2014

ZOMBIES IN BACTERIAL GENOMES: IDENTIFICATION AND ANALYSIS OF PREVIOUSLY VIRULENT PHAGE

Scott Mitchell
scott.dobson.mitchell@gmail.com

ZOMBIES IN BACTERIAL GENOMES:

IDENTIFICATION AND ANALYSIS OF PREVIOUSLY VIRULENT PHAGE

By

Scott Mitchell

(BSc. Biomedical Science, University of Waterloo, 2012)

THESIS PROPOSAL

Submitted to the Department of Biology

Faculty of Science

in partial fulfillment of the requirements for the

Master of Science in Integrative Biology

Wilfrid Laurier University

2014

ABSTRACT

Bacteriophage (or 'phage') are viruses that infect and reproduce within their bacterial hosts. They have a major global impact on bacterial evolution and ecology, and might influence the pathogenicity of their host bacterium by providing virulence factors. Phage can either be described as "virulent" or "temperate"; the distinguishing feature between the two is their method of replication.

This study sought to identify phage sequences within bacterial host genomes and determine the life cycle of the phage, exploring whether there is a connection between defective phage and previously virulent phage. It would normally be expected that any phage sequences identified within a bacterial host would have a temperate life cycle, since only temperate phage enter the lysogenic cycle and insert their DNA into the host as a 'prophage,' while virulent phage replicate via the lytic cycle, in which phage DNA replicates separately from that of the host's and infected cells are lysed.

Defective phage–'zombies' in bacterial genomes–are dormant phage that have become inactive through mutational decay or some other process. It is possible that some of these defective phage are in fact previously virulent phage that have become accidentally inserted within the host genome.

This study detected phage within bacterial genomes using the prophage identification tools PHAge Search Tool (PHAST) and Prophage Finder. Identified sequences were categorized as 'intact,' 'questionable,' or 'incomplete'; questionable and incomplete phage were classified as defective. The lifestyles of the uncovered phage sequences were then determined using PHACTS; six phage were identified as possibly virulent. The life cycles of the phage were further analyzed by assessing the genomic signature distances (GSD) and codon adaptation indexes (CAI) for each phage. Three phage were shown to have a GSD consistent with a virulent life cycle, and the CAI values of four phage corresponded with that of virulent phage. Although previous studies have indicated that some virulent phage may have a temperate lineage, identifying prophage as previously virulent is a novel finding. This has implications for our understanding of phage life cycles and the infection process, as it challenges the idea that only temperate phage insert their DNA into the host genome.

ACKNOWLEDGEMENTS

I'd like to thank Dr. Gabriel Moreno-Hagelsieb for his support, guidance, and insight over the past two years. He not only helped me while I worked to complete the project, but more importantly, helped me find the story that I wanted to tell. Thanks for helping me find my zombies, and for being my Jedi master in bioinformatics. Thank you to Dr. McDonald for agreeing to be on my committee and for her invaluable suggestions and input throughout the process, and for teaching BI601 in a way that helped me become a better scientist. I'd also like to thank committee member Dr. DeWitt-Orr, not only for being on my committee, but for asking the tough questions that made me really stop and think about the research. Every time I responded with, "That's interesting," or "That's a good point," I was buying time to figure out what to say because you'd made me consider something that I hadn't before. A special thank you to the members of the Computational ConSequences lab, including Jennifer Dobson-Mitchell, Marc Del Grande, Thomas Hemmy, Brigitte Hudy, and Karla Natalia Valenzuela Valderas. Thanks to Melanie Whitwell for being the best administrative staff at Laurier, and probably the planet. To my mom, Kathy Dobson, thank you for your support and pretending to find bacterial genomes interesting whenever I'd talk about them. To my sister Jenny, thanks for all your help and support. Especially with the stats, but everything else too. To my friend Andrew Waldie-Porteus, I know it's not a novel but here's an Easter egg in the meantime.

TABLE OF CONTENTS

1. INTRODUCTION: BACKGROUND AND OVERVIEW

1.1. Background on bacteriophage genomics and evolution

Viral infection can be observed within every domain of cellular life, from the largest mammals to the bacteria that live in deep-sea hydrothermal vents (Forterre, 2010). Bacteriophage are viral predators of bacteria, and might be the most abundant biological entity on the planet with an estimated 10^7 particles/mL in seawater and a global population of 10^{31} individuals (Deschavanne *et al*., 2010; Hatfull & Hendrix, 2011). If every phage on the planet were laid out end to end, they would span between the Earth and the Sun 10^{13} times (Hendrix, 2003). With such an enormous population and substantial opportunities for infection and evolutionary interactions between phage and their hosts, a dynamic, constantly changing, genetic structure can be inferred (Hendrix, 2003; Bailly-Bechet *et al*., 2007; Deschavanne *et al*., 2010).

The phage genome is often described as being 'mosaic' in structure; this refers to the fact that when comparisons are made between the DNA of different phage, there are alternating fragments of similarity and divergence (Belcaid *et al*., 2010). This mosaic structure arises from a constant exchange of genetic information, with an estimated 10^{24} phage initiating an infection somewhere on Earth every second (Hatfull, 2008). It has been said that bacteriophage represent the largest number of undiscovered species and the greatest reservoir of unidentified genetic information (Hatfull, 2008).

1.2. Bacteriophage replication: lysogenic and lytic cycles

As mentioned above, a distinction must be made between virulent and temperate phage. Virulent phage reproduce via the lytic cycle, in which the viral genome is injected into

the recipient cell and replication occurs separately from the host DNA (Fig. 1). The viral genome hijacks the cellular machinery of infected cells, directing the assembly of viral particles that are released at the end of the cycle following host cell lysis (Sturino & Klaenhammer, 2006). Conversely, temperate phage incorporate their genome into the bacterial host chromosome as a prophage, replicating alongside the host DNA. Eventually, a temperate phage may progress into the lytic cycle resulting in host cell lysis and the release of virions (Sturino & Klaenhammer, 2006).

However, unlike strictly lytic phage, which do not undergo integration with the host genome, some of these virions can potentially contain both host and bacteriophage DNA, a phenomenon referred to as transduction. When a new host cell is infected, some of the DNA from the first host may be transferred in the process (Sturino & Klaenhammer, 2006) (Fig. 1).

1.3. Horizontal gene transfer: history and mechanisms

Horizontal gene transfer (HGT) is a process by which genetic material is transmitted between organisms, possibly of different species. HGT is the lateral transmission of genetic material; in vertical transfer, genes are transmitted from parental to descendant cells, while HGT involves the passing of genes from one bacterial cell to another (Syvanen, 1994). Frederick Griffith first demonstrated this process in 1928, in an experiment that utilized two strains of *Streptococcus pneumonia*: a highly virulent 'S' strain that caused death when injected into mice, and a temperate 'R' strain that did not cause death. When a mixture of heat-killed S bacteria and living R bacteria was found to be lethal, this lead Griffith to hypothesize that the R strain had somehow been

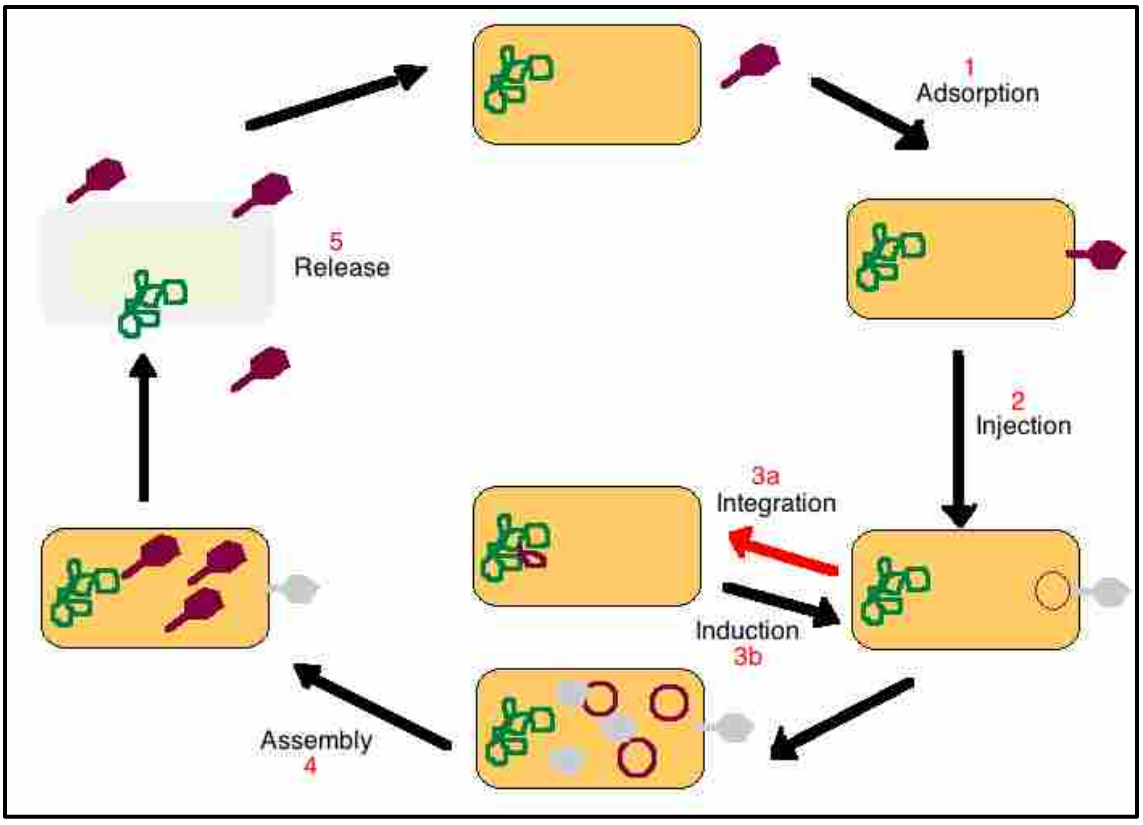

Figure 1: Lytic vs. lysogenic cycle of bacteriophage

Figure 1 depicts the lytic and lysogenic cycles of bacteriophage. In step 1, the bacteriophage attaches to the host cell in a process termed 'adsorption.' In step 2, the phage injects its DNA into the infected host cell. A temperate phage will then progress into step 3a, integration, where its DNA is incorporated into the host genome; this is termed the lysogenic cycle. A virulent phage will proceed directly to step 3b, which involves the hijacking of the host cell machinery; this is termed the lytic cycle. A temperate phage may progress into the lytic cycle through the process of induction if an environmental stressor, such as heat or UV light, is present. In step 4, assembly, phage DNA is packaged into newly-produced virions, which are then released into the surrounding environment in step 5 (created using data from Sturino & Klaenhammer, 2006).

transformed into the virulent form (O'Connor, 2008; Griffith, 1928). This work eventually led to the discovery that genes were made of DNA. The impact of HGT on bacterial evolution gained further recognition following the emergence of multi-drug resistance in the 1950s, as researchers observed rapid acquisition of antibiotic resistance that could not be explained by random point mutation (Ochman *et al*., 2000). Today, HGT is believed to be a primary force in the shaping of prokaryotic genomes (Philippe & Douady, 2003; Zhaxybayeva & Doolittle, 2011). HGT might occur through three mechanisms: transformation, conjugation, and transduction (Acar & Moulin, 2012). Transformation, as observed by Griffith, is the ability of an organism to take up and express exogenous DNA; conjugation involves the passing of genetic information through direct cell-to-cell contact, or through a tube-like structure known as a pilus (Acar & Moulin, 2012).

The third mechanism, transduction, involves the transferring of DNA from one bacterial cell to another by a bacteriophage (Griffiths *et al*., 2000). As mentioned above, following cell lysis, the bacterial chromosome is broken into small pieces that are sometimes incorporated into the phage particles; during subsequent infection, these bacterial genes are injected into the new host along with the phage DNA, and may be incorporated into the new host's genome through recombination. Although only a small proportion of phage carry donor genes (an estimated one in 10,000), transduction is still an important evolutionary force due to the immense number of phage infections and genetic interactions (Griffiths *et al*., 2000). The process of HGT results in an accumulation of changes between distantly related and non-interacting species (Deschavanne *et al*., 2010). Through the use of reference databases and phylogenies that

include sequences of known origin and function, it is possible to understand the evolution, functions and identities of unknown sequences and microorganisms. These kinds of comparative sequence analyses can be used to identify homologous sequences in, for example, genomic datasets, uncovering prophage sequences by recognizing attachment sites, integration genes, and prophage genes.

1.4. Defective phage: 'Zombies' in bacterial genomes

Although most bacterial host genomes have been shown to harbour a number of prophage, not every phage-like element is necessarily an intact, functional phage. Indeed, defective prophage – phage that can be described as existing in a state of mutational decay – have been found in many bacterial genomes, and these remnants often carry genes that are beneficial to the host, including genes with recombination functions, virulence, mutation rate, stress resistance, or toxins that can inhibit the growth of competing bacteria in the environment (Brussow *et al*., 2004; Wang *et al*., 2010; Zhou *et al*., 2011; Panis *et al*., 2012; Rabinovich *et al*., 2012; De Paepe *et al.,* 2014). In addition to serving as a reservoir of genes, including lysis modules and biofilm development, there is a continuous genetic exchange between these defective phage and active, functional phage, which serves to accelerate bacterial evolution (Redfield & Campbell, 1987; Canchay *et al*., 2007). This project refers to these defective phage as 'zombies' in bacterial genomes, as they are classified as nonviable yet still carry out important activities and functions, making them 'undead' remnants within their hosts.

1.5. Current research, applications, and clinical significance of phage genomics: studying the 'arms dealers' of the bacterial world

Investigating and expanding our understanding of phage evolution and interactions among microbial populations and their hosts are difficult tasks due to phage's diverse and mobile nature, yet recent advances in comparative genomic studies can help elucidate the mechanisms by which viruses evolve (Hatfull & Hendrix, 2011). Despite their abundance and clinical significance, knowledge of phage genomics is based on an exceedingly small, biased sample size of the estimated 10^{31} individuals, with approximately 750 phage genomes having been fully sequenced as of 2011 (Hatfull & Hendrix, 2011). As more genome sequences become available for study, a clearer picture of the evolutionary histories, connections, and population structure of bacteriophage can emerge.

HGT is widely believed to be the primary cause of antibiotic resistance (Alonso *et al*., 2002), a growing public health crisis that is causing tens of thousands of deaths and costing billions of dollars every year (Dye, 2009). At least 150,000 deaths worldwide are attributed to tuberculosis caused by multi-drug resistant *Mycobacterium tuberculosis* alone, with many other infectious microorganisms at risk of becoming uncontrollable in the near future (Dye, 2009). For example, methicillin-resistant *Staphylococcus aureus* (MRSA) is resistant to every available antibiotic, making it a global health concern with limited treatment options (McCarthy *et al*., 2012). Phage play an important role in antibiotic resistance; they can be seen as agents of HGT mechanisms, aiding in the proliferation, distribution, and adaptation of antibiotic-resistance genes. They are capable of transferring resistance and virulence genes between bacteria, modifying parameters such as host range and pathogenicity by introducing novel genes or modifying the

expression of pre-existing sequences (McCarthy *et al*., 2012). The characterization of phage life cycles furthers our understanding of phage genomics and population dynamics (McNair *et al.*, 2012; Housby and Mann, 2009).

Although they play a crucial role in the evolution of pathogens, a potential application of phage includes using them as a source of antibiotics that could be used against multi-drug resistant bacteria (O'Flaherty *et al*., 2005). As natural predators of bacteria, phage possess many potential advantages over traditional antibiotics. They are highly specific to their hosts, which would prevent harm to the human body's communities of beneficial bacteria; and because they depend on their bacterial hosts for survival they are self-limiting—once all of the target bacteria are killed, they would naturally die off as well (Morello *et al*., 2011).

There has been reported success in the use of phage for bacterial infections (Morello *et al*., 2011), yet despite the potential benefits of phage therapy, there are still many barriers preventing it from becoming a widespread treatment. Phage particles are quickly removed by the body's phagocytic system, reducing their circulatory time and effectiveness. The majority of studies involving phage therapy have been *in vitro*, which means there are many gaps in our understanding of phage pharmacokinetics. Lastly, the potential scalability of phage therapy is uncertain, as there are many manufacturing, production, and distribution concerns (Keen, 2012). Although these current limitations and problems must be mediated, phage therapy still represents a promising method for the treatment and prevention of bacterial infections and antibiotic resistance (Keen, 2012).

1.6. An integrative approach

Bioinformatics integrates high-throughput technologies, computational analyses and data collection to examine life sciences problems, extracting information from huge amounts of data produced by working with various tools and databases (Thiele *et al*., 2010). Essentially, bioinformatics involves the capturing, integration and analysis of data generated through experiments or gathered from databases to provide insights and shed light on complex biological systems (Thiele *et al*., 2010).

In addition to the integrative nature of the field of bioinformatics, this project specifically used a variety of tools to explore a biological phenomenon that has important implications for many fields and sub-disciplines. As agents of HGT, bacteriophage – including defective phage – have a significant impact on the global microbial community and the crucial functions they perform (Bailly-Bechet *et al*., 2007; Forterre, 2010). Tracking evolutionary exchanges and histories among this immense population is a difficult task, yet may provide valuable insight into the interactions between phage, their host bacterium, and microbial communities at large. This project aimed to utilize computational genomic methods – namely, tracking the distribution of prophage throughout host genomes and analyzing their life cycles – to further our understanding of the evolutionary and ecological mechanisms underlying the complex interactions between these populations, focusing on defective phage – the zombies within bacterial genomes.

