by gentle pipetting and incubated for 2 h at 37°C[10]. After incubation, the samples were centrifuged at 15996 × *g* for 2 minutes using a Sigma 1-15P centrifuge, the supernatant was discarded and 600 µL of the provided Nuclei Lysis Solution were added to each sample. Samples were incubated at 80°C for 5 minutes and subsequently cooled to room temperature. 3 µL of the provided RNase Solution were

[9] According to the manufacturer's instructions, 3.5×10^8 cells are needed. The multiplication factor for these strains, 2×10^8 was determined during the first Strep project.

[10] For more mucous strains, additional steps of mixing by vortex and pipetting were performed prior to incubation.

added and the samples were mixed by inversion. Then, samples were incubated at 37°C for 60 minutes and cooled to room temperature afterwards. Next, 200 µL of the kit's Protein Precipitation Solution were added to the samples which were then vigorously mixed trough vortex for 20 seconds. Following this step, samples were placed on ice for 5 minutes and centrifuged at 15996 × *g* for 3 minutes. The resulting supernatant was transferred to a new microtube containing 600 µL of room temperature isopropanol and mixed by inversion until visible threads of DNA were observed. Samples were subsequently centrifuged at 15996 × *g* for 2 minutes. Supernatants were poured off and the tubes were drained on clean absorbent paper; after drying, 600 µL of room temperature 70% ethanol were added and samples were mixed trough inversion to wash the pellet. Samples were centrifuged at 15996 × *g* for 2 minutes, after which the ethanol was aspirated. Sample tubes were drained on clean absorbent paper and allowed to air dry for 15 minutes. Subsequently, genomic DNA was resuspended in 30 µL of nuclease-free water[11].

2.4 Genomic DNA quality control

2.4.1 DNA quantification

DNA quantification was performed using the Qubit 2.0 fluorometer (Invitrogen) with the dsDNA High-Sensitivity Kit - suitable for samples expected to have between 0.2-100 ng of DNA - and according to the manufacturer's instructions. The volume of sample dispensed for Qubit quantification was 1 µL.

2.4.2 Absorption spectral analysis

To assess the quality of input genomic DNA, 5 µL of each sample were added to 495 µL of nuclease-free water (NFW). The samples' absorbance scans from 200 nm to 400 nm were then taken using a UNICAM UV2 Spectrometer paired with the Vision V3.32 software. Scans were then compared to that of a purified λ phage DNA stock (Invitrogen). The reference, as well as sample scans, can be found in **Appendix C**.

2.4.3 DNA agarose gel electrophoresis

Genomic DNA (30 µL of each sample) was then submitted to electrophoresis in a 0.8% (w/v) agarose (Invitrogen) gel, with 0.5X TBE buffer (40 mM Tris; 45 mM Boric acid; 1 mM EDTA; pH 8.3) and a constant voltage of 5.3 V/cm for 1 h. The gel was stained with Ethidium bromide and revealed in an Alliance 4.7 UV

[11] Resuspension in standard buffers such as TE is not recommended for Nanopore sequencing, given that buffers may interfere with the established current values. Resuspension in water is advised instead.

transilluminator (UVItec) and the image retrieved using the Alliance software. The molecular weight marked used was a "1 kb DNA Ladder" (Invitrogen). Results can be found in **Appendix C**.

2.5 Library preparation

Sequencing efforts for the present work started before ONT's launch of $1D^2$ sequencing options. As a result, strain VSD17 was sequenced using a 1D protocol while strains VSD4, VSD13, VSD19 and GCS-Si were sequenced using the newest $1D^2$ protocol.

2.5.1 1D Genomic DNA by ligation sequencing protocol (using R9.4 chemistry)

Preparation for 1D R9.4 chemistry sequencing was done using Nanopore's SQK-LSK108 kit with a FLO-MIN106 flow cell for the MinION MK 1B; minor alterations to ONT's protocol were made. 2-2.5 µg of genomic DNA obtained in **section 2.4** were diluted in NFW to a final volume of 46 µL. Using a Covaris g-TUBE, DNA was sheared to 8 kb fragments: DNA was placed in the g-TUBE and centrifuged at 6000 × *g* for 1 minute using an Eppendorf 5424 R centrifuge; the g-TUBE was subsequently inverted and centrifuged once more in the same conditions, allowing the user to collect the fragmented DNA at the top section, which was then transferred into a clean microtube. DNA was then end-repaired and dA-tailed using the NEBNext Ultra II End-Repair/dA-Tailing Module (New England Biolabs) by adding to the sample: 7 µL of Ultra II End-prep reaction buffer; 3 µL of Ultra II End-prep enzyme mix; 4 µL of NFW, for a total volume of 60 µL. The tube was mixed by flicking, spinned down and samples were then incubated (10 minutes at 20°C and 10 minutes at 65°C) using a Biometra T Gradient thermocycler. After incubation, 1X volume of magnetic Agencourt AMPure XP beads (Beckman Coulter) were added and DNA cleanup was performed according to the manufacturer's instructions (with a 5 minute incubation period and elution in 31 µL of NFW). After elution, 1 µL of clean end-prepped DNA was quantified using the Qubit 2.0 fluorometer (as specified in **section 2.4.1** of the present chapter) to check whether recovery met the hallmark of 700 ng of DNA. To the leftover 30 µL of end-prepped DNA were added: 20 µL of Adapter Mix (AMX, provided in ONT's kit) and 50 µL of of NEB Blunt / TA Ligase Master Mix, to a final volume of 100 µL. Samples were then mixed by flicking, spinned down and incubated for 10 minutes at room temperature. Subsequently, 40 µL Agencourt AMPure XP beads were added to the reaction, mixed by pipetting and incubated on a rotator mixer for 5 minutes at room temperature. DNA was then placed on a magnetic rack for the beads to pellet and the supernatant was pipetted off; 1X volume of Adapter Bead Binding buffer (ABB, provided with the kit) was added, the beads were resuspended and the tube was then returned to the magnetic rack, allowing beads to pellet. The

addition of ABB, resuspension and pelleting was repeated once more. The resulting bead pellet was then resuspended in 15 µL of Elution Buffer (ELB, provided in ONT's kit) and incubated for 10 minutes at room temperature, after which the sample was placed on the magnetic rack, eluate was removed and transferred to a clean microtube. Again, 1 µL of the adapted library was quantified using the Qubit 2.0 fluorometer to register the quantity of library available at this point.

2.5.2 1D^2 sequencing of genomic DNA protocol (using R9.5 chemistry)

Preparation for 1D^2 R9.5 chemistry sequencing was done using Nanopore's SQK-LSK308 kit with FLO-MIN107 flow cells for the MinION MK 1B; minor alterations to ONT's established protocol were performed. 2-2.5 µg of genomic DNA obtained in **section 2.4** were diluted in NFW to a final volume of 46 µL. Using a Covaris g-TUBE, DNA was sheared to 8 kb fragments: DNA was placed in the g-TUBE and centrifuged at 6000 × *g* for 1 minute using an Eppendorf 5424 R centrifuge; the g-TUBE was subsequently inverted and centrifuged once more in the same conditions, allowing the user to collect the fragmented DNA at the top section, which was then transferred into a clean microtube. DNA was then end-repaired and dA-tailed using the NEBNext Ultra II End-Repair/dA-Tailing Module (New England Biolabs) by adding to the sample: 7 µL of Ultra II End-prep reaction buffer; 3 µL of Ultra II End-prep enzyme mix; 4 µL of NFW, for a total volume of 60 µL. The tube was mixed by flicking, spinned down and samples were then incubated (10 minutes at 20°C and 10 minutes at 65°C) using a Biometra T Gradient thermocycler. After incubation, 1X volume of magnetic Agencourt AMPure XP beads (Beckman Coulter) were added and DNA cleanup was performed according to the manufacturer's instructions (with a 5-minute incubation period and elution in 25 µL of NFW). After elution, 1 µL of clean end-prepped DNA was quantified using the Qubit 2.0 fluorometer to check whether recovery met the hallmark of 700 ng of DNA. About 700-800 ng of end-prepped DNA were diluted in NFW to a final volume of 22.5 µL, and to it were added: 2.5 µL of 1D^2 adapter (provided in the kit) and 25 µL of NEB Blunt / TA Ligase Master Mix. The tube was then mixed by inversion, spinned down and allowed to incubate at room temperature for 30 minutes. After incubation, 20 µL of magnetic Agencourt AMPure XP beads (Beckman Coulter) were added and DNA cleanup was performed according to the manufacturer's instructions (with a 10-minute incubation period and elution in 46 µL of NFW). At this point, 1 µL of the DNA sample was quantified using the Qubit 2.0 fluorometer to register recovery. Afterwards, 5 µL of Barcoded Adapter Mix (BAM, available in the 1D^2 kit) and 50 µL of NEB Blunt / TA Ligase Master Mix were added to the remaining 45 µL of 1D^2 adapted sample (for a total volume of 100 µL). Following gentle mixing by inversion and spinning down, the sample was incubated at room temperature for 30 minutes. Subsequently, 40 µL Agencourt AMPure XP beads were added to the reaction, mixed by pipetting and

incubated on a rotator mixer for 10 minutes at room temperature. DNA was then placed on a magnetic rack for the beads to pellet and the supernatant was pipetted off; 1X volume of Adapter Bead Binding buffer (ABB, provided with the kit) was added, the beads were resuspended and the tube was then returned to the magnetic rack, allowing beads to pellet again. The addition of ABB, resuspension and pelleting was repeated once more. The resulting bead pellet was then resuspended in 15 μL of Elution Buffer (ELB, provided in ONT's kit) and incubated for 10 minutes at room temperature, after which the sample was placed on the magnetic rack, eluate was removed and transferred to a clean microtube. Again, 1 μL of the adapted library was quantified using the Qubit 2.0 fluorometer to assess whether recovery met the aim of about 200 ng of adapted library.

2.6 MinION flow cell set-up

Preparation of the MinION and respective flow cell is a common procedure to both library preparation methodologies, with minor differences. The host computer used for the sequencing run met the requirements to execute the associated MinKNOW software: Windows 10, 16 Gb RAM, SSD, i7 processor, USB 3.0. During sequencing efforts, different versions of the MinKNOW software were used: version 1.5.12 (strain VSD17), version 1.7.10 (strains VSD13, GCS-Si) and version 1.7.14 (strains VSD19, VSD4).

To prepare the MinION for sequencing, the Quality Control protocol should be run first. To this end, the MinION and respective flow cell (FLO-MIN106 for 1D sequencing and FLO-MIN107 for $1D^2$ sequencing) as well as the host computer were assembled as depicted in **Fig. 17.** MinKNOW was then setup to run the Platform QC (executing the NC_Platform_QC.py protocol), validating the integrity of the nanopore array before use and determining the number of available pores for sequencing. Following QC, the flow cell was primed for library loading by adding 800 μL of the priming buffer, (prepared by mixing 480 μL of Running Buffer with Fuel Mix (RBF, provided in the kits) with 520 μL of NFW) through the flow cell's priming port and waiting 5 minutes. During the waiting period, the library obtained either through **step 2.5.1 or 2.5.2** was prepared for loading; to this end, 35 μL of RBF, 25.5 μL of LLB (Library Loading Bead kit, included in the sequencing kits), 12 μL of prepared library and 2.5 μL of NFW were added to a new microtube and mixed gently through pipetting. After the 5 minutes passed, the leftover 200 μL of priming buffer were loaded through the flow cell's priming port. Subsequently, the 75 μL of library were loaded through the SpotON port, in a dropwise fashion. The sequencing script in MinKNOW was then initiated by running either the NC_48Hr_Sequencing_Run_FLO_MIN106_SQK-LSK108.py (for 1D sequencing) or the NC_48Hr_Sequencing_Run_FLO-MIN107_SQK-LSSK308.py (for $1D^2$ sequencing). Live basecalling was not

performed and sequencing runs were stopped when production of a suitable amount of data was detected. Consequently, run times are not uniform throughout different strains.

Fig. 17 - The MinION MK1B structure (A) and setup scheme (B). Adapted from: Oxford Nanopore Technologies

When required, washing protocols (using ONT's Washing Kit) were executed according to the manufacturer's instructions after sequencing to preserve flow cells.

2.7 Nanopore sequencing data analysis

Sequencing data analysis was performed using both local software and server-based tools, as represented in **Fig. 18**. All local software was installed according to the developer's instructions and ran on a command-line based interface on an Ubuntu System 14.04 LTS.

Basecalling was performed after the sequencing run was completed. Albacore v.1.1.2[12] was used for both R9.4 and R9.5 data; R9.4 requires linear basecalling only while R9.5 implies an additional step where linear basecall results are recalled, detecting potential read pairs. For R9.5 data, only the $1D^2$ paired reads were used downstream. Afterwards, NanoPlot v0.17.4[13] was used to assess statistics of sequencing data. Japsa v1.7[14] was then used for read filtering, excluding reads with a quality score (QScore, an indication of how well the raw data fits into the basecalling model that does not fit the usual Phred error rates) below 10 or smaller than 1 000 bp in length. The QScore minimum was defined taking into account the widely used live-basecalling platform Metrichor: when performing the basecalling, Metrichor categorizes reads into

[12] Albacore source code is available at https://github.com/dvera/albacore

[13] NanoPlot source code is available at http://github.com/wdecoster/NanoPlot

[14] Japsa source code is available at https://github.com/mdcao/japsa/ and further documentation can be found at http://japsa.readthedocs.io/en/latest/

"pass" or "fail" bins, depending on whether basecalling is successful and also on the read's QScore, which must be above 9; thus, to make data more comparable with existing results, a threshold of 10 was established (Lu *et al.*, 2016). As for read length, setting a minimum of 1000 bp allows the diminishing of the "noisy" effect from having very small reads as input for assembly, simplifying the process while still retaining a substantial amount of information and improving computational performance (Koren and Phillippy, 2015).

Next, assembly was performed using Canu v1.5[15] in its full pipeline version (comprising read correction, read trimming and unitig construction steps) with standard parameters for uncorrected nanopore reads; the expected genome size – a parameter required for assembly - was estimated to be around 2.2 Mb based on available SDSD and SDSE genomes on NCBI. For quality checkpoints, QUAST v4.5[16] (Gurevich *et al.*, 2013) and the MUMmer v3.23[17] function "dnadiff" were used to compare ongoing assemblies to available SDSD and SDSE reference genomes: *Streptococcus dysgalactiae* subsp. *dysgalactiae* strain ATCC 27957 and *Streptococcus dysgalactiae* subsp. *equisimilis* strain ATCC 12394 (NCBI accession numbers: NZ_AEGO00000000.1 and NC_017567.1, respectively). Following the assembly evaluation, Nanopolish v0.7.0[18] was used for polishing with default parameters, through the "variants --consensus" subprogram. Usage of the Nanopolish algorithm implies previous indexing and aligning using the Burrows-Wheeler Aligner (BWA)[19] (Li, 2013) as well as Sequence Alignment/Map tools (SAMtools)[20] (Li *et al.*, 2009) for necessary file format conversions. A second quality checkpoint was performed. Polished genomes were then annotated using the RAST[21] online server (Rapid Annotation using Subsystem Technology) (Aziz *et al.*, 2008) with the Classic RAST annotation scheme while enabling the frameshift fix and automatic error fix. Subsequently, phage prediction was performed using PhiSpy v3.2[22] without a specified training set and Prophinder v0.4[23] with default parameters. Genome visualization, as well as phage-region sequence retrieval was performed using the Integrative Genomics Viewer v2.3 (IGV)[24] Java application.

