A Thesis
Presented to the
Faculty of
San Diego State University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Biology
with a Concentration in
Molecular Biology

by
Han Suh Kang
Spring 2016
The Undersigned Faculty Committee Approves the

Thesis of Han Suh Kang:

Comprehensive Analysis of Curated Prophage Genomes from PhiSpy for
Assessment of Phage Genome Mosaicism and tRNA Dependencies

Robert A. Edwards, Chair
Department of Biology

Anca Segall
Department of Biology

Barbara A. Bailey
Department of Mathematics and Statistics

4/6/16
Approval Date
ABSTRACT OF THE THESIS

Comprehensive Analysis of Curated Prophage Genomes from PhiSpy for Assessment of Phage Genome Mosaicism and tRNA Dependencies

by

Han Suh Kang

Master of Science in Biology with a Concentration in Molecular Biology
San Diego State University, 2016

Temperate phage genomes are highly variable, mosaic, collections of genes required to enter a bacterial host, integrate into the host's genome, identify when to switch from the lytic to lysogenic cycles and replicate their DNA, package into new virions and escape. Genomes from most Bacterial phyla contain phage genomes integrated as prophages, poised to excise from the genome and release new phage particles. We updated our PhiSpy algorithm to improve detection of prophages, provide a web-based framework for PhiSpy, and used this algorithm to identify 36,488 prophage regions from 11,941 bacterial genomes. Transfer RNA genes were abundant in the prophages, many of which alleviate host codon bias limitations. However, tRNA genes also enable phages to increase translation initiation efficiency and overcome limitations of the hosts translation machinery. We identified integrase genes in 15,765 prophages (43% of the prophages), and used the integrase to orient and align prophage genomes. These alignments demonstrated a highly-conserved phage genome profile that can be used to identify the function of unknown phage genes.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vi</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>MATERIALS AND METHODS</td>
<td>3</td>
</tr>
<tr>
<td>- Phispy Version 3.2 Improvements</td>
<td>3</td>
</tr>
<tr>
<td>- Dataset Generation</td>
<td>3</td>
</tr>
<tr>
<td>- Raw Sequence Phage-Host Correlation Studies</td>
<td>4</td>
</tr>
<tr>
<td>- Genome Heatmaps</td>
<td>4</td>
</tr>
<tr>
<td>- Integrase Phylogeny</td>
<td>4</td>
</tr>
<tr>
<td>- Computational Materials</td>
<td>5</td>
</tr>
<tr>
<td>RESULTS AND DISCUSSION</td>
<td>6</td>
</tr>
<tr>
<td>- Prophage Genometrics</td>
<td>6</td>
</tr>
<tr>
<td>- Phage Integrase</td>
<td>7</td>
</tr>
<tr>
<td>- tRNA Presence in Prophages Compared to tRNA Integration Occurrences</td>
<td>8</td>
</tr>
<tr>
<td>- Genome Architecture</td>
<td>9</td>
</tr>
<tr>
<td>CONCLUSIONS</td>
<td>11</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>22</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1. Frequency of different genes among the prophages. ORFs were indentified from 36, 46 prophages. ................................................................................................11
LIST OF FIGURES

PAGE

Figure 1. Methodology summary flowchart. .................................................................12

Figure 2. Bacterial genomes and the number of prophage regions depicted in blue. ..........12

Figure 3A. Box and whisker plot denoting number of prophages identified per host, grouped by six phyla with most prophages. .................................................................13

Figure 3B. Length (bp) of prophages identified grouped by six phyla with most abundant phyla. ..............................................................................................................14

Figure 4. GC content of prophage shows strong correlation to GC content of host ($R^2 = 0.94735$), but no correlation of prophage GC content with prophage length (upper panel). ..........................................................................................................15

Figure 5A. Residue frequency of the eight-mer motif contained in the catalytic domain of HP1-like integrases. ...................................................................................16