A deeper understanding of phage evolution has important implications for the field of taxonomy. Recent findings on the impact of HGT on the evolution and taxonomic distribution of bacteria have placed a much greater emphasis on the role of lateral

transmission, and it is now believed that key evolutionary transitions – such as DNA replication machinery – may have originated in the viral world (Forterre, 2010). Although this contradicts traditional classification systems and evolutionary histories, which are largely focused on vertical transmission, it presents a clearer picture of the three domains of life. It should come as no surprise that the most abundant biological entity on the planet has played such a central role in microbial evolution. Our knowledge of this complex network of genetic exchange is still in its infancy, yet it has already revolutionized our understanding of the microbial world and lead to important developments in biotechnology (Onodera, 2010). Clearly, just as phage play a crucial role in the evolution of microbial communities, they are driving forward the evolution of fields such as microbiology and genomics.

1.7. Research objectives and hypotheses

This project identifies prophage within bacterial host genomes using the programs PHAST and Prophage Finder, comparing the results to previously conducted studies that used BLAST search. Additionally, it was explored whether there is a connection between the ‘defective’ or ‘incomplete’ phage that were identified, and previously virulent phage that have lost their virulence and become inserted within the genome. This was accomplished by analyzing the life cycles of the phage with the program PHACTS and characterizng the genomic signature distances and codon adaption indices of the phage.

It was expected that PHAST would uncover more prophage sequences than BLAST, as the program analyzes a variety of information including unusual genes, attachment site recognition, and tRNA analysis. It was expected that PHAST would also identify more prophage than Prophage Finder, as it is a newer program that has been

previously shown to have greater sensitivity and more accurate attachment site prediction (Zhou *et al*., 2011). For the prediction of phage life cycles, it was believed that some prophage would be identified as previously virulent by PHACTS, and that the signature distance and CAI of these prophage would be virulent-like.

1.8. Organism of study: Staphylococcus aureus

S. aureus, a gram-positive bacterium, has been recognized as a dangerous pathogen in humans for over 100 years (Lowy, 1998), but it has recently been recognized as a primary cause of skin and soft tissue infections (Klevens *et al*., 2007). Methicillin-resistant *Staphylococcus aureus* (MRSA) has emerged as a global health threat, with widespread community outbreaks and hospital acquired infections; although typically resulting in skin disease, MRSA infections can be fatal (Klevans *et al*., 2007).

S. aureus was selected for this study because previous studies have examined the prevalence of phage within the genomes of different strains, using BLAST for prophage identification, which can be compared to the number of prophage identified using PHAST (McCarthy *et al*., 2012).

INTRODUCTION: PROGRAMS AND PROCEDURES

The following sections provide background information on the programs and procedures that were used in this project, describing traditional approaches and contrasting them with the methods that were used. First, methods of phage identification are outlined, followed by a description of various techniques for characterizing the life cycles of phage.

1.9. Identification of prophage: traditional methodologies

The number of sequenced bacterial genomes has been rapidly increasing as sequencing technologies have progressed and costs have dropped (Hendrix, 2003; Soon *et al*., 2013). However, although our understanding of phage and bacterial genomics has rapidly advanced in recent years, most methods for identifying prophage within bacterial genomes—both experimental and computational—have significant shortcomings and barriers preventing them from locating sequences with high accuracy, reliability, or efficiency (Zhou *et al*., 2011). Experimental approaches, which usually involve exposing host bacteria to UV light to induce them into releasing phage particles, overlooks the presence of defective prophages (Zhou *et al*., 2011). Most computer programs rely on the identification of atypical gene content, unusual tRNAs, and disrupted genes. However, many phage do not reliably insert into the same coding regions or tRNAs (Zhou *et al*., 2011), and the majority of these programs require the bacterial genomes to be annotated. As such, these experimental and computational methodologies typically underestimate the amount of prophage sequences (Zhou *et al*., 2011).

1.10. Identification and classification of prophage: an integrated approach

The most effective method for identifying prophage sequences is an integrated approach that incorporates sequence comparisons to known genes, dinucleotide analysis, detection of disrupted or unusual genes and tRNAs, and hidden Markov model scanning (Zhou *et al*., 2011). The programs PHAST and Prophage Finder utilize these methods to rapidly and accurately locate prophage sequences; additionally, neither requires the input sequence to be well annotated with identified open reading frames (Bose & Barber, 2006; Zhou *et al*., 2011). By not relying on sequence annotation, these programs are more sensitive, as the validity of phage predictions can be affected by the genome annotation process (Bose & Barber, 2006; Zhou *et al*., 2011).

PHAST is a web server that locates prophage sequences within bacterial genomes, annotating and graphically displaying the sequences and indicating prophage 'quality' (Zhou *et al*., 2011). Both raw and annotated input sequences are accepted, and Gene Locater and Interpolated Markox ModelER (GLIMMER) gene prediction is used to identify prophage as well as the position, length, number and boundaries of genes. More specifically, PHAST combines open reading frame prediction through GLIMMER, identification of proteins and phage sequences via BLAST, tRNA analysis, as well as attachment site recognition and gene clustering density readings through Density-Based Spatial Clustering of Applications with Noise.

1.10.1. GLIMMER

GLIMMER is a computational gene finder that uses interpolated Markov models (IMMs) to differentiate between coding regions and non-coding DNA (Delcher *et al*., 1999). A Markov chain is a series of variables in which the probability for each variable depends on the preceding *k* variables, for some constant *k*. For DNA sequence analysis, this means the probability of a base (*b*) depends on the *k* bases preceding it (Delcher *et al*., 1999). These prior k bases as described as the context of base b. IMMs use these contexts to determine the probability of *b*, giving a weight to each context so that the IMM is sensitive to the frequencies of different oligomers in a genome (Delcher *et al*., 1999). PHAST incorporates GLIMMER for accurate prophage identification and to recognize the position, length and boundaries of genes.

1.10.2. DBSCAN

In the simplest terms, the DBSCAN algorithm identifies phage by finding a minimum number of phage-related elements close to each other. The DBSCAN algorithm identifies clusters by analyzing the local density of database elements, and can determine what information is 'noise.' In general terms, clustering algorithms analyze a database *D* composed of *n* objects and identify sets of *k* clusters; this could include data from satellite images, x-ray crystallography, or genomes (Ester *et al*., 1996). Specifically for identifying prophage genomes, DBSCAN can be used to recognize attachment sites and gene clusters by comparing known phage genes to the bacterial genome in question, evaluating the completeness or viability of the prophages according to the local density of phage genes (Zhou *et al*., 2011).

1.11. Identification of phage life cycles

1.11.1. Traditional methodologies

Traditional methodologies of phage life cycle determination include culturing and isolating the phage, which can be difficult due to factors such as time or cost constraints (McNair et al., 2012). However, attempts at determining phage life cycles through computational methods have also often met difficulties due to the previously mentioned 'mosaic' structure of their genomes, as there is no ubiquitous, conserved phage gene that can be analyzed (Hendrix et al., 1999; Deschavanne et al., 2010). One approach used a comparison of structural proteins (Proux et al., 2002), while another made a reticulate classification of phage life cycle according to gene content (Lima-Mendez et al., 2011).

1.11.2. PHACTS

The Phage Classification Tool Set (PHACTS) is a set of programs that utilizes the sequence data of phage genomes to determine phage life cycle, using a similarity algorithm and a supervised Random Forest classifier to predict whether the phage is virulent or temperate (McNair *et al*., 2012). A training set is created from a database of phages with known life cycles by the similarity algorithm. This is used to train a Random Forest that identifies the life cycle of the query phage (McNair *et al*., 2012). The benefits of using PHACTS as opposed to these methods is the speed and accuracy with which results can be attained. Life cycle predictions with PHAST have been shown to have a 99% precision rate when the prediction is deemed confident (McNair *et al.*, 2012).

1.11.3. Genomic signature distance

The characteristic frequency of oligonucleotides in a particular genome sequence is referred to as its 'genomic signature,' with phylogenetically related genomes typically possessing similar signatures (Wang *et al*., 2005). Variations in the signature of a genome have been used to identify horizontally transferred genes, pathogenicity islands, and prophages (Deschavanne *et al*., 2010).

By examining local variations in sequences, a genomic signature distance between a phage and its host can be calculated, with the tetranucleotide frequency of each sequence determining its particular signature (Deschavanne *et al*., 2010). When the genomic signature distances between virulent phage and their hosts are compared to the distances between temperate phage and their hosts, a separation is typically observed, with a shorter host-phage distance for the temperate phage (Deschavanne *et al*., 2010). Thus, if a prophage is shown to have a greater than expected genomic signature distance, it could potentially indicate a virulent ancestry.

1.11.4. Codon adaptation index

Codon bias is defined as an organism using synonymous codons with different frequencies (Hershberg & Petrov, 2008). Examination of codon usage in phage has shown that virulent phage tend to have higher codon usage biases and larger compositional differences to the host genome compared to temperate phage (Bailly-Bechet *et al*., 2007). Codon adaptation index (CAI) is a measurement of synonymous codon usage bias. The relative value of each codon is determined through the use of a reference set of genes with high levels of expression, assigning a score based on the frequencies of codons within those genes. Essentially, CAI measures the degree to which

selection has shaped patterns of codon usage; this means it can be used to assess the extent to which viral genes have adapted to their hosts (Sharp and Li, 1987). A score between 0 and 1 is assigned, with a higher value indicating optimal codon usage, where ‘optimal’ is defined by a reference set of host genes. A temperate phage would be expected to have a CAI closer to 1, compared to a virulent phage, as it would be more adapted to the host genome. Thus, if a prophage is shown to have a virulent-like CAI, it could support the possibility of a virulent ancestry.

METHODS

1.12. Overview

The methodology is outlined in Figure 2. Phage sequences were identified within 41 sequenced *S. aureus* genomes using the program PHAST and the results were compared to the prophage identified through the program Prophage Finder and through a BLAST search of the GenBank database using integrase (*int*) genes, which define the site of prophage integration into the host chromosome (McCarthy *et al*., 2012). The GenBank database and Basic Local Alignment Search Tool (BLAST) are two commonly used tools in computational genomics. GenBank is a publicly available database that contains nucleotide sequences for hundreds of thousands of organisms, consolidating information about DNA and protein sequences, structure, and taxonomy (Benson *et al*., 2008). BLAST is an algorithm that is used for rapid sequence comparison, allowing for the identification and analysis of DNA and protein sequences.

Next, once all of the *S. aureus* host genomes wereanalyzed for the presence of prophage using the methodology described above, the genome sequences of defective and incomplete prophage wereisolated and examined, in order to determine whether these prophage are in fact previously virulent phage that have lost their virulence genes and become inserted into the host genome. This was performed using the program PHACTS, and then further explored by analyzing the genomic signatures and codon usage of the phage.

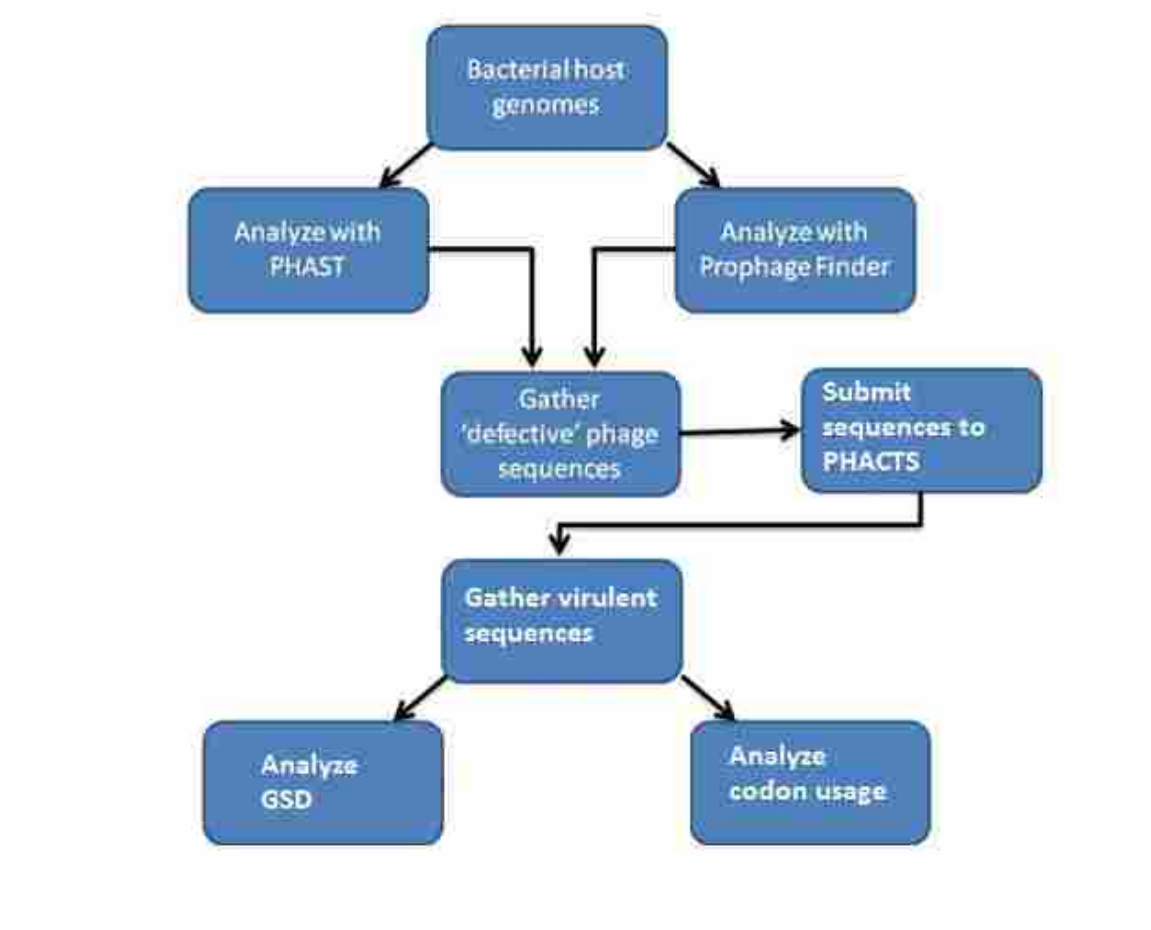

Figure 2: Flowchart of proposed methodology

Figure 2 illustrates the methodology for this study. Bacterial host genomes were analyzed using both PHAST and Prophage Finder to uncover prophage sequences. Next, the 'defective' phage sequences were gathered, and after submitting the sequences to PHAST, the program was used to distinguish between temperate and previously virulent defective phage. Finally, the genomic signature distance (GSD) and codon usage of these phage were analyzed.

1.13. Prophage identification with PHAST: Overview

A python script was used to retrieve every *S. aureus* subspecies host genome sequences from the most recent version of the NCBI GenBank database in FASTA (.fna) format, and then submit these files to the programs PHAST and Prophage Finder.

1.13.1. PHAST identification procedure

PHAST uses a web server to perform a series of database comparisons and phage feature identification analyses, locating and annotating prophage sequences. The sequence database that is used as part of PHAST's prophage identification consists of the NCBI phage database and prophage database. Potential tRNA and tmRNA sites are identified within query sequences, as these can help indicate the location of attachments sites, using the programs tRNAscan-SE and ARAGORN (Zhou *et al*., 2011). Potential phage attachment sites are also located by searching for short nucleotide repeats. Next, phage and phage-like proteins are identified through a BLAST search against the PHAST sequence database. Matched sequences are subsequently assessed for phage density by DBSCAN, which considers the cluster size *n* and the distance *e*. The cluster size *n* establishes the minimum number of phage-like genes in order for a region to be identified as a prophage, and the distance *e* defines the maximum distance between neighbouring genes of the same cluster *n* (Zhou *et al*., 2011). As prophage are typically comprised of more than five proteins, *n* was set to six, and based on the size of identified prophage in the database used by PHACTS, *e* was set to 3,000.

Identified prophage are classified according to their potential viability, being described as either intact, questionable or incomplete. This classification is based on a

'completeness score' that is calculated by one of three different methods, depending on the predicted gene content of the identified prophage or prophage-like element (Fig. 3).

If the region contains only genes of a known phage, then it is automatically assigned a completeness score of 150, the maximum (Zhou *et al*., 2011). If more than 50% of the genes in the region are identified as related to a known phage, then the score is determined according to the size of the region and number of genes:

$S = (B_r / B_p) \times 100 + (G_r / G_p) \times 100$

Where S is the score of the region, B_r is the number of bases in the region, B_p is the number of the bases in the related phage, G_r is the number of genes in the region, and G_p is the number of genes in the related phage (Zhou *et al*., 2011).

If less than 50% of the genes in the region are related to a known phage, four parameters are considered:

i) The number of bases: the region is given +10 towards its score if the number of bases is greater than 30 kb.

ii) The number of genes: the region is given +10 towards its score if the number of genes is greater than 40

iii) The presence of 'cornerstone' genes: the region is given +10 towards its score for each cornerstone gene that is present. Key phage structural genes (such as capsid, tail, and coat genes), DNA regulation genes (such as integrase and terminase genes), and function genes (such as lysin) are cornerstone genes.

iv) The presence of phage-like genes: the region is given +10 towards its score if phage-like genes occupy 70% or more of the region.