---

[15] Canu source code is available at https://github.com/marbl/canu and further documentation can be found at http://canu.readthedocs.io/en/latest/

[16] QUAST source code, as well as available information, is available at http://quast.sourceforge.net/

[17] MUMmer source code, is available at https://github.com/mummer4/mummer and specific information on "dnadiff" can be found at https://github.com/mummer4/mummer/blob/master/docs/dnadiff.README

[18] Nanopolish source code can be found at https://github.com/jts/nanopolish and additional information is available at http://simpsonlab.github.io/blog/

[19] BWA source code is available at https://github.com/lh3/bwa

[20] SAMtools source code is available at http://www.htslib.org/doc/samtools.html

[21] RAST is available at http://rast.nmpdr.org/

[22] PhiSpy source code is available at https://github.com/linsalrob/PhiSpy

[23] Prophinder is available at http://aclame.ulb.ac.be/Tools/Prophinder/

[24] The IGV Java application is available for download at http://software.broadinstitute.org/software/igv/download and further information can be found at http://software.broadinstitute.org/software/igv/userguide

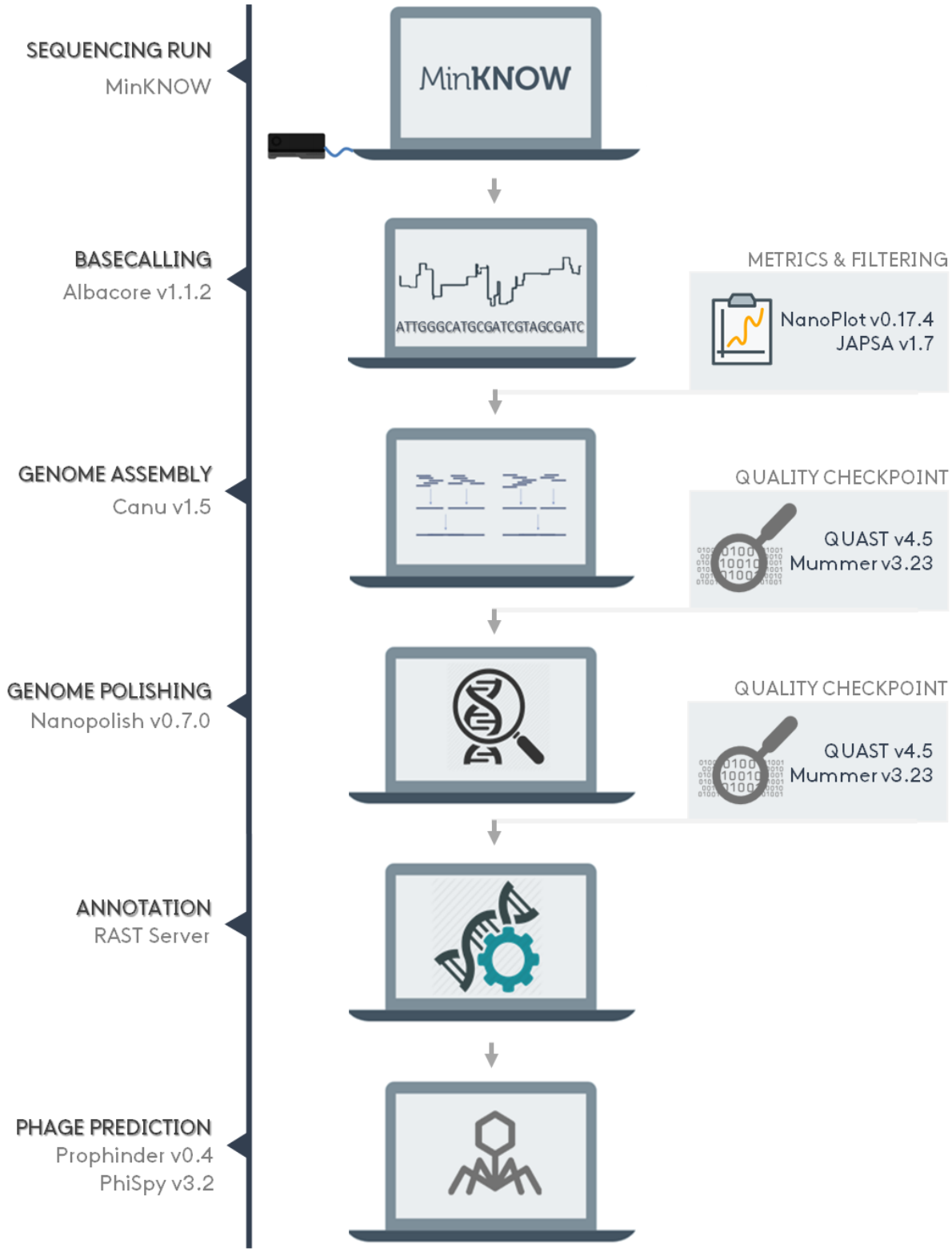


Fig. 18 - Sequencing data analysis workflow. Essential steps from sequencing to prophage sequence detection are listed, as well as the tools used to perform each step. The figure was constructed using images from Oxford Nanopore Technologies

3. RESULTS & DISCUSSION

3.1 Sequencing Metrics

The five strains were sequenced in independent runs, using one R9.4 FLO-MIN106 flow cell and two R9.5 FLO-MIN107 flow cells. Each sequencing run immediately followed library preparation, according to recommendations from ONT's online community. Although the sequencing protocols recommend a final DNA yield of about 200 ng for optimal results, the sequencing run was carried out even when final values did not meet this criterium. Initial sequencing statistics are summarized in **Table 2.** Sequencing data was basecalled after the sequencing run was completed (also undergoing a pairing process, in the case of $1D^2$ sequencing) and was then filtered, generating a subset used for assembly. For strain VSD17, sequenced with R9.4 chemistry, the filtered subset consists of 1D reads according to previously specified criteria (a QScore over 10 and length above 100 bp); for the remaining strains, sequenced with R9.5 chemistry, the filtering process was applied to paired $1D^2$ reads only, consisting of $1D^2$ reads above minimum quality and length thresholds. The NanoPlot script was ran on both sets for all strains and its output allowed to evaluate sequencing in terms of data yield as well as read length and read quality, also providing some insight towards the differences in 1D and $1D^2$ sequencing protocols.

Table 2 - DNA yield, number of active channels and coverage of sequencing runs. Samples VSD4 and VSD19 had final concentrations below the quantification kit detection limit. Coverage was estimated for an expected genome size of about 2.2 Mb and taking into the account the total number of obtained reads (raw data).

Strain	DNA yield	Active Channels	Runtime	Coverage
VSD4	-	393	8 h 23 m	800 x
VSD13	228 ng	430	7 h 32 m	686 x
VSD17	224 ng	478	16 h 17 m	298 x
VSD19	-	477	5 h 38 m	954 x
GCS-Si	333.2 ng	267	9 h 24 m	768 x

The success of nanopore sequencing runs seems to depend on quite a few factors, with DNA quality being of major influence. Overall, according to the results in **Appendix C** and **Table 2**, the most successful DNA extraction seems to be that of strain GCS-Si, both in terms of DNA quality and yield; strains VSD13 and VSD19 follow, with reasonably high quality, while strains VSD4 and VSD17 are lower in quality. VSD4 liquid cultures have a distinctive viscous quality, hindering the DNA extraction process. While not a part of this dissertation, studies of biofilm composition involving these strains are a part of the Strep-hosp project, indicating that this may be a distinguishing feature for these strains. Although Qubit measurements were

not obtained for all strains, sequencing was still carried out, given that instances of successful sequencing runs with less DNA than that recommended by the protocols were previously reported by the MAP community.

Runtimes and the resulting coverages seem to differ substantially between 1D and $1D^2$ sequencing protocols, with strain VSD17 having the longest sequencing run out of all 5 strains (16 h 17 min) and producing the least amount of coverage (298 x). Among $1D^2$ sequenced strains, runtimes and coverage are more uniform, with runtimes varying between 5 h 38 min and 9 h and 24 min, while coverage varies between 600 x and 954 x. Interestingly, strain VSD19 registers the smallest runtime and the highest coverage.

Besides DNA quality, flow cell state also factors in to the success of sequencing runs. Due to their manufacturing process, flow cells are not identical between them, demanding a Quality Control check before initiating sequencing scripts to check the number of available pores. The number of functional pores has been previously found to directly influence data production (Brown, 2015). Additionally, while flow cells are meant to be reused, with the development of specific cleaning protocols, the sequencing process tends to clog a percentage of the pores, directly affecting the following run.

Strain VSD17 was sequenced in a reused R9.4 1D flow cell and yet it showed a number of available channels comparable to those of new R9.5 $1D^2$ flow cells. Strains VSD13 and GCS-Si were sequenced using the same flow cell, as well as strains VSD19 and VSD4, and in both instances the expected drop in available pores (from 430 for VSD13 to 267 for GCS-Si, and from 477 for VSD19 to 393 for VSD4) was observed. Considering each flow cell contains 2,048 pores, the number of active ones seems quite low; however, this is not uncommon. Given that FLO-MIN107, or $1D^2$ sequencing flow cells are relatively new, their durability and endurance through the shipping and storage process may not be fully optimized.

Remarkably, despite the low number of available pores, $1D^2$ sequencing runs still produced a considerable amount of data (represented in **Fig. 19**), especially when compared to the 665 Mb produced for strain VSD17 through 1D sequencing. Strain VSD13 registers the minimum amount R9.5 sequencing data, at 1.5 Gb, while strain VSD19 had the biggest yield, at about 2.1 Gb. The relationship between data yield and the number of reads provides some insight to DNA fragmentation during DNA extraction and library preparation; strain VSD17, for example, whose DNA was found to be quite fragmented and of lower quality, produced a high number of reads (412 253) but the lowest data yield (655 Mb), meaning although more reads were produced, they ought to be quite short, which is not desirable in nanopore sequencing. The remaining strains register much lower read counts and considerably higher data yields, hinting at the production of longer reads.

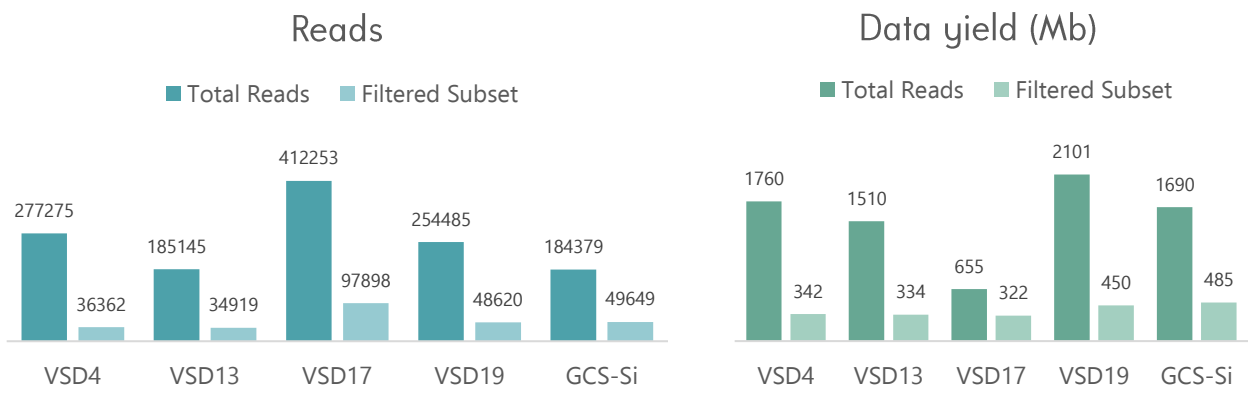
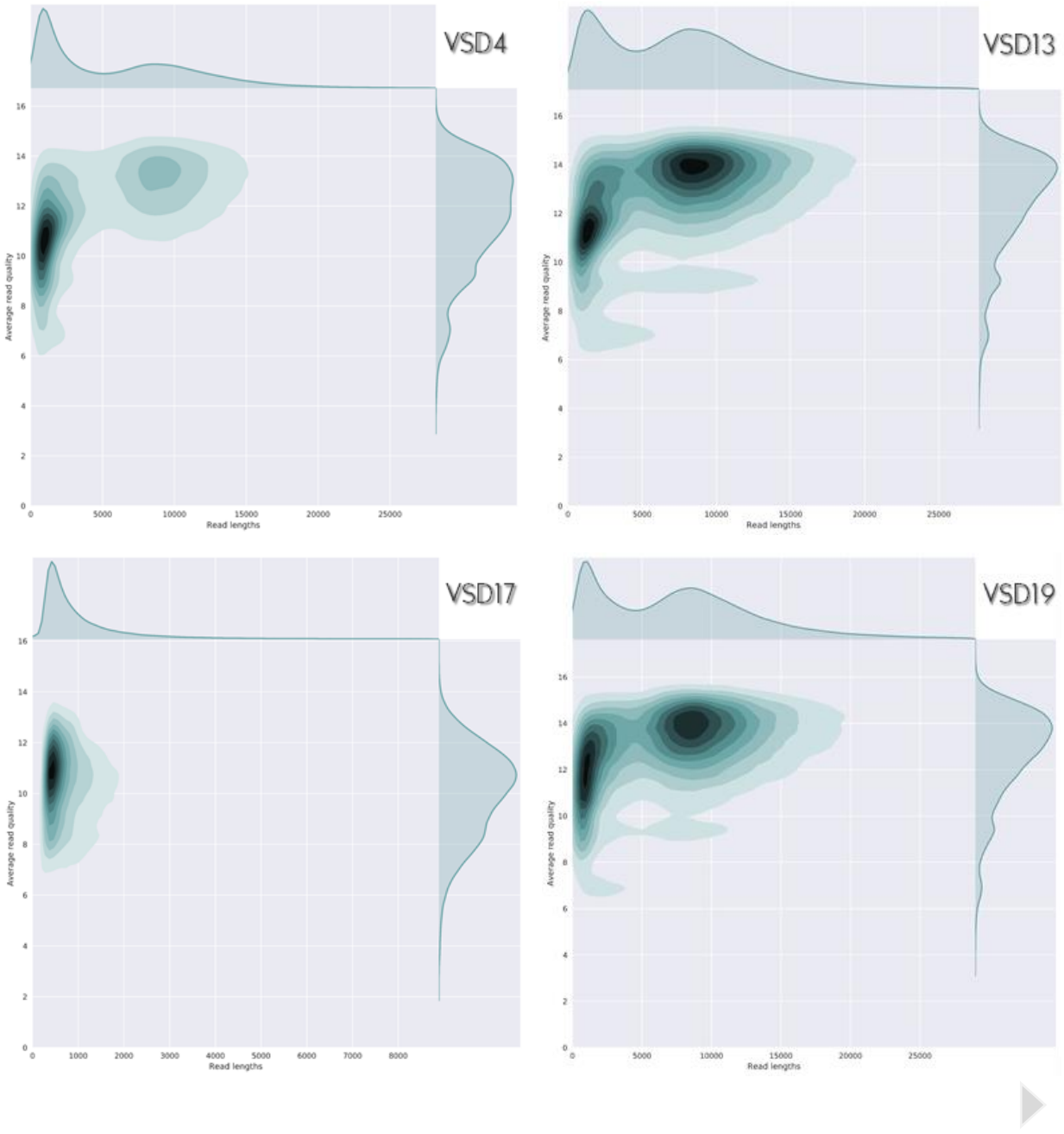


Fig. 19 - Data yield and read number of sequencing runs. The first graph represents the total number of reads and the second one represents data yield in megabases. For each strain, the total obtained reads and the assembly filtered subset are represented.

Unfiltered read length and quality are represented in **Fig. 20,** with bivariate plots showing kernel density estimates. Kernel density estimation, or KDE, is a non-parametric method (because it does not assume an underlying distribution) of estimating the probability density function of a continuous random variable, making it suitable for representation of this kind of data. In KDE, every datum then becomes the center of a kernel function (a probability density function that must be even, like the normal distribution for example), ensuring kernel symmetry – kernel density estimates are "bumps" centered at a given datum and whose size is representative of the probability assigned to the neighbourhood of values that surround the datum (Silverman, 1986). According to the plots in **Fig. 20**, read length and quality seem to directly reflect the quality of input DNA. Even though all strains show a peak of read number at small lengths, due to fragmentation during the protocol's execution and eventual clogging of pores during sequencing, $1D^2$ sequenced strains also show a peak corresponding to a higher read length. This second peak corresponds to about 8 kb in length, which agrees with the protocol's initial shearing step. The peak is highest for strain GCS-Si, the strain with highest DNA quality and lowest for strain VSD4. Although excluded from the plots for clarity, ultra-long reads were produced on all sequencing runs, with a staggering 2 Mb long read for strain VSD4, meaning that about 90% of the genome was present in a single read. Filtered $1D^2$ data subsets feature average read lengths around 9 kb and register over 90% of reads with a QScore above 15. As for strain VSD17, the average read length is about 3 kb and only about 8% of reads in the filtered dataset have a QScore above 15, illustrating once again substantial differences in the sequencing process. Further details on sequencing metrics for both unfiltered and filtered datasets, as well as read length vs. quality bivariate plots for filtered datasets can be found on **Appendix D**. Most available literature on nanopore whole-genome sequencing concerns R7, R7.3 and R9 chemistry, from which both R9.4 and R9.5 seem to be an

improvement in terms of data yield, average read length among other metrics (Batovska *et al.*, 2017; Jain *et al.*, 2017; Salazar *et al.*, 2017).



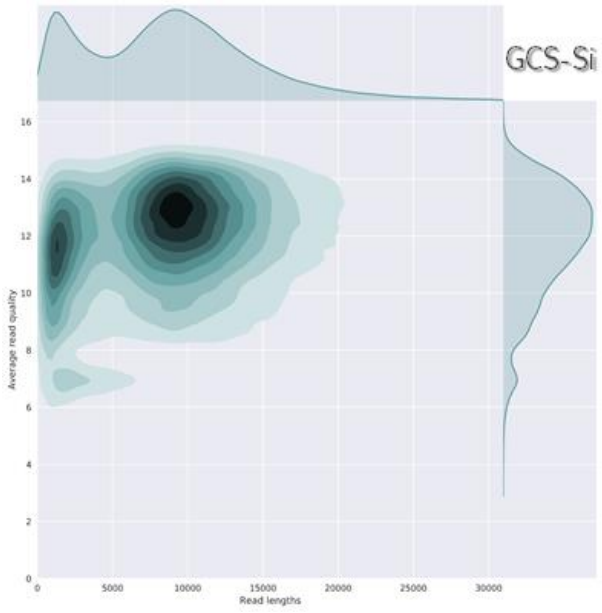


Fig. 20 - Read quality vs. read length distribution of total obtained reads. The bivariate plots, obtained using NanoPlot, show a kernel density estimate (KDE) of the read length compared to the read's QScore. The horizontal axis represents read length (with a maximum of 30 000 bp) and the vertical axis represents average read quality (with a maximum value of 16). For the sake of intelligibility, extremely long outlier reads were excluded from this representation.

