ABSTRACT

Genome-wide analysis of codon usage in Mycobacteriophage and *S. aureus*

Ian Nicholas Boys

Director: Tamarah Adair, Ph.D.

The molecular genetics of microbial life, including viruses and bacteria, is of great relevance to medicine, evolutionary theory, and other areas of science. Here, we study the implications of codon bias, the non-random usage of synonymous codons, and GC3, the frequency of codons with guanine or cytosine in the third nucleotide position. Biased codon usage frequently is the result of adaptation for optimal expression, reflecting tRNA abundance. It has previously been suggested that GC3 is similarly of functional significance, with implications for transcription and translation. We make use of the HHMI's mycobacteriophage genome database to analyze trends in codon bias and GC3 content in a diverse set of viruses with a common host. To this end we utilize genome landscapes to observe trends in genome-wide GC3 usage, and find that patterns in GC3 content can be informative in the difficult process of unraveling the complex relationships between mycobacteriophages. Similarly, we show that codon bias is likewise informative, though it in some instances contradicts the evidence provided by GC3 analysis. In a second portion of the study, we investigate the relationship between codon bias, GC3 content, and gene expression in *S. aureus* BUSA 2288. We find that codon bias is correlated with gene expression in S. aureus, suggesting that it does indeed reflect selection for expression. GC3 is, however, not tied to any trend in expression.

APPROVED BY DIRECTOR OF HONORS THESIS:

_____

Dr. Tamarah Adair, Department of Biology

APPROVED BY THE HONORS PROGRAM:

_____

Dr. Andrew Wisely, Director

DATE: _____

GENOME-WIDE ANALYSIS OF CODON USAGE IN MYCOBACTERIOPHAGE AND *S. AUREUS*

A Thesis submitted to the Faculty of

Baylor University

In Partial Fulfilment of the Requirements for the

Honors Program

By

Ian Nicholas Boys

Waco, TX

May 2015

TABLE OF CONTENTS

CHAPTER ONE

Introduction

General Overview of Investigation

The molecular study of microbial life, including viruses and bacteria, is an important field of science. Basic research in this area has advanced our understanding of as diverse subjects as the mechanisms of disease, the interplay of organisms within the environment, and the means by which organisms are capable of adapting over time. In this study, we investigate codon usage, an integral part of molecular genetic studies, in two microbial systems. Specifically, we study the relationships between codon bias, the non-random usage of synonymous codons, and GC3, the frequency of codons with guanine or cytosine in the third nucleotide position. We make use of the Science Education Alliance's mycobacteriophage genome database, hosted at phagesdb.org, to analyze trends in codon bias and GC3 content in a diverse set of viruses with a common host. In a second portion of the study, we investigate the relationship between codon bias, GC3 content, and gene expression in the bacterial pathogen, *Staphylococcus aureus*.

Phage Genomics

Bacteriophages, a group of bacteria-infecting viruses, are the largest known group of biological entities in the biosphere and may in fact be the largest source of genetic diversity on the earth (Hatfull et al., 2010). Phages have been developing and evolving for perhaps as long as their prokaryotic hosts, allowing such diversity to arise. Additionally, phages frequently exchange genetic material both with each other and with

their bacterial hosts, extending their already-impressive genetic diversity. Besides promoting diversity, the prevalence of such horizontal transfer has given rise to a great degree of mosaicism, or genomic composition resulting from two or more sources of genetic material, within their population as a whole (Pedulla et al., 2003).

Mycobacteriophages are a diverse group of viruses that exhibit the general trends of diversity and mosaicism shown by phages, at large. Mycobacteriophages can infect a wide range of bacterial hosts, including a number of disease-causing bacteria such as *M. leprae* and *M. tuberculosis,* the bacteria respectively responsible for forms of leprosy and tuberculosis, as well as innocuous bacteria, such as the model organism *M. smegmatis*, a common non-pathogenic soil bacterium. All mycobacteriophages isolated and studied thus far are dsDNA phages of the tailed morphotypes siphoviridae and myoviridae (Hatfull et al., 2008). As of early 2015, over 5800 mycobacteriophages have been isolated, over 800 of which have been sequenced, with over 400 annotated and available in GenBank with *Mycobacterium smegmatis* mc2 155 as their common host (phagesdb.org). This collection provides for study a diverse group of phages that have the potential to be in genetic communication (Pope et al., 2011).

The mycobacteriophage population is large and diverse. Individual phage genomes range from just under 41.5 kbp to over 164.5 kbp and contain a large variety of individual genes. As of late 2012, the time of database access for much of the data used in this study, there were 3631 unique groups of related genes, phamilies, which had been identified in the mycobacteriophage population (Cresawn et al., 2011). As more phages are isolated and sequenced, this number continues to increase. Such diversity among the gene phamilies reflects the diversity of the phages themselves. With such a large and

diverse population of phages, it is imperative that some scheme of classification be applied in order to facilitate the accessibility of the genetic information that has been obtained, as well as to assess and explore the genetic relationships between phages and their individual genes. To meet these needs, a cluster-based classification system was established. Traditionally, cluster designations are made primarily with reference to pairwise alignment as well as the number and order of shared genes. Classification of mycobacteriophages has given rise to twenty one clusters of phages, many of which contain subclusters that reflect the varying degrees of similarity between phages in each cluster (Pope et al., 2011)(PhagesDB.org).

With respect to the observed relationships between phages, three categories have been described: first, there are cases where there is evidence of a clear relationship between phages; second, there are cases where there are no easily-assessed relationships between phages; and third, there are cases in which the evidence is mixed, such as when certain portions of phage genomes appear to be related, while others do not (Hatfull et al., 2010). In order to accommodate the wide degree of variability in phage-phage relationships, four primary comparative techniques have been applied in the assessment of inter-phage relationships and phage clustering. These include dotplot analysis, comparison of average nucleotide identity (ANI), assessment of related gene families, and assessment of synteny (conserved gene order). These genome clustering techniques can be grouped in two general categories: nucleotide-level and gene-level analysis.

In the case of nucleotide-level analysis, the primary technique that is used is dotplot analysis, an assessment of overall sequence similarity between two genomes. If a dotplot of two aligned phage genomes indicates similarity that spans for over 50% of the

smaller of the two genomes, they are assigned to the same cluster. A secondary technique is a comparison of ANI between phages, a criterion that generally supports relationships indicated by dotplot analysis. No explicit parameters are established for ANI comparison, though like-clustered phages often share 80-98% ANI, whereas phages from different clusters may share only 53-59% ANI (Hatfull et al., 2010).

With regard to gene-level analysis, the number of phams (related genes) that are shared by different phages is indicative of genomic relationship and in most cases supports the data from dotplot and ANI analysis. A second gene-level analysis is that of genomic synteny, or conserved gene order, which for mycobacteriophages is conducted via Phamerator, a purpose-developed genome browser maintained by Steve Cresawn of James Madison University (Cresawn et al., 2011). The use of pairwise alignment and the correlation of regions of genomic similarity and gene loci provide a means of assessing past transfers of genetic material, an indicator of relatedness (Hatfull et al., 2010).

While these four techniques provide a robust means of exploring the genomic relationships among phages, there are certain instances in which their application still yields ambiguous results. For example, Mycobacteriophage *Patience*, a phage isolated in South Africa by members of the University of Kqazulu-Natal, was designated as a singleton, a phage that is not considered to be a member of any cluster (phagesdb.org). This decision was made because the evidence provided from standard techniques did not provide enough justification for its inclusion in a cluster, though it did show significant similarity to members of cluster H. As a result of an in-depth study of the genomics of *Patience*, it was suggested that the regions of similarity between *Patience* and other phages may reflect recent acquisitions via horizontal transfer (Pope et al., 2014). Due to

this degree of ambiguity, other means of comparing phage genomes, using codon usage and GC3 content, are being explored.

Staphylococcus aureus: Significance, Diversity, and Genomics

Staphylococcus aureus is a low-G+C Gram-positive bacteria in the Phylum Firmicutes, Class Bacilli and Family Staphylococcacea. *S. aureus* is a nonmotile cocci that frequently colonizes the human body and, while most *S. aureus* strains are commensals, at times they may become pathogenic and present a threat to their host. The recent proliferation of multidrug-resistant strains of *S. aureus* has placed great strain on the healthcare industry and has led to an increased incidence of death from staph infections, earning it a designation of a 'serious' hazard level (the second-highest possible) by the CDC (CDC, 2013). In light of this trend towards antibiotic resistant strains, alternative antibacterial treatments, including phage therapy, are being investigated (Rose et al., 2014). Phage therapy, which is a common means of treatment for bacterial diseases in parts of the former USSR, depends upon lytic phages to infect and lyse their bacterial hosts (Thiel, 2004). It is somewhat hindered by the fact that many phages have a very specific host range, necessitating further research into the factors that affect the host specificity and virulence of phages (Rose et al., 2014).

To date, several lineages of *S. aureus* have been identified that exhibit diversity in genetic content and host-specificity. These contain ten that are known to infect humans, six of which can also infect other animals (Sung et al., 2008). *S. aureus* strains contain numerous mobile genetic elements (MGEs), including prophages and plasmids, which in most cases make up 15-20% of the bacterial genome (Lindsay, 2014). Some of these MGEs have been identified as the source of antibiotic resistance observed in some

strains, necessitating their further study (Alibayov et al., 2014). Rapid horizontal transfer of these MGEs has been observed in co-colonization, showing the ease with which resistance genes can be transferred (McCarthy et al., 2014).