The score is then calculated as the sum of these four parameters (Zhou *et al*., 2011).

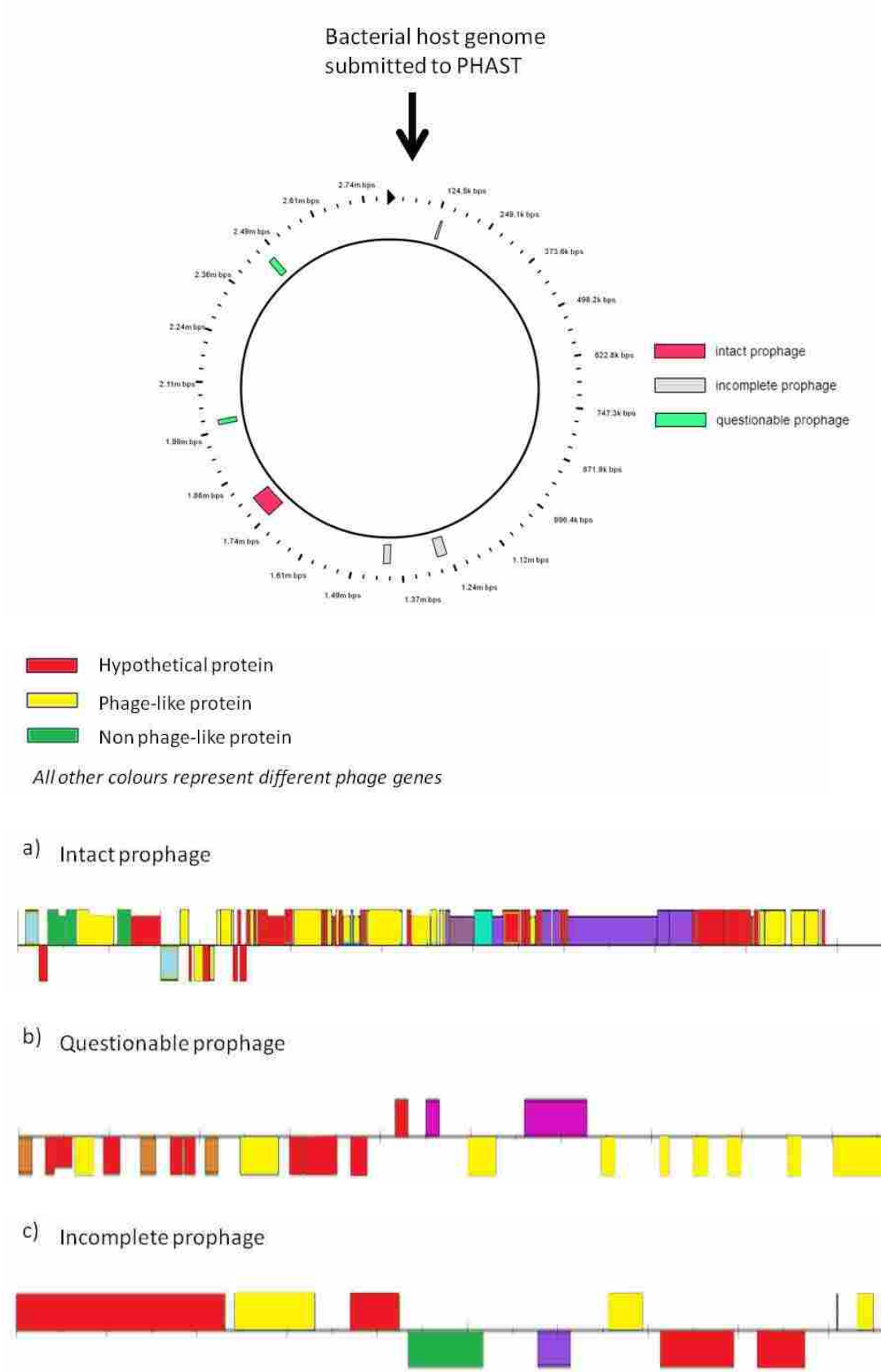
Bacterial host genome
submitted to PHAST
intact prophage
incomplete prophage
questionable prophage
Hypothetical protein
Phage-like protein
Non phage-like protein
All other colours represent different phage genes
a) Intact prophage
b) Questionable prophage
c) Incomplete prophage

Figure 3: Visualization of classification procedure for prophage identified by PHACTS

Figure 3 illustrates the three different types of calculations performed by PHACTS as part of the classification procedure of the program, identifying the prophage as 'intact,' 'questionable,' or 'incomplete.' The circular host genome shown is *S. aureus* subspecies TCH60, which has 6 prophage regions identified by PHAST. As shown in the figure, there is one 'intact' prophage, two 'questionable' prophage, and three 'incomplete' prophage within this host genome. The protein sequences of three of these prophage regions are displayed; an intact prophage in a), a questionable prophage in b), and an incomplete prophage in c). The red regions represent hypothetical proteins, the yellow regions represent phage-like proteins, and the green regions represent non phage-like proteins. These three types of regions are less clearly defined than the other proteins identified by PHAST (represented by other colours). It can be seen that intact prophage tend to have more coding sequences and a smaller proportion of these less defined regions compared to questionable and incomplete prophage.

Once a prophage region has been scored according to one of the three scenarios described above, its 'completeness' is classified. If the score is above 90, the prophage region is classified as intact; between 60 and 90 is classified as questionable; and less than 60 is classified as incomplete (Zhou *et al*., 2011). Both questionable and incomplete prophage represent defective phage. Any prophage region classified as questionable or incomplete was retrieved for further analysis with PHACTS, and its genomic signature and codon usage were characterized (described below). Intact phage were also analyzed using these methods for comparison.

1.14. Prophage Finder

Prophage Finder performs a BLASTX search against the NCBI phage database with a user-defined threshold value (E) to locate potential prophage regions. Next, a Perl program is used to analyze these regions based on user-defined 'hit spacing' and 'hits per prophage.' Hit spacing refers to the maximum number of base pairs between clusters of prophage-like genes, set to 3.5 kb, while hits per prophage is the minimum number of

significant sequence matches (potential genes) for the prophage region, set to six; these parameters were chosen in accordance with those set for PHAST. Fluctuations in GC content, codon usage, and tRNA prediction are also analyzed.

The output of Prophage Finder is a series of text files, including the DNA sequences of the predicted phage loci, a list of genes identified within each phage region, protein sequences, and summary files that describe the locations of each cluster of phage-like genes, GC content calculations, and codon frequency (Bose & Barber, 2006).

1.15. Gathering sequences for further analysis

All prophage sequences found by PHAST were gathered for further analysis, and the defective phage were identified on the basis of their completeness scores. A BLAST search of the GenBank database was performed using integrase (*int*) genes, which define the site of prophage integration into the host chromosome, to compare the data acquired by PHAST to previously published results (McCarthy *et al*., 2012). Prophage sequences found by Prophage Finder were subjected to further analysis, as the program has a higher rate of false positives (Zhou *et al*., 2011).

First, the hits from each potential prophage loci were examined. Predicted prophage with at least 10 hits are very likely to be actual prophages. Gene duplications may occur, and must be subtracted from the total number of hits. For example, if a predicted prophage has 10 hits, but the sixth and seventh hits are duplicates, then the predicted prophage's hits would be corrected to nine.

If a predicted prophage has less than six hits it is likely a false positive (Bose & Barber, 2006). However, because it is possible that these predicted prophage could be an incomplete one, rather than a false positive, the genes for any predicted prophage with

less than six hits were examined; if there were at least two hits that were matches for cornerstone genes in the database (genes for integrase, capside, terminase, methylase, methyltransferase, packaging, helicase, tail, portal, or protease), then the predicted loci was considered an actual prophage region.

Next, the prophage were assessed for their completeness by the same methods as the prophage identified by PHAST: if the region contained all the genes of a known phage, it was automatically assigned a maximum completeness score; if more than 50% of the genes in the region were identified as related to a known phage, then the score was determined according to the size of the region and number of genes; finally, if less than 50% of the genes in the region were related to a known phage, the number of bases, number of genes, presence of cornerstone genes, and presence of phage-like genes were considered.

Known virulent phage that infect *S. aureus* were also retrieved from the NCBI phage database; the nine phage used were 44AHJD, 66, G1, GH15, K, P68, phiSA012, SAP_2, and Twort. These phage were analyzed using PHACTS and their genomic signatures and codon usage were characterized, for comparison to the prophage.

1.16. Identifying life cycles with PHACTS

The phage sequences identified by PHAST and Prophage Finder were submitted to PHACTS, which performed Random Forest calculations to determine the life cycles of the phage. The prediction from a Random Forest calculation uses N known phages, randomly selected from the database for the training set, and M proteins, also randomly selected to generate similarity vectors. $N = 100$ was used, with an equal number of virulent and temperate phage. For every N phage a similarity vector X is created by

aligning the proteins of *N* against every protein *M*. Each protein in *N* is assigned a percent identity score, calculated from the highest scoring pair *S* in $X_N = [SN_1, SN_2, \ldots, SN_M]$ (McNair *et al.*, 2012).

To create the testing set, each protein of the input phage is aligned against every protein *M*. The percent identity score for each protein is calculated in the same manner as the training set, with the creation of a similarity vector, and then the life cycle is predicted using the Random Forest ensemble (McNair *et al.*, 2012).

A series of decision trees is assembled for the Random Forest calculation, selecting *N* cases from the testing set to replace *N* cases from the training set through bootstrapping. A life cycle is predicted by each tree; the final prediction is determined by whichever life cycle was chosen by the majority of trees. A probability score between 0 and 1 is then assigned to the calculation. The number of trees that chose the predicted life cycle is divided by the total number of trees, giving the probability score as the fraction of trees in the Random Forest that correspond with the final predicted life cycle (McNair *et al*., 2012). To account for the variability in predictions that arises from the randomly selected *N* phages and *M* proteins from the training set, 10 replicates are performed, each with a different testing test. The average of the 10 replicates is determined, and the predicted life cycle is the one with the higher average. A prediction is deemed confident when this average probability score is at least two standard deviations away from 0.5 (McNair *et al*., 2012). Enterobacteria phage T4, a known temperate phage, and *S. aureus* phage P68, a known virulent phage, were run through PHACTS to assess the accuracy of the program.

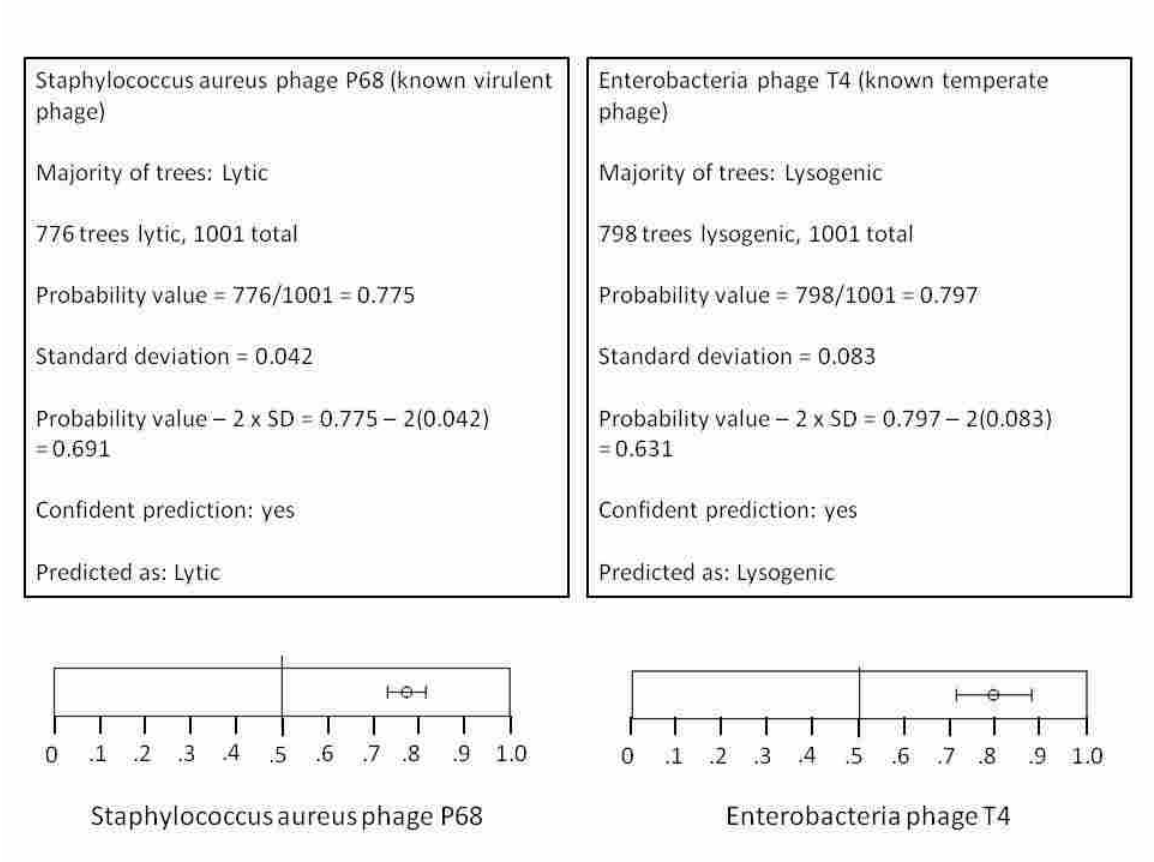

Figure 4: PHACTS analysis of known virulent and temperate phage

Figure 4 displays the results of submitting two phage with known life cycles to PHACTS; *S. aureus* phage P68, a known virulent phage, and *Enterobacteria* phage T4, a known temperate phage. For phage P68, the majority of the trees in the Random Forest classifier predicted a virulent life cycle, as expected. Of the 1001 total trees generated, 776 chose a virulent (lytic) life cycle, giving a probability value of 0.775 and a standard deviation of 0.042. Subtracting 2 x S.D. from the probability value gives a score of 0.691, which means the prediction is confident. For phage T4, the majority of the trees in the Random Forest classifier predicted a temperate life cycle, as expected. Of the 1001 total trees generated, 798 predicted a temperate (lysogenic) life cycle, giving a probability value of 0.797 and a standard deviation of 0.083. Substracting 2 x S.D. from the probability value gives a score of 0.631, which means the prediction is confident.

1.17. Characterization of genomic signature distance

A PHP program was used to determine the oligonucleotide frequencies of every prophage, virulent phage and host genome (Bikandi *et al.*, 2004). These frequencies were then used to compute the genomic signature distances between the host and phage sequences. The olignonucleotide frequency of each sequence was defined by the frequencies of all possible tetranucleotides. Signature distances were determined by measuring the Euclidian distance between the phage and host signatures.

1.18. Characterization of codon adaptation index (CAI)

The codon adaptation index (CAI) was determined for every prophage, virulent phage, and host genome sequence using DAMBE (Data Analysis in Molecular Biology and Evolution), a software package that analyzes sequence data. CAI was measured by determining the degree of translationally favoured codons, defined by comparing the usage of codons in the sequences of the phage to a reference set of host genes (Xia & Xie, 2001; Xia, 2007).

2. RESULTS

2.1. Prophage identification

All 41 *S. aureus* subspecies genomes were analyzed for the presence of prophage, and many strains were shown to contain more prophage than previously indicated. Table A.1 (see Appendix) shows the number of prophage found for each *S. aureus* genome using both PHAST and BLAST. Figure 5 displays the numbers of prophage detected for each strain. PHAST uncovered at least the same number of prophage for every *S. aureus* host genome compared to BLAST, and typically found more. Prophage Finder identified the most prophage, outnumbering PHAST for every *S. aureus* strain. Table A.2 (see Appendix) compares the total number of prophage identified by PHAST compared to Prophage Finder, as well as the number of defective phage. Figure 6 compares the total number of prophage found using PHAST and Prophage Finder, and Figure 7 compares the number of defective phage identified using these two programs. Figure 8 illustrates the percentage of phage that were found to be defective for each strain.

For both PHAST and Prophage Finder, hit spacing was set to 3.5 kb and hits per prophage was set to six. The programs were tested using a set of known prophages; using more permissive parameters did not find additional prophage regions, yet increased the rate of false positives, while more restrictive parameters overlooked certain prophage regions. This was consistent with previous findings for the programs (Zhou *et al.,* 2011; Bose & Barber, 2006).

