Virology By The Numbers

A Quantitative Exploration of Viral Energetics, Genomics, and Ecology

Thesis by

Gita Mahmoudabadi

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2018

(Defended Nov. 20th, 2017)

Gita Mahmoudabadi
ORCID: 0000-0002-8812-7246

To my loving family

ACKNOWLEDGEMENTS

There are many people to whom I am forever indebted. Without them, my journey in science may never have started or would not have been nearly as fulfilling. Without a doubt, my parents' numerous sacrifices have served as the foundation for the incredible educational opportunities that I have received. I want to thank them and my sister for having held a fan on the sparks of my curiosity, and for giving me the courage to chase after my dreams. I am deeply grateful to my grandmother who has always encouraged me to pursue an academic life, believing that passing on one's knowledge is the greatest form of public service.

I am also grateful to many people at GA Tech and Johns Hopkins, where my dream of becoming a scientist was well received and nurtured. In particular, I spent three years working in Valeria Milam's lab. I am very thankful for the guidance that she and Brian Baker provided me. In her lab, I gained valuable insight into the physical properties of materials, and explored the ways in which biopolymers could be used to engineer drug-delivery systems. Furthermore, I am grateful to Mark Styczynski, Joshua Weitz, and Eric Gaucher for mentoring me through an intriguing experience with synthetic biology. I would also like to thank Prakash Kulkarni for exposing me to another facet of biopolymers, namely their capacity for information storage and evolution. Working together on protein evolution, I realized that while experiments serve as the backbone of biology, there is something fundamentally liberating about exploring a topic with nothing more than focused thoughts. This realization was further refined in my interactions with my PhD advisor, Rob Phillips, who asserts that a theory is neither meant as an after-thought to an experiment, nor is it precise unless it is written in the language of mathematics.

Coming to Caltech was a dream come true. I am incredibly privileged for having been at this institution and for having learned from so many passionate scientists throughout the years. Above all, I am very lucky to have met Rob. I decided to join Rob's lab after taking his course on Physical Biology of the Cell, which weaved together topics as seemingly disparate as thermodynamics and transcription. I remember leaving the class feeling inspired, a feeling that has prevailed the passage of time. Those early lectures made lasting impressions on me as is evident from the "street-fighting mathematics" approach employed in estimating the energetic cost of building a virus. Having now been behind the scenes, I know that Rob's seamless performances and engaging books come at a high cost. I am very happy to have had a tough coach whose motto is "bloody drills and bloodless battles".

I am also grateful to Rob for the chance to experience a side of biology that is often overshadowed. I was part of the first generation of spoiled graduate students who went to South Korea to TA for Rob's Evolution course, which later convened in Indonesia for a biogeography field trip that included 4 AM hikes and boat trips to the Wallace Line. Besides the obvious cool-factor of cooking your breakfast at the summit of Mt. Batur using the heat from this volcano, or supping towards Alaska's glaciers past the puzzled seals, Rob's motivation for leading these adventures, in his own words, was to show us that "there is more to biology than PCRs". I am greatly indebted to both the physicist Rob and the naturalist Rob. Much of my curiosity about the global distribution of viruses arises from learning about Wallace, Darwin, and their specimen collections on these trips.

Being the larger than life character that Rob is, he attracts outstanding collaborators. As such, our work has been enriched by interactions with numerous people. I want to especially thank Lea Goentoro, Victoria Orphan and Jared Leadbetter who have been part of my thesis

committee and provided me great feedback. I would also like to thank Ron Milo and Michael Lynch who have been instrumental in our work on bioenergetics. I have also gained invaluable insight from my interactions with Bill Gelbart, Eddy Rubin, Markus Covert and David Baltimore who have reviewed our work on viral genomics. I would like to thank the Boundaries of Life collaborators, in particular Steve Quake, Nathan Wolfe, and Grant Jensen who have introduced me to the Shadow Life hypothesis. I would like to Linas Mazutis and his lab for helping us start single-virus, single-cell infection experiments. I would also like to thank Richard Murray and Michael Elowitz in whose labs I rotated and learned about cell-free synthetic circuits and epigenetic regulation. I am grateful to Niles Pierce and Linda Scott for welcoming my class to Caltech and guiding us through the first year.