Pathogenicity islands, MGEs comprised of prophage-derived, conserved regions of *S. aureus* chromosomal DNA that are responsible for pathogenesis, exist in *S. aureus* and other microbial pathogens (Alibayov et al., 2014). The transfer of MGEs is most-often facilitated by transduction via helper phages, as direct transformation and conjugation are rare in *S. aureus* (Lindsay, 2014). *S. aureus* Pathogenicity islands (SaPi's) most-frequently provide staph with the ability to produce superantigens and other toxins that are responsible for the onset of toxic shock syndrome (Alibayov et al., 2014).

The nature of phages specific to *S. aureus* is two-fold; while they can serve as therapeutic agents, they also are agents in the transfer of genetic material. This duality presents a challenge to phage therapy. It has been suggested that codon bias and other genomic characteristics can be used to identify staph phages that are likely to have high therapeutic efficacy without presenting a risk of unwanted horizontal gene transfer or introduction of virulence factors (Bishal et al., 2012).

Codon Bias: Significance, Applications, and Measures

The study of codon bias, the skewed usage of synonymous codons, is an important component of genomic studies. Codon bias is enabled by the redundancy inherent in genetic code, as there exist 64 possible codons to code for only 20 amino acids. Thus, for most amino acids one of the possible codons can be disproportionately favored in a genetic sequence. It has been demonstrated that this biased codon usage is of

functional significance (Lawrence et al., 2002). Most importantly, it has been linked to selection for optimized expression (through increased translational efficiency), reflecting the pool of available tRNA molecules expressed by a given organism (Ikemura, 1981). Further, codon bias has been shown to result from a combination of factors. These factors include optimization for recombination, mRNA stability, and mutational pressure, the general trend among bacteria towards either GC-rich or GC-poor – but not intermediary – genomes (Behura and Severson, 2013). Codon bias is preserved among evolutionarily-linked organisms, and thus serves as a useful tool in the study of evolutionary and functional processes (Lawrence et al., 2002).

There are several ways to measure the degree by which synonymous codons are used in a non-random fashion. One of the oldest measures, the Effective Number of Codons (Nc or ENc) is a scale that provides a measure of bias across all codons (Wright, 1990). It is calculated for any given coding sequence through the following formula:

$$ENc = 2 + \left(\frac{9}{F_2}\right) + \left(\frac{1}{F_3}\right) + \left(\frac{5}{F_4}\right) + \left(\frac{3}{F_6}\right) \qquad (1)$$

The denominators, F_2, F_3, F_4, and F_6, represent the average homozygosity (analogous to allelic homozygosity in population genetics) of the codons with 2, 3, 4, and 6 synonymous codons, respectively (Wright, 1990). The more-biased the usage, the higher the homozygosity, and thus the lower the ENc. Calculated in this fashion, ENc values can range from 20 (highly biased codon usage) to 61 (neutral, unbiased codon usage) for any given genetic sequence. As a summary statistic, the measure is most-appropriate for genes of a length that gives a good probability that, assuming nonbiased

usage, each codon could be used once (e.g., genes with a length of 61 or longer, permitting all non-stop codons to be used once, assuming completely balanced codon usage of all codons) (Wright, 1990). For shorter genes, bias tends to be overestimated.

ENc can be plotted as a function of GC3 frequency, the frequency of third-position nucleotides in a genetic sequence that are either guanine or cytosine (Wright, 1990). An expected (bias-neutral) ENc value can be calculated for any given genic GC3 value, by the following formula (Wright, 1990). This formula produces a curve that is useful for visually comparing the codon bias of different genetic sequences.

$$ENc_e = 2 + GC3 + \frac{29}{GC3^2 + (1 - GC3)^2} \qquad (2)$$

Where experimental ENc (equation one) deviates from the expected value (equation two), it is suggested that selectional bias is present. ENc, as a measurement of selectional bias, is conserved among viral genes that are of shared origin or are tuned for expression in a specific host (van Hemert et al., 2007a).

ENc is not without limitations, in that it is calculated in a fashion independent of other genomic data. It does not compare any given gene's bias to that of those throughout the rest of the genome, sometimes limiting its comparative use (Lawrence et al., 2002). Thus, it is sometimes favorable to make use of other measures of codon bias that are derived in a different manner. Other measures of bias include the Codon Adaption Index (CAI), and the Adaptive Codon Enrichment (ACE).

CAI is derived from the Relative Synonymous Codon Usage (RSCU) for each amino acid. The RSCU is calculated for a set of highly-expressed genes, which is used as

a benchmark for bias (Comeron and Aguadé, 1998). Other genes' codon usage is compared to this, providing a CAI that represents how biased a gene is for expression by a specific organism. While more-directly tied to expression, and thus of greater functional correlation than ENc, CAI necessitates expression data, and is thus not well-suited for large-scale genomic analyses of organisms for which expression data is not readily available (Lawrence et al., 2002).

ACE circumvents many of the issues inherent to both ENc and CAI, as it provides a statistic that can be tuned to any specific genome without the requirement for expression data (Lawrence et al., 2002). Iterative algorithms can be used to provide a set of genes within a genome that are predicted to be highly expressed, and, since these algorithms can be applied to multiple genomes, ACE can be used to compare intergenomic data (Lawrence et al., 2002). This makes ACE attractive when working with diverse genomic data, as it can provide a means of comparing bias between different genomes in a more-robust manner than can ENc.

However, ENc remains a widely-used indicator of bias, and is well-suited for large-scale analyses of genomic data (Comeron and Aguadé, 1998). Unlike CAI, expression data is not required for its calculation, and it is more-readily calculated than ACE, which depends upon other analyses to predict the population of genes that are optimized for expression in a given genome (Lawrence et al., 2002). Given these advantages, and in light of the fact that mycobacteriophage genomes share much similarity, in part negating the advantage of ACE, we make use of ENc in this study.

GC3: A Genomic Feature of Significance

Studies of GC3, the frequency of codons with guanine or cytosine in the third nucleotide position for any genetic sequence, have suggested that it may be of functional significance. Due to the degeneracy of the genetic code, base substitutions at the third nucleotide position have a lower probability of resulting in missense mutations that would impact protein function. Therefore, alterations to the third base pair are commonly believed to present less of a barrier to a gene's adaptation to reflect overall trends in organismal codon usage (Palidwor et al., 2010). Thus, it has been suggested that biased usage at the third nucleotide position results predominantly as a direct result of codon optimization for organismal tRNA abundancy (Palidwor et al., 2010) (Ikemura, 1981).

Despite this, studies have demonstrated that GC3 exhibits non-random trends at both the genic and genomic level that may be independent of codon bias (Tatarinova et al., 2010). There exist many possible explanations concerning the functional purpose behind such trends in GC3 content, as discussed below.

In their 2010 study, Palidwor et al. developed a model of codon usage that describes a large proportion of GC3 variance for human, prokaryotic, and plant genes as a result of GC content alone (Palidwor et al., 2010). They show that in most cases, GC3 exhibits a linear relationship with overall GC content. They note the exception that for codons with G/C-alternatives in the first nucleotide position, Arginine and Leucine, this trend is not as significant (Palidwor et al., 2010).

In their 2014 study of over 35,000 prokaryotic genomes, Babbitt et al work to explain why degeneracy may have developed as it did, with most degeneracy involving the third nucleotide position (Babbitt et al., 2014). Degeneracy appears to be tied to GC

content in the third position (GC3), with 2- and 3-fold degenerate codons biased most commonly toward C. They suggest that a possible reason for this may be DNA stability, as such bias prefers a more stable phosphate linkage between positions two and three in the nucleotide. Higher GC3 content also increases the size of the minor groove, increasing DNA flexibility. In their study, GC3 was shown to correlate with changes in intrinsic DNA flexibility both in individual genes and across the entire genome, accounting for over 50% of the observed trends in flexibility. The unique nature of the third codon position leads them to suggest that the third nucleotide may have been an addition to an archaic two-nucleotide code that hypothetically existed. Thus, they propose that trends in GC3 usage are the result of a change in the genetic code that is buried deep in evolutionary history, and that optimization of DNA flexibility may be a reason behind observed trends in GC3 content.

In their 2013 study, Takuno et al observe a correlation between GC3 content and methylation patterns in the rice, thale cress, bee, and human genomes. Their analysis showed a strong negative correlation (r = -0.67, $P < 0.0001$) between GC3 and CpG methylation (Tatarinova et al., 2010). Further, they propose that there may be a transcriptional basis for genic trends in GC3 content, as they observe that GC3 frequently increases over the length of a many genes. Specifically, they suggest that increases in GC3 may increase transcriptional efficiency, allowing quick bursts of gene expression. They note that in *A. thaliana*, housekeeping genes have relatively-low GC3 content and are highly-methylated, which has previously been shown to decrease mutation rates (Takuno and Gaut, 2012). The reduction of mutation rates in essential, optimized genes is expected to be advantageous, potentially implicating GC3 in this evolutionary process.

Trends in GC3 usage in viral genomes have been studied to a lesser extent than those of their prokaryotic and eukaryotic counterparts, though a number of telling observations have been made. A study of *Astroviridae*, a genus of ssRNA viruses, suggests that host nucleotide composition (and codon bias) is the driving force behind viral codon bias (van Hemert et al., 2007b). A study of *Begomoviruses*, a genus of dicotyledon-tropic ssDNA viruses, indicated that their codon usage was likewise host-dependent, suggesting that this trend is widespread among viral groups (Xu et al., 2008).

A large-scale study of the four serotypes of dengue virus, a member of the ssRNA family *Flaviviridae*, implicated mutational bias and purifying selection as the driving forces behind viral codon selection (Lara-Ramirez et al., 2014). Specifically, it was noted that the third nucleotide position appeared to be under the greatest pressure, as it exhibited the strongest correlation with codon bias, hinting at possible functional significance (Lara-Ramirez et al., 2014).