3.2 Genome assembly and polishing

Filtered subsets were then used for genome assembly and polishing. Albeit only a small fraction of the obtained data was featured in these subsets, it sufficed to assemble reads into one single contig representing chromosomal DNA in all 5 assembly experiments. During polishing, the total data from each sequencing run is used to polish the previously obtained draft assembly, calculating an improved consensus sequence. All assemblies were evaluated (against both an SDSD and an SDSE reference) before and after the polishing stage to assess Nanopolish's efficiency. Polishing resulted in an overall assembly size increment, as well an increase in the average identity of the aligned sequence blocks. The percentage of aligned bases slightly decreased in some cases but, on the other hand, the size of the longest consecutive alignment possible increased. The number of indels detected also decreased after polishing, while increasing the number of single-nucleotide polymorphisms. Due to the higher error-rate of nanopore sequencing, both indels and single-nucleotide polymorphisms can only be corrected up to a point. Overall, polishing was considered to improve the assembly, and as such, polished assemblies were used for annotation and phage prediction. Results on the effects of assembly polishing can be found on **Appendix E**.

Differences between polished assemblies and references were analyzed quantitatively at first, in terms of the percentage of unaligned bases as well as average base discrepancy, as depicted in **Fig. 21.** It is noteworthy that these results are not free from artifacts left by the assembly process. Strains GCS-Si and

VSD13 seem to differ the most from the references in terms of unaligned bases, although differences are not as clear in what concerns average base discrepancy.

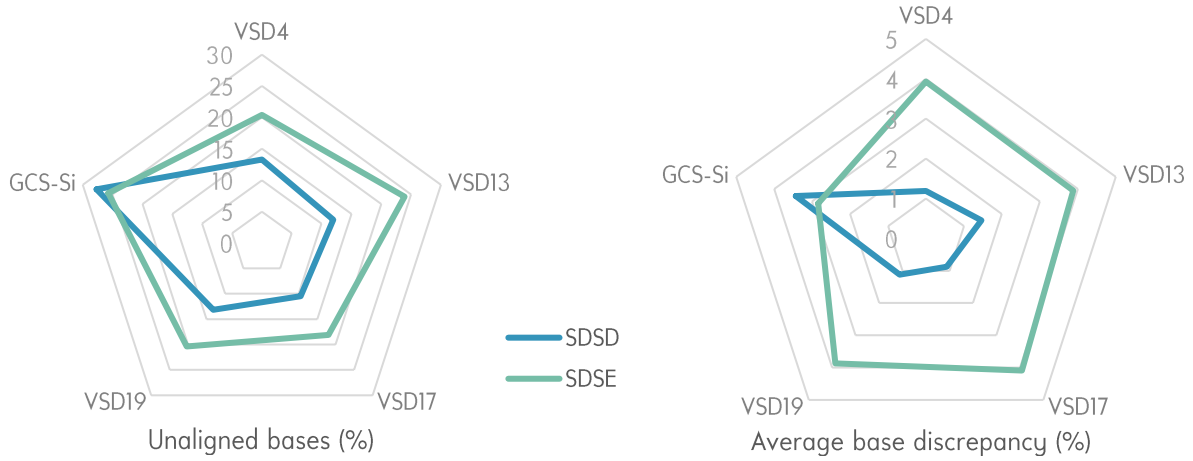


Fig. 21 – Assembly discrepancies with *S. dysgalactiae* subsp. *dysgalactiae* (SDSD) and *S. dysgalactiae* subsp. *equisimilis* (SDSE) reference genomes. Disparities between the five polished assemblies and the SDSD and SDSE reference genomes are represented in terms of the percentage of overall unaligned bases as well as the percentage of discrepant bases between the reference and a given assembly within aligned sequence blocks.

The plots in **Fig. 22**, which are part of the QUAST output, represent the alignment of each assembly with their closest reference. For all VSD strains that is the SDSD genome; strains GCS-Si, however, appears to be closer to the SDSE genome. Strains VSD13 and GCS-Si appear to have the most striking differences from their respective references, as can be seen in **Fig. 21** and **Fig. 22**, with pronounced inverted segments. Interestingly, strain VSD13 is also the subject of *in vitro* and *in vivo* pathogen-host assays in the Strep-hosp project and it has been found to hold remarkable pathogenic potential on *in vivo* assays in zebra fish as well as *in vitro* infection experiments with keratinocytes (Roma-Rodrigues *et al.*, 2016). Strain GCS-Si, as previously mentioned, was isolated from a human host who developed cellulitis after contacting with infected fish (Koh *et al.*, 2009). Thus, strains with increased virulence seem to be diverging the most from reference genomes, which is expectable since references that represent typical behavior for each subspecies were chosen.

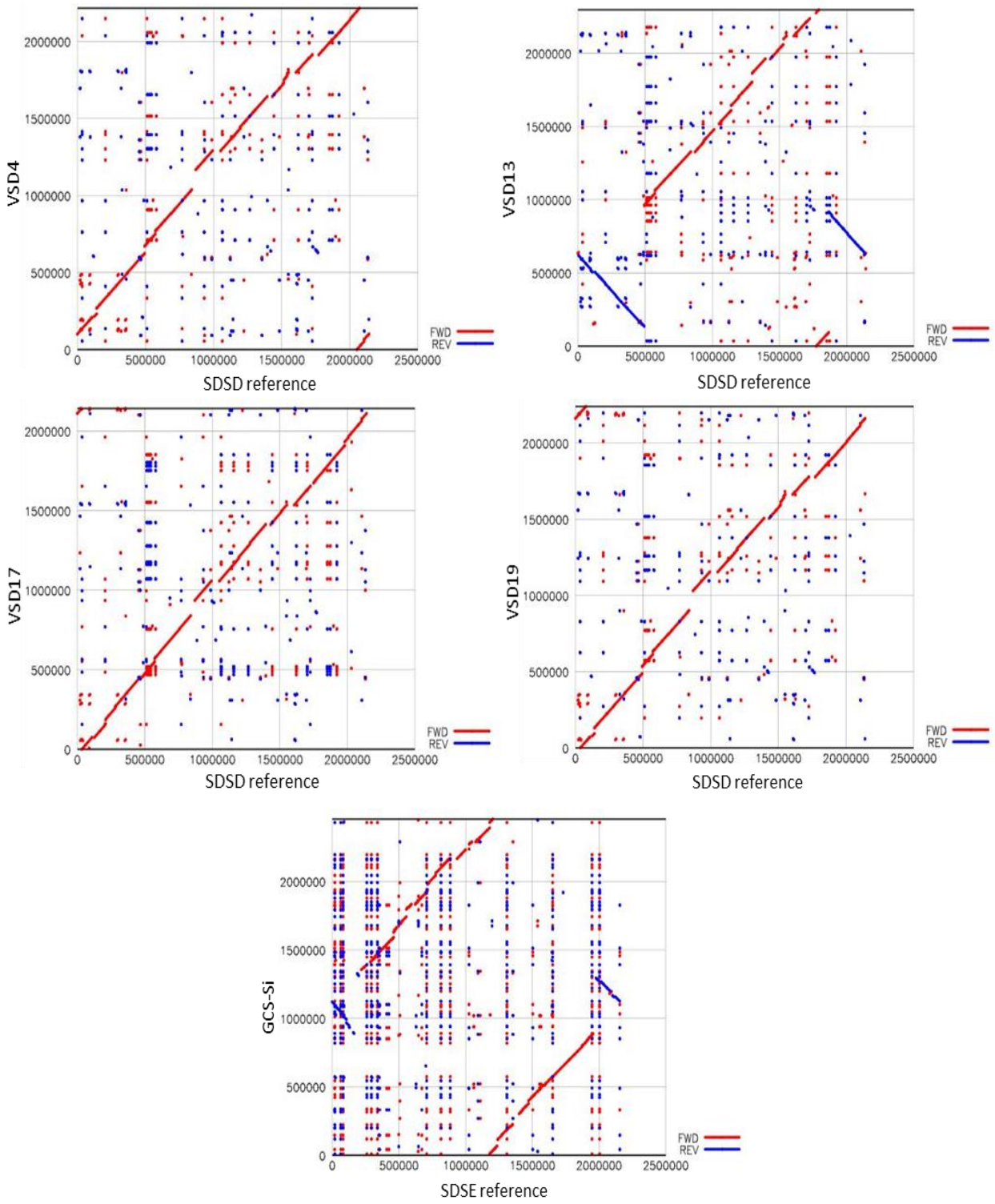


Fig. 22 – Sequence alignments between assemblies and their closest reference genome. Dot-plots were obtained from the QUAST analysis. Red segments represent forward aligned blocks while blue segments represent blocks aligned in the reverse direction and thus indicate inversions between assembly and reference.

3.3 Genome assembly annotation

Polished assemblies were subsequently annotated using RAST, a fully automated tool for the annotation of bacterial and archaeal genomes (Aziz *et al.*, 2008). Results hailing from RAST annotation are depicted in **Fig. 23**. In preliminary testing, three suitable tools for annotation of bacterial genomes were used: Prokka[25] (Seemann, 2014), RAST and Blast2GO[26] (Conesa and Götz, 2008) and RAST presented the best compromise between accuracy, celerity and usability, allowing the analysis to occur in a timely fashion.

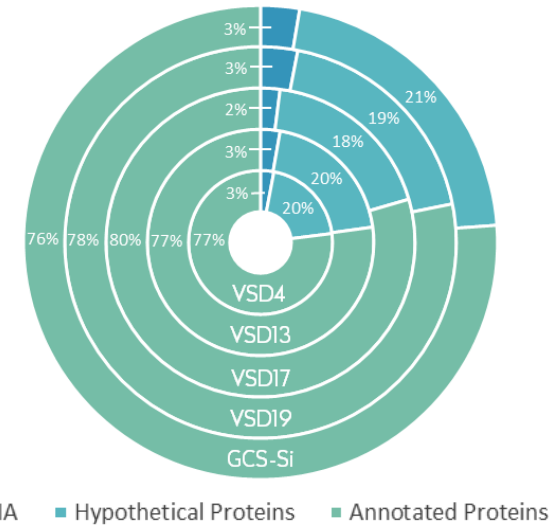


Fig. 23 - RAST annotation results. Percentage of detected RNA coding sequences and protein coding sequences, divided into annotated protein sequences and hypothetical proteins.

The RAST annotation process utilizes its growing library of manually curated subsystems to assign gene functions, making "subsystem-based assertions" when functional variants of subsystems are recognized in query sequences. On the other hand, "nonsubsystem-based assertions" are based on more widely known approaches, integrating results from different tools to produce the assertion. Subsystems are defined by expert curators, integrating literature-bound knowledge into expert assertions that can be projected by this automated tool (Aziz *et al.*, 2008). Sequencing runs were successful enough to yield over 75% annotated proteins for all strains. Between 47-50% of annotated proteins were covered by RAST subsystems, with slight variations in the number of subsystems ticked for each strain (between 321 and 329). Nevertheless, RAST's subsystems do provide insight about the metabolic tendencies of each analyzed genome.

[25] Prokka source code can be found at: https://github.com/tseemann/prokka
[26] Blast2GO is available at: https://www.blast2go.com/

3.4 Prophage prediction and detection of bacteriophage resistome sequences

Annotated genome assemblies were then submitted to prophage prediction, using two different tools: Prophinder and PhiSpy. Prophinder works by translating the coding sequences in the input genome assembly and detecting phage-like coding sequences using gapped BLASTP (protein BLAST) searches against the phage proteins present in the ACLAME database (Lima-Mendez *et al.*, 2008). This database contains a collection of prokaryotic mobile genetic elements hailing from diverse sources, including all known plasmids, transposons as well as phage genomes. ACLAME is also invested in the classification of different functional modules present in MGEs (Leplae *et al.*, 2004). Prophinder aims to detect genomic segments that are statistically enriched in phage-like genes. To do so, the algorithm analyzes a set of *n* consecutive coding sequences (CS) (in which *n* is defined by the user) and models it into a trial series: each CS can either be considered phage-like (success) or not phage-like (failure) (Lima-Mendez *et al.*, 2008). Binomial *P*-values are used to assess the risk of false positives. They define the probability of observing, by chance, *s* or more phage-like CSs in a set, according to the following formula:

$$P\_value = P\ (X \geq s) = \sum_{i=s}^{n} C_n^i p^i (1-p)^{n-i}$$

The probability of success, *p*, is determined by dividing the number of CSs considered phage-like by the total number of CSs on the set, thus inferring the average density of phage-like genes. Because the input genome assembly is screened in windows of *n* coding sequences, evaluating the entire input sequence requires multiple tests, implying the correction of the obtained *P*-values for multi-testing (Lima-Mendez *et al.*, 2008). The resulting expected number of false positives for a set of *T* tests (in which *T* depends on the number of coding sequences on the genome assembly and the user-defined window of analysis) is termed *E*-value, and its logarithmic transformation provides the significance index (*sig*) of the entire tested segment:

$$sig = -\log(E\_value) = -\log(P\_value \times T)$$

These *sig* values are stored in a matrix and negative values are discarded. The matrix is then scanned for detection of local maximum values, validating the corresponding segments as phage-like dense regions. These regions are then sorted in decreasing order of their *sig* values, and precedence of overlapping regions is determined according to predefined rules: regions that contain integrase genes precede over regions that do not, and regions with higher *sig* values precede over those with lower ones (Lima-Mendez *et al.*, 2008).

To counteract the natural tendency of this method to allow small prophages or prophage remnants to go undetected, rounds of selection can be iterative: the same scoring matrix is analyzed each time, but previously detected prophages are masked by setting their *sig* values to -1. Because negative values are not considered for analysis, this process allows for the detection of new maximum values and validation of new prophage sequences. The number of iterations can also be defined by the user (Lima-Mendez *et al.*, 2008).

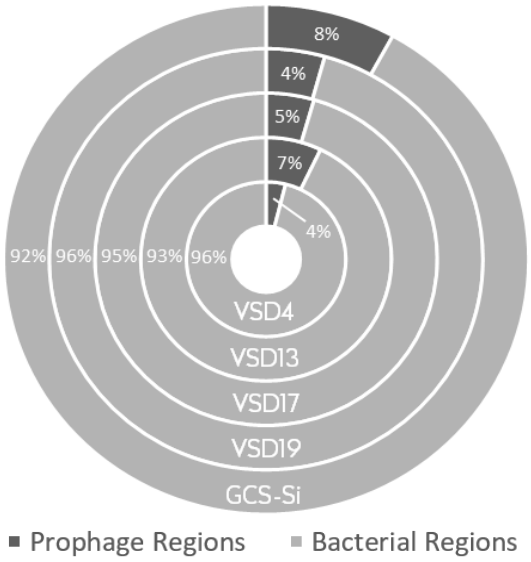
PhiSpy, on the other hand, is a weighted phage detection algorithm more geared towards the *de novo* discovery of phage regions. Rather than relying on homology with known phage homologs, it is based on the ranking of genomic regions by enrichment in predefined distinctive characteristics of prophages: protein length, transcription strand directionality, customized AT and GC skew, abundance of unique phage words[27], phage insertion points and similarity of phage proteins. Only the last two factors require sequence similarity to known phage genes. These metrics are calculated and then fed into a random forest classification algorithm that ranks segments of the input genome. PhiSpy is then less conservative than Prophinder, allowing the detection of more prophage sequences. However, this algorithm is prone to combine several short phage regions into a single large one, as well as reporting a single phage region with more than one detected integrase gene as more than one prophage. Additionally, its current random forest protocol does not yet allow for accurate determination of prophage start and end regions (Akhter *et al.*, 2012).

As expected, the percentage of prophage regions in each bacterial genome according to PhiSpy exceeds that of Prophinder; however, proportions are almost totally maintained, with strain GCS-Si presenting the most phage content, followed by strain VSD13, strains VSD19 and VSD17 (indistinguishable according to Prophinder) and finally strain VSD4. There seems to be some correlation with the results from the alignment of assemblies against their closest reference, where strains GCS-Si and VSD13 diverged the most (for further details consult **Supplementary Fig. 5, Appendix F**).

Considering the different valences of Prophinder and PhiSpy, and to improve the accuracy of phage prediction, only prophage sequences agreed upon by both prediction tools were considered for further analysis. However, because Prophinder provides more detailed information on detected prophages, its output holds primacy over that of PhiSpy in the current analysis. Results of consensual phage prediction are summarized in **Fig. 24.**

[27] A 'word' is defined by the authors as a set of 12 consecutive bp in a sequence. Within the algorithm, libraries for 'bacterial' and 'phage' words were created. Words present in the phage library and absent in the bacterial library are considered unique phage words.

A) Prophage Content



B) Prophage Prediction Summary

Strain	Prophage Designation	Size	Score
VSD4	vsd4_A	37308	0.9200
	vsd4_B	49964	0.7528
VSD13	vsd13_A	39771	1.1743
	vsd13_B	35787	0.9387
	vsd13_C	43489	0.8792
	vsd13_D	45226	0.7181
VSD17	vsd17_A	47068	0.9429
	vsd17_B	50000	0.8791
VSD19	vsd19_A	37014	0.8438
	vsd19_B	50091	0.7206
	vsd19_C	10795	0.3160
GCS-Si	gcs_A	43146	0.7089
	gcs_B	31407	0.6221
	gcs_C	13042	0.5763
	gcs_D	38643	0.5477
	gcs_E	38967	0.4749
	gcs_F	31689	0.0723

Fig. 24 - **Consensual prophage content in bacterial genome assemblies. A)** Percentage of prophage and bacterial regions in each genome assembly; **B)** Designation, size and Prophinder normalized score of detected prophages, in decreasing order of their scores. Prophinder normalized scores represent the *sig* values normalized based on the number of coding sequences in each prediction.

