UNDERSTANDING VIRUS-HOST INTERACTIONS THROUGH SINGLE CELL AND WHOLE GENOME ANALYSIS

A Dissertation
Presented to
The Academic Faculty

by

Shengyun Peng

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biological Sciences

Georgia Institute of Technology
December 2018

UNDERSTANDING VIRUS-HOST INTERACTIONS THROUGH SINGLE CELL AND WHOLE GENOME ANALYSIS

Approved by:

Dr. Joshua S. Weitz, Advisor
School of Biological Sciences & Physics
Georgia Institute of Technology

Dr. Ling Liu
School of Computer Science
Georgia Institute of Technology

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. Justin R. Meyer
Division of Biological Sciences
University of California San Diego

Dr. Frank J. Stewart
School of Biological Sciences
Georgia Institute of Technology

Date Approved: November 5, 2018

To my family and friends

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest thanks and sincere appreciation to my advisor Prof Joshua S. Weitz for his guidance and support during my Ph.D. journey from both the academic side and everyday life. From him, I learned by taking small steps how to think, analyze, solve research problems, as well as how to share and communicate with others as a scientist.

I would like to express my gratitude towards all of my committee members: Prof I. King Jordan, Prof Ling Liu, Prof Justin R. Meyer and Prof Frank J. Stewart for the extensive support and invaluable advice.

I am grateful to all my collaborators including Jacob H. Munson-McGee, Animesh Gupta, Prof Mark J. Young, Prof Rachel J. Whitaker, Samantha Dewerff and Dr. Ramunas Stepanauskas. I truly enjoyed working with such great collaborators. It is not possible to deliver all these results without their dedicated work.

I am also grateful to my friends and colleagues from the Weitz Group for their support and helpful discussions: Dr. Stephen Beckett, Daniel Muratore, Dr. David Demory, Dr. Chung Yin (Joey) Leung, Yu-hui Lin, Guanlin Li, Ashley Coenen, Rogelio Rodriguez, Qi An, Dr. Keith Paarporn, Dr. Charles Wigington, Dr. Luis Jover, Dr. Bradford Taylor, Dr. Ceyhun Eksin, Dr. Hayriye Gulbudak, Rong Jin, Dr. Hend Alrasheed, Walker Gussler and Devika Singh. I want to thank my friends and colleagues for making my graduate life exciting and colorful.

Last but not the least, I would like to thank my family Yihua Peng, Xinhua Han and Lu Wang for their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ANI	Average Nucleotide Identity
ANOVA	Analysis of Variance
ARD	Arms Race Dynamic
BPC	Base Pair Coverage
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
EFF	Efficiency of Phage Infection
EOP	Efficiency of Plaquing
FSD	Fluctuating Selection Dynamic
H:MF	Host-only Mutational Feature
Joint:MF	Combined Feature
MAE	Mean Absolute Error
MDA	Multiple Displacement Amplification
P:MF	Phage-only Mutational Feature
P+H:MF	Phage and Host Mutational Feature
P×H:MF	Phage-cross-host Mutational Feature
PBIN	Phage-bacterial Interactions Network
PCR	Polymerase Chain Reaction
POA	Presence or Absence of Successful Infection
SAG	Single-cell Amplified Genomes
SCG	Single-cell Genomics
SNP	Single Nucleotide Polymorphism
VMR	Virus-to-microbial Cell Ratio
YNP	Yellowstone National Park

SUMMARY

Viruses and their microbial hosts are widely distributed in the environment, including in oceans, soils, fresh water, and even in extreme environments such as the deep ocean, hot springs and the upper atmosphere. Given the ubiquity of viruses of microbes, it is critical to understand virus-host interactions and their effects on ecosystem functioning. My work addresses the problem of virus-host interactions through three motivating questions: 1) to what extent do viruses and hosts interact in a given environment and who interacts with whom, 2) how do interactions shape the coevolutionary dynamics of viruses and hosts and 3) what is the genetic basis for determining both who infects whom and the efficiency of viral infections. Here, I report findings stemming from analysis of virus-host interactions in a natural environment (Yellowstone National Park hot springs) and from an experimental study of coevolution *in vitro*. First, I characterized virus-host interactions in a hot spring's environment, combining evidence from single-amplified genomes and metagenomes to characterize a natural virus-host interaction network, finding that the majority of cells were infected by one (or more) viruses. Second, I developed a new approach to infer the genetic basis for both qualitative and quantitative changes in virus-host interactions unfolding during coevolution. In doing so, I leveraged whole genome analysis to identify novel mutational candidates that could drive large-scale changes in infectivity; the approach can also be applied to characterize the genotype-phenotype map in other phage-host systems. Overall, the findings help deepen our understanding of virus-host interactions and the consequences of infection on complex virus and microbe communities.

CHAPTER 1. INTRODUCTION

1.1 Virus and host

Viruses can infect organisms in different domains from the tree of life, including Eukaryota, Bacteria and Archaea [1]. Since the discovery of the first virus – tobacco mosaic virus – in the 1890's, many different types of viruses have been discovered [2, 3]. These include viruses that infect plants, *i.e.* plant virus [4], viruses that infect bacteria, *i.e.* bacteriophage (or phage) [5], and viruses that infect archaea, *i.e.* archaea virus [6]. Viruses and their hosts can be found across different environments on our planet, such as the ocean, soils, fresh water and also even extreme conditions such as the deep ocean, hot springs and the upper atmosphere [7-15]. Among all hosts of viruses, the microbes – mainly bacteria, archaea and fungi – are the most abundant host types [16, 17]. Many studies have estimated the virus-to-microbial cell ratio (VMR) in different environments and showed that the viruses outnumber their hosts by orders of magnitude. For example, in the ocean, the estimated VMR is about 10:1 [18-20]. A recent study has found a nonlinear, power-law relation better describes the VMR [21].

Given the widespread abundance and distribution of viruses and their microbial hosts, the interactions between the two are also commonly observed in different environments and could have a profound ecological impact [22-24]. In fact, the initial discovery of bacteriophage in 1915 was based on the observed outcomes of phage-host interaction by Frederick Twort and Feilx D'Herelle [25-28]. Viruses mainly interact with hosts through infection. As a result, phage may be able to regulate the population size and density of their hosts. The host distributions, in turn, also determine the phage production and distribution [29]. Recent studies in oceans and lakes have shown that phages and their hosts could impact climate change through the release of biogenic particles and dimethyl sulfide as a result of viral lysis [15, 30].

To systematically evaluate the phage-host interactions, many characteristics of the interactions such as burst size, latent period and lysis-lysogeny decision have been measured and investigated [31-35]. One important life history trait of viruses is the host range, which measures the variety of host cells that a virus can infect. Previous studies have shown that some viruses are generalists, that is they can infect a wide variety of host species, while others are specialists that only infect a few host strains [36-40].

1.2 Virus life cycle

Since viruses do not have their own metabolism system, they depend on their host cells to reproduce. Therefore, each step of viral replication involves interactions with host cells. For a virus to infect a host cell, it first attaches to the cell surface and injects its genome into the host cell [41]. After this step, the virus mainly interacts with its host through two different pathways: the lytic pathway or the lysogenic pathway [42-44]. For viruses that activate the lytic pathway, the virus chromosome integrates into the host genome. Virus genes are turned on and off to actively produce the viral DNA, head and tail proteins, and other components required for viral replication. New virus particles are assembled inside the host cell and eventually released to the environment after lysing of the host. For the lysogenic pathway, most virus genes are turned off after integration. The virus chromosome is passively replicated with the host multiplication. In this case, the host cell will not be 'killed' and the virus in lysogeny mode is described as 'temperate'. The host cell with the virus chromosome integrated into its own genome is called a lysogen and the integrated virus is called a prophage. Studies have shown that the temperate phage can switch from lysogeny to lysis mode when the environment changes, such as introduction of irradiation from UV light [42].

The decision between a lytic cycle or a lysogenic cycle has been extensively studied using bacteriophage λ and its host *E. coli* as the model system [42, 43, 45, 46]. To attach to the host cell, λ phage binds to the cell surface with its *J* protein in the tail fiber. The *J* protein interacts with the *LamB* porin and the phage DNA is injected to the cytoplasm. Afterwards, the lysis-lysogeny decision for λ phage is mainly determined by one factor – the density of a protein that is called λ repressor, which is encoded by the *cI* gene. When its density is high, the phage will go into the lysogenic pathway and when its density is low, the phage will go into the lytic pathway. When UV light is introduced to a lysogen, the host protein *RecA* is activated and cleave the λ repressors under the threat of DNA damage. As a consequence, the density of λ repressor is reduced and the prophage switches from lysogeny mode to lysis mode (Figure 1).

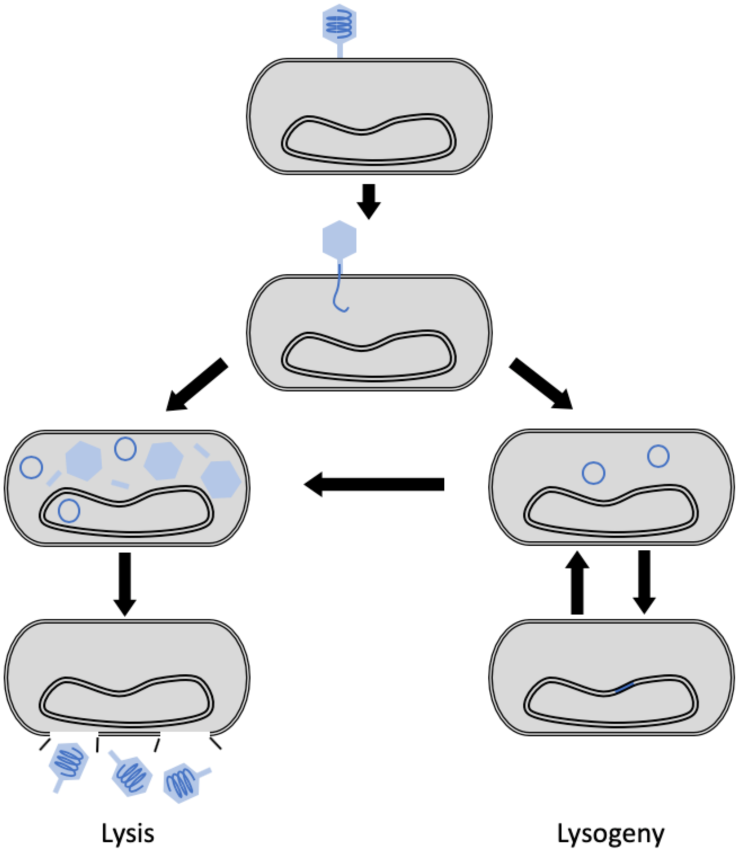

Figure 1 – Life cycle of bacteriophage λ

The λ phage may go either lytic or lysogenic pathway after entering the host cell. The lysogen can be inducted with environmental factor change, such as UV light, and switch to lytic mode. Adapted and remade from Ptashne, M. 2004

1.3 Host defense mechanisms

In response to virus infection, hosts have developed different systems to resist viruses. Extracellular defense mechanisms of the host resist the viral infection through changes in outer membrane receptors caused by genetic mutations. Additionally, adaptive immunity of the host includes various types of mechanisms, including clustered regularly interspaced short palindromic repeats–CRISPR-associated proteins (CRISPR-Cas) [47-54], BREX [55, 56], DISARM [57] and so on. For example, the CRISPR-Cas system is an adaptive immune system of bacteria and archaea. This system is estimated to exist in about

40% of bacterial and 90% of archaeal genomes [48]. The system contains two parts: the CRISPR sequences mainly serve as a biological database for identifying foreign DNA while the *cas* sequences encode proteins that degrade the foreign DNA. There are two major classes for the CRISPR-Cas system, namely class 1 and class 2. They differ by the *cas* genes and the molecular mechanism which generates the CRISPR RNA (crRNA) and cleaves the foreign DNA. The CRISPR sequences comprise three parts: 1. Leader sequence, 2. Repeat sequences and 3. Spacer sequences (Figure 2). The leader sequence, which is located upstream of the CRISPR, is AT-rich and conservative. The repeat sequences are the identical contents to separate the spacer sequences and the length ranges from 23 – 47 *bp*. The spacer sequences, which are captured from phage or plasmid nucleic acid, are the main identifier to recognize the foreign DNA. The length ranges from 21 – 72 *bp*. Each different spacer sequence targets a specific foreign DNA fragment which allows a host to have adaptive immunity to multiple different phages. The common number of repeat-spacer units is less than 50.

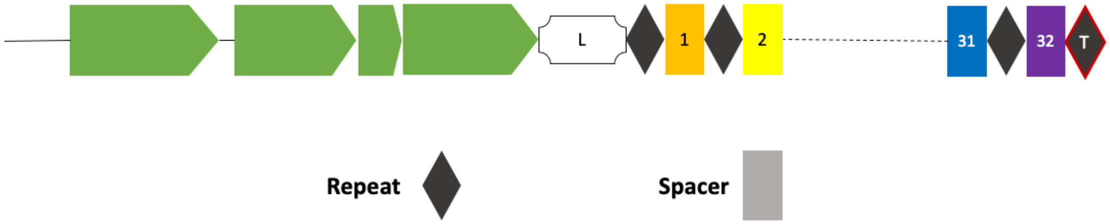

Figure 2 – Schematic of CRISPR-Cas system

The green block indicates the cas genes in the CRISPR locus. L stands for the leader sequences and its typical length is 20 – 534 bp. The black diamonds represent the repeat sequences. The typical length of these invariant repeat sequences is 23 – 47 bp. The last black diamond with red outline indicates the end of CRISPR locus. The colored rectangle shows the spacer sequences. The spacer sequences are highly variable and are originally captured from the foreign DNA. The typical length of spacer sequences is 21 – 72 bp. There can be as many as 375 repeat-spacer units in one CRISPR locus.

1.4 Virus-host interactions in natural environments

In natural environments, virus-host interactions form bipartite interaction networks. In the bipartite network, the viruses and hosts form two disjoint and independent sets of vertices. The edges connect the vertices from one set with the other, rather than within each set. In this case, edges indicate interactions between viruses and hosts. Such networks have different patterns, including modular and nested patterns (Figure 3) [37, 58, 59]. In the modular networks, the edges that connects viruses and hosts tend to occur among distinct groups. In contrast, in the nested networks, the edges that connects viruses and hosts typically forms a hierarchical structure. These interactions have a profound ecological impact [20, 29]. Therefore, it is fundamental to quantitatively characterize virus-microbe interaction networks and understand their impact on nutrient cycles, energy transformation, and ecosystem dynamics. This ‘who infects whom’ question remains one of the fundamental but open questions in studying virus-host interactions.

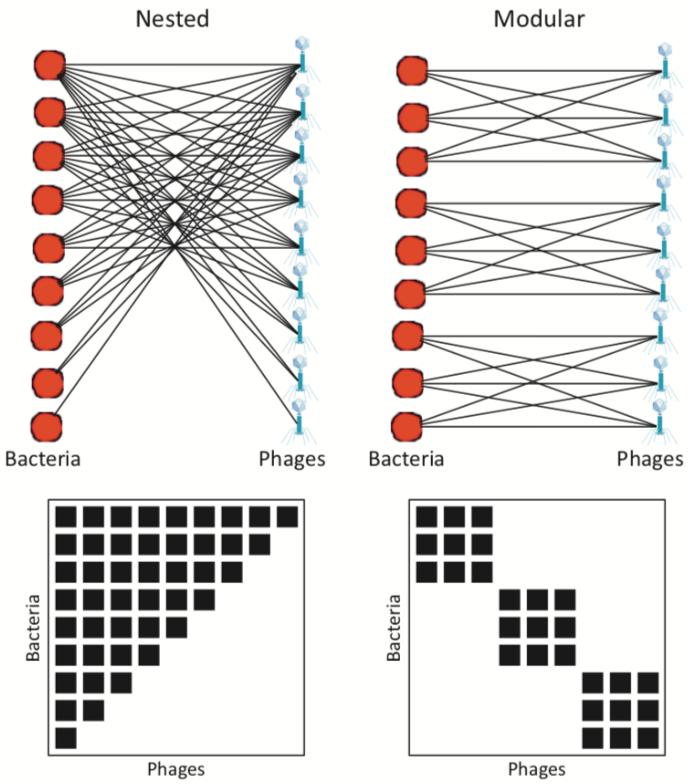

Figure 3 – Patterns of nested and modular bipartite virus-host interaction networks

Schematic showing the nested (left) and modular (right) patterns of the bipartite virus-host interaction networks. Adapted from Weitz et al. Trends in Microbiology, 2013.

Traditional approaches to study virus-host systems depend on laboratory cultures. However, culture-based experiments are limited by the number of culturable virus and host strains and thus do not necessarily recapitulate virus-host interactions in natural environments. It is estimated that only 2% of all microbes on earth can be cultured [60-62]. Additionally, the behavior of the microbes in the cultured condition may not fully reflect their behavior under natural conditions. Since cross-infection experiments need to be done in a pairwise fashion, they require large amount of time and experimental work. In recent years, culture-independent approaches, such as metagenomic based approaches, have also been applied to study virus-host interactions [63-65]. While such approaches provide population level virus-host interactions in natural environments, they often lack the precision to capture within-population diversity.

In Chapter 2, we performed integrated analysis to characterize the structure of virus–host interactions in a Yellowstone National Park (YNP) hot spring microbial community. To reconstruct the virus-host interaction network, we applied bioinformatics approaches to analyze the single cell sequencing data and overlaid evidence at the single-cell level with viral and cellular community structure. We performed three sets of analysis to identify putative virus-host interactions. These analyses were hexanucelotide analysis, network-based analysis based on single cell sequencing and CRISPR-based analysis. Using these approaches, we were able to characterize virus-host interactions in an extreme environment and demonstrated that the virus-host interactions were ubiquitous and complex.

1.5 The linkage between infection/interaction and genetic basis

Host range is an important trait of the virus which can be measured based on virus-host interactions. Such interactions present a strong selection on both the virus and the host. While virus and the host coevolve, both their genomes accumulate mutations that could potentially have an impact on host range. Many different approaches have been used to try to link the changes in host range with their genetic basis [66-71]. Previous studies have been focusing on a limited number of genes or mutations that were known to be involved in phage-host interaction [66, 67]. Recent studies analyzed the association between the host range and the genetic mutations at a genome-wide scale, but only from a static point of view rather than a coevolutionary perspective [68-70].

In chapter 3, we proposed a framework to link the changes in host range as well as the efficiency of phage infection with the changes in host and phage genetic profiles from a 37-day coevolution experiment. We constructed features based on whole-genome mutation profiles of phage and host and systematically evaluated the impact of these changes on host range and efficiency of infection. Our framework revealed both the genes that were previously known to participate in phage-host interactions and ones that could potentially participate. Since our approach is purely data-driven (*i.e.* it does not require prior knowledge on genes or mutations of specific phage or host strains) it could help prioritizing for the downstream validation on the mutations found to be important for virus-host interaction systems, including the ones that are not the same as what we have used.

1.6 Change of infection/interaction over time

The interactions of bacteriophage and their hosts form a complex network [13, 37]. Yet such networks do not remain static over the phage-host coevolution. In fact, both the environment and the underlying genetic basis together shape the network of interactions over time. Under experimental conditions, the interactions between single-species phage and host can be characterized by host range. Two competing theories, namely the arms race dynamic (ARD) and the fluctuating selection dynamic (FSD), have emerged to explain the patterns of phage-host coevolutionary dynamics [72-76]. In ARD, both the host and virus populations accumulate “improved” alleles over time. In FSD, virus populations need to constantly update the allele frequency in order to infect the currently most abundant host genotypes.

In chapter 4, we are not only interested in distinguishing between ARD and FSD based on the observed changes, but also, we are interested in evaluating the dynamics underlying the genetic basis, and how that can be related to the observed phenotypical dynamics. To do so, we investigated the dynamics of genotypes and phenotypes in coevolving virus-microbe, via analysis of full genome sequencing of *Escherichia coli* and bacteriophage λ. In contrast to expectations, we found that the emergence of resistant *E. coli* hosts and host-range mutant λ, in later stages of the experiment arose from rare subpopulations rather than the most recent, dominant lineages. This lineage leap-frog dynamic was enabled by fluctuations in ecological conditions that rescue rare lineages with increasing resistance and infectious genotypes, rather than enabling the progressive genomic changes envisioned in an arms race. We discussed the consequences of leapfrog dynamics for inferring evolutionary dynamics from phenotypes alone, whether in the case of coevolving phage-bacteria systems or in the evolution of human viruses in a changing landscape of adaptive immune cells.

1.7 Thesis summary

In this thesis, I propose to 1) Identify and characterize virus-host interaction networks under extreme environmental conditions, 2) Understand the driving forces in the arms race between the virus and its host by linking infectivity phenotypes with host and viral genomic mutations, and 3) Systematically characterize the evolutionary trajectories of viruses and hosts and identify the coevolutionary dynamics. For part 1, I have leveraged single cell sequencing technology with knowledge from metagenomics to reconstruct the complex virus-host interaction network based on samples from YNP hot springs [77]. By identifying virus-host interactions and characterizing the interaction networks, results from this chapter would improve our fundamental understanding of who infects whom under extreme environmental conditions. For part 2, I have modeled the observed virus-host interaction phenotypes and genetic profiles from a coevolutionary perspective and linked the phenotype and genotype for specific virus-host interactions. Results from this chapter would improve our understanding of the genetic basis for coevolution. For part 3, I have used computational approaches to reconstruct the coevolutionary trajectory of viruses and their hosts based on genotypical changes and phenotypical changes. Results from this chapter would reveal the consistency between both the genotypical and phenotypical coevolution dynamics. Taken together, the results showed that virus-host interactions are ubiquitous in natural environments, including extreme conditions. The virus-host interactions with the ubiquity and complexity, shapes the coevolution trajectory of both virus and host.

CHAPTER 2. A VIRUS OR MORE IN (NEARLY) EVERY CELL: UBIQUITOUS NETWORKS OF VIRUS-HOST INTERACTIONS IN EXTREME ENVIRONMENTS

Adapted from Munson-McGee, Jacob H., Shengyun Peng, Samantha Dewerff, Ramunas Stepanauskas, Rachel J. Whitaker, Joshua S. Weitz, and Mark J. Young. "A virus or more in (nearly) every cell: ubiquitous networks of virus–host interactions in extreme environments." The ISME journal (2018): 1.. Munson-McGee, Jacob H. and Shengyun Peng are the joint first-authors. Shengyun Peng designed the bioinformatics pipeline and performed the analysis for host and virus species identification and classification, as well as the reconstruction of the infection network at cellular and species level. In addition, the statistical test for contamination was conducted by Shengyun Peng.

2.1 Abstract

The application of viral and cellular metagenomics to natural environments has expanded our understanding of the structure, functioning, and diversity of microbial and viral communities. The high diversity of many communities, e.g., soils, surface ocean waters, and animal-associated microbiomes, make it difficult to establish virus-host associations at the single cell (rather than population) level, assign cellular hosts, or determine the extent of viral host range from metagenomics studies alone. Here we combine single-cell sequencing with environmental metagenomics to characterize the structure of virus-host associations in a Yellowstone National Park (YNP) hot spring microbial community. Leveraging the relatively low diversity of the YNP environment, we are able to overlay evidence at the single-cell level with contextualized viral and cellular community structure. Combining evidence from hexanucelotide analysis, single cell read mapping, network-based analytics, and CRISPR-based inference, we conservatively estimate that >60% of cells contain at least one virus type and a majority of these cells contain two or more virus types. Of the detected virus types, nearly 50% were found in more than 2 cellular clades, indicative of a broad host range. The new lens provided by the combination of metaviromics and single-cell genomics reveals virus-host interactions in extreme environments, provides evidence that extensive virus-host associations are common, and further expands the unseen impact of viruses on cellular life.

2.2 Introduction

For most natural environments, we lack a comprehensive inventory of both viruses, their microbial hosts and the virus-host networks they form [78, 79]. A comprehensive understanding is necessary because viruses likely play a central role in controlling microbial community structure and function [80-83]. Culture-based assays have revealed complex networks of infection between bacteriophage and bacterial hosts where a single bacteriophage is able to infect multiple bacterial species, and each bacterial species is a host for multiple different phage types [37, 59, 84, 85]. Comparative genomics of bacterial and archaeal strains also identified the presence of many different proviral elements [86-88]. However, culture-based infection assays and host range determination are limited in scope by the small number of microbial species and their viruses that can presently be cultured.

In recent years, several culture-independent methods have been developed to investigate host–virus associations [65]. These include analysis by metaviromics [13, 89], CRISPR spacer sequences [90-92], phageFISH [93], viral tagging [94, 95], microfluidic digital PCR [96], and single-cell genomics (SCG) [97-100]. Of these methods, SCG has provided some of the most detailed in situ insights into virus-host associations. For example, analysis of 58 single-cell amplified genomes (SAGs) from marine surface bacterioplankton showed that 20 of the SAGs contained viral sequences, some of which were shown to be actively replicating [101]. As a second example, analysis of 127 uncultivated SUP05 bacterial SAGs from an oxygen minimum zone revealed that ~1/3 were infected and that viruses reshaped core cellular metabolism [98]. Yet, few studies combine methods to provide a comprehensive inventory of virus-host associations for the entire microbial community.

2.3 Materials and Methods

2.3.1 Sample site

Water samples (1 mL) were collected from the Nymph Lake 01 (NL01) hot spring in Yellowstone National Park (YNP, Figure 4). At the time of sampling, the hot spring conditions were 83.3˚C, pH 2.45, and 1.085 mS conductivity. Samples were preserved on site with 5% glycerol and immediately flash frozen in a dry ice–ethanol bath. Samples were provided to the Bigelow Single Cell Genomics Center (Boothbay Harbor, ME).

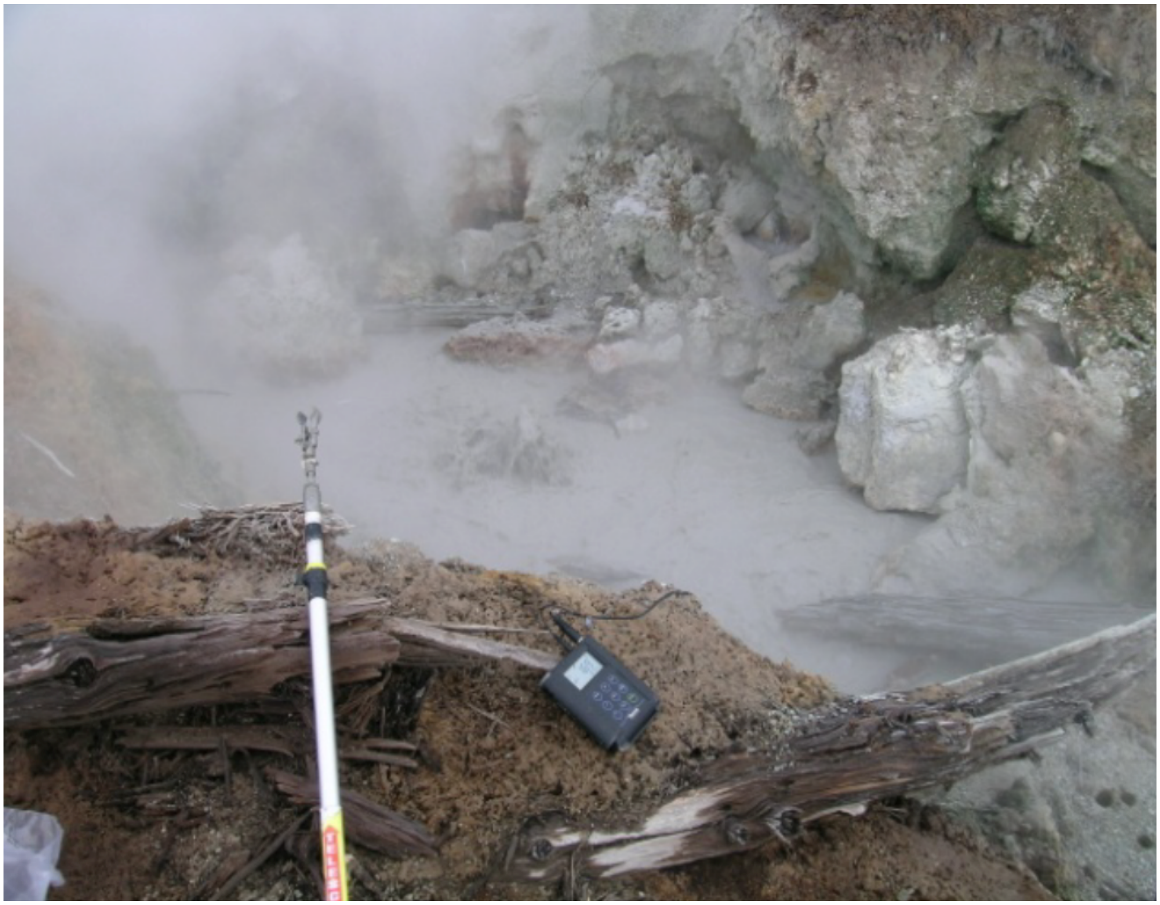

Figure 4 – Picture of the Yellowstone National Park NL01 hot spring from which cells were collected (Photo credit: Mark J. Young)

2.3.2 Single cell genome sequencing

Flow cytometric separation of individual cells and whole genome amplification were performed at the Bigelow Laboratory Single Cell Genomics Center using previously described methods [102, 103]. Based on effective MDA amplification of genetic material, a 384–well plate was selected for low coverage shotgun sequencing with an Ilumina end-paired HiSeq. The obtained reads were trimmed with trimmomatic v0.32 [104], normalized with kmernorm 1.05 (https://sourceforge.net/projects/kmernorm/), and assembled with SPAdes version 3.0.0 [105]. All contigs over 2.2kb were used to estimate genome size and completeness using CheckM [106].

2.3.3 Cellular classification

Cells were classified based on average nucleotide identity (ANI) using an ANI.pl script (https://github.com/chjp/ANI). All cells were compared to previously sequenced single-cell genomes from the same hot spring (Munson-McGee et al., 2015) as well as 18 thermophile reference genomes (Table 1). ANI scores were combined with the percent of SAG base pairs to generate an ANI bar code for every SAG against the 32 reference genomes (https://github.com/psy106616/SAG_hot_spring_YNP). All ANI matches covering <5% of the SAG genome were discarded. SAGs with two or more species present at ≥91% ANI were examined for the presence of double cells. Twelve SAGs showed evidence of having two cells present. Eight of these SAGs were classified as double cells and the remaining 4 were unclassified and removed from further analysis. SAGs with only a single species present at ≥95% ANI using at least 30% of the SAG genome were classified as belonging to the same species as the reference genome(s). SAGs that failed to meet the above categories (≥95% ANI, and or ≥30% coverage) were classified as likely single cells (ANI≥95% coverage <30%) (14 SAGs) or unclassifiable (28 SAGs) and removed from further analysis. ANI results were clustered hierarchically and a heatmap of ANI (Figure 5) and bp coverage (Figure 18) was generated for every classified SAG against every reference genome. 16S rRNA sequences were identified in 8 SAGs and compared to the reference genomes as a means to evaluate the accuracy of ANI-based taxonomic identification.

2.3.4 Hexamer frequency analysis

The contigs from SAGs classified as the same species were grouped together for hexamer frequency analysis. The hexamer frequency distribution of the grouped SAGs as well as a dataset of the viral types present in the NL01 hot spring [13] were generated using VirusHostMatcher [107]. The virus-host pair with the lowest hexamer distance was calculated by d2* [107] and pairs with a distance value <0.3 were used as an indication of a potential virus-host pair.

2.3.5 Viral sequence identification

All sequence reads obtained from SAG sequencing were used as the query of a BLASTn search against the viral database previously described [13]. Reads with a significant match (e-value $<1.0^{-10}$) to the viral database were filtered and classified as having a viral origin if they matched at >95% nucleotide identity over 100 bp. Identified viral reads were subsequently mapped back to their viral group previously established using network analytics [13] using a custom script. Reads that mapped to multiple viral groups were assigned to the viral group with the most reads from that individual SAG to reduce false positives. To test if this mapping protocol resulted in false identification of viruses, controls were performed where the same SAG reads were mapped to the contigs from the Tara Oceans Virome (TOV) datasets (18SUR 66 Mbp and 18DCM 99 Mbp) [89] and a virome from the human gut (6 Mbp) [108] both of which were not expected to contain viruses found in hot spring environments. Additionally, sequence reads from 25 publically available SAGs generated from non-hot spring environments from the JGI IMG (http://jgi.doe.gov/) representing 10 bacterial and two archaeal phyla (703.7 million total reads) were compared against the viral database at the same stringency described above.

We used the following rationale to establish a threshold criteria for identifying virus-host associations within an individual SAG dataset. Since the estimated genome completeness for each SAG varied, we first determined the ratio of identified unique viral sequence reads (average of 150bp in length) to the total unique host base pairs for each SAG. The number of unique viral base pairs was determined by mapping SAG reads to the NL01 viral dataset using BLASTn and removing any overlap to the reference viral genomes. The unique number of host base pairs was calculated using the ANI base composition statistic [109, 110] for each SAG calculated with respect to the 32 reference genomes, minus the unique viral base pairs. These ratios were compared to expected ratios using an average viral genome size of 30 kb, a host genome size from 1.5-3.0 MB, and assuming no sequencing bias towards either virus or host or a 2X bias towards virus or host (arbitrarily chosen to account for variation in amplification). Using this rationale, we determined that a minimum of 2-5 unique 150bp viral sequence reads should be present in an individual SAG dataset if that SAG were in fact infected by a virus.

After determining the profile of viral content in each individual SAG, the dataset was treated as a bipartite network. The BiMat algorithm [111] was applied to the bipartite viral–host network for modularity analysis. The binary network was generated using a minimum cutoff of 2 or 5 unique viral sequence reads from a SAG to the 110 viral groups previously identified in the NL01 hot spring [13].

2.3.6 CRISPR spacer sequence identification

CRISPR spacer sequences were identified in SAG contigs using Piler-CR [112]. Identified CRISPR spacer sequences were extracted and compared against the viral database with virus-host associations assigned to CRISPR spacer sequences that match ≥90% identity over the entire spacer length. Contigs with CRISPR matches were selected and the viral group they belonged to was identified using a custom python script. As controls for the false identification of CRISPR spacer-virus associations, a CRISPR spacer dataset of 966 unique spacers from a human gut microbial community was analyzed against the NL01 viral database. In addition, the SAG CRISPR spacer sequences were compared to the viral dataset of the human gut bacterial community [108] under the same conditions described above.

2.3.7 Statistical test for contamination

To identify the possibility of sample contamination within adjacent wells on the 384-well plates during sample preparation, a statistical approach was used to evaluate the correlation between the physical distance and the sequence similarity between adjacent wells. First, the physical distance between two neighboring wells from the same row or the same column as a unit was defined. A distance matrix with all pairwise distances was computed based on the Euclidean distance between any two wells. Second, the sequence similarity between two wells was calculated based on the number of unique and shared viral groups of the two wells. The Jaccard index of a given pair of wells A and B was calculated as $J = (S_A \cap S_B)/(S_A \cup S_B)$, where S_A denotes the set of viral groups in SAG A and S_B denotes the set of detected viral groups in SAG B. Third, the Spearman's rank correlation was calculated to evaluate the relationship between physical distances of the wells and the Jaccard index. A series of distance cutoffs between 1.5 and 3 were used to calculate the Spearman's correlation of two wells to focus on the cross contamination in nearby wells. Finally, to evaluate the statistical significance of the observed Spearman's correlation coefficients at different distance cutoffs, a permutation test was performed to obtain the null distribution of the Spearman's coefficients. For the permutation test, the plate layout was randomly shuffled 100 times and the Spearman's correlation coefficients were re-calculated at corresponding distance cutoffs. The observed Spearman's correlation coefficients were then compared with the null distributions.

2.4 Results and Discussion

In this study, we combined single-cell genomics and community metagenomics to characterize virus-host interactions. Single cells were randomly isolated directly from hot spring samples, their genomes amplified and sequenced. 109,930,697 total paired end reads were produced from 307 single amplified genomes (SAGs, average ~358,000 reads per cell) with a maximum of 2,015,593 and a minimum of 3,823 reads per SAG (Table 2). A total of 34.1 Mbp was assembled ranging from a minimum total bp of 7,806 to a maximum of 380,184 with an average total assembled length of 110,997 bp per cell. This correlates to an average genome completeness of approximately 9% but ranges from <1% to 44% complete based on CheckM analysis.

In order to determine the cellular identity of each SAG a multistep process was developed (Figure 19). First, the Average Nucleotide Identity (ANI) [109, 110] for all contigs greater than 2kb for each SAG was calculated with respect to 32 reference genomes. The reference genomes consisted of a combination of SAGs previously sequenced at high depth (17-90% genome completeness) from the same hot spring and other complete or near complete thermophilic archaeal and bacterial reference genomes from the NCBI database (WGS release 212, February, 2016). Second, the percentage of sequence homology between a SAG and the reference genomes were determined. SAGs were hierarchically clustered and assigned to their closest cellular species based on ANI score in combination with the percentage of sequence homology between the SAG and its closest reference genome (Figure 5, online Supplemental Table 3). We utilized an ANI score of 95% in combination with 30% sequence coverage to classify the majority of SAGs (253/307 SAGs). The 54 SAGs that were not classified were either double cells of the symbiont Nanoarchaea with its Acidocryptum host (8 examples, discussed below), or 46 SAG cells that failed to meet our classification criteria. These 54 SAGs were removed from further analysis. To further support cellular identification, all SAGs were examined for 16S rRNA gene sequences. 16S rRNA sequences were present in only 8 SAGs and cellular classification based on their 16S rRNA was determined by alignment to reference genomes. In all 8 cases, the 16S rRNA gene and ANI classifications produced the same result.

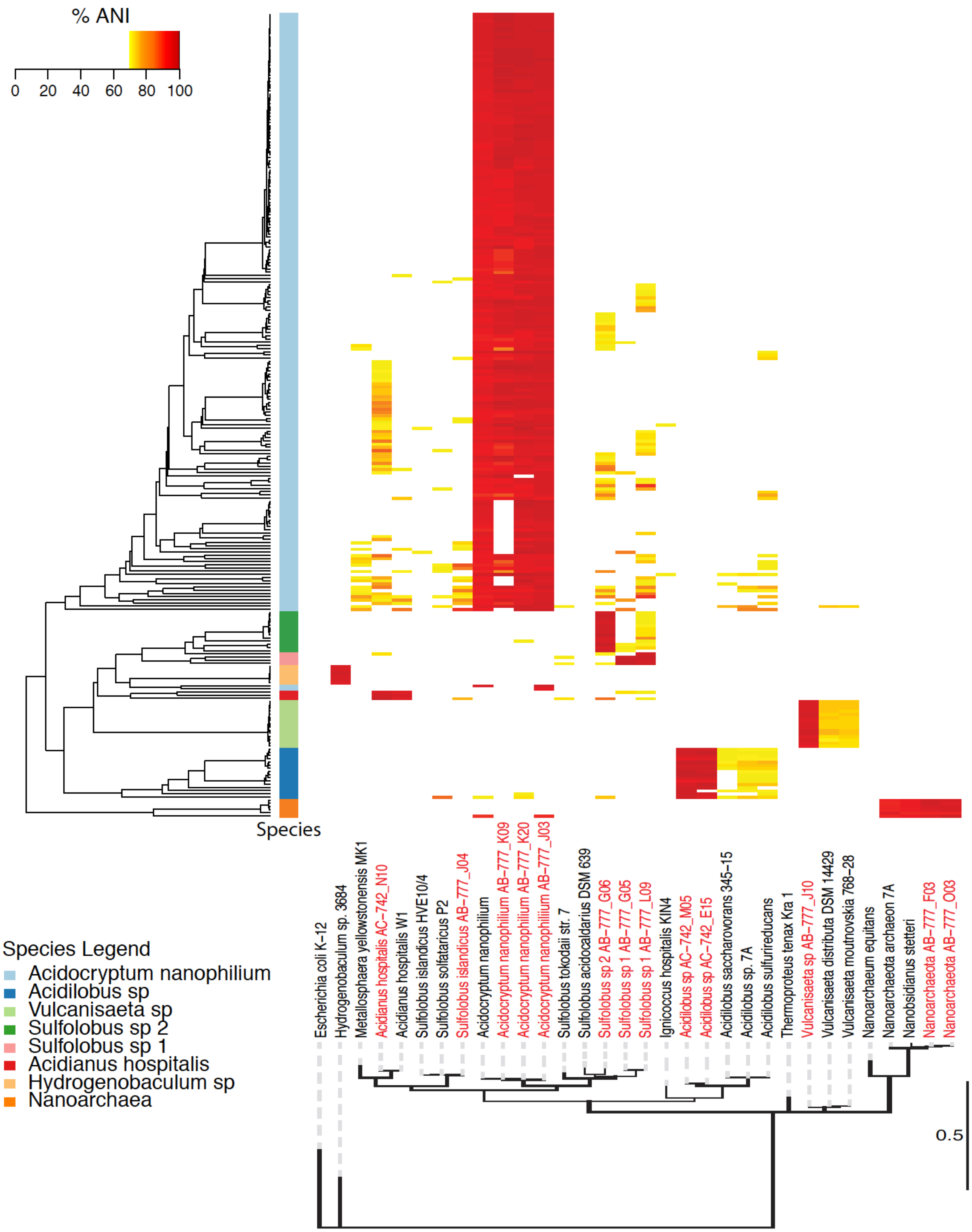

Figure 5 – Cellular classification of SAGs

Heatmap of the average nucleotide identity (ANI) of 253 classified single cell SAGs sequenced in this study compared against 32 reference genomes including 13 SAGs previously sequenced at high coverage from the same hot spring [29] (red text). SAGs were hierarchically clustered using complete linkage (left hierarchical dendrogram). The *column directly to the right of the hierarchical dendrogram indicates classified cell species (color key provided) for all SAGs classified as single cells. Partial length 16S rRNA sequences from the 32 reference genomes were used to construct a maximum likelihood phylogenetic tree and nodes with greater than 95% posterior probability are bolded. The E. coli strain served as the outgroup. The scale bar is in substitutions per site.*

The classification of SAGs revealed a low-diversity microbial community consisting of 8 cellular clades, dominated by Archaea (Figure 5), consistent with our previous studies [113]. The 253 SAGs classified to one of 8 cellular clades. Of these, 247 were classified as one of 7 clades of Archaea (97.6%), 6 were classified as members of a single clade of Bacteria (2.4%), and none were classified as Eukaryotic. The vast majority (98%) of the Archaeal cells are members of the Crenarchaeota (241/247 SAGs) while Nanoarchaeota (6) make up the remaining 2.0%. The only bacterial species detected belonged to the Aquificales. The NL01 microbial community structure was nearly identical to the community structure determined by 16S rRNA amplicon sequencing from a sample taken 12 months previously. Overall, 6 of the 8 clades identified in this study have not been cultured to date, and these 6 uncultured clades comprise 96% of the SAGs in this study (244/253 SAGs).

As a first step in characterizing virus-host associations, we generated a distance matrix based on hexamer nucleotide analysis using the d2* metric [107] of the 8 cellular clades against the 110 viral types previously determined to be present in the hot spring [13] (online Supplemental Table 4). If the smallest measured d2* between a cell type and a virus type was <0.3 it was used as indication of a possible virus-host association. Previous studies have indicated that hexamer nucleotide analysis can be a useful predictor of virus-host associations, given a cutoff of <0.3 as a conservative identification of possible virus-host pairs [107]. Hexamer nucleotide analysis indicated that 61 virus types were associated with the 7 archaeal cell types. The number of virus types associated with a particular archaeal cell type ranged from 28 virus types for the Acidilobus clade to 1 for the *Sulfolobus* sp 1, clade. Controls consisting of 75 bacterial genomes unlikely to serve as hosts for the hot spring viruses along with the grouped sequences from the 8 SAG cellular clades of this study, found no false virus-host associations to the bacterial genomes (online Supplemental Table 4). A limitation of hexanucleotide analysis is that it only suggests a possible virus-host association and does not indicate viral host range [107]. Moreover, hexanucelotide analysis lacks resolution when closely related cellular species/strains are compared [107]. Therefore, this analysis provides an indication of possible virus-host associations and not definitive proof of the association.

Further identification of individual virus types within each SAG was accomplished by mapping sequencing reads from individual SAGs to the 110 viral types present in NL01 previously established by network-based analytics using time-series community viromics data [13]. We first established a rationale for how many viral base pairs would be expected to be detected in given SAGs given the low level of genome completeness obtained (average host genome completion was 9%). This was accomplished by determining the ratio of viral sequence to host base pairs for each SAG (Figure 20) and comparing observed ratios to expectations (see Methods). We estimate that finding two or more unique SAG viral sequences (at least 300 bp) represents a reasonable minimum for detecting virus-host associations. A conservative threshold for virus-host association assumes a two-fold bias in sequence amplification, suggesting a threshold of five or more unique sequence reads (at least 750bp) to a given viral group in a SAG. Using the more conservative requirement of ≥5 SAG viral reads (750bp) matching a virus type, viral sequences were detected in 160 of the 253 classified single cell SAGs (63% of SAGs) (Figure 6, online Supplemental Table 5), virus-host associations identified using the lower value of ≥2 viral reads (300 bp) matching a virus type are provided in online Supplemental Table 5. Viral sequences were detected in all cellular groups except for *Hydrogenobaculum*. Of the 110 viral types, 26, were detected (24% of total vial types) in the 253 SAGs. For example, over 49,851 reads mapped to 34.5kb of continuous sequence represented on the entirety of 3 contigs assembled from a single *Acidocryptum nanophilum* SAG (AD-903-K19). This 34.5kb segment likely represents the near-full length genome of a new archaeal virus.

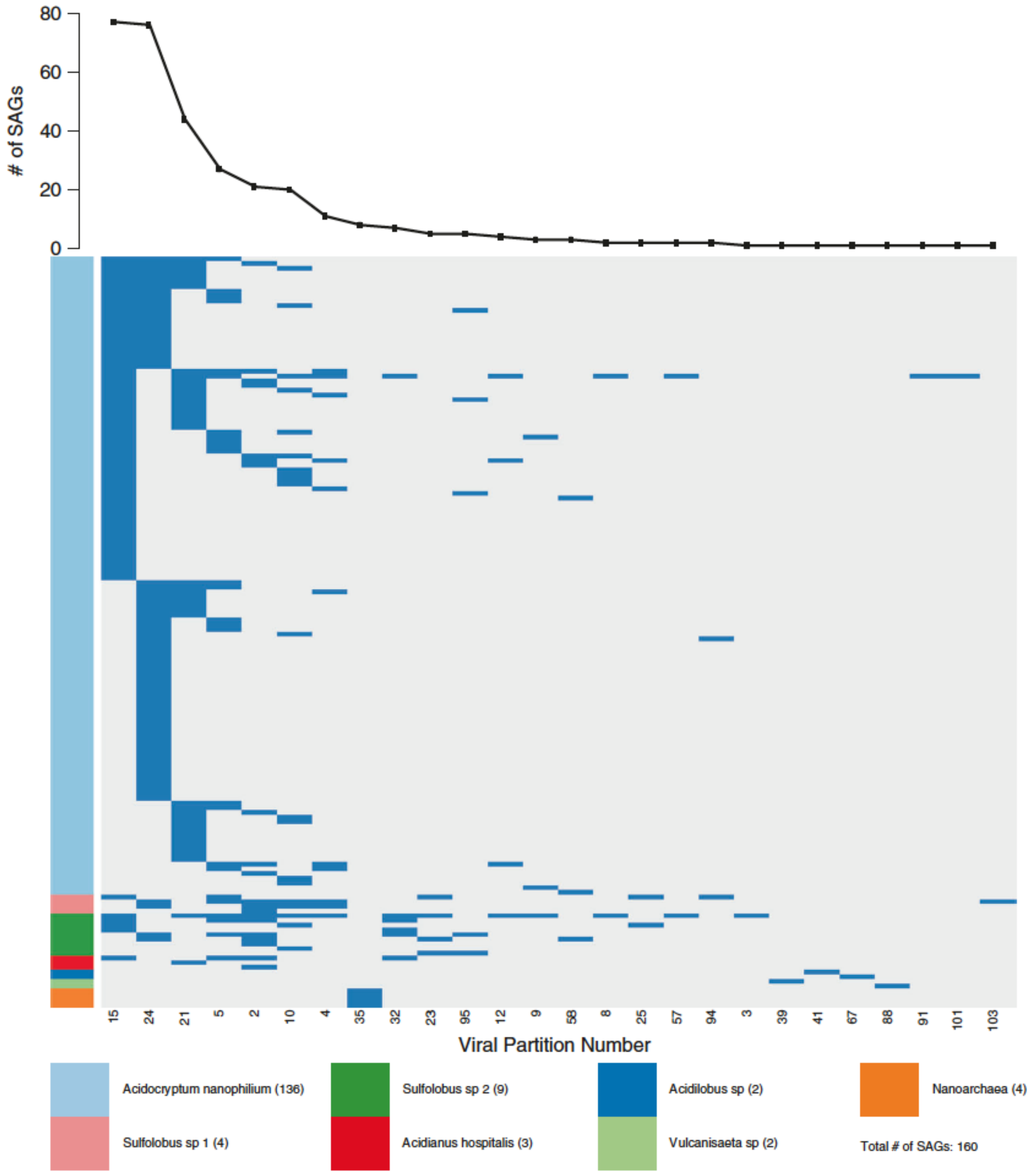

Figure 6 – Detection of viral types in 160 SAGs

26 of 110 virus types were detected by BLASTn identification of SAG sequencing reads to NL01 viral community. Viral group numbers are taken from Bolduc et al.. Blue indicates the detection of a viral group in a SAG and white indicates that a viral group was not detected in a SAG. SAGs are grouped by cell type (vertical axis, a color key for cell the type is provided) and viral groups (horizontal axis) are ordered by detection frequency (top graph)

Next, we examined the number of virus types found in each infected SAG. Surprisingly, more than one viral type was detected in a majority of the cells. Of the 160 SAGs where viral reads were detected, 95 (59%) had ≥750 bp sequence reads from 2 or more viral types, with an average of 2 viral types detected per cell (Figure 6). This data suggests that co-infection may be common in the hot spring environment. Indeed, 63% of cells randomly sampled by SAG analysis had evidence of virus association. Given the low depth of average SAG genome coverage (approx. 9%), we anticipate that actual association levels are much higher, suggesting that (nearly) all cells in the hot spring interact with viruses. This work extends the scope of virus associations measured in previous reports in marine environments where viral sequences were found in 30–50% of cells [98, 101].