In their 2008 study of a diverse set of bacteriophages that infect *E. coli*, *P. aeruginosa*, and *L. lactis*, Lucks et al showed that viral codon bias and GC3 may be linked to transcriptional efficiency or mRNA stability, and thus expression of viral genes, during infection (Lucks et al., 2008). They found that structural proteins showed the greatest similarity to host GC3 and codon usage, suggesting that genes that must be highly-expressed may be under pressure to reflect such characteristics. This conclusion is supported by a study that showed that GC3 correlates with a drastic increase in transcriptional efficiency in human cell culture (Plotkin and Kudla, 2011).

Aims of Investigation

It has been demonstrated that GC3 and codon bias are of functional significance. In this study, we explore the implications of both genomic features in Mycobacteriophages and *S. aureus*. We hypothesize that both GC3 and codon bias (as measured by ENc) are of functional significance in phages, and that they will therefore exhibit non-random trends in related phages. As part of exploring this hypothesis, we investigate the relationship of singleton *Mycobacteriophage Patience* and cluster H phages with a focus on GC3 and ENc. Further, we hypothesize that GC3 and ENc will influence gene expression in *S. aureus*.

CHAPTER TWO

Methods

Source of Sequences and Database Versions

Phage genome data used in this study were sourced from the Phamerator database, a phage-specific genomic database maintained by Steve Cresawn of James Madison University (Cresawn et al., 2011). For the phage genome-based portion of this study, the date of access for the database was May 19th, 2012. Genomic data were retrieved from the database in FastA format through the use of a Python script. Only coding sequences were used in this study. For the pham and gene-based segments of this study, data was retrieved on various dates during the fall of 2014 using the standalone Phamerator program (available freely at phagesdb.org) or through use of NCBI BLAST (Altschul et al., 1990).

Genomic data for *S. aureus* strain N315 were retrieved from EnsemblBacteria (EBI). The sequence used was submitted by Kuroda et al (Kuroda et al., 2001). RNA-Seq data were obtained from a previous study of blue-light sensitivity of *S. aureus* isolate BUSA2288. RNA was extracted, enriched, and purified from two sets of 24 1mL dark-grown (control) cultures that were pooled before submission to OSHSC for analysis on an Illumia MiSeq sequencer. RNA-Seq data were imported into GeneSifter (PerkinElmer), an online microarray and sequence analysis package for organization and analysis.

Calculation and Analysis of Codon Usage Statistics and Codon Bias

Open-source and purpose-developed tools (code available upon request) were used to analyze both *S. aureus* and phage genomic data. Codon usage frequencies were calculated at the genomic level by summing the codon usage in all annotated genes using self-developed Perl and C++ programs. Raw frequencies were compared without any form of weighting to compensate for amino acid abundancy. GC3, the number of codons with guanine or cytosine in the third nucleotide position, was calculated in the same manner for all called genes for all phages that were analyzed. Codon usage statistics for each gene, genome, and cluster of phages were analyzed. The purpose-developed codon analysis suite, Codoninator, was used to identify genes that differed from mean GC3 values by a given margin (typically set at plus or minus two standard deviations), facilitating identification of genes of interest.

For every phage or *S. aureus* gene, the Effective Number of Codons (ENc) was calculated through the use of CodonW (John Peden, maintained at http://sourceforge.net/projects/codonw), a program that conducts codon-based analysis of genetic sequences. There was a subset of *S. aureus* genes (16 of the total of 2951) for which the ENc could not be calculated, owing to the presence of an internal stop codon. This set of genes was considered irrelevant for the current study and excluded from analysis.

S. aureus transcripts were separated into expression brackets based on the order of magnitude of their transcript count (detection number). Since overall expression levels varied between the two replicates by nearly an order of magnitude for most genes, the

replicate with the greatest range of transcript levels was selected for analysis. This provided the best resolution when comparing transcript levels. The separation of data into brackets provided sets of 64, 576, 1180, 543, and 157 transcripts (10^5 to 10^1) for analysis. A chi-squared analysis of these expression brackets, comparing observed to expected ENc, as calculated by the below formula, was performed. Expected ENc, a value specific to each genic GC3 value, assumes completely balanced usage of synonymous codons, and provides a benchmark against which to measure observed codon bias (Wright, 1990).

$$ENc_e = 2 + GC3 + \frac{29}{GC3^2 + (1 - GC3)^2} \qquad (2)$$

GraphPad Prism was used to perform a correlation analysis of variance from expected ENc, GC3, and expression level. Correlations were assessed for significance by two-tailed t-tests with a cutoff of $P < 0.05$.

A dotplot was generated using the program Gepard (Jan Krumseik, maintained at http://www.helmholtz-muenchen.de/en/mips/services/analysis-tools/gepard/index.html).

Genome Landscape Generation and Description

Genome landscapes are a measure of the relative change in a specific measurement throughout an entire genome (Lucks et al., 2008). In this study, they served as a means of comparing genomes and assessing the relatedness of phages. Since GC3 genome landscapes are calculated with reference to a reading frame, only coding sequences are included as input. The same technique applies to the generation of GC1,

GC2, and overall GC landscapes, though for GC landscapes the calculation must be performed on the nucleotide, not codon, level, to account for GC at each position. The interspersed non-coding sequences, which in bacteriophages make up only a small portion of the genome, are ignored. GC3 Genome Landscapes are generated by assigning a binary value to each codon: 0 if its third nucleotide position is comprised of adenine or thymine (AT) and 1 if it is comprised of guanine or cytosine (GC). For each each codon, the binary value is added to a cumulative genome-wide sum up to that point, and the genome-wide average GC3 frequency is subtracted from the sum (Lucks et al., 2008).

$$cGC3 = \sum_{i=1}^{m} \left(GC3i - \overline{GC3} \right) \qquad (3)$$

This summary statistic generates a graph in which upward slopes are indicative of regions of relatively-high GC3, when compared to the genome-wide average, and in which downward slopes indicate lower-than-average GC3. An example of GC3 landscape generation can be found in Appendix A. In this study, the GC1/2/3 (collectively referred to as GCX) landscapes were generated through the use of a Perl script. Data in the form of a list of cGCX values (the raw value, representative of change in GCX for each location in the genome) were output for statistical processing. Cluster-wide cGCX averages were generated for each cluster, and a comparable control, comprised of sixteen randomly-selected phages from the HHMI database, was used as a benchmark for statistical comparison.

In this study, landscapes were normalized based on genome length in order to facilitate comparison of different viral genomes. All landscapes were treated as if they

were the same length, permitting comparison between landscapes of different lengths. Landscapes were produced by sampling the cGC3 level one-hundred times at regular intervals. Landscapes with more-frequent samplings were also explored, and were not found to be of improved comparative utility. Alternately, raw genome landscapes can be produced for comparison of specific features in genomes of different lengths.

Cluster average landscapes were produced by taking the mean value for each landscape of the hundred loci in all phages within a cluster. Similarity of landscapes within a cluster was determined by calculating the standard deviation for each landscape locus and then averaging all hundred samplings that produced the overall landscape to attain a value representative of the entire genome.

Analysis of S. aureus strain N315 Transcript Levels, GC3, and ENc

Transcription levels were linked to CodonW data (GC3, ENc, and other bias statistics) through the ID conversion utility included as a part of DAVID (NIAID, NIH), a bioinformatics package (Huang et al., 2007). The cross-referenced data were analyzed using a combination of Microsoft Excel and GraphPad Prism. Except where otherwise noted, Pearson's correlations were tested for statistical significance using a two-tailed t-test with a cutoff at $P < 0.05$.

CHAPTER THREE

Results

Comparative Bacteriophage Codon Usage Analysis

In order to compare overall codon usage between phages and clusters, codon usage frequencies were calculated for each phage. The frequencies were found to be similar among like-clustered phages, likely reflecting an increase in shared genomic content of such groups. With regard to singleton *Patience*, *Patience's* codon usage frequencies were compared with the average frequencies of each cluster. A mean percent difference (MPD) among codon frequencies was calculated by averaging the percent differences between *Patience* and the averages of the various clusters for each individual codon (Table 1). An example of calculating the mean percent difference between two codon usage tables can be found in Appendix B. The difference in codon usage between Patience and cluster H was 38.8%, whereas for the next-closest cluster, L, the difference was 46.1%. Other clusters differed even further than cluster L, with some, such as clusters B, C, I, K, O, P, and Q differing by more than 70%.

Table 1. *Mean Percent Difference of all Codon Usage Ratios between Patience and all Clusters*. This analysis shows that *Patience*'s codon usage was most similar to those of cluster H.

Cluster	MPD	Cluster	MPD
A	71.649	J	52.900
B	81.192	K	77.384
C	81.433	L	46.089
D	50.252	M	64.524
E	68.373	N	75.109
F	50.495	O	79.049
G	76.374	P	83.069
H	38.842	Q	79.565
I	79.497		

In order to assess the similarity of the ENc values of selected clusters, the overall variance ($\boldsymbol{\sigma^2}$) of genome-wide ENc values (calculated per gene) for all phages in each cluster was calculated (Table 2). An artificial control cluster consisting of a sample of sixteen randomly-selected phages was used to assess the significance of the calculated variance. Most clusters, such as D, G, L, and P, exhibited extensive similarity among their ENc values, as indicated by their having a variance among their ENc values that was separated from that of the control group by an entire order of magnitude or more. Some clusters, however, had a variance that was within an order of magnitude of that of the control, indicative of less similarity. Following this approach, it was determined that clusters A, B, I, and K were more diverse than the others. The variance among all of the ENc values of cluster H changed from 0.94 to 13.25 when *Patience* was included as a

member, a jump of two orders of magnitude to a value that was barely below that of the control, 14.67.

Table 2. *ENc Variance for all Clusters*. Clusters differed widely in the range of similarity in ENc values among their members.