Consensual prophage content holds similar values to the previsions of Prophinder alone, reinforcing the idea that Prophinder seems to be more reliable than PhiSpy. In fact, Prophinder-only predictions were only detected in strains GCS-Si and VSD19, while PhiSpy had unique predictions on all strains. This does not imply inaccuracy *per se*, given that these predictions might indeed represent novel prophage sequences or phage remnants. Prophage content values of this magnitude are not unusual for species of the *Streptococcus* genus, with *S. pyogenes* strains containing up to 12% prophage sequences on their genomes (Canchaya *et al.*, 2003).

Most putative prophage sequences have sizes within the expected range for *Siphoviridae*, with some of its smallest members scoring about 21 000 bp in size (Hatfull and Hendrix, 2011). Predictions gcs_C and vsd19_C, however, are much smaller and may represent phage remnants present in their respective genomes. As for Prophinder score values, although predictions gcs_F, vsd19_C and gcsd_E present lower values compared to the remaining sequences, they were maintained throughout the analysis because they were confirmed by both tools.

Alongside prophage detection, RAST annotation files were also scanned for functions related to the bacteriophage resistome (such as the previously mentioned restriction-modification systems, CRISPR/Cas systems and abortive infection systems).

As represented in **Fig. 27**, both putative prophage sequences and resistome-associated sequences are widely distributed throughout their respective host genome assemblies. The presence of phage resistance mechanisms suggests interplay between the two counterparts, pointing towards an additional hypothesis as to why productive infection seems so elusive within these strains. To assess whether the failure in lysogeny lies in the defective nature of phage tails or simply in the success of defensive mechanisms, determining the completeness of both prophage sequences and resistome-associated sequences is a crucial step.

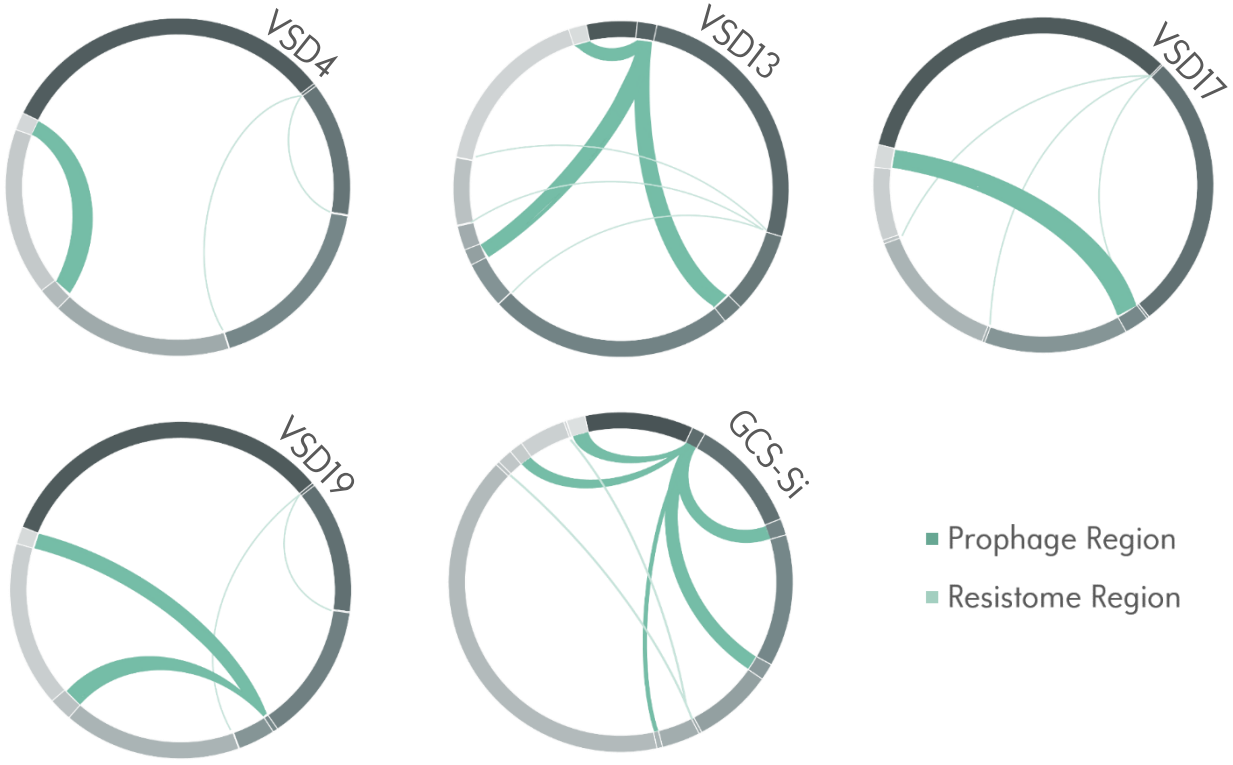


Fig. 25 – Distribution of prophage and resistome regions within bacterial genome assemblies. Putative prophage regions are connected through darker green links and regions associated with the bacteriophage resistome are connected through lighter green links. Size proportions between the highlighted regions and the genome assembly were maintained.

The prophage prediction process encompasses querying the genome assemblies against the specialized ACLAME database, which assigns functions to some of its hits. Thus, if all required lysogeny-associated functions were detected, the prophage sequence should be considered complete. However, as inferred from **Fig. 26**, the ratio between coding sequences with effective hits on the database and those with assigned functions is quite low, suggesting the need for additional analysis.

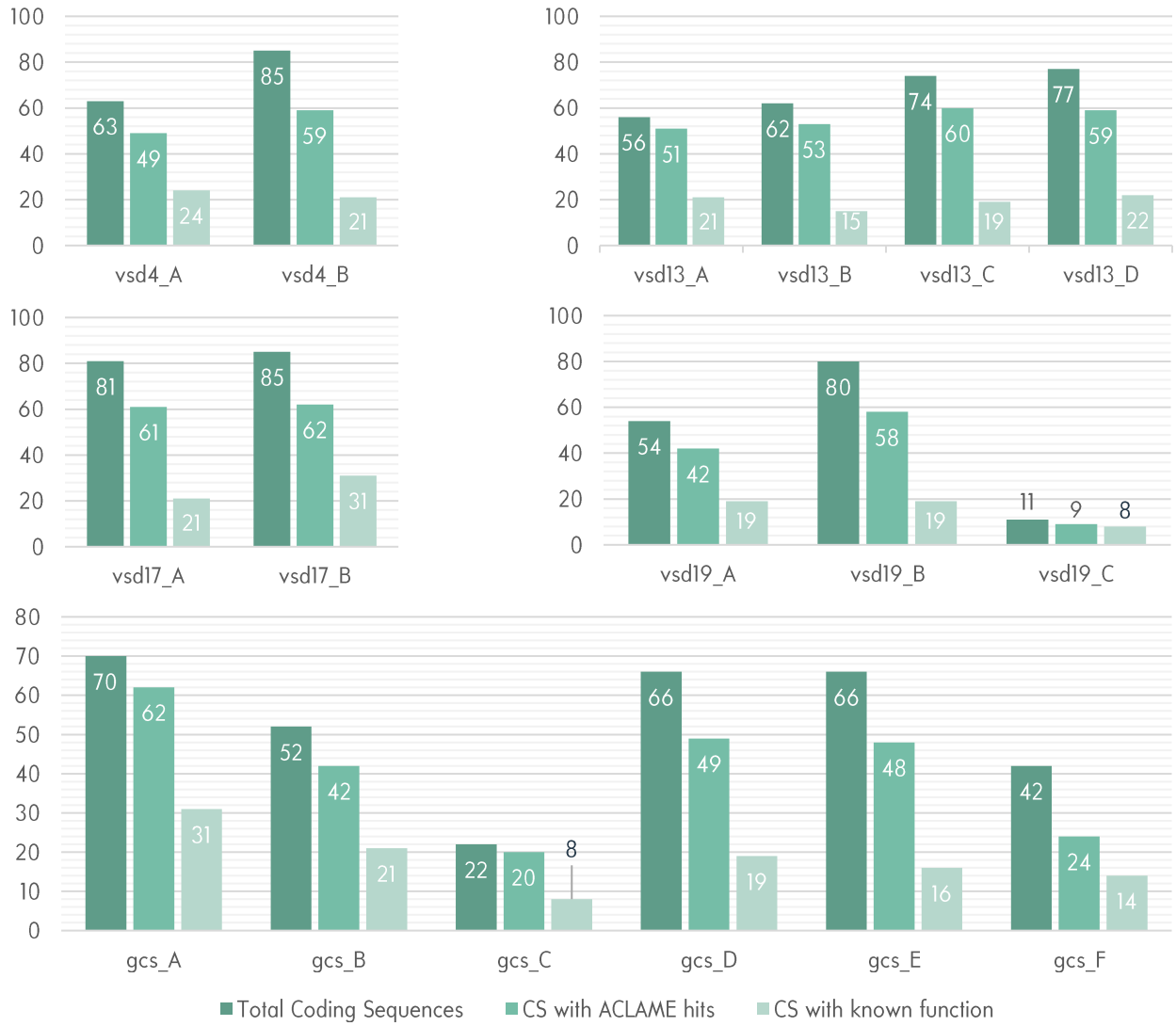


Fig. 26 - Coding sequences within each prophage. Overview of total coding sequences, sequences with hits in the ACLAME database as well as sequences with known functions within the ACLAME database.

3.5 Assessing completeness of putative prophages and resistome-associated sequences

To better comprehend the phage-host interplay within the five sequenced strains, annotation analysis, as well as sequence homology searches[28] were performed both on detected prophage sequences as well as bacteriophage resistome associated regions. Assessment of phage sequence completeness was performed by scanning annotation for all expected phage modules, as well as checking for homology with

[28] Sequences were queried using Nucleotide BLAST optimized for highly similar sequences (megablast). BLAST is available at: https://blast.ncbi.nlm.nih.gov/Blast.cgi

functional bacteriophages (Canchaya *et al.*, 2003). Results from the experiments in **Chapter III** indicate that phage genomes appear to be able to replicate and carry out the lysogeny cycle and that their encapsidation also occurs as expected, placing the main focus of this analysis on the integrity of tail modules. Results from the analysis of prophage completeness are represented in **Fig. 27**, with details on the modular integrity of each prediction detailed in **Fig. 28**.

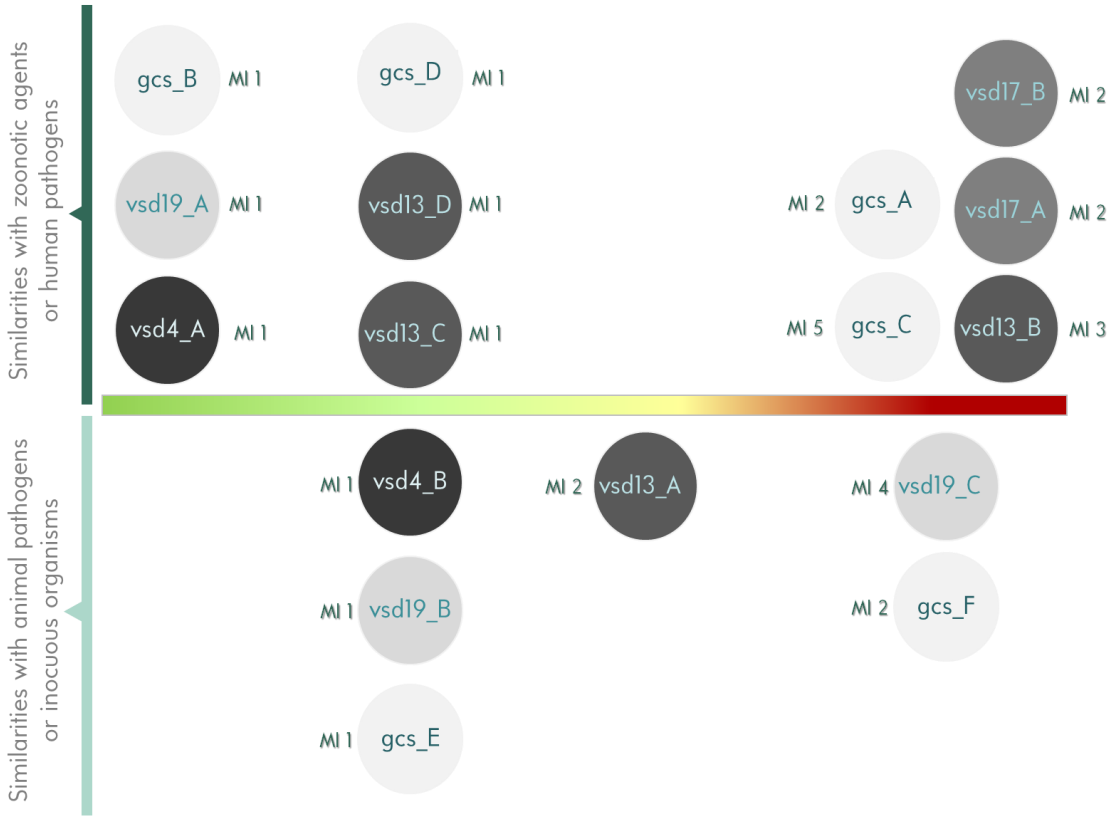


Fig. 27 - Integrity of putative prophage sequences. Sequence integrity is located in a color scale where red designates presumably defective phages (one or more functional modules missing from the sequence and no substantial homology with efficient prophage sequences was detected); yellow indicates phages with missing modules but substantial homology to fully functional prophages; light green indicates presumably functional phages (with annotated representatives for all required modules); darker green specifies phages with all present modules and homology to known functional bacteriophages, and thus the highest probability of proving fully functional. States of modular integrity (MI) are detailed on **Fig. 28**. Additionally, sequences were also divided based on their similarities to zoonotic agents and human pathogens or animal pathogens and inocuous organisms, to better assess their potential in aiding SDSD to cross the zoonotic agent barrier.

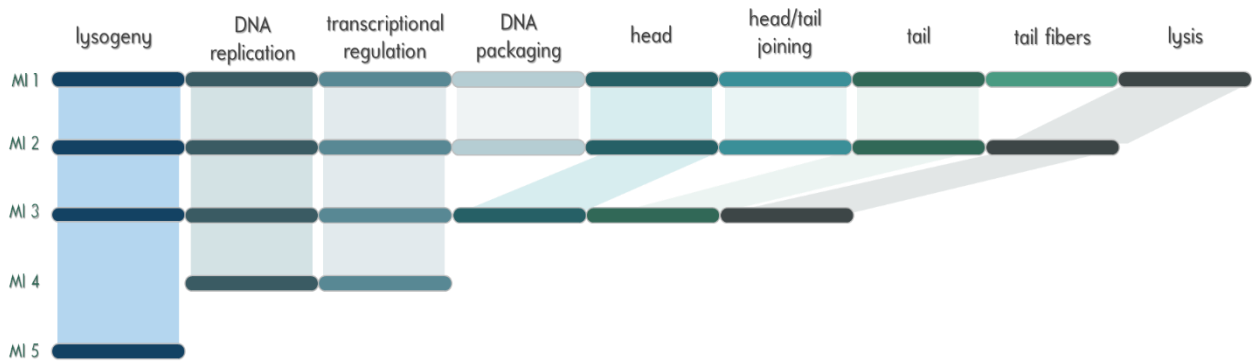


Fig. 28 – Modular integrity of predicted prophage sequences. Above are represented all required prophage functional modules and the five states of completion found through this analysis. State **MI 1** corresponds to complete phage sequences and encompasses predictions vsd4_A, vsd4_B, vsd13_C, vsd13_D, vsd19_A, vsd19_B, gcs_B, gcs_D and gcs_E; state **MI 2** was the most common one among incomplete phage sequences including predictions vsd13_A, vsd17_A, vsd17_B, gcs_A and gcs_F; state **MI 3** corresponds to prediction vsd13_B; state **MI 4** corresponds to prediction vsd19_C and state **MI 5** corresponds to prediction gcs_C.

In light of this analysis, strain VSD4 appears to harbor only functional sequences. Strain VSD13, contains three seemingly functional sequences and one prediction with missing modules. Strain VSD19 has two functional prophages and one sequence that is most likely a phage remnant, considering these results and the sequence size. Strain GCS-Si's phage patrimony appears to encompass half functional and half defective phages (with gcs_C being a phage remnant). Lastly, strain VSD17 appears to have no functional phage sequences. Several degrees of modular completeness were detected throughout the strains, with predictions vsd19_C and gcs_C showing the most drastic lack of functional modules. Lack of tail fibers alone hinders the process of infection and has been found to rend bacteriophage particles unable to successfully carry out their life cycle (Crawford and Goldberg, 1980).