Several lines of evidence indicate that the detected virus-host associations are biologically relevant and not a consequence of random associations. First, no sequencing reads from any of the 307 SAGs were recruited onto two much larger marine viral metagenomic or a human gut viral metagenomic datasets using the identical mapping stringency conditions (Table 3). Additionally, sequencing reads from 25 publicly available non-hot spring associated SAGs from the JGI IMG (https://img.jgi.doe.gov/) representing 10 bacterial and two archaeal phyla were compared against the viral database used in this study. These SAG’s isolated from other environments, totaling 703.7 million reads, did not match any of the 110 viral groups used in this study at the same stringency settings (Table 4). These controls support the conclusion that the conditions used in this study strike a balance between viral detection sensitivity and stringency sufficient to detect biologically relevant virus-host associations in individual SAGs. Future targeted virus RTqPCR analysis on single cells should clarify if the detected viruses are actively replicating.

Analysis of CRISPR spacer sequences were used to detect additional virus-host associations. CRISPR spacer sequences were extracted from SAGs and mapped to the 110 viral types (online Supplemental Table 8). A total of 2,321 unique CRISPR spacer sequences were detected in 135 SAGs. Spacer sequences were found in all cell types except for the *Nanobsidianus*. Previous studies had also failed to identify CRISPR sequences in *Nanobsidianus* sp from YNP hot springs [113, 114]. CRISPR spacer-virus matches were found for 695 (30%) spacer sequences to 38 of the 110 viral types from 121 SAGs (90% of spacer-containing SAGs). The majority of spacers with matches were found in *Acidocryptum* cells (541/695). Twenty-two viral types were identified by both read mapping and by CRISPR spacer matching to the same cellular species. As expected, controls of comparing 966 non-relevant CRISPR spacer sequences derived from the human gut microbial community to the 110 hot springs viral types failed to detect any virus-host associations under the same conditions. Overall, 47 of the 110 viral types (42%) were detected by either mapping of SAG reads or by SAG CRISPR spacer matching. Furthermore, 18 of these 47 virus types were predicted by hexamer distance analysis to the same host. Taken together, these 3 independent measures support the conclusion that virus-host associations are a common feature in this hot spring environment.

It is worthwhile to retrospectively consider how useful it is to relay on ANI to accurately connect viruses to potential hosts. In this work we have the advantage of having internal standards of viral sequences present within individual SAGs to compare against ANI analysis at different threshold cut offs. We observe that ANI cut off values of <0.3 are reasonable values reduce detecting false positives while maintaining the detection of meaningful host-virus pairs.

The contextualized virus-host associations and CRISPR spacer analysis (Figure 7, online Supplemental Tables 8) provide complementary information on the realized and potential host range of viruses, respectively. By combining these two lines of evidence we asked: what is the host range of individual virus types? Twenty-four viruses infected only a single cellular clade. In contrast, 23 virus types were detected in >2 host genera within the *Sulfolobaceae* family. Every previously characterized virus detected was found in at least one new host species. For example, STIV previously shown to infect *S. solfataricus* [115], was also detected in *Acidocryptum* cells. These results demonstrate that culture-independent approaches can be used to investigate the host range of uncultured viruses across the entire microbial community. Despite finding multiple new associations, it is important to recognize that reported host ranges remain *lower bounds*, i.e., increased depth of sampling could reveal even more virus types within classified SAGs.

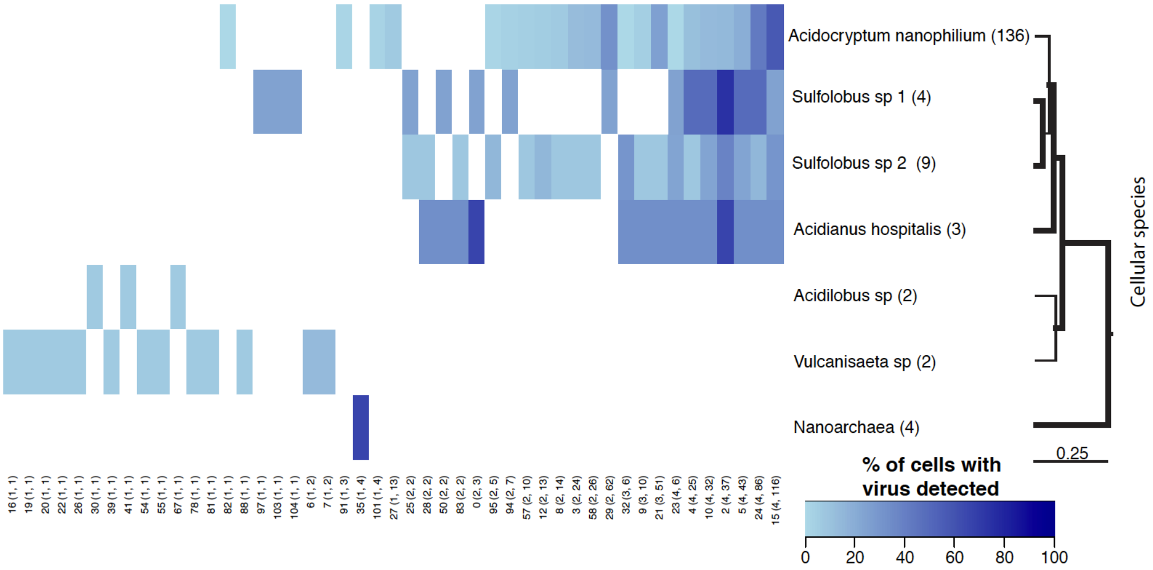

Figure 7 – Ubiquitous interaction of multiple viruses with cells

The heatmap indicates the detection frequency of 47 viral groups detected by BLASTn analysis or the matching of CRISPR spacer sequences. Viral groups are arranged from least frequently detected to the most frequently detected. Numbers below the heatmap are viral group numbers taken from [16] and numbers in parenthesis indicate the number of *species and cells that a group was detected in. The number after the species name on the right hand side is the number of cells classified as members of that species. Partial length 16S sequences from representative genomes were used to make a ML tree and nodes with greater than 0.95 posterior probability are bolded. The scale bar is in substitutions per base. Detected viral groups with described members are: group 0 = SIRV1,2, group 23 = ASV1, SSV1,2, 4–9, group 26 = ATV, group 28 = AFV1, group 29 = STIV1,2 and group 32 = STST1,2 and ARSV1*

The inference methods in the present analysis are made possible by network-based analytics that determine viral groups but also limited by relatively low SAG coverage (~9%). As a consequence, we cannot easily distinguish actively replicating viruses within individual SAGs, define their viral lifestyles (lytic, lysogenic, or chronic) or define individual viruses at the species level. Despite these limitations, it is remarkable that we detect *in situ* the majority of host and viral types – currently identifiable from whole community sequencing projects – and their associations within a relatively low number of SAGs.

This work shows the advantage of combining single-cell genomics with metagenomics to establish a comprehensive understanding of virus-host associations in a focal environment. Unlike previous studies of virus-microbe interactions, we are able to contextualize virus-host infection networks and link the identity of viruses found in different cells. In doing so, we both identify the hosts and host-range of virus types. Guided by the knowledge of the overall virus community, the incorporation of SAG analysis – including contextualized community network mapping and CRISPR detection – allows for the identification of individual hosts and the host range of an individual virus type in a culture-independent fashion. This study shows that nearly all cells in the NL01 hot spring interact with viruses, that multiple, concurrent interactions are common, and that a broad spectrum of virus types from specialists to generalists coexist in a relatively low-diversity community. These results should encourage the development of more robust empirical methods and theoretical models to assess the relevance of superinfection and a diversity of viral lifestyles in shaping natural communities.

CHAPTER 3. LINKING GENOTYPE WITH PHENOTYPE IN THE BACTERIOPHAGE LAMBDA AND ESCHERICHIA COLI INTERACTION NETWORK

This chapter is being prepared for publication as: Shengyun Peng, Chung Yin (Joey) Leung, Animesh Gupta, Justin R. Meyer and Joshua S. Weitz. 'Linking genotype with phenotype in the bacteriophage lambda and host interaction network'.

3.1 Abstract

Characterization of species interaction networks has led to a better understanding of microbial community structure and function. Interaction networks are typically established by phenotypic assays, little is known regarding the link between phenotypic changes and underlying changes in genotypes. Previous approaches and theories developed to address this question relies on prior knowledge of the functional role of the gene or mutation, and thus were typically limited by prior knowledge. In this study, we proposed a data-driven framework that systematically evaluated such link between phage-host interaction phenotype and genotype. We measured the changes in host range and efficiency of infection for bacteriophage λ strains sampled from a 37-day coevolution experiment. We also characterized the changes in the genetic profiles of both the phage strains and host strains based on whole genome sequencing data. A two-step framework was built to link the phenotypical changes in terms of the host range and efficiency of infection with the changes in the genetic profiles. Overall, our framework systematically evaluated the genetic basis for phage-host interaction phenotypes, identified several important genes that

have been experimentally validated to participate in phage-host interactions and also revealed new genes that could potentially participate in the phage-host interaction.

3.2 Introduction

Next-generation sequencing technology has revealed widespread diversity in microbial communities [63, 77]. In parallel, the development of analytical tools to characterize species interaction networks has led to a better understanding of microbial community structure and function [116-118]. Despite the parallel rise of these fields, there have been relatively few exchanges between the two. Interaction networks are typically established by phenotypic assays and not genome sequences. Theoretically, it should be possible to predict the interaction network of microbial species from genome sequences alone, since their genetics determine traits which, in turn, modulate the identity, mode, and quantitative rate of interactions with other microbes. For example, a bacteriophage (phage) can only infect bacterial strains they can adhere to [119-121]; such adsorption requires expression of a cell-surface receptor (e.g., protein, lipid, carbohydrate). Despite significant progress in linking microbial genotype to phenotype, less progress has been made in linking pairs of microbial genotypes to an interaction phenotype [23, 37, 58, 71, 122-127].

The problem of understanding the genetic basis for interactions requires the development of new computational approaches to construct a genotype-by-phenotype map. Current approaches to estimate this map try to correlate phenotypic differences with genetic variation (e.g., this is true for the broad scope of work in genome-wide associated studies [128-130]). The challenge for inferring interaction-associated phenotype, is that such interactions arise due to the interaction of multiple genotypes, e.g., phage and host genotypes. For example, mutation-based association approaches have been used to find the combination of virus and host mutations that are associated with the virus-host interaction phenotype [68-70]. Such approaches have similarities to the more general problem of studying complex traits that are affected by gene by gene (G x G) interactions and gene by environment (G x E) interactions. The importance of such interactions may explain the "missing heritability" problem where genetic effects discovered by association analysis do not sum to the estimated heritability of the trait [131-133].

Predicting virus-microbe interactions is highly dependent on taxonomic scale. For example, computational approaches are increasingly used to predict the host range of viruses, e.g., leveraging tetranucleotide frequencies and other sequence-specific information (reviewed in *Edwards et al.* and *Dutilh et al.* [88, 134]). However, predicting strain-specific interactions remains poorly understood, particularly in light of the fact that taxonomic markers are a poor proxy for infection profiles [135]. Prior work on microevolutionary changes in infectivity have focused on changes to genes or proteins with known functions in model organisms [66, 67, 136]. Such approaches are dependent on the existing annotation of genes or mutations, and thus are limited by both the quality and quantity of annotations available. Such a dependence limits our ability to identify novel loci that could modulate infection phenotypes.

Here, our work aims to link whole genome-wide changes in both the phage and host with the observed changes in interaction phenotypes. We do so leveraging measurements of whole genotypes and phenotypes amongst coevolving populations of *Escherichia coli* B strain REL606 and bacteriophage λ strain cI26 during a 37 day experiment. By jointly measuring phenotypes and genotypes, we set out to develop a framework that could identify the link between genotypes and phenotypes. In doing so, we also address the question: do host mutations, virus mutations, or some combination, serve as better predictors of infection outcome?

3.3 Materials and Methods

3.3.1 Experimental setup and data collection

The *Escherichia coli* B strain REL606 and bacteriophage λ strain cI26 were used as ancestral strain for host and virus respectively (Figure 8). Phage and host were cocultured for a 37-day period. Samples were taken on checkpoint days for pairwise quantitative plaque assays as described in Chapter 4. The EOP value measures the efficiency of a phage infecting a derived host strain relative to that for infecting the ancestral strain. The EOP value for a phage, j, infecting a host, i, is computed as

$$e_{ij} = \frac{q_{(i,j)}}{q_{(anc,j)}} \times d^{s_{(i,j)} - s_{(anc,j)}} , \quad (1)$$

where $q_{(i,j)}$ is the number of plaques on the petri dish for phage j against host i, $q_{(anc,j)}$ is the number of plaques on the petri dish for phage j against the ancestral host strain, $s_{(i,j)}$ is the number of dilutions performed to get distinguishable and countable clear plaques for phage j against host i, $s_{(anc,j)}$ is the number of dilutions performed to get distinguishable and countable clear plaques for phage j against the ancestral host strain and d is the dilution ratio which is 5 in our experiment. A positive EOP value from the cross-infection plaque assay indicates a successful infection event for a given phage-host pair. In contrast, a zero EOP value indicates the absence of the infection event for a phage-host pair. A larger EOP value from the cross-infection plaque assay indicates that the phage can infect a given host more efficiently than infecting the ancestral host strain.

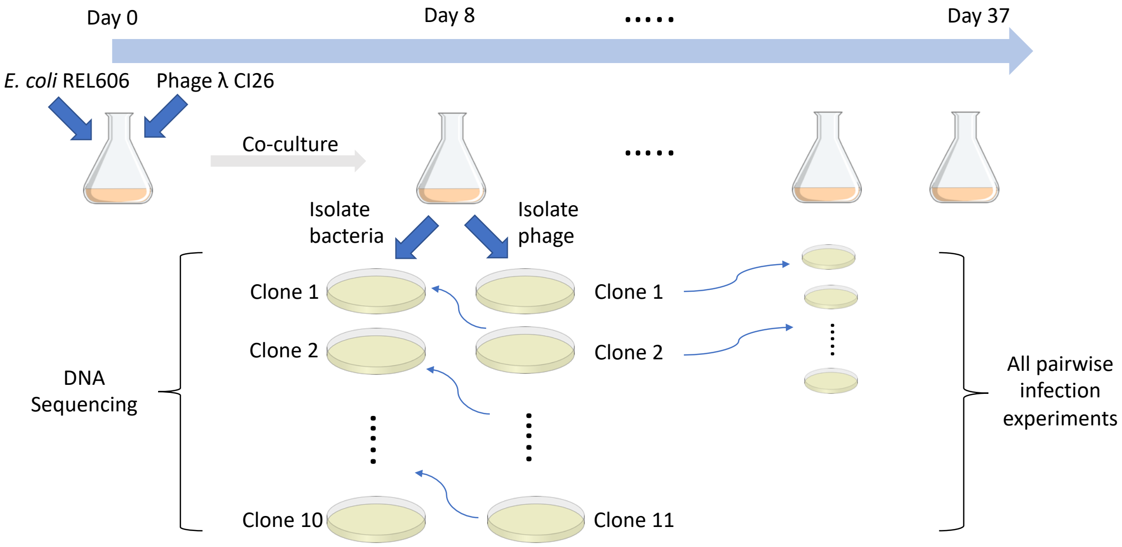

Figure 8 – Experimental design of the cross-infection plaque assay

For each phage and host samples taken from each checkpoint, the DNA extraction, library preparation and sequencing experiment was performed as described in Chapter 4. Mutation profiles based on the genome sequencing data were constructed using *breseq* as described in Chapter 4. In addition to the mutations revealed by *breseq* results, for both host and phage we created an artificial mutation as the indicator for the ancestral strain in order to add the ancestral strain into the mutation profile table. For this artificial mutation, only the ancestral strain is indicated to have this mutation. All other strains were shown to not have this mutation in the mutation profile table.

3.3.2 Feature construction

For a total number of U host samples and Vphage samples, we denote the EOP value for the i-th host against j-th phage as e_{ij} where $i \in [1, U]$ and $j \in [1, V]$. Let N be the total number of unique mutations observed for the host and M be the total number of unique mutations observed for the phage, the host mutation profile H is a matrix of dimension U by N, and the phage mutation profile P is a matrix of dimension V by M. Let h_{il} be an element from H, then $h_{il} = 1$ corresponds to the presence of the l-th mutation in the i-th host whereas $h_{il} = 0$ corresponds to the absence of the l-th mutation in the ith host. Similarly, let p_{jk} be an element from P, then $p_{jk} = 1$ corresponds to the presence of the k-th mutation in j-th phage whereas $p_{jk} = 0$ corresponds to the absence of the k-th mutation in the j-th phage.

Five sets of features were constructed based on the mutation profiles of the host and phage. The H:MF is constructed based on only the host mutation profiles. Model Φ that utilizes the H:MF can be represented as:

$$\phi_{ij}^{(1)} = \gamma_1 + \sum_{l=1}^{N} \alpha_l h_{il} , \qquad (2)$$

where γ_1 represents a scalar of the bias term and α_l is the coefficient for the l-th host mutation. γ_1 and α_l will be learned from the model. The model utilizing H:MF can also be represented in matrix form as:

$$\Phi^{(1)} = \Gamma_1 + H \cdot R_\alpha , \qquad (3)$$

where Γ_1 is a U by V matrix by repeating γ_1, i.e. $\Gamma_1 = [\gamma_1]_{U\times V}$, R_α is a N by V matrix by stacking the same coefficient vector α horizontally, i.e. $[\alpha|\alpha| \cdots |\alpha|\alpha]_{N\times V}$.

The P:MF is constructed based on only the phage mutation profiles. Model Φ that utilizes the P:MF can be represented as:

$$\phi_{ij}^{(2)} = \gamma_2 + \sum_{k=1}^{M} \tilde{\alpha}_k p_{jk} , \qquad (4)$$

where γ_2 represents a scalar of the bias term and $\tilde{\alpha}_k$ is the coefficient for the k-th phage mutation. γ_2 and $\tilde{\alpha}_k$ will be learned from the model. The model utilizing P:MF can also be represented in matrix form as:

$$\Phi^{(2)} = \Gamma_2 + [P \cdot R_{\tilde{\alpha}}]^T , \qquad (5)$$

where Γ_2 is a U by V matrix by repeating γ_2 and $R_{\tilde{\alpha}}$ is a M by U matrix by stacking the same coefficient vector $\tilde{\alpha}$ horizontally, i.e. $[\tilde{\alpha}|\tilde{\alpha}|\cdots|\tilde{\alpha}|\tilde{\alpha}]_{M\times U}$.

Model Φ that utilizes P+H:MF can be represented as:

$$\phi_{ij}^{(3)} = \gamma_3 + \sum_{l=1}^{N} \alpha_l h_{il} + \sum_{k=1}^{M} \tilde{\alpha}_k p_{jk} \ , \qquad (6)$$

where γ_3 represents a scalar of the bias term, α_l is the coefficient for the l-th host mutation and $\tilde{\alpha}_k$ is the coefficient for the k-th phage mutation. γ_3, α_l and $\tilde{\alpha}_k$ will be learned from the model. The model utilizing P+H:MF can also be represented in matrix form as:

$$\Phi^{(3)} = \Gamma_3 + H \cdot R_{\alpha} + [P \cdot R_{\tilde{\alpha}}]^T \ , \qquad (7)$$

where Γ_3 is a U by V matrix by repeating γ_3, i.e. $\Gamma_3 = [\gamma_3]_{U\times V}$, R_{α} is a N by V matrix by stacking the same coefficient vector α horizontally, i.e. $[\alpha|\alpha|\cdots|\alpha|\alpha]_{N\times V}$ and $R_{\tilde{\alpha}}$ is a M by U matrix by stacking the same coefficient vector $\tilde{\alpha}$ horizontally, i.e. $[\tilde{\alpha}|\tilde{\alpha}|\cdots|\tilde{\alpha}|\tilde{\alpha}]_{M\times U}$. The assumption for P+H:MF is that the impact of mutations from both the phage or host have additive effects on the observed outcome.

Model Φ that utilizes P×H:MF as the input can be represented as:

$$\phi_{ij}^{(4)} = \gamma_4 + \sum_{l=1}^{N} \sum_{k=1}^{M} \beta_{lk} h_{il} p_{jk} \ , \qquad (8)$$

where γ_4 represents a scalar of the bias term, β_{lk} denotes the coefficient for the l-th host mutation and k-th phage mutation in the corresponding i-th host and j-th phage pair. γ_4 and β_{lk} will be learned from the model. The model utilizing P×H:MF can also be represented in the matrix form as:

$$\Phi^{(4)} = \Gamma_4 + H \cdot B \cdot P^T \ , \qquad (9)$$

where Γ_4 is a U by V matrix by repeating γ_4, i.e. $\Gamma_4 = [\gamma_4]_{U\times V}$, B is the N by M coefficient matrix. The assumption for the P×H:MF is that the impact of the genetic mutations on the observed outcome comes from the additive effects of co-occurring phage-host mutation pairs. In other words, $h_{il}p_{jk} = 1$ only when both the host i has mutation l and phage j has mutation k.

Based on the definition of P+H:MF and P×H:MF, it is natural to combine both features to get a more sophisticated input feature, Joint:MF, by adding up both effects. Model Φ that utilizes the Joint:MF can be represented as:

$$\phi_{ij}^{(5)} = \gamma_5 + \sum_{l=1}^{N} \alpha_l h_{il} + \sum_{k=1}^{M} \tilde{\alpha}_k p_{jk} + \sum_{l=1}^{N} \sum_{k=1}^{M} \beta_{lk} h_{il} p_{jk} \,. \tag{10}$$

The matrix form of Joint:MF is:

$$\Phi^{(5)} = \Gamma_5 + H \cdot R_{\alpha} + [P \cdot R_{\tilde{\alpha}}]^T + H \cdot B \cdot P^T \,, \tag{11}$$

where Γ_5 is a U by V matrix by repeating γ_5, i.e. $\Gamma_5 = [\gamma_5]_{U \times V}$.

3.3.3 Framework design

In order to link the phenotypical changes of phage-host interactions with their genotypes, we designed a framework comprised of two steps. This is because the capability of a phage to infect a host and the efficiency of a phage infecting a host may have different underlying molecular mechanisms. The first step of our framework is designed for predicting the existence of phage infectivity. The step 1 model tries to find the set of features that can best distinguish between the successful infections and the failed ones by using classification models. The second step is based on the subset of phage-host pairs where the host is susceptible to the phage (EOP > 0). The step 2 model of our framework is designed to evaluate the potential impact of the genotype on this observed phenotype by modeling the efficiency of the phage in infecting a host.

3.3.4 Model for predicting existence of phage infectivity

For a given phage-host pair, in order to determine the presence or absence of a successful infection event, we binarized the EOP values e_{ij} into 0 and 1, i.e.

$$d_{ij} = \mathbf{1}_{\{e_{ij} > 0\}} , \qquad (12)$$

where $d_{ij} = 0$ indicates a failure of the infection and $d_{ij} = 1$ indicates success. As a result, this problem became a classification problem. Here we used logistic regression to model the relationship between mutation profiles and the existence of successful infection in phage-host pairs, that is

$$\phi_{ij}^{(\cdot)} = \ln\left(\frac{d_{ij}}{1-d_{ij}}\right) . \qquad (13)$$

Each of the five sets of features, namely H:MF, P:MF, P+H:MF, P×H:MF and Joint:MF, were used as the input features for the models $\phi_{ij}^{(1)}$, $\phi_{ij}^{(2)}$, $\phi_{ij}^{(3)}$, $\phi_{ij}^{(4)}$ and $\phi_{ij}^{(5)}$ respectively. In practice, we used LASSO for feature selection and regularization. The penalty term parameter for LASSO was determined by using 10-fold cross-validation on the training data. Finally, the prediction classification error, calculated as $\frac{\#\, False\ Positives\, +\, \#\, Fasle\ Negatives}{\#\, Test\ Samples}$, was used to assess the performance for this model. The mean classification error was calculated by taking the mean of classification error from 200 runs.

3.3.5 Model for predicting infection efficiency

Since the EOP values are continuous, neither the zero-inflated Poisson or negative binomial models are appropriate for modeling the outcomes. As a result, we applied a log transformation on the positive EOP values to make the distribution more normal-like. For a given phage-host pair where a successful infection event is present, that is $e_{ij} > 0$, we denote the natural log transformed EOP value as:

$$e'_{ij} = \ln(e_{ij}) . \qquad (14)$$

Shapiro-Wilk test was performed to check the normality of the distribution of e'_{ij}.

Linear regression was used to model the relationship between mutation profiles and the intensity of successful infections in phage-host pairs, that is

$$\phi_{ij}^{(\cdot)} = e'_{ij} \, . \qquad (15)$$

Each of the five sets of features, namely H:MF, P:MF, P+H:MF, P×H:MF and Joint:MF, were used as the input features for the models $\phi_{ij}^{(1)}$, $\phi_{ij}^{(2)}$, $\phi_{ij}^{(3)}$, $\phi_{ij}^{(4)}$ and $\phi_{ij}^{(5)}$ respectively. For the linear model, we also used LASSO for feature selection and regularization. The penalty term parameter for LASSO was determined by using 10-fold cross-validation on the training data. Finally, the MAE was used to evaluate the performance of the model.

3.3.6 Train-validation split and feature evaluation

To assess the performance of different features for the logistic regression model, we performed 200 bootstrap runs to predict the existence of phage infection. Specifically, in each run the training set was generated by randomly select $U \times V$ samples from the entire dataset with replacement. The d_{ij} values that were not selected as training samples form the validation set. As a control, for each run, a null model was built to predict the outcomes by randomly sample d_{ij} values from a Bernoulli distribution $Bern(\,\hat{p})$ where $\hat{p}$ is the maximum likelihood estimator (MLE) of the proportion of successful infection from the training set of that run. After the 200 runs, the training and validation prediction error were compared between pairs of the models including the null model and models based on H:MF, P:MF, P+H:MF, P×H:MF and Joint:MF.

Similarly, we also performed 200 bootstrap runs for the linear model to predict the infection efficiency. Specifically, in each run the training set was generated by randomly sample e'_{ij} with replacement. The size of e'_{ij} sampled as the training set in each run matches the total number of the e'_{ij}. The e'_{ij} that were not selected in the training set forms the validation set. As a control, for each run, a null model was built by always predicting the efficiency of infection as the mean e'_{ij} of the training set for that run. After the 200 runs, the training and validation MAEs were compared between pairs of the models including the null model and models based on H:MF, P:MF, P+H:MF, P×H:MF and Joint:MF.

3.3.7 Final model and predictions

After comparing the training and validation performance of models based on H:MF, P:MF, P+H:MF, P×H:MF and Joint:MF with 200 bootstrap runs, a final model, which contains both the step 1 and step 2 model was constructed. The penalty term parameter for each of the step 1 and step 2 models was chosen as the mean of the best penalty term parameter from each of the 200 bootstrap runs. After model fitting, the predicted outcome $\tilde{d}_{ij}$ for step 1 and $\tilde{e}'_{ij}$ for step 2. For each step of the final models, the importance of feature was measured by the absolute value of coefficients learned from each step.

3.4 Results

3.4.1 The mutation and cross-infection matrices for phage and host

To quantify the relative quantity of plaques formed by a phage strain infecting a host strain, we computed the efficiency of plating (EOP) values for all phage-host pairs sampled during the 37-day coevolution experiments. The EOP value measures the relative quantity of plaques formed by a phage strain infecting a host strain. Details of the EOP calculation are described in the Materials and Methods section. The resulting EOP values exhibited a skewed distribution with 95% of values ranging from 0 to 1.5. At the beginning of the experiment, the ancestral host strain was susceptible to all phage strains (EOP > 0), while at the end of the experiment, the majority of the host samples from day 37 were resistant to all phage strains (EOP = 0) (Figure 22). Overall, the EOP matrix showed the complexity of the observed phenotype from phage-host interactions (Figure 9). A total of 2295 phage-host interaction pairs were observed, among which 913 pairs denoted successful phage infections (EOP > 0) and 1382 denoted unsuccessful infections (EOP = 0). Since the observed positive EOP values span a wide range and has a long-tailed distribution, there was large variance in the observed phenotype in terms of the efficiency of phage infection (Figure 22). For the observed genotypes, the mutation profiles of the host and phage revealed a number of changes in their genomes, including 18 and 176 unique mutations for the host and phage, respectively (Figure 9, Table 5). As a result, we set out to develop a framework that links the changes in phage-host interactions to their respective genotypes.

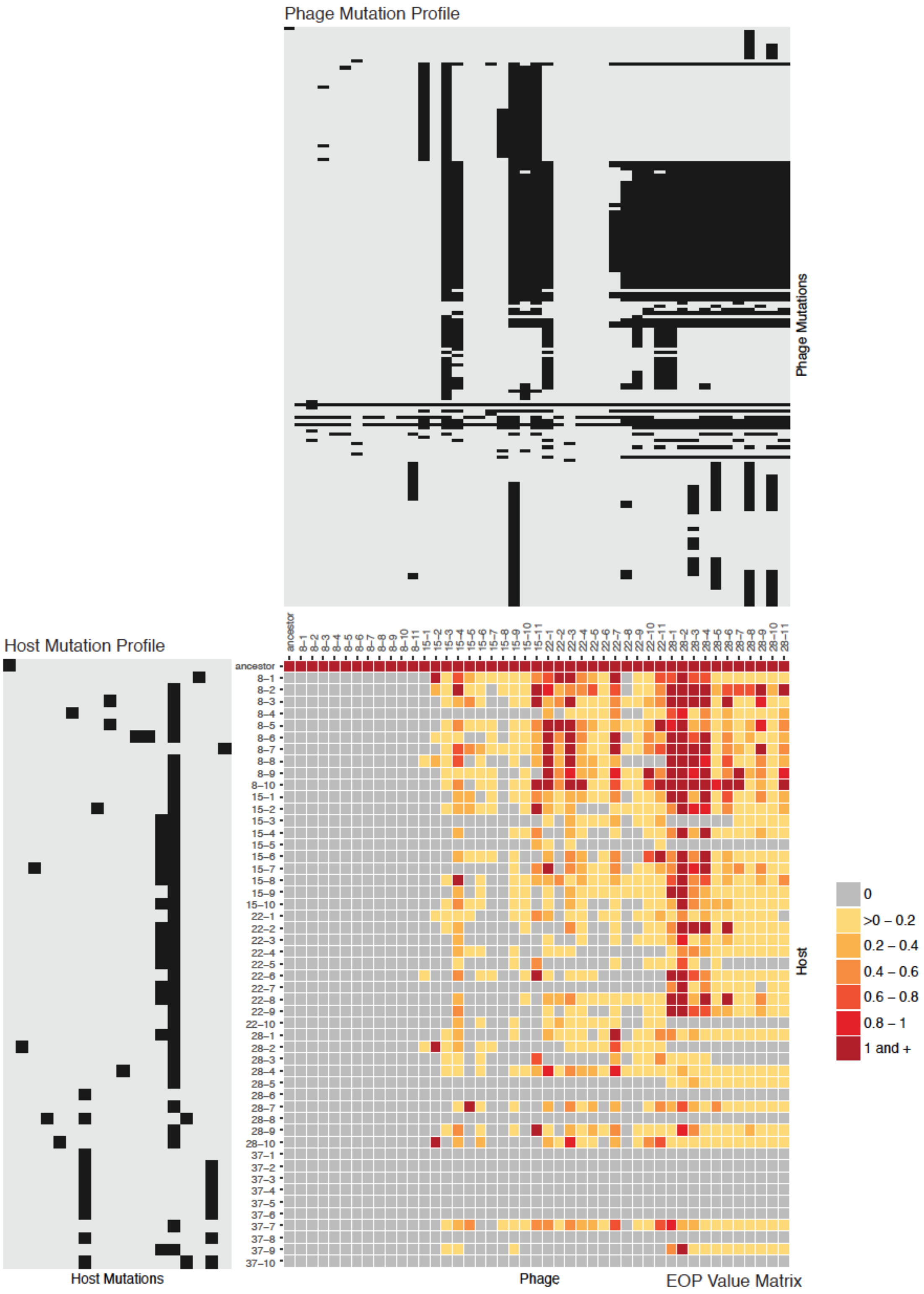

Figure 9 – Heatmaps showing the EOP value matrix as well as host and phage mutation profiles

The upper panel is showing phage mutation profile. The left panel is showing host mutation profile. Black cell indicates the presence of a mutation. Gray cell represents the absence of a mutation. The heatmap is showing the EOP value bands. The color key showing the color and the corresponding EOP value range.

3.4.2 Model for predicting phage-host interaction network

We developed a framework for predicting the effect of genetic mutations on the presence or absence of successful infection (POA) of phage-host pairs embedded in a phage-microbe interaction network. We began with logistic regression models that utilize mutations as features to predict qualitative variation in the infection network, i.e., 'whom infects whom'. We classified different models in terms of the distinct feature sets that underlie predictions, including a host-only mutational feature (H:MF), a phage-only mutational feature (P:MF), and an additive phage and host mutational feature (P+H:MF). All of these models leveraged differences in phage or host genotypes. However, it is possible that combinations of mutations of phage and host act in a nonlinear way to impact phenotype. For that reason we also included the phage-cross-host mutational feature set (P×H:MF) as well as models that include both 'first-order' (phage and host mutations) and 'second-order' (phage-cross-host mutations) effects (i.e., the combined feature set model, Joint:MF). These features were constructed based on the genetic mutation profiles of the host and phage. By comparing the performance of the logistic regression models built based on different sets of features, we found that the additive phage and host model (P+H:MF) outperforms all other features on the validation set ($P < 9.44e-5$) with a mean classification error of 15.07% (Figures 10 and 23). Our results showed that the P+H:MF contains the best set of features for predicting the POA for a given phage-host pair. One explanation for this result could be that each of the important mutations that occurred during the coevolution process have sufficiently large effect size to impact the presence or absence of the interaction. Overall, we built a final model based on P+H:MF for step 1 (Figure 11). Feature importance analysis revealed 7 host mutations and 27 phage mutations that were shown to have a positive effect on the observed phage infection, comparing with 5 host mutations and 15 phage mutations that were shown to have a negative effect (Table 7).

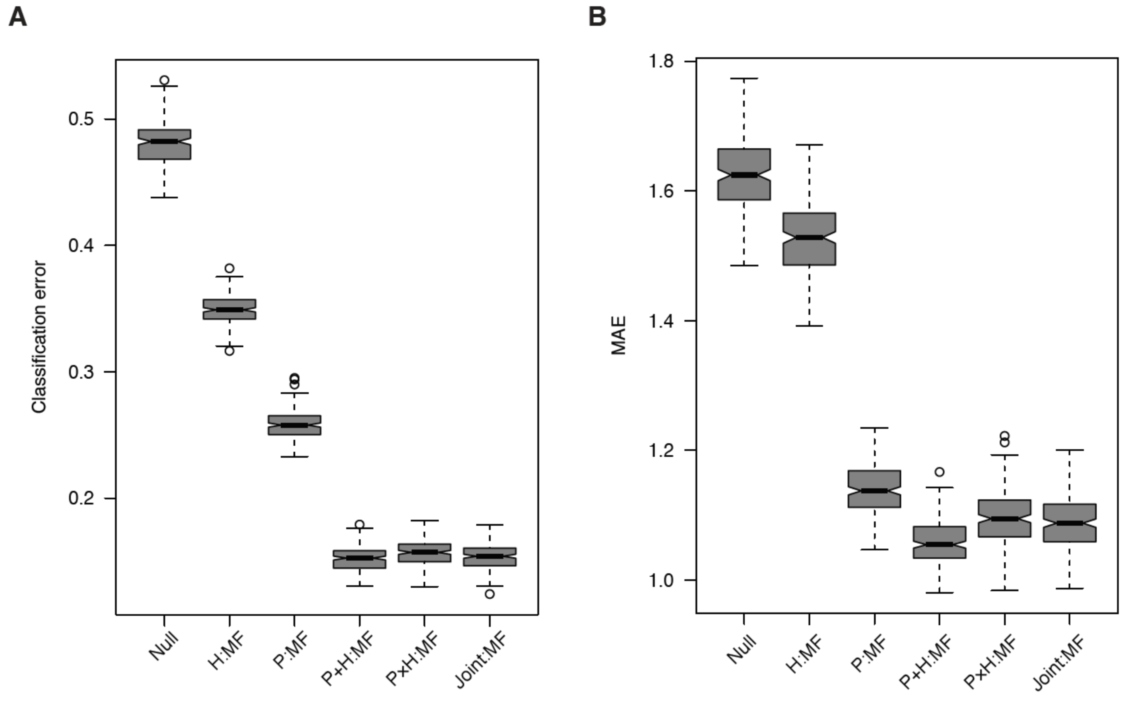

Figure 10 – Model performance for different feature sets on validation set

(A) Boxplot for validation set classification error for step 1 on 200 bootstrap runs for null model and models based on H:MF, P:MF, P+H:MF, P×H:MF and Joint:MF. (B) Boxplot for validation set MAE for step 2 on 200 bootstrap runs for null model and models based on H:MF, P:MF, P+H:MF, P×H:MF and Joint:MF.

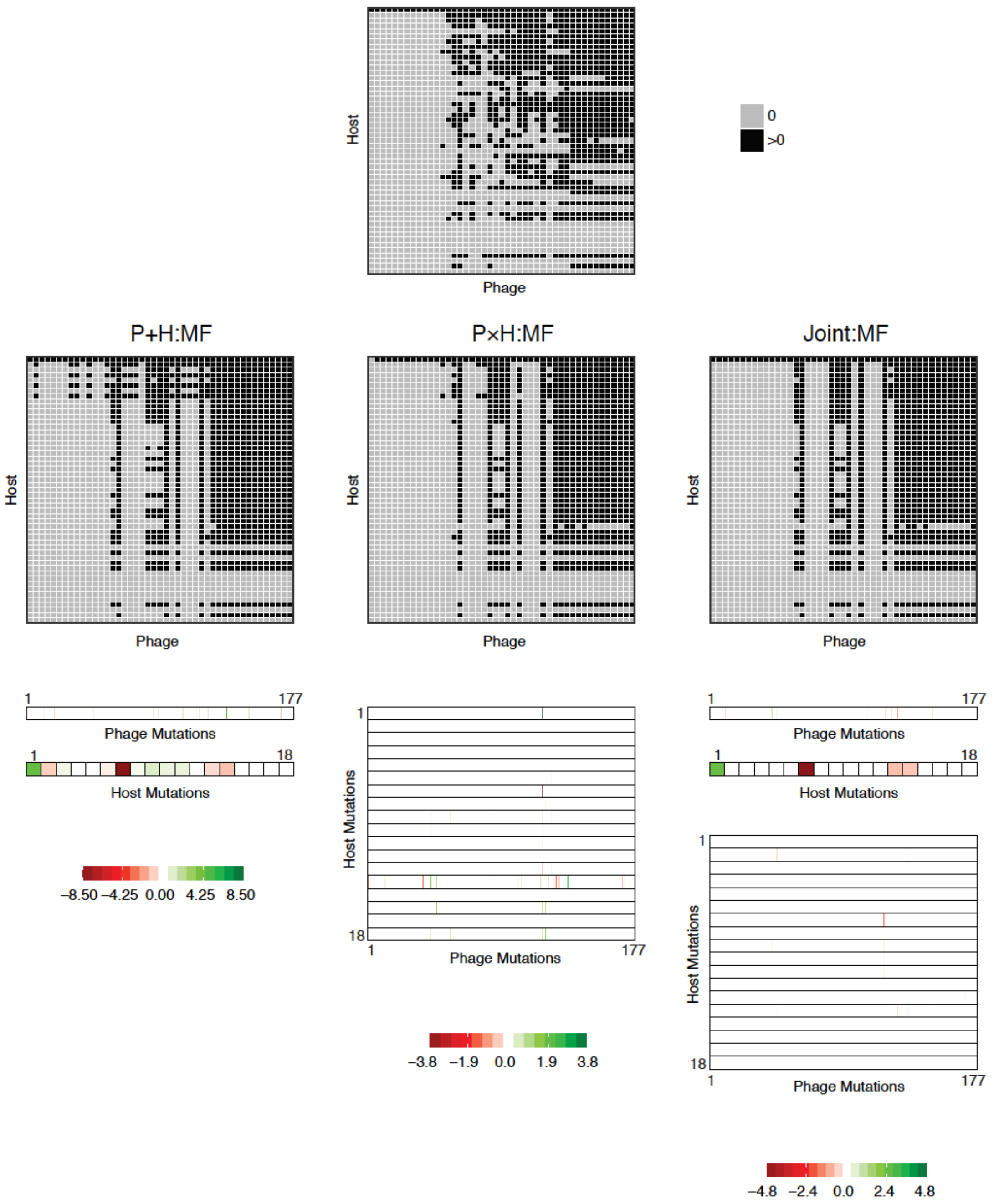

Figure 11 – Results from final model for step 1 based on P+H:MF, P×H:MF and Joint:MF

Top panel: The true phage-host interaction network based on observed EOP from experiment. Middle panel: The predicted interaction network based on P+H:MF, P×H:MF and Joint:MF, respectively. Bottom panel: The coefficients learned from the P+H:MF, P×H:MF and Joint:MF features, respectively.

3.4.3 Model for predicting the efficiency of infection

As the next step in our framework, we extended the prior prediction framework so as to identify phage and host mutations that have large impacts on the efficiency of phage infection (EFF) with the presence of phage infection (EOP > 0). Since the log-transformed positive EOP values followed a normal distribution (P = 3.283e-8), here we used linear regression to model the quantitative impact of mutations on EFF (Figure 24). We examined models based on five sets of features, namely the H:MF, P:MF, P+H:MF, P×H:MF and Joint:MF. Model performances were compared based on the validation mean absolute error (MAE). The results showed that the linear regression model with the additive feature set (P+H:MF) gives the lowest validation MAE (P < 3.95e-14) with median MAE to be 1.05 (Figures 10 and 23). Overall, we built a final linear model based on P+H:MF for step 2 (Figure 12). Feature importance analysis revealed that there were 7 host mutations and 34 phage mutations that were shown to have a positive effect on promoting the efficiency of phage infection, compared with 7 host mutations and 33 phage mutations that were predicted to have a negative effect (Table 8).

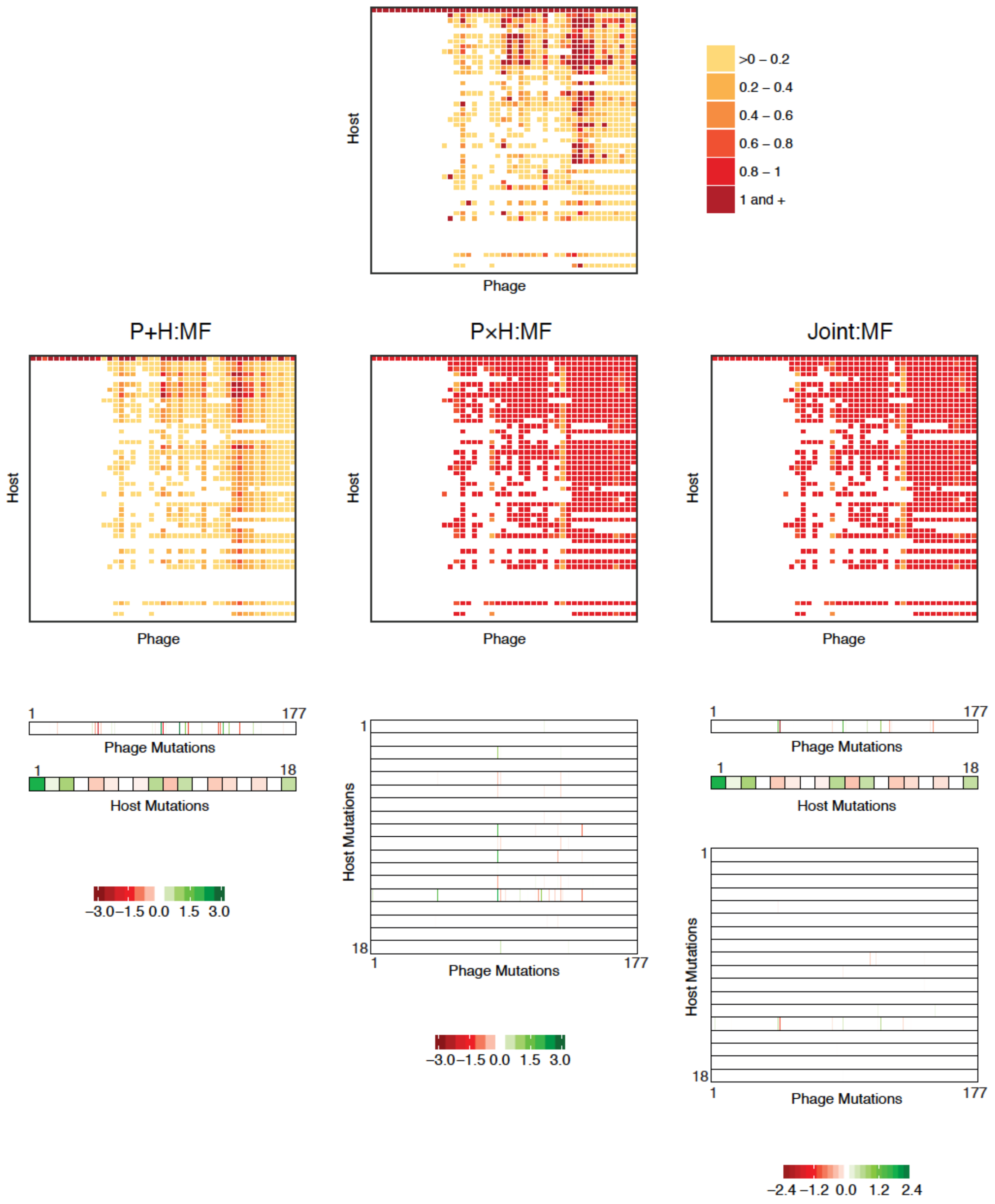

Figure 12 – Results from final model for step 2 based on P+H:MF, P×H:MF and Joint:MF

Top panel: The true phage infection efficiency based on the observed positive EOP from experiment. Middle panel: The predicted infection efficiency based on P+H:MF, P×H:MF and Joint:MF, respectively. Bottom panel: The coefficients learned from the P+H:MF, P×H:MF and Joint:MF features, respectively.