Cluster	**ENc σ^2**	**Cluster**	**ENc σ^2**
A	3.84	K	1.85
B	4.24	L	0.14
C	0.22	M	0.58
D	0.01	N	0.29
E	0.20	O	0.18
F	0.25	P	0.10
G	0.08	Q	0.05
H	0.94	H + *Patience*	13.25
I	5.20	Control	14.67
J	0.29		

The average of all the observed ENc values (formula 1) was compared to the average of all the expected ENc values, as calculated from the genic GC3 frequencies (formula 2), in order to assess selectional bias within the clusters (Table 3). All clusters besides A, E, M, N, and Q had average ENc values that deviated from their expected ENc values by more than the control group did from its expected value. This deviation suggests the presence of some form of selectional bias in the codon usage of many phages.

Table 3. *Expected and Actual ENc Values for all Clusters.* Difference from expected ENc is an indicator of codon bias.

Cluster/Singleton	Average ENc	Expected ENc	Difference
A	38.92	38.22	0.70
B	35.59	36.88	-1.29†
C	34.78	36.54	-1.76†
D	44.66	42.05	2.60†
E	38.50	38.38	0.12
F	46.08	42.70	3.39†
G	36.67	37.86	-1.19†
H	48.36	46.93	1.43†
I	36.18	37.11	-0.93†
J	45.11	41.42	3.69†
K	36.46	37.23	-0.77†
L	46.00	44.58	1.42†
M	40.56	39.94	0.63
N	37.64	37.83	-0.19
O	35.99	36.92	-0.93†
P	35.92	37.00	-1.08†
Q	36.81	37.34	-0.54
Patience	55.45	60.73	5.28†
Control Group	39.99	39.29	0.70

† - Values that are of a greater magnitude than the control

A visual comparison of *Patience* and cluster H ENc and GC3 values highlights this difference in codon usage between the phages (Figure 1). Typically, *Patience's* genes are of a lower GC3 content and higher ENc, when compared to those of cluster H phages. This indicates that while *Patience* may be most similar to cluster H phages, it still differs from cluster H in terms of codon bias and GC3 content.

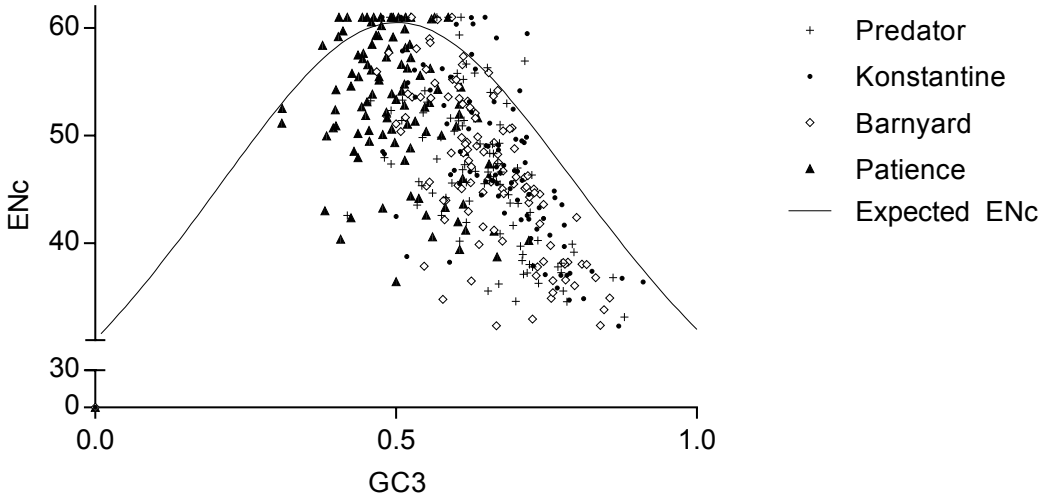

Figure 1. ENc v GC3 Plots for Cluster H Phages and Patience A comparison of the ENc vs. GC3 plots for all genes of *Patience* and cluster H phages highlights some of the differences between *Patience* and other cluster H phages.

Genome Landscapes as a Tool for the Study of Phage Genomics

While our primary interest was GC3, we additionally explored the use of GC1 and GC2 landscapes in visualizing phage genomes. Landscapes were calculated using the summary statistic described in formula 3 for GC1, GC2, and GC3. A comparison of the standard deviations among the different landscapes within a cluster was used to determine which landscape would be best-suited for comparative purposes (Table 4). There was no consistent pattern found for the GCX landscapes in terms of amount of variability as calculated by averaging variance at each landscape locus within a cluster. In most cases, such as in clusters C, E, and O, GC3 was more variable than GC1 or GC2, however occasionally (Cluster Q) GC1 or GC2 had similar variability. GC3 was selected as the primary landscape for study, since GC3 has been shown to be functionally-relevant

in previous studies (Zheng et al., 2014)(Tatarinova et al., 2010)(Babbitt et al., 2014)(Lara-Ramirez et al., 2014).

Table 4. *Standard Deviation of GC1/2/3 for Select Clusters*

	Cluster										
	C	D	E	F	G	J	M	N	O	P	Q
GC1	17.74	9.92	12.09	24.68	9.03	31.41	24.66	18.02	9.96	10.85	4.01
GC2	21.09	8.80	17.50	17.23	6.51	41.01	14.07	11.63	9.79	10.01	9.95
GC3	42.33	11.52	33.02	38.21	11.34	56.97	66.65	20.46	21.14	14.74	9.56

The GC3 genome landscapes of like-clustered phages exhibited similarity. Cluster F is one of the most diverse clusters and was thus chosen as a point of focus when investigating the utility of genome landscapes for clustering (Hatfull et al., 2008). A comparison of the graph of the average GC3 landscape of cluster F and those of F-cluster members Llij and Dorothy shows that, while individual phages differ to a degree from the average, as a whole their upward and downward trends generally reflect that of the cluster, indicating a similar distribution of GC3-rich regions throughout their genomes (Figure 2).

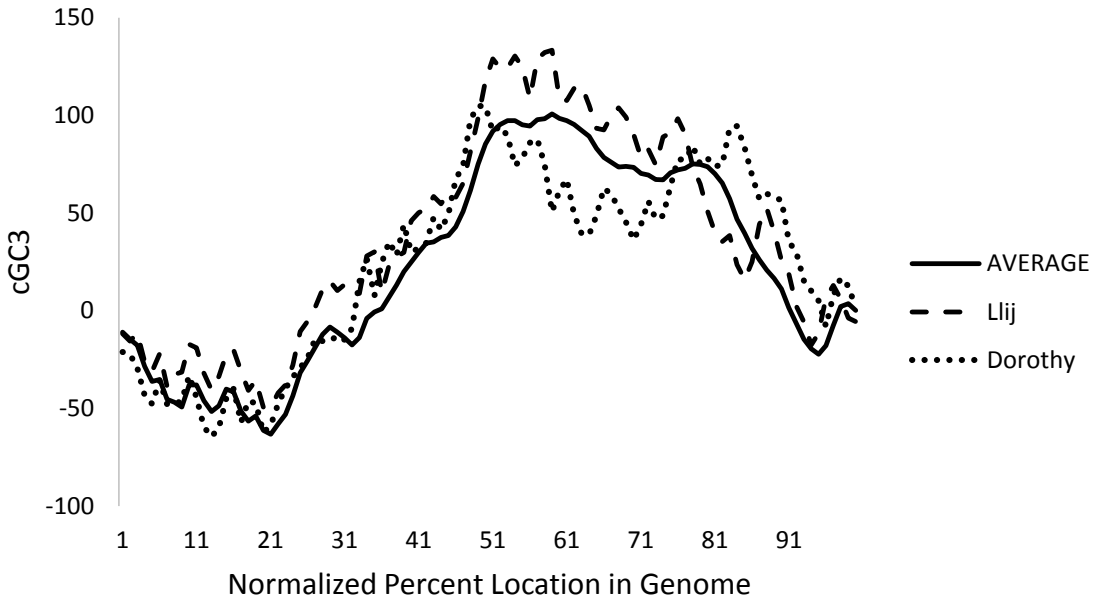

Figure 2. *Cluster F GC3 Genome Landscapes: Genomic Similarity* A comparison of cluster F's average GC3 landscape and those of two of its members, Llij and Dorothy, demonstrates that cluster average landscapes are representative of those of individual cluster members, preserving many of their defining characteristics. Note that when reading GC3 genome landscapes, upward slopes indicate genomic regions with higher-than-average GC3, whereas downward slopes indicate regions of lower-than-average GC3.

Some clusters have a greater degree of diversity than others, which has prompted the further division of their members into subclusters. When the cluster B subcluster GC3 landscapes are superimposed, both the diversity among and similarity between the different subclusters are apparent (Figure 3). While the graphs deviate from one another substantially, the slopes of the different subclusters are often correlated. For example, in the 60% to 85% range, all subclusters experience a steep upward slope, indicative of high genomic GC3 content, which is then followed by a steep downward slope towards the 90% range of the genome, indicative of lower-than-average GC3 content. In this and other cases, the relationships between subclusters become apparent upon close examination of their landscapes.