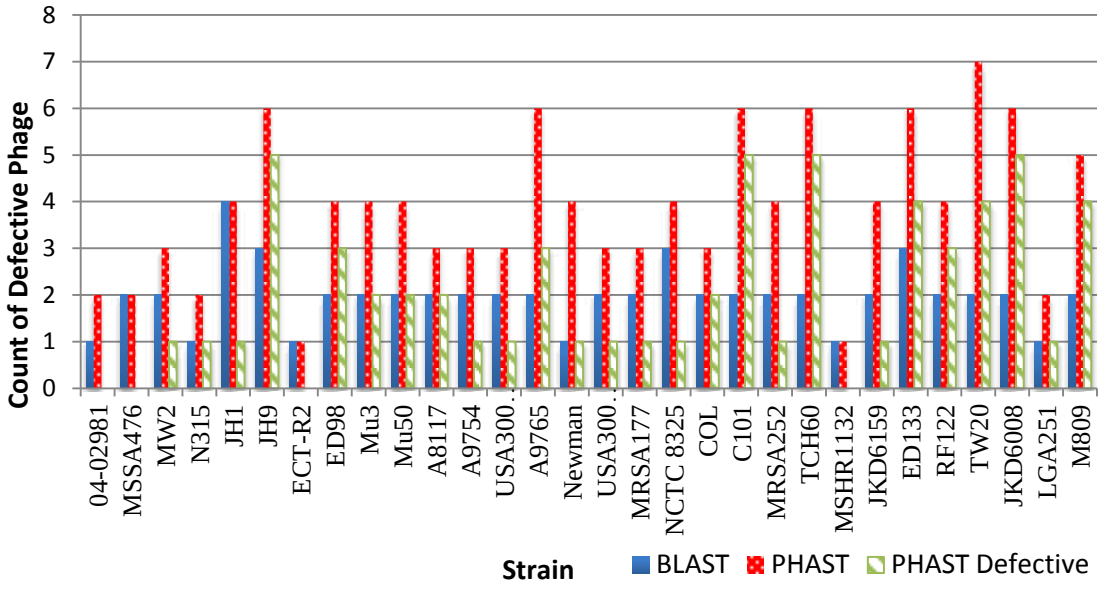

Figure 5: Graphical representation of detected prophage in S. aureus strains.

Figure 5 displays the numbers of prophage detected for strains of *S. aureus* using BLAST and PHAST. The red bars show the total number of prophage detected by PHAST (intact, incomplete and questionable) while the green bars show the number of defective prophage detected by PHAST (incomplete and questionable). In every strain, the total number of prophage detected by PHAST was higher than the number found by BLAST.

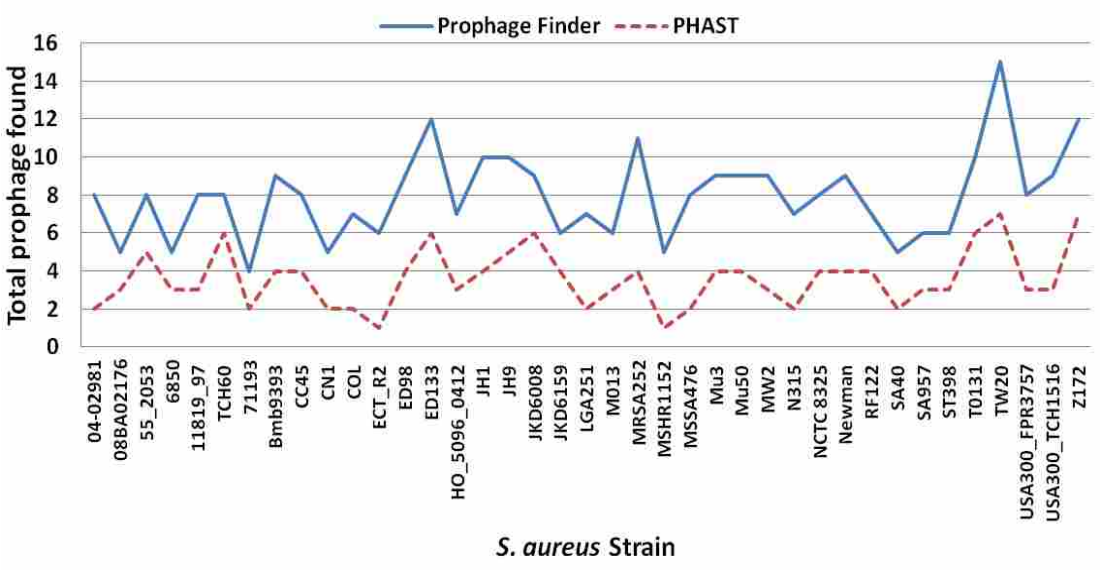

Figure 6: Total number of prophage found using PHAST and Prophage Finder.

Figure 6 compares the total number of prophage detected for strains of *S. aureus* using PHAST and Prophage Finder. The red region shows the total number of prophage detected by PHAST while the blue region shows the number of total number of prophage detected by Prophage Finder. The data for each program is shown independent of the other; for instance, 04-02981 has 2 prophage identified by PHAST, and 8 identified by Prophage Finder. In every strain, the total number of prophage detected by Prophage Finder was higher than the number found by PHAST.

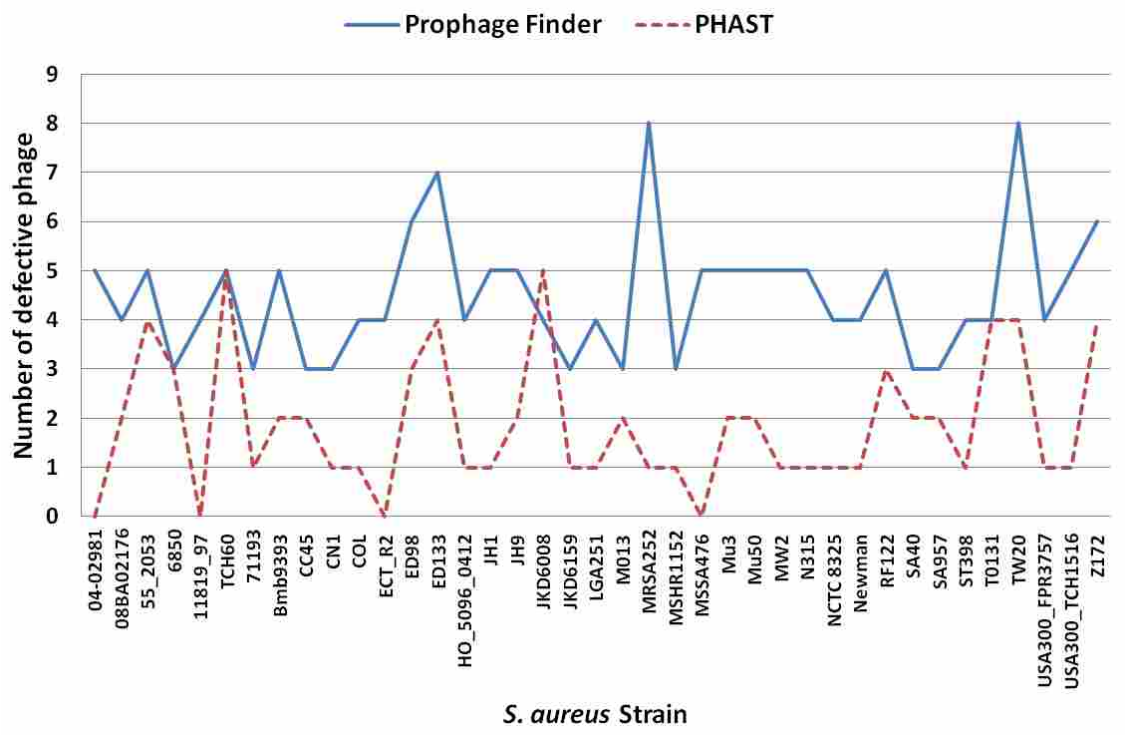

Figure 7: Number of defective phage identified by PHAST and Prophage Finder.

Figure 7 compares the number of defective phage identified by PHAST (red) and Prophage Finder (blue) for every *S. aureus* host strain. For the majority of strains, Prophage Finder uncovered a higher number of defective phage.

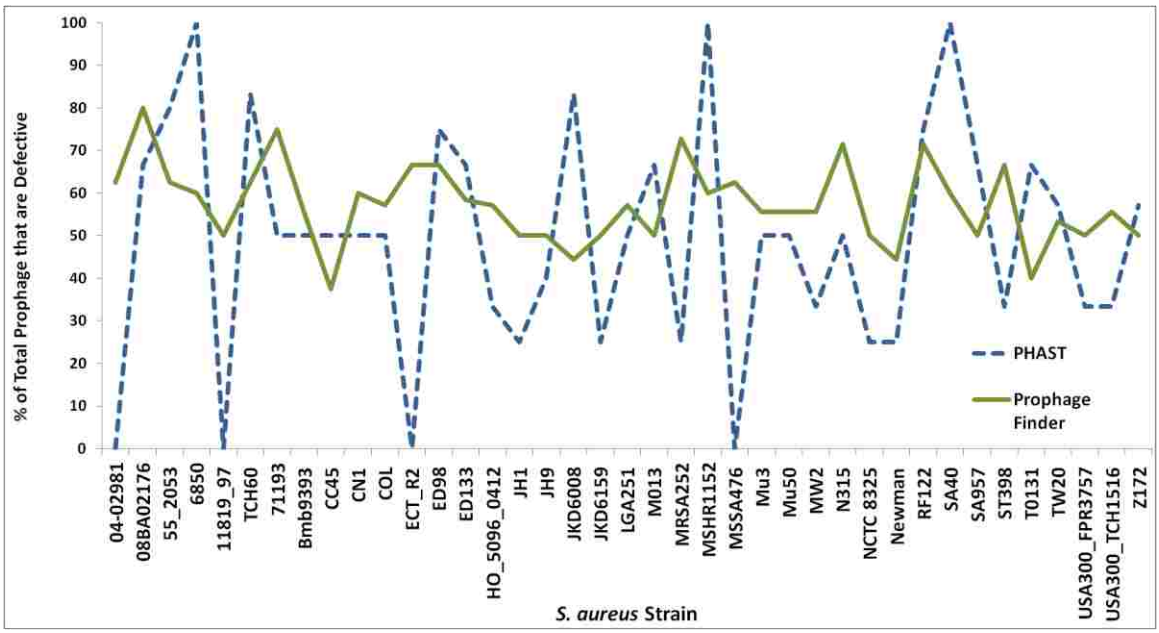

Figure 8: Defective percentage of total identified phage for PHAST and Prophage Finder. Figure 8 shows what percentage of the phage identified for each host strain by PHAST (blue line) and Prophage Finder (green line) were found to be defective. The percentage of defective phage found for Prophage Finder was more consistent and typically above 50%, with the lowest being 37% defective phage for strain CC45. The percentage for PHAST fluctuated much more, ranging from 0% to 100%.

Although Prophage Finder detected more prophage, for many strains it did not identify prophage otherwise found by PHAST. Using TCH60 as an example, PHAST identified more prophage than BLAST, uncovering three prophage as well as three 'defective' phage, while BLAST only found two prophage; Prophage Finder identified eight prophage including three not found by PHAST, while one of the six prophage uncovered by PHAST was not found by Prophage Finder. Figure 9 compares the results found by all three methods (PHAST, BLAST and Prophage Finder) for strain TCH60 alone, and figure 10 compares the results for all three methods across every *S. aureus* strain.

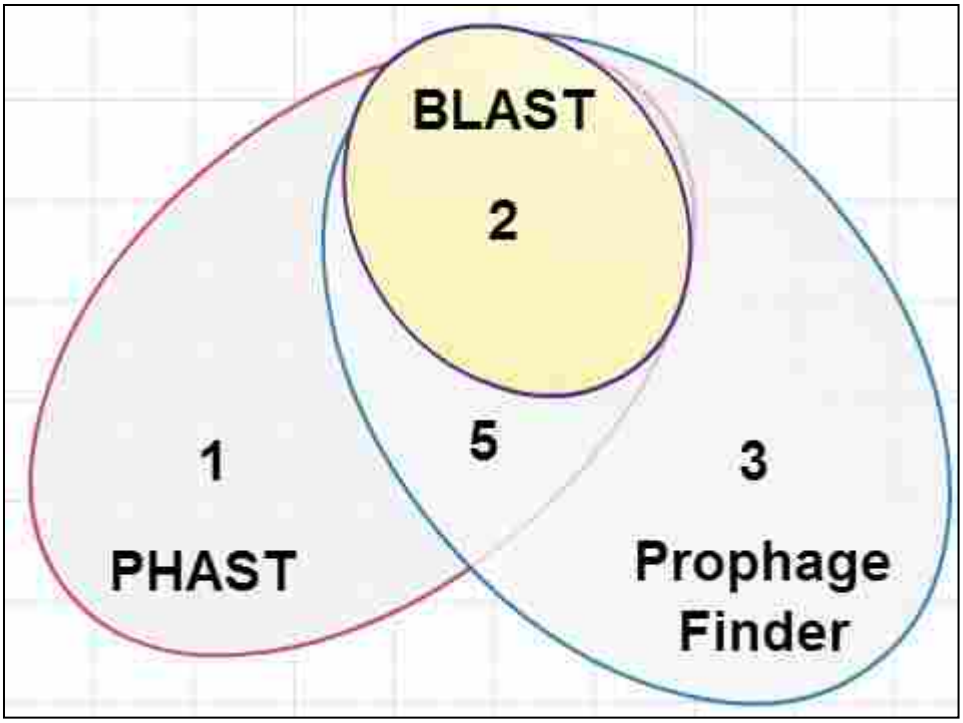

Figure 9: Euler diagram of prophage detected in TCH60 genome. Figure 9 displays the distribution of detected prophage in the *S. aureus* TCH60 genome using PHAST, BLAST and Prophage Finder. BLAST identified two prophage, both of which were detected by PHAST and Prophage Finder. PHAST detected six prophage, including one that was overlooked by Prophage Finder. Prophage Finder uncovered eight prophage, three of which were not detected by PHAST.

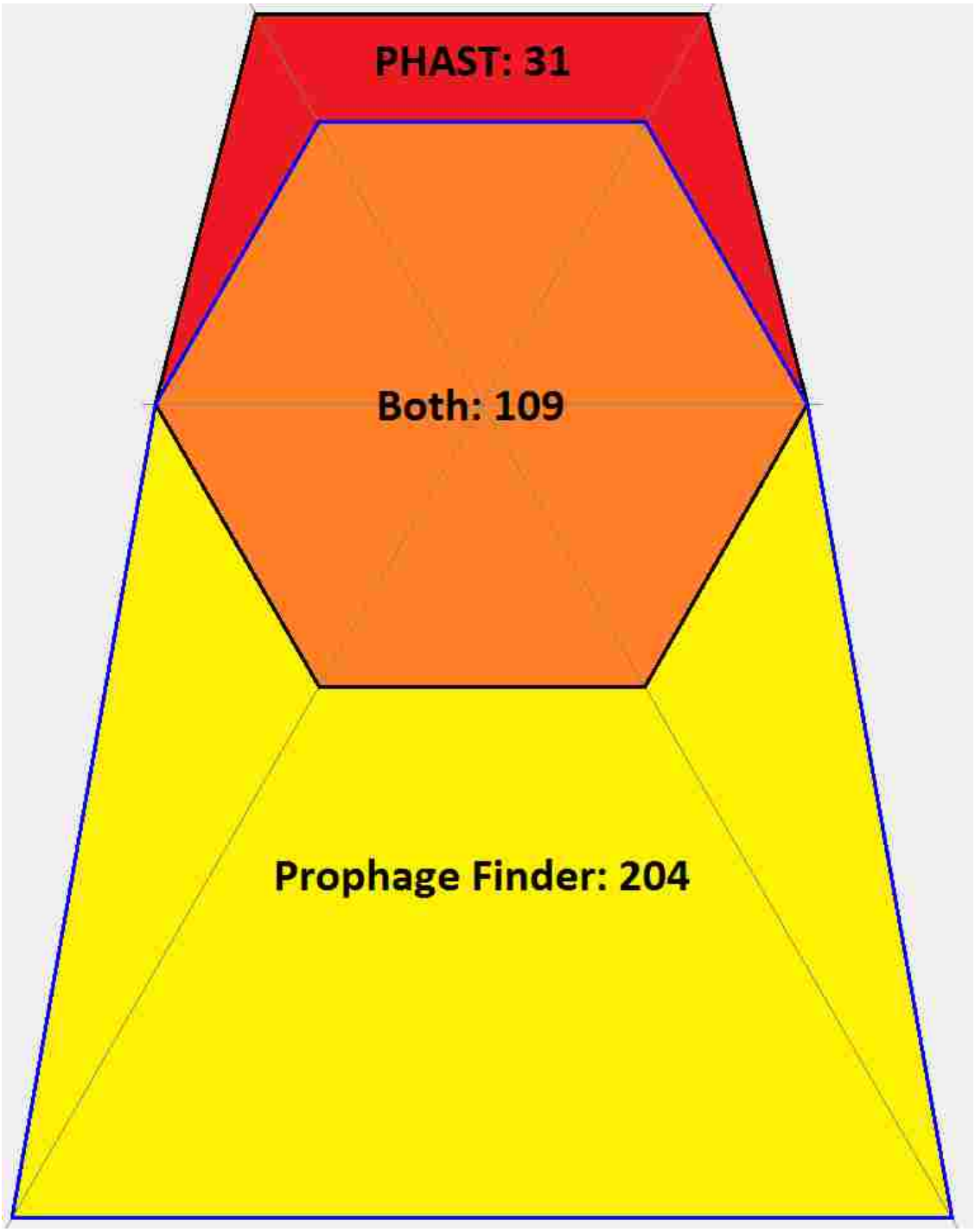

Figure 10: Diagram of prophage detected by PHAST versus Prophage Finder. Figure 10 is an area-proportional diagram comparing the prophage found by PHAST and Prophage Finder. The red area labeled 'PHAST' shows the number of prophage identified only by PHAST (31), the yellow area labeled 'Prophage Finder' shows the number of prophage identified by only Prophage Finder, and the orange area labeled 'Both' shows the number of common prophage that both programs identified.