3.4.4 Molecular mechanism behind the important features

Several putatively important mutations were revealed by the feature analysis using final predictive models for step 1 and step 2 (Figure 25). The top five important features that contributed to the increase of POA includes the indicator variable for the ancestral host strain, one point mutation in the phage *S* gene region, two mutations in the phage *J* gene region and one mutation in the *bor* gene region. For the decrease of POA, the top five important features included a 16 *bp* deletion in the host *manXYZ* gene region, three point mutations in the phage *J* gene region and one point mutation in the phage intergenic region between the lambda *p79* gene and the end of the genome. Similarly, the top five important features that contributed to the increase of EFF includes the indicator variable for the ancestral host strain and four mutations on the phage *J* gene region. For the decrease of EFF, the top five important features included one mutation in the intergenic region between *bor* and lambda *p78* gene region and four mutations in the phage *J* gene region.

A 16 *bp* deletion was found to be the most important feature for predicting POA, but was not found to be important for predicting EFF. The mutation profile table showed that this mutation was shared by 10 host strains, 2 of which were sampled from day 28 and 8 were from day 37. These 10 host strains were super-resistant, that is, the 10 host strains were resistant to the ancestral strain and all the phage isolates from the experiment. This mutation was located in the region of the host *ManXYZ* gene, which encodes the PTS mannose transporter subunit IID. This protein could be exploited by the phage to inject their DNA into the host. Our findings were consistent with a previous study that showed that the mutations in *ManXYZ* lead to the host super-resistant phenotype [66].

Another important feature was the ancestor indicator variable that was found to be important for the increase of both the POA and EFF. This was consistent with the fact that the ancestral host strain is susceptible to the ancestral phage strain as well as all the phage samples collected during the experiment. Finally, several mutations located in the phage *J* gene region were found to be important for both POA and EFF. The *J* gene encodes the tail fiber of phage λ which participated in the process of injecting phage DNA into the host. Thus, it played an important role in the host-phage interaction and the mutations in the J gene region could have a large impact on phage-host interaction [120, 137, 138]. This was consistent with our model predicting the mutations to be important for both POA and EFF.

3.5 Discussion

In this study, we developed a computational framework for predicting the network and efficiency of phage-host interactions by linking phenotypes with the genetic mutation profiles of both phage and host. The basis for our inference was an assumption that mutations can contribute directly, or via mutational-interactions, to changes in phenotype. Our comparative analysis revealed that an additive model that incorporates mutational effects of phage and host separately had the highest predictive value in linking genotype to phenotype. In doing so, the framework identified gene regions already recognized in mediating phage-bacteria infections for bacteriophage λ and *E.* coli. The model also identified important features that were located in gene regions that could potentially participate in phage-host not previously known to contribute to the phage-host interaction. Hence, the framework has the potential to identify novel genes and mutations that modulate virus-microbe interactions.

For example, based on the feature importance analysis, we identified one mutation located in the phage *S* gene region that is found to be uniquely important for predicting the presence (or not) of infection. This gene encodes holin which is a small inner membrane protein required for phage-induced host lysis [139]. Notably, the phage-host interaction network observed in our experiment is based on the quantitative plaque assay, in which clearings (plaques) would appear where bacterial cells were infected and lysed by the phage [140, 141]. Thus, it was possible for the mutation in the S gene to have a direct impact on the lysis of the host cells, which would then have an impact on the final observed phenotype.

Another mutation that occurred in the phage *lom* gene region was exclusively important for the quantitative infection efficiency. The *lom* gene encodes an outer membrane protein that is putatively associated with the host's ability to adhere to human buccal epithelial cells [142]. Although this protein is not currently known to be directly involved in the process of phage infecting the host, the fact that it encodes an outer membrane protein and that it has an impact on the host phenotype suggest that it could have potential role in the phage-host interaction.

Although our analysis suggested that individual mutations act independently, rather than together, to determine infection outcome, we recognized that this finding may reflect the nature of our training and test sets. During the model construction, regularization terms were used for each of the five models built based on H:MF, P:MF, P+H:MF, P×H:MF and Joint:MF. At the training stage, P+H:MF did not outperform the P×H:MF and Joint:MF models both in step 1 and step 2. However, at the test stage, the P+H:MF model outperformed both the P×H:MF and Joint:MF models. Nevertheless, it was possible that the performance of P×H:MF and Joint:MF models was limited by the number of samples observed. There were many possible combinations of phage-host mutation pairs in the feature space of P×H:MF and Joint:MF, but majority of them were not observed. Although expanding the feature space allows the model to capture the interaction between host and phage mutation pairs, however, when more features were introduced to the linear model, due to the limited number of samples, the system became under-determined. Even with the penalty terms, the solution was still suboptimal. It may be worthwhile to consider the P×H:MF or Joint:MF models in future work, particularly given a larger number of samples.

Our inference framework could detect the importance of previously identified adaptive mutations that modify phage-host interactions. However, we must be cognizant of the potential for both false positives and false negatives. False detection may arise due to evolutionary effects including genetic hitchhiking of neutral mutations, recombination, and identification of adaptive mutations that are unrelated to the infection process. Moreover, we did not expect the identification of adaptive mutations to be comprehensive. We linked genotype to phenotypic changes arising in a specific coevolutionary process as measured by a subset of clonal phage and host isolates, hence there will be significant regimes of mutational space left unexplored.

In summary, we have developed a framework for predicting genotypic drivers of both the qualitative and quantitative nature of host-pathogen interactions. In doing so the framework recapitulated the finding of mutations known to influence infection outcome as well as novel sites. In doing so, this framework could help prioritize molecular work to identify novel drivers of infection. Although we applied this framework in the context of phage-bacteria coevolutionary dynamics, the data-driven approach does not necessarily require prior knowledge on specific genes or mutations and can be applied to other host-pathogen coevolution systems as well.

CHAPTER 4. GENOME SEQUENCING REVEALS A DISCONNECT BETWEEN COEVOLUTIONARY PATTERN AND PROCESS

This chapter is being prepared for publication as: Animesh Gupta, Shengyun Peng*, Chung Yin Leung, Joshua S. Weitz and Justin R. Meyer. 'Genome sequencing reveals a disconnect between coevolutionary pattern and process'. Animesh Gupta*, Shengyun Peng* contributed equally. Shengyun Peng performed bioinformatics analysis for the coevolution dynamic and phylodynamic analysis, including constructing the host and phage mutation profile from raw sequencing data, performing the time-shift analysis and reconstructing the phylodynamic trees, running tests for selection on phage samples and jointly proposing the leap-frog dynamic hypothesis.*

4.1 Abstract

New analytical techniques have revealed that ecological networks, whether they are between antagonists like hosts and parasites or cooperators like pollinators and flowers, possess similar nonrandom patterns. The first step to understanding why these network structures exist is to understand how they evolved in the first place. Here we studied *E. coli* and bacteriophage λ's coevolution under controlled laboratory settings. The experiment was initiated with isogenic strains, but they rapidly evolved to form a rich interaction network. Like most phage-bacterial interactions networks (PBINs), the structure was nested such that the host-range of an ancestral phage fell within the more derived genotypes. This pattern has been predicted to occur through arms race dynamics, where bacteria gain ever increasing resistance and phages expand their host ranges to infect the

resistant bacteria. Full genome sequencing revealed a much more complex progression. Multiple lineages of the bacteria and phage coexist and the lineages that dominate late in the arms race evolve from cryptic subpopulations rather the dominant lineage. These findings help resolve the mechanisms underlying PBIN structure and provide a cautionary example of the pitfalls with applying parsimony to interpreting evolutionary process from pattern.

4.2 Introduction

Phage and their bacterial hosts are ubiquitous in nature and play a key role in regulating microbial ecosystems [37, 89, 143, 144]. These viruses have multifaceted effects: Phages drive mortality which can regulate bacterial population size and enhance nutrient cycling [145, 146]. The mortality also triggers bacteria to evolve resistance through a number of mechanisms including resistance mutations or even the develop diverse anti-phage defense systems, including CRISPR-Cas and restriction-modification proteins [147]. The proliferation of defense strategies can impact bacterial diversity [84, 148], which can feedback to trigger the evolution of phage counter defenses and drive their diversification too [147, 149-151]. As a result, such interactions between antagonistically coevolving host and phage can drive the formation of complex interaction networks [58, 59]. These eco-evolutionary dynamics often have profound impacts on the larger ecosystems the microbes are embedded in [22-24].

One common way to study the complex networks that develop between phage and bacteria is to construct a phage-bacteria interaction networks (PBINs) [37]. PBINs are bipartite matrices with values that describe how well each phage can infect each bacterial strain. The data for the matrices is typically collected by challenge experiments, where an array of different hosts is subjected to infection by an array of phage types. PBINs have been used to generate hypothesis for the types of coevolutionary dynamics that occur between phages and their hosts. For example, the most common PBIN structure observed is called nested [59], where phage host ranges fall one within another like a set of Russian dolls. Nested structures are thought to arise from arms race dynamics (ARD) where bacteria evolve resistance, and phages counter by expanding their host-range to include the new resistant type [152, 153]. Phage continue to evolve towards a broader host-range (and, similarly, host towards increasing the number of phages they are resistant to) giving rise to nestedness [154, 155].

An alternative structure is modular where phages have more specialized host-ranges. The phrase modularity stems from the observation that groups of phages tend to infect the same bacteria creating dense clumps of interactions in the network. Modularity is thought to arise from an alternative coevolutionary sequence known as fluctuating selection dynamics (FSD) [74-76, 152]. Under this dynamic, bacteria evolve resistance and when the phage counters it, it loses infectivity on other bacteria, resulting in narrow host-ranges. The dynamic is fluctuating because a range of hosts and phages can be maintained by negative frequency-dependent selection that leads to kill-the-winner fluctuations [156]. While ARD and FSD are two examples, the patterns in the network can be more complex and even share characteristics of both [157].

The different coevolutionary dynamics are thought to arise from the underlying genetic architecture of their interactions. ARD is commonly referred to as gene-for-gene because it is thought that the interaction between the phage and bacteria depends on a number of different genes. Bacteria evolve resistance through disrupting one locus, and then phage respond by not requiring that locus for infection. By reducing the number of host genes required for infection the coevolved phages will be able to infect the contemporary and ancestral bacterial genotypes. FSD is often called allele-for-allele (or matching alleles) because it is thought that this type of coevolution occurs when the interaction is controlled by a single locus. For example, the bacteria could evolve resistance by altering the phage receptor to deflect infection, and then the phages could evolve to exploit the new receptor at the cost of losing function on the ancestral form. This is often referred to as lock and key dynamics, where there are specialized keys that open specific locks.

Ideally, in order to determine how different coevolutionary dynamics yield different PBINs, times series of the changing interactions would be measured, as well as full genome sequencing to determine the genetic architecture of their coevolution. Previous studies have used phenotypic assays to determine how host range and resistance change over coevolutionary time [158]. Others have attempted to analyze the genetic basis of coevolution by linking mutations in the host or virus to resistance or host range expansion, respectively [159]. To the best of our knowledge, no study has measured PBINs and also sequenced full genomes of both the host and bacteria.

To provide a more comprehensive understanding of the formation of PBINs, we measured the changes in cross-infectivity using pairwise quantitative plaque assay amongst 51 host and 45 phage strains sampled at different times in a 37-day coevolution experiment. We constructed the PBINs to identify if they show any patterns of modularity on nestedness and then confirmed the type of coevolutionary dynamics at play using time-shift analysis.

We also sequenced the whole genomes of isolated phage and host strains to understand the genomics of coevolution.

4.3 Methods

4.3.1 Experimental setup and sample isolation

Meyer et. al [40] performed the original coevolution experiment with the strain REL606 of *Escherichia coli* B and an obligatory lytic strain of λ. Both, *E. coli* and λ, were co-cultured in a carbon-limited minimal glucose media at 37°C and allowed to evolve for 37 days by transferring 1% of the community to fresh medium at the end of each day. Periodically, 2 ml of community was preserved by adding ~15% of glycerol and freezing the mixture at -80 °C.

We randomly isolated ten host and eleven phage clones from frozen stocks of a population from days 8, 15, 22, 28 and 37. In total, 50 strains of *E.* coli and 44 λs were isolated from the coevolution experiment (no phage were detected on day 37). To isolate bacterial clones, a small amount of frozen population was diluted in 0.9% *wv* sodium chloride solution and then spread on a Luria-Bertani (LB) agar plates [41]. The plates were then incubated at 37 °C for 24 h to pick individual colonies. The picked colonies were re-streaked and grown two more times on LB agar plates in the same manner to get rid of any phage particles. Finally, ten colonies from each day-timepoint were picked at random and grown overnight at 37 °C to run pairwise infection assays. These isolated clones were also preserved with ~15% of glycerol at -80 °C.

Phage clones were isolated by first mixing appropriate dilution (in sodium chloride) of frozen community with 4 ml of molten (~50 °C) soft agar (LB agar except with only

0.8% *wv* agar) and ~5 x 10^8 cells of bacterial strain REL606, and then pouring the mixture over an LB agar plate. The plates were dried and incubated overnight at 37 °C to pick 11 individual plaques at random. Clonal phage stocks were made by growing these picked plaques overnight with ~5 x 10^8 bacterial cells in 4 ml of the evolution medium shaken at 220 rpm and 37 °C. Stocks were created the next morning by removing cells with centrifugation and treatment with 100 µl chloroform. 2 ml of phage was also preserved with 15% of glycerol at -80 °C.

4.3.2 Pairwise infection assays

Pairwise quantitative infection assays were performed for all the combination of host strains and phage strains isolated (online Supplemental Table 1 at https://github.com/speng32/thesis_supp_files). Specifically, 7 serial 1/10th dilutions were made of each phage culture. 2 µl of each dilution plus the full-strength phage stock was spotted on top of *E. coli* lawns. Bacterial lawns were made for every single genotype and REL606, meaning 17,952 spots were plated. Efficiency of plaquing (EOP) was calculated as the phage density calculated on a coevolved isolate divided by the density calculated on the sensitive REL606 ancestor. This method provides a quantitative measurement for the infectivity of a given phage on a specific host.

4.3.3 Analysis of Nestedness and Modularity

BiMat [111] was used to assess the nestedness of the PBIN. The raw EOP value matrix was binarized into 0 for EOP = 0 and 1 for EOP > 0. Two preprocessing setting were applied on the input EOP matrix. In the first setting (setting 1), the rows and columns that contain all zeros were removed. In the second setting (setting 2), a row with all 1's was added to the EOP value matrix to represent that the ancestral host strain can be infected by all phage strains. *BiMat* was ran with each of the two preprocessed EOP matrix as input with default settings and revealed qualitatively similar results. Here we report on results from setting 1.

4.3.4 Resistance and infectivity calculation and statistical test

For a total number of n host samples and m phage samples, we denote the EOP value for the ith host sample against jth phage sample as e_{ij} where $i \in [1, n]$ and $j \in [1, m]$. We denote the five checkpoint days of day 8, 15, 22, 28 and 37 for host by $k,$ where $k = 1,2,3,4,5$, and the four checkpoint days of day 8, 15, 22 and 28 for phage by l where $l = 1,2,3,4$. Host resistance for a host sample i is calculated as

$$r_i = \sum_{j=1}^{m} \mathbf{1}_{\{e_{ij}>0\}}, \quad (16)$$

which measures the number of phage strains that the host is resistant to. The host range of a phage sample j is calculated as

$$h_j = \sum_{i=1}^{n} \mathbf{1}_{\{e_{ij}>0\}}, \quad (17)$$

which measures the number of host strains that the phage can successfully infect. The resistance percentage for each checkpoint of host is calculated as

$$RP_k = \frac{\sum_{i \in A_k} r_i}{m \times |A_k|}, \quad (18)$$

where A_k denotes the range of the host sample that belongs to the kth checkpoint and $|A_k|$ denotes the cardinality of the set A_k, i.e. the number of host samples at the kth checkpoint. The host range percentage for each checkpoint of phage is calculated as

$$HP_l = \frac{\sum_{j \in B_l} h_j}{n \times |B_l|}, \quad (19)$$

where B_l denotes the range of the phage sample that belongs to the lth checkpoint and $|A_k|$ denotes the cardinality of the set B_l, i.e. the number of phage samples at the lth checkpoint.

To evaluate the changes of in the resistance of host and the host range of phage, we used Analysis of Variance (ANOVA) to compare these measurements across different sampling days.

4.3.5 Time-shift analysis

We performed time-shift analysis to compare the mean EOP values of samples when they interact with their past, contemporary and future counterparts. For the host sample i, the average EOP value from interactions with phages from checkpoint l is calculated as

$$EB_{il} = \frac{\sum_{j \in B_l} e_{ij}}{|B_l|}. \quad (20)$$

Each data point on the host time-shift curve represents an EB_{il} value and the values from the same host were connected with dotted lines. For the phage sample j, the average EOP value from interactions with hosts from checkpoint k is calculated as

$$EP_{jk} = \frac{\sum_{i \in A_k} e_{ij}}{|A_k|}. \quad (21)$$

Each data point on the phage time-shift curve represents an EP_{jk} value and the values from the same phage were connected with dotted lines.

To test if there is a significant increasing trend in the host time-shift curves, we performed one-sided paired t-test by comparing the average EOP values from the last phage checkpoint – day 28 – against that from each previous checkpoint, namely day 8, 15 and 22. Similarly, to test if there is a significant decreasing trend in the phage time-shift curves, we also performed one-sided paired t-test by comparing the average EOP values from the initial host checkpoint – day 8 – against that from each later checkpoint, namely day 15, 22, 28 and 37.

4.3.6 Whole Genome Sequencing for λ and E. coli clones and pre-analysis

4.3.6.1 Preparing clonal λ stocks for DNA extraction

λ clones from each timepoint were revived by growing ~3 µl of frozen stocks overnight with 100µl of ~5x10^9 cells of strain DH5α (a *E. coli* K-12 derivative) at 37 °C in 4 ml of LBM9 medium shaken at 220 rpm supplemented with 40 µl of additional 1M magnesium sulphate to facilitate λ growth, where LBM9 is 10 g tryptone, 5 g yeast extract, 12.8 g sodium phosphate heptahydrate, 3 g potassium phosphate monobasic, 0.5 g sodium chloride, 1 g ammonium chloride, 1.2 g magnesium sulphate, 11 mg calcium chloride per L water. 100 µl of chloroform was added to the overnight cultures to kill the host cells, and then centrifuged at 3900rpm for 10 min to pellet the cells and debris. λ lysates obtained were filtered and stored at 4°C with 2% chloroform. 10 µl of these λ lysates were again grown overnight with DH5α in the same manner to propagate high phage densities for genomic DNA extraction. Final λ stocks were obtained by centrifuging the overnight λ cultures at 3900 rpm for 10 min and then filtering it with 0.22 µm filter tips to remove all cells and debris.

4.3.6.2 Removal of any bacterial DNA

Any remaining bacterial DNA was first removed from λ stocks before extracting λ DNA. 1 mL of the λ stocks was added to 200 µL of ice cold L2 buffer (PEG6000/NaCl from TekNova Cat #P4168) in 1.5 ml centrifuge tubes and mixed well by inverting the tubes. These were incubated for 1 h before centrifuging tubes at 4°C for 10 min at 12,000 g. Supernatant was discarded, and tubes were dried by inverting for 10 min. 100 µl of DNase solution (65 µl molecular biology grade water with 10 µl of 10x DNase I buffer and

25 µl of DNase I (RNase free) from New England Biolabs) was carefully pipetted into the tubes to resuspend the pellets. The suspended solution was incubated for 1 hr at 37°C before a heat shock of 10 min at 75°C after which tubes were placed on ice before extracting DNA.

4.3.6.3 Extraction of λ genomic DNA

We used Invitrogen's PureLink Pro 96 Genomic DNA kit (Catalog no. K1821-04A) to extract λ genomic DNA. Purified λ from above was transfer into wells of 96 Deep Well Block provided in kit and kit protocol was followed from step 3 of 'Preparing lysates for gram negative bacterial cells'.

4.3.6.4 Preparing clonal E. coli stocks for DNA extraction

E. coli clonal stocks were revived by growing ~3 µl of frozen stocks overnight in LB.

4.3.6.5 Extraction of E. coli genomic DNA

Invitrogen's PureLink Pro 96 Genomic DNA kit (Catalog no. K1821-04A) was used to extract genomic DNA from overnight cultures of *E. coli* clonal stocks.

4.3.6.6 Preparation of genomic library and sequencing

We used ref. [46] for both *E. coli* and λ to prepare genomic libraries. Sequencing was done at UC San Diego IGM Genomics using paired-end Illumina HiSeq 4000 platform.

4.3.6.7 Pre-analysis of sequenced reads

After collecting the raw reads, the adapters were removed using cutadapt [160] and quality control (QC) was performed for each isolated strain using FastQC [161].

4.3.7 Mutation profile tables for isolated host and phage clones

The QC filtered sequencing reads were then analyzed using the *breseq* (v0.32.1) pipeline [162]. We ran the pipeline in the consensus mode with default parameters except for the consensus-frequency-cutoff, which was set to 0.5. The *breseq* pipeline first aligns the reads to the reference genome using bowtie2 [163]. It then analyzes the mapped reads to identify mutations based on new junction, missing coverage and read alignment evidences. Finally, it generates a summary mutation profile table with a list of mutations and corresponding evidence (online Supplemental Table 2 at https://github.com/speng32/thesis_supp_files). The same *breseq* settings were used to analyze both host and phage data.

4.3.8 Test for selection on phage samples

The D_N/D_S ratio was computed for phage whole genome as well as phage J protein region to test for the presence of selection [164, 165]. We only performed this test for phage since their evolution was dominated by nucleotide substitutions in protein coding genes, and the host mutation profiles consisted of many large indels and intergenic changes. To compute the D_N/D_S ratio, a pseudo count of $\alpha = 0.5$ was added to both the D_N and D_S counts to avoid dividing by zeros.

4.3.9 Phylogenetic reconstruction

Due to the prevalence of large insertions and deletions in the host genomes, conventional nucleotide substitution models were not suitable for estimating the host phylogenetic tree. However, such models are still suitable for estimating the maximum-likelihood phylogenetic tree for phage genomes. As a result, two different approaches were taken to reconstruct the evolutionary trajectories of the host and virus.

To construct the phage phylogeny, multiple sequence alignments were performed for all recovered genomes and the ancestral genome using *mafft* (v7.305b) [166] with default settings except that retree was set to 2 and maxiterate was set to 1000. A maximum likelihood tree was constructed using *raxml-ng* [167]. We performed root-to-tip regression analysis to confirm the existence of temporal signal in the maximum likelihood tree (Figures 29 and 30). This was done by regressing tip distance from the root against the sample time. The significance of correlation between tip distance from the root and the sample time was evaluated by comparing the observed with the null distribution of coefficient of determination (R^2). The null distribution of R^2 was generated by randomly permuting the sample times for 500 times. Finally, the *TreeTime* [168] program was used to generate the phylogenetic tree.

To reconstruct the host evolutionary trajectory, a pairwise Hamming distance matrix was first computed using the R packages *e1071* and *phangorn* [169]. Specifically, the hamming distance between a pair of host genomes was calculated as the number of different mutations from the two genomes. This approach is different from the approaches used by nucleotide substitution models where each base pair change in the two genomes was counted as a single mutation event. The neighbor-joining (NJ) trees were then built based on the hamming distance matrix using *T-REX* [170]. Similar root-to-tip regression analysis was performed to confirm the temporal signal as described in the previous paragraph. Finally, the *TreeTime* program was used to build the host phylodynamic tree.

4.3.10 Genomic analyses of whole community from Day 8

120 µl of frozen stock of whole community was grown for 24 h in 10 ml of media similar to the original coevolution experiment [40] to revive the population. Phage and

bacteria were then separated, and their genomic DNA was extracted in the same manner as described above for clonal stocks. Genomic library was prepared using NexteraXT kit at UC San Diego IGM Genomics. IGM also sequenced the samples using 75 base single reads on the Illumina HiSeq 4000 platform. *breseq* v0.32.0 was used to analyze whole population sequencing data of Day 8. We ran *breseq* in polymorphism mode with default settings to construct the mutation profile tables.

4.4 Results

4.4.1 Coevolutionary changes in resistance and infectivity

To study the coevolutionary arms race between *E. coli* and λ, we quantified changes in cross-infectivity amongst multiple host and phage strains sampled at different timepoints from the coevolution experiment (Figure 8). We isolated ten host and eleven phage clones from populations preserved at Day 8, 15, 22, 28 and 37 (no phage at day 37 due to extinction), and performed quantitative pairwise plaque assays between them (online Supplemental Table 1 at https://github.com/speng32/thesis_supp_files). The cross-infection matrix revealed a complex but ordered pattern of nestedness as is typically observed in most phage-bacterial interaction networks (PBINs) (Figure 13) [43]. Additionally, we did not uncover evidence for a modular pattern based on *bimat* result (data not shown). The ordered pattern of nestedness emerges when an arms race between bacteria and phage leads to bacteria evolving resistance and phage evolving counter-resistance to it while retaining the ability to infect the previous sensitive host.

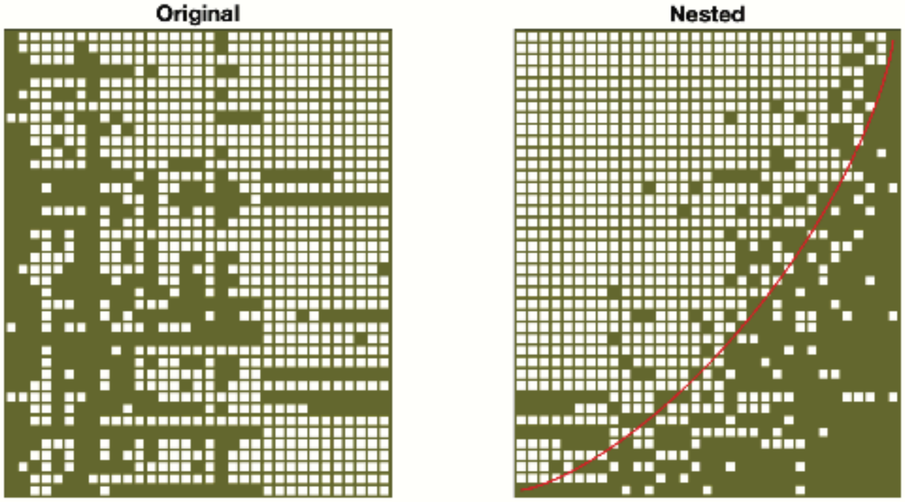

Figure 13 – Phage (columns) and bacterial (rows) interaction network

The Filled squares indicate a combination of host and phage that result in successful interactions. The original network was reassembled to maximize nestedness using the software BiMat. The red line highlights the isocline using the NTC algorithm. The nestedness value of the network based on NODF algorithm is 0.839. Null models based on 200 random shuffles have a mean of 0.638 and std of 0.011.

Note that although all isolated hosts on Day 8 were resistant to all Day 8 phage clones (Figure 14), the phage population did not go extinct in the coevolution experiment due to “leaky-resistance” of host [42]. This is a phenomenon where a small fraction of susceptible host cells is maintained because of a high rate of genetic reversion from resistant to susceptible. The reversion rate is high enough to sustain the phage population through daily serial dilution transfers, but lower than what we can sample from picking individual colonies. Eventually, resistance levels had reached such high levels and the reversion rate was low enough that the phage went extinct sometime between days 28 and 37.

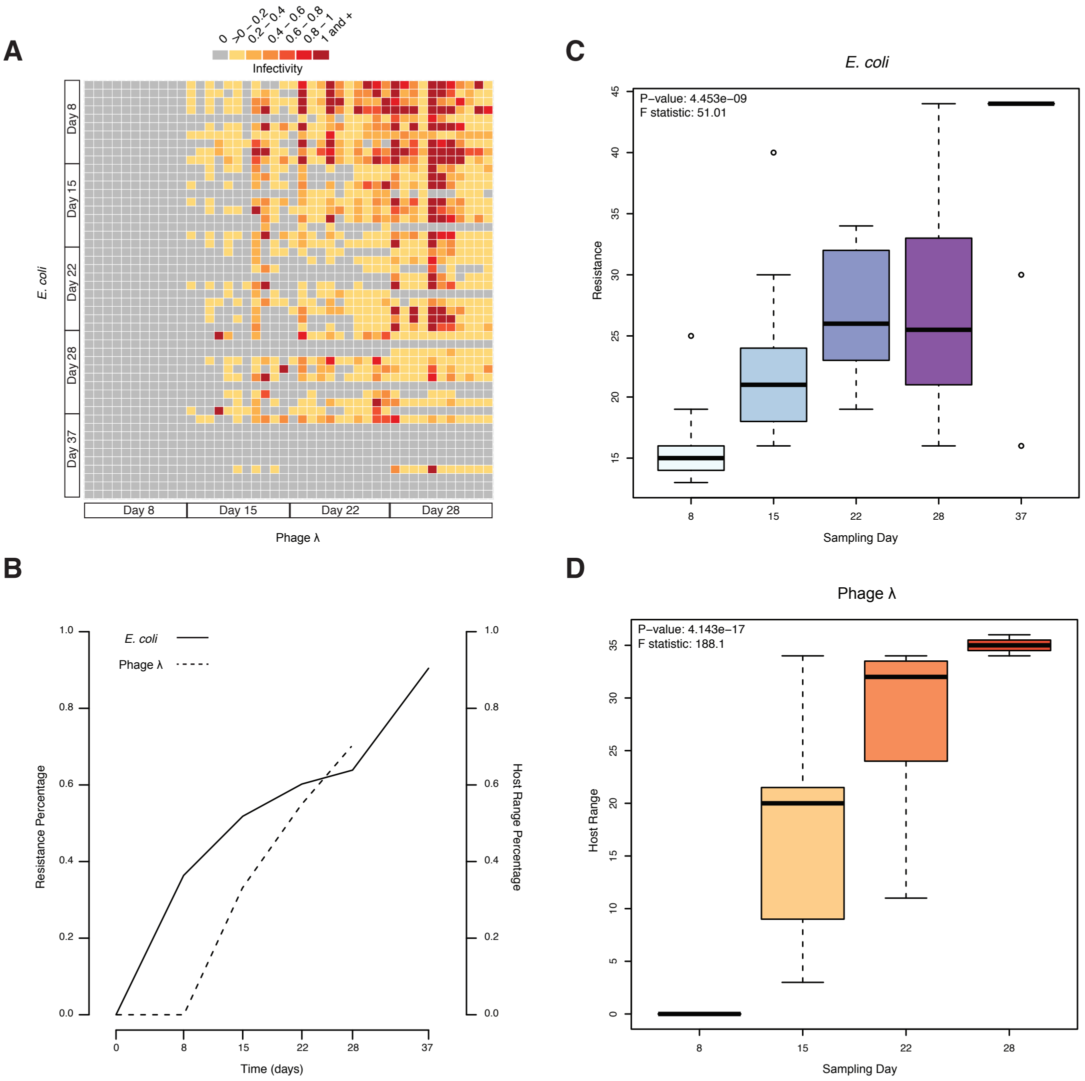

Figure 14 – Host resistance and phage infectivity measured by pairwise plaque assay

(A) Heatmap showing the plaque assay result where grey cells represent no infection, yellow represents low infectivity and red represents high infectivity. (B) Line plot showing the resistance percentage of host and the host range percentage of phage at each checkpoint. (C) Boxplot showing the average resistance of hosts from the same sampling day across five checkpoints. (D) Boxplot showing the average infectivity of phages from the same sampling day across four checkpoints. The statistical significance of the difference between the average resistance and host range from different checkpoints were evaluated using ANOVA.

In line with the nested pattern, Figure 14B shows the average increase in host-resistance by *E.* coli and average increase in host range by λ with time. For *E. coli*, the resistance percentage – the proportion of host genomes from a given sampling day that are resistant to infection –increases monotonically as the coevolution time increases; and for λ too, the host range percentage – the proportion of host genomes that can be infected by phage sampled at a given day – also increases with time. ANOVA results show that the resistance of the host increased significantly (P = 4.453e-09, F = 51.01) during the experiment (Figure 14C). Similarly, by comparing the infectivity of the phage samples from different days, we also observe significant changes (P = 4.143e-17, F = 188.81) in host-range (Figure 14D) over the course of the coevolution experiment.

4.4.2 Time-shift analysis and signatures of coevolutionary dynamics

To further dissect the complex network of cross-infection, we zoomed in on each sampling day and performed a time-shift analysis on host and phage clones isolated from that day against their counterparts from the past, contemporary and the future. Specifically, we compared the EOP values that quantifies the interaction between hosts and phage isolated from any two given days. A higher EOP value implies lower host resistance or higher phage infectivity. A mean EOP value was calculated for each host isolate from its EOP values with all the phage isolates from a given day. These mean EOP values of host clones isolated from a given day were then plotted over time (Figure 15B). Host samples from Day 8 showed increased susceptibility to λ isolated from future days when compared with λ clones isolated from Day 8 (P < 2.546e-4). For days 15 and 22, hosts had higher EOP for phage samples from the future versus that from the past and contemporary (P < 2.883e-3 and P < 1.923e-4). Hosts isolated from Day 28 and 37 showed similar resistance to previous days; no future phage population were present for hosts isolated from Day 28 and 37. Similar analysis was performed for phage isolates, where mean EOP values of all phage isolates from a given day were plotted for different days (Figure 15C). Since all isolated hosts were resistant to all Day 8 phages, all EOP values were zero for Day 8 phages. No statistically significant difference was observed in mean EOP values across time for phage isolates from day 15. However, for phage samples from day 22 and 28, infectivity on past hosts were higher than that from contemporary and the future (*P* < 3.173e-7 and *P* < 2.417e-4). This pattern is consistent with the arms race dynamics (ARD), where the infectivity of the evolved phage on hosts from the past is always higher than that on hosts from the future [158].

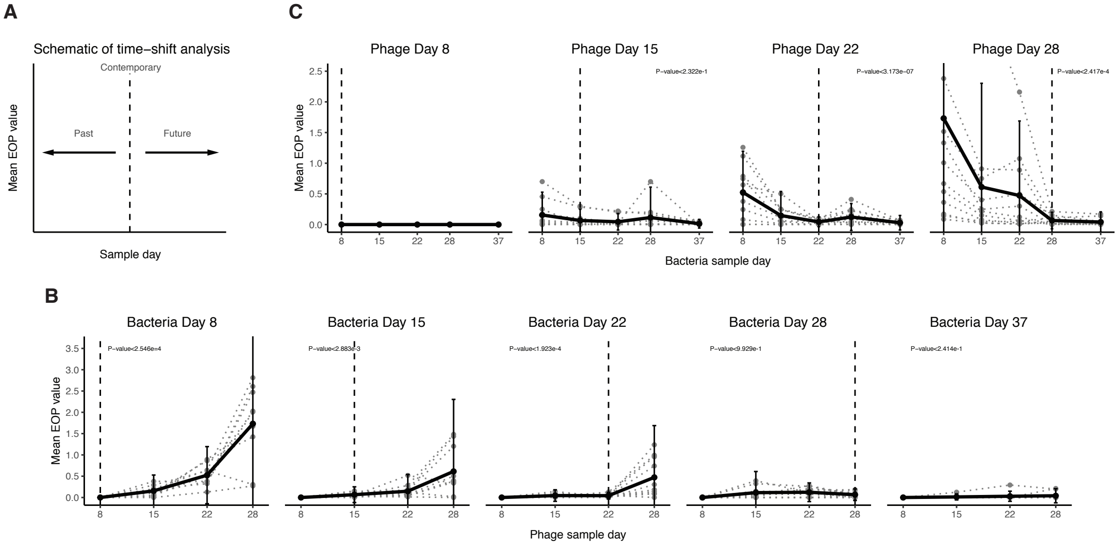

Figure 15 – Time-shift analysis results from different checkpoints

(A) Schematic for the time-shift analysis that compares the mean EOP from hosts or phages interacting with their counterparts from the past, contemporary and the future. (B) Time-shift results from host checkpoints day 8, 15, 22, 28 and 37, respectively. The gray dotted line shows the time-shift curve for each individual host and the black line shows the average. The vertical dashed line represents the host sample day. The P-values shown here are the maximum P-value from one-sided paired t tests comparing the final checkpoints with each of the previous checkpoints. (C) Time-shift results from phage checkpoints day 8, 15, 22 and 28 respectively. The gray dotted line shows the time-shift curve for each individual phage and the black line shows the average. The vertical dashed line represents the phage sample day. The P-values shown here are the maximum P-value from one-sided paired t tests comparing the initial checkpoints with each of the later checkpoints.

4.4.3 Bacteria and phage whole-genome sequence analysis

Whole genome sequencing revealed a total of 18 and 176 unique mutations for the host and phage strains respectively, resulting in 15 unique host genotypes and 34 unique phage genotypes (Figures 31, 32 and online Supplemental Tables 2 and 3 at https://github.com/speng32/thesis_supp_files). For *E. coli*, the 18 unique mutations consist of 7 nonsynonymous point mutations, 1 intergenic point mutation, 7 deletions and 3 duplications. These 18 unique mutations collectively affected a total of 1,021 nucleotides in the ancestral genome. The most abundant mutation that occurred in 38 out of 50 host genomes was a frameshift mutation caused by a 25-base duplication in the *malT* gene. This is consistent with the previously observed mutations from the coevolution experiment [40]. MalT is a positive regulator of an outer-membrane LamB protein of *E. coli* that λ uses to infect *E. coli*. The mutation in the *malT* gene of *E. coli* interferes with the expression of *lamB,* and confers resistance to phage. A frameshift mutation in the *manZ* gene emerges later in the experiment which was previously shown to confer high levels of resistance [40]. It appears to have the same affect here, all of the host with this mutation are resistant to all λ genotypes. *manZ* encodes an inner-membrane pore protein which transports λ's DNA across *E. coli*'s inner membrane. Another common mutation was a 777 bp deletion that was detected in 15 genomes. This mutation caused by the excision of an IS element and is known to occur at a high rate in REL606 [171]. None of the affected genes (ECB_RS14915 which encodes the SDR family oxidoreductase, ECB_RS14920 which encodes the IS1 family transposase and ECB_RS14925 which encodes a hypothetical protein) are known to have any effect on λ resistance [172]. This mutation is likely just a genomic hitchhiker that occurs because of its high mutation rate.

In λ isolates, a total of 176 unique mutations consisting of 53 nonsynonymous SNPs, 87 synonymous SNPs, 2 insertions, 3 deletions and 31 intergenic mutations, affecting a total of 182 nucleotides were identified. All the insertions and deletions detected were small indels that involved only 1 or 2 bases. Out of all mutations, 116 were in the *J* gene which encodes the host recognition protein of λ. J protein initiates infection by binding to *E. coli*'s LamB protein and some of these J mutations have been shown to increase adsorption rates to LamB and allow λ to exploit a novel receptor, OmpF [173, 174]. During the coevolution, we observed strongest selection for phage on Day 8 (Figure 33) and as phage population approached extinction by Day 37, the D_N/D_S ratio decreased. Overall, the high D_N/D_S shows that the phage experienced strong selection throughout the study in line with the ARD model.

4.4.4 Phylogenomics of coevolving phage and bacteria

A typical ARD pattern was observed in the λ-*E. coli* interaction network, but was it driven by the gene-for-gene model of coevolution at the genomic level? To answer this, we reconstructed the phylogenetic trees for both host and phage from whole genome sequences sampled at different days (Figure 16). Due to the prevalence of large insertions and deletions in the host genomes, conventional substitution models were not suitable to estimate phylogenetic trees for the host. The temporal signal was checked (Figures 29 and 30). As a result, we used an alternative approach as described in the Methods. We consider the ancestral strain as the root and all samples collected between the root and the last sample day as derived strains. Samples on the last day are described as the final strains. A typical ARD pattern at the genomic level would result in a directed phylogenetic tree where at each timestep the most dominant genotype is carried forward by accumulating more mutation in response to higher selection pressure by phage. This would result in the derived strains of Day 37 (tip of the tree) to be the furthest away from the ancestral strain (root of the tree). But interestingly, the phylogenomic pattern of host indicates a much more complex dynamic. We see that the strain with the highest level of resistance occurs at Day 37 (marked in red), but it is in fact most closely related to the sensitive ancestor. None of the intermediate derived strains were predicted to be the ancestor for the most dominant types present at the end of the coevolution. We hypothesize that this lineage had evolved early on in the experiment, but had remained at low levels until later in the experiment when broad host-range phages evolve and apply more pressure on the bacteria. We call this a ‘leap-frog’ dynamic where a rare lineage overtakes a dominant lineage later during coevolution.

A similar leap-frog dynamic was observed from the phylogenomics of λ (Figure 16B). None of the derived strains from Day 8 were predicted to be the ancestor of the final strains sampled on Day 28. When we compared the number of derived strains on the early dominant branch (green) versus the dominant later branch (blue), there was a gradual shift from day 8 and 28. The majority of the genotypes on Day 8 were located on the green branch, whereas by Day 22, about half the population had shifted to blue branch. Finally, all the genotypes of Day 28 were located on the blue branch.

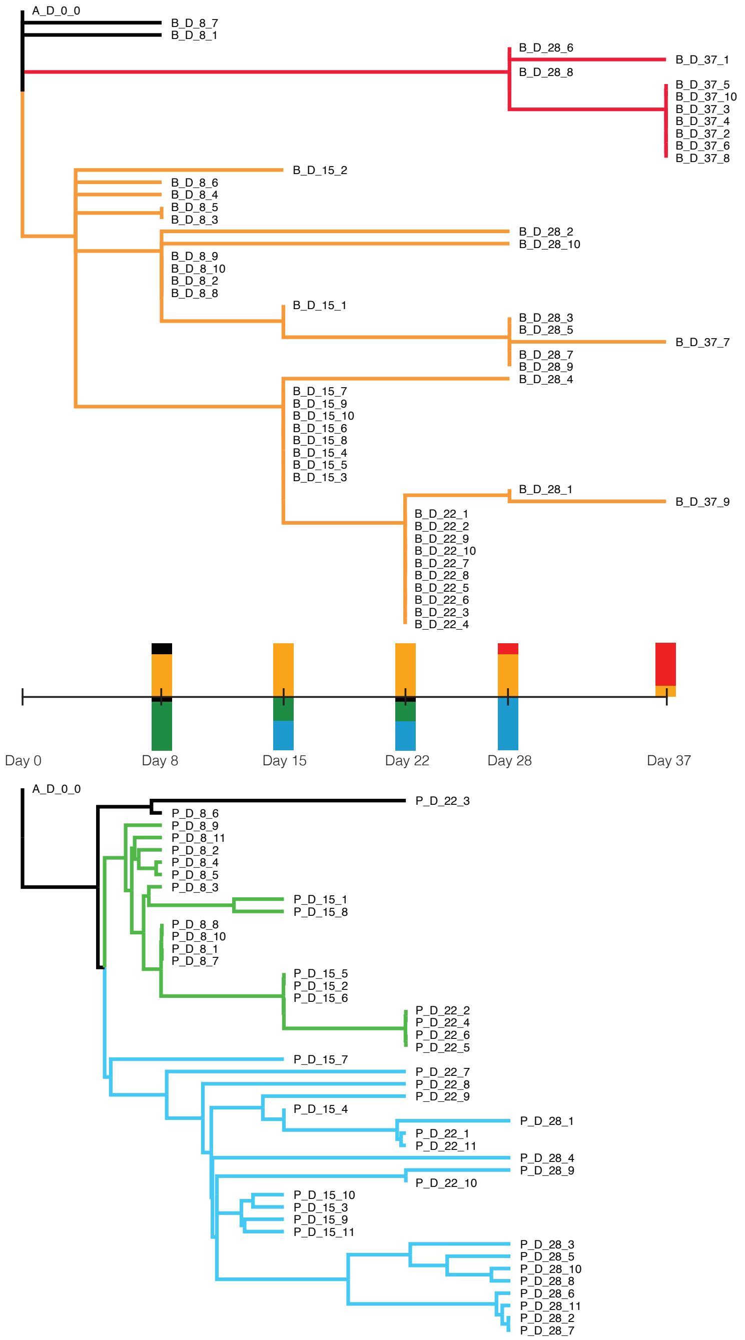

Figure 16 – Reconstructed phylogenomic trees of the hosts and phage

(A) The host phylodynamic tree reconstructed based on host mutation profiles. All super-resistant host strains are located on the red branch. The bar above the time scale represents the proportion of host strains from each colored branch across different checkpoints. (B) The phage phylodynamic tree reconstructed based on the phage mutation profiles. All day 28 phage strains are located on the blue branch. The bar below the time

scale represents the proportion of host strains from each colored branch across different checkpoints.

4.4.5 Whole population sequencing of the early community

To test whether the later dominant lineages were present earlier, we sequenced full genomes of *E. coli* and λ extracted from the mixed community on day 8. We predicted that we would be able to detect mutations that were on the late-dominant lineages that we not observed in the early-dominant lineages. Indeed, we uncovered the 16-*base* deletion in the *manZ* gene for *E. coli* and the single base substitution in *H* gene of λ which defined the final dominant clade in the coevolution (Figure 17, Tables 9 and 10). This confirms our lineage leap-frog dynamic hypothesis where a rare lineage from earlier timesteps emerges later in the arms race. Notably, the population sequencing revealed many more mutations than observed by sequencing isolates (Figures 34 and 35), suggesting that there are high levels of cryptic genetic variants in this coevolving population. As seen for the *manX* and H mutations, this variation can provide the genetic 'ammunition' important for later stages of the arms race.

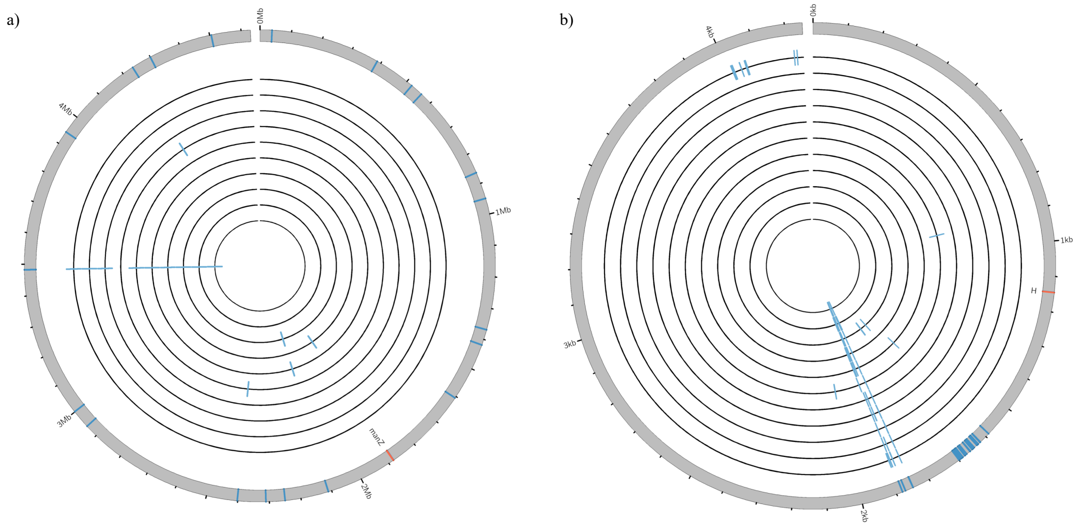

Figure 17 – Genomic diversity in whole population versus isolated clones on Day 8

The outer gray ring represents the whole population and the inner black circles represent all the isolated clones at Day 8 for a) E. coli b) λ. All the marks show different mutations present in them. The mutations marked in red (in gene manZ for E. coli and H for λ) is in the lineage dominant at the end of coevolution but whose evidence is found only in whole population sequencing.

4.4.6 Molecular mechanism underlying leap-frog dynamic

In order to study the molecular mechanism underlying the observed coevolutionary dynamics, we analyzed the gene functional annotation of several key players in the phage-host interaction. The ancestral phage strain uses the J protein to target the host porin LamB and injects the phage DNA into the periplasm [175, 176]. One positive regulator of the LamB porin is the HTH-type transcriptional regulator *malT*. As a result, mutations in the host malT protein downregulates the expression of LamB and affects phage-host interaction by mitigating λ's ability to exploit LamB.