It was a pleasure to mentor several undergraduate students: Adam Catching, Alison Cheung, Brian Silver, Matt Morgan, and Ana Mahmoudabadi. I thoroughly enjoyed and substantially benefited from these one-on-one mentoring experiences. I would also like to thank many former and current members of the Phillips lab whom I am lucky to have met and learned from: Maja Bialecka-Fornal, Kelsey Homyk, Franz Weinert, James Boedicker, Geoff Lovely, Lacramioara Bintu, Hernan Garcia, David Van Valen, Stephanie Johnson, Mattias Rydenfelt, Heun Jin Lee, Robert Brewster, Justin Bois, Yi-Ju Chen, Arbel Tadmor, Daniel Jones, Katie Miller, Celene Barerra, Jonathan Gross, Manuel Razo, Stephanie Barnes, Nathan Belliveau, Helen Foley, Griffin Chure, Soichi Hirokawa, Bill Ireland, Suzy Beeler, Vahe Galstyan, David Angeles, and Muir Morrison.

Furthermore, I am greatly indebted to Mysha Sarwar, Morgane Grivel, Pablo Guerrero, Marcello Gori, Lee Wilson, Matt Gethers, Merri Sheffield, Batool Nekooie, Reza Saadein, Hassan Jafarzadeh, Tami Khazaei, Albert Sun, Amira Pettus, Saima Mehmood, Maha Hosain,

Stephanie Tzouanas, Alborz Mahdavi, and many others for their friendship. I would like to especially thank Parisa Farmand, Melika Motamedi, Zahra Daftarian, Sheliza Bhanjee, Aliyah and Nadia Saadein, Mariya Akbashev, Ghazal and Ava Motakef, Lynn Tsai, and Sam Amin for their help as citizen scientists on my experiments.

I don't even know where to begin with thanking Reza Khiabani, my best friend and the love of my life. We would not have made it through several years of long-distance relationship had it not been for his unwavering love and encouragement, so I will end with a quote from Hafez. Thank you Reza for lighting up my world.

Even after all this time

the Sun never says to the Earth "you owe me."

Look what happens with a love like that

it lights the whole sky.

ABSTRACT

Over the past couple of decades, technological advancements in sequencing and imaging have unequivocally proven that the world of viruses is far bigger and more consequential than previously imagined. There are 10^{31} viruses estimated to inhabit our planet, outnumbering even bacteria. Despite their astronomical numbers and staggering sequence diversity, environmental viruses are poorly characterized. In this thesis we will demonstrate our three-pronged exploration of viruses through the lenses of energetics (Chapters 2 and 3), genomics (Chapter 4) and ecology (Chapter 5). We will first focus on one of the defining features of viruses, namely their reliance on their host for energy, and demonstrate the energetic cost of building a virus and mounting an infection. In our second study, we present one of the largest surveys of complete viral genomes, providing a comprehensive and quantitative snapshot of viral genomic trends for thousands of viruses. In our third study, we shift our focus towards ecological questions surrounding the large number of commensal phages inhabiting the human body. We discovered that phage community composition could serve as a fingerprint, or a "phageprint" – highly personal and stable over time. To our knowledge, this study is one of the largest studies of human phages and the first to demonstrate the feasibility of human identification based on phage sequences.

PUBLISHED CONTENT AND CONTRIBUTIONS

Mahmoudabadi G, Milo R, & Phillips R. (2017) Energetic cost of building a virus. Proceedings of the National Academy of Sciences 114(22):E4324-E4333. doi:10.1073/pnas.1701670114

G.M. initiated the project, carried out energetic cost estimates, prepared the data, and participated in the writing of the manuscript.

Mahmoudabadi, G., & Phillips, R. (2018) A comprehensive and quantitative exploration of thousands of viral genomes. eLife 7:e31955. doi: 10.7554/eLife.31955.001

G.M. initiated the project, developed bioinformatics pipeline, carried out data analysis, and participated in the writing of the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS AND/OR TABLES

Number *Page*

Chapter II

Figures

SI Dataset

Chapter III

Figures

Tables

SI Figures

SI Tables

SI Datasets

Chapter IV

Figures

Tables

SI Figures

SI Tables

SI Text Files

Chapter V

Figures

SI Tables

Chapter I

Introduction

Over the past couple of decades, technological advancements in sequencing and imaging have unequivocally proven that the world of viruses is far bigger and more consequential than previously imagined (1-7). There are 10^{31} viruses estimated to inhabit our planet, outnumbering even bacteria (8, 9). Viruses have been shown to impact biogeochemical cycles (10-13) and evolution of host organisms (4, 14-16).