2.2. PHACTS life cycle predictions

Six defective phage were identified as virulent by PHACTS; three were 'confident' predictions and three were 'non-confident.' Figure 11 compares the probability scores of the six phage.

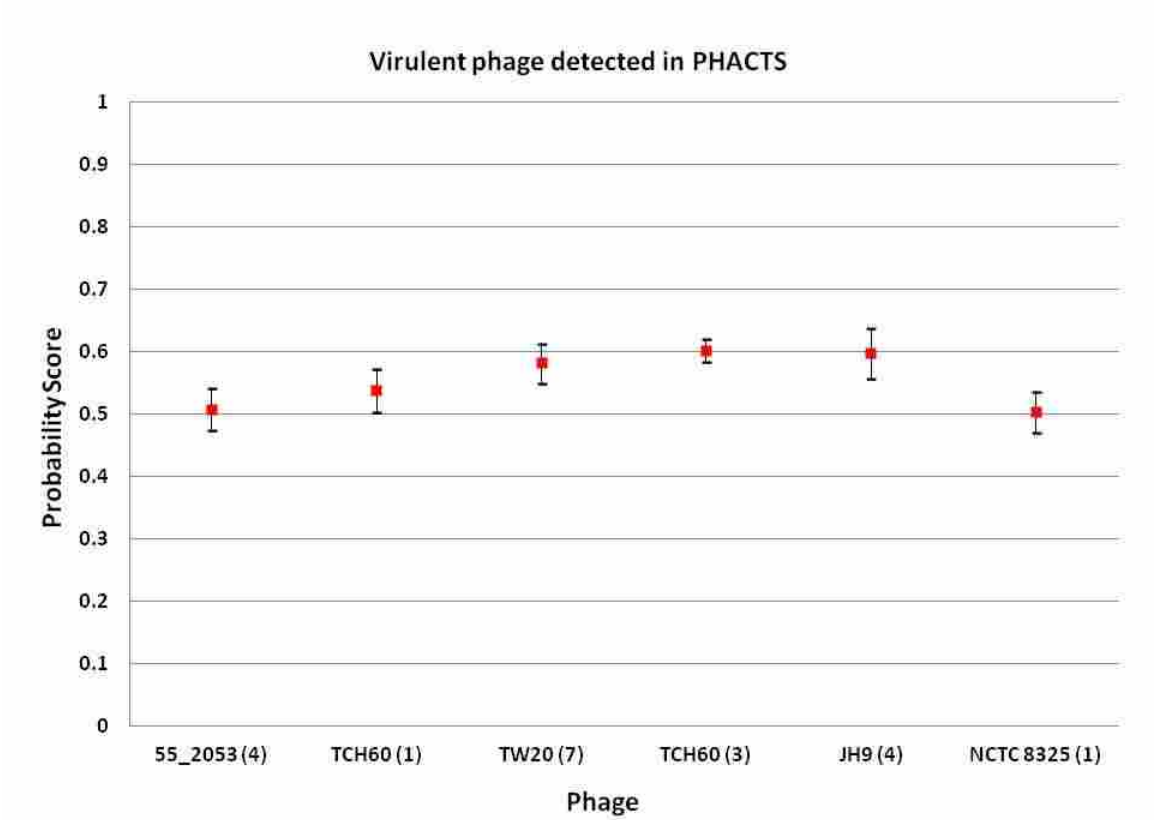

Figure 11: Comparison of probability scores for potential 'zombie' phage. Figure 11 compares the probability scores of the six potential 'zombie' phage identified by PHACTS. The bars extending from each probability score represent one standard deviation.

2.3. Genomic signature distances

To further test if the putative defective virulent phage belonged into the virulent category, we measured the similarity of their DNA signature and that of the host. Virulent phage are expected to have a larger signature distance to that of the host compared to temperate ones, as the temperate phage – due to their integration within the host genome – acquire a more similar signature over time, and therefore a shorter signature distance to the host (Deschavanne *et al*., 2010). The genomic signature distances were determined for every phage identified by PHAST and Prophage Finder, as well as the nine virulent *S. aureus* phage from the NCBI database. As expected, a clear separation was observed between the known temperate and virulent phage, with a smaller distance observed between temperate phage and hosts than between the virulent phage and hosts. The signature distances for the potential zombie phage are shown in table A.3 in the Appendix, alongside the signature distances of the temperate phage from the same hosts and the known virulent phage. Three of the six potential zombie phage possessed a signature distance that clearly resembled that of a virulent phage, rather than a temperate phage: NCTC 8325 prophage 1, TW20 prophage 7, and TCH60 prophage 3. TCH60 prophage 1 had a signature distance that was greater than the known temperate phage, yet still less than all the virulent phage. Figures 12 – 17 illustrate the signature distances of the six potential zombie phage alongside the known temperate and virulent phage.

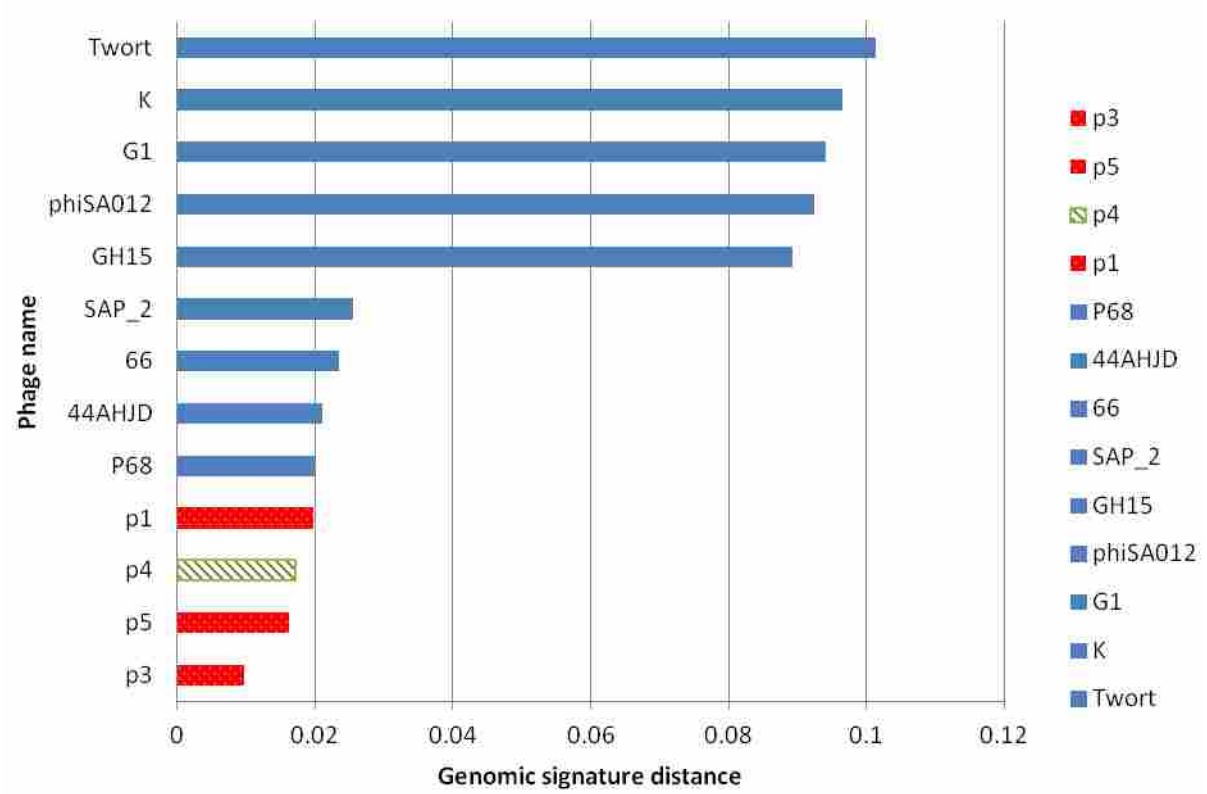

Figure 12: Signature distances for S. aureus 55-2053

Figure 12 shows the signature distances for the potential zombie phage alongside the temperate phage for host strain 55_2053 and the known virulent phage. The green striped bar represents the potential zombie phage, the red dotted bars represent the temperate phage, and the blue bars represent the known virulent phage. A separation is observed between the known virulent and temperate phage. The potential zombie phage has a temperate-like signature distance.

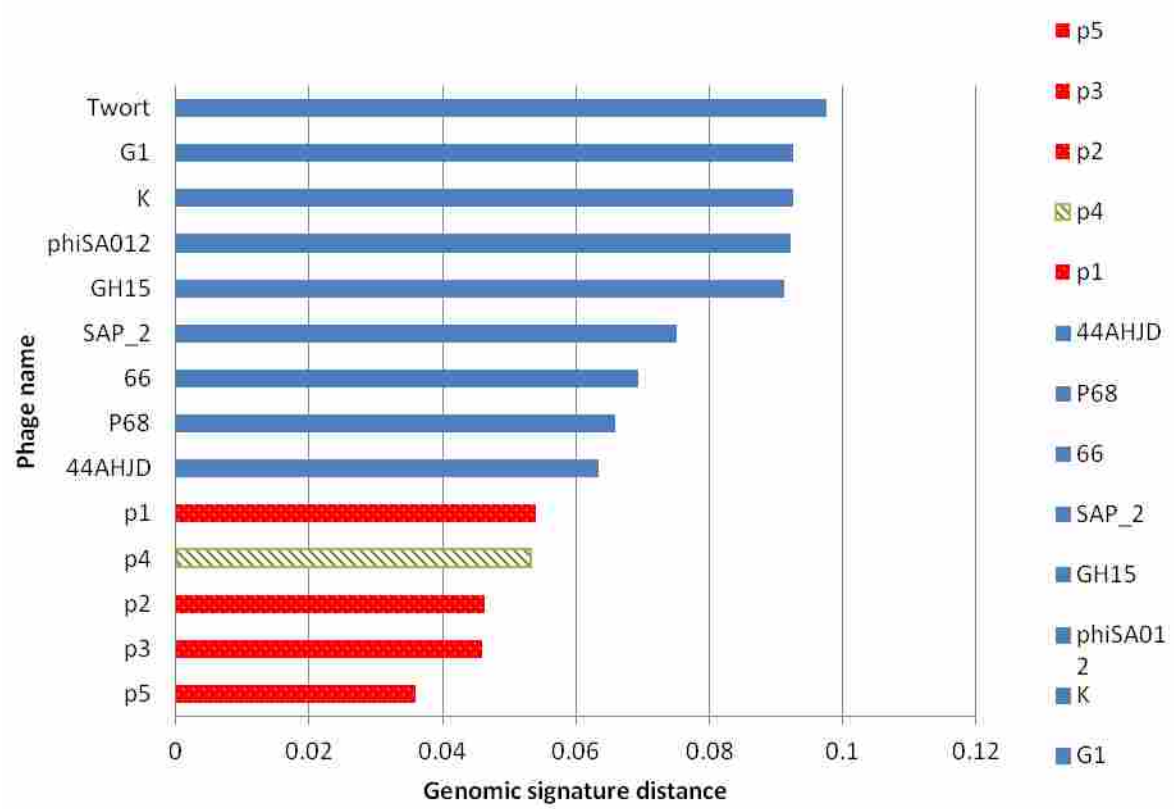

Figure 13: Signature distances for S. aureus JH9

Figure 13 shows the signature distances for the potential zombie phage alongside the temperate phage for host strain JH9 and the known virulent phage. The green striped bar represents the potential zombie phage, the red dotted bars represent the temperate phage, and the blue bars represent the known virulent phage. A separation is observed between the known virulent and temperate phage. The potential zombie phage has a temperate-like signature distance.

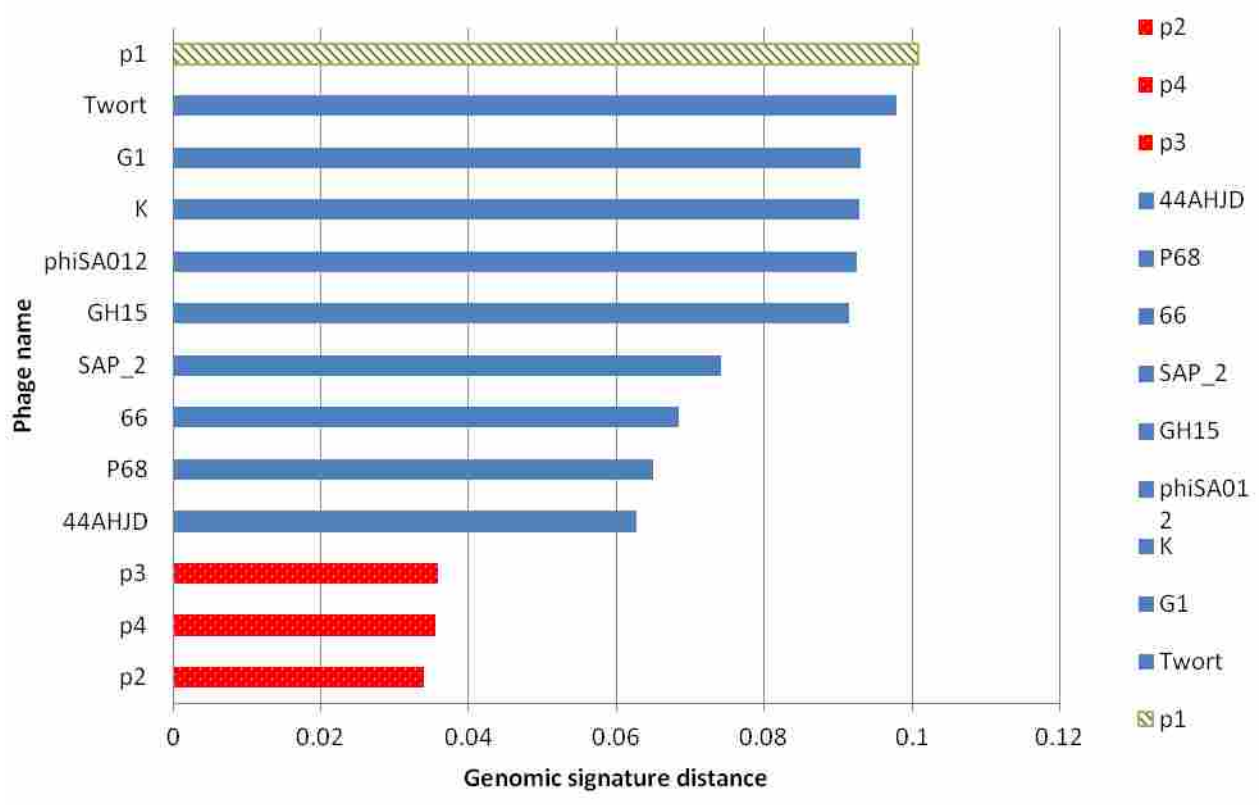

Figure 14: Signature distances for S. aureus NCTC 8325

Figure 14 shows the signature distances for the potential zombie phage alongside the temperate phage for host strain NCTC_8325 and the known virulent phage. The green striped bar represents the potential zombie phage, the red dotted bars represent the temperate phage, and the blue bars represent the known virulent phage. A separation is observed between the known virulent and temperate phage. The potential zombie phage has a virulent-like signature distance.

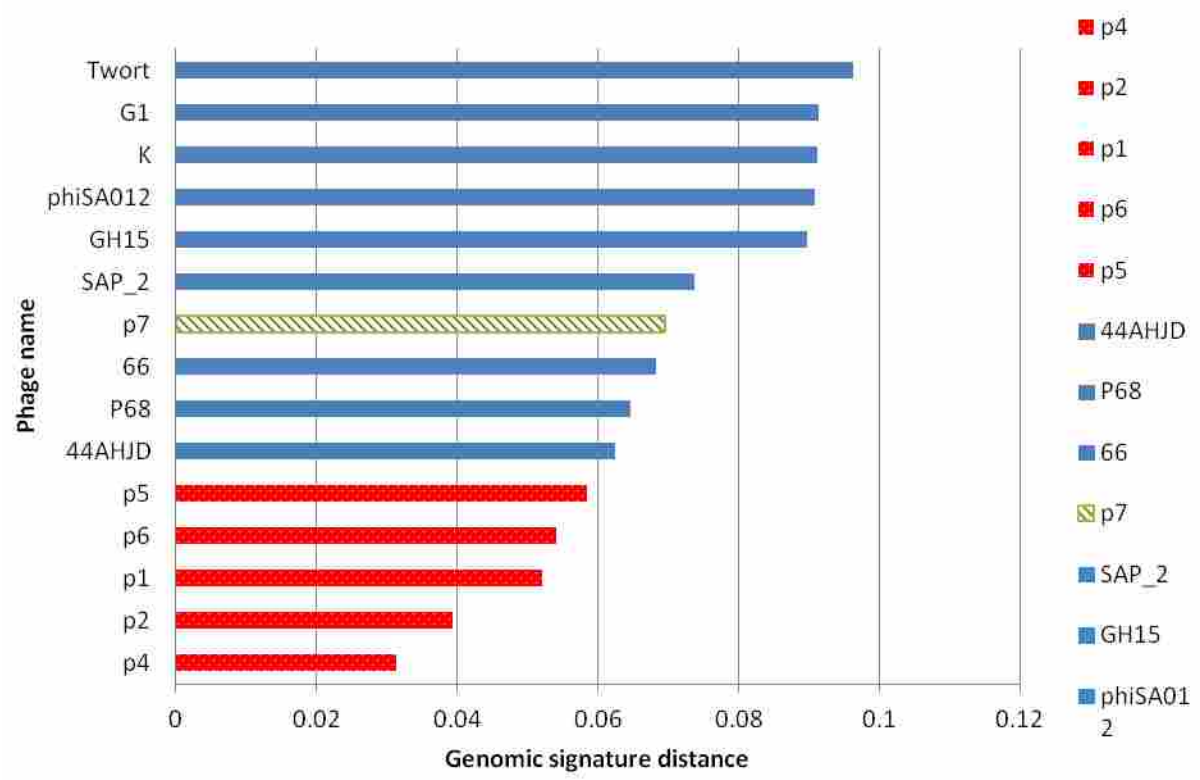

Figure 15: Signature distances for S. aureus TW20

Figure 15 shows the signature distances for the potential zombie phage alongside the temperate phage for host strain TW20 and the known virulent phage. The green striped bar represents the potential zombie phage, the red dotted bars represent the temperate phage, and the blue bars represent the known virulent phage. A separation is observed between the known virulent and temperate phage. The potential zombie phage has a virulent-like signature distance.