Our results show that during the early stage of our experiment, the most common mutation in host genotypes – the 25-base duplication within the gene region that encodes *malT* – occurs amongst many of the day 8 host strains. As the coevolution plays forward, the majority of the derived host strains from later days, including all Day 15 and Day 22 derived strains, also carries this duplication. In contrast, none of the super-resistant strains of Day 37 have this mutation. Instead, they have a common 16-base deletion in the *manXYZ* gene. This gene encodes a permease for mannose, which is an inner membrane protein that λ uses to finally inject its DNA into the cytoplasm of the cell after attaching to an outer membrane protein of *E. coli* [177, 178]. Mutations in *manXYZ* have been shown to lead to the super-resistant phenotype in host strains [66]. But *manXYZ* gene is also shown to help *E. coli* uptake glucose, so mutation in this gene should hinder *E. coli's* growth rate in our experimental conditions. Alternatively, *malT* mutants have been shown to confer a slight benefit to growth rate in glucose medium [45]. Thus, the hosts with *manXYZ* mutations were overshadowed by *malT* which experienced high levels of cost-free resistance. As λ evolved to use a new receptor and increase its infectivity, *manZ* mutant's superior levels of resistance began to payoff. Cryptic genetic variation that arose early during the arms race were selected for at later stages when the ecology of the system, namely phage infectivity, change to favor its rise.

4.5 Discussion

To comprehensively understand the dynamics of λ-*E. coli* coevolution at different levels, we constructed the PBINs at phenotypic level and analyzed whole genomes of both λ and *E. coli*. We measured cross-infectivity amongst 51 hosts and 45 viruses sampled at 5 different days that coevolved over the course of a 37-day coevolution experiment and performed time-shift analysis on the observed changes. We then also sequenced all host-phage strains used to construct the PBINs and whole community of host-phage from a single day of an early timepoint of coevolution to relate interactions at phenotypic level with dynamics at genomic level.

The nested pattern of λ-*E. coli* PBIN revealed a typical ARD between phage and its bacterial host. However, the genomic data revealed that the arms race was not driven by this model's predicted steady accumulation of resistance or host-range mutations. Instead, the genomic data revealed 'leap-frog' dynamics for both the host and virus where an "old" lineage is maintained in the population for long duration until the ecological conditions change to favor it and drive it to dominance. The genomic data are more in line with FSD, where a large number of variants can be maintained in a population and different types are selected at varying times during coevolution. Reality falls somewhere in the middle of these two coevolutionary models.

The assumption of parsimony led to the misinterpretation of the dynamics that yield nestedness. A single evolving lineage is much more likely than a huge diversity of contending lineages. However, the reality is that the eco-evolutionary dynamics observed here yield the emergence and maintenance of vast genetic diversity and much more complex dynamics. This realization in line with other recent genomic-based studies that have reveal much more rare genetic variation than previously anticipated [179]. Our result for viruses is particularly important because the parsimonious assumption that modern lineages stem from previously observed lineages is also made for constructing phylogenies of human viruses such as influenza [180]. If this assumption is flawed for influenza, then researchers may misinterpret the number of molecular changes and its evolutionary dynamics. This would interfere with the analysis of its genomic evolution and subsequently, predictions for future strains and vaccine development.

CHAPTER 5. CONCLUSION

5.1 Summary of research advances

5.1.1 Research advance 1

An integrated analysis based on single cell sequencing, metagenomics and bioinformatics approaches was applied to evaluate virus-host interaction in a Yellowstone National Park (YNP) hot spring. The recovered virus-host relationships at both cell and species levels illustrated the ubiquity and complexity of the virus-host interaction network. Specifically, the results shown that the majority of the hosts in the environment contain viruses. Furthermore, most host cells contain viruses from multiple different viral partitions. In turn, within the relatively low-diversity community, the coexistence of a broad spectrum of virus types from specialists to generalists was observed. Taken together, these results should inspire new methods to assess the relevance of superinfection and the variation in the viral lifestyles in natural environments.

5.1.2 Research advance 2

During a coevolutionary experiment, the phenotype of phage-host interactions was quantified using quantitative plaque assays. Whole genome sequencing was performed for the isolated strains at different time points to reveal the genotypical variations that had occurred and accumulated. Machine learning algorithms were applied to link the phenotypical changes and genotypical changes. Quantitative models were built based on a two-step modelling framework and different sets of features. The outcomes revealed important genes, some of which have been experimentally validated for their roles in phage-host interactions, while others were genes that could potentially be involved. The flexibility of this framework allows for application on data from other host-pathogen system to reveal the most impactful mutations during the coevolution process in a quantitative way.

5.1.3 Research advance 3

Time-shift analysis was performed based on the host range of phage during the coevolutionary experiment. The arms-race dynamic (ARD) pattern was observed from the result of time-shift analysis. The phylodynamic trees for both host and phage were reconstructed based on the mutation profiles and sampling day to provide a comprehensive understanding of the coevolutionary process. The phylodynamic trees revealed a leap-frog dynamic which suggested that the current populations arose from rare subpopulations rather than the most recent, dominant lineages. The different conclusions based on phenotype and genotype evidences reveals that coevolutionary dynamics are much more complex than simple models can explain. The assumptions of linear genomic evolution could lead to misinterpretations of the evolutionary pattern and process.

5.2 The ubiquitous of viral-host interactions

In Chapter 2, we characterized the structure of virus-host interactions in a Yellowstone National Park (YNP) hot spring microbial community to quantitatively measure the extend of virus-host interactions in natural environments. By performing an integrated hexanucelotide, single cell sequencing and CRISPR-based analysis, we conservatively estimated that >60% of host cells contain at least one virus type. The majority of these cells contain two or more virus types. In conclusion, in the published work, we found that nearly all cells in the YNP NL01 hot spring interact with viruses, that multiple, concurrent interactions are common and that a broad spectrum of virus types from specialists to generalists coexist in a relatively low-diversity community [77].

These results should encourage the development of more robust empirical methods and theoretical models to assess the relevance of superinfection and a diversity of viral lifestyles in shaping natural communities. Current single-cell sequencing results do not fully capture the diverse sequences found in a cell due to coverage limitations. Higher-coverage sequencing data would provide more confidence and possibly new insights for investigating superinfection. Beyond the ubiquity of the virus-host interaction network in the hot spring, the viral lifestyles can also be further characterized across different spatial and temporal scales. Time series samples can be used to further investigate the dynamics of the virus-host interaction network. If we consider different hot springs as independent systems, by including samples from other similar hot springs, we could assess the diversity and similarity of the virus-host interaction networks.

5.3 The link between host range and genetic basis

Given a pair of virus and host that is known to interact with each other, in this case bacteriophage λ and *Escherichia coli*, we measured the changes in host range and the genetic profiles of both phage and *E. coli*. We proposed a two-step framework to link the phenotypical changes in terms of the host range and efficiency of infection with the changes in the genetic profiles. Overall, our framework confirmed several genes that were consistent with experimental validations, suggesting that our framework is capable of identifying the mutations in canonical genes that were known to involve in phage-host interactions. Our framework also revealed several genes that could potentially participate in such interactions, suggesting that it is capable of discovering novel genes that could participate in phage-host interactions. Although downstream experimental validation on the mutation or mutation pairs found are still necessary to confirm our newly identified sites, our framework can help prioritize experiments that genetically manipulate phage and host genomes.

For future work, experimental validations could be performed to evaluate the role of the novel genes predicted to be involved with the infection process (S and lom). Also, it is possible that the models which we term P×H:MF and Joint:MF have not yet reached their full potential due to the limited number of samples. These models could be refined given more sample data Finally, since our framework is very flexible, the logistic regression and linear regression used in the two steps can be replace by other models that also generate classification and regression results.

5.4 The genotypical and phenotypical coevolution dynamic

Under experimental conditions, samples taken at different checkpoints not only allow us to observe the genotypical and phenotypical changes, but also allow us to track the patterns of coevolution dynamic. Therefore, we investigated the dynamics of genotypes and phenotypes in coevolving virus-microbe, via analysis of full genome sequencing of *Escherichia coli* and bacteriophage λ. In contrast, we found that the phenotypical changes support the arms race dynamic. We also found that the emergence of resistant *E. coli* hosts and host-range mutant λ phage in later stages of the experiment arose from rare subpopulations rather than recent, dominant lineages. This lineage leap-frog dynamic is enabled by fluctuations in ecological conditions that rescue rare lineages with increasing resistance and infectious genotypes, rather than enabling the progressive genomic changes envisioned in an arms race.

Due to the limit number of samples taken at each checkpoint, we were not able to the shift in allele frequency spectrum in either host or phage. By performing metagenomic sequencing and analysis, such results would provide additional evidence to support the phage-host interaction dynamic.

5.5 Perspective

Taken together, our results showed that virus-host interactions are ubiquitous in natural environments, including extreme conditions. The observed virus-host interaction network that consists virus species that are generalists and specialists is highly complex. The observed changes in phage-host interactions can be tied to the genetic basis. And the theoretical framework based on genotypical changes, in turn, can also reveal potential genes that could participate in phage-host interactions. From a coevolutionary stand point, the observed phenotypical changes support the arms race dynamic while the genotypical changes supports the leap-frog dynamic. This shows the complexity in virus-host coevolution dynamic. In conclusion, virus-host interactions with the ubiquity and complexity, shape the coevolution trajectory of both virus and host and have a profound impact on the ecology of various environments.

APPENDIX A.
SUPPLEMENTARY INFORMATION FOR CHAPTER 2

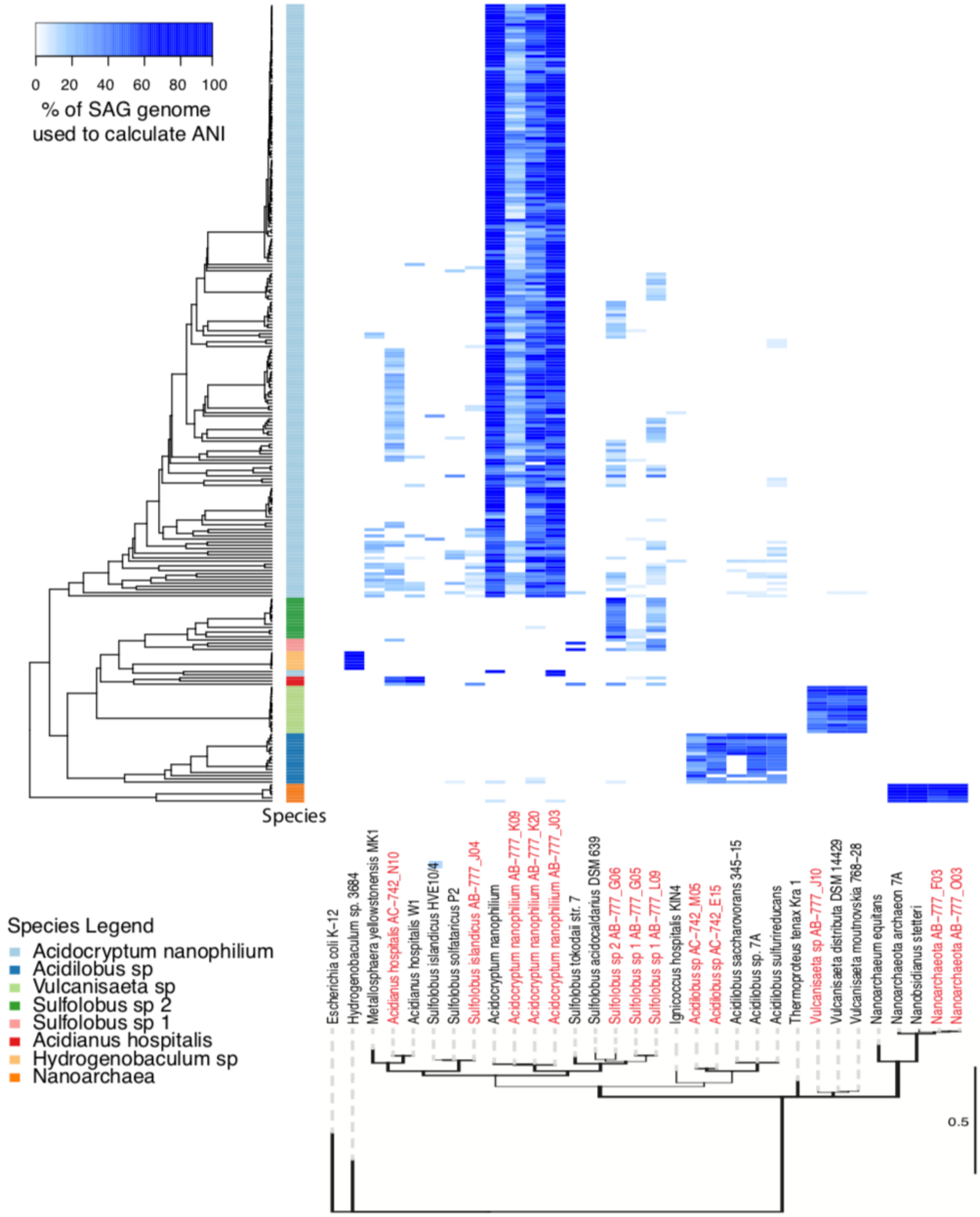

Figure 18 – Heatmap of the percent of the SAG genome used to calculate ANI for all classified SAGs against 32 reference genomes

SAGs are in the same order as Figure 5. Matches where less then 5% of the SAG genome was used were removed as were matches with a corresponding ANI <70%.

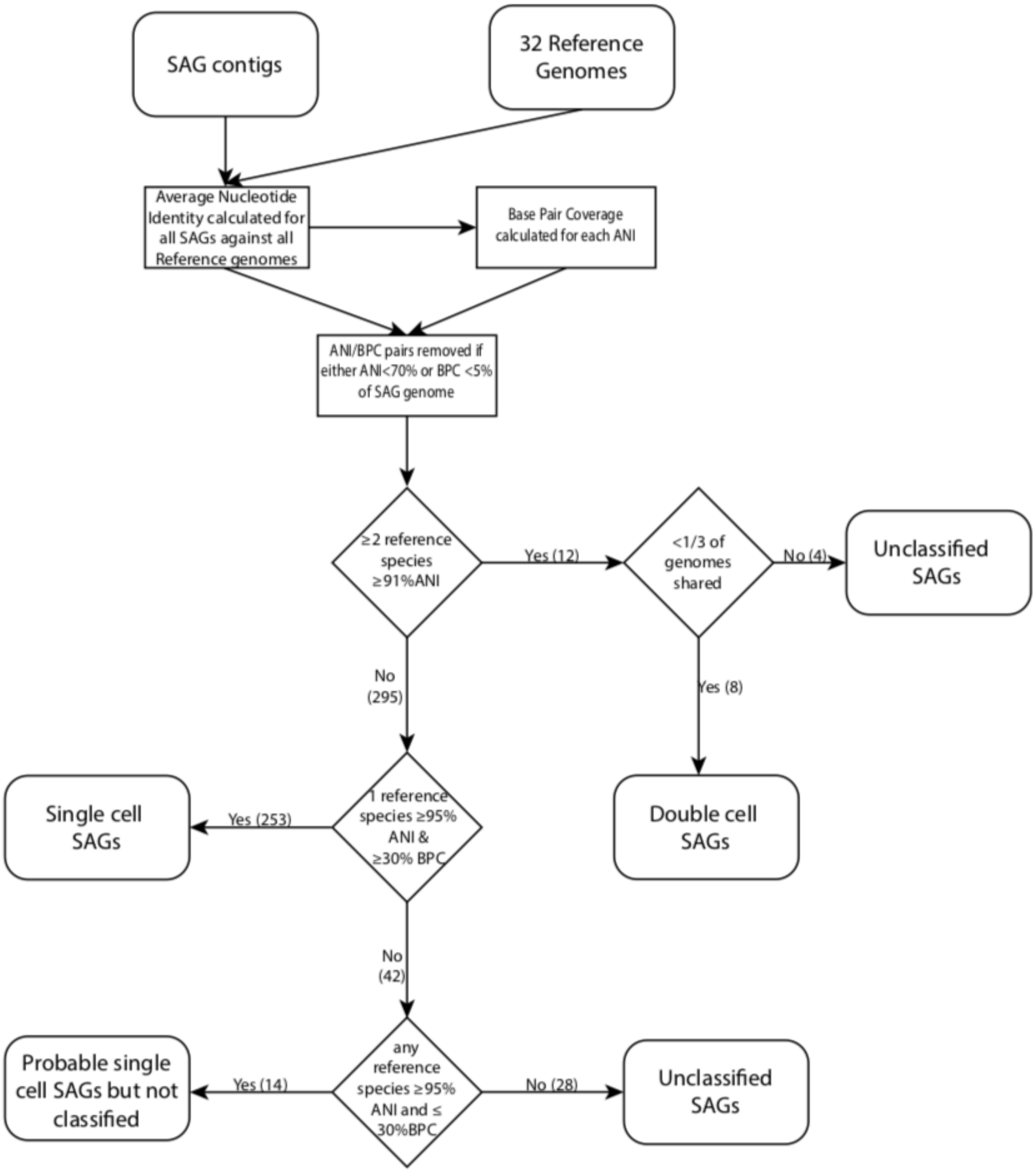
SAG contigs
32 Reference Genomes
Average Nucleotide Identity calculated for all SAGs against all Reference genomes
Base Pair Coverage calculated for each ANI
ANI/BPC pairs removed if either ANI<70% or BPC <5% of SAG genome
≥2 reference species ≥91%ANI
Yes (12)
<1/3 of genomes shared
No (4)
Unclassified SAGs
Yes (8)
Double cell SAGs
No (295)
1 reference species ≥95% ANI & ≥30% BPC
Yes (253)
Single cell SAGs
No (42)
any reference species ≥95% ANI and ≤ 30%BPC
Yes (14)
Probable single cell SAGs but not classified
No (28)
Unclassified SAGs

Figure 19 – Schematic overview of the logic pipeline used to classify single amplified genomes (SAG)

The average nucleotide identity (ANI) was calculated using the script provided here (https://github.com/chjp/ANI) and the base pair coverage (BPC) was calculated using a custom perl script. Numbers in parenthesis indicate the number of SAGs at each step of the pipeline.

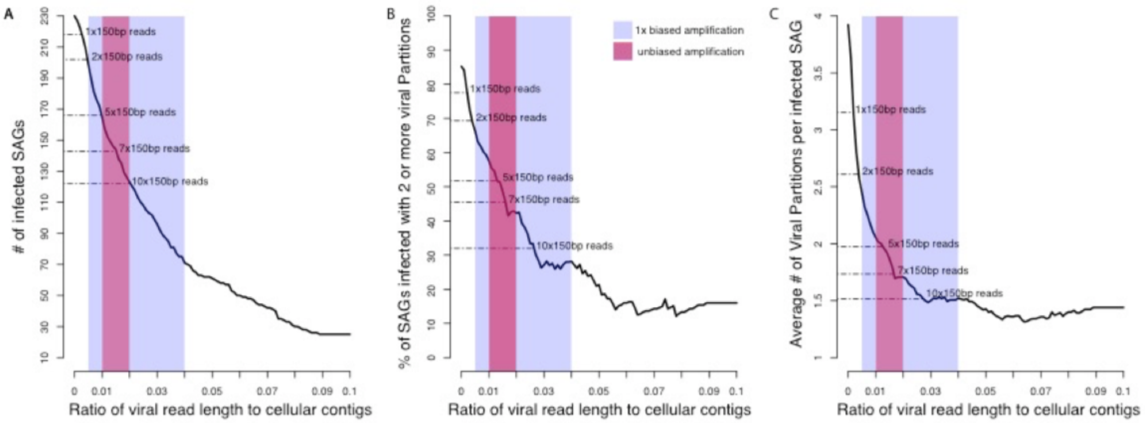

Figure 20 – Graphical representation showing the ratio of viral reads to assembled cellular contigs

The boxes showing expected biases were calculated using 30kb as the average size of thermophilic Archaeal viral genomes and an average thermophilic Archaeal genome size of 1.5-2Mbp. On all graphs different read cutoff levels from 1-10 150bp are shown. ***A.*** *The number of infected SAGs,* ***B****. the percentage of infected SAGs with two or more viral types present, and* ***C.*** *the average number of viral partitions present per infected SAG.*

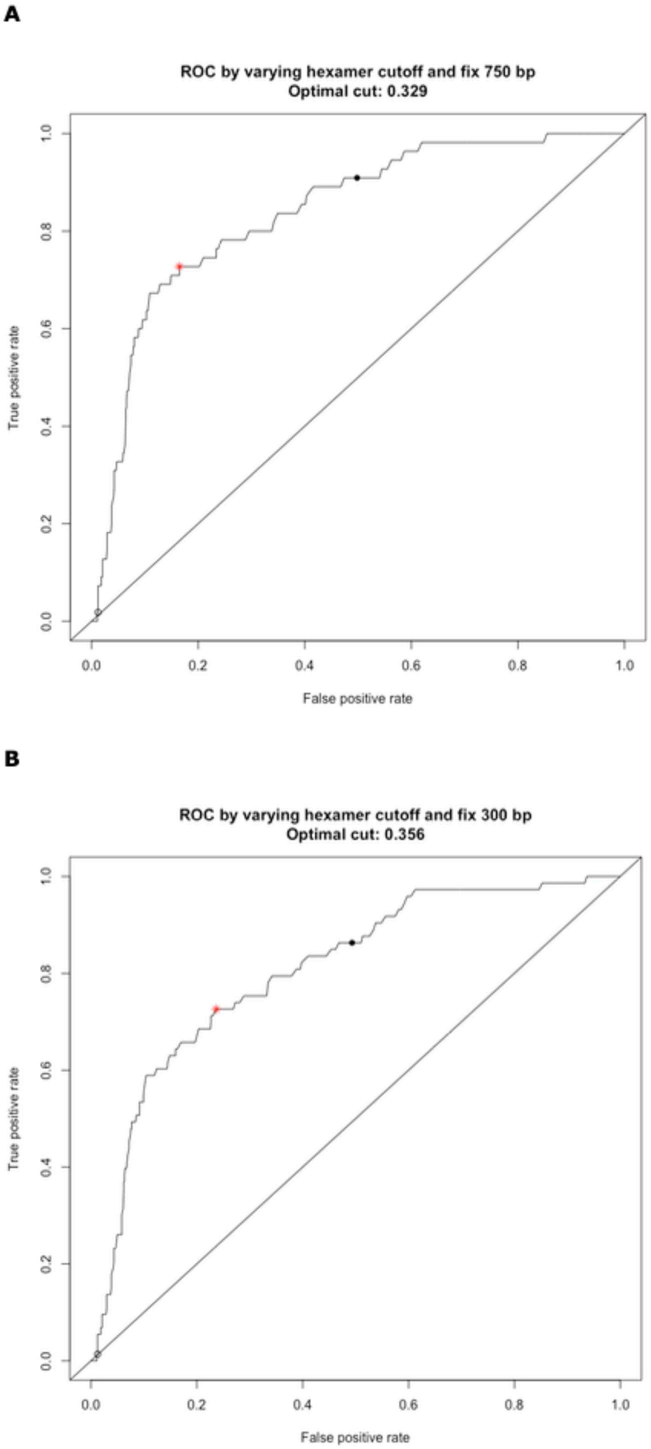

Figure 21 – Receiver operating characteristic (ROC) curves assuming A. 5 viral sequence reads (750bp) or B. 2 viral sequence reads (300bp). Optimal hexanucleotide analysis cut off values are indicated.

Table 1 – Reference genomes used in this study and a reference for each

Reference Genome	Reference
Hydrogenobaculum sp. 3684	GCA_000213785.1
Metallosphaera yellowstonensis MK1	GCA_000243315.1
Nanoarchaeum equitans	GCA_000008085.1
Nanodsidianus stetteri	GCA_000387965.1
Sulfolobus acidocaldarius DSM 639	GCA_000012285.1
Sulfolobus islandicus HVE10/4	GCA_000189575.1
Sulfolobus solfataricus P2	GCA_000007005.1
Sulfolobus tokodaii str. 7	GCA_000011205.1
Vulcanisaeta distributa DSM 14429	GCA_000148385.1
Vulcanisaeta moutnovskia 768-28	GCA_000190315.1
Nanoarchaeota archaeon 7A	GCA_001552015.1
Acidilobus sp. 7A	CP010515.1
Ignicoccus hospitalis KIN4/I	GCA_000017945.1
Acidilobus sulfurireducans	636559880
Acidilobus saccharovorans 345-15	GCA_000144915.1
Acidianus hospitalis W1	GCA_000213215.1
Acidocryptum nanophilium	GCA_000389735.1
Escherichia coli str. K-12 substr. MDS42	GCA_000350185.1

Table 2 – SAG sequencing and assembly statistics

SAG	Raw read count	# reads used for detection of viral sequences	kmernorm normalized read count	# contigs	# passing contigs (2200 bp)	Max contig length	Total contig length	%GC	estimated compleateness (CheckM)
AD-903-A01	100471	89488	23257	48	8	22270	69548	47.498	12.15
AD-903-A03	119883	107965	36931	50	14	19244	138597	45.916	7.94
AD-903-A04	128054	113889	45843	34	10	23309	119997	45.235	13.39
AD-903-A05	140374	123098	33594	81	17	28943	155754	47.693	22.92
AD-903-A06	200777	179648	23292	21	3	38033	48525	39.571	0
AD-903-A07	73829	65901	28417	42	9	24903	89210	56.845	13.5
AD-903-A08	92509	82440	37684	48	9	32925	107131	44.941	9.35
AD-903-A10	187508	165630	45300	50	15	25746	116872	48.437	5.03
AD-903-A11	191486	171882	64955	155	34	40775	280405	37.899	5.66
AD-903-A13	107595	99068	19382	34	4	27307	52237	45.066	5.69

AD-903-A14	200800	179664	52037	39	12	39271	139785	45.908	16.27
AD-903-A15	150544	133078	15413	31	3	22397	38521	55.622	2.54
AD-903-A16	208077	184523	35396	51	13	11594	90030	44.547	0
AD-903-A17	135403	116492	13566	16	1	21878	21878	43.628	0
AD-903-A18	240202	215165	44136	59	13	25371	102516	45.649	5.3
AD-903-A19	207275	188757	26692	41	9	19952	72682	43.639	4.98
AD-903-A20	272779	242016	13726	22	4	14200	31359	42.482	0
AD-903-A21	205241	187093	44402	43	8	36796	106701	41.939	0.72
AD-903-A22	204937	186319	69606	74	21	30858	206855	36.373	8.33
AD-903-A23	179304	164471	57417	139	23	30586	203424	35.947	0
AD-903-B02	678240	641140	60562	113	24	20845	135497	24.499	16.74
AD-903-B03	419087	383585	9446	8	1	11972	11972	25.351	0
AD-903-B04	698749	628444	23512	41	6	7096	27250	42.631	0
AD-903-B05	272043	250119	38854	53	11	37509	118992	48.498	22.1

AD-903-B06	360490	323260	39360	36	8	12818	73617	46.406	9.35
AD-903-B07	265267	241217	23829	32	7	17958	63586	49.144	16.67
AD-903-B08	546145	503149	34381	43	11	21764	80526	46.666	9.71
AD-903-B09	398828	363926	70584	99	28	14508	166277	45.413	16.18
AD-903-B10	497786	435310	43734	54	15	17211	98022	45.363	4.17
AD-903-B11	574522	518263	76753	153	34	21309	264582	37.449	19.81
AD-903-B13	3823	3501	2468	6	2	11753	18891	45.895	0
AD-903-B14	1104794	1033018	25785	33	5	20960	39537	38.276	0
AD-903-B15	490446	456758	92052	165	28	46830	274936	33.885	13.21
AD-903-B16	595155	543906	56018	33	9	24865	99206	47.164	5.61
AD-903-B17	155163	139639	66610	97	19	38687	207869	46.849	22.32
AD-903-B18	427723	386263	90459	63	18	43199	219613	45.31	18.22
AD-903-B19	244343	220125	81805	99	22	26223	216512	49.244	10.28
AD-903-B20	444430	402788	44790	65	17	21756	117780	49.216	6.54

AD-903-B21	306411	278599	79289	99	22	32741	204363	48.261	22.43
AD-903-B22	653923	614168	126355	326	52	45690	366447	28.984	29.52
AD-903-B23	412057	377888	24206	42	7	17878	50838	46.392	4.67
AD-903-C02	214880	197180	16364	45	10	15217	69577	45.463	6.54
AD-903-C03	173582	160366	51786	92	24	28739	191594	35.874	17.45
AD-903-C04	155712	138807	17869	20	5	18744	55234	47.547	13.69
AD-903-C05	110731	97279	30638	75	7	34889	76705	44.959	6.54
AD-903-C06	331090	300127	22066	22	3	18963	39872	46.15	5.61
AD-903-C07	155551	140629	17684	22	4	23898	38678	39.692	0.93
AD-903-C08	211324	193542	33802	39	11	30031	93715	45.679	12.15
AD-903-C09	143290	133869	40446	123	23	20259	178139	35.687	17.81
AD-903-C10	342216	307971	35259	68	17	12241	93523	49.162	11.01
AD-903-C11	266981	238079	48236	56	14	21750	111818	48.796	12.22
AD-903-C13	106411	97919	34849	18	8	25050	87387	46.35	10.52

AD-903-C14	405952	369975	39466	57	14	19776	105023	45.923	11.21
AD-903-C15	124592	111532	55906	130	26	43223	235079	44.815	15.03
AD-903-C16	313830	285444	56573	95	23	21973	161555	47.363	14.56
AD-903-C17	247672	222957	25088	13	3	29173	48566	42.104	0
AD-903-C18	276997	256121	21194	42	10	16207	66105	43.34	4.36
AD-903-C19	278192	251536	61529	84	22	24982	191138	47.051	24.06
AD-903-C20	223445	203679	26520	17	3	34110	66280	42.877	7.94
AD-903-C22	321288	292488	14661	41	5	16664	42617	43.83	5.61
AD-903-C23	424486	382351	15500	34	8	10132	37344	46.318	0.93
AD-903-D02	358701	329361	86518	120	22	43911	253179	46.945	21.27
AD-903-D03	497286	440098	31395	27	6	20822	57368	50.058	12.26
AD-903-D04	509470	456977	81984	111	18	41701	179617	48.787	18.99
AD-903-D05	232802	211935	46831	56	8	30706	108462	56.536	18.04
AD-903-D06	1207228	1109461	81002	61	13	24555	157470	34.732	8.33

AD-903-D07	277844	250140	40538	79	10	35057	131076	45.093	12.62
AD-903-D08	388666	351823	47120	102	19	21399	158407	45.595	2.34
AD-903-D09	465403	428817	44123	96	23	15986	145689	43.392	5.49
AD-903-D10	550113	493731	70384	66	14	34362	159368	49.719	21.03
AD-903-D11	407415	365038	69505	72	16	45433	196464	48.064	20.54
AD-903-D13	640209	598110	54973	85	17	35574	171761	36.521	7.33
AD-903-D14	1029988	943540	84028	78	11	41453	173492	46.808	20.63
AD-903-D15	600335	543229	27240	35	7	14964	48798	44.832	6.25
AD-903-D16	1034983	947564	33255	26	6	15176	52624	43.773	7.94
AD-903-D17	311585	279031	53613	60	9	40365	114927	56.563	16.04
AD-903-D18	806707	731863	93289	100	27	30348	263014	48.342	21.49
AD-903-D19	445725	403206	54393	39	12	38608	126888	47.346	11.29
AD-903-D20	1004202	915334	52131	41	8	40171	97823	55.646	7.55
AD-903-D21	757993	680668	14516	40	4	13659	39813	52.073	3.74

AD-903-D22	529452	487279	24885	25	3	24699	35515	39.271	0
AD-903-D23	689545	645657	77472	178	30	18043	188113	35.906	22.92
AD-903-E02	160046	145886	35583	55	14	23508	98517	49.073	10.28
AD-903-E03	248566	232073	17525	34	8	16163	54650	45.096	0
AD-903-E04	140335	128464	50949	108	15	36477	157606	34.587	13.58
AD-903-E05	139390	123375	58242	38	15	20374	148891	45.377	14.8
AD-903-E06	176576	160847	30170	37	7	14006	59642	43.977	6.54
AD-903-E07	96517	87317	11120	12	2	16701	18826	48.778	9.52
AD-903-E08	198605	182642	31853	42	8	21524	75856	47.925	6.11
AD-903-E09	99414	93514	45185	81	13	35365	152846	36.32	22.64
AD-903-E10	185324	168282	39262	55	13	28947	95607	48.145	3.89
AD-903-E11	144388	131658	37267	26	10	26002	109422	44.169	15.18
AD-903-E13	502627	453201	19436	26	6	9589	21681	44.62	0
AD-903-E15	153356	139459	15191	28	8	14447	49257	49.043	6.54

AD-903-E16	384001	354566	26034	31	4	28687	46411	44.793	0
AD-903-E17	76615	69564	19338	23	7	24751	52419	48.257	6.54
AD-903-E18	208233	189281	8121	11	2	5179	10046	50.09	0
AD-903-E20	269328	245110	44793	52	12	16693	95808	50.302	9.35
AD-903-E21	130674	121697	75294	192	49	32313	360212	37.608	31.15
AD-903-E22	192756	174977	22696	17	4	15980	48739	42.734	8.93
AD-903-E23	121473	112499	16657	30	3	19061	33299	43.224	2.8
AD-903-F02	384592	349081	40053	76	22	25737	127512	46.951	9.66
AD-903-F03	193704	168091	20919	35	6	24254	67716	47.947	14.88
AD-903-F04	342714	303574	54546	66	15	29813	145560	47.035	10.28
AD-903-F05	290196	271770	43797	97	23	12408	125027	25.123	19.63
AD-903-F06	1761666	1559582	80868	79	15	30351	125715	55.924	15.57
AD-903-F07	219813	200477	57106	166	24	17190	171349	34.267	13.39
AD-903-F08	393941	362402	82284	181	39	17036	265234	37.701	16.87
AD-903-F09	211734	194878	26931	43	8	20921	63605	43.945	0.93
AD-903-F10	570622	504889	58310	79	17	30367	146839	46.676	20.87
AD-903-F11	192671	170625	33788	35	11	19234	80909	43.263	0
AD-903-F13	296118	273143	37829	27	1	65464	65464	40.156	0

AD-903-F14	458383	407926	34429	40	5	19954	58644	43.846	0
AD-903-F15	221212	196514	50569	83	18	17413	133435	46.315	4.67
AD-903-F16	541642	487115	44305	31	6	37934	83506	45.752	0
AD-903-F17	239646	216313	34649	76	10	26410	106880	47.049	15.26
AD-903-F18	728266	680409	26078	76	13	9231	59029	24.603	12.31
AD-903-F19	420418	375799	26595	38	7	21600	66406	47.663	11.75
AD-903-F20	321815	286017	72206	159	22	73839	210397	47.462	9.35
AD-903-F21	391605	352309	36301	38	16	10759	81255	45.298	7.48
AD-903-F22	362619	324977	83879	120	18	38177	197529	47.095	21.43
AD-903-F23	214444	193334	30753	35	4	25837	75172	45.091	12.5
AD-903-G02	172898	159307	63343	101	19	45218	243318	37.823	31.13
AD-903-G03	251049	230029	17769	14	6	9980	36494	44.988	7.54
AD-903-G04	152293	137937	53429	91	21	48775	229489	35.762	21.79
AD-903-G05	156927	140623	17938	33	3	27264	39301	36.424	0
AD-903-G06	194630	172556	72241	161	39	27293	266786	45.157	13.08
AD-903-G07	127442	116837	37474	81	16	35628	135407	37.8	5.66
AD-903-G08	365418	330186	39312	64	9	28117	100051	46.122	14.68
AD-903-G09	44757	40936	26925	75	13	22922	118586	45.518	15.26

AD-903-G10	253657	223187	76776	69	17	29746	176496	46.919	8.41
AD-903-G11	89343	79135	29175	39	10	20451	78339	44.489	4.67
AD-903-G13	175445	161673	8093	10	3	5564	12082	42.617	0
AD-903-G14	195057	174484	56369	44	13	25500	148492	48.768	8.94
AD-903-G15	102803	93996	17084	60	8	16008	53281	37.664	3.77
AD-903-G16	378895	343248	33887	38	7	23061	75413	42.295	0
AD-903-G17	155264	138557	20310	36	4	24476	42931	45.773	0.93
AD-903-G18	310565	281809	42038	28	6	38586	84366	45.189	7.48
AD-903-G20	391458	359264	41624	74	9	44732	120937	45.622	7.48
AD-903-G21	226349	205898	30258	42	8	24405	89584	45.632	11.21
AD-903-G22	160932	146551	66204	79	18	40689	182973	37.218	9.72
AD-903-G23	330023	304912	18527	13	2	26312	29585	41.163	0
AD-903-I02	288621	264212	28304	35	9	21788	81517	44.954	0
AD-903-I03	130512	120241	37987	33	7	38184	106953	55.618	9.18
AD-903-I04	271626	250230	11152	16	6	15520	30452	39.239	0
AD-903-I05	172321	155696	64535	74	15	27992	186668	48.399	7.94

AD-903-I06	140309	127171	20105	15	4	21881	38216	49.887	4.67
AD-903-I07	46031	39318	15613	3	2	21123	23775	46.17	0
AD-903-I08	224189	209431	27155	50	11	34057	89612	44.626	0.93
AD-903-I09	189422	180364	46066	125	32	10135	142641	25.375	17.45
AD-903-I10	445873	411186	95211	165	40	37033	331117	34.25	19.82
AD-903-I11	183599	164719	31557	42	11	12540	80126	49.991	12.79
AD-903-I13	93771	87081	24213	24	8	13810	63847	48.807	14.78
AD-903-I14	464941	438446	63077	98	26	27852	153465	28.898	14.05
AD-903-I15	251530	231283	19088	39	9	7762	43880	44.344	4.67
AD-903-I16	353196	328685	22919	27	5	14637	44672	43.884	9.13
AD-903-I17	116059	106030	35540	61	13	28847	118374	47.68	0
AD-903-I18	341901	314932	87368	72	19	35812	236476	47.297	13.08
AD-903-I19	235402	218126	81324	109	28	27732	278422	38.601	11.32
AD-903-I20	291173	269208	51244	28	9	27969	135345	48.227	12.15
AD-903-I21	128388	119452	39094	86	9	51173	134118	35.297	9.51
AD-903-I22	310485	287508	32154	33	8	17457	66439	46.67	9.52
AD-903-I23	222178	205664	16085	9	3	19065	25261	54.942	1.89
AD-903-J02	583686	531378	36596	60	13	14925	95673	46.94	10.28
AD-903-J03	276137	242504	47620	80	18	34967	171712	47.404	22.52
AD-903-J04	354770	316604	31009	11	1	53005	53005	43.181	0
AD-903-J05	259664	235907	23000	9	2	23596	41467	51.639	3.16
AD-903-J06	635879	583363	39819	67	8	41761	98023	35.412	8.33

AD-903-J07	481745	434995	63278	59	14	27551	126728	48.573	0
AD-903-J08	875503	795203	49723	51	11	23695	102757	50.113	11.9
AD-903-J09	377959	350989	56911	70	19	16511	163305	46.939	17.4
AD-903-J10	785150	697419	39918	16	3	42210	69350	43.448	0
AD-903-J11	704524	633100	47961	76	17	25298	139257	47.473	17.76
AD-903-J13	446549	410286	77162	112	26	43070	213926	47.594	13.1
AD-903-J14	1155188	1041988	73183	82	19	29636	165668	48.562	8.33
AD-903-J15	414357	373795	77618	55	12	40090	166189	55.53	23.66
AD-903-J16	725972	658696	79051	55	16	32109	176760	47.783	22.62
AD-903-J17	612359	555445	46343	72	14	30104	120616	46.352	11.21
AD-903-J18	21234	19145	13909	23	5	12461	38071	57.561	3.77
AD-903-J19	726416	638406	62342	63	18	22795	135363	48.453	16.51
AD-903-J20	895294	825646	28347	50	11	16497	73674	43.683	1.87
AD-903-J21	551418	500716	42088	80	15	46822	125846	45.185	0
AD-903-J22	969594	882855	91881	68	17	31999	198375	47.609	11.21
AD-903-J23	751614	697008	16348	39	5	13281	33575	49.388	0
AD-903-K02	142318	130163	32200	78	15	15744	110768	45.614	11.9
AD-903-K03	155540	143199	57480	104	23	20579	199191	37.876	14.88
AD-903-K04	363412	331708	20829	30	2	39603	46085	42.79	1.87
AD-903-K05	168591	150428	30926	69	12	29469	100219	45.107	5.61

AD-903-K06	307281	277715	48550	71	14	20605	115914	45.246	11.21
AD-903-K07	191609	174497	36481	21	7	29777	74573	46.254	16.94
AD-903-K08	176366	158452	18701	32	9	13205	45550	40.806	7.94
AD-903-K09	181645	169743	35579	40	7	25045	89720	44.877	2.18
AD-903-K10	448568	399951	82108	70	13	54045	185528	56.769	21.17
AD-903-K11	201878	178950	48087	66	15	19380	96268	49.34	14.98
AD-903-K13	100175	91914	18049	15	4	31484	39427	48.594	13.1
AD-903-K14	226359	203920	40667	44	13	29842	116922	44.579	14.29
AD-903-K15	78010	70106	25600	27	13	13701	85305	51.672	12.77
AD-903-K16	311608	282014	51504	56	12	28961	119492	44.159	10.32
AD-903-K17	124693	111382	2884	6	3	4502	9162	33.617	0
AD-903-K18	194219	174575	44111	62	10	21429	85259	54.713	8.39
AD-903-K19	333253	302286	21559	29	6	19556	50710	44.405	0
AD-903-K20	324255	292367	12382	25	4	6596	18701	50.495	3.12

AD-903-K21	217414	200912	87227	245	46	26068	331809	37.962	27.38
AD-903-K22	193016	172718	33439	37	6	15581	61540	43.596	0
AD-903-K23	320532	293695	45897	83	13	30556	139295	48.752	15.89
AD-903-L02	1031057	957966	15642	31	5	4324	15022	44.455	0
AD-903-L03	259575	228332	45295	86	16	22533	144408	44.144	6.85
AD-903-L04	633934	582705	61349	88	17	19064	127134	29.154	14.49
AD-903-L05	364420	331305	31578	21	11	16009	63799	47.766	13.1
AD-903-L06	7246	6434	2762	6	2	4659	7806	54.151	0
AD-903-L07	354781	329153	56341	128	32	25347	237751	36.786	23.1
AD-903-L08	913628	842961	97360	145	35	32330	258152	36.171	21.39
AD-903-L09	1651580	1530395	42422	47	14	9858	67935	48.991	0
AD-903-L10	2015593	1816814	45220	50	11	21903	79617	60.051	0
AD-903-L11	386390	351533	28368	61	14	37547	110313	44.357	7.48
AD-903-L13	380432	353314	19022	24	4	19815	45091	43.155	0
AD-903-L14	524921	480910	22227	34	10	11359	54512	48.731	0
AD-903-L16	702670	626832	20084	29	4	13739	45012	51.52	0
AD-903-L17	182860	160818	47843	98	11	38458	133646	47.808	7.48
AD-903-L18	446119	408052	75412	52	17	36135	172690	47.641	16.2
AD-903-L19	503493	455435	53138	54	11	21107	128501	46.478	5.61
AD-903-L20	367342	333693	42806	86	16	21669	122537	47.351	18.15

AD-903-L21	733442	676716	13873	19	4	8800	21278	42.579	0
AD-903-L22	477737	434902	35461	42	9	23131	76889	46.238	4.21
AD-903-L23	750709	688190	26418	38	10	16328	55660	47.738	0
AD-903-M02	169464	156403	38924	79	16	24940	139857	37.831	18.13
AD-903-M03	132914	119162	17943	22	2	21824	29929	56.029	14.15
AD-903-M04	158179	138892	23510	24	4	20383	50222	43.164	0
AD-903-M05	115250	102395	65870	123	34	29070	289651	37.999	20.75
AD-903-M06	216539	194704	40597	41	9	28133	113094	44.673	3.74
AD-903-M07	112558	101557	36101	47	9	26422	115275	47.094	10.71
AD-903-M08	146360	131583	37319	57	16	32278	119958	46.163	11.92
AD-903-M10	398458	353985	15842	27	6	6007	24274	37.641	0
AD-903-M11	116366	102484	27352	37	8	19743	58770	46.275	0
AD-903-M13	139637	127891	34997	76	13	23444	83371	45.264	2.08
AD-903-M14	239933	213388	45428	59	16	13506	115537	48.312	13.55
AD-903-M15	130497	115868	27196	13	2	65053	75355	42.724	0

AD-903-M16	190885	169716	39103	74	12	26339	124587	44.577	6.54
AD-903-M17	300376	263294	32232	43	9	10216	51795	47.806	0
AD-903-M18	197109	175202	49988	56	14	26321	122825	47.308	0
AD-903-M19	106814	94572	19262	29	7	14485	45665	46.101	8.93
AD-903-M20	347919	324664	55406	141	29	24329	170645	29.873	22.66
AD-903-M21	129482	116594	57691	62	13	40681	180122	46.044	23.36
AD-903-M23	609372	546464	15583	20	3	4989	12295	51.354	0
AD-903-N02	354188	325445	74849	148	17	52128	222312	37.177	43.55
AD-903-N03	219203	195663	32911	46	12	18611	96956	46.102	7.48
AD-903-N04	1074787	944591	17731	35	3	7339	13117	55.729	0.94
AD-903-N05	426180	398308	62382	190	38	19172	214087	25.146	26.01
AD-903-N06	391922	349295	38888	52	10	29285	93324	47.65	16.31
AD-903-N07	681087	628377	48273	52	12	36759	113524	45.113	18.89
AD-903-N08	610591	550303	26532	19	4	17358	42076	47.783	11.21

AD-903-N09	484178	452504	38840	45	11	16144	86783	47.815	9.35
AD-903-N10	724359	644001	57090	94	18	16649	117738	47.441	10.28
AD-903-N11	772965	681375	58542	85	17	20383	126658	48.807	10.75
AD-903-N13	791804	732858	33938	52	10	31134	79046	44.011	8.88
AD-903-N14	593392	536564	36501	59	11	48140	88495	47.504	15.11
AD-903-N16	351243	314714	31435	32	9	24721	73396	49.121	0
AD-903-N17	1023551	930502	54856	76	16	16721	93659	34.997	3.57
AD-903-N18	141077	125366	7269	9	3	7741	17604	58.322	0
AD-903-N19	613095	555753	33385	29	10	17952	77547	47.397	14.49
AD-903-N20	604042	553075	25975	20	3	28410	44806	42.276	6.07
AD-903-N21	372124	338631	78256	58	16	45532	177121	49.061	15.42
AD-903-N22	586216	536275	122424	240	45	40433	380184	36.395	24.29
AD-903-N23	344755	316816	57190	60	16	29644	170110	48.679	26.88
AD-903-O02	219015	200272	49742	87	17	20058	149094	45.858	20.04

AD-903-O03	120603	109314	21390	13	3	22521	46723	46.746	6.54
AD-903-O04	172416	152680	36019	112	16	20070	125398	36.499	7.23
AD-903-O05	200554	180655	23152	34	6	19893	53344	43.69	0
AD-903-O06	374728	335747	20369	42	7	8986	34249	46.766	3.74
AD-903-O07	214368	197567	61612	118	36	12959	192528	34.106	8.33
AD-903-O08	125055	113036	29077	31	9	21161	78800	46.393	13.99
AD-903-O09	57262	51840	20623	61	11	24196	78191	45.794	0
AD-903-O10	289296	255751	39001	45	11	20076	71545	41.894	0
AD-903-O11	244715	215696	33301	49	9	13262	56953	49.469	9.35
AD-903-O13	142917	130992	25433	49	14	10114	77071	48.677	0
AD-903-O14	329748	301931	61630	35	11	36387	166676	46.23	10.75
AD-903-O15	160923	144062	37136	29	4	31118	77599	57.74	13.21
AD-903-O16	448961	399671	68069	92	26	14051	146318	44.295	4.67
AD-903-O17	53137	46964	18112	40	6	14658	46921	48.253	0