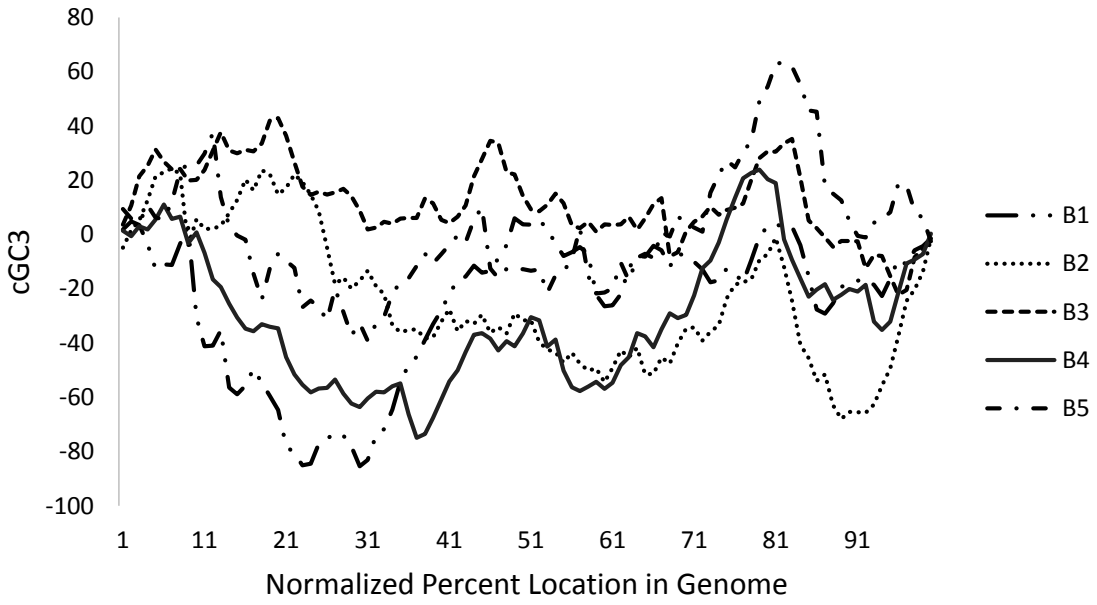

Figure 3. *Cluster B Genome Landscapes by Subcluster: Related Landscapes* A comparison of all subcluster GC3 Landscapes of cluster B demonstrates that subclusters share similar genomic features, despite their increased genetic distance.

Since the genome landscapes of like-clustered phages are similar, it follows that they can be used to evaluate relationships between phages that are not included in the same cluster, potentially highlighting genomic links that may not be suggested by a phage's cluster designation. One such case is the relationship between singleton *Patience* and cluster H. A graph of the GC3 landscapes of *Patience* and cluster H subcluster averages shows a high degree of similarity between the GC3 landscapes of *Patience* and those of cluster H (Figure 4). There are, however, regions in which trends in GC3 content are nearly the opposite of those of cluster H phages, such as in the region surrounding 40% of the distance through the genome.

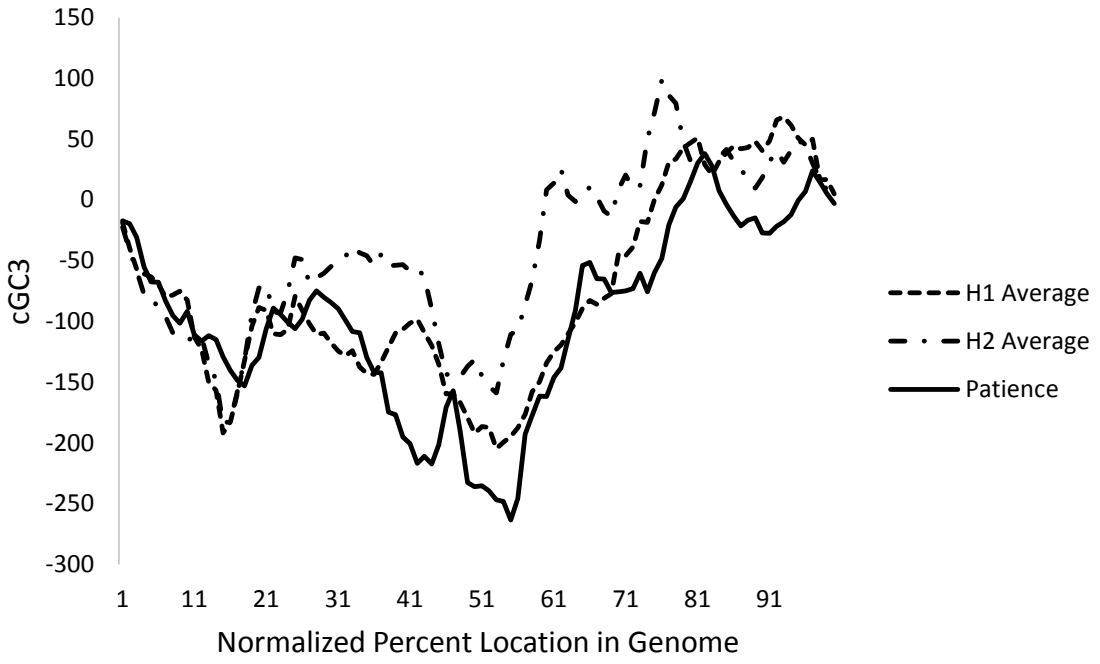

Figure 4. *Cluster H and Patience Genome Landscapes: Evidence for Relationships* The GC3 landscape of phage *Patience* is similar to those of cluster H throughout much of its genome, lending support to the notion that *Patience* may be related to cluster H phages.

This suggests that GC3 landscapes may be a useful tool in identifying genomic relationships that may not be obvious through the use of other techniques. A measurement of the standard deviation of each point along the normalized GC3 landscapes was taken for each cluster (Table 5). With *Patience* added to cluster H, the standard deviation within cluster H shifted from 34.754 to 37.470, a value that was still less than that exhibited by certain other clusters and subclusters. Comparison with a control cluster of sixteen randomly-selected phages validated this technique as a measurement of relatedness of genomes, as the control exhibited a standard deviation of 73.234, which was outside of the bounds of all established clusters and subclusters.

Table 5. *GC3 Landscape Standard Deviations*. Analysis of the GC3 landscapes showed that Patience's GC3 landscape was similar to those of the phages in cluster H.

Cluster and Subcluster GC3 Landscape Standard Deviations

	Full	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	+P
A	66.547†	28.247	29.900	14.765	11.372	33.645	18.565	19.065	14.226		24.337	
B	23.852	13.482	7.539	15.723	14.647							
C	42.326†	38.231†										
D	11.529											
E	33.022											
F	38.212†	33.187	21.797									
G	11.339											
H	34.754	27.236										37.470
I	37.159	12.448										
J	56.965†											
K	25.929	14.905	28.342	28.342	13.906	24.222						
L	25.986	9.360	22.513									
M	66.654†											
N	20.456											
O	21.142											
P	14.743											
Q	9.569											
Ctrl	73.234†											

† – value that is higher than that of cluster H with *Patience* included as a member
+P – with *Patience* included in the cluster

Comparative Analysis of Pham GC, GC1, GC2, and GC3

In order to ascertain whether GC3 and GC are conserved at the genic level, average pham (gene family) GC/GC3 was compared with the pham size (indicative of the degree of conservation of a pham among known phages) by means of Pearson correlations. A strong correlation (r = 0.414, P = 0.0134) between the number of pham members and GC3 was detected. Similarly, a strong correlation (r = 0.445, P = 0.0074) between GC and number of pham members was detected. A weaker correlation (r =

0.367, $P = 0.03$) was found between number of members and GC2, and none was found for GC1. Further, while a strong correlation between standard deviation of average pham GCX values and number of pham members was found for both GC1 and GC2 ($r = 0.468$, $P = 0.0046$ and $r = 0.552$, $P = 0.0006$, respectively), no such trend was observed for either GC or GC3.

Genome-Wide Pham Synteny and Dotplot Analysis

Further evidence of a relationship between *Patience* and cluster H phages comes from the application of traditional clustering techniques. Classic techniques, such as dotplot analysis of *Patience* and cluster H phages and an analysis of genome-wide pham synteny indicated a weak relationship between *Patience* and cluster H. A dotplot demonstrates a degree of base-pair similarity between *Patience* and cluster H phages (Figure 5). *Patience* has regions of its genome that share identity with those of both *Predator*, an H1 phage, and *Barnyard*, an H2 phage. Such regions of alignment, however, do not span at least 50% of the shorter phage's genome, as would be required to designate it a member of cluster H under the current clustering guidelines.

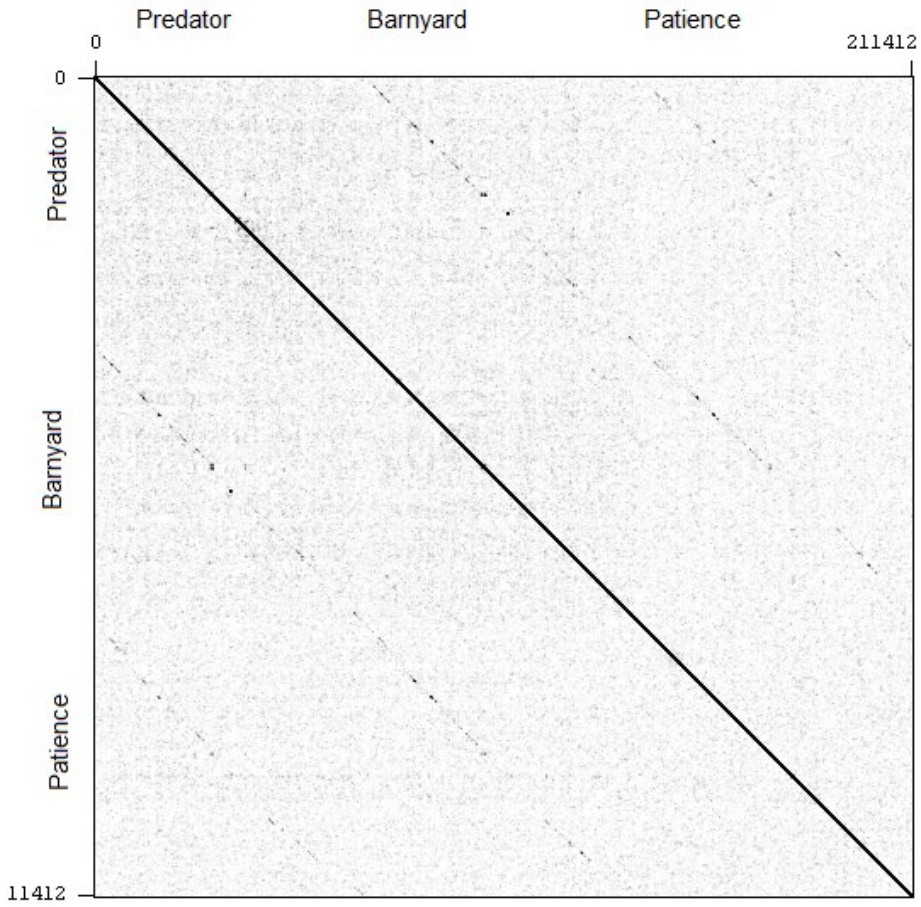

Figure 5. *Dotplot of Patience and cluster H Phages Predator and Barnyard* A dotplot analysis of *Patience* and cluster H phages *Predator* and *Barnyard* indicates a degree of similarity in between *Patience's* genome and those of H-cluster phages. Diagonal lines indicate genomic regions in which base pairs match between two genomes.