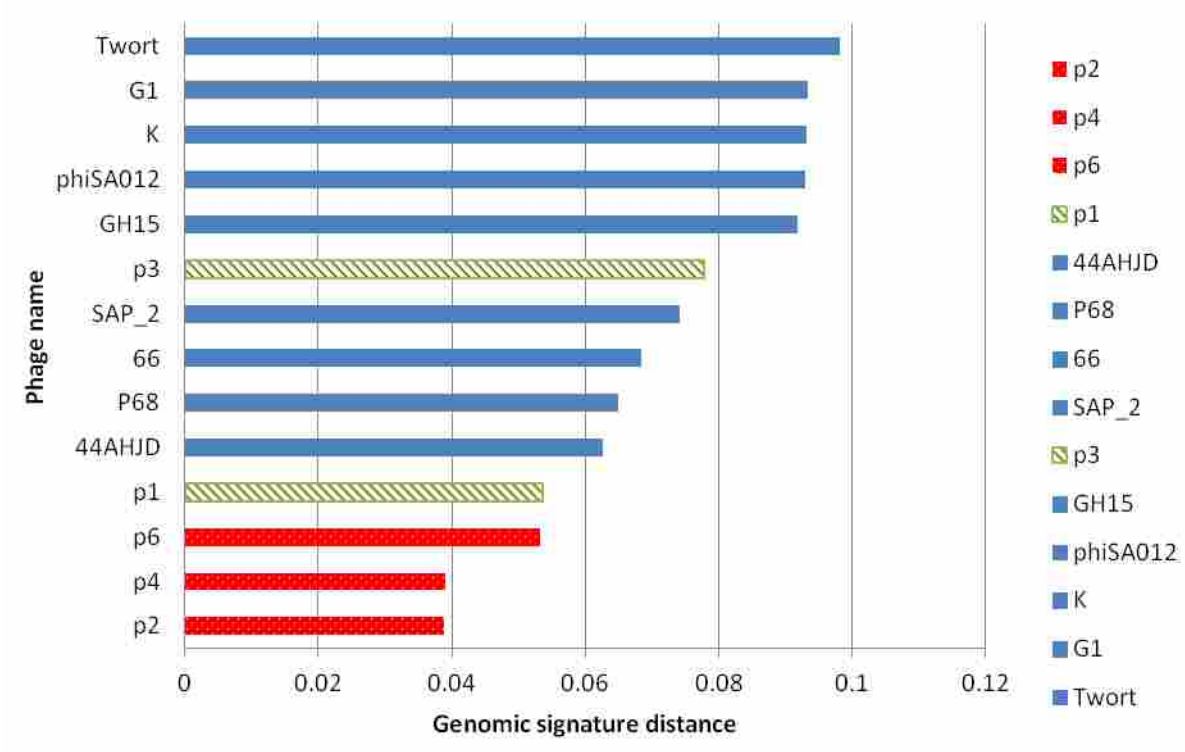

Figure 16: Signature distances for S. aureus TCH60

Figure 16 shows the signature distances for the two potential zombie phage alongside the temperate phage for host strain TCH60 and the known virulent phage. The green bars represents the potential zombie phage, the red bars represent the temperate phage, and the blue bars represent the known virulent phage. A separation is observed between the known virulent and temperate phage. One potential zombie phage (p3) has a virulent-like signature distance, while the other potential zombie phage (p1) has a temperate-like signature distance.

2.4. Codon usage

Like DNA signatures, coding genes from virulent phage are expected to have more dissimilar codon usages to that of their hosts than coding genes from temperate phage. The codon adaptation index (CAI) values were determined by DAMBE for the six potential zombie phage, the temperate phage of the corresponding host, and the known virulent phage. For CAI a value between 0 and 1 is assigned; a higher value indicates optimal codon usage, where 'optimal' is defined by a reference set of host genes. As expected, the known temperate phage for each host had CAI values that were closer to 1, compared to the known virulent phage. The known virulent phage had an average CAI of 0.415736. For NCTC 8325, the three temperate phage had an average CAI of 0.641423. The zombie phage, NCTC 8325 prophage 1, had a CAI of 0.28402. Similar results were observed for zombie phage JH9 prophage 4, TW20 prophage 7, and TCH60 prophage 3, which each had a CAI that was closer to that of the known virulent phage than the known temperate phage. For zombie phage 55_2053 prophage 4 and TCH60 prophage 1, the CAI values were not closely associated with the virulent phage.

Table A.4 in the Appendix displays the CAI values for every potential zombie phage, the corresponding host genomes and temperate phage, and known virulent phage. Figures 17 – 22 are graphical representations of these data. Figures 17 – 21 show the CAI for the six potential zombie phage alongside the temperate phage of the corresponding hosts, and the known virulent phage. A clear separation is observed between the known virulent and temperate phage. Four of the six potential zombie phage have a CAI that groups them among the virulent phage (NCTC8325 prophage 1, TCH60 prophage 3, JH9 prophage 4, and TW20 prophage 7).

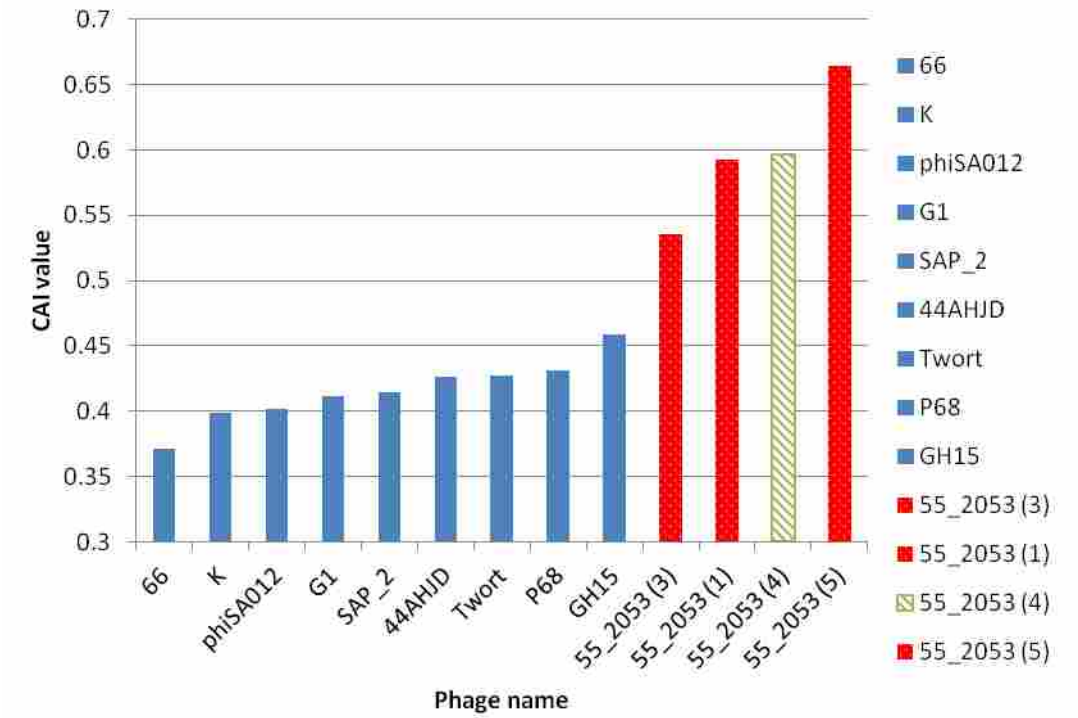

Figure 17: CAI values for 55_2053 phage

Figure 17 shows the CAI for the potential zombie phage alongside the temperate phage of host strain 55_2053, and the known virulent phage. The green striped bar represents the potential zombie phage, the red dotted bars represent the temperate phage, and the blue bars represent the known virulent phage.

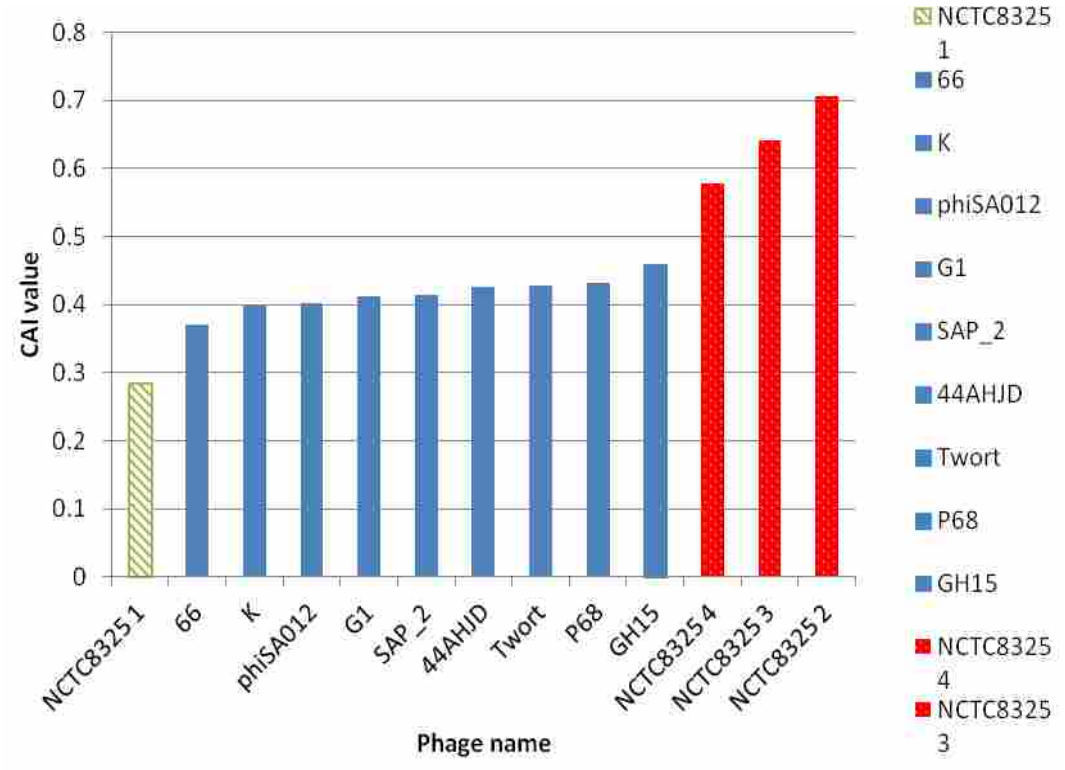

Figure 18: CAI values for NCTC8325 phage

Figure 18 shows the CAI for the potential zombie phage alongside the temperate phage of host strain NCTC 8325, and the known virulent phage. The green striped bar represents the potential zombie phage, the red dotted bars represent the temperate phage, and the blue bars represent the known virulent phage.

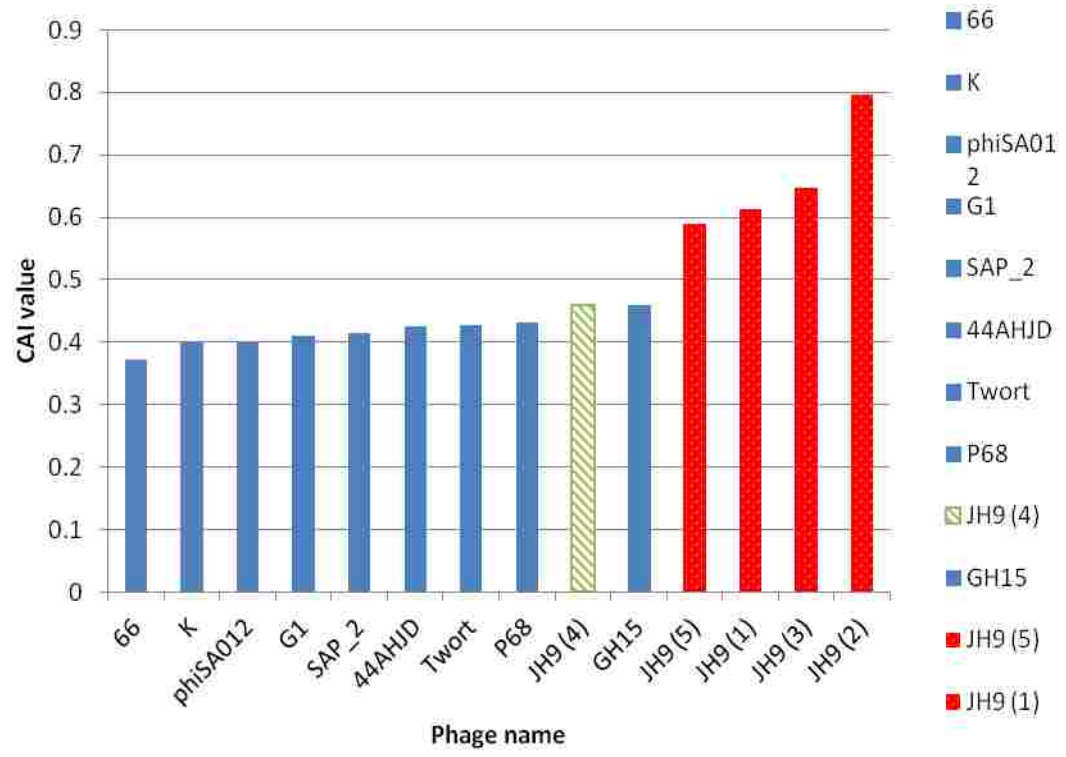

Figure 19: CAI values for JH9 phage

Figure 19 shows the CAI for the potential zombie phage alongside the temperate phage of host strain JH9, and the known virulent phage. The green striped bar represents the potential zombie phage, the red dotted bars represent the temperate phage, and the blue bars represent the known virulent phage.

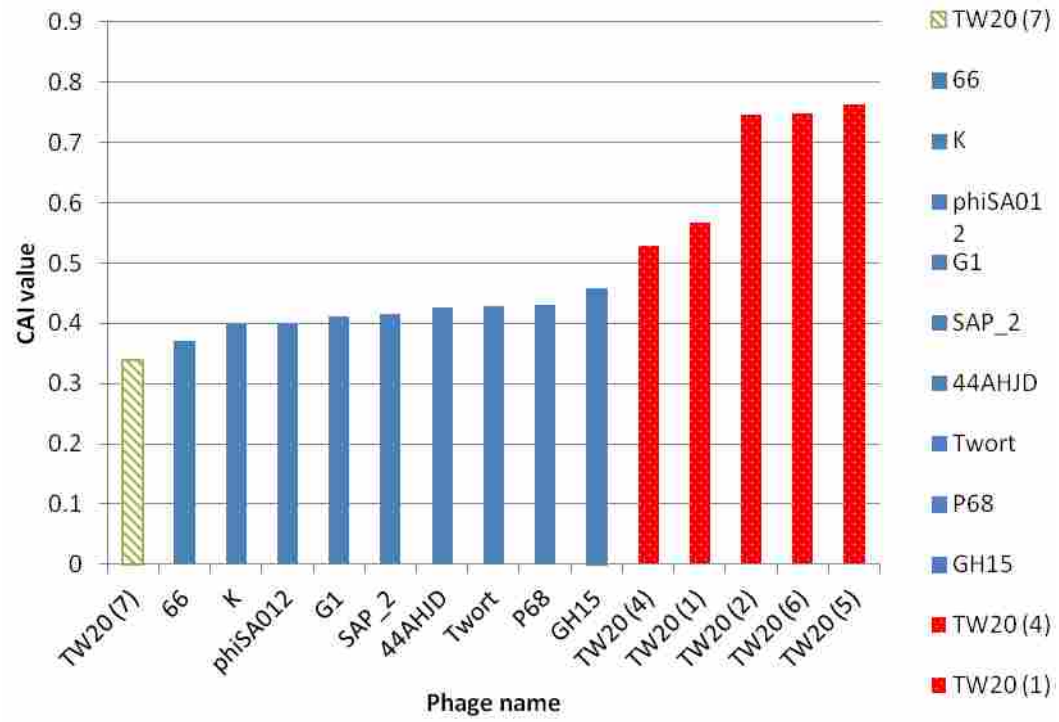

Figure 20: CAI values for TW20 phage

Figure 20 shows the CAI for the potential zombie phage alongside the temperate phage of host strain TW20, and the known virulent phage. The green striped bar represents the potential zombie phage, the red dotted bars represent the temperate phage, and the blue bars represent the known virulent phage.