As for these elements' role in bacterial pathogenicity towards humans, strains VSD13 and GCS-Si would be the most affected ones, since they report the most sequences related to human pathogens or zoonotic agents (*S. pyogenes, S. agalactiae, S. suis, S. dysgalactiae* subsp. *equisimilis*), followed by strains VSD17 and then VSD4 and VSD19. These predictions agree with previously mentioned findings about the increased virulence of strains GCS-Si and VSD13.

Even if these sequences exist only as genome-integrated phage remnants, they can still impact host fitness if the virulence genes present prove to be functional. Consequently, and to complete previous predictions, all sequences were scanned for the presence of possible virulence factors. Strain VSD13 has two phage-encoded copies of the *speK* gene, which encodes a streptococcal pyrogenic exotoxin, one of the main streptococcal superantigens, as well as a copy of streptodornase D (a streptococcal deoxyribonuclease) and an extracellular nuclease; strain VSD19 records a single *speK* copy as well as an extracellular nuclease;

strain GCS-Si displays a pathogenicity island (SAPIn2)[29]; strain VSD17 has two phage-encoded extracellular nucleases and a gene encoding the zeta toxin[30]; finally, strain VSD4 shows no phage-encoded sequences with virulence attributes. These results strengthen the idea that, although bacteriophages are not entirely responsible for a strain's virulence repertoire, they do contribute towards enriching host fitness and pathogenicity.

Integrity of resistome-associated sequences, on the other hand, was confirmed by checking annotation files against expected system structures (Blumenthal and Cheng, 2002; Makarova *et al.*, 2012; O'Connor *et al.*, 1999). Results from this analysis are depicted on **Fig. 29**.

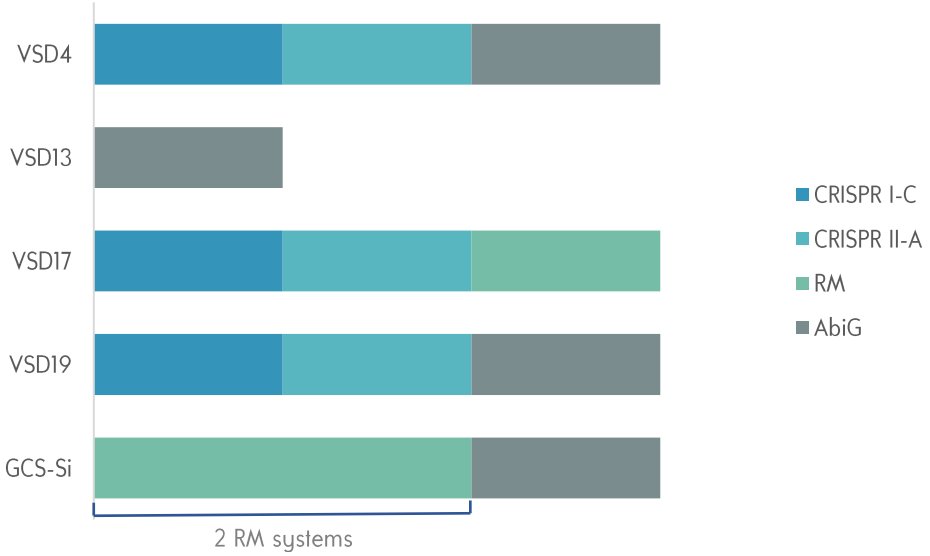


Fig. 29 - Bacteriophage resistome of SDSD strains. Representation of the bacteriophage resistome inferred through annotation file analysis and BLAST-mediated homology searches.

Restriction-modification mechanisms are possibly the most ubiquitous of all those related to the bacteriophage resistome, and thus the easiest for phages to circumvent given that the probability of exposure to these mechanisms is substantially high (Labrie *et al.*, 2010). In fact, bacteriophages have developed a plethora of ways to override this defensive action, including the encoding of methylases; phage-encoded methylases have been found in predictions vsd4_A, vsd17_A, vsd19_A, gcs_E and gcs_F.

CRISPR/Cas systems, for their mode of action and scarcer nature, might represent more of a challenge for bacteriophages; nevertheless, ways to bypass them do exist (as mentioned in **Chapter I**) such

[29] The SAPIn2 pathogenicity island has been found in *Staphylococcus aureus* and *Streptococcus pyogenes* genomes, and has been shown to encode proteins of the superantigen superfamily (Arcus *et al.*, 2002).

[30] The zeta toxin represents the toxin module a Toxin-Antitoxin system belonging to the epsilon/zeta TA family. Members of this family are routinely found in the genomes of pathogenic bacteria and are able to promote the host's virulence (Mutschler *et al.*, 2011).

as point mutations in spacer sequences, which due to their precise nature, could not be object of the present analysis.

As for the AbiG system, which affects phage RNA transcription (as described in **Chapter I**), although no specific counter-resistance mechanisms were described, phage susceptibility to AbiG action is variable (O'Connor *et al.*, 1999).

While these three system categories are considered fundamental in bacterial resistance to phage infection, there are additional ways through which bacterial hosts defend themselves. Production of extracellular matrix such as hyaluronan, not only protects bacteria against severe environmental conditions, but provides a barrier between phages and bacterial receptors. This is quite common amongst streptococci, resulting in phage evolution towards the production of hyaluronidases to counteract this defensive mechanism (Labrie *et al.*, 2010). These phage-encoded enzymes were found throughout all sequenced genomes (predictions vsd4_B, vsd13_D, vsd17_A, vsd17_B, vsd19_B, gcs_A). An overview of the different putative prophage sequences as well as their virulence and resistance features can be found on **Supplementary Table 3, Appendix F**.

Because prophage genomes are fairly flexible, their evolution in response to the selective pressures of resistance mechanisms is fast, meaning that no resistance mechanism is universally efficient. As such, the best defensive approach maybe the rotation between different mechanisms and no fixed combination of resistome sequences outperforms others indefinitely (Durmaz and Klaenhammer, 1995). Attempts to theoretically predict strain resistance to bacteriophages from this data alone are then limited. For example, strains VSD4 and VSD19 share the same bacteriophage resistome and nonetheless, strain VSD4 was found to be resistant to all bacteriophages in infection assays, while strain VSD19 acted as a successful host for infection in some of the experiments performed during the first Strep project. Their resistome relies heavily on CRISPR systems, whose efficiency highly depends on the host's previous exposure to viral infection, which could help explain the differences in actual resistance to phage infection. However, based on the diversity of mechanisms alone, strains VSD13 and GCS-Si should prove less resistant given that they encode only one and two mechanisms respectively; these results agree with prophage content predictions, which determined that these two strains have the biggest share of prophage sequences within their genome.

4. CONCLUSIONS

WGS was performed on SDSD strains to answer the question left by **Chapter III** results: is the lack of productive phage infection caused by defective phage tails or other factors? Sequencing results revealed indeed putative prophage sequences lacking tail components (and additional functional modules, in some

cases), as well as sequences appearing to be fully functional at a genomic level. Beyond analyzing prophage sequences, performing WGS on the tested strains allowed a glimpse at the host's side of phage infection, stressing the complexity of the phage-host evolutionary arms-race. By utilizing a third-generation sequencing methodology, recovery of a more trustworthy phage modular structure without sequencing individual phage genomes was achieved. Beyond gauging the genomic state of integrated prophage sequences, WGS allowed the characterization of the host's bacteriophage resistome; the variety in terms of prophage range, as well as combinations of resistome-associated systems, attests to high plasticity of streptococcal genomes. Sequencing results suggest that lack of productive infection can then be attributed to not one, but two main causes: phage defectiveness and lack of phage counter-resistance to bacterial defenses.

Although WGS provides a preliminary outlook into this multi-layered question, it appears to be an informative one, given that a degree of correlation between sequencing data and experimental observations can already be established and is at its strongest with data concerning strains VSD13 and GCS-Si. Their weaker bacteriophage resistome and higher prophage content are in agreement, as are their bolder phage-encoded virulence content and reports of increased virulence in comparison to more typical SDSD strains. Moreover, this data also suggests that strains VSD17 (although somewhat permissive) and VSD19 were not suitable as hosts for infection assays, given their phage repertoire and resistome content. Most of all, sequencing data points towards the hypothesis posed during **Chapter I**: that crosstalk between known streptococcal human pathogens, zoonotic agents and SDSD strains does occur and can indeed enhance their pathogenic potential towards new mammalian hosts. As anticipated, this MGE interplay seems to involve *S. pyogenes* and well-known elements of its virulence gene repertoire and seems to substantially influence the pathogenicity of SDSD strains involved.

It is worthy of notice that, even for well-known *Streptococcus* phages, such as phage A25 for example, genome annotation is not extensively detailed, hindering the process of comparison and its results. As such, and considering the direct effect that sequencing error has on downstream analysis, awareness to the use of multiple tools and data mining strategies was maintained throughout this work. Setbacks related to the error-prone character of the sequencing methodology were expected and are particularly visible at the annotation level, where the indels and frameshift errors accentuate the miscalls and redundancy of this process.

Although a considerable number of sequences with unattributed function remained, this is not solely the reflex of sequencing limitations, but also of the untapped potential that lies within phage genomes. As mentioned in **Chapter I**, phages are thought to represent the largest reservoir of unexplored genes available,

and the need to further pursue research in this area became clear in the course of this work. WGS data does, nevertheless, provide very helpful input as to which strains and prophage sequences seem to be the most promising, establishing important guidelines on phage infection and bacterial virulence on tested strains and highlighting the dynamic nature of phage-bacterium interactions.

These WGS experiments and analysis thus far are but the basis for the characterization of MGE within these strains. Steps to improve data quality, such as manual hybrid correction with second-generation short high-quality reads, can be implemented to diminish annotation mishaps and redundancy. Manual parsing of annotation results, as well as the inclusion of additional phage prediction tools and alternative analytical pipelines, may also contribute towards taking full advantage of the generated sequencing data.

As mentioned in **Chapter I** of this dissertation, bacteriophages can also acquire plasmid form, although integration into the bacterial chromosome is more common. In order to assess their existence, the unassembled files resulting from Canu assembly would have to be annotated and manually parsed before they could undergo phage prediction analysis, as well as homology searches with other bacteriophages. Given that for every strain there are over 250 unassembled sequences, this would signify applied the aforementioned pipeline to well over 1250 sequences. Due to time constraints, it was not possible to conduct such an analysis within the bounds of this dissertation. Even so, its interest should not be undermined, as it could complement these results in terms of MGE characterization.

Because of its inherent versatility, this data can also be exploited for purposes other than the study of phages, representing a substantial asset in the study of these SDSD strains. The evolution of ONT's sequencing technology and its accompanying analysis tools is remarkably fast and promising, with a substantial jump in performance being observed throughout sequencing experiments within this dissertation. It is expectable that, in a not-so-distant future, issues that currently represent analytical hurdles for nanopore data will be overcome by technological improvement, further extending the potential of nanopore sequencing.

CHAPTER V. General conclusions and future remarks

Throughout this dissertation, several different approaches were employed in an attempt to understand and characterize phage-bacterium dynamics in the given SDSD population. Although no major deviations from the broad themes and goals of the dissertation outlines occurred, results obtained in **Chapter II** forced the reassessment of the initially proposed strategy.

The usage of classical induction/infection assays proved to be extremely time-consuming and quite laboursome, considering the array of conditions and types of assays tested. While results obtained in this manner are quite conclusive when they are positive – confirming not only the presence of phages but successful lysogeny as well – a negative outcome should not be held as definitive, as proven in the course of this work. These methods also have the disadvantage of relying mostly on visual confirmation of phage plaques (with an exception, of course, for essays performed in liquid media), which introduces an underlying degree of subjectivity to assay results. Consequently, disparities between different experiments, as those observed between the first and current Strep projects and within the current attempts themselves, may have their roots in this issue. These experiments did, however, raise the main question explored during the present work and proved useful in guiding subsequent efforts.

Bacteriophage DNA extraction and AFM visualization were valuable in determining phage presence and physical integrity. Results from the previous Chapter guided alterations done to standard DNA extraction and microscopy preparation protocols, which elongated the process up to 5 consecutive days but proved mostly effective in preserving bacteriophages. Even with extended protocols, this approach was not nearly as time consuming as the first one and demonstrated to be equally as informative. The possibility to observe phage particles without the extensive adulterations required by most standard microscopy preparation protocols proved to be an important asset in evaluating integrity as close to physiological conditions as possible. Seemingly contradictory but complementary results between **Chapters II** and **III** represented the turning point at which the complexity of this theme became blatantly clear, as did the need to integrate different strategies to better address this question.

WGS was by far the approach that provided the largest wealth of information. Because of the non-directed nature of this methodology, it provides data that goes beyond the user-defined scope - this is a major advantage, given that this dissertation is part of a larger project which aims to characterize the studied strains beyond phage interactions. Nanopore sequencing, particularly, seemed to fit in quite well with the purpose of the present work: its long-read generation abilities mesh well with the modular-based evolutionary mechanisms of bacteriophages, allowing an easier recovery of correct sequence structure when

compared to second-generation technologies. The MinION is also fast, user-friendly and easy to implement, giving the user more control over the sequencing process and subsequent data analysis and the freedom to tailor the process according to the task at hand. The flexibility of this system opens new possibilities, such as the coupling of the phage DNA extraction protocol defined during this dissertation with a sequencing run. This would allow sequencing of the phage genome in its virion state and a subsequent comparison to the prophage genome. A run of viral genomes alone would also mean that, without the burden of a bacterial genome, phage DNA would be sequenced with a lot more depth than it was during WGS, providing further insight into the phages' genomic structure. This does, however, require optimization of DNA extraction and library preparation protocols. Revisiting infection assays and phage particle isolation in the light of WGS acquired information, using it to guide the selection of suitable hosts, could also prove interesting.

It is not the the potential of each employed approach by itself, but rather the integration of both classical and computational-based methodologies that shows true promise in the characterization of phage-host dynamics, and in unveiling its complexity. Far more than a question of lethargy of infection or failure to infect, bacteriophage interactions investigated during this dissertation proved to be a mix of both at the very least. However, such an assessment was only possible when looking at different sets of results as a whole. Although sequencing data represents a vast repository of information, it should be a stepping stone in the study of phage-bacterium interactions. Phage-host interplay is, of course, not exclusively conditioned by the genomic integrity of both counterparts, but depends on the correct expression, assembly and interaction of the components that mediate this process, most of which remains uncharacterized for these SDSD strains. An integrated omics approach could then be a fitting strategy to further study these strains not only in terms of phage interaction, but also beyond the bounds of MGE repertoire.

References

Abdelsalam, M., Asheg, A., and Eissa, A.E. (2013). *Streptococcus dysgalactiae*: An emerging pathogen of fishes and mammals. Int. J. Vet. Sci. Med. *1*, 1–6.

Abedon, S.T. (2009). Phage evolution and ecology (Elsevier Inc.).

Abedon, S.T., and Yin, J. (2009). Volume 1: Isolation, Characterization and Interactions. In Bacteriophages: Methods and Protocols, M.R.J. Clokie, and A.M. Kropinski, eds. (Humana Press), p.

Akhter, S., Aziz, R.K., and Edwards, R.A. (2012). PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. Nucleic Acids Res. *40*, 1–13.

Aksyuk, A.A., Bowman, V.D., Kaufmann, B., Fields, C., Klose, T., Holdaway, H.A., Fischetti, V.A., and Rossmann, M.G. (2012). Structural investigations of a *Podoviridae* streptococcus phage C1, implications for the mechanism of viral entry. Proc. Natl. Acad. Sci. *109*, 14001–14006.

Alessandrini, A., and Facci, P. (2005). AFM: a versatile tool in biophysics. Meas. Sci. Technol. *16*, R65–R92.

Anderson, B., Rashid, M.H., Carter, C., Pasternack, G., Rajanna, C., Revazishvili, T., Dean, T., Senecal, A., and Sulakvelidze, A. (2011). Enumeration of bacteriophage particles: Comparative analysis of the traditional plaque assay and real-time QPCR- and nanosight-based assays. Bacteriophage *1*, 86–93.

Arcus, V.L., Langley, R., Proft, T., Fraser, J.D., and Baker, E.N. (2002). The Three-dimensional structure of a superantigen-like protein, SET3, from a pathogenicity island of the *Staphylococcus aureus* genome. J. Biol. Chem. *277*, 32274–32281.

Arkhangelsky, E., and Gitis, V. (2008). Effect of transmembrane pressure on rejection of viruses by ultrafiltration membranes. Sep. Purif. Technol. *62*, 619–628.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., et al. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics *9*, 75.

Banks, D.J., Porcella, S.F., Barbian, K.D., Beres, S.B., Philips, L.E., Voyich, J.M., DeLeo, F.R., Martin, J.M., Somerville, G.A., and Musser, J.M. (2004). Progress toward Characterization of the Group A *Streptococcus* Metagenome: Complete Genome Sequence of a Macrolide-Resistant Serotype M6 Strain. J. Infect. Dis. *190*, 727–738.

Banks, D.J., Lei, B., and Musser, J.M. (2003). Prophage Induction and Expression of Prophage-Encoded Virulence Factors in Group A *Streptococcus* Serotype M3 Strain MGAS315. Infect. Immun. *71*, 7079–7086.

Batovska, J., Lynch, S.E., Rodoni, B.C., Sawbridge, T.I., and Cogan, N.O. (2017). Metagenomic arbovirus detection using MinION nanopore sequencing. J. Virol. Methods *249*, 79–84.