Additionally, a comparison of the pham maps of *Patience* and H-cluster phages showed that *Patience* shares 39 of its 110 phams with at least one cluster H phage (Figure 6). Perhaps more striking is that, of these 39 phams, 18 are unique to *Patience* and cluster H, appearing in no other clusters, a factor that supports the possibility of their having been in genetic communication, either directly or through an intermediary.

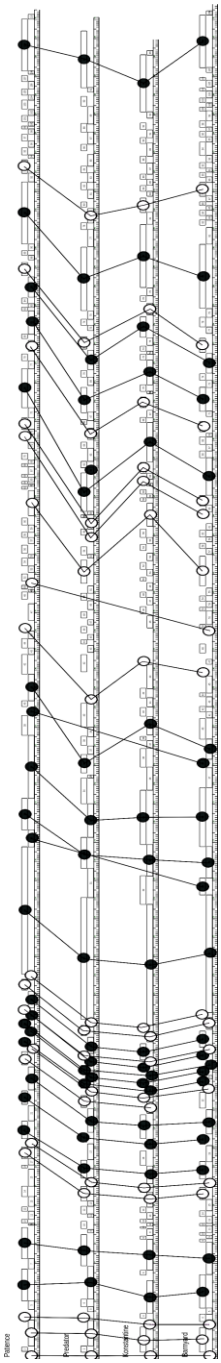

Figure 6. Annotated Pham Map of Patience and Cluster H This annotated pham map shows the synteny of shared phams between Patience and cluster H. Portrayed is a map of the full genomes of Patience and cluster H phages Predator, Konstantine, and Barnyard. Lines connect common phams, and circles indicate the exclusivity of the shared phams. Phams marked by a white circle are phams that are unique to Patience and cluster H. Phams marked by a black circle are those that are shared with at least one phage from another cluster.

S. aureus Gene Expression Compared to Genomic Features

Our investigation of *S. aureus* genomics centered on comparison between mRNA transcription levels (representative of gene expression), genomic measures of codon bias (ENc), and GC3. Pearson's correlations were calculated for each set of measures (GC3 and transcript count, ENc and transcript count, etc.) and significance was tested by a two-tailed t-test.

A weak, negative correlation between observed ENc and expression level was detected (r = -0.1551 P <0.0001). This supports the idea that codon bias, as measured by ENc, reflects optimization for gene expression in *S. aureus*. To further investigate this trend, the relationship between variance from expected ENc and transcription levels, a measure of selectional bias, and expression was assessed, and a weak positive correlation was observed (r = 0.1925 P < 0.0001).

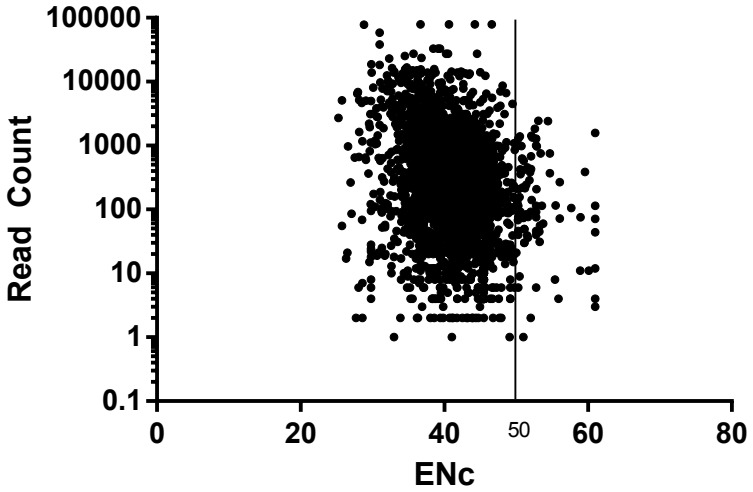

Figure 7. Comparison of read count (representative of transcript abudance) and ENc for all expressed *S. aureus* genes. The group of genes with a high ENc of 50 or more, representative of very unbiased codon usage, do not contain any genes that are expressed at as high of levels as those with lower ENc.

A chi-squared analysis of ENc values for *S. aureus* N315 transcripts was conducted (Table 6). Transcripts were separated into five categories based on expression level, by order of magnitude. Expected ENc was calculated based on transcript GC3. A significant ($P < 0.01$) deviation from the expected ENc was calculated for the most-highly expressed set of genes, indicating the presence of selectional bias in highly-expressed genes. No other groups deviated significantly from the expected ENc.

Table 6. *Summary of Chi-squared Analysis of ENc for Genes Grouped by Expression Level.* Only the transcripts that were detected at the highest level were found to differ significantly from the expected value. While not part of a general trend, individual genes in lower expression brackets did exhibit substantial variation from expected ENc.

Bracket	**n**	***P* < 0.01**
log5	64	yes
log4	576	-
log3	1180	-
log2	543	-
log1	157	-

A similar analysis of GC3 showed no correlation between GC3 and expression ($r = 0.02081$, $P = 0.2859$) (Figure 8). Most genes, regardless of expression level, fell within a GC3 range of roughly 0.1 to 0.3.

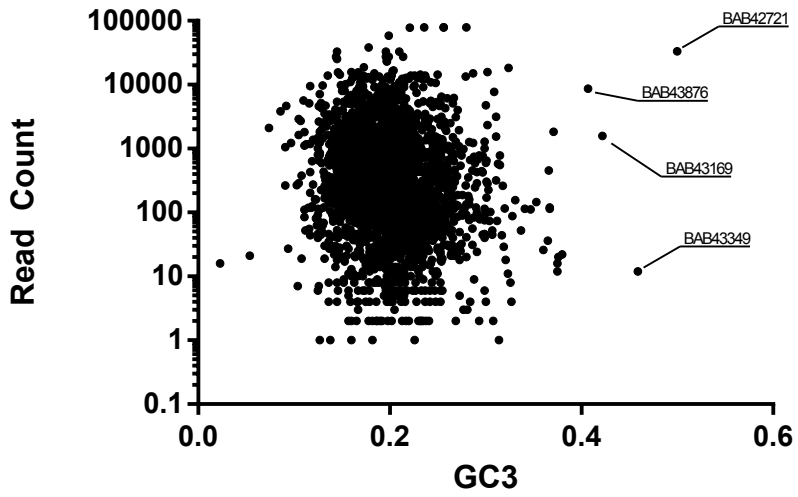

Figure 8. *Comparison of Read Count (Transcript Abudance) and GC3 for All Expressed S. aureus Genes.* No significant correlation was detected, though there are some outliers of potential interest.

CHAPTER FOUR

Discussion

Strengths and Weaknesses of Mycobacteriophage Clustering

The relationships between *Mycobacteriophages* are of a complex nature. Many phages exhibit extensive synteny and base-per-base nucleotide similarity, indicative of either shared evolutionary history or genetic exchange, but it is pertinent that other techniques for evaluating the relatedness of phages be explored. The cluster-based approach to categorizing the relationships between phages is limited in that it in some cases cannot account for the degree of mosaicism observed in different groups of phages that are related at some level, either by common ancestry or through horizontal gene transfer. We believe that the additional, codon-based methods outlined in this study can be used in conjunction with standard techniques to further enhance the investigation of inter-phage relationships, either as a component of, or separately from, the established clustering system. As an example, traditional techniques provide insufficient evidence to classify *Patience* as a cluster H phage, but, when taken along with evidence provided by analysis of codon usage and GC3 landscapes, they provide support for a stronger relationship between *Patience* and cluster H.

A 2014 study of Mycobacteriophage *Patience* concluded that it likely recently acquired via horizontal transfer the means of infecting mycobacterial species. The authors believe that *Patience* evolved in a low-GC environment before acquiring the required mechanisms for host expansion. They identify signs of recent, rapid adaptation –

including frameshifts to produce novel proteins, and codon bias in highly-expressed genes that reflects that of mycobacterial hosts (Pope et al., 2014). This rapid adaptation to a new host range could explain the conflicting nature of the evidence for a relationship between *Patience* and other mycobacteriophages, such as those of cluster H.

Genome Landscapes as a Tool to Identify Cryptic Relationship among Bacteriophages

In this study, we show that genome landscapes present a novel way of visually comparing phage genomes, and that they can provide information pertinent to the investigation of the relationships between phages. In order to facilitate the use of genome landscapes for clustering, it would be useful to develop statistical techniques which could better quantify the data presented by the landscapes, so that objective comparison could be implemented. The aforementioned method of using the deviation in cGC3 at various locations across the genome as a measure of relatedness is a starting point in this process, but a more sophisticated technique that perhaps takes into account the mosaicism that is observed in many phages would be preferable. If a specific, concrete guideline for comparison, comparable to those used in existing clustering methods (such as the 50% nucleotide similarity requirement in dotplot analysis), could be developed, genome landscapes could serve a much-more definitive purpose in genomic clustering. One promising approach would be the development of profile-Hidden Markov Models (profile-HMMs) that operate similarly to gene family predictors but instead utilize changes in GC3 in order to model families of related phages (Yoon, 2009).