AD-903-O18	230345	207981	45941	54	14	29811	120471	46.301	13.75
AD-903-O19	137196	122701	33481	43	12	28933	91366	45.526	17.46
AD-903-O21	367031	334230	40981	48	15	21239	103822	45.916	7.67
AD-903-O20	335531	306907	37321	69	13	13301	88898	45.445	1.87
AD-903-O22	309077	279079	27079	58	8	29197	73173	44.668	8.33
AD-903-O23	151651	135782	46975	71	15	29209	161141	47.406	4.67
AD-903-P01	177384	159269	23422	36	3	35471	42582	43.821	5.61
AD-903-P02	355904	325681	82174	81	19	28179	194991	48.449	19.62
AD-903-P03	208122	184752	32534	72	12	22917	93763	35.154	4.25
AD-903-P04	369542	327899	60826	73	22	24351	138248	47.347	0
AD-903-P05	168232	149411	29218	29	6	18405	52686	46.117	0
AD-903-P06	420810	316700	30433	42	4	35113	64385	46.06	2.8
AD-903-P07	184389	165046	32008	39	7	27388	60793	44.847	6.07
AD-903-P08	460642	425782	40291	55	15	17262	128854	48.563	8.1

AD-903-P09	281557	261285	25357	20	2	28636	47325	43.776	5.61
AD-903-P10	435351	390696	73962	69	22	29678	200362	49.967	20
AD-903-P11	280341	252227	44059	47	5	31499	82231	43.411	0
AD-903-P13	324655	297116	36584	61	7	22236	71387	46.557	9.52
AD-903-P14	592297	538728	65782	61	21	21439	166972	48.328	23.63
AD-903-P15	445704	419644	97790	149	42	24686	283077	24.604	35.9
AD-903-P16	666411	623249	83571	149	37	14587	192530	25.715	29.55
AD-903-P17	218524	191929	48603	69	14	21806	129102	48.411	17.32
AD-903-P18	565945	507474	14250	17	2	10993	17335	47.967	0
AD-903-P19	454206	400825	51279	75	13	17777	115701	49.791	0
AD-903-P20	712766	647426	30095	47	10	22197	91972	47.006	4.67
AD-903-P21	192733	170873	27876	37	5	18109	51469	46.607	0
AD-903-P22	304463	272864	58367	49	9	38267	126585	47.489	0
AD-903-P23	260512	233978	54430	76	8	60577	113015	48.52	22.1

109930	99777	12933		738	340763
697	660	961		39	09
358080.					110997
4463					.749
201559					
3					380184
3823					7806

Table 3 – Recruitment of reads from SAGs used in this study onto publicly available viral metagenomes from other environments at 95% ID over 100bp

SAG	Species	Number of SAG reads	TOV 18 SUR	TOV 18DCM	Human gut
AD-903-A01	A. nanophilium	89488	0	0	0
AD-903-A03	A. nanophilium	107965	0	0	0
AD-903-A04	A. nanophilium	113889	0	0	0
AD-903-A05	A. nanophilium	123098	0	0	0
AD-903-A06	Unclassified	179648	0	0	0
AD-903-A07	Acidilobus sp	65901	0	0	0
AD-903-A08	Vulcanisaeta sp	82440	0	0	0
AD-903-A10	A. nanophilium	165630	0	0	0
AD-903-A11	Sulfolobus sp 2	171882	0	0	0
AD-903-A13	A. nanophilium	99068	0	0	0
AD-903-A14	A. nanophilium	179664	0	0	0
AD-903-A15	Acidilobus sp	133078	0	0	0
AD-903-A16	Likely Vulcanisaeta sp	184523	0	0	0
AD-903-A17	Unclassified	116492	0	0	0
AD-903-A18	A. nanophilium	215165	0	0	0
AD-903-A19	A. nanophilium	188757	0	0	0
AD-903-A20	Unclassified	242016	0	0	0
AD-903-A21	A. nanophilium	187093	0	0	0
AD-903-A22	Likely Sulfolobus sp 1	186319	0	0	0
AD-903-A23	Sulfolobus sp 1	164471	0	0	0
AD-903-B02	Unclassified	641140	0	0	0
AD-903-B03	Unclassified	383585	0	0	0
AD-903-B04	Vulcanisaeta sp	628444	0	0	0
AD-903-B05	A. nanophilium	250119	0	0	0
AD-903-B06	A. nanophilium	323260	0	0	0
AD-903-B07	A. nanophilium	241217	0	0	0
AD-903-B08	A. nanophilium	503149	0	0	0
AD-903-B09	Vulcanisaeta sp	363926	0	0	0
AD-903-B10	A. nanophilium	435310	0	0	0
AD-903-B11	Sulfolobus sp 2	518263	0	0	0
AD-903-B13	Likely A. nanophilium & Metallosphaera	3501	0	0	0

	yellowstonensis MK1				
AD-903-B14	A. nanophilium & Nanoarchaea	1033018	0	0	0
AD-903-B15	Acidianus hospitalis W1	456758	0	0	0
AD-903-B16	A. nanophilium	543906	0	0	0
AD-903-B17	A. nanophilium	139639	0	0	0
AD-903-B18	Vulcanisaeta sp	386263	0	0	0
AD-903-B19	A. nanophilium	220125	0	0	0
AD-903-B20	A. nanophilium	402788	0	0	0
AD-903-B21	A. nanophilium	278599	0	0	0
AD-903-B22	Nanoarchaea & Sulfolobus sp 1	614168	0	0	0
AD-903-B23	Unclassified	377888	0	0	0
AD-903-C02	A. nanophilium	197180	0	0	0
AD-903-C03	Sulfolobus sp 1	160366	0	0	0
AD-903-C04	A. nanophilium	138807	0	0	0
AD-903-C05	Vulcanisaeta sp	97279	0	0	0
AD-903-C06	A. nanophilium	300127	0	0	0
AD-903-C07	A. nanophilium	140629	0	0	0
AD-903-C08	A. nanophilium	193542	0	0	0
AD-903-C09	Likely Sulfolobus sp 1	133869	0	0	0
AD-903-C10	A. nanophilium	307971	0	0	0
AD-903-C11	A. nanophilium	238079	0	0	0
AD-903-C13	A. nanophilium	97919	0	0	0
AD-903-C14	A. nanophilium	369975	0	0	0
AD-903-C15	Vulcanisaeta sp	111532	0	0	0
AD-903-C16	A. nanophilium	285444	0	0	0
AD-903-C17	A. nanophilium	222957	0	0	0
AD-903-C18	A. nanophilium	256121	0	0	0
AD-903-C19	A. nanophilium	251536	0	0	0
AD-903-C20	A. nanophilium	203679	0	0	0
AD-903-C22	A. nanophilium	292488	0	0	0
AD-903-C23	A. nanophilium	382351	0	0	0
AD-903-D02	A. nanophilium	329361	0	0	0
AD-903-D03	A. nanophilium	440098	0	0	0
AD-903-D04	A. nanophilium	456977	0	0	0
AD-903-D05	Acidilobus sp	211935	0	0	0

AD-903-D06	Hydrogenobaculum sp. 3684	1109461	0	0	0
AD-903-D07	Vulcanisaeta sp	250140	0	0	0
AD-903-D08	Vulcanisaeta sp	351823	0	0	0
AD-903-D09	A. nanophilium & Nanoarchaea	428817	0	0	0
AD-903-D10	A. nanophilium	493731	0	0	0
AD-903-D11	A. nanophilium	365038	0	0	0
AD-903-D13	Likely Sulfolobus sp 1	598110	0	0	0
AD-903-D14	A. nanophilium	943540	0	0	0
AD-903-D15	A. nanophilium	543229	0	0	0
AD-903-D16	A. nanophilium	947564	0	0	0
AD-903-D17	Acidilobus sp	279031	0	0	0
AD-903-D18	A. nanophilium	731863	0	0	0
AD-903-D19	A. nanophilium	403206	0	0	0
AD-903-D20	Acidilobus sp	915334	0	0	0
AD-903-D21	A. nanophilium	680668	0	0	0
AD-903-D22	A. nanophilium	487279	0	0	0
AD-903-D23	Nanoarchaea & Vulcanisaeta sp	645657	0	0	0
AD-903-E02	A. nanophilium	145886	0	0	0
AD-903-E03	A. nanophilium	232073	0	0	0
AD-903-E04	Hydrogenobaculum sp. 3684	128464	0	0	0
AD-903-E05	Likely Vulcanisaeta sp	123375	0	0	0
AD-903-E06	A. nanophilium	160847	0	0	0
AD-903-E07	A. nanophilium	87317	0	0	0
AD-903-E08	A. nanophilium	182642	0	0	0
AD-903-E09	Sulfolobus sp 2	93514	0	0	0
AD-903-E10	A. nanophilium	168282	0	0	0
AD-903-E11	A. nanophilium	131658	0	0	0
AD-903-E13	A. nanophilium	453201	0	0	0
AD-903-E15	A. nanophilium	139459	0	0	0
AD-903-E16	A. nanophilium	354566	0	0	0
AD-903-E17	A. nanophilium	69564	0	0	0
AD-903-E18	Unclassified	189281	0	0	0
AD-903-E20	A. nanophilium	245110	0	0	0
AD-903-E21	Sulfolobus sp 2	121697	0	0	0
AD-903-E22	A. nanophilium	174977	0	0	0

AD-903-E23	Likely A. nanophilium & Sulfolobus sp 1	112499	0	0	0
AD-903-F02	A. nanophilium	349081	0	0	0
AD-903-F03	A. nanophilium	168091	0	0	0
AD-903-F04	A. nanophilium	303574	0	0	0
AD-903-F05	Nanoarchaea	271770	0	0	0
AD-903-F06	Acidilobus sp	1559582	0	0	0
AD-903-F07	Hydrogenobaculum sp. 3684	200477	0	0	0
AD-903-F08	Sulfolobus sp 2	362402	0	0	0
AD-903-F09	Unclassified	194878	0	0	0
AD-903-F10	A. nanophilium	504889	0	0	0
AD-903-F11	Likely A. nanophilium & Sulfolobus sp 1	170625	0	0	0
AD-903-F13	Unclassified	273143	0	0	0
AD-903-F14	Unclassified	407926	0	0	0
AD-903-F15	Likely Vulcanisaeta sp	196514	0	0	0
AD-903-F16	A. nanophilium	487115	0	0	0
AD-903-F17	A. nanophilium	216313	0	0	0
AD-903-F18	Nanoarchaea	680409	0	0	0
AD-903-F19	A. nanophilium	375799	0	0	0
AD-903-F20	Unclassified	286017	0	0	0
AD-903-F21	A. nanophilium	352309	0	0	0
AD-903-F22	A. nanophilium	324977	0	0	0
AD-903-F23	A. nanophilium	193334	0	0	0
AD-903-G02	Sulfolobus sp 2	159307	0	0	0
AD-903-G03	A. nanophilium	230029	0	0	0
AD-903-G04	Likely Sulfolobus sp 1	137937	0	0	0
AD-903-G05	Acidianus hospitalis W1	140623	0	0	0
AD-903-G06	Vulcanisaeta sp	172556	0	0	0
AD-903-G07	Sulfolobus sp 2	116837	0	0	0
AD-903-G08	A. nanophilium	330186	0	0	0
AD-903-G09	Vulcanisaeta sp	40936	0	0	0
AD-903-G10	A. nanophilium	223187	0	0	0
AD-903-G11	Unclassified	79135	0	0	0
AD-903-G13	Unclassified	161673	0	0	0

AD-903-G14	A. nanophilium	174484	0	0	0
AD-903-G15	Likely Sulfolobus sp 2	93996	0	0	0
AD-903-G16	A. nanophilium	343248	0	0	0
AD-903-G17	A. nanophilium	138557	0	0	0
AD-903-G18	A. nanophilium	281809	0	0	0
AD-903-G20	A. nanophilium	359264	0	0	0
AD-903-G21	A. nanophilium	205898	0	0	0
AD-903-G22	Likely Sulfolobus sp 1	146551	0	0	0
AD-903-G23	A. nanophilium	304912	0	0	0
AD-903-I02	Vulcanisaeta sp	264212	0	0	0
AD-903-I03	Acidilobus sp	120241	0	0	0
AD-903-I04	Unclassified	250230	0	0	0
AD-903-I05	A. nanophilium	155696	0	0	0
AD-903-I06	A. nanophilium	127171	0	0	0
AD-903-I07	A. nanophilium	39318	0	0	0
AD-903-I08	A. nanophilium	209431	0	0	0
AD-903-I09	Unclassified	180364	0	0	0
AD-903-I10	Hydrogenobaculum sp. 3684	411186	0	0	0
AD-903-I11	A. nanophilium	164719	0	0	0
AD-903-I13	A. nanophilium	87081	0	0	0
AD-903-I14	A. nanophilium & Nanoarchaea	438446	0	0	0
AD-903-I15	A. nanophilium	231283	0	0	0
AD-903-I16	A. nanophilium	328685	0	0	0
AD-903-I17	A. nanophilium	106030	0	0	0
AD-903-I18	A. nanophilium	314932	0	0	0
AD-903-I19	Sulfolobus sp 2	218126	0	0	0
AD-903-I20	A. nanophilium	269208	0	0	0
AD-903-I21	Hydrogenobaculum sp. 3684	119452	0	0	0
AD-903-I22	A. nanophilium	287508	0	0	0
AD-903-I23	Acidilobus sp	205664	0	0	0
AD-903-J02	A. nanophilium	531378	0	0	0
AD-903-J03	A. nanophilium	242504	0	0	0
AD-903-J04	Unclassified	316604	0	0	0
AD-903-J05	Unclassified	235907	0	0	0
AD-903-J06	Sulfolobus sp 1	583363	0	0	0
AD-903-J07	A. nanophilium	434995	0	0	0

AD-903-J08	A. nanophilium	795203	0	0	0
AD-903-J09	A. nanophilium	350989	0	0	0
AD-903-J10	Unclassified	697419	0	0	0
AD-903-J11	A. nanophilium	633100	0	0	0
AD-903-J13	A. nanophilium	410286	0	0	0
AD-903-J14	A. nanophilium	1041988	0	0	0
AD-903-J15	Acidilobus sp	373795	0	0	0
AD-903-J16	A. nanophilium	658696	0	0	0
AD-903-J17	A. nanophilium	555445	0	0	0
AD-903-J18	Acidilobus sp	19145	0	0	0
AD-903-J19	A. nanophilium	638406	0	0	0
AD-903-J20	A. nanophilium	825646	0	0	0
AD-903-J21	A. nanophilium	500716	0	0	0
AD-903-J22	A. nanophilium	882855	0	0	0
AD-903-J23	A. nanophilium	697008	0	0	0
AD-903-K02	A. nanophilium	130163	0	0	0
AD-903-K03	Sulfolobus sp 2	143199	0	0	0
AD-903-K04	A. nanophilium	331708	0	0	0
AD-903-K05	A. nanophilium	150428	0	0	0
AD-903-K06	A. nanophilium	277715	0	0	0
AD-903-K07	A. nanophilium	174497	0	0	0
AD-903-K08	Likely A. nanophilium	158452	0	0	0
AD-903-K09	A. nanophilium	169743	0	0	0
AD-903-K10	Acidilobus sp	399951	0	0	0
AD-903-K11	A. nanophilium	178950	0	0	0
AD-903-K13	A. nanophilium	91914	0	0	0
AD-903-K14	A. nanophilium	203920	0	0	0
AD-903-K15	A. nanophilium	70106	0	0	0
AD-903-K16	A. nanophilium	282014	0	0	0
AD-903-K17	A. nanophilium	111382	0	0	0
AD-903-K18	Acidilobus sp	174575	0	0	0
AD-903-K19	A. nanophilium	302286	0	0	0
AD-903-K20	A. nanophilium	292367	0	0	0
AD-903-K21	Sulfolobus sp 2	200912	0	0	0
AD-903-K22	Unclassified	172718	0	0	0
AD-903-K23	A. nanophilium	293695	0	0	0
AD-903-L02	A. nanophilium	957966	0	0	0
AD-903-L03	Vulcanisaeta sp	228332	0	0	0
AD-903-L04	Nanoarchaea	582705	0	0	0

AD-903-L05	A. nanophilium	331305	0	0	0
AD-903-L06	Unclassified	6434	0	0	0
AD-903-L07	Sulfolobus sp 2	329153	0	0	0
AD-903-L08	Likely Sulfolobus sp 1	842961	0	0	0
AD-903-L09	A. nanophilium	1530395	0	0	0
AD-903-L10	Unclassified	1816814	0	0	0
AD-903-L11	A. nanophilium	351533	0	0	0
AD-903-L13	Likely A. nanophilium & Sulfolobus sp 1	353314	0	0	0
AD-903-L14	A. nanophilium	480910	0	0	0
AD-903-L16	A. nanophilium	626832	0	0	0
AD-903-L17	Unclassified	160818	0	0	0
AD-903-L18	A. nanophilium	408052	0	0	0
AD-903-L19	A. nanophilium	455435	0	0	0
AD-903-L20	A. nanophilium	333693	0	0	0
AD-903-L21	A. nanophilium & Sulfolobus sp 1	676716	0	0	0
AD-903-L22	A. nanophilium	434902	0	0	0
AD-903-L23	A. nanophilium	688190	0	0	0
AD-903-M02	Sulfolobus sp 2	156403	0	0	0
AD-903-M03	Acidilobus sp	119162	0	0	0
AD-903-M04	A. nanophilium	138892	0	0	0
AD-903-M05	Likely Sulfolobus sp 2	102395	0	0	0
AD-903-M06	A. nanophilium	194704	0	0	0
AD-903-M07	A. nanophilium	101557	0	0	0
AD-903-M08	A. nanophilium	131583	0	0	0
AD-903-M10	Unclassified	353985	0	0	0
AD-903-M11	A. nanophilium	102484	0	0	0
AD-903-M13	A. nanophilium & Nanoarchaea	127891	0	0	0
AD-903-M14	A. nanophilium	213388	0	0	0
AD-903-M15	Unclassified	115868	0	0	0
AD-903-M16	Vulcanisaeta sp	169716	0	0	0
AD-903-M17	A. nanophilium	263294	0	0	0
AD-903-M18	A. nanophilium	175202	0	0	0
AD-903-M19	A. nanophilium	94572	0	0	0
AD-903-M20	Nanoarchaea & Sulfolobus sp 2	324664	0	0	0

AD-903-M21	A. nanophilium	116594	0	0	0
AD-903-M23	A. nanophilium	546464	0	0	0
AD-903-N02	Sulfolobus sp 2	325445	0	0	0
AD-903-N03	A. nanophilium	195663	0	0	0
AD-903-N04	Acidilobus sp	944591	0	0	0
AD-903-N05	Nanoarchaea	398308	0	0	0
AD-903-N06	A. nanophilium	349295	0	0	0
AD-903-N07	A. nanophilium	628377	0	0	0
AD-903-N08	A. nanophilium	550303	0	0	0
AD-903-N09	A. nanophilium	452504	0	0	0
AD-903-N10	A. nanophilium	644001	0	0	0
AD-903-N11	A. nanophilium	681375	0	0	0
AD-903-N13	A. nanophilium	732858	0	0	0
AD-903-N14	A. nanophilium	536564	0	0	0
AD-903-N15			0	0	0
AD-903-N16	A. nanophilium	314714	0	0	0
AD-903-N17	Hydrogenobaculum sp. 3684	930502	0	0	0
AD-903-N18	Acidilobus sp	125366	0	0	0
AD-903-N19	A. nanophilium	555753	0	0	0
AD-903-N20	A. nanophilium	553075	0	0	0
AD-903-N21	A. nanophilium	338631	0	0	0
AD-903-N22	Sulfolobus sp 1	536275	0	0	0
AD-903-N23	A. nanophilium	316816	0	0	0
AD-903-O02	A. nanophilium	200272	0	0	0
AD-903-O03	A. nanophilium	109314	0	0	0
AD-903-O04	Likely Sulfolobus sp 1	152680	0	0	0
AD-903-O05	A. nanophilium	180655	0	0	0
AD-903-O06	A. nanophilium	335747	0	0	0
AD-903-O07	Acidianus hospitalis W1	197567	0	0	0
AD-903-O08	A. nanophilium	113036	0	0	0
AD-903-O09	Vulcanisaeta sp	51840	0	0	0
AD-903-O10	Likely A. nanophilium	255751	0	0	0
AD-903-O11	A. nanophilium	215696	0	0	0
AD-903-O13	A. nanophilium	130992	0	0	0
AD-903-O14	A. nanophilium	301931	0	0	0
AD-903-O15	Acidilobus sp	144062	0	0	0
AD-903-O16	Vulcanisaeta sp	399671	0	0	0

AD-903-O17	A. nanophilium	46964	0	0	0
AD-903-O18	A. nanophilium	207981	0	0	0
AD-903-O19	A. nanophilium	122701	0	0	0
AD-903-O20	A. nanophilium	334230	0	0	0
AD-903-O21	A. nanophilium	306907	0	0	0
AD-903-O22	A. nanophilium	279079	0	0	0
AD-903-O23	A. nanophilium	135782	0	0	0
AD-903-P01	A. nanophilium	159269	0	0	0
AD-903-P02	A. nanophilium	325681	0	0	0
AD-903-P03	Unclassified	184752	0	0	0
AD-903-P04	Unclassified	327899	0	0	0
AD-903-P05	A. nanophilium	149411	0	0	0
AD-903-P06	A. nanophilium	316700	0	0	0
AD-903-P07	A. nanophilium	165046	0	0	0
AD-903-P08	A. nanophilium	425782	0	0	0
AD-903-P09	A. nanophilium	261285	0	0	0
AD-903-P10	A. nanophilium	390696	0	0	0
AD-903-P11	A. nanophilium	252227	0	0	0
AD-903-P13	A. nanophilium	297116	0	0	0
AD-903-P14	A. nanophilium	538728	0	0	0
AD-903-P15	Nanoarchaea	419644	0	0	0
AD-903-P16	Nanoarchaea	623249	0	0	0
AD-903-P17	A. nanophilium	191929	0	0	0
AD-903-P18	Unclassified	507474	0	0	0
AD-903-P19	A. nanophilium	400825	0	0	0
AD-903-P20	A. nanophilium	647426	0	0	0
AD-903-P21	A. nanophilium	170873	0	0	0
AD-903-P22	Unclassified	272864	0	0	0
AD-903-P23	A. nanophilium	233978	0	0	0

Table 4 – Recruitment of reads from publically available SAGs onto the NL01 viral dataset

metagenome used in this study at 95% ID over 100bp	# reads	% GC	Max read Length	# reads match NL10 viral	% of reads that match NL10 viral network	# major partitions hit	# reads mapping to major partitions
Acidobacteria_bacterium_SCGC_AAA001-I23	24558598	49	157	349	1.42E-03	0	0
Acidobacteria_bacterium_SCGC_AAA003-J17	17423082	56	157	14	8.04E-05	0	0
actinobacterium SCGC AAA027-D23	28218510	46	150	1	3.54E-06	0	0
alpha proteobacterium SCGC AAA027-C06	32051398	27	150	0	0.00E+00	0	0
Bacteroidetes_bacterium_SCGC_AD-308-D03v2	15207438	34	150	0	0.00E+00	0	0
Bacteroidetes_bacterium_SCGC_AD-311-C03v2	28743844	32	150	0	0.00E+00	0	0
beta proteobacterium SCGC AAA024-K11	33398800	48	150	1	2.99E-06	0	0
candidate division OP8 bacterium SCGC AC-335-L06	28508108	35	150	0	0.00E+00	0	0
Chloroflexi_bacterium_SCGC_AC-312_J06v2	18122298	51	150	0	0.00E+00	0	0
Colwellia_sp_SCGC_AC281-C22	20007424	35	150	2	1.00E-05	0	0
Deferribacteres_bacterium_SCGC_AC-312_E04v2	20664154	40	146	0	0.00E+00	0	0
Deltaproteobacteria_bacterium SCGC_AC-312_D19v2	28601930	40	151	0	0.00E+00	0	0
Desulfovibrionales_bacterium_SCGC_AC-335-L09	28069186	41	150	0	0.00E+00	0	0
Epsilonproteobacteria_bacterium_SCGC_AD-305-P03v2	73384752	37	146	0	0.00E+00	0	0
Eudoraea_sp_SCGC_5250	30067226	37	150	1	3.33E-06	0	0
Euryarchaeota_archaeon_SCGC_AB-633-I06	31597216	34	157	0	0.00E+00	0	0
Firmicutes_bacterium_SCGC_AC-699-C23	26355166	49	150	55	2.09E-04	0	0
Firmicutes_bacterium_SCGC_AC-699-M18	29592720	48	150	0	0.00E+00	0	0
Gammaproteobacteria_bacterium_SCGC_AAA003-E02	23779320	45	150	48	2.02E-04	0	0
Gemmatimonadetes_bacterium_SCGC_AAA007-L19	30129182	52	157	22	7.30E-05	0	0

Halothiobacillaceae_bacterium_SCGC_AB-674-E03	30027042	46	150	1	3.33E-06	0	0
Ignavibacteriaceae_bacterium_SCGC_AB-674-D06	29949740	27	150	58	1.94E-04	0	0
Lentisphaerae_bacterium_SCGC_AAA283-D08	22829088	44	150	0	0.00E+00	0	0
Nitrospirae_bacterium_SCGC_AB-219-C22	29809482	46	157	0	0.00E+00	0	0
Thaumarchaeota_archaeon_SCGC_AAA287-E17	22559998	34	150	0	0.00E+00	0	0
Total	703655702			552	7.84E-05	0	0
Average	28146228.08	41		22.08		0	0

APPENDIX B.
SUPPLEMENTARY INFORMATION FOR CHAPTER 3

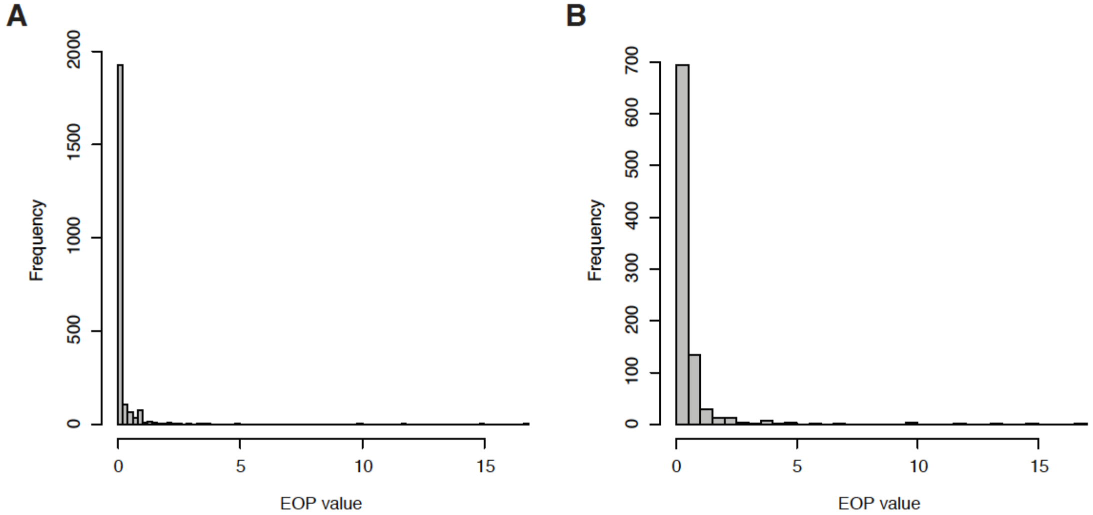

Figure 22 – Distribution of the observed EOP values

(A) Overall distribution of the EOP values. (B) Distribution of positive EOP values only.

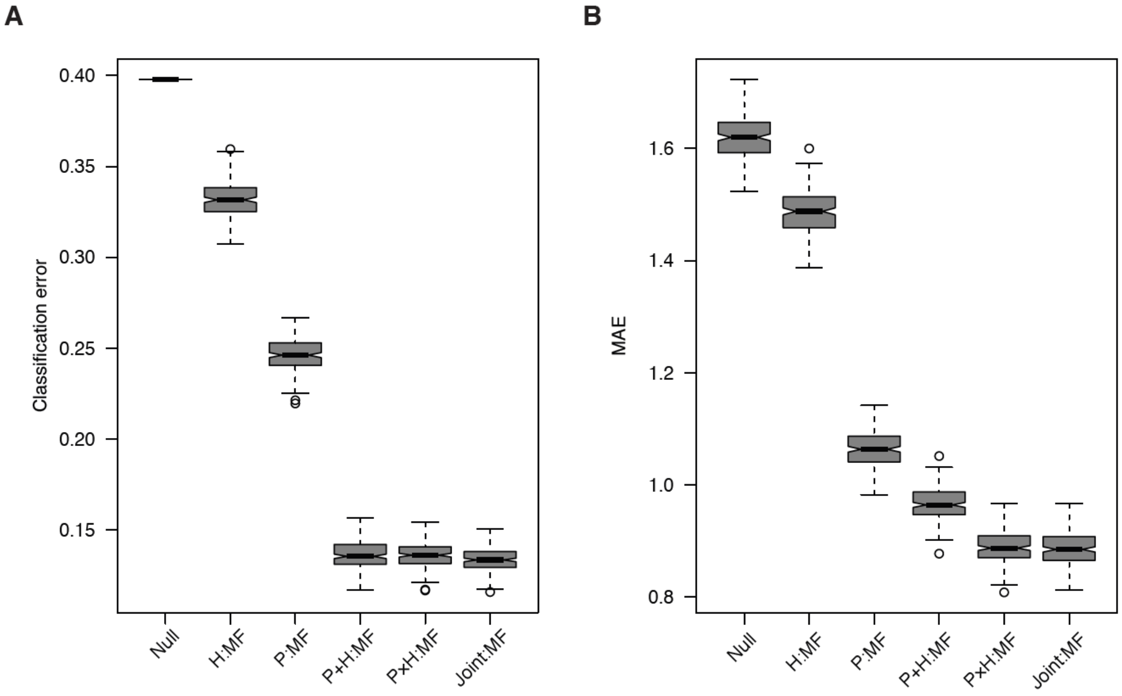

Figure 23 – Model performance for different feature sets on training set

(A) Boxplot of training set classification error for step 1 based on 200 bootstrap runs for null model and models based on H:MF, P:MF, P+H:MF, P×H:MF and Joint:MF. (B) Boxplot of training set MAE for step 2 on 200 bootstrap runs for null model and models based on H:MF, P:MF, P+H:MF, P×H:MF and Joint:MF.

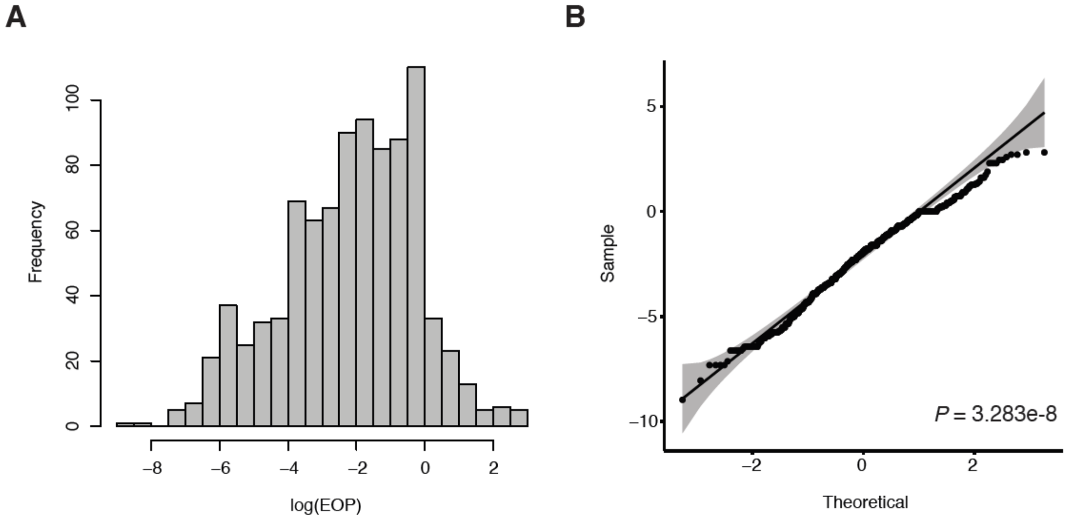

Figure 24 – Log transformed positive EOP value distribution

(A) Distribution of the log positive EOP values (B) Q-Q plot for log positive EOP values against normal quantiles. P value calculated from Shapiro-Wilk test.

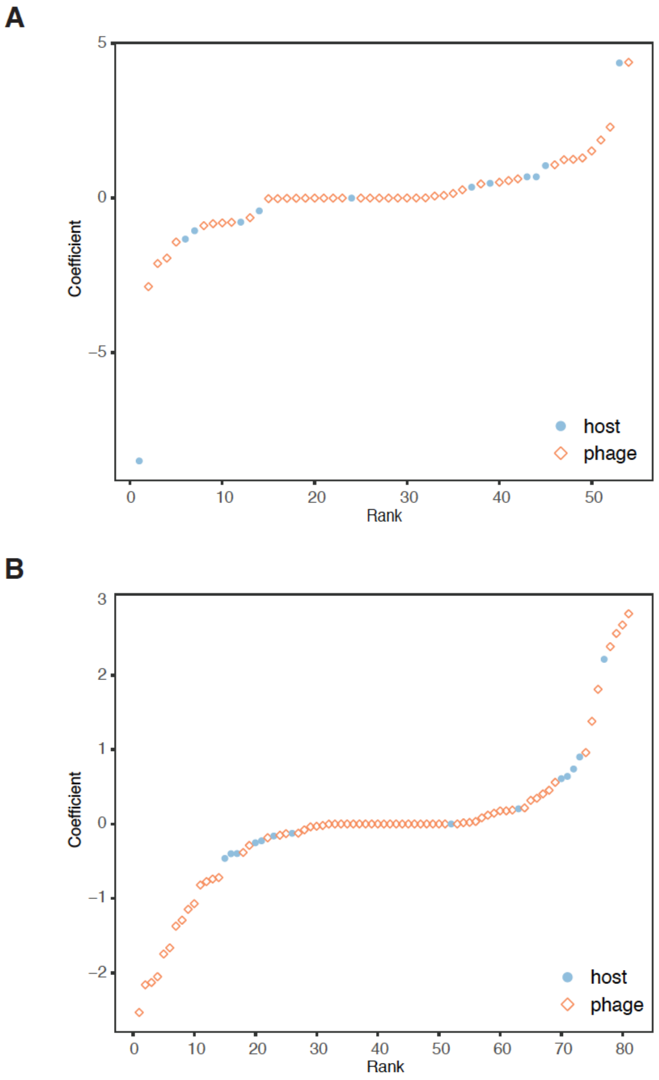

Figure 25 – Rank ordered coefficients from the final step 1 model (A) and step 2 model (B) based on P+H:MF

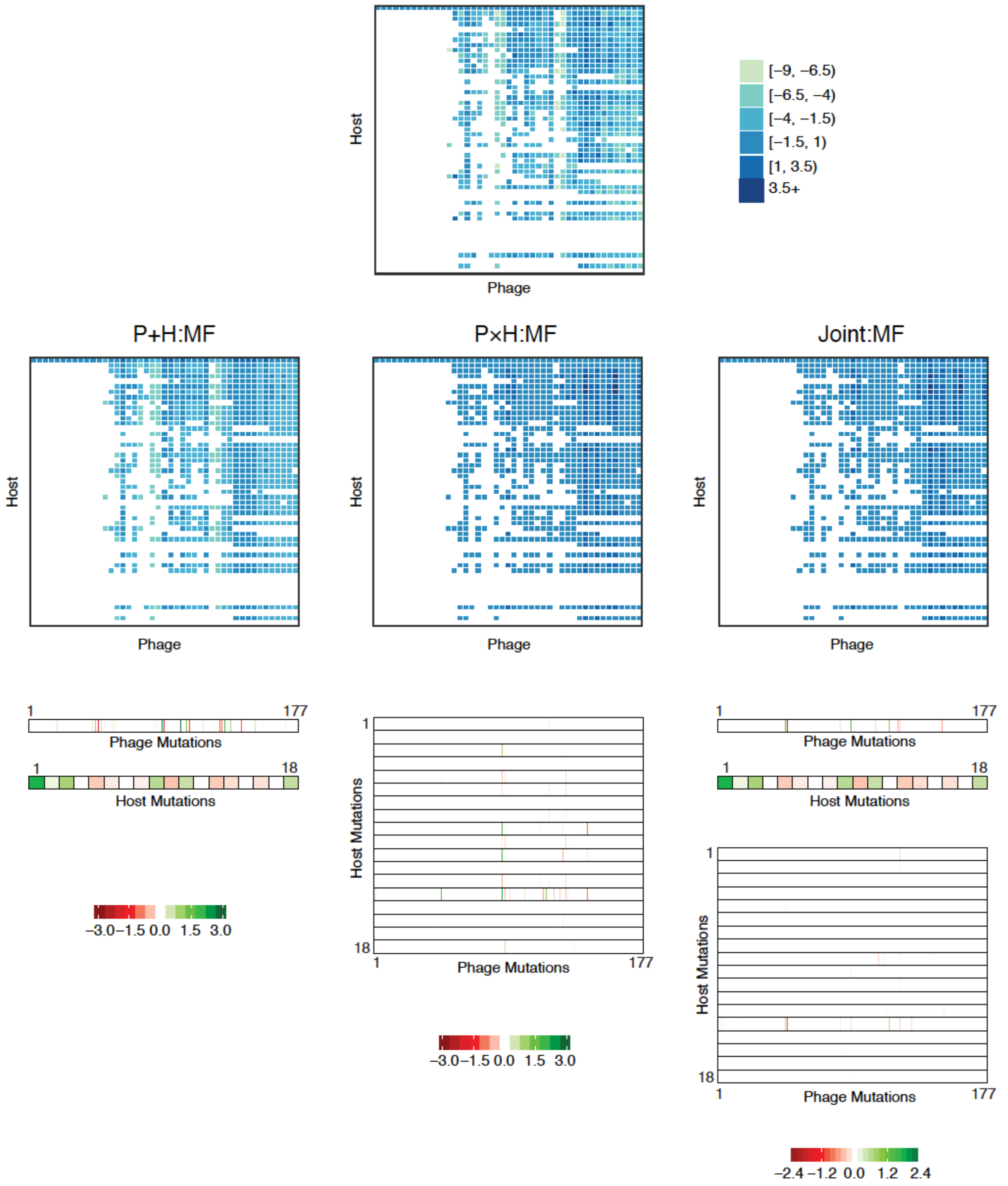

Figure 26 – Results from final model for step 2 based on P+H:MF, P×H:MF and Joint:MF in log scale

Top panel: The true log transformed phage infection efficiency based on observed positive EOP from experiment. Middle panel: The predicted log transformed phage infection efficiency based on P+H:MF, P×H:MF and Joint:MF, respectively. Bottom panel: The coefficients learned from the P+H:MF, P×H:MF and Joint:MF features, respectively.

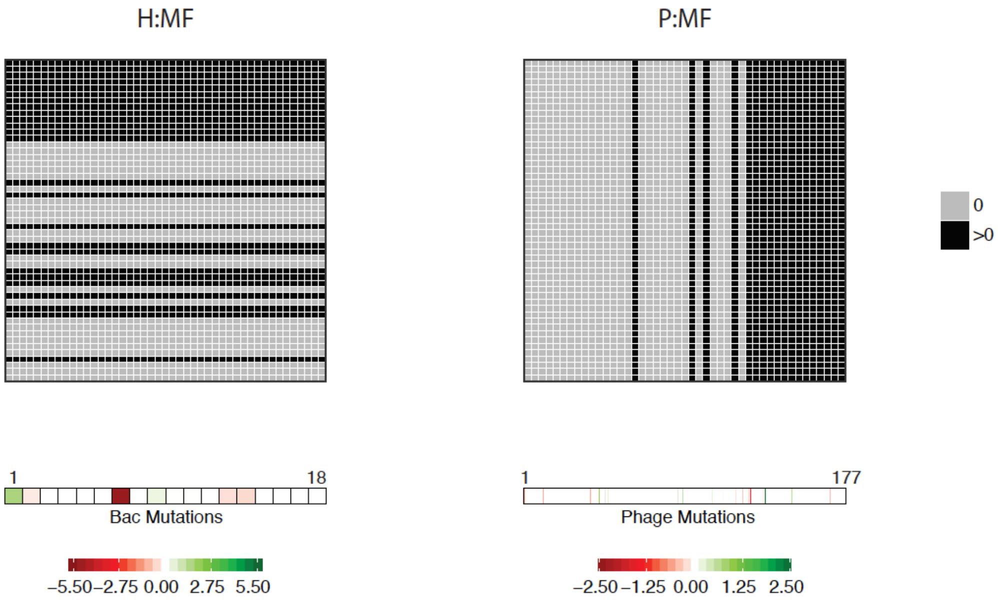

Figure 27 – Results from final model for step 1 based on H:MF and P:MF

Top panel: The predicted interaction network based on H:MF and P:MF, respectively. Bottom panel: The coefficients learned from the H:MF and P:MF features, respectively.

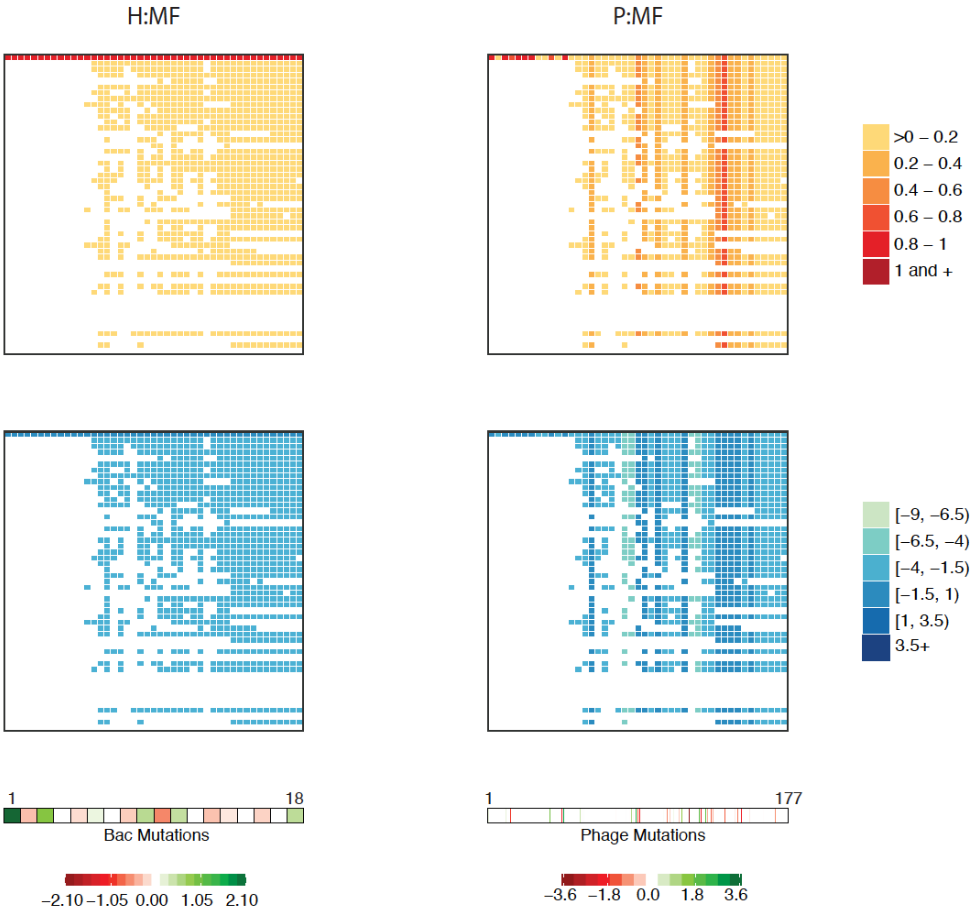

Figure 28 – Results from final model for step 2 based on H:MF and P:MF

Top panel: The predicted infection efficiency based on H:MF and P:MF, respectively. Mid panel: The predicted log transformed phage infection efficiency based on H:MF and P:MF, respectively. Bottom panel: The coefficients learned from the H:MF and P:MF features, respectively.

Table 5 – Mutation profile tables for host

position	mutation	B_D_8_1	B_D_8_2	B_D_8_3	B_D_8_4	B_D_8_5
1,003,271	G→T					
1,004,191	A→C					
1,027,154	C→A					
1,173,078	G→A					
1,368,326	C→A					
1,881,802	Δ10 bp				100%	
1,882,915	Δ16 bp					
2,103,918	(CCAG)$_{7\to8}$					
2,103,918	(CCAG)$_{7\to10}$			100%		100%
2,247,493	Δ1 bp					
2,401,525	3 bp→AA					
2,401,529	A→T					
3,023,945	Δ777 bp					
3,482,706	(AGTGGGAACTGGCGGCGGAGCTGCC)$_{1\to2}$		100%	100%	100%	100%
3,482,802	Δ141 bp					
3,482,943	A→C	100 %				
4,214,272	Δ12 bp					
4,228,027	Δ1 bp					
position	B_D_8_6	B_D_8_7	B_D_8_8	B_D_8_9	B_D_8_10	B_D_15_1
1,003,271						

1,004,191						
1,027,154						
1,173,078						
1,368,326						
1,881,802						
1,882,915						
2,103,918						
2,103,918						
2,247,493						
2,401,525	100%					
2,401,529	100%					
3,023,945						
3,482,706	100%		100%	100%	100%	100%
3,482,802						
3,482,943						
4,214,272						
4,228,027		100%				
position	**B_D_15_2**	**B_D_15_3**	**B_D_15_4**	**B_D_15_5**	**B_D_15_6**	**B_D_15_7**
1,003,271						
1,004,191						
1,027,154						100%
1,173,078						

1,368,326						
1,881,802						
1,882,915						
2,103,918	100%					
2,103,918						
2,247,493						
2,401,525						
2,401,529						
3,023,945		100%	100%	100%	100%	100%
3,482,706	100%	100%	100%	100%	100%	100%
3,482,802						
3,482,943						
4,214,272						
4,228,027						
position	**B_D_15_8**	**B_D_15_9**	**B_D_15_10**	**B_D_22_1**	**B_D_22_2**	**B_D_22_3**
1,003,271						
1,004,191						
1,027,154						
1,173,078						
1,368,326						
1,881,802						
1,882,915						

2,103,918						
2,103,918						
2,247,493						
2,401,525						
2,401,529						
3,023,945	100%	?	100%	?	100%	100%
3,482,706	100%	100%	100%	100%	100%	100%
3,482,802						
3,482,943						
4,214,272						
4,228,027						
position	**B_D_22_4**	**B_D_22_5**	**B_D_22_6**	**B_D_22_7**	**B_D_22_8**	**B_D_22_9**
1,003,271						
1,004,191						
1,027,154						
1,173,078						
1,368,326						
1,881,802						
1,882,915						
2,103,918						
2,103,918						
2,247,493						

2,401,525						
2,401,529						
3,023,945	100%	100%	?	100%	100%	?
3,482,706	100%	100%	100%	100%	100%	100%
3,482,802						
3,482,943						
4,214,272						
4,228,027						

position	**B_D_22_10**	**B_D_28_1**	**B_D_28_2**	**B_D_28_3**	**B_D_28_4**	**B_D_28_5**
1,003,271						
1,004,191			100%			
1,027,154						
1,173,078						
1,368,326						
1,881,802						
1,882,915						
2,103,918						
2,103,918						
2,247,493					100%	
2,401,525						
2,401,529						
3,023,945	?	100%			?	