Figure 5B. Randomly selected integrases shows clear clusterings of integrase containing the eight-mer motif (highlighted). .................................................................17

Figure 6. tRNAs identified within prophage region. .......................................................18

Figure 7A. Conserved prophage genome organization...............................................19

Figure 7B. Phage gene heatmap aligned using a model Clustal alignment algorithm using +0 mismatch, +1 match scoring rules shows a clear order in which genes occur ..............................................................................................................20
INTRODUCTION

Phages are major drives in global biogeochemical cycles through the control they exert on bacterial populations. Despite their abundance in the environment they are under-represented in sequence databases and knowledge-bases. The two life cycles, the lytic and lysogenic cycles, provide a unique and different window into phage genomics. During the lysogenic cycle, temperate phages integrate their genomes into their host’s genome in a non-lethal fashion – either as a plasmid (1) or by directly recombining with the host’s genome to generate a region of DNA known as a prophage (2). The phage replicates in the host genome until external signals promote a transition to the lytic cycle, at which point the phage hijacks the host’s cellular DNA replication machinery in order to replicate the phage genome, and hijacks the host’s protein synthesis machinery to create new phage particles. Eventually the phage lyses the host’s cell membrane to release infective particles to find new hosts to attack.

The activation and movement of a prophage from one host to another enables the phage to act as mediators of horizontal gene transfer, via either generalized or specialized transduction (3). This phage mediated horizontal gene transfer may account for up to $10^{15}$ transduction events per second (4, 5). Phage mediated horizontal gene transfer events are responsible for the spread of novel functions including the development of pathogenicity in bacterial strains such as *Escherichia coli* O157:H7 (4), *Vibrio cholerae* El Tor N16961 (6), *Streptococcus pyogenes* (7) and many *Staphylococcus aureus* strains (8). Phages may mediate the spread of antibiotic resistance in the mammalian intestine (9), and also transfer more benign genes such as those involved in photosynthesis (10).

We recently proposed a new model for phage host interactions that suggests at higher host densities ecological considerations favor the temperate lifestyle. The increased energy available at high host densities allows bacteria to support the replication of prophages, which in turn provide beneficial protection to the host via super-infection exclusion. The phages are
using their host for replication, but are also providing protection for their host from other invading phages (11).

Identification of prophages in bacterial genomes is complicated by the problem that many phage genes are not homologous to other proteins. In some cases a newly discovered phage does not carry a single gene with any sequence similarity to any other gene in the available databases, highlighting how little we know about these viruses (12). The underrepresentation of well-annotated phage genes in available databases makes it difficult to assign functions to newly sequenced genomes. This is exacerbated as genome sequencing increases at a near-exponential rate requiring us to rely on automated annotation software that is dependent on primary sequence homology to assign functions to new genes. Most prophage identification algorithms rely on sequence homology to known phage genes to identify prophage regions (13-17). This leads to a paradoxical problem of not being able to identify a prophage that does not contain any regions homologous to known phage proteins. Our prophage-finding program, PhiSpy, uses a novel algorithm that is capable of identifying a prophage even if there is no homology to any of the proteins in the databases by analyzing a suite of genomic characteristics that separate phage-encoding regions from the bacterial core of the genome (18).

Here we present an analysis of over 11,000 bacterial genomes from which we identified 36,488 prophages. We used this data to identify a core phage genome organizational profile and demonstrate that the majority of phage genomes adhere to this model. This genome structure allows us to predict the function of genes with no known homology solely based on their location in the genome. We identified the integrase genes from 15,162 prophages, and aligned those genes to identify phylogenetic relationships between prophages, and we explored the connection between phage and host genomic characteristics to enable predictions of the bacteria that phage are infecting.
MATERIALS AND METHODS