One issue with the method developed here is that it relies upon normalized landscapes, and thus compares genomes of varied lengths with no direct consideration for their differences. Non-normalized clusters can easily be generated, but were not well-

suited for this study. The control cluster assay suggested that it is improbable that phages of different lengths, when normalized, will seem artificially similar, since trends in GC3 usage are unique phage-to-phage, but this would need to be addressed in any formalized comparative model.

As they stand, landscapes are most useful as an alternate method of presenting genomic data in order to aid in subjective decisions. As shown in the case of *Patience* and cluster H, genome landscapes can provide evidence for relationships between phages that may otherwise not be apparent. By providing a novel means of visualizing relationships between phages, genome landscapes may help indicate the existence of relationships between genomes that have on the nucleotide level been separated by a significant degree of divergence.

Additionally, genome landscapes and codon-bias analysis may provide useful information in the exploration of the relationships between related clusters. For example, Phamerator indicates a likely relationship between clusters I and P over large segments of their genomes – analysis of genome landscapes and codon usage and bias within such related regions could prove to be informative in the larger investigation of bacteriophage genomics.

Interpreting ENc Values as Indicators of Relatedness

Similarity between the ENc values of phages generally reflects and affirms traditionally determined cluster relationships. However, an analysis of ENc did not indicate a relationship between *Patience* and cluster H, as its ENc value deviates greatly from those of other H-cluster phages. This is likely related to the fact that *Patience* has very low GC content, which results in an average GC3 frequency of 0.498, which is

much lower than that of all other phages referenced in this study. Most clusters have average GC3 frequencies that range from 0.75 to 0.85. Cluster H phages also have low GC3 frequencies, with an average of 0.655, a value that, while low, is not nearly as low as that of *Patience.*

When looking at ENc as an indicator of codon bias by checking deviation from expected ENc, it was shown that *Patience* exhibited a great degree of such deviation, indicating that biased codon usage was indeed present throughout its genome. Cluster H, however, deviated by a lesser degree than *Patience*, possibly suggesting divergent selectional pressures, again placing ENc-based analysis in contrast with the evidence provided by the codon ratios and genome landscapes that suggest a strong relationship between *Patience* and cluster H. This contradiction further underscores the complexity of the relationships between phages.

Of interest, it was observed that clusters with less variance in ENc (Table 2) among all their genes generally had a greater level of bias, as indicated by deviation from expected ENc (Table 3). Initial analysis points to a possible negative correlation (r = -0.2616, P = 0.2481), though more-robust analysis would be required to confirm this relationship. Such a relationship, if verified, could implicate maintenance of codon bias among phages, as measured by ENc, as a driving force in their evolution.

Evidence for Pham-Wide Conservation of GC3 and GC

An analysis of GCX and GC statistics for phams that are shared by *Patience* with at least one other phage (n = 48) supports the hypothesis that GC3 content is under selective pressure. Strong correlations between the degree by which a pham is shared and both GC (r = 0.445, P = 0.0074) and GC3 (r = 0.414, P = 0.0134) were found, indicating

that GC3 is higher among more widely-conserved phams, suggesting a functional significance. A weaker correlation was found for GC2, and none was found for GC1, suggesting less of a functional role for these two nucleotide positions. GC1 and GC2 were found to vary more among phams with more members, where no such relationship was found for GC or GC3. This lack of an increase in variation with pham size suggests a non-random distribution of GC and GC3 within phams, again suggesting possible functional significance. While this relationship merits further investigation, it should be noted that all phage genomes are not sequenced, so data concerning the degree by which phams are shared within the current sample could be an inaccurate representation of the greater population of Mycobacteriophages. Sampling bias, either resulting from the geographic distribution of collection sites or from the implemented isolation technique, could potentially skew the representation towards a specific subpopulation of phages.

Functional Implications of GC3 and ENc in S. aureus

Our analysis did not reveal any relationship between transcript abundancy and genic GC3 in *S. aureus*. However, the hypothesis that GC3 may impact translational efficiency remains untested, and in our opinion merits investigation. A relationship was, however, detected between codon bias (as indicated by deviation from expected ENc) and transcript abundancy. Most genes within the highly-expressed bracket are housekeeping genes associated with transcription, translation, metabolism, and maintenance of homeostasis. Other highly-expressed genes included Protein A (an IgG-binding virulence factor), a cation antiporter, and cold-shock protein A. All of these are genes of significance to normal biological function of *S. aureus*, which follows their showing a significant deviation from their expected ENc.

We looked for correlations between GC3/ENc and transcript abundance with no regard for differential expression among different classes of genes. Housekeeping genes, virulence genes, prophage genes, and other classes of genes would all be expected to be expressed at different levels, regardless of GC3 and ENc. Further, different classes of genes would be separately up or down-regulated depending upon growth conditions. It could be informative to test for correlations between GC3/ENc and transcript abundance within such classes of genes, so as to reduce the probability that any impact on expression is masked by differences in gene expression as a result of gene function.

As previously mentioned, when comparing GC3 with transcript abudancy, no overt trend was discovered. However, there exist some outliers, such as BAB42721, a high-GC3, highly-expressed (GC3 of 0.5, count of 33470), short (33 residues), gene of no known function. BLAST results indicate similarity to short sequences found in other *S. aureus* genomes. Another outlier, BAB43876, has a high GC3 (0.407), is short (55 residues) and is translated at medium-high levels (count of 8702). It shows a high deviation (variance of 1.831) from expected ENc, indicating possible selectional pressure. As noted, however, ENc can be skewed for short (less than 61-residue) peptides, a factor that could explain this deviation. It displays wide homology among *Staphylococcus* species, and some homologous proteins are suggested to be toxins (S. aureus Pepa1 Chain A) (Sayed et al., 2012). Another outlier is BAB43169, which also has a high GC3 (0.422), yet shows completely-balanced codon usage (ENc = 61). It is transcribed at medium levels (count of 1577) and is identified as a ligase. Another outlier with high GC3 (0.459), high codon bias (ENc variance of 4.68), yet low expression (transcript count of 12) is BAB43349, a hypothetical protein of no known function.

While not providing evidence for a general role for GC3 in *S. aureus*, such outliers may indicate certain cases in which GC3 does in some way influence expression.

Shortcomings and Further Steps

While this study provides some evidence for significance of GC3 within Mycobacteriophages, primarily through analysis of conservation of patterns in GC3 among phages, it fails to suggest why this may be the case. It would be of benefit to conduct a controlled analysis of the influence of GC3 upon transcription and translation in phages. An experiment utilizing tools to create recombinant phages with GC3 variants (high, low, and control) of a known gene would provide a means of assessing the influence of GC3 on transcription and translation. Bacteriophage Recombineering of Electroporated DNA (BRED), a recombineering technique, has been used to insert EGFP into D29 phage, and a similar experiment with GC3-high and GC3-low EGFP variants would provide a simple means of assaying such effects, if any (Marinelli et al., 2008) (da Silva et al., 2013).

The section of the study concerning *S. aureus* provided preliminary evidence that there is no link between GC3 and translation in the bacterium. More replicates, as well as techniques to assay any impact on translation (such as MS) would further the study. A recombineering experiment similar to the one proposed above would also be beneficial, in that it would allow for a controlled means of assaying the impact of GC3 on expression in isolation.

APPENDICES

APPENDIX A

Genome Landscape Example

Genome landscapes are visual representations of trends in a measurement, such as GC3, throughout a genome. The following example will demonstrate the generation of a GC3 genome landscape for a short coding sequence. For reference, formula 3 is included below.

$$cGC3 = \sum_{i=1}^{m} \left(GC3i - \overline{GC3} \right) \qquad (3)$$

Given the sequence AAT.GGC.GTG.GGC.GCG.ATA.GTA.GAA, an algorithm to generate GC3 landscapes will first calculate the overall GC3 frequency of the sequence. In this case, four of the eight codons are GC3, so the mean GC3 is 0.5.

The assignment of binary values for each codon will then proceed. A GC3 codon is assigned a 1, and an AT3 codon is assigned a 0.

Table A.1 *Example Codon Binary Assignments*

Position	0	1	2	3	4	5	6	7
Codon	AAT	GGC	GTG	GGC	GCG	ATA	GTA	GAA
Value	0	1	1	1	1	0	0	0

The cGC3 (landscape value) is then calculated for each position along the landscape by equation 3.

Table A.2 *Example Landscape Generation*

Position	**Value**	**GC3i – $\overline{GC3}$**	**cGC3**
0	0	(0 + 0) – 0.5	-0.5
1	1	(-0.5 + 1) – 0.5	0.0
2	1	(0 + 1) – 0.5	0.5
3	1	(0.5 + 1) – 0.5	1.0
4	1	(1 + 1) – 0.5	1.5
5	0	(1.5 + 0) – 0.5	1.0
6	0	(1.0 + 0) – 0.5	0.5
7	0	(0.5 + 0) – 0.5	0.0

The above example generates the following landscape.