3,482,706	100%	100%	100%	100%	100%	100%
3,482,802						
3,482,943						
4,214,272						
4,228,027						
position	**B_D_28_6**	**B_D_28_7**	**B_D_28_8**	**B_D_28_9**	**B_D_28_10**	**B_D_37_1**
1,003,271	100%		100%			100%
1,004,191						
1,027,154						
1,173,078			100%			
1,368,326					100%	
1,881,802						
1,882,915	100%		100%			100%
2,103,918						
2,103,918						
2,247,493						
2,401,525						
2,401,529						
3,023,945						
3,482,706		100%		100%	100%	
3,482,802			100%			
3,482,943						

4,214,272						
4,228,027						
position	**B_D_37_2**	**B_D_37_3**	**B_D_37_4**	**B_D_37_5**	**B_D_37_6**	**B_D_37_7**
1,003,271	100%	100%	100%	100%	100%	
1,004,191						
1,027,154						
1,173,078						
1,368,326						
1,881,802						
1,882,915	100%	100%	100%	100%	100%	
2,103,918						
2,103,918						
2,247,493						
2,401,525						
2,401,529						
3,023,945						
3,482,706						100%
3,482,802				?		
3,482,943						
4,214,272	100%	100%	100%	100%	100%	
4,228,027						

position	**B_D_37_8**	**B_D_37_9**	**B_D_37_10**	**annotation**	**gene**	**description**
1,003,271	100%		100%	N268K (AAC →AAA)	*ECB_RS04930* ←	phosphoporin PhoE
1,004,191				intergenic (-117/+485)	*ECB_RS04930* ← / ← *ECB_RS04935*	phosphoporin PhoE/asparagine--tRNA ligase
1,027,154				L34M (CTG →ATG)	*ECB_RS05030* →	ABC transporter ATP-binding protein
1,173,078				W214* (TGG →TAG)	*ECB_RS05820* →	PTS glucose EIICB component
1,368,326				N90K (AAC →AAA)	*ECB_RS06835* →	thiosulfate sulfurtransferase PspE
1,881,802				coding (142-151/801 nt)	*ECB_RS09445* →	PTS mannose/fructose/sorbose transporter subunit IIC
1,882,915	100%		100%	coding (442-457/852 nt)	*ECB_RS09450* →	PTS mannose transporter subunit IID
2,103,918				coding (185/216 nt)	*ECB_RS23820* →	hypothetical protein
2,103,918				coding (185/216 nt)	*ECB_RS23820* →	hypothetical protein
2,247,493				coding (141/624 nt)	*ECB_RS11220* ←	cytochrome c biogenesis ATP-binding export protein CcmA
2,401,525				coding (1297-1299/2145 nt)	*ECB_RS11915* ←	multifunctional fatty acid oxidation complex subunit alpha
2,401,529				I432N (ATC →AAC)	*ECB_RS11915* ←	multifunctional fatty acid oxidation complex subunit alpha
3,023,945		100%			*[ECB_RS14915]–[ECB_RS14925]*	*[ECB_RS14915], ECB_RS14920, [ECB_RS14925]*
3,482,706		100%		coding (1022/2706 nt)	*ECB_RS17295* →	transcriptional regulator MalT
3,482,802			100%	coding (1118-1258/2706 nt)	*ECB_RS17295* →	transcriptional regulator MalT
3,482,943				Q420P (CAA →CCA)	*ECB_RS17295* →	transcriptional regulator MalT
4,214,272	100%		100%	coding (1584-1595/1650 nt)	*ECB_RS20720* →	glucose-6-phosphate isomerase
4,228,027				coding (1125/1341 nt)	*lamB* →	maltoporin

Table 6 – Mutation profile tables for phage

position	mutation	P_D_8_1	P_D_8_2	P_D_8_3	P_D_8_4	P_D_8_5
175	T→G					
327	C→T					
332	A→G					
384	G→A					
412	G→A					
429	A→G					
483	A→G					
489	G→A					
583	C→A					
9,067	T→C					
11,451	C→T					
15,890	A→G					100%
16,218	G→T					
16,227	T→C					
16,299	A→G					
16,318	2 bp→CC					
16,350	T→C					
16,449	C→T			100%		
16,485	G→C			?		
16,497	A→G					
16,524	C→T					
16,596	G→A			?		
16,599	G→A					
16,606	2 bp→GT					
16,725	C→T					
16,774	2 bp→CT					
16,791	T→C					
16,794	T→C					
16,866	A→G					
16,869	A→G					

16,893	T→C					
16,902	C→G					
16,905	C→T					
16,908	A→C					
16,938	T→C					
16,972	A→C			100%		
16,980	T→C			?		
16,983	T→G			?		
16,986	T→C					
16,998	G→A			100%		
17,049	C→T					
17,055	T→C					
17,059	G→A					
17,081	+G					
17,082	A→C					
17,085	Δ1 bp					
17,088	C→G					
17,090	A→G					
17,136	A→G					
17,160	T→C					
17,183	A→G					
17,200	C→T					
17,211	A→C					
17,280	G→A					
17,328	A→C					
17,334	T→C					
17,343	G→A					
17,391	T→C					
17,409	T→C					
17,421	G→C					
17,424	A→C					
17,430	C→T					
17,433	A→G					

17,457	T→C					
17,466	C→T					
17,469	T→C					
17,478	2 bp→GG					
17,487	C→T					
17,494	A→C					
17,502	G→A					
17,535	A→T					
17,547	G→A					
17,556	G→T					
17,586	G→A					
17,613	T→G					
17,652	A→G					
17,659	2 bp→CA					
17,673	G→A					
17,679	C→G					
17,712	C→G					
17,721	C→T					
17,759	A→G					
17,775	A→G					
17,788	+CA					
17,793	G→A					
17,795	Δ2 bp					
17,796	T→C					
17,797	Δ1 bp					
17,805	T→C					
17,862	C→T					
17,868	T→C					
17,913	C→T					
17,916	C→T					
17,919	T→C					
17,921	G→A					
17,923	G→C					

17,928	C→T					
17,937	2 bp→AT					
17,937	4 bp→ATCC					
17,940	A→C					
17,943	T→C					
17,946	C→T					
17,950	G→A					
17,964	2 bp→AG					
18,255	G→T					
18,257	2 bp→GT					
18,265	A→G					
18,267	C→T					
18,285	C→A					
18,297	4 bp→ATAT					
18,309	C→T					
18,330	C→T					
18,342	C→A					
18,463	A→G		100%			
18,503	C→T	100%	100%	100%	100%	100%
18,535	A→C		100%			
18,538	A→G					
18,731	C→T					
18,734	T→C	100%	100%	100%	100%	100%
18,814	C→T					
18,823	G→A	100%	100%	100%	100%	100%
18,825	T→A					
18,825	T→G		100%			
18,868	A→C				100%	100%
18,868	A→G					
18,868	A→T		100%			
18,884	T→C					
19,260	T→C					
19,791	C→G					

20,200	A→G					
39,183	$(G)_{5\to6}$					
39,198	G→A					
40,140	T→C					
40,158	G→A					
40,161	C→G					
40,166	C→A					
40,189	G→A					
40,194	T→G					
40,434	T→C					
40,601	G→A					
40,612	T→C					
40,616	C→T					
40,625	A→C					
40,637	A→G					
40,663	C→T					
40,672	C→T					
40,683	2 bp→CC					
40,723	2 bp→TT					
40,898	G→C					
40,905	T→C					
40,909	T→A					
40,912	2 bp→GT					
40,919	Δ1 bp					
40,929	C→T					
40,931	T→C					
40,933	+T					
40,939	G→T					
40,946	C→G					
40,957	T→C					
40,973	A→C					
42,104	2 bp→AC					
42,115	C→T					

42,120	T→A					
42,129	T→C					
42,131	2 bp→GG					
42,165	C→T					
42,207	G→A					
42,300	C→A					
42,432	C→G					
42,434	2 bp→AG					
42,437	C→T					
42,449	T→C					
42,464	C→T					
42,472	C→T					
42,476	A→G					
42,491	T→C					
position	**P_D_8_6**	**P_D_8_7**	**P_D_8_8**	**P_D_8_9**	**P_D_8_10**	**P_D_8_11**
175						
327						
332						
384						
412						
429						
483						
489						
583						
9,067	100%					
11,451						
15,890						
16,218						
16,227						
16,299						
16,318						
16,350						

16,449						
16,485						
16,497						
16,524						
16,596						
16,599						
16,606						
16,725						
16,774						
16,791						
16,794						
16,866						
16,869						
16,893						
16,902						
16,905						
16,908						
16,938						
16,972						
16,980						
16,983						
16,986						
16,998						
17,049						
17,055						
17,059						
17,081						
17,082						
17,085						
17,088						
17,090						
17,136						
17,160						

17,183						
17,200						
17,211						
17,280						
17,328						
17,334						
17,343						
17,391						
17,409						
17,421						
17,424						
17,430						
17,433						
17,457						
17,466						
17,469						
17,478						
17,487						
17,494						
17,502						
17,535						
17,547						
17,556						
17,586						
17,613						
17,652						
17,659						
17,673						
17,679						
17,712						
17,721						
17,759						
17,775						

17,788						
17,793						
17,795						
17,796						
17,797						
17,805						
17,862						
17,868						
17,913						
17,916						
17,919						
17,921						
17,923						
17,928						
17,937						
17,937						
17,940						
17,943						
17,946						
17,950						
17,964						
18,255						
18,257						
18,265						
18,267						
18,285						
18,297						
18,309						
18,330						
18,342						
18,463						
18,503	100%	100%	100%	100%	100%	100%
18,535						

18,538						
18,731						
18,734		100%	100%		100%	100%
18,814						
18,823		100%	100%	100%	100%	100%
18,825						
18,825						
18,868						100%
18,868						
18,868						
18,884	100%					
19,260						
19,791						
20,200	100%					
39,183						
39,198						
40,140						100%
40,158						100%
40,161						100%
40,166						100%
40,189						100%
40,194						100%
40,434						100%
40,601						100%
40,612						100%
40,616						100%
40,625						100%
40,637						100%
40,663						
40,672						
40,683						
40,723						
40,898						

40,905						
40,909						
40,912						
40,919						
40,929						
40,931						
40,933						
40,939						
40,946						
40,957						
40,973						
42,104						
42,115						
42,120						
42,129						
42,131						
42,165						
42,207						100%
42,300						100%
42,432						
42,434						
42,437						
42,449						
42,464						
42,472						
42,476						
42,491						
position	**P_D_15_1**	**P_D_15_2**	**P_D_15_3**	**P_D_15_4**	**P_D_15_5**	**P_D_15_6**
175						
327						
332						
384						

412						
429						
483						
489						
583						
9,067						
11,451	100%		100%	100%		
15,890	100%		100%			
16,218	100%		100%			
16,227	100%		100%			
16,299	100%		100%			
16,318	100%		100%			
16,350	100%		100%			
16,449	100%		100%			
16,485	100%		100%			
16,497	100%		100%			
16,524	100%		100%			
16,596	100%		100%			
16,599	100%		100%			
16,606	100%		100%			
16,725	100%		100%			
16,774	100%		100%			
16,791	100%		100%			
16,794	100%		100%			
16,866	100%		100%			
16,869	100%		100%			
16,893	100%		100%			
16,902	100%		100%			
16,905	100%		100%			
16,908	100%		100%			
16,938	100%		100%			
16,972	100%		100%			
16,980	100%		100%			

16,983	100%		100%			
16,986	100%		100%			
16,998	100%		100%			
17,049			100%	100%		
17,055			100%	100%		
17,059			100%	100%		
17,081			100%	100%		
17,082			100%	100%		
17,085			100%	100%		
17,088			100%	100%		
17,090			100%	100%		
17,136			100%	100%		
17,160			100%	100%		
17,183			100%	100%		
17,200			100%	100%		
17,211			100%	100%		
17,280			100%	100%		
17,328			100%	100%		
17,334			100%	100%		
17,343			100%	100%		
17,391			100%	100%		
17,409			100%	100%		
17,421			100%	100%		
17,424			100%	100%		
17,430			100%	100%		
17,433			100%	100%		
17,457			100%	100%		
17,466			100%	100%		
17,469			100%	100%		
17,478			100%	100%		
17,487			100%	100%		
17,494			100%	100%		
17,502			100%	100%		

17,535			100%	100%		
17,547			100%	100%		
17,556			100%	100%		
17,586			100%	100%		
17,613			100%	100%		
17,652			100%	100%		
17,659			100%	100%		
17,673			100%	100%		
17,679			100%	100%		
17,712			100%			
17,721			100%	100%		
17,759			100%	100%		
17,775			100%	100%		
17,788						
17,793				?		
17,795						
17,796				100%		
17,797			100%			
17,805			100%	100%		
17,862			100%	100%		
17,868			100%	100%		
17,913			100%	100%		
17,916			100%	100%		
17,919			100%	100%		
17,921			100%	100%		
17,923			100%	100%		
17,928			100%	100%		
17,937				100%		
17,937			100%			
17,940				100%		
17,943			100%	?		
17,946			100%	?		
17,950			100%	100%		

17,964			100%	?		
18,255			100%	?		
18,257			100%	?		
18,265			100%	100%		
18,267			100%	100%		
18,285			100%	100%		
18,297			100%	100%		
18,309			100%			
18,330			100%			
18,342			100%			
18,463						
18,503	100%	100%	100%	100%	100%	100%
18,535						
18,538	100%		100%	100%		
18,731						
18,734	100%	100%	100%		100%	100%
18,814	100%		100%	100%		
18,823	100%	100%	100%	100%	100%	100%
18,825	100%		100%	100%		
18,825						
18,868			100%			
18,868	100%					
18,868				100%		
18,884						
19,260						
19,791						
20,200						
39,183						
39,198						
40,140						
40,158						
40,161						
40,166						

40,189						
40,194						
40,434						
40,601						
40,612						
40,616						
40,625						
40,637						
40,663						
40,672						
40,683						
40,723						
40,898						
40,905						
40,909						
40,912						
40,919						
40,929						
40,931						
40,933						
40,939						
40,946						
40,957						
40,973						
42,104						
42,115						
42,120						
42,129						
42,131						
42,165						
42,207						
42,300						
42,432						

42,434						
42,437						
42,449						
42,464						
42,472						
42,476						
42,491						
position	**P_D_15_7**	**P_D_15_8**	**P_D_15_9**	**P_D_15_10**	**P_D_15_11**	**P_D_22_1**
175						
327						
332						
384						
412						
429						
483						
489						
583						
9,067						
11,451	100%		100%		100%	100%
15,890			100%	100%	100%	
16,218			100%	100%	100%	
16,227			100%	100%	100%	
16,299			100%	100%	100%	
16,318			100%	100%	100%	
16,350			100%	100%	100%	
16,449			100%	100%	100%	
16,485			100%	100%	100%	
16,497			100%	100%	100%	
16,524			100%	100%	100%	
16,596			100%	100%	100%	
16,599			100%	100%	100%	
16,606			100%	100%	100%	

16,725		100%	100%	100%	100%	
16,774		100%	100%	100%	100%	
16,791		100%	100%	100%	100%	
16,794		100%	100%	100%	100%	
16,866		100%	100%	100%	100%	
16,869		100%	100%	100%	100%	
16,893		100%	100%	100%	100%	
16,902		100%	100%	100%	100%	
16,905		100%	100%	100%	100%	
16,908		100%	100%	100%	100%	
16,938		100%	100%	100%	100%	
16,972		100%	100%	100%	100%	
16,980		100%	100%	100%	100%	
16,983		100%	100%	100%	100%	
16,986		100%	100%	100%	100%	
16,998			100%	100%	100%	
17,049			100%	100%	100%	100%
17,055			100%	100%	100%	100%
17,059			100%	100%	100%	100%
17,081			100%		100%	100%
17,082			100%	100%	100%	100%
17,085			100%	100%	100%	100%
17,088			100%	100%	100%	100%
17,090			100%	100%	100%	100%
17,136			100%	100%	100%	100%
17,160			100%	100%	100%	100%
17,183			100%	100%	100%	100%
17,200			100%	100%	100%	100%
17,211			100%	100%	100%	100%
17,280			100%	100%	100%	100%
17,328			100%	100%	100%	100%
17,334			100%	100%	100%	100%
17,343			100%	100%	100%	100%

17,391			100%	100%	100%	100%
17,409			100%	100%	100%	100%
17,421			100%	100%	100%	100%
17,424			100%	100%	100%	100%
17,430			100%	100%	100%	100%
17,433			100%	100%	100%	100%
17,457			100%	100%	100%	100%
17,466			100%	100%	100%	100%
17,469			100%	100%	100%	100%
17,478			100%	100%	100%	100%
17,487			100%	100%	100%	100%
17,494			100%	100%	100%	100%
17,502			100%	100%	100%	100%
17,535			100%	100%	100%	100%
17,547			100%	100%	100%	100%
17,556			100%	100%	100%	100%
17,586			100%	100%	100%	100%
17,613			100%	100%	100%	100%
17,652			100%	100%	100%	100%
17,659			100%	100%	100%	100%
17,673			100%	100%	100%	100%
17,679			100%	100%	100%	100%
17,712			100%	100%	100%	
17,721			100%	100%	100%	100%
17,759			100%	100%	100%	100%
17,775			100%	100%	100%	100%
17,788			100%		100%	
17,793						
17,795			100%		100%	
17,796			Δ	100%	Δ	?
17,797						
17,805			100%	100%	100%	100%
17,862			100%	100%	100%	100%

17,868			100%	100%	100%	100%
17,913				?		100%
17,916				?		100%
17,919						100%
17,921				?		100%
17,923				?		100%
17,928				?		100%
17,937						
17,937						100%
17,940						
17,943						100%
17,946						100%
17,950						100%
17,964						100%
18,255						100%
18,257						100%
18,265				?		100%
18,267				?		100%
18,285				100%		100%
18,297				100%		100%
18,309			100%	100%	100%	
18,330				100%		
18,342				100%		
18,463						
18,503	100%	100%	100%	100%	100%	100%
18,535						
18,538	100%	100%	100%	100%	100%	100%
18,731	100%					
18,734		100%	100%		100%	
18,814		100%	100%		100%	100%
18,823	100%	100%	100%	100%	100%	100%
18,825		100%	100%		100%	100%
18,825						

18,868			100%		100%	
18,868						
18,868						100%
18,884						
19,260						100%
19,791		100%		100%		
20,200						
39,183		100%				100%
39,198						
40,140						
40,158						
40,161						
40,166						
40,189						
40,194						
40,434			100%			
40,601			100%			
40,612			100%			
40,616			100%			
40,625			100%			
40,637			100%			
40,663			100%			
40,672			100%			
40,683			100%			
40,723			100%			
40,898			100%			
40,905			100%			
40,909			100%			
40,912			100%			
40,919			100%			
40,929			100%			
40,931			100%			
40,933			100%			

40,939			100%			
40,946			100%			
40,957			100%			
40,973			100%			
42,104			100%			
42,115			100%			
42,120			100%			
42,129			100%			
42,131			100%			
42,165			100%			
42,207			100%			
42,300			100%			
42,432			100%			
42,434			100%			
42,437			100%			
42,449			100%			
42,464			100%			
42,472			100%			
42,476			100%			
42,491			100%			
position	**P_D_22_2**	**P_D_22_3**	**P_D_22_4**	**P_D_22_5**	**P_D_22_6**	**P_D_22_7**
175						
327						
332						
384						
412						
429						
483						
489						
583						
9,067						
11,451						100%

15,890						
16,218						
16,227						
16,299						
16,318						
16,350						
16,449						
16,485						
16,497						
16,524						
16,596						
16,599						
16,606						
16,725						
16,774						
16,791						
16,794						
16,866						
16,869						
16,893						
16,902						
16,905						
16,908						
16,938						
16,972						
16,980						
16,983						
16,986						
16,998						
17,049						100%
17,055						100%
17,059						?
17,081						

17,082						
17,085						
17,088						
17,090						
17,136						
17,160						
17,183						
17,200						
17,211						
17,280						100%
17,328						?
17,334						100%
17,343						100%
17,391						100%
17,409						100%
17,421						100%
17,424						100%
17,430						100%
17,433						100%
17,457						100%
17,466						100%
17,469						100%
17,478						100%
17,487						100%
17,494						100%
17,502						100%
17,535						100%
17,547						100%
17,556						100%
17,586						100%
17,613						
17,652						
17,659						

17,673						?
17,679						?
17,712						
17,721						100%
17,759						100%
17,775						?
17,788						
17,793						
17,795						
17,796						
17,797						
17,805						
17,862						100%
17,868						100%
17,913						
17,916						
17,919						
17,921						
17,923						
17,928						
17,937						
17,937						
17,940						
17,943						
17,946						
17,950						
17,964						
18,255						
18,257						
18,265						
18,267						
18,285						
18,297						

18,309						
18,330						
18,342						
18,463						
18,503	100%	100%	100%	100%	100%	100%
18,535						
18,538						100%
18,731						
18,734	100%		100%	100%	100%	100%
18,814						
18,823	100%		100%	100%	100%	100%
18,825						
18,825						
18,868						
18,868						
18,868						
18,884		100%				
19,260						
19,791						
20,200						
39,183						
39,198		100%				
40,140						
40,158						
40,161						
40,166						
40,189						
40,194						
40,434						
40,601						
40,612						
40,616						
40,625						

40,637						
40,663						
40,672						
40,683						
40,723						
40,898						
40,905						
40,909						
40,912						
40,919						
40,929						
40,931						
40,933						
40,939						
40,946						
40,957						
40,973						
42,104						
42,115						
42,120						
42,129						
42,131						
42,165						
42,207						
42,300						
42,432						
42,434						
42,437						
42,449						
42,464						
42,472						
42,476						
42,491						

position	**P_D_22_8**	**P_D_22_9**	**P_D_22_10**	**P_D_22_11**	**P_D_28_1**	**P_D_28_2**
175						
327						
332						
384						
412						
429						
483						
489						
583						
9,067						
11,451	100%	100%	100%	100%	100%	100%
15,890						
16,218						
16,227						
16,299						
16,318						
16,350						
16,449						
16,485						
16,497						
16,524						
16,596						
16,599						
16,606						
16,725						
16,774						
16,791						
16,794						
16,866						
16,869						
16,893						

16,902						
16,905						
16,908						
16,938						
16,972						
16,980						
16,983						
16,986						
16,998						
17,049	100%	100%	100%	100%	100%	100%
17,055	100%	100%	100%	100%	100%	100%
17,059	100%	100%	100%	100%	100%	100%
17,081		100%	100%		100%	100%
17,082		100%	100%	100%	100%	100%
17,085		100%	100%	100%	100%	100%
17,088	100%	100%	100%	100%	100%	100%
17,090	100%	100%	100%	100%	100%	100%
17,136	100%	100%	100%	100%	100%	100%
17,160	100%	100%	100%	100%	100%	100%
17,183	100%	100%	100%	100%	100%	100%
17,200	100%	100%	100%	100%	100%	100%
17,211	100%	100%	100%	100%	100%	100%
17,280	100%	100%	100%	100%	100%	100%
17,328	100%	100%	100%	100%	100%	100%
17,334	100%	100%	100%	100%	100%	100%
17,343	100%	100%	100%	100%	100%	100%
17,391	100%	100%	100%	100%	100%	100%
17,409	100%	100%	100%	100%	100%	100%
17,421	100%	100%	100%	100%	100%	100%
17,424	100%	100%	100%	100%	100%	100%
17,430	100%	100%	100%	100%	100%	100%
17,433	100%	100%	100%	100%	100%	100%
17,457	100%	100%	100%	100%	100%	100%

17,466	100%	100%	100%	100%	100%	100%
17,469	100%	100%	100%	100%	100%	100%
17,478	100%	100%	100%	100%	100%	100%
17,487	100%	100%	100%	100%	100%	100%
17,494	100%	100%	100%	100%	100%	100%
17,502	100%	100%	100%	100%	100%	100%
17,535	100%	100%	100%	100%	100%	100%
17,547	100%	100%	100%	100%	100%	100%
17,556	100%	100%	100%	100%	100%	100%
17,586	100%	100%	100%	100%	100%	100%
17,613	100%	100%	100%	100%	100%	100%
17,652	100%	100%	100%	100%	100%	100%
17,659	100%	100%	100%	100%	100%	100%
17,673	100%	100%	100%	100%	100%	100%
17,679	100%	100%	100%	100%	100%	100%
17,712						
17,721	100%	100%	100%	100%	100%	100%
17,759	100%	100%	100%	100%	100%	100%
17,775	100%	100%	100%	100%	100%	100%
17,788						100%
17,793				100%	?	
17,795			100%			100%
17,796	?		Δ	100%	100%	Δ
17,797		100%				
17,805	100%	100%	100%	100%	100%	100%
17,862	100%	100%	100%	100%	100%	100%
17,868	100%	100%	100%	100%	100%	100%
17,913	?	100%		100%	100%	
17,916		100%		100%	100%	
17,919		100%		100%	100%	
17,921		100%		100%	100%	
17,923		100%		100%	100%	
17,928		100%		100%	100%	

17,937		?				
17,937		?		100%	100%	
17,940		?				
17,943		?		100%	100%	
17,946		?		100%	100%	
17,950		?		100%	100%	
17,964				100%	100%	
18,255		100%		100%	100%	
18,257		100%		100%	100%	
18,265		100%		100%	100%	
18,267		100%		100%	100%	
18,285	100%	100%		100%	100%	
18,297	100%	100%		100%	100%	
18,309						
18,330						
18,342						
18,463						
18,503	100%	100%	100%	100%	100%	100%
18,535						
18,538	100%	100%	100%	100%	100%	100%
18,731						
18,734	100%	100%	100%			
18,814	100%	100%	100%	100%	100%	100%
18,823	100%	100%	100%	100%	100%	100%
18,825		100%	100%	100%	100%	100%
18,825						
18,868		100%	100%			
18,868						
18,868				100%	100%	100%
18,884						
19,260		100%		100%		100%
19,791						
20,200						

39,183	100%	100%	100%	100%	100%	100%
39,198						
40,140						
40,158						
40,161						
40,166						
40,189						
40,194						
40,434						
40,601						
40,612						
40,616						
40,625						
40,637						
40,663	100%					
40,672	100%					
40,683	?					
40,723						
40,898						
40,905						
40,909						
40,912						
40,919						
40,929						
40,931						
40,933						
40,939						
40,946						
40,957						
40,973						
42,104						
42,115						
42,120						

42,129						
42,131						
42,165	100%					
42,207	100%					
42,300	100%					
42,432						
42,434						
42,437						
42,449						
42,464						
42,472						
42,476						
42,491						
position	**P_D_28_3**	**P_D_28_4**	**P_D_28_5**	**P_D_28_6**	**P_D_28_7**	**P_D_28_8**
175						100%
327						100%
332						100%
384						100%
412						100%
429						100%
483						100%
489						100%
583						100%
9,067						
11,451	100%	100%	100%	100%	100%	100%
15,890						
16,218						
16,227						
16,299						
16,318						
16,350						
16,449						

16,485						
16,497						
16,524						
16,596						
16,599						
16,606						
16,725						
16,774						
16,791						
16,794						
16,866						
16,869						
16,893						
16,902						
16,905						
16,908						
16,938						
16,972						
16,980						
16,983						
16,986						
16,998						
17,049	100%	100%	100%	100%	100%	100%
17,055	100%	100%	100%	100%	100%	100%
17,059	100%	100%	100%	100%	100%	100%
17,081	100%	100%	100%	100%	100%	100%
17,082	100%	100%	100%	100%	100%	100%
17,085	100%	100%	100%	100%	100%	100%
17,088	100%	100%	100%	100%	100%	100%
17,090	100%	100%	100%	100%	100%	100%
17,136	100%	100%	100%	100%	100%	100%
17,160	100%	100%	100%	100%	100%	100%
17,183	100%	100%	100%	100%	100%	100%

17,200	100%	100%	100%	100%	100%	100%
17,211	100%	100%	100%	100%	100%	100%
17,280	100%	100%	100%	100%	100%	100%
17,328	100%	100%	100%	100%	100%	100%
17,334	100%	100%	100%	100%	100%	100%
17,343	100%	100%	100%	100%	100%	100%
17,391	100%	100%	100%	100%	100%	100%
17,409	100%	100%	100%	100%	100%	100%
17,421	100%	100%	100%	100%	100%	100%
17,424	100%	100%	100%	100%	100%	100%
17,430	100%	100%	100%	100%	100%	100%
17,433	100%	100%	100%	100%	100%	100%
17,457	100%	100%	100%	100%	100%	100%
17,466	100%	100%	100%	100%	100%	100%
17,469	100%	100%	100%	100%	100%	100%
17,478	100%	100%	100%	100%	100%	100%
17,487	100%	100%	100%	100%	100%	100%
17,494	100%	100%	100%	100%	100%	100%
17,502	100%	100%	100%	100%	100%	100%
17,535	100%	100%	100%	100%	100%	100%
17,547	100%	100%	100%	100%	100%	100%
17,556	100%	100%	100%	100%	100%	100%
17,586	100%	100%	100%	100%	100%	100%
17,613	100%	100%	100%	100%	100%	100%
17,652	100%	100%	100%	100%	100%	100%
17,659	100%	100%	100%	100%	100%	100%
17,673	100%	100%	100%	100%	100%	100%
17,679	100%	100%	100%	100%	100%	100%
17,712						
17,721	100%	100%	100%	100%	100%	100%
17,759	100%	100%	100%	100%	100%	100%
17,775	100%	100%	100%	100%	100%	100%
17,788					100%	

17,793		?	100%			
17,795		100%		100%	100%	100%
17,796	100%	Δ	100%	Δ	Δ	Δ
17,797						
17,805	100%	100%	100%	100%	100%	100%
17,862	100%	100%	100%	100%	100%	100%
17,868	100%	100%	100%	100%	100%	100%
17,913						
17,916						
17,919						
17,921						
17,923						
17,928						
17,937						
17,937						
17,940						
17,943						
17,946						
17,950						
17,964						
18,255						
18,257						
18,265						
18,267						
18,285		100%				
18,297		100%				
18,309						
18,330						
18,342						
18,463						
18,503	100%	100%	100%	100%	100%	100%
18,535						
18,538	100%	100%	100%	100%	100%	100%

18,731						
18,734		100%	100%	100%		100%
18,814	100%	100%	100%	100%	100%	100%
18,823	100%	100%	100%	100%	100%	100%
18,825	100%	100%	100%	100%	100%	100%
18,825						
18,868		100%	100%	100%		100%
18,868						
18,868	100%				100%	
18,884						
19,260	100%	100%	100%	100%		100%
19,791						
20,200						
39,183	100%	100%	100%	100%	100%	100%
39,198						
40,140			100%			100%
40,158			100%			100%
40,161			100%			100%
40,166			100%			100%
40,189			100%			100%
40,194			100%			100%
40,434			100%			100%
40,601	100%		100%			100%
40,612	100%		100%			100%
40,616	100%		100%			100%
40,625	100%		100%			100%
40,637	100%		100%			100%
40,663	100%		100%			100%
40,672	100%		100%			100%
40,683	100%		100%			100%
40,723	100%					
40,898	?					
40,905	?					

40,909	?					
40,912	?					
40,919	100%					
40,929	?					
40,931	?					
40,933	100%					
40,939	100%					
40,946	100%					
40,957	100%					
40,973	?					
42,104	?					
42,115	100%		100%			
42,120	100%		100%			
42,129	100%		100%			
42,131	100%		100%			
42,165	100%		100%			100%
42,207	100%		100%			100%
42,300			100%			100%
42,432			100%			100%
42,434			100%			100%
42,437			100%			100%
42,449						100%
42,464						100%
42,472						100%
42,476						100%
42,491						100%
position	**P_D_28_9**	**P_D_28_10**	**P_D_28_11**	**annotation**	**gene**	**description**
175				intergenic (–/-15)	– / → *nu1*	–/DNA packaging protein
327				V46V (GTC→GTT)	*nu1* →	DNA packaging protein
332				K48R (AAA→AGA)	*nu1* →	DNA packaging protein
384				E65E (GAG→GAA)	*nu1* →	DNA packaging protein
412		100%		A75T (GCA→ACA)	*nu1* →	DNA packaging protein

429		100%		G80G (GGA→GGG)	*nu1* →	DNA packaging protein
483		100%		E98E (GAA→GAG)	*nu1* →	DNA packaging protein
489		100%		K100K (AAG→AAA)	*nu1* →	DNA packaging protein
583		100%		L132I (CTC→ATC)	*nu1* →	DNA packaging protein
9,067				R38R (CGT→CGC)	*V* →	tail component
11,451	100%	100%	100%	A304V (GCA→GTA)	*H* →	tail component
15,890				D129G (GAC→GGC)	*J* →	tail:host specificity protein
16,218				L238L (CTG→CTT)	*J* →	tail:host specificity protein
16,227				R241R (CGT→CGC)	*J* →	tail:host specificity protein
16,299				K265K (AAA→AAG)	*J* →	tail:host specificity protein
16,318				coding (814-815/3399 nt)	*J* →	tail:host specificity protein
16,350				H282H (CAT→CAC)	*J* →	tail:host specificity protein
16,449				G315G (GGC→GGT)	*J* →	tail:host specificity protein
16,485				A327A (GCG→GCC)	*J* →	tail:host specificity protein
16,497				T331T (ACA→ACG)	*J* →	tail:host specificity protein
16,524				S340S (AGC→AGT)	*J* →	tail:host specificity protein
16,596				P364P (CCG→CCA)	*J* →	tail:host specificity protein
16,599				S365S (TCG→TCA)	*J* →	tail:host specificity protein
16,606				coding (1102-1103/3399 nt)	*J* →	tail:host specificity protein
16,725				N407N (AAC→AAT)	*J* →	tail:host specificity protein
16,774				coding (1270-1271/3399 nt)	*J* →	tail:host specificity protein
16,791				N429N (AAT→AAC)	*J* →	tail:host specificity protein
16,794				V430V (GTT→GTC)	*J* →	tail:host specificity protein
16,866				T454T (ACA→ACG)	*J* →	tail:host specificity protein
16,869				E455E (GAA→GAG)	*J* →	tail:host specificity protein
16,893				D463D (GAT→GAC)	*J* →	tail:host specificity protein
16,902				V466V (GTC→GTG)	*J* →	tail:host specificity protein
16,905				G467G (GGC→GGT)	*J* →	tail:host specificity protein
16,908				A468A (GCA→GCC)	*J* →	tail:host specificity protein
16,938				V478V (GTT→GTC)	*J* →	tail:host specificity protein
16,972				S490R (AGC→CGC)	*J* →	tail:host specificity protein
16,980				G492G (GGT→GGC)	*J* →	tail:host specificity protein
16,983				G493G (GGT→GGG)	*J* →	tail:host specificity protein

16,986				R494R (CGT→CGC)	*J* →	tail:host specificity protein
16,998				V498V (GTG→GTA)	*J* →	tail:host specificity protein
17,049	100%	100%	100%	S515S (TCC→TCT)	*J* →	tail:host specificity protein
17,055	100%	100%	100%	G517G (GGT→GGC)	*J* →	tail:host specificity protein
17,059	100%	100%	100%	A519T (GCG→ACG)	*J* →	tail:host specificity protein
17,081	100%	100%	100%	coding (1577/3399 nt)	*J* →	tail:host specificity protein
17,082	100%	100%	100%	G526G (GGA→GGC)	*J* →	tail:host specificity protein
17,085	100%	100%	100%	coding (1581/3399 nt)	*J* →	tail:host specificity protein
17,088	100%	100%	100%	G528G (GGC→GGG)	*J* →	tail:host specificity protein
17,090	100%	100%	100%	N529S (AAT→AGT)	*J* →	tail:host specificity protein
17,136	100%	100%	100%	V544V (GTA→GTG)	*J* →	tail:host specificity protein
17,160	100%	100%	100%	G552G (GGT→GGC)	*J* →	tail:host specificity protein
17,183	100%	100%	100%	E560G (GAG→GGG)	*J* →	tail:host specificity protein
17,200	100%	100%	100%	L566L (CTG→TTG)	*J* →	tail:host specificity protein
17,211	100%	100%	100%	R569R (CGA→CGC)	*J* →	tail:host specificity protein
17,280	100%	100%	100%	V592V (GTG→GTA)	*J* →	tail:host specificity protein
17,328	100%	100%	100%	E608D (GAA→GAC)	*J* →	tail:host specificity protein
17,334	100%	100%	100%	S610S (AGT→AGC)	*J* →	tail:host specificity protein
17,343	100%	100%	100%	V613V (GTG→GTA)	*J* →	tail:host specificity protein
17,391	100%	100%	100%	T629T (ACT→ACC)	*J* →	tail:host specificity protein
17,409	100%	100%	100%	Y635Y (TAT→TAC)	*J* →	tail:host specificity protein
17,421	100%	100%	100%	A639A (GCG→GCC)	*J* →	tail:host specificity protein
17,424	100%	100%	100%	R640R (CGA→CGC)	*J* →	tail:host specificity protein
17,430	100%	100%	100%	D642D (GAC→GAT)	*J* →	tail:host specificity protein
17,433	100%	100%	100%	T643T (ACA→ACG)	*J* →	tail:host specificity protein
17,457	100%	100%	100%	S651S (AGT→AGC)	*J* →	tail:host specificity protein
17,466	100%	100%	100%	L654L (CTC→CTT)	*J* →	tail:host specificity protein
17,469	100%	100%	100%	R655R (CGT→CGC)	*J* →	tail:host specificity protein
17,478	100%	100%	100%	coding (1974-1975/3399 nt)	*J* →	tail:host specificity protein
17,487	100%	100%	100%	D661D (GAC→GAT)	*J* →	tail:host specificity protein
17,494	100%	100%	100%	S664R (AGT→CGT)	*J* →	tail:host specificity protein
17,502	100%	100%	100%	R666R (CGG→CGA)	*J* →	tail:host specificity protein
17,535	100%	100%	100%	T677T (ACA→ACT)	*J* →	tail:host specificity protein

17,547	100%	100%	100%	T681T (ACG→ACA)	*J* →	tail:host specificity protein
17,556	100%	100%	100%	A684A (GCG→GCT)	*J* →	tail:host specificity protein
17,586	100%	100%	100%	A694A (GCG→GCA)	*J* →	tail:host specificity protein
17,613	100%	100%	100%	D703E (GAT→GAG)	*J* →	tail:host specificity protein
17,652	100%	100%	100%	A716A (GCA→GCG)	*J* →	tail:host specificity protein
17,659	100%	100%	100%	coding (2155-2156/3399 nt)	*J* →	tail:host specificity protein
17,673	100%	100%	100%	T723T (ACG→ACA)	*J* →	tail:host specificity protein
17,679	100%	100%	100%	G725G (GGC→GGG)	*J* →	tail:host specificity protein
17,712				A736A (GCC→GCG)	*J* →	tail:host specificity protein
17,721	100%	100%	100%	D739D (GAC→GAT)	*J* →	tail:host specificity protein
17,759	100%	100%	100%	Q752R (CAG→CGG)	*J* →	tail:host specificity protein
17,775	100%	100%	100%	R757R (AGA→AGG)	*J* →	tail:host specificity protein
17,788				coding (2284/3399 nt)	*J* →	tail:host specificity protein
17,793	100%	?		T763T (ACG→ACA)	*J* →	tail:host specificity protein
17,795			100%	coding (2291-2292/3399 nt)	*J* →	tail:host specificity protein
17,796	100%	100%	Δ	R764R (CGT→CGC)	*J* →	tail:host specificity protein
17,797				coding (2293/3399 nt)	*J* →	tail:host specificity protein
17,805	100%	100%	100%	G767G (GGT→GGC)	*J* →	tail:host specificity protein
17,862	100%	100%	100%	Y786Y (TAC→TAT)	*J* →	tail:host specificity protein
17,868	100%	100%	100%	Y788Y (TAT→TAC)	*J* →	tail:host specificity protein
17,913				A803A (GCC→GCT)	*J* →	tail:host specificity protein
17,916				V804V (GTC→GTT)	*J* →	tail:host specificity protein
17,919				G805G (GGT→GGC)	*J* →	tail:host specificity protein
17,921				R806Q (CGG→CAG)	*J* →	tail:host specificity protein
17,923				A807P (GCG→CCG)	*J* →	tail:host specificity protein
17,928				S808S (AGC→AGT)	*J* →	tail:host specificity protein
17,937				coding (2433-2434/3399 nt)	*J* →	tail:host specificity protein
17,937				coding (2433-2436/3399 nt)	*J* →	tail:host specificity protein
17,940				E812D (GAA→GAC)	*J* →	tail:host specificity protein
17,943				G813G (GGT→GGC)	*J* →	tail:host specificity protein
17,946				Y814Y (TAC→TAT)	*J* →	tail:host specificity protein
17,950				D816N (GAT→AAT)	*J* →	tail:host specificity protein
17,964				coding (2460-2461/3399 nt)	*J* →	tail:host specificity protein

18,255				V917V (GTG→GTT)	*J* →	tail:host specificity protein
18,257				coding (2753-2754/ 3399 nt)	*J* →	tail:host specificity protein
18,265				N921D (AAC→GAT)	*J* →	tail:host specificity protein
18,267				N921D (AAC→GAT)	*J* →	tail:host specificity protein
18,285				D927E (GAC→GAA)	*J* →	tail:host specificity protein
18,297				coding (2793-2796/ 3399 nt)	*J* →	tail:host specificity protein
18,309				A935A (GCC→GCT)	*J* →	tail:host specificity protein
18,330				G942G (GGC→GGT)	*J* →	tail:host specificity protein
18,342				A946A (GCC→GCA)	*J* →	tail:host specificity protein
18,463				T987A (ACG→GCG)	*J* →	tail:host specificity protein
18,503	100%	100%	100%	A1000V (GCG→GTG)	*J* →	tail:host specificity protein
18,535				S1011R (AGC→CGC)	*J* →	tail:host specificity protein
18,538	100%	100%	100%	S1012G (AGT→GGT)	*J* →	tail:host specificity protein
18,731				A1076V (GCG→GTG)	*J* →	tail:host specificity protein
18,734	100%	100%	100%	V1077A (GTA→GCA)	*J* →	tail:host specificity protein
18,814	100%	100%	100%	H1104Y (CAT→TAT)	*J* →	tail:host specificity protein
18,823	100%	100%	100%	D1107K (GAT→AAG)	*J* →	tail:host specificity protein
18,825	100%	100%	100%	D1107K (GAT→AAA)	*J* →	tail:host specificity protein
18,825				D1107K (GAT→AAG)	*J* →	tail:host specificity protein
18,868	100%	100%		I1122L (ATT→CTT)	*J* →	tail:host specificity protein
18,868				I1122V (ATT→GTT)	*J* →	tail:host specificity protein
18,868			100%	I1122F (ATT→TTT)	*J* →	tail:host specificity protein
18,884				L1127P (CTG→CCG)	*J* →	tail:host specificity protein
19,260		100%	100%	L99P (CTG→CCG)	*lom* →	outer host membrane
19,791				R48G (CGT→GGT)	*orf-401* →	Tail fiber protein
20,200				E184G (GAA→GGA)	*orf-401* →	Tail fiber protein
39,183	100%	100%	100%	intergenic (+364/-7)	*orf-64* → / → *S*	hypothetical protein/anti-holin
39,198				M3I (ATG→ATA)	*S* →	anti-holin
40,140				R57R (CGT→CGC)	*Rz* →	cell lysis protein
40,158				A63A (GCG→GCA)	*Rz* →	cell lysis protein
40,161				L64L (CTC→CTG)	*Rz* →	cell lysis protein
40,166				A66E (GCA→GAA)	*Rz* →	cell lysis protein
40,189		100%		D74N (GAT→AAT)	*Rz* →	cell lysis protein

40,194		100%		A75A (GCT→GCG)	*Rz* →	cell lysis protein
40,434		100%		intergenic (+3/+29)	*Rz* → / ← *bor*	cell lysis protein/Bor protein precursor
40,601		100%		V52V (GTC→GTT)	*bor* ←	Bor protein precursor
40,612		100%		K49E (AAG→GAG)	*bor* ←	Bor protein precursor
40,616		100%		G47G (GGG→GGA)	*bor* ←	Bor protein precursor
40,625		100%		S44S (TCT→TCG)	*bor* ←	Bor protein precursor
40,637		100%		H40H (CAT→CAC)	*bor* ←	Bor protein precursor
40,663		100%		A32T (GCA→ACA)	*bor* ←	Bor protein precursor
40,672		100%		A29T (GCA→ACA)	*bor* ←	Bor protein precursor
40,683		100%		coding (73-74/294 nt)	*bor* ←	Bor protein precursor
40,723				coding (33-34/294 nt)	*bor* ←	Bor protein precursor
40,898				intergenic (-142/+149)	*bor* ← / ← *lambdap78*	Bor protein precursor/putative envelope protein
40,905				intergenic (-149/+142)	*bor* ← / ← *lambdap78*	Bor protein precursor/putative envelope protein
40,909				intergenic (-153/+138)	*bor* ← / ← *lambdap78*	Bor protein precursor/putative envelope protein
40,912				intergenic (-156/+134)	*bor* ← / ← *lambdap78*	Bor protein precursor/putative envelope protein
40,919				intergenic (-163/+128)	*bor* ← / ← *lambdap78*	Bor protein precursor/putative envelope protein
40,929				intergenic (-173/+118)	*bor* ← / ← *lambdap78*	Bor protein precursor/putative envelope protein
40,931				intergenic (-175/+116)	*bor* ← / ← *lambdap78*	Bor protein precursor/putative envelope protein
40,933				intergenic (-177/+114)	*bor* ← / ← *lambdap78*	Bor protein precursor/putative envelope protein
40,939				intergenic (-183/+108)	*bor* ← / ← *lambdap78*	Bor protein precursor/putative envelope protein
40,946				intergenic (-190/+101)	*bor* ← / ← *lambdap78*	Bor protein precursor/putative envelope protein
40,957				intergenic (-201/+90)	*bor* ← / ← *lambdap78*	Bor protein precursor/putative envelope protein
40,973				intergenic (-217/+74)	*bor* ← / ← *lambdap78*	Bor protein precursor/putative envelope protein
42,104				intergenic (+155/–)	*lambdap79* → / –	hypothetical protein/–
42,115				intergenic (+166/–)	*lambdap79* → / –	hypothetical protein/–
42,120				intergenic (+171/–)	*lambdap79* → / –	hypothetical protein/–
42,129				intergenic (+180/–)	*lambdap79* → / –	hypothetical protein/–
42,131				intergenic (+182/–)	*lambdap79* → / –	hypothetical protein/–
42,165		100%		intergenic (+216/–)	*lambdap79* → / –	hypothetical protein/–
42,207		100%		intergenic (+258/–)	*lambdap79* → / –	hypothetical protein/–
42,300		100%		intergenic (+351/–)	*lambdap79* → / –	hypothetical protein/–
42,432		100%		intergenic (+483/–)	*lambdap79* → / –	hypothetical protein/–
42,434		100%		intergenic (+485/–)	*lambdap79* → / –	hypothetical protein/–

42,437		100%		intergenic (+488/–)	*lambdap79* → / –	hypothetical protein/–
42,449		100%		intergenic (+500/–)	*lambdap79* → / –	hypothetical protein/–
42,464		100%		intergenic (+515/–)	*lambdap79* → / –	hypothetical protein/–
42,472		100%		intergenic (+523/–)	*lambdap79* → / –	hypothetical protein/–
42,476		100%		intergenic (+527/–)	*lambdap79* → / –	hypothetical protein/–
42,491		100%		intergenic (+542/–)	*lambdap79* → / –	hypothetical protein/–