Beniac, D.R., Siemens, C.G., Wright, C.J., and Booth, T.F. (2014). A filtration based technique for simultaneous SEM and TEM sample preparation for the rapid detection of pathogens. Viruses *6*, 3458–3471.

Bentley, R.W., Leigh, J.A., and Collins, M.D. (1991). Intrageneric structure of *Streptococcus* based on comparative analysis of small-subunit rRNA sequences. Int J Syst Bacteriol *41*, 487–494.

Blumenthal, R.M., and Cheng, X. (2002). Restriction-Modification Systems (New York, USA: John Wiley & Sons, Inc.).

Botstein, D. (1980). A Theory of Modular Evolution for Bacteriophages. Ann. N. Y. Acad. Sci. *354*, 484–491.

Boyd, E.F., and Brüssow, H. (2002). Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. Trends Microbiol. *10*, 521–529.

Boža, V., Brejová, B., and Vinař, T. (2016). DeepNano: Deep Recurrent Neural Networks for Base Calling in MinION Nanopore Reads. PLoS One *12*, e0178751.

Brandt, C.M., and Spellerberg, B. (2009). Human Infections Due to *Streptococcus dysgalactiae* subsp. *equisimilis*. Clin. Infect. Dis. *49*, 766–772.

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

Calvinho, L.F., Almeida, R.A., and Oliver, S.P. (1998). Potential virulence factors of *Streptococcus dysgalactiae* associated with bovine mastitis. Vet. Microbiol. *61*, 93–110.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brüssow, H. (2003). Prophage genomics. Microbiol. Mol. Biol. Rev. *67*, 238–276, table of contents.

Conesa, A., and Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int. J. Plant Genomics *2008*, 619832.

Crawford, J.T., and Goldberg, E.B. (1980). The function of tail fibers in triggering baseplate expansion of bacteriophage T4. J. Mol. Biol. *139*, 679–690.

Cumby, N., Davidson, A.R., and Maxwell, K.L. (2012). The moron comes of age. Bacteriophage *2*, 225–228.

Davies, M.R., McMillan, D.J., Van Domselaar, G.H., Jones, M.K., and Sriprakash, K.S. (2007). Phage 3396 from a *Streptococcus dysgalactiae* subsp. *equisimilis* pathovar may have its origins in *Streptococcus pyogenes*. J. Bacteriol. *189*, 2646–2652.

Deschamps, S., Mudge, J., Cameron, C., Ramaraj, T., Anand, A., Fengler, K., Hayes, K., Llaca, V., Jones, T.J., and May, G. (2016). Characterization, correction and de novo assembly of an Oxford Nanopore genomic dataset from *Agrobacterium tumefaciens*. Sci. Rep. *6*, 28625.

Desiere, F., McShan, W.M., van Sinderen, D., Ferretti, J.J., and Brüssow, H. (2001). Comparative genomics reveals close genetic relationships between phages from dairy bacteria and pathogenic streptococci: evolutionary implications for prophage-host interactions. Virology *288*, 325–341.

Durmaz, E., and Klaenhammer, T.R. (1995). A Starter Culture Rotation Strategy Incorporating Paired Restriction/ Modification and Abortive Infection Bacteriophage Defenses in a Single *Lactococcus lactis* Strain. Appl. Environ. Microbiol. *61*, 1266–1273.

Facklam, R. (2002). What happened to the streptococci: overview of taxonomic and nomenclature changes. Clin. Microbiol. Rev. *15*, 613–630.

Flusberg, B.A., Webster, D., Lee, J., Travers, K., Olivares, E., Clark, A., Korlach, J., Turner, S.W., Biosciences, P., Drive, A., et al. (2010). Direct detection of DNA methylation during single-molecule, real- time sequencing. Nat. Methods *7*, 461–465.

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

Fortier, L.-C., and Sekulovic, O. (2013). Importance of prophages to evolution and virulence of bacterial pathogens. Virulence *4*, 354–365.

Fulde, M., and Valentin-Weigand, P. (2012). Epidemiology and Pathogenicity of Zoonotic Streptococci. In Current Topics in Microbiology and Immunology, pp. 49–81.

Genteluci, G.L., Silva, L.G., Souza, M.C., Glatthardt, T., de Mattos, M.C., Ejzemberg, R., Alviano, C.S., Figueiredo, A.M.S., and Ferreira-Carvalho, B.T. (2015). Assessment and characterization of biofilm formation among human isolates of *Streptococcus dysgalactiae* subsp. *equisimilis*. Int. J. Med. Microbiol. *305*, 937–947.

Gera, K., and McIver, K.S. (2013). Laboratory growth and maintenance of *Streptococcus pyogenes* (The Group A *Streptococcus*, GAS). Curr. Protoc. Microbiol. 1–14.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351.

Goyal, P., Krasteva, P. V, Van Gerven, N., Gubellini, F., Van den Broeck, I., Troupiotis-Tsaïlaki, A., Jonckheere, W., Péhau-Arnaudet, G., Pinkner, J.S., Chapman, M.R., et al. (2014). Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. Nature *516*, 250–253.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.

Haenni, M., Saras, E., Bertin, S., Leblond, P., Madec, J.Y., and Payot, S. (2010). Diversity and mobility of integrative and conjugative elements in bovine isolates of *Streptococcus agalactiae*, *S. dysgalactiae* subsp. *dysgalactiae*, and *S. uberis*. Appl. Environ. Microbiol. *76*, 7957–7965.

Halasa, T., Huijps, K., Østerås, O., and Hogeveen, H. (2007). Economic effects of bovine mastitis and mastitis management: a review. Vet. Q. *29*, 18–31.

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

Hatfull, G.F., and Hendrix, R.W. (2011). Bacteriophages and their genomes. Curr. Opin. Virol. *1*, 298–303.

Heather, J.M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics *107*, 1–8.

Heng, N.C.K., Ragland, N.L., Swe, P.M., Baird, H.J., Inglis, M.A., Tagg, J.R., and Jack, R.W. (2006). Dysgalacticin: A novel, plasmid-encoded antimicrobial protein (bacteriocin) produced by *Streptococcus dysgalactiae* subsp. *equisimilis*. Microbiology *152*, 1991–2001.

Holden, M.T.G., Heather, Z., Paillot, R., Steward, K.F., Webb, K., Ainslie, F., Jourdan, T., Bason, N.C., Holroyd, N.E., Mungall, K., et al. (2009). Genomic evidence for the evolution of *Streptococcus equi* host restriction, increased virulence, and genetic exchange with human pathogens. PLoS Pathog. *5*.

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. Science *327*, 167–170.

Hulo, C., Masson, P., Toussaint, A., Osumi-Sutherland, D., De Castro, E., Auchincloss, A.H., Poux, S., Bougueleret, L., Xenarios, I., and Le Mercier, P. (2017). Bacterial Virus Ontology; Coordinating across databases. Viruses *9*.

Hyman, P., and Abedon, S.T. (2010). Bacteriophage host range and bacterial resistance (Elsevier Inc.).

Ingrey, K.T., Ren, J., and Prescott, J.F. (2003). A fluoroquinolone induces a novel mitogen-encoding bacteriophage in *Streptococcus canis*. Infect. Immun. *71*, 3028–3033.

Ivanovska, I., Wuite, G., Jönsson, B., and Evilevitch, A. (2007). Internal DNA pressure modifies stability of WT phage. Proc. Natl. Acad. Sci. U. S. A. *104*, 9603–9608.

Iwasaki, H., Takahagi, M., Shiba, T., Nakata, A., and Shinagawa, H. (1991). *Escherichia coli* RuvC protein is an endonuclease that resolves the Holliday structure. EMBO J. *10*, 4381–4389.

Iyer, V.N., and Szybalski, W. (1963). A Molecular Mechanism of Mitomycin Action: Linking of Complementary Dna Strands. Proc. Natl. Acad. Sci. U. S. A. *50*, 355–362.

Jain, M., Tyson, J.R., Loose, M., Ip, C.L.C., Eccles, D.A., O'Grady, J., Malla, S., Leggett, R.M., Wallerman, O., Jansen, H.J., et al. (2017). MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. F1000Research *6*, 760.

Jalili, N., and Laxminarayana, K. (2004). A review of atomic force microscopy imaging systems: Application to molecular metrology and biological sciences. Mechatronics *14*, 907–945.

Jensen, A., and Kilian, M. (2012). Delineation of *Streptococcus dysgalactiae*, its subspecies, and its clinical

and phylogenetic relationship to *Streptococcus pyogenes*. J. Clin. Microbiol. *50*, 113–126.

Jordal, S., Glambek, M., Oppegaard, O., and Kittang, B.R. (2015). New tricks from an old cow: Infective endocarditis caused by *Streptococcus dysgalactiae* subsp. *dysgalactiae*. J. Clin. Microbiol. *53*, 731–734.

Koh, T.H., Sng, L.H., Yuen, S.M., Thomas, C.K., Tan, P.L., Tan, S.H., and Wong, N.S. (2009). Streptococcal cellulitis following preparation of fresh raw seafood. Zoonoses Public Health *56*, 206–208.

Koren, S., and Phillippy, A.M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr. Opin. Microbiol. *23*, 110–120.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2016). Canu : scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. 1–35.

Kroemer, S., Galland, D., Guérin-Faublée, V., Giboin, H., and Woehrlé-Fontaine, F. (2012). Survey of marbofloxacin susceptibility of bacteria isolated from cattle with respiratory disease and mastitis in Europe. Vet. Rec. *170*, 53.

Kuhl, S., Abedon, S.T., and Hyman, P. (2012). Diseases caused by phages. In Bacteriophages in Health and Disease, P. Hyman, and S.T. Abedon, eds. (Wallingford: CABI), pp. 21–33.

Kuznetsov, Y.G., Chang, S.-C., Credaroli, A., and McPherson, A. (2013). Unique Tail Appendages of Marine Bacteriophages. Adv. Microbiol. *3*, 55–59.

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

Lavezzo, E., Barzon, L., Toppo, S., and Palù, G. (2016). Third generation sequencing technologies applied to diagnostic microbiology: benefits and challenges in applications and data analysis. Expert Rev. Mol. Diagn. *16*, 1011–1023.

Leplae, R., Hebrant, A., Wodak, S.J., and Toussaint, A. (2004). ACLAME: a CLAssification of Mobile genetic Elements. Nucleic Acids Res. *32*, D45-9.

Levine, M. (1961). Effect of Mitomycin C on Interactions between Temperate Phages and Bacteria. Virology *13*, 493–499.

Levy, S.E., and Myers, R.M. (2016). Advancements in Next-Generation Sequencing. Annu. Rev. Genomics Hum. Genet. *17*, 95–115.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008). Prophinder: A computational tool

for prophage prediction in prokaryotic genomes. Bioinformatics *24*, 863–865.

Loman, N.J., Quick, J., and Simpson, J.T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Doi.org 15552.

Los, M., and Wegrzyn, G. (2012). Pseudolysogeny. In Bacteriophages: Part A, M. Łobocka, and W.T. Szybalski, eds. (Academic Press), p. 396.

Lu, H., Giordano, F., and Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics. Proteomics Bioinformatics *14*, 265–279.

Lux, T., Nuhn, M., Hakenbeck, R., and Reichmann, P. (2007). Diversity of bacteriocins and activity spectrum in Streptococcus pneumoniae. J. Bacteriol. *189*, 7741–7751.

Madigan, M.T., Martinko, J.M., Bender, K.S., Buckley, D.H., and Stahl, D.A. (2015). Brock Biology of Microorganisms (Essex: Pearson Education).

Madigan, M.T., Martinko, J.M., Stahl, D., and Clark, D.P. (2010). Brock Biology of Microorganisms (13th Edition) (Benjamin Cummings).

Magi, A., Semeraro, R., Mingrino, A., Giusti, B., and D'Aurizio, R. (2017). Nanopore sequencing data analysis: state of the art, applications and challenges. Brief. Bioinform. 1–17.

Makarova, K.S., Brouns, S.J.J., Horvath, P., Sas, D.F., and Wolf, Y.I. (2012). Evolution and classification of the CRISPR-Cas systems. Nat. Rev. … *9*, 467–477.

Maruyama, F., Watanabe, T., and Nakagawa, I. (2016). *Streptococcus pyogenes* Genomics. In *Streptococcus pyogenes*: Basic Biology to Clinical Manifestations, J.J. Ferretti, D.L. Stevens, and V.A. Fischetti, eds. (Oklahoma City), pp. 136–203.

McShan, W.M., and Nguyen, S. V. (2016). The Bacteriophages of *Streptococcus pyogenes*. In *Streptococcus pyogenes*: Basic Biology to Clinical Manifestations, J.J. Ferretti, D.L. Stevens, and V.A. Fischetti, eds. (Oklahoma City), pp. 204–228.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nat. Rev. Genet. *11*, 31–46.

Mims, C., Playfair, J., Roitt, I., Wakelin, D., and Williams, R. (1998). Medical Microbiology (Mosby).

Mullan, W.M.A. (2002). Factor Affecting Plaque Formation.

Muralt, P. (2000). Ferroelectric thin films for micro-sensors and actuators: a review. J. Micromechanics Microengineering *10*, 136–146.

Murray, P.R., Rosenthal, K.S., and Pfaller, M.A. (2009). Medical Microbiology (Philadelphia: Mosby/Elsevier).

Mutschler, H., Gebhardt, M., Shoeman, R.L., and Meinhart, A. (2011). A Novel Mechanism of Programmed Cell Death in Bacteria by Toxin–Antitoxin Systems Corrupts Peptidoglycan Synthesis. PLoS Biol. *9*, e1001033.

Nguyen, S. V, and McShan, W.M. (2014). Chromosomal islands of *Streptococcus pyogenes* and related streptococci: molecular switches for survival and virulence. Front. Cell. Infect. Microbiol. *4*, 109.

O'Connor, L., Tangney, M., and Fitzgerald, G.F. (1999). Expression, regulation, and mode of action of the AbiG abortive infection system of *Lactococcus lactis* subsp. *cremoris* UC653. Appl. Environ. Microbiol. *65*, 330–335.

Park, M.J., Eun, I.-S., Jung, C.-Y., Ko, Y.-C., Kim, Y.-J., Kim, C.-K., and Kang, E.-J. (2012). *Streptococcus dysgalactiae* subspecies *dysgalactiae* infection after total knee arthroplasty: a case report. Knee Surg. Relat. Res. *24*, 120–123.

Pelkonen, S., Lindahl, S.B., Suomala, P., Karhukorpi, J., Vuorinen, S., Koivula, I., Väisänen, T., Pentikäinen, J., Autio, T., and Tuuminen, T. (2013). Transmission of *Streptococcus equi* subspecies *zooepidemicus* infection from horses to humans. Emerg. Infect. Dis. *19*, 1041–1048.

Pillet, F., Chopinet, L., Formosa, C., and Dague, É. (2014). Atomic Force Microscopy and pharmacology: From microbiology to cancerology. Biochim. Biophys. Acta - Gen. Subj. *1840*, 1028–1050.

Rato, M.G., Bexiga, R., Nunes, S.F., Vilela, C.L., and Santos-Sanches, I. (2010). Human group A streptococci virulence genes in bovine group C streptococci. Emerg. Infect. Dis. *16*, 116–119.

Rato, M.G., Bexiga, R., Florindo, C., Cavaco, L.M., Vilela, C.L., and Santos-Sanches, I. (2013). Antimicrobial resistance and molecular epidemiology of streptococci from bovine mastitis. Vet. Microbiol. *161*, 286–294.

Rato, M.G., Nerlich, A., Bergmann, R., Bexiga, R., Nunes, S.F., Vilela, C.L., Santos-Sanches, I., and Chhatwal, G.S. (2011). Virulence gene pool detected in bovine group C *Streptococcus dysgalactiae* subsp. *dysgalactiae* isolates by use of a group A *S. pyogenes* virulence microarray. J. Clin. Microbiol. *49*, 2470–2479.

Reuter, J.A., Spacek, D. V., and Snyder, M.P. (2015). High-Throughput Sequencing Technologies. Mol. Cell *58*, 586–597.

Richards, V.P., Palmer, S.R., Bitar, P.D.P., Qin, X., Weinstock, G.M., Highlander, S.K., Town, C.D., Burne, R.A., and Stanhope, M.J. (2014). Phylogenomics and the dynamic genome evolution f the genus *Streptococcus*. Genome Biol. Evol. *6*, 741–753.

Roma-Rodrigues, C., Alves-Barroco, C., Raposo, L.R., Costa, M.N., Fortunato, E., Baptista, P.V., Fernandes, A.R., and Santos-Sanches, I. (2016). Infection of human keratinocytes by *Streptococcus dysgalactiae* subspecies *dysgalactiae* isolated from milk of the bovine udder. Microbes Infect. *18*, 290–293.