Despite their astronomical numbers and staggering sequence diversity (17, 18), we have large gaps in our understanding of environmental viruses. This is in part due to the field being at its infancy. It is also due to the conceptual and technical challenges that are unique to the study of viruses. For example, in contrast to ribosomal RNA sequences that are conserved across cellular genomes and serve as the basis for taxonomic and evolutionary studies, there are no universally conserved sequences within viral genomes. As a result, we have not yet been able to develop a genomic classification for viruses. Moreover, our understanding of the deep-time evolutionary history of viruses remains limited.

However, these alienating differences between viruses and cells can also serve to significantly broaden our understanding of biology and overturn what we may have accepted as dogma. Viruses, by virtue of encoding their genomes in double- and single-stranded versions of both RNA and DNA, provide us a window into an alternative biology. What I am alluding to is already a big part of biology's recent history. For example, the central dogma was turned on its head by Baltimore (19) Temin and Mizutani (20) through their

discovery of viral reverse transcription. The future of biology will also likely be impacted by the study of viruses that are caught breaking biology's perceived rules.

Viruses have also offered us the platform and the tools for some of biology's most infamous experiments. Hershey and Chase used a phage infection experiment to contest the popular assumption that proteins, rather than DNA, were responsible for inheritance (21). Luria and Delbruck also used a phage experiment to challenge Lamarckian evolution (22). Furthermore, Viruses have significantly contributed to our molecular biology toolbox (23). In lieu of recent technological advances, the study of viruses will inevitably result in even greater number of tools and insights.

In this thesis we will demonstrate our three-pronged exploration of viruses through the lenses of energetics (Chapters 2 and 3), genomics (Chapter 4), and ecology (Chapter 5). We will first focus on one of the defining features of viruses, namely their reliance on their host for energy, and demonstrate the energetic cost of building a virus and mounting an infection. Although many experimental studies make it clear that viruses are parasitic entities that hijack the molecular resources of the host, a detailed estimate for the energetic cost of viral synthesis was largely lacking. To quantify the energetic cost of viruses to their hosts, we first had to develop a framework for describing cellular energetic costs and their evolutionary consequences (Chapter 2).

With the energetic cost of molecular building blocks in hand, we then enumerated the costs associated with two very distinct but representative DNA and RNA viruses, namely, T4 and influenza (Chapter 3). We found that for these viruses, translation of viral proteins is the most energetically expensive process. Interestingly, we found the cost of building a T4 phage and a single influenza virus were nearly the same. Due to influenza's higher burst size, however, the overall cost of a T4 phage infection is only a small fraction of the cost of an

influenza infection. The costs of these infections relative to their host's estimated energy budget during the infection reveal that a T4 infection consumes about a third of its host's energy budget, whereas an influenza infection consumes only about 1%.

Building on our estimates for T4, we show how the energetic costs of double-stranded DNA phages scale with the capsid size, revealing that the dominant cost of building a virus can switch from translation to genome replication above a critical size. Lastly, using our predictions for the energetic cost of viruses, we provide estimates for the strengths of selection and genetic drift acting on newly incorporated genetic elements in viral genomes, under conditions of energy limitation. This novel, physical approach to the study of viruses provides a promising path towards a deeper understanding of factors governing viral burst sizes, life-cycle strategies, and evolutionary trajectories.

In our second study we turned to another hallmark feature of viruses, namely their genomic diversity. We developed one of the largest surveys of complete viral genomes, providing a comprehensive and quantitative snapshot of viral genomic trends for thousands of viruses (Chapter 4). We explored the diversity and biases of the NCBI viral database and provided distributions of viral genome length, gene length, gene density, noncoding DNA (or RNA) percentage, and abundances of functional gene categories across thousands of viral genomes. We also created a coarse-grained method for visualizing viral genome organization. Because existing viral classification systems were developed prior to the sequencing era, we present our analysis in light of different classification systems in order to assess the utility of each classification in capturing genomic trends.