PHISPY VERSION 3.2 IMPROVEMENTS

Several improvements were made to the PhiSpy program, and have released a new version of the software and a web site where people can use the program. The most important change we made was to allow PhiSpy to process genomes that were composed of multiple contigs, as the original version of PhiSpy would treat proteins from different contigs as adjoining proteins. The original PhiSpy used a scanning window that looked 40 proteins ahead of the current protein being classified, but to remove the inherent direction bias, we changed the window to be 20 proteins on either side of the protein being classified. For proteins near the edges of contigs, which had less than 20 proteins on a side, we included all possible proteins. To update the list of prophage protein annotation keywords, all the phage genomes from PhAnToMe, and all the complete prokaryotic genomes from GenBank were downloaded. We compared the annotations of each set, and found 12 keywords unique to phage protein annotations: endolysin, terminase, baseplate, base plate, virion, antirepressor, excisionase, Cro-like repressor, CI-like repressor, rIIIA lysis, rI lysis, and rIIB lysis. These keywords were incorporated into PhiSpy for protein classification. In addition to these major changes, many other minor bug fixes were applied. A website was also built so that users can run PhiSpy online, without having to download and install the program and it’s dependencies: http://edwards.sdsu.edu/PhiSpy/. The source code for the updated version of PhiSpy is available from that website and from GitHub: https://github.com/linsalrob/EdwardsLab/tree/master/PhiSpy.

DATASET GENERATION

All bacterial genomes were extracted from the SEED (19,20) database which contains both sequence and annotation information that is used by PhiSpy in making prophage designations. The bacterial genomes contained a representative of every genus currently known. The updated version of PhiSpy (version 3.2) was run on all bacterial genomes using
the generic bacterial training set option, and the outputs were parsed to identify all predicted prophage regions. All PhiSpy predictions are available from https://edwards.sdsu.edu/PhiSpy and https://figshare.com/articles/PhiSpy_prophage_predictions/3146656

**RAW SEQUENCE PHAGE-HOST CORRELATION STUDIES**

Prophage coordinates from PhiSpy were used to extract nucleotide sequence data from the original bacterial genome. The nucleotide sequences are available as fasta files from https://edwards.sdsu.edu/PhiSpy and https://figshare.com/articles/PhiSpy_prophage_predictions/3146656. Putative tRNA-encoding genes were identified from those genomes using tRNAscan-SE (21). Modal codon usage was calculated from both the predicted prophage and the host genomes as described previously (22), and tRNA use and modal codon use were compared using the $\chi^2$-test. G+C content of the phage and host nucleotide sequences was calculated using custom written Python scripts.

**GENOME HEATMAPS**

High-quality prophage genome predictions were defined as those predictions where the annotations contain at least one of each of an integrase, a capsid, a portal, a terminase, and one or more tail proteins. The annotated genes in all the prophages were tabulated to identify a suite of genes that occur frequently in phage genome annotations (integrase, portal, tail, holin, transposase, terminase, endonuclease, capsid, antitermination, holliday junction resolvase, baseplate assembly protein, primosomal protein, protease-like, lysin, toxic,-encoding lysozyme).

To generate the genome alignment, each gene was assigned a unique letter indicating its function or a letter indicating that the function was a minority function and/or was unknown. A modified Clustal Omega alignment algorithm that scores +1 for matches and 0 for mismatches was used to align the annotated genomes.

**INTEGRASE PHYLOGENY**

A total of 20,443 integrase gene sequences were extracted from the high-quality manually curated dataset and clustered using CD-HIT (23) at a 90% similarity threshold (an approximate virus “species” demarcation (13, 15) to produce 8,536 protein clusters. A representative of each cluster was used in the phylogenetic assessment of the integrase genes.
In addition, short motifs were identified in the integrase genes by using $k$-mer profiles for $k$ from 4 amino acids to 7 amino acids long.

Short polypeptides allowed us to identify shared amino acid motifs including the eight residue motif at the active site of the tyrosine recombinases. One hundred randomly selected integrases were aligned using Clustal Omega (24,25) and visualized using figTree (26).