Table 7 – Ordered features with non-zero coefficients from final model for step 1 based on P+H:MF

name	coef_val	position	mutation	annotation	gene	description	init_appear_day
bac_mut_6	-8.498125911	1882915	Δ16 bp	coding (442-457/852 nt)	ECB_RS09450 →	PTS mannose transporter subunit IID	28
phage_mut_124	-2.860562595	18868	A→C	I1122L (ATT→CTT)	J →	tail:host specificity protein	8
phage_mut_126	-2.114939761	18868	A→T	I1122F (ATT→TTT)	J →	tail:host specificity protein	8
phage_mut_36	-1.940609809	16972	A→C	S490R (AGC→CGC)	J →	tail:host specificity protein	8
phage_mut_168	-1.422548593	42300	C→A	intergenic (+351/–)	lambdap79 → / –	hypothetical protein/–	8
bac_mut_13	-1.325544733	3482706	(AGTGGGAACTGGCGGCGGAGCTGCC)1→2	coding (1022/2706 nt)	ECB_RS17295 →	transcriptional regulator MalT	8
bac_mut_1	-1.056102573	1004191	A→C	intergenic (-117/+485)	ECB_RS04930 ← / ← ECB_RS04935	phosphoporin PhoE/asparagine--tRNA ligase	28
phage_mut_10	-0.892727015	9067	T→C	R38R (CGT→CGC)	V →	tail component	8
phage_an	-0.826503428	NA	NA	NA	NA	phage ancestor indicator	NA
phage_mut_18	-0.801645409	16449	C→T	G315G (GGC→GGT)	J →	tail:host specificity protein	8
phage_mut_120	-0.78290646	18814	C→T	H1104Y (CAT→TAT)	J →	tail:host specificity protein	15
bac_mut_12	-0.776945648	3023945	Δ777 bp		[ECB_RS14915]–[ECB_RS14925]	[ECB_RS14915], ECB_RS14920, [ECB_RS14925]	15
phage_mut_114	-0.633761706	18463	A→G	T987A (ACG→GCG)	J →	tail:host specificity protein	8
bac_mut_5	-0.413595493	1881802	Δ10 bp	coding (142-151/801 nt)	ECB_RS09445 →	PTS mannose/fructose/sorbose transporter subunit IIC	8

phage_mut_40	-0.018633463	16998	G→A	V498V (GTG→GTA)	J →	tail:host specificity protein	8
phage_mut_123	-0.018002224	18825	T→G	D1107K (GAT→AAG)	J →	tail:host specificity protein	8
phage_mut_89	-0.007357429	17805	T→C	G767G (GGT→GGC)	J →	tail:host specificity protein	15
phage_mut_109	-0.005089204	18285	C→A	D927E (GAC→GAA)	J →	tail:host specificity protein	15
phage_mut_110	-0.002745251	18297	4 bp→ATAT	coding (2793-2796/3399 nt)	J →	tail:host specificity protein	15
phage_mut_116	-0.001384401	18535	A→C	S1011R (AGC→CGC)	J →	tail:host specificity protein	8
phage_mut_57	1.02E-15	17343	G→A	V613V (GTG→GTA)	J →	tail:host specificity protein	15
phage_mut_56	4.06E-15	17334	T→C	S610S (AGT→AGC)	J →	tail:host specificity protein	15
phage_mut_102	5.35E-15	17946	C→T	Y814Y (TAC→TAT)	J →	tail:host specificity protein	15
bac_mut_11	1.97E-14	24015529	A→T	I432N (ATC→AAC)	ECB_RS11915 ←	multifunctional fatty acid oxidation complex subunit alpha	8
phage_mut_42	2.00E-14	17055	T→C	G517G (GGT→GGC)	J →	tail:host specificity protein	15
phage_mut_46	3.89E-14	17085	Δ1 bp	coding (1581/3399 nt)	J →	tail:host specificity protein	15
phage_mut_101	4.58E-14	17943	T→C	G813G (GGT→GGC)	J →	tail:host specificity protein	15
phage_mut_70	4.64E-06	17502	G→A	R666R (CGG→CGA)	J →	tail:host specificity protein	15
phage_mut_54	0.001941074	17280	G→A	V592V (GTG→GTA)	J →	tail:host specificity protein	15
phage_mut_69	0.002827943	17494	A→C	S664R (AGT→CGT)	J →	tail:host specificity protein	15
phage_mut_141	0.00402645	40612	T→C	K49E (AAG→GAG)	bor ←	Bor protein precursor	8
phage_mut_113	0.00852083	18342	C→A	A946A (GCC→GCA)	J →	tail:host specificity protein	15
phage_mut_115	0.065650324	18503	C→T	A1000V (GCG→GTG)	J →	tail:host specificity protein	8

phage_mut_104	0.087045829	17964	2 bp→AG	coding (2460-2461/3399 nt)	J →	tail:host specificity protein	15
phage_mut_140	0.149463173	40601	G→A	V52V (GTC→GTT)	bor ←	Bor protein precursor	8
phage_mut_112	0.266685101	18330	C→T	G942G (GGC→GGT)	J →	tail:host specificity protein	15
bac_mut_7	0.35254696	2103918	(CCAG)7→8	coding (185/216 nt)	ECB_RS23820 →	hypothetical protein	15
phage_mut_11	0.45617881	11451	C→T	A304V (GCA→GTA)	H →	tail component	15
bac_mut_2	0.479136152	1027154	C→A	L34M (CTG→ATG)	ECB_RS05030 →	ABC transporter ATP-binding protein	15
phage_mut_99	0.510910114	17937	4 bp→ATCC	coding (2433-2436/3399 nt)	J →	tail:host specificity protein	15
phage_mut_44	0.565097986	17081	+G	coding (1577/3399 nt)	J →	tail:host specificity protein	15
phage_mut_41	0.622699572	17049	C→T	S515S (TCC→TCT)	J →	tail:host specificity protein	15
bac_mut_9	0.687768643	2247493	Δ1 bp	coding (141/624 nt)	ECB_RS11220 ←	cytochrome c biogenesis ATP-binding export protein CcmA	28
bac_mut_10	0.687774776	2401525	3 bp→AA	coding (1297-1299/2145 nt)	ECB_RS11915 ←	multifunctional fatty acid oxidation complex subunit alpha	8
bac_mut_8	1.049413308	2103918	(CCAG)7→10	coding (185/216 nt)	ECB_RS23820 →	hypothetical protein	8
phage_mut_87	1.077549856	17796	T→C	R764R (CGT→CGC)	J →	tail:host specificity protein	15
phage_mut_103	1.240548309	17950	G→A	D816N (GAT→AAT)	J →	tail:host specificity protein	15
phage_mut_117	1.252385866	18538	A→G	S1012G (AGT→GGT)	J →	tail:host specificity protein	15
phage_mut_84	1.298861158	17788	+CA	coding (2284/3399 nt)	J →	tail:host specificity protein	15
phage_mut_147	1.52352789	40683	2 bp→CC	coding (73-74/294 nt)	bor ←	Bor protein precursor	15
phage_mut_119	1.87691919	18734	T→C	V1077A (GTA→GCA)	J →	tail:host specificity protein	8
phage_mut_45	2.295777471	17082	A→C	G526G (GGA→GGC)	J →	tail:host specificity protein	15
bac_an	4.368971965	NA	NA	NA	NA	host ancestor indicator	NA

phage_mut_132	4.390950904	39198	G→A	M3I (ATG →ATA)	S →	anti-holin	22

Table 8 – Ordered features with non-zero coefficients from final model for step 2 based on P+H:MF

name	coef_val	position	mutation	annotation	gene	description	init_appear_day
phage_mut_149	-2.532013543	40898	G→C	intergenic (-142/+149)	bor ← / ← lambdap78	Bor protein precursor/putative envelope protein	15
phage_mut_89	-2.161152065	17805	T→C	G767G (GGT→GGC)	J →	tail:host specificity protein	15
phage_mut_118	-2.130066439	18731	C→T	A1076V (GCG→GTG)	J →	tail:host specificity protein	15
phage_mut_45	-2.051360951	17082	A→C	G526G (GGA→GGC)	J →	tail:host specificity protein	15
phage_mut_88	-1.747164292	17797	Δ1 bp	coding (2293/3399 nt)	J →	tail:host specificity protein	15
phage_mut_169	-1.663968518	42432	C→G	intergenic (+483/–)	lambdap79 → / –	hypothetical protein/–	15
phage_mut_105	-1.37255746	18255	G→T	V917V (GTG→GTT)	J →	tail:host specificity protein	15
phage_mut_139	-1.293045344	40434	T→C	intergenic (+3/+29)	Rz → / ← bor	cell lysis protein/Bor protein precursor	8
phage_mut_125	-1.147054441	18868	A→G	I1122V (ATT→GTT)	J →	tail:host specificity protein	15
phage_mut_126	-1.070655153	18868	A→T	I1122F (ATT→TTT)	J →	tail:host specificity protein	8
phage_mut_131	-0.820498193	39183	(G)5→6	intergenic (+364/-7)	orf-64 → / → S	hypothetical protein/anti-holin	15
phage_mut_86	-0.774992148	17795	Δ2 bp	coding (2291-2292/3399 nt)	J →	tail:host specificity protein	15
phage_mut_43	-0.739131464	17059	G→A	A519T (GCG→ACG)	J →	tail:host specificity protein	15
phage_mut_12	-0.720051132	15890	A→G	D129G (GAC→GGC)	J →	tail:host specificity protein	8

bac_mut_9	-0.462576568	2247493	Δ1 bp	coding (141/624 nt)	ECB_RS11220 ←	cytochrome c biogenesis ATP-binding export protein CcmA	28
bac_mut_4	-0.400226247	1368326	C→A	N90K (AAC→AAA)	ECB_RS06835 →	thiosulfate sulfurtransferase PspE	28
bac_mut_12	-0.397827389	3023945	Δ777 bp		[ECB_RS14915]–[ECB_RS14925]	[ECB_RS14915], ECB_RS14920, [ECB_RS14925]	15
phage_mut_47	-0.383998668	17088	C→G	G528G (GGC→GGG)	J →	tail:host specificity protein	15
phage_mut_18	-0.288244273	16449	C→T	G315G (GGC→GGT)	J →	tail:host specificity protein	8
bac_mut_13	-0.252538224	3482706	(AGTGGGAACTGGCGGCGGAGCTGCC)1→2	coding (1022/2706 nt)	ECB_RS17295 →	transcriptional regulator MalT	8
bac_mut_15	-0.226213255	3482943	A→C	Q420P (CAA→CCA)	ECB_RS17295 →	transcriptional regulator MalT	8
phage_mut_55	-0.185559827	17328	A→C	E608D (GAA→GAC)	J →	tail:host specificity protein	15
bac_mut_5	-0.161507783	1881802	Δ10 bp	coding (142-151/801 nt)	ECB_RS09445 →	PTS mannose/fructose/sorbose transporter subunit IIC	8
phage_mut_75	-0.151752332	17613	T→G	D703E (GAT→GAG)	J →	tail:host specificity protein	15
phage_mut_168	-0.130027394	42300	C→A	intergenic (+351/–)	lambdap79 → / –	hypothetical protein/–	8
bac_mut_7	-0.12545278	2103918	(CCAG)7→8	coding (185/216 nt)	ECB_RS23820 →	hypothetical protein	15
phage_mut_83	-0.124104287	17775	A→G	R757R (AGA→AGG)	J →	tail:host specificity protein	15
phage_mut_154	-0.079258297	40929	C→T	intergenic (-173/+118)	bor ← / ← lambdap78	Bor protein precursor/putative envelope protein	15
phage_mut_85	-0.038923329	17793	G→A	T763T (ACG→ACA)	J →	tail:host specificity protein	22
phage_mut_160	-0.031345329	40973	A→C	intergenic (-217/+74)	bor ← / ← lambdap78	Bor protein precursor/putative envelope protein	15

phage_mut_13	-0.020517309	16218	G→T	L238L (CTG→CTT)	J →	tail:host specificity protein	15
phage_mut_129	-0.00154779	19791	C→G	R48G (CGT→GGT)	orf-401 →	Tail fiber protein	15
phage_mut_170	-4.19E-14	42434	2 bp→AG	intergenic (+485/–)	lambdap79 → / –	hypothetical protein/–	15
phage_mut_40	-3.11E-15	16998	G→A	V498V (GTG→GTA)	J →	tail:host specificity protein	8
phage_mut_150	-1.34E-15	40905	T→C	intergenic (-149/+142)	bor ← / ← lambdap78	Bor protein precursor/putative envelope protein	15
phage_mut_161	-1.05E-15	42104	2 bp→AC	intergenic (+155/–)	lambdap79 → / –	hypothetical protein/–	15
phage_mut_76	-9.91E-16	17652	A→G	A716A (GCA→GCG)	J →	tail:host specificity protein	15
phage_mut_155	-5.88E-16	40931	T→C	intergenic (-175/+116)	bor ← / ← lambdap78	Bor protein precursor/putative envelope protein	15
phage_mut_15	-5.15E-16	16299	A→G	K265K (AAA→AAG)	J →	tail:host specificity protein	15
phage_mut_14	-4.47E-17	16227	T→C	R241R (CGT→CGC)	J →	tail:host specificity protein	15
phage_mut_101	2.66E-17	17943	T→C	G813G (GGT→GGC)	J →	tail:host specificity protein	15
phage_mut_42	4.48E-17	17055	T→C	G517G (GGT→GGC)	J →	tail:host specificity protein	15
phage_mut_102	5.51E-17	17946	C→T	Y814Y (TAC→TAT)	J →	tail:host specificity protein	15
phage_mut_157	1.84E-16	40939	G→T	intergenic (-183/+108)	bor ← / ← lambdap78	Bor protein precursor/putative envelope protein	15
phage_mut_82	2.16E-16	17759	A→G	Q752R (CAG→CGG)	J →	tail:host specificity protein	15
phage_mut_91	2.19E-16	17868	T→C	Y788Y (TAT→TAC)	J →	tail:host specificity protein	15
phage_mut_58	2.24E-16	17391	T→C	T629T (ACT→ACC)	J →	tail:host specificity protein	15
phage_mut_116	2.97E-16	18535	A→C	S1011R (AGC→CGC)	J →	tail:host specificity protein	8
phage_mut_134	3.90E-16	40158	G→A	A63A (GCG→GCA)	Rz →	cell lysis protein	8
phage_mut_2	5.01E-16	327	C→T	V46V (GTC→GTT)	nu1 →	DNA packaging protein	28
phage_mut_141	4.02E-15	40612	T→C	K49E (AAG→GAG)	bor ←	Bor protein precursor	8

bac_mut_11	2.05E-14	2401529	A→T	I432N (ATC→AAC)	ECB_RS11915 ←	multifunctional fatty acid oxidation complex subunit alpha	8
phage_mut_123	1.37E-07	18825	T→G	D1107K (GAT→AAG)	J →	tail:host specificity protein	8
phage_mut_167	0.016733222	42207	G→A	intergenic (+258/–)	lambdap79 → / –	hypothetical protein/–	8
phage_mut_1	0.020234009	175	T→G	intergenic (–/-15)	– / → nu1	–/DNA packaging protein	28
phage_mut_156	0.032709773	40933	+T	intergenic (-177/+114)	bor ← / ← lambdap78	Bor protein precursor/putative envelope protein	15
phage_mut_153	0.082708364	40919	Δ1 bp	intergenic (-163/+128)	bor ← / ← lambdap78	Bor protein precursor/putative envelope protein	15
phage_mut_90	0.117863411	17862	C→T	Y786Y (TAC→TAT)	J →	tail:host specificity protein	15
phage_mut_81	0.144121108	17721	C→T	D739D (GAC→GAT)	J →	tail:host specificity protein	15
phage_mut_109	0.175913807	18285	C→A	D927E (GAC→GAA)	J →	tail:host specificity protein	15
phage_mut_56	0.176228514	17334	T→C	S610S (AGT→AGC)	J →	tail:host specificity protein	15
phage_mut_140	0.186721057	40601	G→A	V52V (GTC→GTT)	bor ←	Bor protein precursor	8
bac_mut_1	0.203591311	1004191	A→C	intergenic (-117/+485)	ECB_RS04930 ← / ← ECB_RS04935	phosphoporin PhoE/asparagine--tRNA ligase	28
phage_mut_54	0.215488832	17280	G→A	V592V (GTG→GTA)	J →	tail:host specificity protein	15
phage_mut_41	0.3167206 3	17049	C→T	S515S (TCC→TCT)	J →	tail:host specificity protein	15
phage_mut_114	0.345624215	18463	A→G	T987A (ACG→GCG)	J →	tail:host specificity protein	8
phage_mut_124	0.404002965	18868	A→C	I1122L (ATT→CTT)	J →	tail:host specificity protein	8
phage_mut_133	0.453165474	40140	T→C	R57R (CGT→CGC)	Rz →	cell lysis protein	8
phage_mut_148	0.559233722	40723	2 bp→TT	coding (33-34/294 nt)	bor ←	Bor protein precursor	15
bac_mut_10	0.608510822	2401525	3 bp→AA	coding (1297-1299/2145 nt)	ECB_RS11915 ←	multifunctional fatty acid oxidation complex subunit alpha	8
bac_mut_17	0.639539386	4228027	Δ1 bp	coding (1125/1341 nt)	lamB →	maltoporin	8

bac_mut_8	0.738693216	2103918	(CCAG)7→10	coding (185/216 nt)	ECB_RS23820 →	hypothetical protein	8
bac_mut_2	0.900183276	1027154	C→A	L34M (CTG→ATG)	ECB_RS05030 →	ABC transporter ATP-binding protein	15
phage_mut_132	0.958400064	39198	G→A	M3I (ATG→ATA)	S →	anti-holin	22
phage_mut_103	1.379223634	17950	G→A	D816N (GAT→AAT)	J →	tail:host specificity protein	15
phage_mut_128	1.808889793	19260	T→C	L99P (CTG→CCG)	lom →	outer host membrane	22
bac_an	2.212627851	NA	NA	NA	NA	host ancestor indicator	NA
phage_mut_84	2.38285428	17788	+CA	coding (2284/3399 nt)	J →	tail:host specificity protein	15
phage_mut_87	2.559949754	17796	T→C	R764R (CGT→CGC)	J →	tail:host specificity protein	15
phage_mut_44	2.67311362	17081	+G	coding (1577/3399 nt)	J →	tail:host specificity protein	15
phage_mut_99	2.824210284	17937	4 bp→ATCC	coding (2433-2436/3399 nt)	J →	tail:host specificity protein	15

APPENDIX C.
SUPPLEMENTARY INFORMATION FOR CHAPTER 4

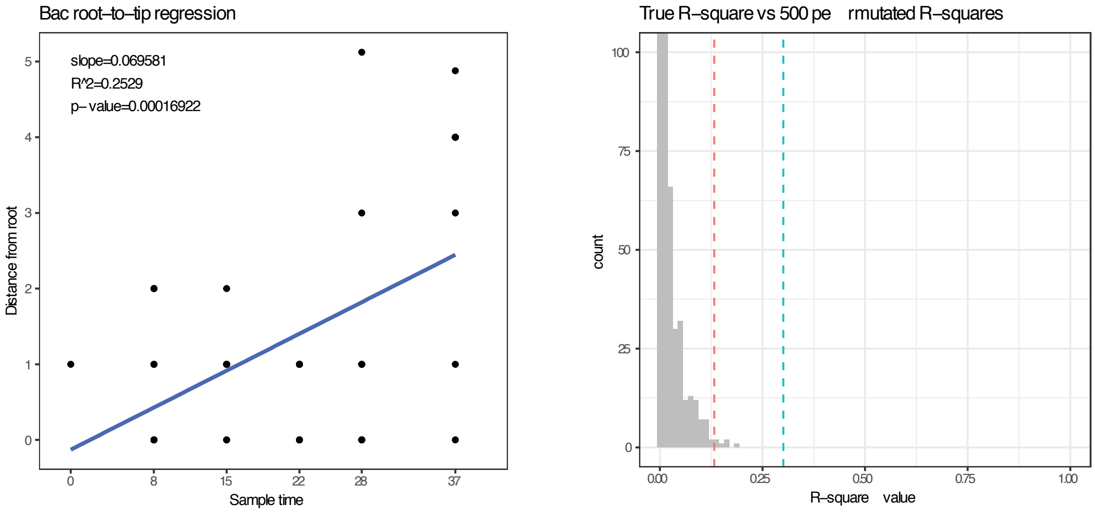

Figure 29 – Temporal signal analysis for the host phylodynamic tree

(A) Root-to-tip regression analysis results from the neighbor-joining tree based on hamming distance matrix for E. coli. (B) Significance level assessed by comparing the fitted R squared value versus 500 random permuted ones.

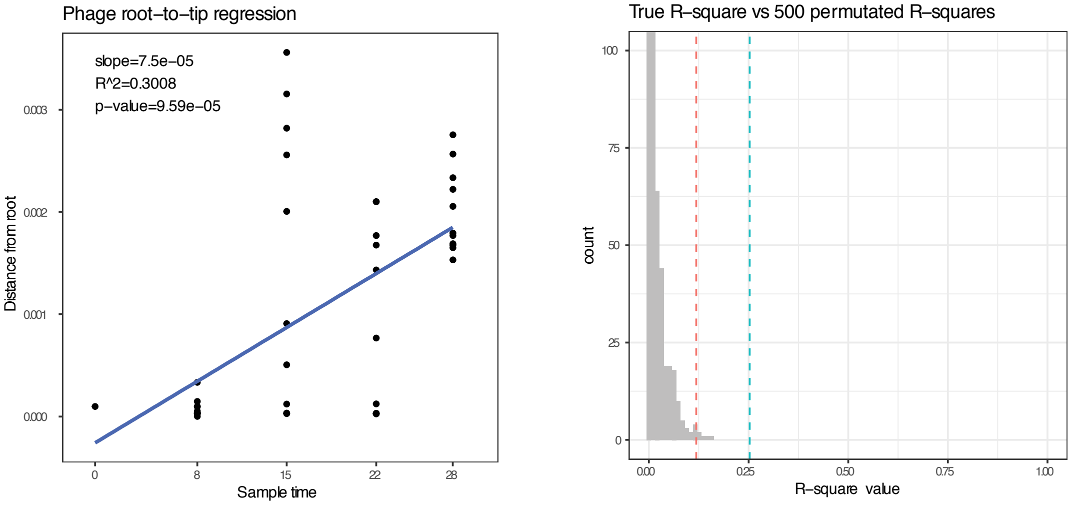

Figure 30 – Temporal signal analysis on the phage phylodynamic tree

(A) Root-to-tip regression analysis results from the maximum likelihood tree built for phage. (B) Significance level assessed by comparing the fitted R squared value versus 500 random permuted ones.

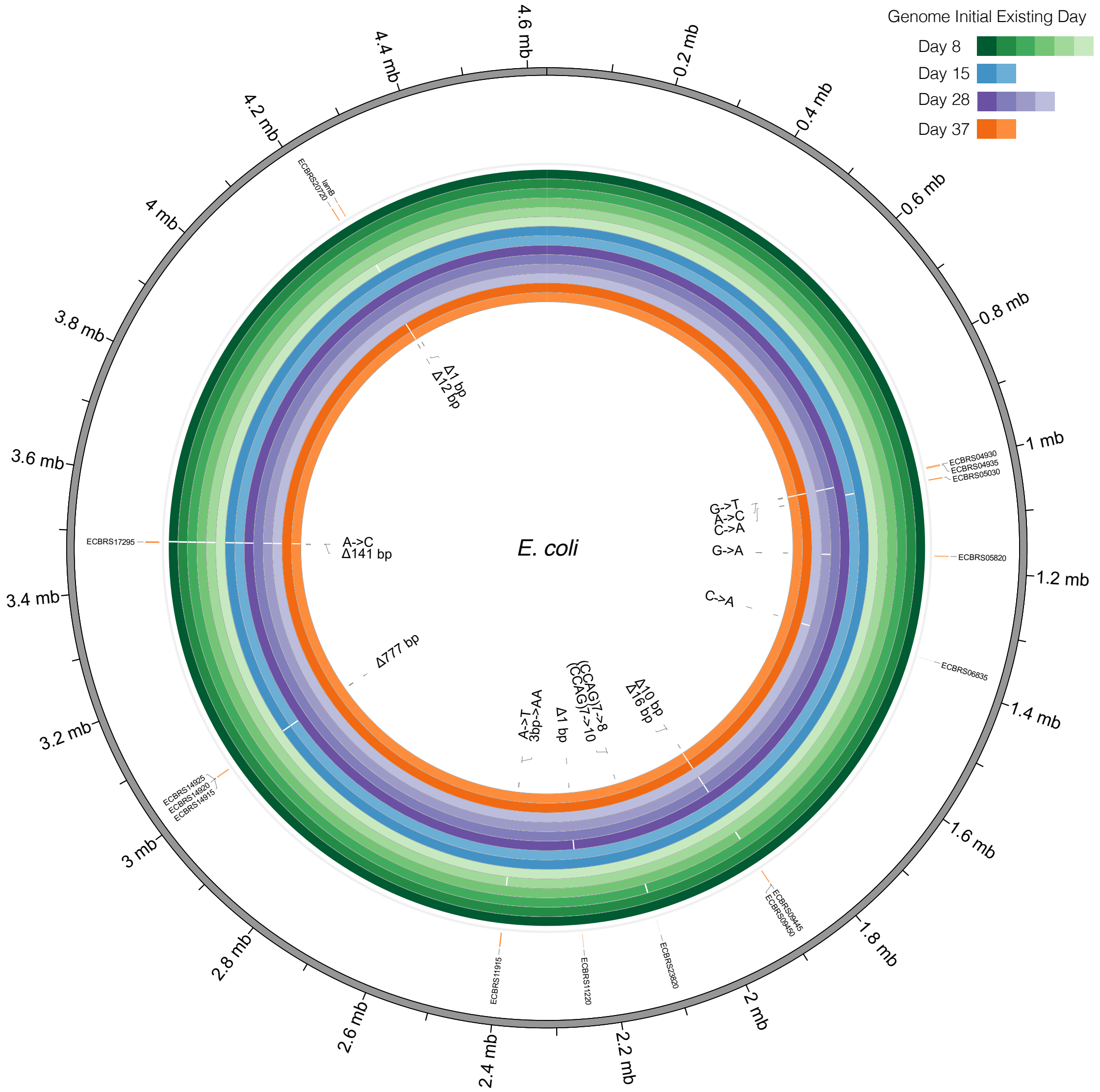

Figure 31 – Recovered unique genomes for *E. coli*

The outer gray ring represents the reference host genome. The orange bars represent the genes that harbors the observed mutation. The colored rings represent samples taken during the experiment. The color groups represent the sampling days. Inner grey bars represent the unique mutations observed from all samples. Different shades of the same color represent different unique genotypes from the same sampling day. White gaps in the genome rings indicate the location of observed mutations.

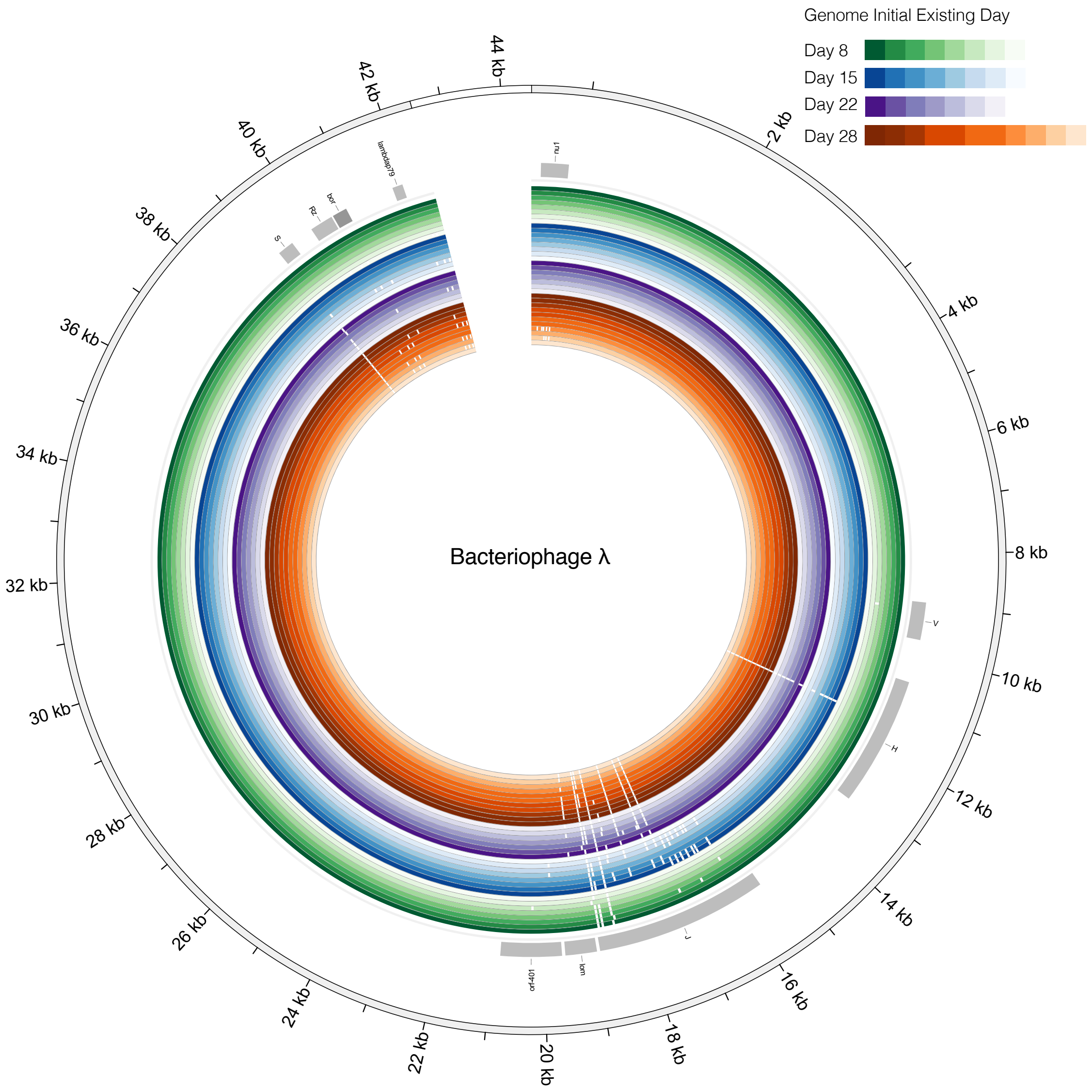

Figure 32 – Recovered unique genomes for the bacteriophage λ

The outer gray ring represents the reference phage genome. The inner grey bars represent the genes that harbors the observed mutations. The colored rings represent samples taken during the experiment. The color groups represent the sampling days. Different shades of the same color represent different unique genotypes from the same sampling day. White gaps in the genome rings indicate the location of observed mutations.

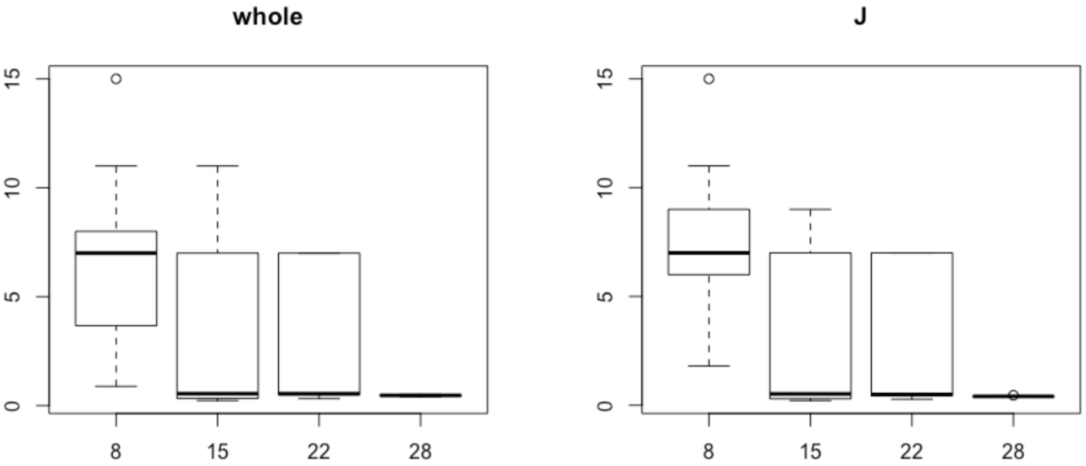

Figure 33 – D_N/D_S ratios for phage whole genome (A) and J protein region (B) across sampling days

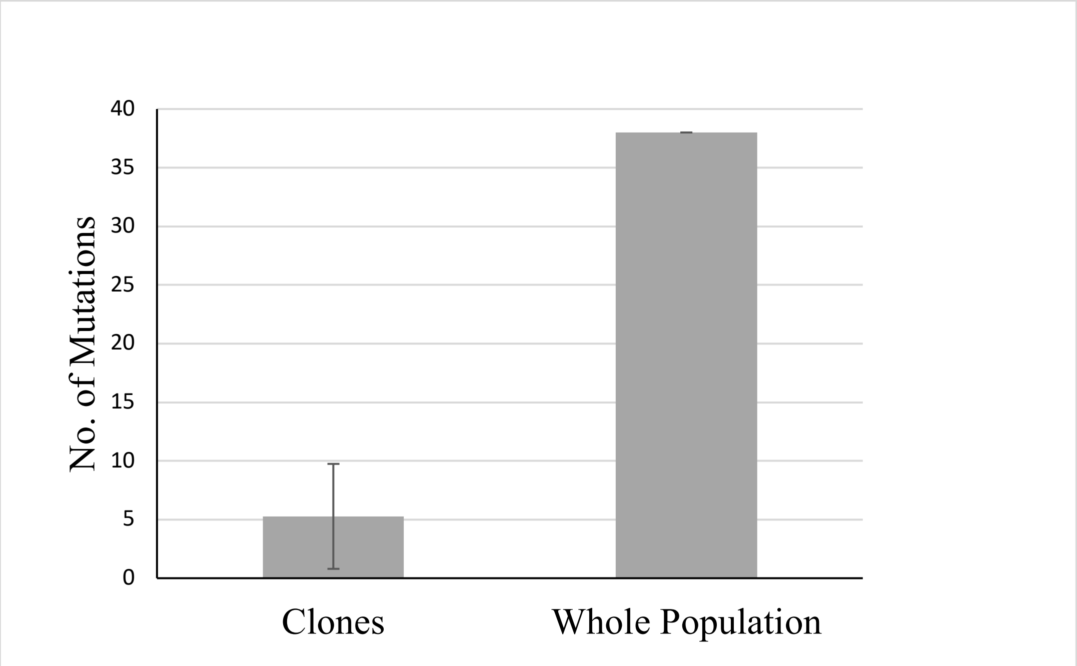

Figure 34 – Difference in genomic variation observed between whole population sequencing and 11 isolated clones of λ on Day 8

The large error bar for clones is because of a recombination event between prophage and a single clone isolated on Day 8.

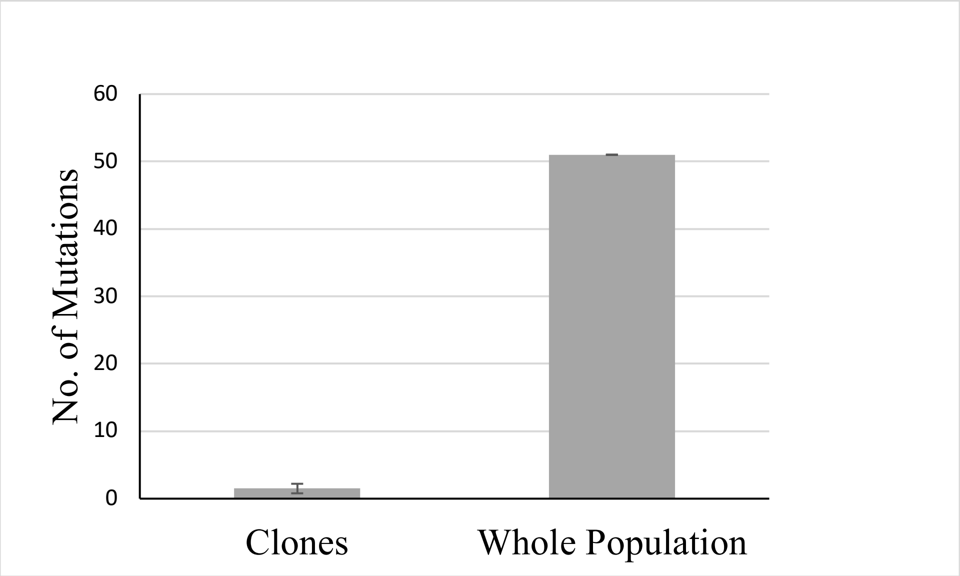

Figure 35 – Difference in genomic variation observed between whole population sequencing and 10 isolated clones of *E. coli* on Day 8

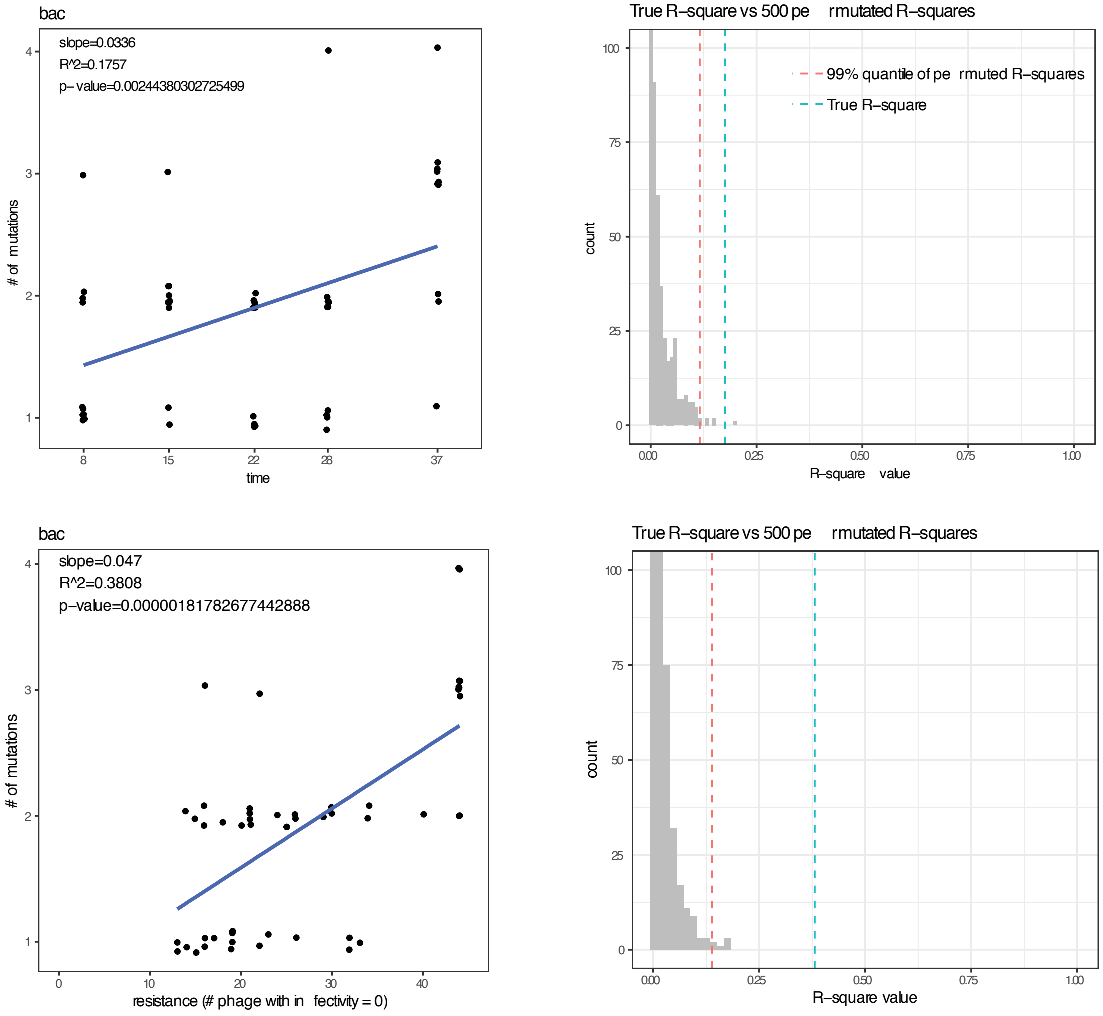

Figure 36 – Regression analysis of host genotype against coevolution time and phenotype

(A) Regression of the number of mutations in E. coli samples against sampling time (B) Regression of the number of mutations against host resistance. Jittering is applied for better visualization. Significance level assessed by comparing the fitted R squared value vs 500 random permuted ones for the regression against time (C) and regression against phenotype (D).

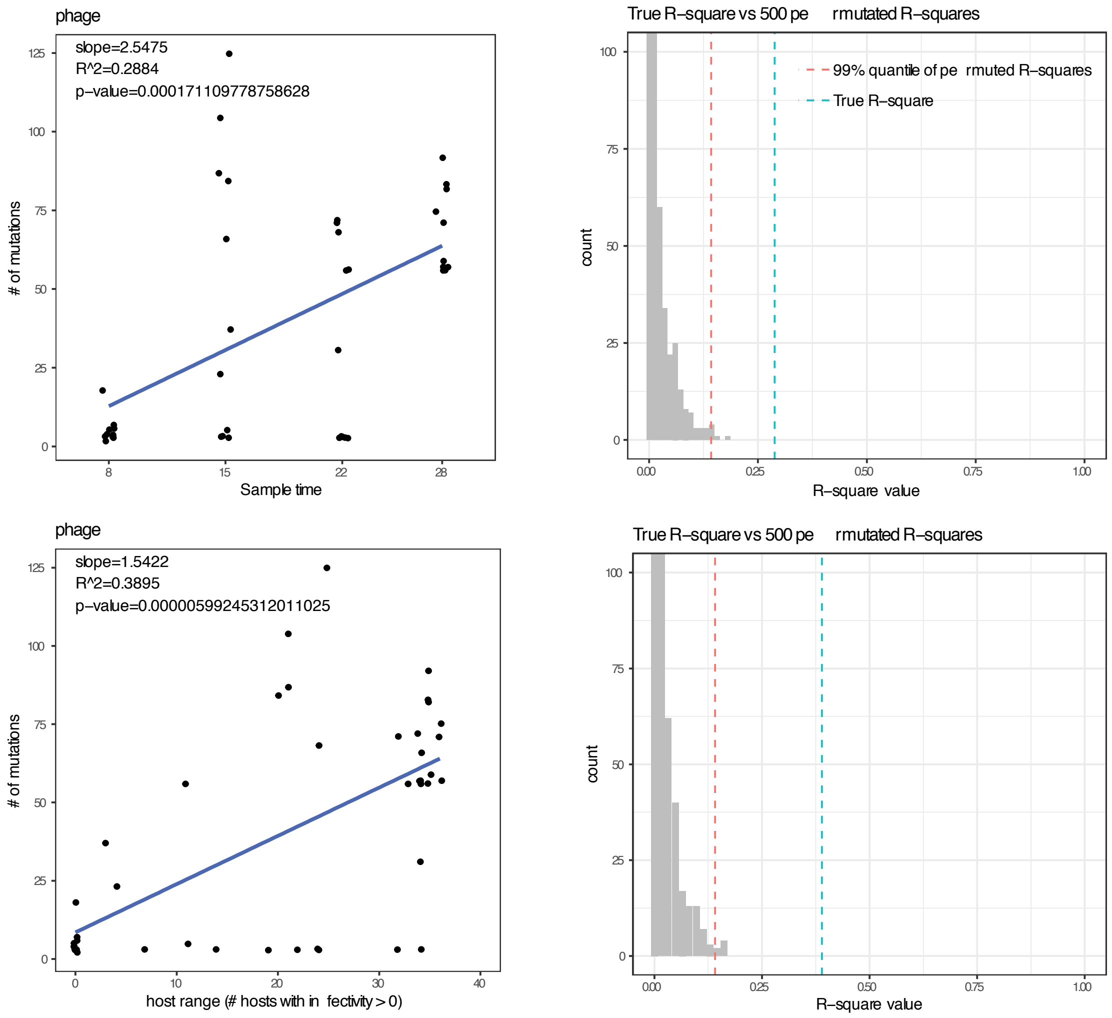

Figure 37 – Regression analysis of phage genotype against coevolution time and phenotype

(A) Regression of the number of mutations in bacteriophage λ samples against sampling time. (B) Regression of the number of mutations against host resistance. Jittering is applied for better visualization. Significance level assessed by comparing the fitted R squared value vs 500 random permuted ones for the regression against time (C) and regression against phenotype (D).

Table 9 – Genomic variation present in the phage population on Day 8 of the coevolution experiment as compared to the ancestral λ strain cI26 used in the study

Genome Location	Mutatio n	Amino acid change	Gen e	Product
11,445	C→T	A->V	*H* →	Tail component
11,451	**C→T**	**A->V**	***H*** **→**	**Tail component**
15,890	A→G	D->G	*J* →	Tail- host specificity protein
16,218	G→T		*J* →	Tail- host specificity protein
16,227	T→C		*J* →	Tail- host specificity protein
16,299	A→G		*J* →	Tail- host specificity protein
16,318	A→C	M->L	*J* →	Tail- host specificity protein
16,319	T→C	M->T	*J* →	Tail- host specificity protein
16,350	T→C		*J* →	Tail- host specificity protein
16,449	C→T		*J* →	Tail- host specificity protein
16,485	G→C		*J* →	Tail- host specificity protein
16,497	A→G		*J* →	Tail- host specificity protein
16,524	C→T		*J* →	Tail- host specificity protein
16,596	G→A		*J* →	Tail- host specificity protein
16,599	G→A		*J* →	Tail- host specificity protein
16,606	A→G	T->A	*J* →	Tail- host specificity protein
16,607	C→T	T->M	*J* →	Tail- host specificity protein
16,725	C→T		*J* →	Tail- host specificity protein
16,774	G→C	A->P	*J* →	Tail- host specificity protein
16,775	C→T	A->V	*J* →	Tail- host specificity protein
16,791	T→C		*J* →	Tail- host specificity protein
16,794	T→C		*J* →	Tail- host specificity protein
16,866	A→G		*J* →	Tail- host specificity protein

16,869	A→G		*J* →	Tail- host specificity protein
16,893	T→C		*J* →	Tail- host specificity protein
16,902	C→G		*J* →	Tail- host specificity protein
16,905	C→T		*J* →	Tail- host specificity protein
16,908	A→C		*J* →	Tail- host specificity protein
16,938	T→C		*J* →	Tail- host specificity protein
16,972	A→C	S->R	*J* →	Tail- host specificity protein
16,980	T→C		*J* →	Tail- host specificity protein
16,983	T→G		*J* →	Tail- host specificity protein
16,986	T→C		*J* →	Tail- host specificity protein
16,998	G→A		*J* →	Tail- host specificity protein
18,503	*C→T*	*A->V*	*J* →	*Tail- host specificity protein*
18,734	*T→C*	*V->A*	*J* →	*Tail- host specificity protein*
18,823	*G→A*	*D->N*	*J* →	*Tail- host specificity protein*
18,868	*A→T*	*I->F*	*J* →	*Tail- host specificity protein*

Table 10 – Genomic variation present in *E. coli* population on Day 8 compared to ancestral genome (GenBank: CP000819.1)

Genome Location	Mutation	Gene	Product
38,192	G→T	*carB* → / → *caiF*	carbamoyl-phosphate synthase large subunit/DNA-binding transcriptional activator
38,193	C→T	*carB* → / → *caiF*	carbamoyl-phosphate synthase large subunit/DNA-binding transcriptional activator
38,194	C→T	*carB* → / → *caiF*	carbamoyl-phosphate synthase large subunit/DNA-binding transcriptional activator
38,195	C→T	*carB* → / → *caiF*	carbamoyl-phosphate synthase large subunit/DNA-binding transcriptional activator
38,196	A→T	*carB* → / → *caiF*	carbamoyl-phosphate synthase large subunit/DNA-binding transcriptional activator
38,199	A→T	*carB* → / → *caiF*	carbamoyl-phosphate synthase large subunit/DNA-binding transcriptional activator
38,200	A→T	*carB* → / → *caiF*	carbamoyl-phosphate synthase large subunit/DNA-binding transcriptional activator
386,921	C→T	*phoR* →	sensory histidine kinase in two-component regulatory system with PhoB
519,803	G→T	*fdrA* →	membrane protein FdrA
519,808	A→C	*fdrA* →	membrane protein FdrA
560,154	C→T	*ompT* ←	outer membrane protease VII (outer membrane protein 3b)
863,867	G→A	*yliC* →	predicted peptide transporter subunit: membrane component of ABC superfamily
863,868	G→T	*yliC* →	predicted peptide transporter subunit: membrane component of ABC superfamily
863,873	A→C	*yliC* →	predicted peptide transporter subunit: membrane component of ABC superfamily
863,874	T→C	*yliC* →	predicted peptide transporter subunit: membrane component of ABC superfamily
949,387	G→A	*trxB* ← / → *lrp*	thioredoxin reductase, FAD/NAD(P)-binding/DNA-binding transcriptional dual regulator, leucine-binding
1,368,412:1	$(T)_{9\rightarrow10}$	*pspE* → / → *ycjM*	thiosulfate:cyanide sulfurtransferase (rhodanese)/predicted glucosyltransferase
1,418,284	G→T	*rzpR* →	predicted defective peptidase
1,605,635	Δ1 bp	*stfR* ←	predicted tail fiber protein
1,605,636	T→G	*stfR* ←	predicted tail fiber protein
1,605,637	G→T	*stfR* ←	predicted tail fiber protein
1,605,637:1	+T	*stfR* ←	predicted tail fiber protein
1,881,837	Δ1 bp	*manY* →	mannose-specific enzyme IIC component of PTS
1,881,838	Δ1 bp	*manY* →	mannose-specific enzyme IIC component of PTS
1,882,021	C→T	*manY* →	mannose-specific enzyme IIC component of PTS

1,882,908:1	+A	*manZ* →	mannose-specific enzyme IID component of PTS
1,882,915	**Δ16 bp**	***manZ* →**	**mannose-specific enzyme IID component of PTS**
2,111,270	C→A	*ECB_01999* →	putative phage protein
2,250,122:1	+G	*napG* ←	quinol dehydrogenase periplasmic component
2,250,126	A→C	*napG* ←	quinol dehydrogenase periplasmic component
2,250,129	Δ1 bp	*napG* ←	quinol dehydrogenase periplasmic component
2,310,865	G→A	*yfaZ* ← / → *yfaO*	predicted outer membrane porin protein/predicted NUDIX hydrolase
2,310,868	C→A	*yfaZ* ← / → *yfaO*	predicted outer membrane porin protein/predicted NUDIX hydrolase
2,401,525	Δ1 bp	*yfcX* ←	fused enoyl-CoA hydratase and epimerase and isomerase/3-hydroxyacyl-CoA dehydrogenase
2,401,526	G→A	*yfcX* ←	fused enoyl-CoA hydratase and epimerase and isomerase/3-hydroxyacyl-CoA dehydrogenase
2,401,527	C→A	*yfcX* ←	fused enoyl-CoA hydratase and epimerase and isomerase/3-hydroxyacyl-CoA dehydrogenase
2,401,529	A→T	*yfcX* ←	fused enoyl-CoA hydratase and epimerase and isomerase/3-hydroxyacyl-CoA dehydrogenase
2,940,619	A→T	*ygfB* ←	hypothetical protein
3,000,508	C→G	*flu* →	antigen 43 (Ag43) phase-variable biofilm formation autotransporter
3,482,706:1	25-bp duplication	*malT* →	transcriptional regulator MalT
3,483,094:1	+C	*malT* →	transcriptional regulator MalT
3,483,094:2	+T	*malT* →	transcriptional regulator MalT
3,942,902	T→A	*yifK* →	predicted transporter
4,236,155	T→A	*lexA* →	LexA repressor
4,236,156	T→A	*lexA* →	LexA repressor
4,236,158	C→G	*lexA* →	LexA repressor
4,236,160	T→A	*lexA* →	LexA repressor
4,236,161	T→A	*lexA* →	LexA repressor
4,300,483	C→T	*phnG* ←	carbon-phosphorus lyase complex subunit
4,504,878	T→A	*insA-25* → / → *ECB_04162*	IS1 protein InsA/hypothetical protein
4,537,685	A→T	*yjiC* ← / → *yjiD*	hypothetical protein/DNA replication/recombination/repair protein

REFRENCES

1. Koonin EV, Senkevich TG, Dolja VV. The ancient Virus World and evolution of cells. Biol Direct. 2006;1:29. Epub 2006/09/21. doi: 10.1186/1745-6150-1-29. PubMed PMID: 16984643; PubMed Central PMCID: PMCPMC1594570.

2. Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 2005;13(6):278-84. Epub 2005/06/07. doi: 10.1016/j.tim.2005.04.003. PubMed PMID: 15936660.

3. Lecoq H. [Discovery of the first virus, the tobacco mosaic virus: 1892 or 1898?]. C R Acad Sci III. 2001;324(10):929-33. Epub 2001/09/26. PubMed PMID: 11570281.

4. Khelifa M, Masse D, Blanc S, Drucker M. Evaluation of the minimal replication time of Cauliflower mosaic virus in different hosts. Virology. 2010;396(2):238-45. Epub 2009/11/17. doi: 10.1016/j.virol.2009.09.032. PubMed PMID: 19913268.

5. Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Ruger W. Bacteriophage T4 genome. Microbiol Mol Biol Rev. 2003;67(1):86-156, table of contents. Epub 2003/03/11. PubMed PMID: 12626685; PubMed Central PMCID: PMCPMC150520.

6. Prangishvili D, Forterre P, Garrett RA. Viruses of the Archaea: a unifying view. Nat Rev Microbiol. 2006;4(11):837-48. Epub 2006/10/17. doi: 10.1038/nrmicro1527. PubMed PMID: 17041631.

7. Suttle CA. Viruses in the sea. Nature. 2005;437(7057):356-61. Epub 2005/09/16. doi: 10.1038/nature04160. PubMed PMID: 16163346.

8. Paul JH, Sullivan MB, Segall AM, Rohwer F. Marine phage genomics. Comp Biochem Physiol B Biochem Mol Biol. 2002;133(4):463-76. Epub 2002/12/10. PubMed PMID: 12470812.

9. Paul JH, Sullivan MB. Marine phage genomics: what have we learned? Curr Opin Biotechnol. 2005;16(3):299-307. Epub 2005/06/18. doi: 10.1016/j.copbio.2005.03.007. PubMed PMID: 15961031.

10. Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell. 2012;148(6):1258-70. Epub 2012/03/20. doi: 10.1016/j.cell.2012.01.035. PubMed PMID: 22424233; PubMed Central PMCID: PMCPMC5050011.

11. Kinross JM, Darzi AW, Nicholson JK. Gut microbiome-host interactions in health and disease. Genome Med. 2011;3(3):14. Epub 2011/03/12. doi: 10.1186/gm228. PubMed PMID: 21392406; PubMed Central PMCID: PMCPMC3092099.

12. Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D. Assembly of viral metagenomes from yellowstone hot springs. Appl Environ Microbiol.

2008;74(13):4164-74. Epub 2008/04/29. doi: 10.1128/AEM.02598-07. PubMed PMID: 18441115; PubMed Central PMCID: PMCPMC2446518.

13. Bolduc B, Wirth JF, Mazurie A, Young MJ. Viral assemblage composition in Yellowstone acidic hot springs assessed by network analysis. ISME J. 2015;9(10):2162-77. Epub 2015/07/01. doi: 10.1038/ismej.2015.28. PubMed PMID: 26125684; PubMed Central PMCID: PMCPMC4579470.

14. Peterson JF. Electron microscopy of soil-borne wheat mosaic virus in host cells. Virology. 1970;42(2):304-10. Epub 1970/10/01. PubMed PMID: 4099066.

15. DeLeon-Rodriguez N, Lathem TL, Rodriguez RL, Barazesh JM, Anderson BE, Beyersdorf AJ, et al. Microbiome of the upper troposphere: species composition and prevalence, effects of tropical storms, and atmospheric implications. Proc Natl Acad Sci U S A. 2013;110(7):2575-80. Epub 2013/01/30. doi: 10.1073/pnas.1212089110. PubMed PMID: 23359712; PubMed Central PMCID: PMCPMC3574924.

16. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A. 1998;95(12):6578-83. Epub 1998/06/17. PubMed PMID: 9618454; PubMed Central PMCID: PMCPMC33863.

17. Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S. Global distribution of microbial abundance and biomass in subseafloor sediment. Proc Natl Acad Sci U S A. 2012;109(40):16213-6. Epub 2012/08/29. doi: 10.1073/pnas.1203849109. PubMed PMID: 22927371; PubMed Central PMCID: PMCPMC3479597.

18. Weinbauer MG. Ecology of prokaryotic viruses. FEMS Microbiol Rev. 2004;28(2):127-81. Epub 2004/04/28. doi: 10.1016/j.femsre.2003.08.001. PubMed PMID: 15109783.

19. Milo R, Jorgensen P, Moran U, Weber G, Springer M. BioNumbers--the database of key numbers in molecular and cell biology. Nucleic Acids Res. 2010;38(Database issue):D750-3. Epub 2009/10/27. doi: 10.1093/nar/gkp889. PubMed PMID: 19854939; PubMed Central PMCID: PMCPMC2808940.

20. Thingstad TF. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnology and Oceanography. 2000;45(6):1320-8.

21. Wigington CH, Sonderegger D, Brussaard CP, Buchan A, Finke JF, Fuhrman JA, et al. Re-examination of the relationship between marine virus and microbial cell abundances. Nat Microbiol. 2016;1:15024. Epub 2016/08/31. doi: 10.1038/nmicrobiol.2015.24. PubMed PMID: 27572161.

22. Bohannan BJM, Lenski RE. The Relative Importance of Competition and Predation Varies with Productivity in a Model Community. Am Nat. 2000;156(4):329-40. Epub 2000/10/01. doi: 10.1086/303393. PubMed PMID: 29592139.

23. Koskella B, Brockhurst MA. Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. FEMS Microbiol Rev. 2014;38(5):916-31. Epub 2014/03/13. doi: 10.1111/1574-6976.12072. PubMed PMID: 24617569; PubMed Central PMCID: PMCPMC4257071.

24. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7(11):828-36. Epub 2009/10/17. doi: 10.1038/nrmicro2235. PubMed PMID: 19834481.

25. Twort FW. Further Investigations on the Nature of Ultra-Microscopic Viruses and their Cultivation. J Hyg (Lond). 1936;36(2):204-35. Epub 1936/06/01. PubMed PMID: 20475326; PubMed Central PMCID: PMCPMC2170983.

26. Keen EC. A century of phage research: bacteriophages and the shaping of modern biology. Bioessays. 2015;37(1):6-9. Epub 2014/12/19. doi: 10.1002/bies.201400152. PubMed PMID: 25521633; PubMed Central PMCID: PMCPMC4418462.

27. Taylor MW. The Discovery of Bacteriophage and the d'Herelle Controversy. Viruses and Man: A History of Interactions: Springer; 2014. p. 53-61.

28. Duckworth DH. " Who discovered bacteriophage?". Bacteriological reviews. 1976;40(4):793.

29. Chibani-Chennoufi S, Bruttin A, Dillmann ML, Brussow H. Phage-host interaction: an ecological perspective. J Bacteriol. 2004;186(12):3677-86. Epub 2004/06/04. doi: 10.1128/JB.186.12.3677-3686.2004. PubMed PMID: 15175280; PubMed Central PMCID: PMCPMC419959.

30. Danovaro R, Corinaldesi C, Dell'anno A, Fuhrman JA, Middelburg JJ, Noble RT, et al. Marine viruses and global climate change. FEMS Microbiol Rev. 2011;35(6):993-1034. Epub 2011/01/06. doi: 10.1111/j.1574-6976.2010.00258.x. PubMed PMID: 21204862.

31. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, et al. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J Bacteriol. 2008;190(4):1390-400. Epub 2007/12/11. doi: 10.1128/JB.01412-07. PubMed PMID: 18065545; PubMed Central PMCID: PMCPMC2238228.

32. Abedon ST, Herschler TD, Stopar D. Bacteriophage latent-period evolution as a response to resource availability. Appl Environ Microbiol. 2001;67(9):4233-41. Epub 2001/08/30. PubMed PMID: 11526028; PubMed Central PMCID: PMCPMC93152.

33. Bohannan BJ, Lenski RE. Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. Ecology letters. 2000;3(4):362-77.

34. Hadas H, Einav M, Fishov I, Zaritsky A. Bacteriophage T4 development depends on the physiology of its host Escherichia coli. Microbiology. 1997;143 (Pt 1):179-85. Epub 1997/01/01. doi: 10.1099/00221287-143-1-179. PubMed PMID: 9025292.

35. Lederberg EM, Lederberg J. Genetic Studies of Lysogenicity in Escherichia Coli. Genetics. 1953;38(1):51-64. Epub 1953/01/01. PubMed PMID: 17247421; PubMed Central PMCID: PMCPMC1209586.

36. Dekel-Bird NP, Sabehi G, Mosevitzky B, Lindell D. Host-dependent differences in abundance, composition and host range of cyanophages from the Red Sea. Environ Microbiol. 2015;17(4):1286-99. Epub 2014/07/22. doi: 10.1111/1462-2920.12569. PubMed PMID: 25041521.

37. Weitz JS, Poisot T, Meyer JR, Flores CO, Valverde S, Sullivan MB, et al. Phage-bacteria infection networks. Trends Microbiol. 2013;21(2):82-91. Epub 2012/12/19. doi: 10.1016/j.tim.2012.11.003. PubMed PMID: 23245704.

38. Elena SF, Agudelo-Romero P, Lalic J. The evolution of viruses in multi-host fitness landscapes. Open Virol J. 2009;3:1-6. Epub 2009/07/03. doi: 10.2174/1874357900903010001. PubMed PMID: 19572052; PubMed Central PMCID: PMCPMC2703199.

39. Sullivan MB, Waterbury JB, Chisholm SW. Cyanophages infecting the oceanic cyanobacterium Prochlorococcus. Nature. 2003;424(6952):1047-51. Epub 2003/08/29. doi: 10.1038/nature01929. PubMed PMID: 12944965.

40. Woolhouse ME, Taylor LH, Haydon DT. Population biology of multihost pathogens. Science. 2001;292(5519):1109-12. Epub 2001/05/16. PubMed PMID: 11352066.

41. Orlova EV. How viruses infect bacteria? EMBO J. 2009;28(7):797-8. Epub 2009/04/09. doi: 10.1038/emboj.2009.71. PubMed PMID: 19352408; PubMed Central PMCID: PMCPMC2670874.

42. Herskowitz I, Hagen D. The lysis-lysogeny decision of phage lambda: explicit programming and responsiveness. Annu Rev Genet. 1980;14:399-445. Epub 1980/01/01. doi: 10.1146/annurev.ge.14.120180.002151. PubMed PMID: 6452089.

43. Erez Z, Steinberger-Levy I, Shamir M, Doron S, Stokar-Avihail A, Peleg Y, et al. Communication between viruses guides lysis-lysogeny decisions. Nature. 2017;541(7638):488-93. Epub 2017/01/19. doi: 10.1038/nature21049. PubMed PMID: 28099413; PubMed Central PMCID: PMCPMC5378303.

44. Weitz JS, Beckett SJ, Brum JR, Cael BB, Dushoff J. Lysis, lysogeny and virus-microbe ratios. Nature. 2017;549(7672):E1-E3. Epub 2017/09/22. doi: 10.1038/nature23295. PubMed PMID: 28933438.

45. Ptashne M. A genetic switch: phage lambda revisited: Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY:; 2004.

46. Ackers GK, Johnson AD, Shea MA. Quantitative model for gene regulation by lambda phage repressor. Proc Natl Acad Sci U S A. 1982;79(4):1129-33. Epub 1982/02/01. PubMed PMID: 6461856; PubMed Central PMCID: PMCPMC345914.

47. Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010;468(7320):67-71. Epub 2010/11/05. doi: 10.1038/nature09523. PubMed PMID: 21048762.

48. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Science. 2010;327(5962):167-70. Epub 2010/01/09. doi: 10.1126/science.1179555. PubMed PMID: 20056882.

49. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, et al. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol. 2011;9(6):467-77. Epub 2011/05/10. doi: 10.1038/nrmicro2577. PubMed PMID: 21552286; PubMed Central PMCID: PMCPMC3380444.

50. Bhaya D, Davison M, Barrangou R. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu Rev Genet. 2011;45:273-97. Epub 2011/11/09. doi: 10.1146/annurev-genet-110410-132430. PubMed PMID: 22060043.

51. Deveau H, Garneau JE, Moineau S. CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol. 2010;64:475-93. Epub 2010/06/10. doi: 10.1146/annurev.micro.112408.134123. PubMed PMID: 20528693.

52. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. 2015;13(11):722-36. Epub 2015/09/29. doi: 10.1038/nrmicro3569. PubMed PMID: 26411297; PubMed Central PMCID: PMCPMC5426118.

53. Barrangou R, Marraffini LA. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. Mol Cell. 2014;54(2):234-44. Epub 2014/04/29. doi: 10.1016/j.molcel.2014.03.011. PubMed PMID: 24766887; PubMed Central PMCID: PMCPMC4025954.

54. Rath D, Amlinger L, Rath A, Lundgren M. The CRISPR-Cas immune system: biology, mechanisms and applications. Biochimie. 2015;117:119-28. Epub 2015/04/15. doi: 10.1016/j.biochi.2015.03.025. PubMed PMID: 25868999.

55. Barrangou R, van der Oost J. Bacteriophage exclusion, a new defense system. EMBO J. 2015;34(2):134-5. Epub 2014/12/17. doi: 10.15252/embj.201490620. PubMed PMID: 25502457; PubMed Central PMCID: PMCPMC4337066.

56. Goldfarb T, Sberro H, Weinstock E, Cohen O, Doron S, Charpak-Amikam Y, et al. BREX is a novel phage resistance system widespread in microbial genomes. EMBO J. 2015;34(2):169-83. Epub 2014/12/03. doi: 10.15252/embj.201489455. PubMed PMID: 25452498; PubMed Central PMCID: PMCPMC4337064.

57. Ofir G, Melamed S, Sberro H, Mukamel Z, Silverman S, Yaakov G, et al. DISARM is a widespread bacterial defence system with broad anti-phage activities. Nat Microbiol. 2018;3(1):90-8. Epub 2017/11/01. doi: 10.1038/s41564-017-0051-0. PubMed PMID: 29085076; PubMed Central PMCID: PMCPMC5739279.

58. Beckett SJ, Williams HT. Coevolutionary diversification creates nested-modular structure in phage-bacteria interaction networks. Interface Focus. 2013;3(6):20130033. Epub 2014/02/12. doi: 10.1098/rsfs.2013.0033. PubMed PMID: 24516719; PubMed Central PMCID: PMCPMC3915849.

59. Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS. Statistical structure of host-phage interactions. Proc Natl Acad Sci U S A. 2011;108(28):E288-97. Epub 2011/06/29. doi: 10.1073/pnas.1101595108. PubMed PMID: 21709225; PubMed Central PMCID: PMCPMC3136311.

60. Hurwitz BL, Sullivan MB. The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. PLoS One. 2013;8(2):e57355. Epub 2013/03/08. doi: 10.1371/journal.pone.0057355. PubMed PMID: 23468974; PubMed Central PMCID: PMCPMC3585363.

61. Edwards RA, Rohwer F. Viral metagenomics. Nat Rev Microbiol. 2005;3(6):504-10. Epub 2005/05/12. doi: 10.1038/nrmicro1163. PubMed PMID: 15886693.

62. Rappe MS, Giovannoni SJ. The uncultured microbial majority. Annu Rev Microbiol. 2003;57:369-94. Epub 2003/10/07. doi: 10.1146/annurev.micro.57.030502.090759. PubMed PMID: 14527284.

63. Aylward FO, Boeuf D, Mende DR, Wood-Charlson EM, Vislova A, Eppley JM, et al. Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. Proc Natl Acad Sci U S A. 2017;114(43):11446-51. Epub 2017/10/27. doi: 10.1073/pnas.1714821114. PubMed PMID: 29073070; PubMed Central PMCID: PMCPMC5663388.

64. Tsementzi D, Wu J, Deutsch S, Nath S, Rodriguez RL, Burns AS, et al. SAR11 bacteria linked to ocean anoxia and nitrogen loss. Nature. 2016;536(7615):179-83. Epub 2016/08/04. doi: 10.1038/nature19068. PubMed PMID: 27487207; PubMed Central PMCID: PMCPMC4990128.

65. Brum JR, Sullivan MB. Rising to the challenge: accelerated pace of discovery transforms marine virology. Nat Rev Microbiol. 2015;13(3):147-59. Epub 2015/02/03. doi: 10.1038/nrmicro3404. PubMed PMID: 25639680.

66. Meyer JR, Dobias DT, Weitz JS, Barrick JE, Quick RT, Lenski RE. Repeatability and contingency in the evolution of a key innovation in phage lambda. Science. 2012;335(6067):428-32. Epub 2012/01/28. doi: 10.1126/science.1214449. PubMed PMID: 22282803; PubMed Central PMCID: PMCPMC3306806.

67. Lobo FP, Mota BE, Pena SD, Azevedo V, Macedo AM, Tauch A, et al. Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts. PLoS One. 2009;4(7):e6282. Epub 2009/07/21. doi: 10.1371/journal.pone.0006282. PubMed PMID: 19617912; PubMed Central PMCID: PMCPMC2707012.

68. MacPherson A, Otto SP, Nuismer SL. Keeping Pace with the Red Queen: Identifying the Genetic Basis of Susceptibility to Infectious Disease. Genetics. 2018;208(2):779-89. Epub 2017/12/11. doi: 10.1534/genetics.117.300481. PubMed PMID: 29223971; PubMed Central PMCID: PMCPMC5788537.

69. Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, Clark TG, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet. 2009;41(6):657-65. Epub 2009/05/26. doi: 10.1038/ng.388. PubMed PMID: 19465909; PubMed Central PMCID: PMCPMC2889040.

70. Scanlan PD, Hall AR, Lopez-Pascua LD, Buckling A. Genetic basis of infectivity evolution in a bacteriophage. Mol Ecol. 2011;20(5):981-9. Epub 2010/11/16. doi: 10.1111/j.1365-294X.2010.04903.x. PubMed PMID: 21073584.

71. de Jonge PA, Nobrega FL, Brouns SJJ, Dutilh BE. Molecular and Evolutionary Determinants of Bacteriophage Host Range. Trends Microbiol. 2018. Epub 2018/09/06. doi: 10.1016/j.tim.2018.08.006. PubMed PMID: 30181062.

72. Agrawal A, Lively CM. Infection genetics: gene-for-gene versus matching-alleles models and all points in between. Evolutionary Ecology Research. 2002;4(1):91-107.

73. Woolhouse ME, Webster JP, Domingo E, Charlesworth B, Levin BR. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nature genetics. 2002;32(4):569.

74. Gandon S, Buckling A, Decaestecker E, Day T. Host-parasite coevolution and patterns of adaptation across time and space. J Evol Biol. 2008;21(6):1861-6. Epub 2008/08/23. doi: 10.1111/j.1420-9101.2008.01598.x. PubMed PMID: 18717749.

75. Sasaki A. Host-parasite coevolution in a multilocus gene-for-gene system. Proc Biol Sci. 2000;267(1458):2183-8. Epub 2001/06/21. doi: 10.1098/rspb.2000.1267. PubMed PMID: 11413631; PubMed Central PMCID: PMCPMC1690804.

76. Frank SA. Coevolutionary Genetics of Plants and Pathogens. Evolutionary Ecology. 1993;7(1):45-75. doi: Doi 10.1007/Bf01237734. PubMed PMID: WOS:A1993KH13200004.

77. Munson-McGee JH, Peng S, Dewerff S, Stepanauskas R, Whitaker RJ, Weitz JS, et al. A virus or more in (nearly) every cell: ubiquitous networks of virus-host interactions in extreme environments. ISME J. 2018;12(7):1706-14. Epub 2018/02/23. doi: 10.1038/s41396-018-0071-7. PubMed PMID: 29467398; PubMed Central PMCID: PMCPMC6018696.

78. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature. 2016;536(7617):425-30. Epub 2016/08/18. doi: 10.1038/nature19094. PubMed PMID: 27533034.

79. Fuhrman JA, Schwalbach M. Viral influence on aquatic bacterial communities. Biol Bull. 2003;204(2):192-5. Epub 2003/04/18. doi: 10.2307/1543557. PubMed PMID: 12700152.

80. Rohwer F, Thurber RV. Viruses manipulate the marine environment. Nature. 2009;459(7244):207-12. Epub 2009/05/16. doi: 10.1038/nature08060. PubMed PMID: 19444207.

81. Breitbart M. Marine viruses: truth or dare. Ann Rev Mar Sci. 2012;4:425-48. Epub 2012/03/31. doi: 10.1146/annurev-marine-120709-142805. PubMed PMID: 22457982.

82. Needham DM, Chow CE, Cram JA, Sachdeva R, Parada A, Fuhrman JA. Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. ISME J. 2013;7(7):1274-85. Epub 2013/03/01. doi: 10.1038/ismej.2013.19. PubMed PMID: 23446831; PubMed Central PMCID: PMCPMC3695287.

83. Sullivan MB, Weitz JS, Wilhelm S. Viral ecology comes of age. Environ Microbiol Rep. 2017;9(1):33-5. Epub 2016/11/27. doi: 10.1111/1758-2229.12504. PubMed PMID: 27888577.

84. Malki K, Kula A, Bruder K, Sible E, Hatzopoulos T, Steidel S, et al. Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. Virol J. 2015;12:164. Epub 2015/10/11. doi: 10.1186/s12985-015-0395-0. PubMed PMID: 26453042; PubMed Central PMCID: PMCPMC4600314.

85. Comeau AM, Buenaventura E, Suttle CA. A persistent, productive, and seasonally dynamic vibriophage population within Pacific oysters (Crassostrea gigas). Appl Environ Microbiol. 2005;71(9):5324-31. Epub 2005/09/10. doi: 10.1128/AEM.71.9.5324-5331.2005. PubMed PMID: 16151121; PubMed Central PMCID: PMCPMC1214601.

86. Winstanley C, Langille MG, Fothergill JL, Kukavica-Ibrulj I, Paradis-Bleau C, Sanschagrin F, et al. Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of Pseudomonas aeruginosa. Genome Res. 2009;19(1):12-23. Epub 2008/12/03. doi: 10.1101/gr.086082.108. PubMed PMID: 19047519; PubMed Central PMCID: PMCPMC2612960.

87. Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ. Biogeography of the Sulfolobus islandicus pan-genome. Proc Natl Acad Sci U S A. 2009;106(21):8605-10.

Epub 2009/05/14. doi: 10.1073/pnas.0808945106. PubMed PMID: 19435847; PubMed Central PMCID: PMCPMC2689034.

88. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage-host relationships. FEMS Microbiol Rev. 2016;40(2):258-72. Epub 2015/12/15. doi: 10.1093/femsre/fuv048. PubMed PMID: 26657537; PubMed Central PMCID: PMCPMC5831537.

89. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science. 2015;348(6237):1261498. Epub 2015/05/23. doi: 10.1126/science.1261498. PubMed PMID: 25999515.

90. Snyder JC, Bateson MM, Lavin M, Young MJ. Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. Appl Environ Microbiol. 2010;76(21):7251-8. Epub 2010/09/21. doi: 10.1128/AEM.01109-10. PubMed PMID: 20851987; PubMed Central PMCID: PMCPMC2976250.

91. Anderson RE, Brazelton WJ, Baross JA. Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. FEMS Microbiol Ecol. 2011;77(1):120-33. Epub 2011/03/18. doi: 10.1111/j.1574-6941.2011.01090.x. PubMed PMID: 21410492.

92. Berg Miller ME, Yeoman CJ, Chia N, Tringe SG, Angly FE, Edwards RA, et al. Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. Environ Microbiol. 2012;14(1):207-27. Epub 2011/10/19. doi: 10.1111/j.1462-2920.2011.02593.x. PubMed PMID: 22004549.

93. Allers E, Moraru C, Duhaime MB, Beneze E, Solonenko N, Barrero-Canosa J, et al. Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. Environ Microbiol. 2013;15(8):2306-18. Epub 2013/03/16. doi: 10.1111/1462-2920.12100. PubMed PMID: 23489642; PubMed Central PMCID: PMCPMC3884771.

94. Deng L, Gregory A, Yilmaz S, Poulos BT, Hugenholtz P, Sullivan MB. Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as revealed by viral tagging. MBio. 2012;3(6). Epub 2012/11/01. doi: 10.1128/mBio.00373-12. PubMed PMID: 23111870; PubMed Central PMCID: PMCPMC3487772.

95. Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, et al. Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. Nature. 2014;513(7517):242-5. Epub 2014/07/22. doi: 10.1038/nature13459. PubMed PMID: 25043051.

96. Tadmor AD, Ottesen EA, Leadbetter JR, Phillips R. Probing individual environmental bacteria for viruses by using microfluidic digital PCR. Science.

2011;333(6038):58-62. Epub 2011/07/02. doi: 10.1126/science.1200758. PubMed PMID: 21719670; PubMed Central PMCID: PMCPMC3261838.

97. Stepanauskas R. Single cell genomics: an individual look at microbes. Curr Opin Microbiol. 2012;15(5):613-20. Epub 2012/10/03. doi: 10.1016/j.mib.2012.09.001. PubMed PMID: 23026140.

98. Roux S, Hawley AK, Torres Beltran M, Scofield M, Schwientek P, Stepanauskas R, et al. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. Elife. 2014;3:e03125. Epub 2014/08/31. doi: 10.7554/eLife.03125. PubMed PMID: 25171894; PubMed Central PMCID: PMCPMC4164917.

99. Labonte JM, Field EK, Lau M, Chivian D, Van Heerden E, Wommack KE, et al. Single cell genomics indicates horizontal gene transfer and viral infections in a deep subsurface Firmicutes population. Front Microbiol. 2015;6:349. Epub 2015/05/09. doi: 10.3389/fmicb.2015.00349. PubMed PMID: 25954269; PubMed Central PMCID: PMCPMC4406082.

100. Wilson WH, Gilg IC, Moniruzzaman M, Field EK, Koren S, LeCleir GR, et al. Genomic exploration of individual giant ocean viruses. ISME J. 2017;11(8):1736-45. Epub 2017/05/13. doi: 10.1038/ismej.2017.61. PubMed PMID: 28498373; PubMed Central PMCID: PMCPMC5520044.

101. Labonte JM, Swan BK, Poulos B, Luo H, Koren S, Hallam SJ, et al. Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. ISME J. 2015;9(11):2386-99. Epub 2015/04/08. doi: 10.1038/ismej.2015.48. PubMed PMID: 25848873; PubMed Central PMCID: PMCPMC4611503.

102. Stepanauskas R, Sieracki ME. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. Proc Natl Acad Sci U S A. 2007;104(21):9052-7. Epub 2007/05/16. doi: 10.1073/pnas.0700496104. PubMed PMID: 17502618; PubMed Central PMCID: PMCPMC1885626.

103. Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, et al. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. Science. 2011;333(6047):1296-300. Epub 2011/09/03. doi: 10.1126/science.1203690. PubMed PMID: 21885783.

104. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. Epub 2014/04/04. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PubMed Central PMCID: PMCPMC4103590.

105. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77. Epub 2012/04/18. doi: 10.1089/cmb.2012.0021. PubMed PMID: 22506599; PubMed Central PMCID: PMCPMC3342519.

106. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25(7):1043-55. Epub 2015/05/16. doi: 10.1101/gr.186072.114. PubMed PMID: 25977477; PubMed Central PMCID: PMCPMC4484387.

107. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res. 2017;45(1):39-53. Epub 2016/12/03. doi: 10.1093/nar/gkw1002. PubMed PMID: 27899557; PubMed Central PMCID: PMCPMC5224470.

108. Manrique P, Bolduc B, Walk ST, van der Oost J, de Vos WM, Young MJ. Healthy human gut phageome. Proc Natl Acad Sci U S A. 2016;113(37):10400-5. Epub 2016/08/31. doi: 10.1073/pnas.1601060113. PubMed PMID: 27573828; PubMed Central PMCID: PMCPMC5027468.

109. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106(45):19126-31. Epub 2009/10/27. doi: 10.1073/pnas.0906412106. PubMed PMID: 19855009; PubMed Central PMCID: PMCPMC2776425.

110. von Neubeck M, Huptas C, Gluck C, Krewinkel M, Stoeckel M, Stressler T, et al. Pseudomonas helleri sp. nov. and Pseudomonas weihenstephanensis sp. nov., isolated from raw cow's milk. Int J Syst Evol Microbiol. 2016;66(3):1163-73. Epub 2015/12/18. doi: 10.1099/ijsem.0.000852. PubMed PMID: 26675012.

111. Flores CO, Poisot T, Valverde S, Weitz JS. BiMat: a MATLAB package to facilitate the analysis of bipartite networks. Methods Ecol Evol. 2016;7(1):127-32.

112. Edgar RC. PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics. 2007;8:18. Epub 2007/01/24. doi: 10.1186/1471-2105-8-18. PubMed PMID: 17239253; PubMed Central PMCID: PMCPMC1790904.

113. Munson-McGee JH, Field EK, Bateson M, Rooney C, Stepanauskas R, Young MJ. Nanoarchaeota, Their Sulfolobales Host, and Nanoarchaeota Virus Distribution across Yellowstone National Park Hot Springs. Appl Environ Microbiol. 2015;81(22):7860-8. Epub 2015/09/06. doi: 10.1128/AEM.01539-15. PubMed PMID: 26341207; PubMed Central PMCID: PMCPMC4616950.

114. Podar M, Makarova KS, Graham DE, Wolf YI, Koonin EV, Reysenbach AL. Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. Biol Direct. 2013;8:9. Epub 2013/04/24. doi: 10.1186/1745-6150-8-9. PubMed PMID: 23607440; PubMed Central PMCID: PMCPMC3655853.

115. Rice G, Tang L, Stedman K, Roberto F, Spuhler J, Gillitzer E, et al. The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans

all domains of life. Proc Natl Acad Sci U S A. 2004;101(20):7716-20. Epub 2004/05/05. doi: 10.1073/pnas.0401773101. PubMed PMID: 15123802; PubMed Central PMCID: PMCPMC419672.

116. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10(8):538-50. Epub 2012/07/17. doi: 10.1038/nrmicro2832. PubMed PMID: 22796884.

117. Flannick J, Novak A, Srinivasan BS, McAdams HH, Batzoglou S. Graemlin: general and robust alignment of multiple large interaction networks. Genome Res. 2006;16(9):1169-81. Epub 2006/08/11. doi: 10.1101/gr.5235706. PubMed PMID: 16899655; PubMed Central PMCID: PMCPMC1557769.

118. Berry D, Widder S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Front Microbiol. 2014;5:219. Epub 2014/06/07. doi: 10.3389/fmicb.2014.00219. PubMed PMID: 24904535; PubMed Central PMCID: PMCPMC4033041.

119. Neurath AR, Kent SB, Strick N, Parker K. Identification and chemical synthesis of a host cell receptor binding site on hepatitis B virus. Cell. 1986;46(3):429-36. Epub 1986/08/01. PubMed PMID: 3015414.

120. Wang J, Hofnung M, Charbit A. The C-terminal portion of the tail fiber protein of bacteriophage lambda is responsible for binding to LamB, its receptor at the surface of Escherichia coli K-12. J Bacteriol. 2000;182(2):508-12. Epub 2000/01/12. PubMed PMID: 10629200; PubMed Central PMCID: PMCPMC94303.

121. Chatterjee S, Rothenberg E. Interaction of bacteriophage l with its E. coli receptor, LamB. Viruses. 2012;4(11):3162-78. Epub 2012/12/04. doi: 10.3390/v4113162. PubMed PMID: 23202520; PubMed Central PMCID: PMCPMC3509688.

122. Bajić D, Vila JC, Blount ZD, Sanchez A. On the deformability of an empirical fitness landscape by microbial evolution. bioRxiv. 2018:293407.

123. Buckling A, Rainey PB. Antagonistic coevolution between a bacterium and a bacteriophage. Proc Biol Sci. 2002;269(1494):931-6. Epub 2002/05/25. doi: 10.1098/rspb.2001.1945. PubMed PMID: 12028776; PubMed Central PMCID: PMCPMC1690980.

124. Elena SF, Lenski RE. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet. 2003;4(6):457-69. Epub 2003/05/31. doi: 10.1038/nrg1088. PubMed PMID: 12776215.

125. Poullain V, Gandon S, Brockhurst MA, Buckling A, Hochberg ME. The evolution of specificity in evolving and coevolving antagonistic interactions between a bacteria and its phage. Evolution. 2008;62(1):1-11. Epub 2007/11/17. doi: 10.1111/j.1558-5646.2007.00260.x. PubMed PMID: 18005153.

126. Kaltz O, Shykoff J. Within-and among-population variation in infectivity, latency and spore production in a host–pathogen system. Journal of Evolutionary Biology. 2002;15(5):850-60.

127. Gurney J, Aldakak L, Betts A, Gougat-Barbera C, Poisot T, Kaltz O, et al. Network structure and local adaptation in co-evolving bacteria-phage interactions. Mol Ecol. 2017;26(7):1764-77. Epub 2017/01/17. doi: 10.1111/mec.14008. PubMed PMID: 28092408.

128. Horton MW, Bodenhausen N, Beilsmith K, Meng D, Muegge BD, Subramanian S, et al. Genome-wide association study of Arabidopsis thaliana leaf microbial community. Nat Commun. 2014;5:5320. Epub 2014/11/11. doi: 10.1038/ncomms6320. PubMed PMID: 25382143; PubMed Central PMCID: PMCPMC4232226.

129. Falush D. Bacterial genomics: Microbial GWAS coming of age. Nat Microbiol. 2016;1:16059. Epub 2016/08/31. doi: 10.1038/nmicrobiol.2016.59. PubMed PMID: 27572652.

130. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet. 2017;18(1):41-50. Epub 2016/11/15. doi: 10.1038/nrg.2016.132. PubMed PMID: 27840430.

131. Wei WH, Hemani G, Haley CS. Detecting epistasis in human complex traits. Nat Rev Genet. 2014;15(11):722-33. Epub 2014/09/10. doi: 10.1038/nrg3747. PubMed PMID: 25200660.

132. An P, Mukherjee O, Chanda P, Yao L, Engelman CD, Huang CH, et al. The challenge of detecting epistasis (G x G interactions): Genetic Analysis Workshop 16. Genet Epidemiol. 2009;33 Suppl 1:S58-67. Epub 2009/11/20. doi: 10.1002/gepi.20474. PubMed PMID: 19924703; PubMed Central PMCID: PMCPMC3692280.

133. Gibson G. A primer of human genetics: Sinauer Associates, Incorporated, Publishers; 2015.

134. Dutilh BE, Reyes A, Hall RJ, Whiteson KL. Editorial: Virus Discovery by Metagenomics: The (Im)possibilities. Front Microbiol. 2017;8:1710. Epub 2017/09/26. doi: 10.3389/fmicb.2017.01710. PubMed PMID: 28943867; PubMed Central PMCID: PMCPMC5596103.

135. Sullivan NJ, Geisbert TW, Geisbert JB, Xu L, Yang ZY, Roederer M, et al. Accelerated vaccination for Ebola virus haemorrhagic fever in non-human primates. Nature. 2003;424(6949):681-4. Epub 2003/08/09. doi: 10.1038/nature01876. PubMed PMID: 12904795.

136. Modi SR, Lee HH, Spina CS, Collins JJ. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. Nature. 2013;499(7457):219-22. Epub 2013/06/12. doi: 10.1038/nature12212. PubMed PMID: 23748443; PubMed Central PMCID: PMCPMC3710538.

137. Werts C, Michel V, Hofnung M, Charbit A. Adsorption of bacteriophage lambda on the LamB protein of Escherichia coli K-12: point mutations in gene J of lambda responsible for extended host range. J Bacteriol. 1994;176(4):941-7. Epub 1994/02/01. PubMed PMID: 8106335; PubMed Central PMCID: PMCPMC205142.

138. Wang J, Michel V, Hofnung M, Charbit A. Cloning of the J gene of bacteriophage lambda, expression and solubilization of the J protein: first in vitro studies on the interactions between J and LamB, its cell surface receptor. Res Microbiol. 1998;149(9):611-24. Epub 1998/11/25. PubMed PMID: 9826917.

139. Chang CY, Nam K, Young R. S gene expression and the timing of lysis by bacteriophage lambda. J Bacteriol. 1995;177(11):3283-94. Epub 1995/06/01. PubMed PMID: 7768829; PubMed Central PMCID: PMCPMC177022.

140. Anderson B, Rashid MH, Carter C, Pasternack G, Rajanna C, Revazishvili T, et al. Enumeration of bacteriophage particles: Comparative analysis of the traditional plaque assay and real-time QPCR- and nanosight-based assays. Bacteriophage. 2011;1(2):86-93. Epub 2012/02/16. doi: 10.4161/bact.1.2.15456. PubMed PMID: 22334864; PubMed Central PMCID: PMCPMC3278645.

141. Sambrook J, Russell DW, Sambrook J. The condensed protocols from Molecular cloning : a laboratory manual. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press; 2006. v, 800 p. p.

142. Vica Pacheco S, Garcia Gonzalez O, Paniagua Contreras GL. The lom gene of bacteriophage lambda is involved in Escherichia coli K12 adhesion to human buccal epithelial cells. FEMS Microbiol Lett. 1997;156(1):129-32. Epub 1997/11/22. PubMed PMID: 9368371.

143. Suttle CA. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol. 2007;5(10):801-12. Epub 2007/09/15. doi: 10.1038/nrmicro1750. PubMed PMID: 17853907.

144. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. Science. 2008;320(5879):1034-9. Epub 2008/05/24. doi: 10.1126/science.1153213. PubMed PMID: 18497287.

145. Fuhrman JA, Noble RT. Viruses and protists cause similar bacterial mortality in coastal seawater. Limnology and Oceanography. 1995;40(7):1236-42.

146. Gobler CJ, Hutchins DA, Fisher NS, Cosper EM, Saňudo-Wilhelmy SA. Release and bioavailability of C, N, P Se, and Fe following viral lysis of a marine chrysophyte. Limnology and Oceanography. 1997;42(7):1492-504.

147. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010;8(5):317-27. Epub 2010/03/30. doi: 10.1038/nrmicro2315. PubMed PMID: 20348932.

148. Buckling A, Rainey PB. The role of parasites in sympatric and allopatric host diversification. Nature. 2002;420(6915):496-9. Epub 2002/12/06. doi: 10.1038/nature01164. PubMed PMID: 12466840.

149. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. Science. 2018;359(6379). Epub 2018/01/27. doi: 10.1126/science.aar4120. PubMed PMID: 29371424.

150. Stern A, Sorek R. The phage-host arms race: shaping the evolution of microbes. Bioessays. 2011;33(1):43-51. Epub 2010/10/28. doi: 10.1002/bies.201000071. PubMed PMID: 20979102; PubMed Central PMCID: PMCPMC3274958.

151. Dy RL, Richter C, Salmond GP, Fineran PC. Remarkable Mechanisms in Microbes to Resist Phage Infections. Annu Rev Virol. 2014;1(1):307-31. Epub 2014/11/03. doi: 10.1146/annurev-virology-031413-085500. PubMed PMID: 26958724.

152. Agrawal A, Lively CM. Infection genetics: gene-for-gene versus matching-alleles models and all points in between. Evolutionary Ecology Research. 2002;4(1):79-90. PubMed PMID: WOS:000174068800005.

153. Woolhouse MEJ, Webster JP, Domingo E, Charlesworth B, Levin BR. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nature Genetics. 2002;32(4):569-77. doi: 10.1038/ng1202-569. PubMed PMID: WOS:000179593000007.

154. Pelosi L, Kuhn L, Guetta D, Garin J, Geiselmann J, Lenski RE, et al. Parallel changes in global protein profiles during long-term experimental evolution in Escherichia coli. Genetics. 2006;173(4):1851-69. Epub 2006/05/17. doi: 10.1534/genetics.105.049619. PubMed PMID: 16702438; PubMed Central PMCID: PMCPMC1569701.

155. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS One. 2015;10(5):e0128036. Epub 2015/05/23. doi: 10.1371/journal.pone.0128036. PubMed PMID: 26000737; PubMed Central PMCID: PMCPMC4441430.

156. Winter C, Bouvier T, Weinbauer MG, Thingstad TF. Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "killing the winner" hypothesis revisited. Microbiol Mol Biol Rev. 2010;74(1):42-57. Epub 2010/03/04. doi: 10.1128/MMBR.00034-09. PubMed PMID: 20197498; PubMed Central PMCID: PMCPMC2832346.

157. Flores CO, Valverde S, Weitz JS. Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. ISME J. 2013;7(3):520-32. Epub 2012/11/28. doi: 10.1038/ismej.2012.135. PubMed PMID: 23178671; PubMed Central PMCID: PMCPMC3578562.

158. Betts A, Kaltz O, Hochberg ME. Contrasted coevolutionary dynamics between a bacterial pathogen and its bacteriophages. Proc Natl Acad Sci U S A. 2014;111(30):11109-14. Epub 2014/07/16. doi: 10.1073/pnas.1406763111. PubMed PMID: 25024215; PubMed Central PMCID: PMCPMC4121802.

159. Hall AR, Scanlan PD, Morgan AD, Buckling A. Host-parasite coevolutionary arms races give way to fluctuating selection. Ecol Lett. 2011;14(7):635-42. Epub 2011/04/28. doi: 10.1111/j.1461-0248.2011.01624.x. PubMed PMID: 21521436.

160. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17(1):pp. 10-2.

161. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.

162. Deatherage DE, Barrick JE. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol Biol. 2014;1151:165-88. Epub 2014/05/20. doi: 10.1007/978-1-4939-0554-6_12. PubMed PMID: 24838886; PubMed Central PMCID: PMCPMC4239701.

163. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9. Epub 2012/03/06. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PubMed Central PMCID: PMCPMC3322381.

164. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977;267(5608):275-6. Epub 1977/05/19. PubMed PMID: 865622.

165. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 2000;15(12):496-503. Epub 2000/12/15. PubMed PMID: 11114436.

166. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059-66. Epub 2002/07/24. PubMed PMID: 12136088; PubMed Central PMCID: PMCPMC135756.

167. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3. Epub 2014/01/24. doi: 10.1093/bioinformatics/btu033. PubMed PMID: 24451623; PubMed Central PMCID: PMCPMC3998144.

168. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4(1):vex042. Epub 2018/01/18. doi: 10.1093/ve/vex042. PubMed PMID: 29340210; PubMed Central PMCID: PMCPMC5758920.

169. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592-3. Epub 2010/12/21. doi: 10.1093/bioinformatics/btq706. PubMed PMID: 21169378; PubMed Central PMCID: PMCPMC3035803.

170. Makarenkov V. T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. Bioinformatics. 2001;17(7):664-8. Epub 2001/07/13. PubMed PMID: 11448889.

171. Cooper VS, Schneider D, Blot M, Lenski RE. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of Escherichia coli B. J Bacteriol. 2001;183(9):2834-41. Epub 2001/04/09. doi: 10.1128/JB.183.9.2834-2841.2001. PubMed PMID: 11292803; PubMed Central PMCID: PMCPMC99500.

172. Maynard ND, Birch EW, Sanghvi JC, Chen L, Gutschow MV, Covert MW. A forward-genetic screen and dynamic analysis of lambda phage host-dependencies reveals an extensive interaction network and a new anti-viral strategy. PLoS Genet. 2010;6(7):e1001017. Epub 2010/07/16. doi: 10.1371/journal.pgen.1001017. PubMed PMID: 20628568; PubMed Central PMCID: PMCPMC2900299.

173. Burmeister AR, Lenski RE, Meyer JR. Host coevolution alters the adaptive landscape of a virus. Proc Biol Sci. 2016;283(1839). Epub 2016/09/30. doi: 10.1098/rspb.2016.1528. PubMed PMID: 27683370; PubMed Central PMCID: PMCPMC5046904.

174. Maddamsetti R, Johnson DT, Spielman SJ, Petrie KL, Marks DS, Meyer JR. Gain-of-function experiments with bacteriophage lambda uncover residues under diversifying selection in nature. Evolution. 2018. Epub 2018/08/29. doi: 10.1111/evo.13586. PubMed PMID: 30152871.

175. Berkane E, Orlik F, Stegmeier JF, Charbit A, Winterhalter M, Benz R. Interaction of bacteriophage lambda with its cell surface receptor: an in vitro study of binding of the viral tail protein gpJ to LamB (Maltoporin). Biochemistry. 2006;45(8):2708-20. Epub 2006/02/24. doi: 10.1021/bi051800v. PubMed PMID: 16489764.

176. Grayson P, Han L, Winther T, Phillips R. Real-time observations of single bacteriophage lambda DNA ejections in vitro. Proc Natl Acad Sci U S A. 2007;104(37):14652-7. Epub 2007/09/07. doi: 10.1073/pnas.0703274104. PubMed PMID: 17804798; PubMed Central PMCID: PMCPMC1976217.

177. Erni B, Zanolari B, Kocher HP. The mannose permease of Escherichia coli consists of three different proteins. Amino acid sequence and function in sugar transport, sugar phosphorylation, and penetration of phage lambda DNA. J Biol Chem. 1987;262(11):5238-47. Epub 1987/04/15. PubMed PMID: 2951378.

178. Esquinas-Rychen M, Erni B. Facilitation of bacteriophage lambda DNA injection by inner membrane proteins of the bacterial phosphoenol-pyruvate: carbohydrate phosphotransferase system (PTS). J Mol Microbiol Biotechnol. 2001;3(3):361-70. Epub 2001/05/22. PubMed PMID: 11361066.

179. Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature.

2015;519(7542):181-6. Epub 2015/03/04. doi: 10.1038/nature14279. PubMed PMID: 25731169; PubMed Central PMCID: PMCPMC4426284.

180. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proc Natl Acad Sci U S A. 2016;113(12):E1701-9. Epub 2016/03/10. doi: 10.1073/pnas.1525578113. PubMed PMID: 26951657; PubMed Central PMCID: PMCPMC4812706.