In our third study, we shifted our focus towards ecological questions surrounding viruses. Just as they are abundant in the oceans (24), viruses are abundant in the human body. For example, up to 10^8 viruses can be found in just a milliliter of human saliva (5). For

this reason, we chose to specifically target human oral phages to explore their communities as a function of space and time. Considering the lack of universal phage markers, we aimed to discover environment-specific markers so that we could study previously unexplored phage families. As we will demonstrate, this marker-based approach reveals phage community composition with a resolution that cannot be achieved through typical metagenomic studies. We will further demonstrate that at this resolution, phage community composition can serve as a fingerprint, or a "phageprint" – highly unique to each individual and stable over at least 30 days.

By creating sample collection kits and instructional videos, we crowd-sourced sample collection, thereby gathering ~700 samples from individuals living in different parts of the world, including samples from genetically related individuals and couples. To our knowledge, our study is one of the largest studies of human oral phages and the first to demonstrate the feasibility of human identification based on phage sequences. It highlights yet again the astounding viral sequence diversity that underlies the highly complex and personal phage communities.

Whether it is through energetics, genomics, ecology, or any other lens, viruses offer us a unique view of biology, and one that perhaps deserves greater attention. We often look up at the night sky to be amazed and inspired by the possibility of life elsewhere. We have the desire to explore these uncharted territories in part because we want to know whether we, as life forms with a DNA-written history, are unique or simply one of the many possibilities. Yet, here on our own planet and even in our own bodies, there are just as equally amazing and unexplored worlds with strange and unfamiliar inhabitants. In this thesis, we hope to shed some light on these fascinating biological entities that have overrun our planet.

1.1 References

1. Breitbart M & Rohwer F (2005) Here a virus, there a virus, everywhere the same virus? *Trends in microbiology* 13(6):278-284.
2. Suttle CA (2007) Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology* 5(10):801-812.
3. De Paepe M, Leclerc M, Tinsley CR, & Petit M-A (2014) Bacteriophages: an underestimated role in human and animal health? *Frontiers in cellular and infection microbiology* 4:39.
4. Quirós P*, et al.* (2014) Antibiotic resistance genes in the bacteriophage DNA fraction of human fecal samples. *Antimicrobial agents and chemotherapy* 58(1):606-609.
5. Wahida A, Ritter K, & Horz H-P (2016) The Janus-Face of bacteriophages across human body habitats. *PLoS pathogens* 12(6):e1005634.
6. Edlund A, Santiago-Rodriguez TM, Boehm TK, & Pride DT (2015) Bacteriophage and their potential roles in the human oral cavity. *Journal of oral microbiology* 7.
7. Norman JM*, et al.* (2015) Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160(3):447-460.
8. Suttle CA, Chan AM, & Cottrell MT (1990) Infection of phytoplankton by viruses and reduction of primary productivity. *Nature* 347(6292):467-469.
9. Bergh Ø, BØrsheim KY, Bratbak G, & Heldal M (1989) High abundance of viruses found in aquatic environments. *Nature* 340(6233):467-468.
10. Motegi C*, et al.* (2009) Viral control of bacterial growth efficiency in marine pelagic environments. *Limnology and Oceanography* 54(6):1901-1910.
11. Nagata T (2008) Organic matter–bacteria interactions in seawater. *Microbial Ecology of the Oceans, Second Edition*:207-241.

12. Wilhelm SW & Suttle CA (1999) Viruses and nutrient cycles in the sea: viruses play critical roles in the structure and function of aquatic food webs. *Bioscience* 49(10):781-788.

13. Weitz JS & Wilhelm SW (2012) Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 biology reports* 4.

14. Modi SR, Lee HH, Spina CS, & Collins JJ (2013) Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* 499(7457):219-222.

15. Stern A & Sorek R (2011) The phage‐host arms race: shaping the evolution of microbes. *Bioessays* 33(1):43-51.

16. Stern A, Mick E, Tirosh I, Sagy O, & Sorek R (2012) CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome research* 22(10):1985-1994.

17. Youle M, Haynes M, & Rohwer F (2012) Scratching the surface of biology's dark matter. *Viruses: Essential agents of life*, (Springer), pp 61-81.

18. Paez-Espino D*, et al.* (2016) Uncovering Earth's virome. *Nature* 536(7617):425-430.

19. Baltimore D (1970) RNA-dependent DNA polymerase in virions of RNA tumour viruses.

20. Temin HM & Mizutani S (1970) RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* 226(5252):1211-1213.

21. Hershey AD & Chase M (1952) INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE. *The Journal of General Physiology* 36(1):39-56.