**Computational Materials**

Computational work was conducted on the Sun Grid Engine. Standard python libraries were used for file parsing and biopython version 1.65 was used to parse GenBank files. Scripts written for file parsing and data analysis is available in the github repository: https://github.com/hkang408/ProphageGenomics and from https://github.com/EdwardsLab
RESULTS AND DISCUSSION

Prophage Genometrics

At the time of analysis, the SEED contained 11,941 bacterial genomes spanning 34 different phyla from Archeae and Bacteria (Figure 2). A total of 36,465 prophages were identified in 9,883 bacterial hosts (82.76% of the bacteria have at least one prophage). Just under one half of the prophages that were identified (15,765 prophages) contained an annotated integrase protein, suggesting they are integrative phages, while the mechanism of integration of the remaining phages could not be easily identified. A total of 1,511,644 prophage ORFs were identified – of these, 45.7% or 798,235 ORFs had no known function, highlighting how much there is still to learn about phages in general. A total of 7,341 portal genes, 6,228 holin genes, 14,771 capsid genes, and 59,010 tail-related genes were identified in the prophage dataset (Table 1).

As we have shown previously, some families of genes are easier to identify than others, and there remain gaps in our understanding of families of core phage proteins that we cannot readily identify just from sequence homology alone(27).

Across the most represented phyla, there were an average of between 2.7 (Actinobacteria) and 3.8 (Proteobacteria) prophages found per genome. Within phyla there were also differences (although not statistically significant) in the phage genome characteristics. For example, Sodalis glossinidius, a proteobacteria had the most prophages at 29, while several phyla Caldiserica, Deinococcus–Thermus, Dictyoglomi, Elusimicrobia, and Fibrobacteres had no prophages, Spirochaetes had the shortest prophages on average (25,595 bp) and Bacteroidetes had the longest prophages on average (46,059 bp) (Figure 3A, 3B). Across all of the prophages, the mean and median prophage genome length was 36,804bp and 32,352bp, respectively. Shorter prophage genomes may indicate that the phage is no longer viable and is in the recombination-driven process of being removed from the genome.
(28). Alternatively, different genome lengths may represent different packaging strategies used by prophages associated with different phyla (29).

Several prophage genome characteristics correlate with the host’s characteristics (30). For example, the prophage’s and host's GC content strongly correlated ($R^2 = 0.94735$; Figure 4). As we have shown previously, the GC content of phages can be accurately used to predict the phyla of the host associated with a phage (to >95% accuracy), however GC content can not be accurately used to predict the phages’ host at lower taxonomic levels. Even though %GC content correlates at the phyla level, there are not enough degrees of freedom in %GC calculations to separate thousands of phages and their hosts. Percent GC is basically a $k$-mer DNA profile with $k = 1$. Higher $k$-mer DNA profiles increase the correlation between phage and host, but when $k$ exceeds 7 nucleotides the data becomes sparse and overfits the predictions (31).

**PHAGE INTEGRASE**

Phage integrases are responsible for catalyzing the site-specific recombination event that inserts the phages into the hosts’ genome (32-36). A total of 15,765 prophages (43%) contained an integrase-like protein from one of the two main families – the tyrosine and serine recombinases – named after the catalytic residue involved in recombination. The former are similar to the integrase employed by phage $\lambda$ to insert into the chromosome, while the latter are exemplified by *Streptomyces* phage C31 (37) and *Mycobacteria* phage Bxb1 (38). Aside from the catalytic residue in which they’re named after, there are no conserved domains found across all members of either class of integrase.