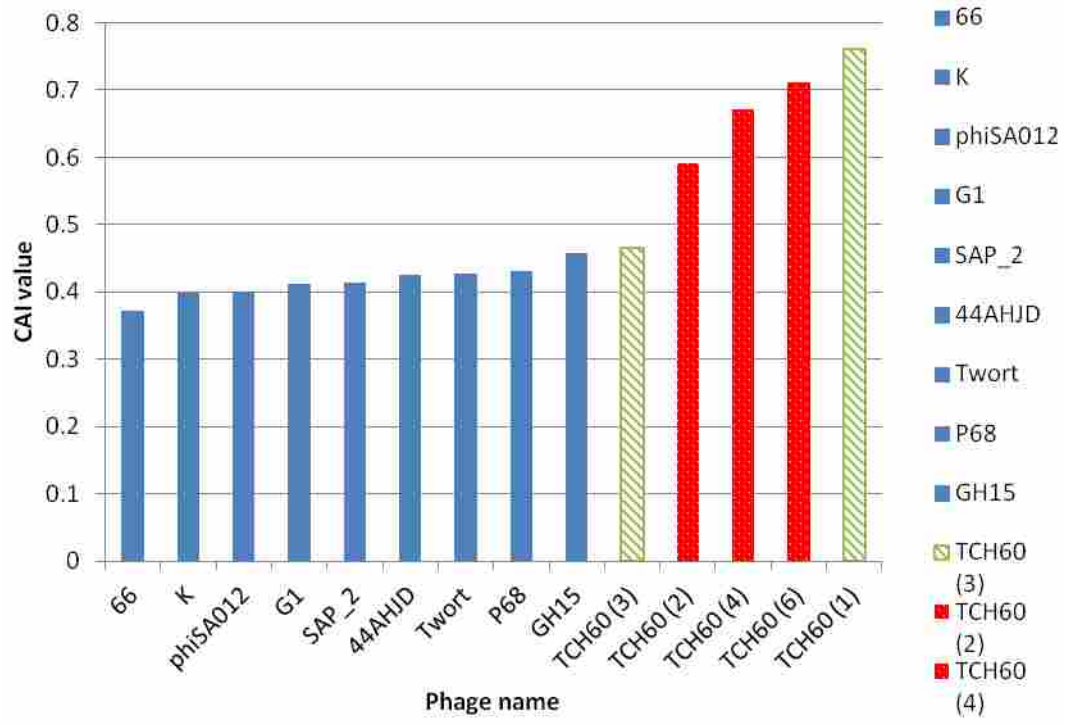

Figure 21: CAI values for TCH60 phage

Figure 21 shows the CAI for the potential zombie phage alongside the temperate phage of host strain TCH60, and the known virulent phage. The green striped bars represent the potential zombie phage, the red dotted bars represent the temperate phage, and the blue bars represent the known virulent phage.

3. DISCUSSION

3.1. Overview

As expected, for all host genomes analyzed PHAST uncovered more prophage than BLAST. It was also expected that PHAST would consistently identify more prophage than Prophage Finder, as the latter is an older program and attachment site prediction and the sensitivity of Prophage Finder have previously been found to be less accurate and efficient than that of PHAST (Zhou *et al*., 2011). However, Prophage Finder consistently identified more prophage than PHAST, with a larger portion of its identified prophage classified as defective. PHAST and Prophage Finder both identified phage that were not uncovered by the other program, necessitating that both were used to ensure that the identification of phage sequences was as thorough and complete as possible. Many studies use BLAST to identify prophage within host genomes, yet these results indicate that PHAST and Prophage Finder – especially when used together – are much more effective, uncovering significantly more sequences.

PHACTS identified six of these predicted phage as virulent. Although temperate phage may enter the lytic cycle through induction, virulent phage do not insert their sequences into the host genome and become a prophage. The two life cycles are distinct (McNair *et al.*, 2012). As such it would normally be expected that any prophage found, by the very nature of their location within the host genome, would be temperate phage. Yet, the results gathered from this program suggest that a few virulent phage have somehow become inserted within the host genome like prophage. As previously mentioned, some virulent phage are believed to have a temperate lineage; possibly through an HGT event, the phage may have lost genes required for integration or gained

genes involved in prompt lysis (Deschavanne *et al*., 2010). Indeed, virulent phage T1 is thought to have an evolutionary history tracing back to either temperate phage N15, HK022 or HK97 (Deschavanne *et al*., 2010). The results of this project support the possibility of the opposite event, whereby some temperate phage may in fact be previously virulent phage that somehow became inserted within the host genome. One way this could happen is if they acquired an integrase. If this were the case, these phage would most likely not necessarily be characterized as defective, due to the nature and time span of their integration with the host genome. Two of the sequences of these virulent and defective phage contained an integrase: TCH60 Prophage 3 and NCTC Prophage 1. For the sequences that did not contain an integrase, another way in which virulent phage could be integrated is if they infected a host that contained active integrases from a temperate phage, and their DNA was thus accidentally inserted into their hosts in a similar way in which retroinsertion accidentally creates pseudogenes from abundant RNA molecules in Eukaryotes (Zhang *et al.*, 2002). Integrated virulent phage would not have a mechanism for going back into a lytic stage and would thus become defective with time. Indeed, all six phage that were identified as virulent by PHACTS were classified as defective by PHAST and Prophage Finder. Additionally, it may be possible that for a virulent phage to become inserted during transduction, inducing host genome breakdown and incorporating its DNA during reassembly. To examine the potential life cycles of these phage further, their DNA sequence signatures and codon usages were characterized. It has been shown that DNA signatures can identify a virulent phage that has recently acquired a module for lysogeny (De Paepe *et al.*, 2014). Temperate phage typically have a much smaller signature distance to the host genome's

signature, compared to virulent phage, as they are integrated as a prophage and subject to the same selective pressures and mutations (De Paepe *et al.*, 2014). This can be seen as analogous to the 'amelioration' process that, over time, results in horizontally acquired genes assuming similar molecular characteristics to the host sequence (Lawrence and Ochman, 1997). As demonstrated by Marri and Golding (2008), genes can be somewhat differentiated by their relative residency times within a genome. As such, a virulent phage that recently 'converted' to a temperate life cycle would have a larger signature distance than expected, compared to other temperate phage, due to its relatively shorter residency time within the host genome (Deschavanne *et al*., 2010; De Paepe *et al.*, 2014). It has been hypothesized that this is a result of temperate phage becoming closely associated with the host genome during the prophage state, adopting the characteristics of the surrounding host sequence over time, similar to what has been observed with horizontally acquired genes (Deschavanne *et al*., 2010). It has also been suggested that analyzing genomic signature distances could help identify a former temperate phage that has recently lost the ability to insert itself within the host genome. An example of this is lytic phage T1, which has a genomic signature distance that resembles that of a temperate phage. Furthermore, a temperate lineage for this phage is supported by phylogenetic analysis (Deschavanne *et al*., 2010).

Of the six phage that were identified as virulent by PHACTS—the potential zombie phage—three were shown to have a signature distance that resembled virulent phage: TW20 prophage 7 and TCH60 prophage 3, which were classified as confident predictions, and NCTC 8325 prophage 1, which was classified as a non-confident prediction. The DNA signature distances of these three phage to their host genomes

signatures were clearly much higher than those of the temperate phage from the corresponding hosts (Figures 12 – 16).

Examination of codon usage in phage has shown that virulent phage tend to have higher codon usage biases and larger compositional differences to the host genome compared to temperate phage (Bailly-Bechet *et al*., 2007). It has been suggested that this is due to temperate phages having the same mutational biases as the host, as they are inserted within the genome as a prophage. This typically results in a much closer genomic composition between a temperate phage and the host, compared to a virulent phage and its host. Additionally, the higher codon usage bias in virulent phage may allow for the faster replication and efficient translation that is necessary for the lytic life cycle (Bailly-Bechet *et al*., 2007).

CAI, a measurement of synonymous codon usage bias, essentially measures the degree to which selection has shaped patterns of codon usage; this means it can be used to assess the extent to which viral genes have adapted to their hosts (Sharp and Li, 1987). It is believed that temperate phage exhibit a codon usage more similar to that of the host as the prophage state shares the same mutation spectrum as the host genome; further, a prophage – due to the increased time span of association with the host, compared to a strictly virulent phage – has a much higher chance of recombining or acquiring host genes (Chithambaram *et al.*, 2014).

As expected, the difference in codon adaptation index (CAI) values between hosts and temperate phage was less than the difference between hosts and known virulent phage (see Table A.4 and Figures 17– 21). For NCTC prophage 1, a potential zombie phage, the difference in CAI value did not correspond with the temperate phage, and was

actually lower than all known virulent phage. This is consistent with the PHACTS life cycle identification. Interestingly, NCTC prophage 1 also had the greatest genomic signature distance from its host. TCH60 prophage 3, TW20 prophage 7, and JH9 prophage 4, three other potential zombie phage, also had CAI value that more closely resembled virulent phage than temperate phage, while 55_2053 prophage 4 and TCH60 prophage 1 (the two remaining potential zombie phage) did not have CAI values that corresponded more closely with those of the known virulent phage.

Host strains MRSA177 and MRSA252 – of interest due to the fact that MRSA, as previously mentioned, has become a global health concern – were each found to contain a number of prophage, both defective and intact. Three prophage were identified for MRSA177 (one defective, two intact) and four prophage were identified for MRSA252 (one defective, three intact). None of the prophage identified for either strain were found to be potential zombie phage.

Examining all the data together, for the six potential zombie phage identified by PHAST, three (NCTC 8325 prophage 1, TW20 prophage 7, and TCH60 prophage 3) were shown to have signature distances that corresponded to what would be expected for a previously virulent phage. The CAI value for NCTC 8325 prophage 1, TCH60 prophage 3, TW20 prophage 7, and JH9 prophage 4 corresponded with this. This means that for NCTC 8325 prophage 1, TCH60 prophage 3, TW20 prophage 7, and JH9 prophage 4, the results of PHACTS, the characterization of genomic signature distance and CAI values are all consistent with a previously virulent life cycle. The remaining two potential zombie phage had conflicting results across the three methods. This does not necessarily indicate that they were not previously virulent phage – it is possible that their

insertion into the host genome occurred earlier than for the other zombie phage, and thus the data for signature distance and codon usage is not as clear.

3.2. Possible limitations

Although the results of this project indicate that the classification of phage life cycles may not be as distinct as previously believed, it is difficult, and perhaps not possible, to distinguish a phage as previously virulent if the insertion did not occur relatively recently, as a longer time span would result in signature distances and CAI values that resemble that of a temperate phage.

3.3. Next steps

Further analysis could focus on several lines of study. Alternative hypotheses for the defective zombie phage could be explored; perhaps rather than being previously virulent phage, some other genetic process or evolutionary event could account for their detection by PHACTS. Additional methods could be used for assessing the life cycles of the phage. For instance, GeneMarkS is a self-training program that can be used to predict genes in unknown sequences, using Markov chain models of both coding and non-coding DNA sequences to identify gene starts; the parameters for these predictive models are defined by training sets of sequences of known type (Besemer *et al.,* 2001). Thus, using this program it could be possible to examine the defective phage identified by PHAST and Prophage Finder to distinguish between prophage and previously virulent phage that have become dormant. The program would first be 'trained' by inputting the sequences of known temperate and virulent phage, and then through the identification of sequence motifs the defective phage could be characterized as either prophage or previously virulent phage. Phylogenetic trees could be constructed to assess the evolutionary

histories of the zombie phage (Roberts *et al.*, 2004; Deschavanne *et al*., 2010). Additionally, the defective phage could be compared with respective "prototype" phage to characterize the genetic defects, such as frame-shift mutations, deletions and insertions (Asadulghani *et al*., 2009). This could be performed through a multiple sequence alignment using CLUSTALW, and visualized by the multiple alignment editor Jalview (see Figure 22). Finally, an experiment could be designed to test the various models of virulent phage insertion suggested here, determining whether the incorporation of an integrase is a viable mechanism, or if virulent phage may insert into the host genome during transduction.

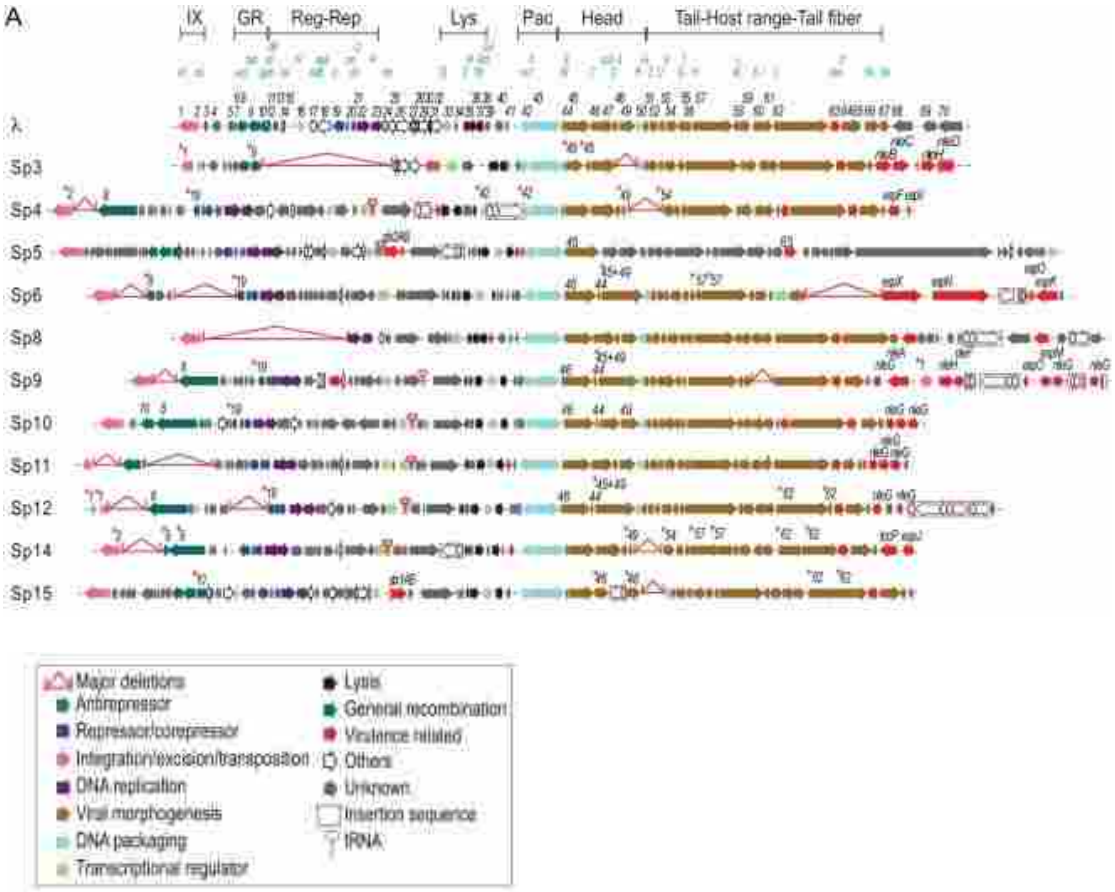

Figure 22: Characterization of phage defects through multiple sequence alignment

Figure 22 displays a multiple sequence alignment in which the genomic organization of various defective lambdoid phage from *Escherichia coli* O157 are compared to a corresponding prototypical phage genome, identifying deletions, insertions, and other genetic defects (Figure edited from Asadulghani et al., 2009).

4. SUMMARY

It is possible that some temperate phage are in fact previously virulent phage that erroneously inserted themselves within the host genome. Despite the belief that phage life cycles are largely distinct – with virulent phage never inserting themselves in the host genome – defective phage may be the remnants of past evolutionary events in which a virulent phage acquired a lysogeny module, such as an integrase gene. Defective phage, previously believed to be biologically inert 'garbage' DNA, have been shown to have an enormous ecological and evolutionary impact. The characterization of phage life cycles – including these cryptic zombie phage – thus has important consequences for our understanding of microbial communities. Through analysis with PHACTS and characterization of genomic signature distance and codon usage, the life cycles of these zombies were examined. The data supported the possibility that at least some of these zombies are defective, previously virulent phage; this is a novel finding as it may contradict the belief that virulent phage never insert their DNA into the host genome. Further, it was shown that PHAST and Prophage Finder are significantly more effective at uncovering prophage sequences than BLAST.

5. LAY SUMMARY

Phage play an integral role in microbial evolution, with an immense, constantly shifting population. This constitutes a complex, dynamic web of evolutionary interactions between phage, their hosts, and the bacterial world at large. This study identified prophage sequences within bacterial host genomes using the programs PHAST and Prophage Finder, which uncovered more prophage than indicated by previously published results, as these programs are notably more accurate and efficient than other prophage identification methods (such as BLAST). *S. aureus*, the host organism that was analyzed in this study, is a dangerous human pathogen with important clinical significance. Additionally, it was shown that there is potentially a connection between 'defective' and 'incomplete' phage and previously virulent phage that have lost their virulence genes.