22. Luria SE & Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28(6):491.

23. Rittié L & Perbal B (2008) Enzymes used in molecular biology: a useful guide. *Journal of cell communication and signaling* 2(1-2):25-45.

24. Suttle CA (2005) Viruses in the sea. *Nature* 437(7057):356-361.

Chapter II

A Quantitative Framework for Estimating Cellular Energetic Costs and their Evolutionary Consequences

2.1 Introduction

The possible interplay between the energetic cost of a cellular structure and its evolutionary fate is a subject that will become increasingly important as evolutionary cell biology matures as a science. All of biology starts at the level of the cell, which houses a myriad biochemical processes, information storage mechanisms, and physical substructures. To fully understand the mechanisms of evolution, it is ideal to start at the cellular level, where we can tease apart the ways in which complex molecular structures emerge from the assembly or transformation of building blocks such as amino acids, nucleotides, and lipid molecules. Thanks to the advances in biochemistry, much of the information to accomplish this task is either in hand or within reach. As a result, we argue it is time to outline a logical framework for quantifying the relationships between the energetic costs of cellular structures and their susceptibility to establishment and modification by the processes of evolution. Here, the phrase "cellular structure" will be used as an umbrella term broadly referring to any cellular entity built from building blocks, be that entity a few base pairs of DNA, an entire chromosome or an organelle.

At the heart of this subject are two important concepts: 1) all cellular structures have strict energetic costs of construction and maintenance; and 2) the energetic costs and benefits of these cellular structures can translate into fitness differences that influence long-term evolutionary trajectories. Here, we attempt to provide a quantitative framework for addressing

these issues. We first aim to provide a clear depiction of the definitions and subtleties in performing energetic censuses of the cell. We then attempt to examine the evolutionary consequences of cellular structures based on their energetic cost. Finally, we highlight some of the remaining conceptual uncertainties in this area.

2.2 The energetic cost of a cellular structure

We start with the total cost of a simple cellular structure, e.g., a protein molecule or complex, an information-bearing molecule at the DNA or RNA level, or a membrane. The basic principle here is that all cellular structures, regardless of their fitness costs or benefits, entail some baseline energetic costs of construction and maintenance. In accounting for cellular energetic costs, we could report the costs using several different units, but the underlying premise is that the hydrolysis of ATP and ATP-equivalent molecules serves as the universal currency of bioenergetics across the different domains of life (1, 2). We could, for example, report energetic costs in units of Joules. Under physiological conditions, ATP hydrolysis and conversion into adenosine diphosphate (ADP) and orthophosphate (P_i) results in about -50 kJ/mol free energy change (1). However, the actual change in free energy depends on the exact concentrations of reactants and products. It is usually much more convenient to enumerate energetic costs in units of numbers of ATP hydrolyses (or their equivalent), which is also in keeping with previous efforts (3-8). To remain consistent, we will use the symbol P with different subscripts as a shorthand notation to represent an ATP (or an ATP-equivalent) hydrolysis event in the context of different energetic cost definitions (8-10). As will be discussed below, even with this seemingly straightforward approach, there still remain critical subtleties.

Cellular structures are assembled from molecular building blocks such as amino acids, nucleotides, lipids, and carbohydrates. If not provided by the outside environment, these building blocks must be synthesized within the cell by processes requiring carbon skeletons and the expenditure of energy. In fact, in the context of the metabolism of many bacteria, all building blocks as well as coenzymes and prosthetic groups can be synthesized from a small number of precursor metabolites (11). If some building blocks are available externally, the biosynthetic costs will be diminished, but there will still be costs of transformation to arrive at the full set of internal building blocks (the cost of converting one amino acid to another, for example). Here, we will assume that all molecular building blocks are derived from one carbon source, namely glucose, and further assume that sources of inorganic nitrogen and other trace elements are provided in excess within the growth media. These assumptions are especially applicable to growth conditions in the laboratory.

Moreover, the assembly cost of a cellular structure is obtained by adding up the requirements for construction of that structure from its molecular building blocks, e.g., the necessities for polymerizing a protein from its constituent amino acids, adding post-translational modifications, and folding the subsequent chain into the appropriate globular form. Finally, there will often be maintenance costs, e.g., accommodation of molecular turnover, and identification and elimination of cumulative errors.