The 20,443 integrase sequences were clustered at a 90% sequence identity threshold using CD-HIT which resulted in 8,536 clusters. Representatives from each cluster were used to generate a smaller dataset in which to work with. Five-mer profiles of the amino acid sequences were used to identify an eight-residue motif in the catalytic region of tyrosine recombinases with a consensus sequence of HDLRHTFA (Figure 5A). The 5 strongly conserved amino acids suggest that they are either directly or indirectly responsible for the catalytic action of the tyrosine active site. Alignment of one hundred randomly selected integrases from the condensed integrase set shows that the integrases that contained the motif
were grouped together (Figure 5B), offering a potential route in identifying tyrosine recombinases without having to identify the catalytic residue involved in recombination.

**tRNA Presence in Prophages Compared to tRNA Integration Occurrences**

Many phages carry tRNA genes in their genomes that enhance phage fecundity (39,40). Several hypotheses have been proposed to explain the presence of tRNA genes in phage genomes, including providing tRNAs to compensate for missing or lowly-expressed genes in the host (41). A previous study of 37 phage genomes from 15 bacterial genomes demonstrated that phage tRNA genes compensate for codon bias between host and phage, especially when a codon is common in the phage and rare in the host (42). In addition, tRNA genes carried by the phage may be used to replace those interrupted during phage insertion into the genome at tRNA genes (43,44). It is also possible that carrying multiple tRNA genes in a phage genome provides alternative attachment sites for the phage to integrate into the host's genome in case the preferred attachment site has mutated.

A total of 17,539 tRNA genes and 3,833 tRNA genes were identified from the phages in our total and condensed dataset respectively. As expected the tRNA gene profiles in the phages reflects the number of synonymous codons for each tRNA. However, we found that tRNA-Met is one of the most commonly carried tRNA genes among the prophages we studied. We have recently shown that phages often carry the peptide deformylase that acts to deformylate newly synthesized proteins (45). We propose that the methionine tRNA genes encoded by the prophages are included to increase the overall rate of translation when the phage enters the lytic cycle. Together, the peptide deformylase and tRNA-Met suggest that phage protein synthesis saturates the host's protein synthesis machinery and that optimal phage replication requires protein translation in excess of the maximum that the host can provide.

We explored the hypothesis that the tRNA genes compensate for codon bias between the host and the phage. For a tRNA gene found in a prophage, the anticodon was converted to the genomic DNA codon it’s associated with. This codon’s modal usage was calculated for the prophage’s genome as well as for the host’s genome without the prophage sequence. The codon’s modal usage in the phage and host was compared using a chi-square test. We found
that 41.3% of the codons associated with the tRNA found in a prophage had statistically higher modal usage in the phage than in the host.

We then applied the same approach to compare the phage’s modal usage of the same codon with an expected usage assuming an even distribution given the number of synonymous codons available. This comparison revealed that 31% of the codon associated with the tRNA found in a prophage had a higher usage in the phage than is expected assuming an even usage rate given the number of synonymous codons possible.

**GENOME ARCHITECTURE**

Phages have a mosaic genome architecture(46,47), in which the genome is organized into groups or clusters of genes. Mosaicism has been proposed to be driven by illegitimate recombination, but it is not clear from previous genomic analyses whether the mosaic boundaries are between clusters of related genes or between individual genes (28,45,48).

We identified the eight most commonly occurring and well annotated genes within the prophage dataset. The majority (66.9%; 10,542/15,765) of these predicted prophages had the integrase gene located near either end of the genome (since the prophage could be on either strand of the genome sequence). These were normalized by reverse-complementing half of the genomes to place the integrase gene consistently at the left end of every genome. The remainder was normalized by recircularizing the genome in silico, and introducing a break that placed the integrase at the left end of the genome. Thus, we generated a genome data set where all the genomes are organized with the integrase at the same end of the genome, allowing us to compare across genome architecture. A conserved gene order was shown across the prophage genomes (Figure 7A), suggesting that there are clusters of related functions. To highlight those clusters, we used a modified Smith-Watermann approach, essentially ascribing a single letter to each of the functions and then aligning the genomes based on the order of the functions with a +1 score for a match and no penalty for a mismatch (Figure 7B). This shows that the gene order across 1,537 phage genomes is highly conserved, reminiscent of the early—middle—late gene organization of many well characterized phages (49-52). Based on this alignment we propose that the mosaicism in phage genomes is driven at the gene-cluster level, and that phage genomes are not comprised of random hybrids of many genomes, but that the recombination (illegitimate or otherwise) is constrained to result
in a highly conserved gene order where the early genes including entry and DNA replication precede DNA packaging, which in turn precede tail formation and host lysis. The phage genome is therefore not a collection of genes that have been acquired at random by recombination, but analogous to a computer motherboard, where pieces and parts may be replaced provided they are slotted into the appropriate gene expression framework, and a viable phage will only be produced if genes are placed in the appropriate modules.
CONCLUSIONS