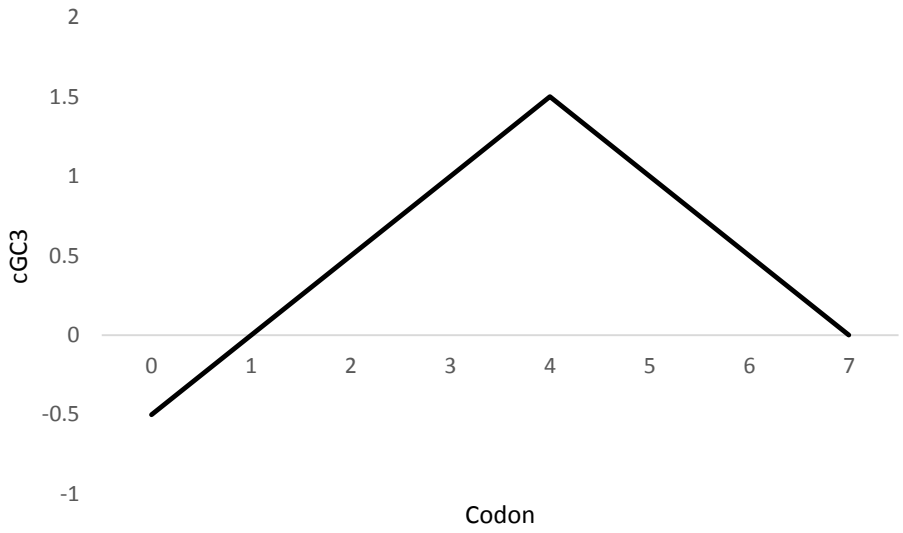

Figure A.1 *Example GC3 Genome Landscape*

APPENDIX B

Mean Percent Difference Example

The mean percent difference in codon frequencies was used to compare codon usage among phages. The following is a simplified example of the calculation, detailing its use in a hypothetical four-codon genetic system.

Table B.1 *Example Genome Codon Frequencies*

	Genome 1				**Genome 2**			
Codon	A	B	C	D	A	B	C	D
Frequency	20%	15%	55%	10%	15%	25%	40%	20%

First, the percent difference would be calculated for each pair of codons, per equation four, below.

$$\%Diff = \frac{|x-y|}{\left(\frac{x+y}{2}\right)} \qquad (4)$$

For the example data set, this provides the following result.

Table B.2 Example *Percent Difference Calculation*

Percent Difference				
Codon	A	B	C	D
Freq. 1	20.0%	15.0%	55.0%	10.0%
Freq. 2	15.0%	25.0%	40.0%	20.0%
% Diff.	28.6%	50.0%	31.6%	66.7%

Next, the mean percent difference is calculated by taking the arithmetic mean of all calculated percent differences. In this example, the result is 44.23%.

BIBLIOGRAPHY

Alibayov, B., Baba-Moussa, L., Sina, H., Zdeňková, K., and Demnerová, K. (2014). Staphylococcus aureus mobile genetic elements. Mol. Biol. Rep. *41*, 5005–5018.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Babbitt, G.A., Alawad, M.A., Schulze, K.V., and Hudson, A.O. (2014). Synonymous codon bias and functional constraint on GC3-related DNA backbone dynamics in the prokaryotic nucleoid. Nucleic Acids Res. *42*, 10915–10926.

Behura, S.K., and Severson, D.W. (2013). Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. Biol. Rev. Camb. Philos. Soc. *88*, 49–61.

Bishal, A.K., Saha, S., and Sau, K. (2012). Synonymous codon usage in forty staphylococcal phages identifies the factors controlling codon usage variation and the phages suitable for phage therapy. Bioinformation *8*, 1187–1194.

CDC, Antibiotic Resistance Threats (2013). http://www.cdc.gov/drugresistance/pdf/ar-threats-2013-508.pdf

Comeron, J.M., and Aguadé, M. (1998). An Evaluation of Measures of Synonymous Codon Usage Bias. J. Mol. Evol. *47*, 268–274.

Cresawn, S.G., Bogel, M., Day, N., Jacobs-Sera, D., Hendrix, R.W., and Hatfull, G.F. (2011). Phamerator: a bioinformatic tool for comparative bacteriophage genomics. BMC Bioinformatics *12*, 395.

Hatfull, G.F., Cresawn, S.G., and Hendrix, R.W. (2008). Comparative genomics of the mycobacteriophages: insights into bacteriophage evolution. Res. Microbiol. *159*, 332–339.

Hatfull, G.F., Jacobs-Sera, D., Lawrence, J.G., Pope, W.H., Russell, D.A., Ko, C.-C., Weber, R.J., Patel, M.C., Germane, K.L., Edgar, R.H., et al. (2010). Comparative genomic analysis of sixty mycobacteriophage genomes: Genome clustering, gene acquisition and gene size. J. Mol. Biol. *397*, 119–143.

Van Hemert, F.J., Berkhout, B., and Lukashov, V.V. (2007a). Host-related nucleotide composition and codon usage as driving forces in the recent evolution of the Astroviridae. Virology *361*, 447–454.

Van Hemert, F.J., Lukashov, V.V., and Berkhout, B. (2007b). Different rates of (non-)synonymous mutations in astrovirus genes; correlation with gene function. Virol. J. *4*, 25.

Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C., et al. (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. *35*, W169–175.

Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. *151*, 389–409.

Kuroda, M., Ohta, T., Uchiyama, I., Baba, T., Yuzawa, H., Kobayashi, I., Cui, L., Oguchi, A., Aoki, K., Nagai, Y., et al. (2001). Whole genome sequencing of meticillin-resistant Staphylococcus aureus. Lancet *357*, 1225–1240.

Lara-Ramirez, Rez, E.E., Salazar, M.I., L&#xf3, pez-L&#xf3, Pez, M., Jes&#xfa, A.D., s, Salas-Benito, J.S., S&#xe1, et al. (2014). Large-Scale Genomic Analysis of Codon Usage in Dengue Virus and Evaluation of Its Phylogenetic Dependence. BioMed Res. Int. *2014*, e851425.

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

Lindsay, J.A. (2014). Staphylococcus aureus genomics and the impact of horizontal gene transfer. Int. J. Med. Microbiol. IJMM *304*, 103–109.

Lucks, J.B., Nelson, D.R., Kudla, G.R., and Plotkin, J.B. (2008). Genome landscapes and bacteriophage codon usage. PLoS Comput. Biol. *4*, e1000001.

Marinelli, L.J., Piuri, M., Swigonová, Z., Balachandran, A., Oldfield, L.M., van Kessel, J.C., and Hatfull, G.F. (2008). BRED: a simple and powerful tool for constructing mutant and recombinant bacteriophage genomes. PloS One *3*, e3957.

McCarthy, A.J., Loeffler, A., Witney, A.A., Gould, K.A., Lloyd, D.H., and Lindsay, J.A. (2014). Extensive horizontal gene transfer during Staphylococcus aureus co-colonization in vivo. Genome Biol. Evol. *6*, 2697–2708.

Palidwor, G.A., Perkins, T.J., and Xia, X. (2010). A general model of codon bias due to GC mutational bias. PloS One *5*, e13431.

Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N.R., et al. (2003). Origins of highly mosaic mycobacteriophage genomes. Cell *113*, 171–182.

Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. *12*, 32–42.

Pope, W.H., Jacobs-Sera, D., Russell, D.A., Peebles, C.L., Al-Atrache, Z., Alcoser, T.A., Alexander, L.M., Alfano, M.B., Alford, S.T., Amy, N.E., et al. (2011). Expanding the Diversity of Mycobacteriophages: Insights into Genome Architecture and Evolution. PLoS ONE *6*.

Pope, W.H., Jacobs-Sera, D., Russell, D.A., Rubin, D.H.F., Kajee, A., Msibi, Z.N.P., Larsen, M.H., Jacobs, W.R., Lawrence, J.G., Hendrix, R.W., et al. (2014). Genomics and Proteomics of Mycobacteriophage Patience, an Accidental Tourist in the Mycobacterium Neighborhood. mBio *5*.

Rose, T., Verbeken, G., Vos, D.D., Merabishvili, M., Vaneechoutte, M., Lavigne, R., Jennes, S., Zizi, M., and Pirnay, J.-P. (2014). Experimental phage therapy of burn wound infection: difficult first steps. Int. J. Burns Trauma *4*, 66–73.

Sayed, N., Nonin-Lecomte, S., Réty, S., and Felden, B. (2012). Functional and structural insights of a Staphylococcus aureus apoptotic-like membrane peptide from a toxin-antitoxin module. J. Biol. Chem. *287*, 43454–43463.

Da Silva, J.L., Piuri, M., Broussard, G., Marinelli, L.J., Bastos, G.M., Hirata, R.D.C., Hatfull, G.F., and Hirata, M.H. (2013). Application of BRED technology to construct recombinant D29 reporter phage expressing EGFP. FEMS Microbiol. Lett. *344*, 166–172.

Sung, J.M.-L., Lloyd, D.H., and Lindsay, J.A. (2008). Staphylococcus aureus host specificity: comparative genomics of human versus animal isolates by multi-strain microarray. Microbiol. Read. Engl. *154*, 1949–1959.

Takuno, S., and Gaut, B.S. (2012). Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. Mol. Biol. Evol. *29*, 219–227.

Tatarinova, T.V., Alexandrov, N.N., Bouck, J.B., and Feldmann, K.A. (2010). GC3 biology in corn, rice, sorghum and other grasses. BMC Genomics *11*, 308.

Thiel, K. (2004). Old dogma, new tricks--21st Century phage therapy. Nat. Biotechnol. *22*, 31–36.

Wright, F. (1990). The Effective Number of Codons Used in a Gene. Gene *87*, 23–29.

Xu, X., Liu, Q., Fan, L., Cui, X., and Zhou, X. (2008). Analysis of synonymous codon usage and evolution of begomoviruses. J. Zhejiang Univ. Sci. B *9*, 667–674.

Yoon, B.-J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. Curr. Genomics *10*, 402–415.

Zheng, Q., Jiao, N., Zhang, R., Wei, J., and Zhang, F. (2014). The evolutionary divergence of psbA gene in Synechococcus and their myoviruses in the East China Sea. PloS One *9*, e86644.