LITERATURE CITED

Acar, J.F., and Moulin, G. (2012). Antimicrobial resistance: a complex issue. Revue scientifique et technique 31:1, 23-31.

Alonso, A., Sanchez, P., and Martinez, J.L. (2001). Environmental selection of antibiotic resistance genes. Environmental microbiology 3(1): 1-9.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology 215(3): 403-410.

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

Belcaid M., Bergeron A., Poisson G. (2010). Mosaic graphs and comparative genomics in phage communities. Journal of Computational Biology 17: 1315-1326.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2008). GenBank. Nucleic Acids Research 36: D25-D30.

Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucelic Acids Research 29(12): 2607-2618.

Bikandi, J., San Millá, R., Rementeria, A., and Garaizar, J. (2004). In Silico analysis of complete bacterial genomes: PCR, AFLP-PCR, and endonuclease restriction. Bioinformatics 20: 798-799. DOI: 10.1093/bioinformatics/btg491.

Bose, M. and Barber, R.D. (2006). Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. In Silico Biology 6 (3): 223-7.

Brussow, H., Canchaya, C., and Hardt, W. (2004). Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. Microbiology and molecular biology reviews 68(3): 560-602.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brussow, H. (2003). Prophage genomics. Microbiology and molecular biology reviews 67(2): 238-276.

Canchaya, C.A., Ventura, M., and van Sinderen, D. "Bacteriophage bioinformatics and genomics." *Bacteriophage: Genetics and Molecular Biology*. Eds. Stephen McGrath and Douwe van Sinderen. Norfollk: Horizon Scientific Press, 2007. Print.

Chithambaram, S., Prabhakaran, R., and Xia, X. (2014). Differential codon adaptation between dsDNA and ssDNA phages in Escherichia coli. Molecular Biology and Evolution 31(6): 1606-1617.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999). Improved microbial gene identification with GLIMMER. Nucleic Acids Research 27(23): 4636-4641.

De Paepe, M., Hutinet, G., Son, O., Amarir-Bouhram, J., Schbath, S., and Petit, M. (2014). Temperate Phages Acquire DNA from Defective Prophages by Relaxed Homologous Recombination: The Role of Rad52-Like Recombinases. PLoS Genet 10(3): e1004181. doi:10.1371/journal.pgen.1004181

Deschavanne, P., DuBow, M.S., and Regeard, C. (2010). The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. Virology Journal 2010;7:163.

Dye, C. (2009). Doomsday postponed? Preventing and reversing epidemics of drug-resistant tuberculosis. Nature Reviews Microbiology 7, 81-87.

Elhai, J., Hailan, L., and Arnaud, T. (2012). Detection of horizontal transfer of individual genes by anomalous oligomer frequencies. BMC genomics 13: 245.

Ester, M., Kriegel, H., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD-96 Proceedings, Institute for Computer Science, University of Munich: 226-231.

Fett, Boba (1977). Star Wars Episode IV: A New Hope, 66-501.

Fonterre, P. (2010). Defining Life: The Virus Viewpoint. Origins of life and evolution of the biosphere 40(2): 151-160.

Frost, L.S., Leplae, R., Summers, A.O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. Nature Reviews Microbiology 3, 722-732.

Golding, G.B., and Marri, P.R. (2008). Gene amelioration demonstrated: the journey of nascent genes in bacteria. Genome 51(2): 164-168.

Griffith, F. (1928). The Significance of Pneumococcal Types. Journal of Hygiene 27 (2): 113–159.

Griffiths, A.J.F., Miller, J.H., Suzuki D.T., *et al*. An Introduction to Genetic Analysis. 7th edition. New York: W.H. Freeman; 2000. Transduction.

Hatfull, G.F. and Hendrix, R.W. (2011). Bacteriophages and their genomes. Current Opinions in Virology 2011;1:298-303.

Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E., and Hatfull, G.F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: All the

world's a phage. Proceedings of the National Academy of Sciences of the United States of America 96: 2192–2197.

Hendrix, R.W. (2002). Bacteriophages: Evolution of the majority. Theoretical population biology 61, 471-480.

Hendrix, R.W. (2003). Bacteriophage genomics. Current Opinion in Microbiology 6: 506-511.

Hershberg, R., and Petrov, D.A. (2008). Selection on codon bias. Annual Review of Genetics 42: 287-299.

Housby, J.N., and Mann, N.H. (2009). Phage therapy. Drug Discovery Today 14: 536-540.

Keen, E.C. (2012). Phage therapy: concept to cure. Frontiers in microbiology 3: 238.

Klevens, R.M., Morrison, M.A., Nadle, J., Petit, S., Gershman, K., Ray, S., Harrison, L.H., Lynfield, R., Dumyati, G., Townes, J.M., Craig, A.S., Zell, E.R., Fosheim, G.E., McDougal, L.K., Carey, R.B., and Fridkin, S.K. (2007). Invasive Methicillin-Resistant *Staphylococcus aureus* infections in the United States. The Journal of the American Medical Association 298(15): 1763-1771.

Koonin, E.V., Makaroca, K.S., and Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. Annual Review of Microbiology 55: 709-742.

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

Lowy, F.D. (1998). Is *Staphylococcus aureus* an intracellular pathogen. Trends in microbiology 8: 341-344.

McCarthy, A.J., Witney, A.A., and Lindsay, J.A. (2012). *Staphylococcus aureus* temperate bacteriophage: carriage and horizontal gene transfer is lineage associated. Frontiers in Cellular and Infection Microbiology 2(6): 1-10.

McNair, K., Bailey, B.A., and Edwards, R.A. (2012). PHACTS, a computational approach to classifying the lifestyle of phages. Bioinformatics 28(5): 614-618.

Morello, E., Saussereau, E., Maura, D., Huerre, M., Touqui, L., *et al*. (2011). Pulmonary Bacteriophage Therapy on Pseudomonas aeruginosa Cystic Fibrosis Strains: First Steps Towards Treatment and Prevention. PLoS ONE 6(2): e16963. doi:10.1371/journal.pone.0016963.

Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299-304.

O'Connor, C. (2008). Isolating hereditary material: Frederick Griffith, Oswald Avery, Alfred Hershey, and Martha Chase. Nature Education 1(1).

O'Flaherty, S., Ross, R.P., Meaney, W., Fitzgerald, G.F., Elbreki, M.F., and Coffey, A. (2005). Potential of the Polyvalent Anti-Staphylococcus Bacteriophage K for Control of Antibiotic-Resistant Staphylococci from Hospitals. Applications of Environmental Microbiology 71(4): 1836-1842.

Onodera, K.K. (2010). Molecular biology and biotechnology of bacteriophage. Advances in biochemical engineering/biotechnology 119: 17-43.

Panis, G., Franche, N., Mejean, V., and Ansaldi, M. (2012). Insights into the Functions of a Prophage Recombination Directionality Factor. Viruses 4(11): 2417-2431.

Philippe, H., and Douady, C.J. (2003). Horizontal gene transfer and phylogenetics. Current Opinions in microbiology 6(5): 498-505.

Proux, C., van Sinderen, D., Suarez, J., Garcia, P., Ladero, V., Fitzgerald, G.F., Desiere, F., and Brüssow, H. (2002). The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria. Journal of Bacteriology 184(21): 6026-36.

Putonti, C., Luo, Y., Katili, C., Chumakov, S., Fox, G.E., Graur, D., and Fofanov, Y. (2006). A computational tool for the genomic identification of regions of unusual compositional properties and its utilization in the detection of horizontally transferred sequences. Molecular Biology and Evolution 23(10), 1863-1868.

Rabinovich, L., Sigal, N., Borovok, I., Nir-Paz, R., and Herskovits, A.A. (2012). Prophage Excision Activates Listeria Competence Genes that Promote Phagosomal Escape and Virulence. Cell 150(4): 792-802.

Roberts, M.D., Martin, N.L., and Kropinski, A.M. (2004). The genome and proteome of coliphage T1. Virology 318(1):245-66.

Sharp, P. M., and Li, W.H. (1987). The codon adaptation index- a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research 15: 1281-1295.

Soon, W. W.; Hariharan, M.; Synder, M.P. (2013). High-throughput sequencing for biology and medicine. Molecular Systems Biology 9: 640-54.

Sturino, J.M., and Klaenhammer, T.R. (2006). Engineered bacteriophage-defence systems in bioprocessing. Nature Reviews microbiology 4: 395-404.

Syvanen, M. (1994). Horizontal gene transfer: evidence and possible consequences. Annual Review of Genetics 28: 237-261.

Thiele, H., Glandorf, J., and Hufnagel, P. (2010). Bioinformatics strategies in life sciences: from data processing and data warehousing to biological knowledge extraction. Journal of Integrative Bioinformatics 7(1): 141.

Wang, Y., Hill, K., Singh, S., and Kari, L. (2005). The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene 346: 173-185.

Wang, X., Kim, Y., Ma, Q., Hong, S.H., Pokusaeva, K., Sturino, J.M., and Wood, T.K. (2010). Cryptic prophages help bacteria cope with adverse environments. Nature communications 1:147 doi: 10.1038/ncomms1146.

Xia, X. and Xie, Z. (2001). DAMBE: Software Package for Data Analysis in Molecular Biology and Evolution. Journal of Heredity 92 (4): 371-373.

Xia, X. (2007). An Improved Implementation of Codon Adaptation Index. Evoluationary Bioinformatics Online 3: 53-58.

Zhou, Y., Liang, Y., Lynch, K.H., Dennis, J.J., and Wishart, D.S. (2011). PHAST: A Fast Phage Search Tool. Nucleic Acids Research 1-6.

Zhang, Z.,Harrison, P., and Gerstein, M. (2002). Identification and Analysis of Over 2000 Ribosomal Protein Pseudogenes in the Human Genome. Genome Research 12: 1466 - 1482.

Zhaxybayeva, O., and Doolittle, W.F. (2011). Lateral gene transfer. Current Biology 21(7): R242-R246.

APPENDIX

Table A.1: Number of prophage identified using PHAST and BLAST for S. aureus strains

	Number of prophage detected		
Strain	**BLAST**	**PHAST total**	**PHAST 'defective'**
04-02981	1	2	0
MSSA476	2	2	0
MW2	2	3	1
N315	1	2	1
JH1	4	4	1
JH9	3	6	5
ECT-R2	1	1	0
ED98	2	4	3
Mu3	2	4	2
Mu50	2	4	2
A8117	2	3	2
A9754	2	3	1
USA300 TCH1516	2	3	1
A9765	2	6	3
Newman	1	4	1
USA300 FPR3757	2	3	1
MRSA177	2	3	1
NCTC 8325	3	4	1
COL	2	3	2
C101	2	6	5
MRSA252	2	4	1
TCH60	2	6	5
MSHR1132	1	1	0
JKD6159	2	4	1
ED133	3	6	4
RF122	2	4	3
TW20	2	7	4
JKD6008	2	6	5
LGA251	1	2	1
M809	2	5	4

This table shows the number of phage detected using PHAST and BLAST for 30 *S. aureus* genomes. The total number of phage detected by PHAST is displayed, along with the number of 'defective' phage.

Table A.2: Number of prophage identified using PHAST and Prophage Finder

Strain	PHAST total	PHAST defective	Prophage Finder total	Prophage Finder defective
04-02981	2	0	8	5
08BA02176	3	2	5	4
55_2053	5	4	8	5
6850	3	3	5	3
11819_97	3	0	8	4
TCH60	6	5	8	5
71193	2	1	4	3
Bmb9393	4	2	9	5
CC45	4	2	8	3
CN1	2	1	5	3
COL	2	1	7	4
ECT_R2	1	0	6	4
ED98	4	3	9	6
ED133	6	4	12	7
HO_5096_0412	3	1	7	4
JH1	4	1	10	5
JH9	5	2	10	5
JKD6008	6	5	9	4
JKD6159	4	1	6	3
LGA251	2	1	7	4
M013	3	2	6	3
MRSA252	4	1	11	8
MSHR1152	1	1	5	3
MSSA476	2	0	8	5
Mu3	4	2	9	5
Mu50	4	2	9	5
MW2	3	1	9	5
N315	2	1	7	5
NCTC 8325	4	1	8	4
Newman	4	1	9	4
RF122	4	3	7	5
SA40	2	2	5	3
SA957	3	2	6	3
ST398	3	1	6	4
T0131	6	4	10	4
TW20	7	4	15	8
USA300_FPR3757	3	1	8	4
USA300_TCH1516	3	1	9	5
Z172	7	4	12	6

Table A.2 shows the total number of phage and the number of ‘defective’ phage detected using PHAST and Prophage Finder for every *S. aureus* genome.

Table A.3: Genomic signature distances

	55_2053	JH9	NCTC 8325	TCH60	TW20
Prophage 1	0.0197	0.0541	0.1009	0.0536	0.052
Prophage 2	NA	0.0464	0.0339	0.0388	0.0394
Prophage 3	0.0097	0.0459	0.0358	0.0778	NA
Prophage 4	0.0172	0.0533	0.0355	0.039	0.0314
Prophage 5	0.0163	0.0359	NA	NA	0.0584
Prophage 6	NA	NA	NA	0.0532	0.0541
Prophage 7	NA	NA	NA	NA	0.0695
44AHJD	0.021	0.0635	0.0627	0.0627	0.0624
66	0.0235	0.0694	0.0685	0.0684	0.0683
G1	0.0941	0.0926	0.093	0.0933	0.0913
GH15	0.0893	0.0912	0.0915	0.0919	0.0898
K	0.0966	0.0925	0.0929	0.0932	0.0912
P68	0.0199	0.0659	0.065	0.0649	0.0647
SAP_2	0.0255	0.0752	0.0742	0.0742	0.0738
Twort	0.1013	0.0975	0.0978	0.0982	0.0962
phiSA012	0.0924	0.0922	0.0925	0.0929	0.0908
Avg. Temp.	0.01523333	0.045575	0.03506667	0.043666667	0.04706
Avg. Virul.	0.06262222	0.08222222	0.08201111	0.082188889	0.0809444
"Zombie"	0.0172	0.0533	0.1009	0.0536, 0.0778	0.0695

Table A.3 shows the genomic signature distances for the prophage (temperate phage), virulent phage, and zombie phage for S. aureus strains 55_2053, JH9, NCTC 8325, TCH60, and TW20 (the five strains that were shown by PHACTS to potentially harbor zombie phage). The zombie phage are highlighted green, and the average signature distance is shown for temperate and virulent phage for each host strain. For values that are not available ('NA'), the host strain may not contain that prophage (for instance, strain 55_2053 only contains five prophage, and thus prophage 6 and 7 are shown as 'NA'). NA values that are highlighted red have been discarded due to insufficient data (calculations for signature distance are not reliable for phage with less than 20 coding sequences).

Table A.4: CAI values

	55_2053	JH9	NCTC 8325	TCH60	TW20
Prophage 1	0.5924	0.61252	0.28402	0.76153	0.56733
Prophage 2	NA	0.79685	0.70662	0.59212	0.74538
Prophage 3	0.5351	0.64721	0.64016	0.46491	NA
Prophage 4	0.59603	0.45847	0.57749	0.67151	0.52873
Prophage 5	0.66391	0.58877	NA	NA	0.76333
Prophage 6	NA	NA	NA	0.71162	0.74898
Prophage 7	NA	NA	NA	NA	0.33897
44AHJD	0.42629	---	---	---	---
66	0.37142	---	---	---	---
G1	0.41117	---	---	---	---
GH15	0.45872	---	---	---	---
K	0.39876	---	---	---	---
P68	0.43145	---	---	---	---
SAP_2	0.41479	---	---	---	---
Twort	0.42759	---	---	---	---
phiSA012	0.40143	---	---	---	---
Avg. Temp.	0.597137	0.661338	0.641423	0.658417	0.67075
Avg. Virul.	0.415736	---	---	---	---
"Zombie"	0.59603	0.45847	0.28402	0.76153, 0.46491	0.33897

Table A.4 shows the CAI for the potential zombie phage, the other prophage, and the known virulent phage. For values that are not available ('NA'), the host strain may not contain that prophage (for instance, strain 55_2053 only contains five prophage, and thus prophage 6 and 7 are shown as 'NA'). NA values that are highlighted red have been discarded due to insufficient data (calculations for codon usage and signature distance are not reliable for phage with less than 20 coding sequences). The average CAI for the virulent phage, and for the temperate phage of each host, are listed at the bottom alongside the zombie phage CAI for easy comparison of values.

Table A.5: Zombie phage data summary

	55_2053 (4)	JH9 (4)	NCTC_8325 (1)	TCH60 (1)	TCH60 (3)	TW20 (7)
Probability score	0.507	0.572	0.503	0.537	0.598	0.565
Confident prediction?	No	Yes	No	No	Yes	Yes
GSD	0.0172	0.0533	0.1009	0.0536	0.0778	0.0695
Virulent-like GSD?	No	No	Yes	No	Yes	Yes
CAI	0.59603	0.45847	0.28402	0.76153	0.46491	0.33897
Virulent-like CAI?	No	Yes	Yes	No	Yes	Yes
Prophage sequence length (Kb)	20.5	36	72	30.5	63.1	18.5

Table A.5 summarizes the information gathered on every potential zombie phage, including whether the prediction by PHACTS was confident or non-confident, the probability score, whether the signature distance and CAI were virulent-like, and the sequence length.