Here we present an analysis of over 11,000 bacterial genomes from which we identified 36,488 prophages. We demonstrate that the prophages across the entire Bacterial and Archaeal kingdoms, and are optimized to replicate and escape from their hosts. Many phages are limited by the initiation of translation by the host's machinery, and attempt to increase translation rates by carrying their own tRNA genes, increasing the availability of tRNAs loaded with methionine presumably for translation initiation, and increasing the availability of peptide deformylase. We have also demonstrated a highly conserved gene order across prophages, that suggests phage genome mosacism is limited to clusters of conserved genes.

Table 1. Frequency of different genes among the prophages. ORFs were indentified from 36, 46 prophages.

<table>
<thead>
<tr>
<th>Function</th>
<th>Number of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrase</td>
<td>20,514</td>
</tr>
<tr>
<td>Portal</td>
<td>7,341</td>
</tr>
<tr>
<td>Capsid protein</td>
<td>14,771</td>
</tr>
<tr>
<td>Tail protein</td>
<td>59,010</td>
</tr>
<tr>
<td>Holin</td>
<td>6,228</td>
</tr>
<tr>
<td>Cytotoxic repressor</td>
<td>132</td>
</tr>
<tr>
<td>Endonuclease</td>
<td>4,043</td>
</tr>
<tr>
<td>Terminase</td>
<td>15,657</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>737,992</td>
</tr>
<tr>
<td>Other</td>
<td>645,956</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1,511,644</strong></td>
</tr>
</tbody>
</table>
Figure 1. Methodology summary flowchart.

Figure 2. Bacterial genomes and the number of prophage regions depicted in blue.
Figure 3A. Box and whisker plot denoting number of prophages identified per host, grouped by six phyla with most prophages. Centerline denotes median while box edges denotes the first and third quartile.
Figure 3B. Length (bp) of prophages identified grouped by six phyla with most abundant phyla. Center line denotes median while box edges denotes the first and third quartile.
Figure 4. GC content of prophage shows strong correlation to GC content of host ($R^2 = 0.94735$), but no correlation of prophage GC content with prophage length (upper panel).
Figure 5A. Residue frequency of the eight-mer motif contained in the catalytic domain of HP1-like integrases. The bars represent the frequency with which each amino acid was found at each of the eight positions in the motif. The consensus sequence is HDLRHTFA.
Figure 5B. Randomly selected integrases shows clear clusterings of integrase containing the eight-mer motif (highlighted).
Figure 6. tRNAs identified within prophage region. Colors represent different synonymous codons.
Figure 7A. Conserved prophage genome organization. Proteins in the phage genomes are colored by their annotation (integrase, portal, tail, holin, transposase, terminase, endonuclease, capsid, antitermination, holliday junction resolvase, baseplate assembly protein, primosomal, protease-like, lysin, toxic ??, lysozyme), and the phage genomes were normalized to the largest prophage.
Figure 7B. Phage gene heatmap aligned using a model Clustal alignment algorithm using +0 mismatch, +1 match scoring rules shows a clear order in which genes occur.
REFERENCES