Salazar, A.N., Vries, A.R.G. De, Broek, M. Van Den, De, P., Cortés, T., Brickwedde, A., Brouwers, N., Salazar, A.N., Vries, A.R.G. De, Broek, M. Van Den, et al. (2017). Nanopore sequencing enables near-complete de novo assembly of *Saccharomyces cerevisiae* reference strain CEN . PK113-7D considered co-first authors . 1 . Delft Bioinformatics Lab , Delft University of Technology , Delft , The 2 . Department of Biotechnolog. 1–35.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

Shimomura, Y., Okumura, K., Murayama, S.Y., Yagi, J., Ubukata, K., Kirikae, T., and Miyoshi-Akiyama, T. (2011). Complete genome sequencing and analysis of a Lancefield group G *Streptococcus dysgalactiae* subsp. *equisimilis* strain causing streptococcal toxic shock syndrome (STSS). BMC Genomics *12*, 17.

Silverman, B.W. (1986). Density estimation for statistics and data analysis (Chapman and Hall).

Snyder, L., Henkin, T.M., Peters, J.E., and Champness, W. (2013). Bacteriophages: Lytic Development, Genetics, and Generalized Transduction. In Molecular Genetics of Bacteria, (Washington: American Society of Microbiology), pp. 265–320.

Stern, A., and Sorek, R. (2012). The phage-host arms-race : Shaping the evolution of microbes. Bioessays *33*, 43–51.

Stoiber, M., and Brown, J. (2017). BasecRAWller: Streaming Nanopore Basecalling Directly from Raw Signal. Doi.org 133058.

Suzuki, H., Lefébure, T., Hubisz, M.J., Bitar, P.P., Lang, P., Siepe, A., and Stanhope, M.J. (2011). Comparative genomic analysis of the *Streptococcus dysgalactiae* species group: Gene content, molecular adaptation, and promoter evolution. Genome Biol. Evol. *3*, 168–185.

Szermer-Olearnik, B., Drab, M., Mąkosa, M., Zembala, M., Barbasz, J., Dąbrowska, K., and Boratyński, J. (2017). Aggregation/dispersion transitions of T4 phage triggered by environmental ion availability. J. Nanobiotechnology *15*, 32.

Takahashi, T., Ubukata, K., and Watanabe, H. (2011). Invasive infection caused by *Streptococcus dysgalactiae* subsp. *equisimilis*: Characteristics of strains and clinical features. J. Infect. Chemother. *17*, 1–10.

Tomasz, M. (1995). Mitomycin C: small, fast and deadly (but very selective). Chem. Biol. *2*, 575–579.

Vandamme, P., Pot, B., Falsen, E., Kersters, K., and Devriese, L. a (1996). Taxonomic study of lancefield streptococcal groups C, G, and L (Streptococcus dysgalactiae) and proposal of *S. dysgalactiae* subsp. *equisimilis* subsp. nov. Int. J. Syst. Bacteriol. *46*, 774–781.

Vasu, K., and Nagaraja, V. (2013). Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol. Mol. Biol. Rev. *77*, 53–72.

Verweij, J., and Pinedo, H.M. (1990). Mitomycin C: mechanism of action, usefulness and limitations. Anticancer. Drugs *1*, 5–13.

Vieira, V. V, Teixeira, L.M., Zahner, V., Momen, H., Facklam, R.R., Steigerwalt, A.G., Brenner, D.J., and Castro, A.C. (1998). Genetic relationships among the different phenotypes of *Streptococcus dysgalactiae* strains. Int. J. Syst. Bacteriol. *48 Pt 4*, 1231–1243.

Vojtek, I., Pirzada, Z. a, Henriques-Normark, B., Mastny, M., Janapatla, R.P., and Charpentier, E. (2008).

Lysogenic transfer of group A *Streptococcus* superantigen gene among Streptococci. J. Infect. Dis. *197*, 225–234.

Wagner, P.L., and Waldor, M.K. (2002). Bacteriophage Control of Bacterial Virulence. Infect. Immnunity *70*, 3985–3993.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*.

Wescombe, P.A., Heng, N.C., Burton, J.P., Chilcott, C.N., and Tagg, J.R. (2009). Streptococcal bacteriocins and the case for *Streptococcus salivarius* as model oral probiotics. Futur. Microbiol *4*, 819–835.

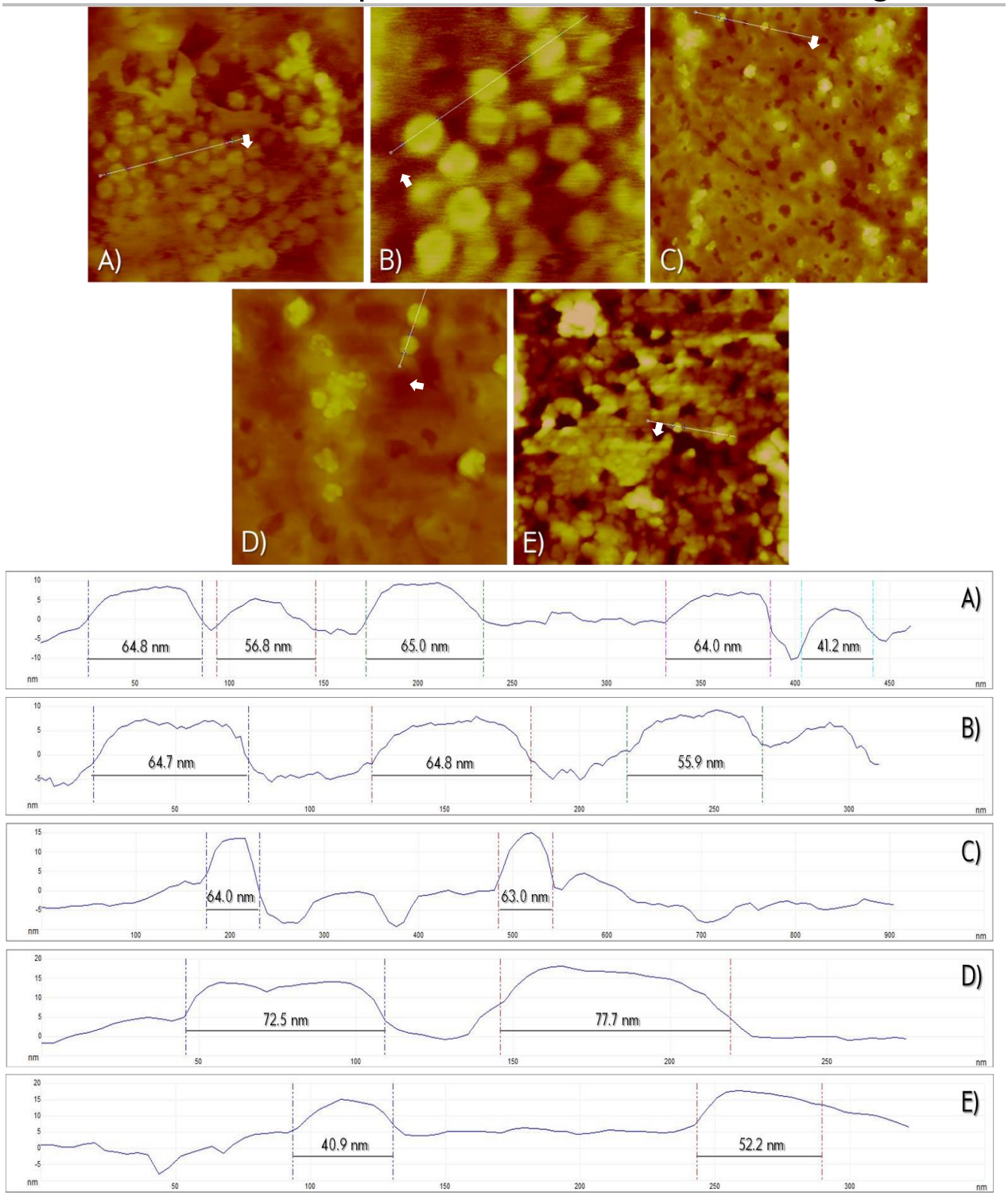
Whiley, R.A., and Hardie, J.M. (2009). Volume 3: The *Firmicutes* - Genus I *Streptococcus*. In Bergey's Manual of Systematic Bacteriology, P. Vos, G. Garrity, D. Jones, N.R. Krieg, W. Ludwig, F. Rainey, K.-H. Schleifer, and W.B. Whitman, eds. (New York, NY: Springer New York), pp. 655–711.

APPENDIX A. Bacterial Strain Information

Supplementary Table 1 – Bacterial strain information for Chapters II, III and IV gathered during the first Strep project. In the third column, SPYO stands for *S. pyogenes*, SDSD stands for *S. dysgalactiae* subsp. *dysgalactiae*. Virulence gene presence was assessed through PCR. The last column represents strain performance in terms of their permissiveness as hosts (H) and the infectivity of the phages isolated from said strains (Φ), in an assay with a total of 25 strains analyzed.

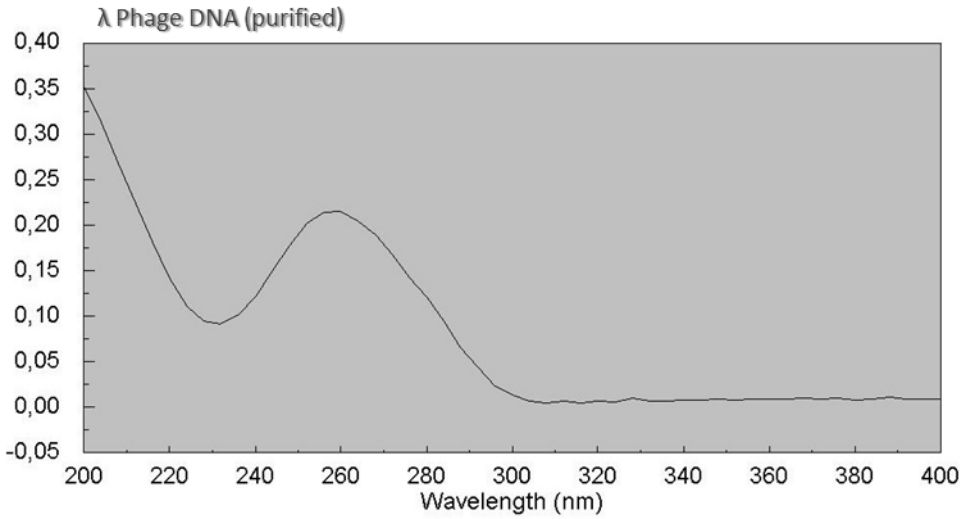
Chapter	Code	Species	Lancefield Group	Hemolysis	Host	Virulence Genes	Strep I infection performance
II	GAP8	SPYO	GAS	β-hemolysis	Human (child, 4 years old)	*speJ, speG, speB*	ϕ (23/25) H (22/25)
II	GAP58	SPYO	GAS	β-hemolysis	Human (adult, 21 years old)	*speJ, speA, smeZ, speG, speB, speF*	ϕ (24/25) H (25/25)
II	GAP88	SPYO	GAS	β-hemolysis	Human (adult, 33 years old)	*speJ, speA, speG, speB, speF*	ϕ (22/25) H (25/25)
II	GAP826	SPYO	GAS	β-hemolysis	Human (nd, nd years old)	ND	ϕ (25/25) H (23/25)
III and IV	VSD4	SDSD	GCS	α-hemolysis	Bovine (Holstein Friesian)	*speC, speK, spd1*	ϕ (25/25) H (0/25)
II	VSD5	SDSD	GCS	α-hemolysis	Bovine (Holstein Friesian)	*sdn*	ϕ (24/25) H (25/25)
II	VSD9	SDSD	GCS	α-hemolysis	Bovine (Holstein Friesian)	-	ϕ (22/25) H (25/25)
II, III and IV	VSD13	SDSD	GCS	α-hemolysis	Bovine (Holstein Friesian)	*speC, speK, speL, speM, spd1*	ϕ (23/25) H (25/25)
II, III and IV	VSD17	SDSD	GCS	α-hemolysis	Bovine (Holstein Friesian)	-	ϕ (23/25) H (25/25)
II, III and IV	VSD19	SDSD	GCS	α-hemolysis	Bovine (Holstein Friesian)	*speC, speK, spd1*	ϕ (25/25) H (24/25)
II, III and IV	GCS-Si	SDSD	GCS	α-hemolysis	Human (adult, nd years old)	*sagA*	ϕ (25/25) H (19/25)

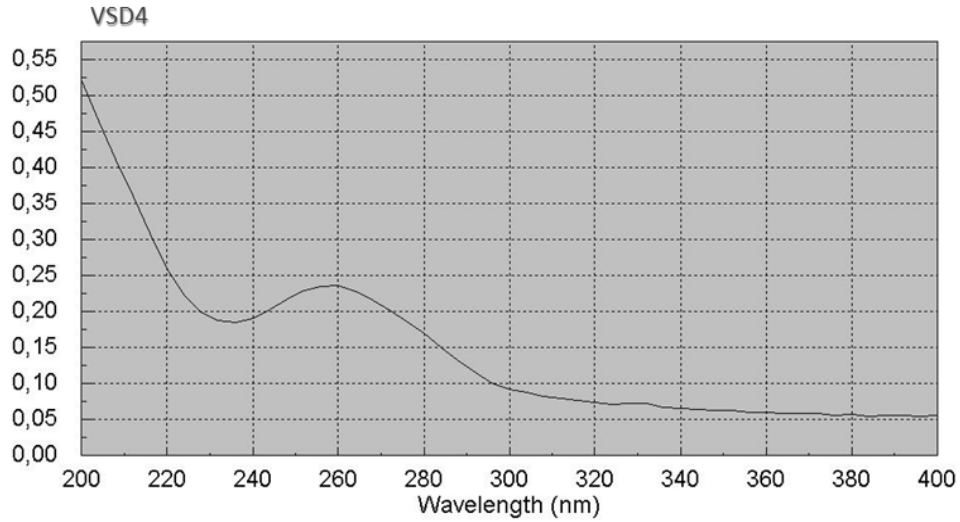
APPENDIX B. Capsid size determination through AFM

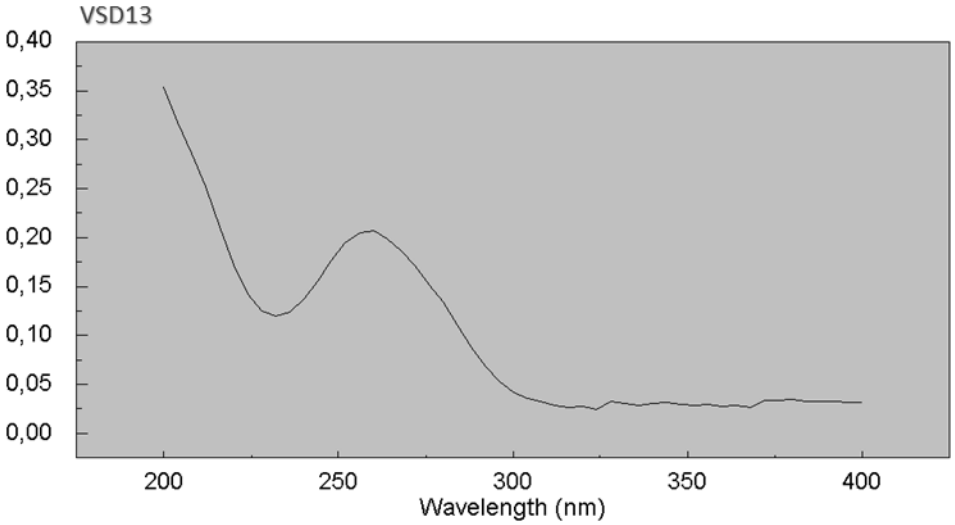


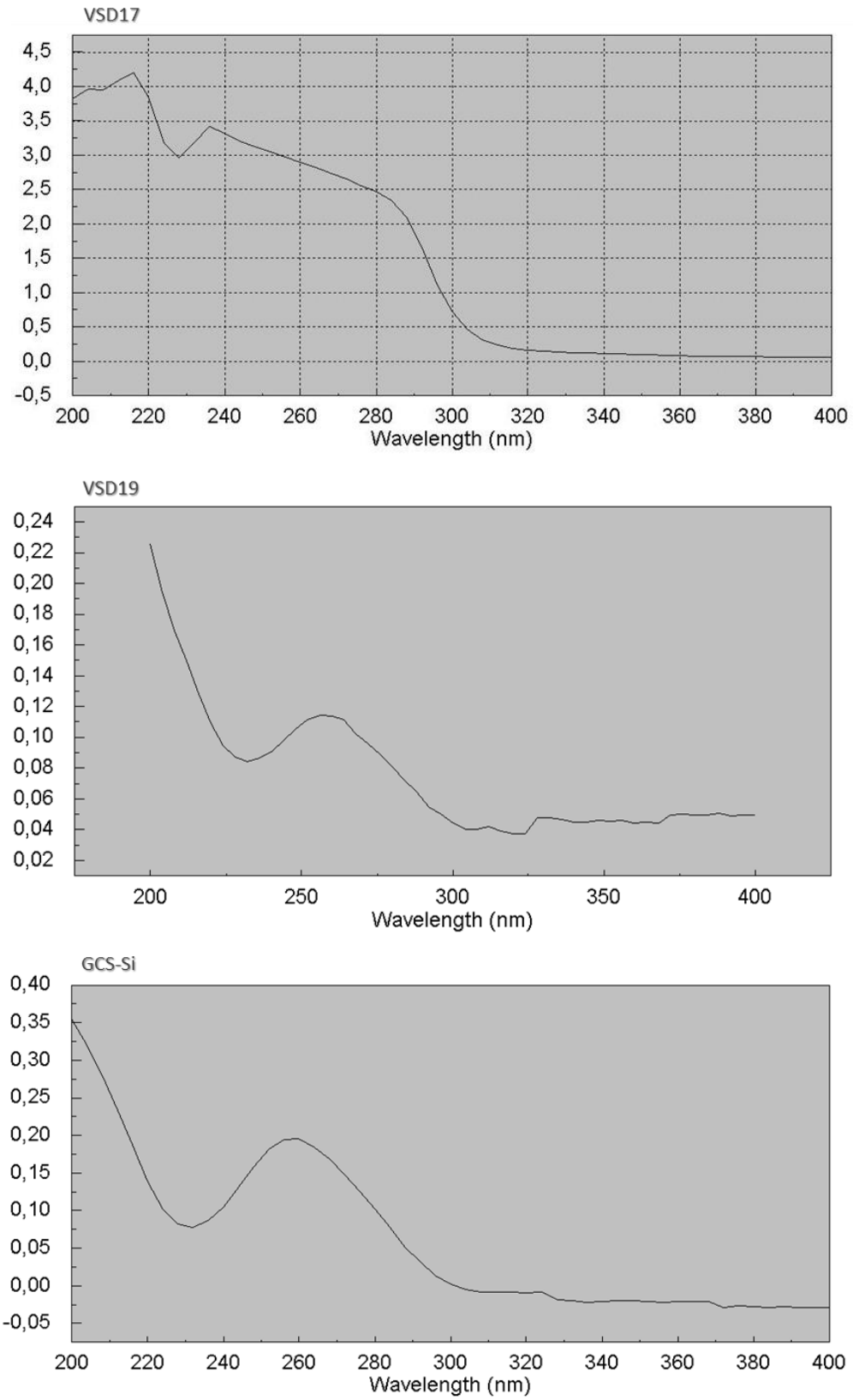
Supplementary Fig. 1 - Capsid size analysis. The average capsid size between all samples is 60,47 nm. Above are represented the images shown in the main text (with the sections used for measurements highlighted with white arrows) and corresponding surface graphs below. Different images were taken with different window size settings, which results in seemingly different images and graphs. Sets **A)** and **B)** correspond to sample VSD13 (average capsid size: 59,65 nm); sets **C)** and **D)** correspond to sample VSD17 (average capsid size: 69,3 nm); set **E)** corresponds to the T7 phage sample (estimated average capsid size; 46,55 nm; expected capsid size: 50 nm).

APPENDIX C. Genomic DNA quality control results

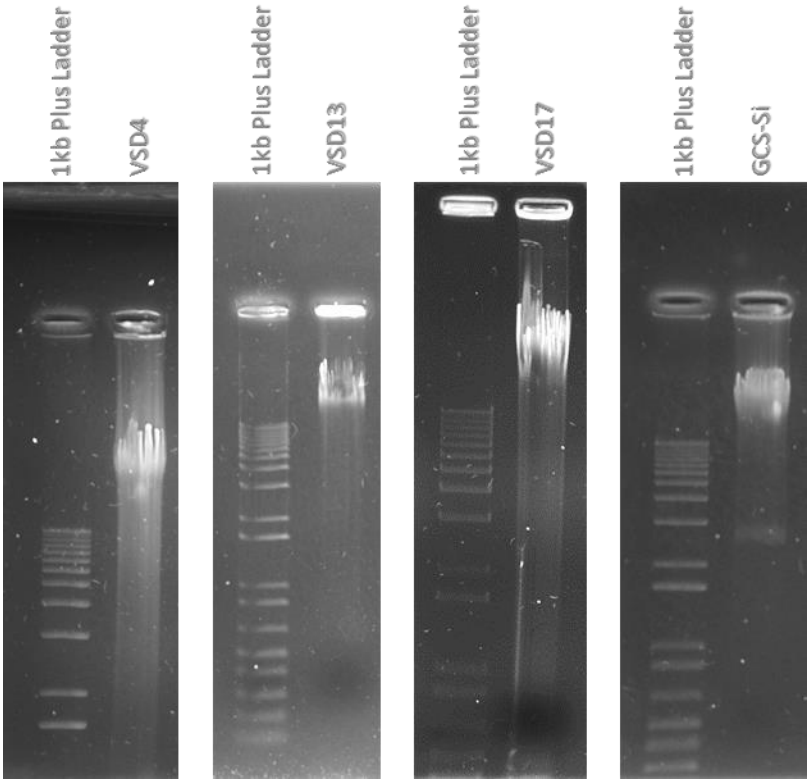








Supplementary Fig. 2 – Genomic DNA absorbance scans. Absorbance scans were taken as described in **section 2.4.2** of **Chapter IV.**

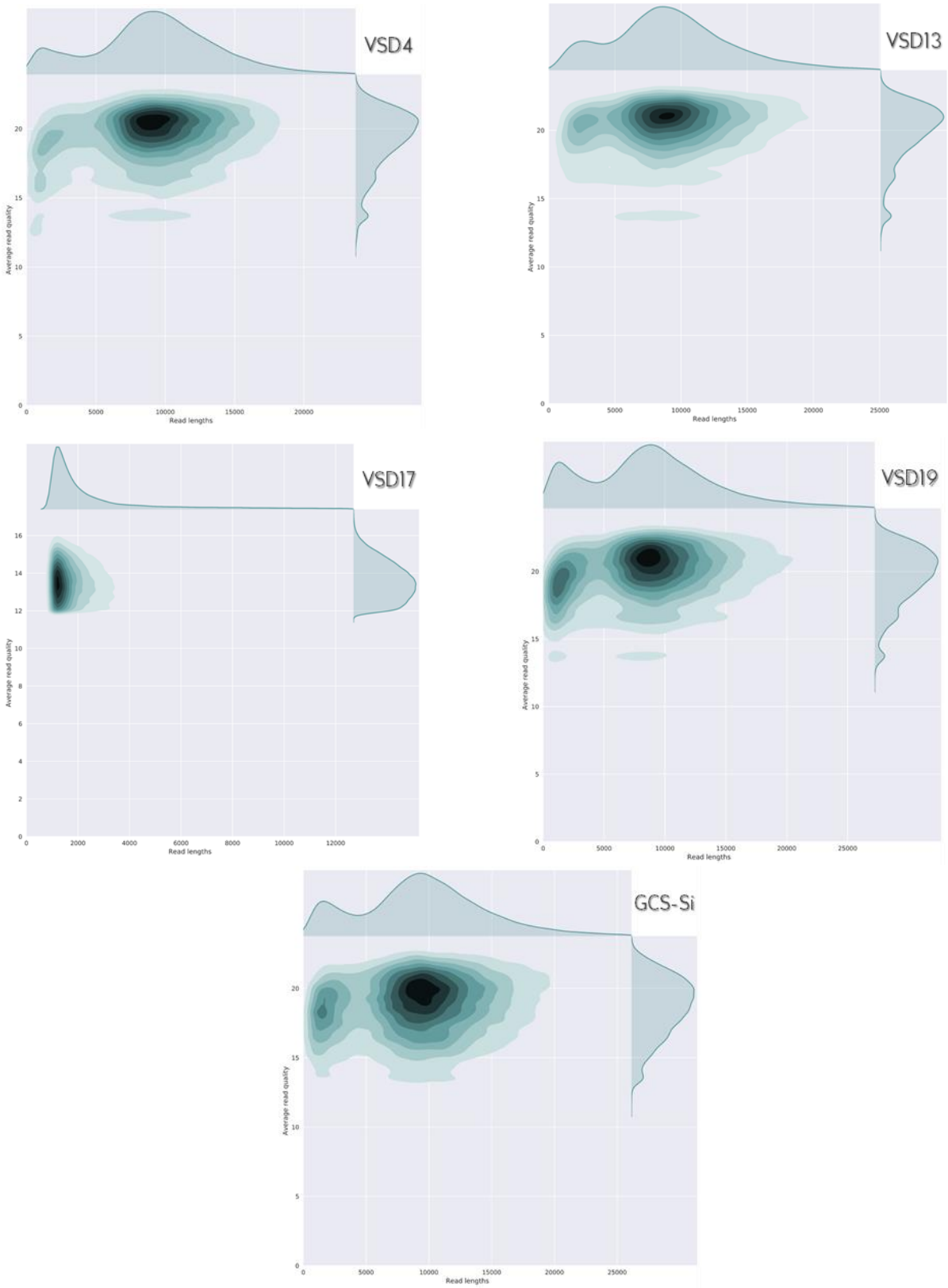


Supplementary Fig. 3 – Genomic DNA agarose gel electrophoresis. Genomic DNA from strains VSD4, VSD13, VSD17 and GCS-Si underwent AGE as specified in **section 2.4.3** of **Chapter IV**. As for strain VSD19, although the quality of genomic DNA was up to par (as can be verified through the absorbance scan) the extraction process had a low yield and as such the total volume of genomic DNA was used for library preparation.

APPENDIX D. Supplementary sequencing metrics

Supplementary Table 2 - Additional sequencing metrics for total obtained reads and filtered subsets. Total obtained reads correspond to data obtained prior to read pair detection while the filtered subset corresponds to paired $1D^2$ reads that meet the quality and length criteria. The read length N50 represents, within a set of sequences of varying lengths, the shortest sequence length to cover 50% of the total bases present within the set. Read quality is represented by the percentage of reads whose QScore is above the specified value (for filtered datasets, the minimum acceptable value was Q10, so all reads have QScores above 10). Longest reads do not correspond to the highest quality reads.

Strain	Dataset	Data yield		Read Length				Read Quality					Reads in filtered subset (%)
		Reads	Mb	Mean	Median	N50	Longest read (bp)	> Q5	> Q10	> Q15	> Q20	Highest quality read	
VSD4	Total Reads	2 77 275	1 760	6 346.77	5 115	10 559	2 035 561	99.82%	77.52%	0.84%	0%	16.72	13.11%
	Filtered 1D²	36 362	342	9 407.49	9 345	11 148	55 802	-	-	93.77%	41.28%	23.92	
VSD13	Total Reads	185 145	1 510	8 153.77	7 598	11 074	475 762	99.89%	85.22%	3.87%	0%	17.08	18.86%
	Filtered 1D²	34 919	334	9 574.72	9 053	11 162	118 631	-	-	95.21%	55.75%	24.40	
VSD17	Total Reads	412 253	655	1 588.01	732	3 301	324 185	100.00%	99.50%	5.54%	0%	16.07	23.75%
	Filtered 1D	97 898	322	3 291.61	1 787	5 494	83 135	-	-	8.98%	0%	17.39	
VSD19	Total Reads	254 485	2 101	8 255.79	7 658	11 523	463 174	99.94%	87.76%	5.40%	0%	17.63	19.11%
	Filtered 1D²	48 620	450	9 246.63	8 784	11 570	73 764	-	-	96.05%	48.40%	24.67	
GCS-Si	Total Reads	184 379	1 690	9 166.61	8 813	11 914	965 540	99.93%	81.41%	0.95%	0%	16.74	26.93%
	Filtered 1D²	49 649	485	9 774.84	9 509	11 760	87 419	-	-	93.35%	28.99%	23.78	



Supplementary Fig. 4 - **Read quality vs. read length distribution of filtered subsets.** The bivariate plots, obtained using NanoPlot, show a kernel density estimate (KDE) of the read length compared to the read's QScore. The horizontal axis represents read length and the vertical axis represents average read quality (with a maximum value of 16 for VSD17 and 20 for the remaining strains). For the sake of intelligibility, extremely long outlier reads were excluded from this representation.

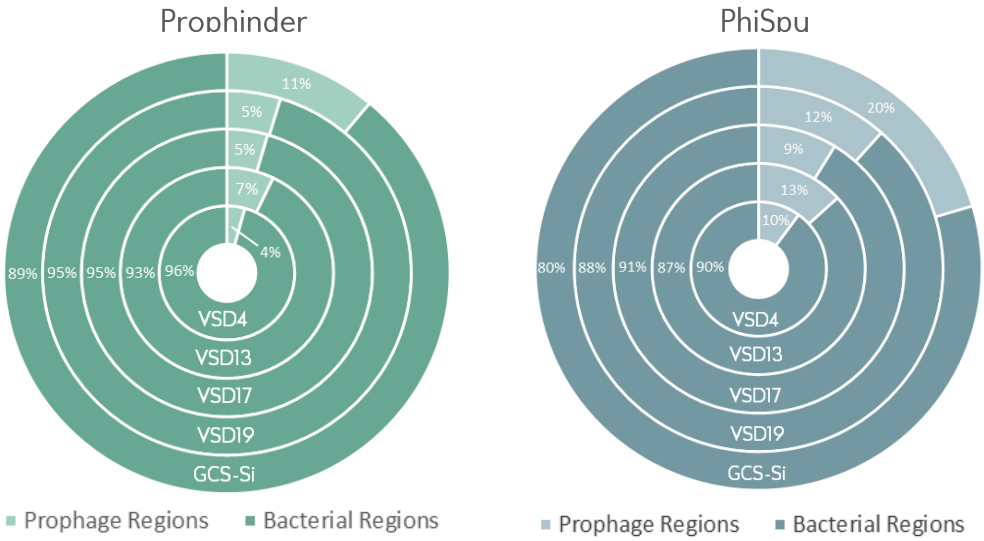
APPENDIX E. Assembly evaluation: effects of polishing draft assemblies

Supplementary Table 3 - Alignment of polished and unpolished assemblies with reference SDSD and SDSE genomes. Longest alignment length registers the size of the biggest consecutive alignment block; average identity represents the percentage of identical bases within aligned sequence blocks; SNPs represents the number of single-nucleotide polymorphisms.

Strain	Assembly stage	Size	% G+C	SDSD (2.14 Mb; 39.36% G+C)					
				Aligned Bases (%)	Longest alignment Length (bp)	Average Identity (%)	SNPs	Indels	Insertion Sum
VSD4	Unpolished	2.207 Mb	39.28	86.73	1 895 372	98.42	14 964	13 428	331 385
	Polished	2.216 Mb	39.23	86.71	1 903 919	98.81	15 331	5 894	332 245
VSD13	Unpolished	2.288 Mb	39.32	88.05	1 957 821	98.21	20 518	12 455	357 577
	Polished	2.296 Mb	39.26	88.04	1 961 041	98.54	20 800	5 792	362 105
VSD17	Unpolished	2.103 Mb	39.42	89.46	1 814 082	97.24	10 123	40 464	307 017
	Polished	2.142 Mb	39.43	89.52	1 886 060	99.12	8 535	7 497	276 331
VSD19	Unpolished	2.232 Mb	39.28	86.85	1 920 305	98.58	15 026	10 341	332 752
	Polished	2.240 Mb	39.22	86.84	1 923 672	98.88	15 197	4 559	336 011
GCS-Si	Unpolished	2.448 Mb	39.61	72.22	1 661 280	96.38	46 878	12 824	813 449
	Polished	2.455 Mb	39.59	72.31	1 665 485	96.57	47 607	8 681	817 556

Strain	Assembly stage	SDSE (2.16 Mb; 39.5% G+C)					
		Aligned Bases (%)	Longest alignment Length (bp)	Average Identity (%)	SNPs	Indels	Insertion Sum
VSD4	Unpolished	79.57	1 733 464	95.76	57 769	13 683	506 624
	Polished	79.6	1 733 333	96.07	58 769	7 220	509 561
VSD13	Unpolished	76.07	1 686 515	95.84	57 772	12 035	634 239
	Polished	76.1	1 692 860	96.12	58 575	6 547	637 758
VSD17	Unpolished	81.7	1 685 831	94.26	58 932	39 231	447 395
	Polished	81.91	1 722 515	95.91	60 463	9 524	453 736
VSD19	Unpolished	79.63	1 716 697	95.89	58 817	11 139	549 526
	Polished	79.62	1 715 486	96.13	59 488	6 225	552 255
GCS-Si	Unpolished	74.3	1 703 189	96.93	39 674	12 724	824 759
	Polished	74.28	1 705 179	97.16	40 063	8 454	829 656

APPENDIX F. Supplementary phage prediction results



Supplementary Fig. 5 - Percentage of prophage and bacterial regions according to both phage detection tools.

Supplementary Table 4 - **Overview of putative prophage sequences and their respective features.** Phage features include virulence related sequences as well as counter-resistance associated sequences.

Strain	Prophage Designation	Features
VSD4	vsd4_A	Methylase
	vsd4_B	Hyaluronidase
VSD13	vsd13_A	-
	vsd13_B	Streptodornase D
	vsd13_C	Streptococcal pyrogenic exotoxin K
	vsd13_D	Hyaluronidase Streptococcal pyrogenic exotoxin K Streptococcal extracellular nuclease 2
VSD17	vsd17_A	Hyaluronidase Methylase Streptococcal extracellular nuclease 3
	vsd17_B	Hyaluronidase Streptococcal extracellular nuclease 3 Toxin Zeta
VSD19	vsd19_A	Methylase
	vsd19_B	Hyaluronidase Streptococcal pyrogenic exotoxin K Streptococcal extracellular nuclease 2
	vsd19_C	-
GCS-Si	gcs_A	Hyaluronidase Pathogenicity Island SaPIn2
	gcs_B	-
	gcs_C	-
	gcs_D	-
	gcs_E	Methylase
	gcs_F	Methylase