The sum of costs noted above represents the baseline investment that must be made in a cellular structure regardless of its benefit to the host cell (Figure 1). Given the near universality of many biosynthetic pathways and enzyme-reaction mechanisms, the assembly and maintenance costs can generally be calculated from information in the literature. The ability to make such calculations is a highly desirable complementary approach to laborious experimental approaches (12, 13), such as modifications of gene-expression levels, as these can

have additional side effects (e.g., promiscuous binding or aggregation) that are difficult to quantify and irrelevant to construction/maintenance costs. What has been summarized in the paragraphs above, however, are the **direct costs** of a cellular structure, which do not fully describe the energetic consequences for the cell. We use P_D to symbolize the unit of direct cost.

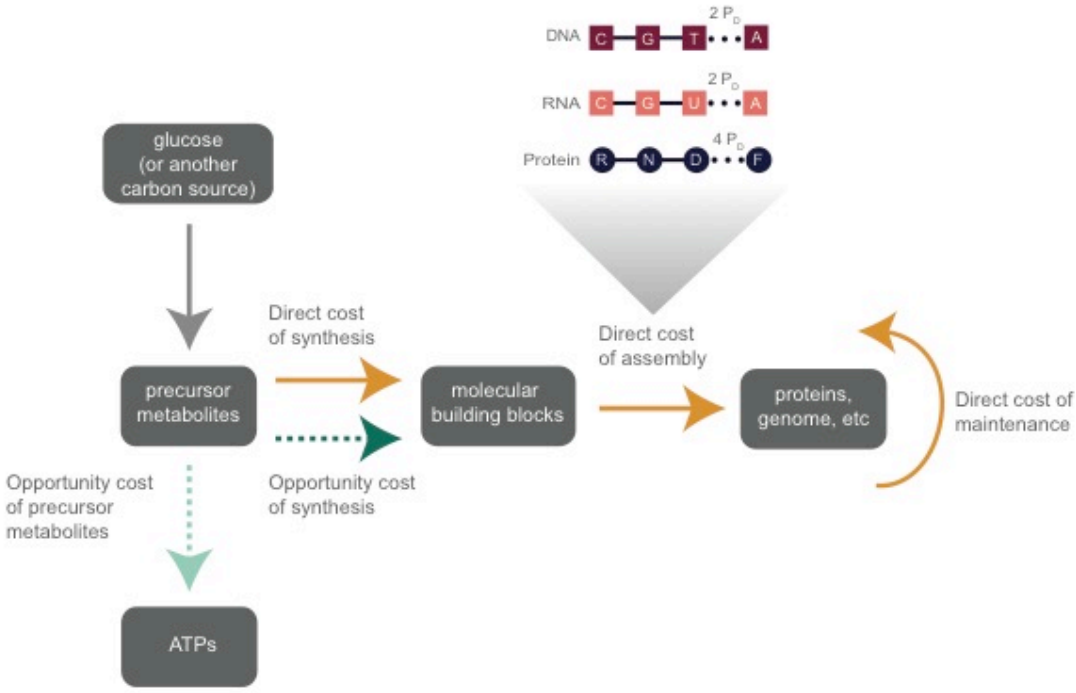


Figure 1. The distinction between direct and opportunity costs associated with synthesizing molecular building blocks. As glucose is partially metabolized into precursor metabolites, the energy that could have been captured from the complete metabolism of glucose is referred to as the opportunity cost of precursor metabolites (light green arrow). As precursor metabolites get converted to molecular building blocks, the conversion consumes electron carrier molecules such as NADH. If not used during the synthesis of molecular building blocks, these electron carriers would result in the generation of ATP, and thus the conversion of precursor metabolites to building blocks incurs additional opportunity cost (dark green arrow). The conversion also consumes ATP, which we count as the direct cost of synthesis. The assembly

of macromolecules such as proteins from building blocks requires additional post-synthesis costs such as the cost of polymerization and maintenance. The polymerization costs per nucleotide or amino acid are denoted. Direct costs are shown by solid orange arrows, whereas the opportunity costs are denoted in shades of green and dotted lines. The same color scheme is used in Figure 2.

The construction and maintenance of a cellular structure represents a drain on resources that could otherwise be allocated to other cellular functions. When metabolic precursors that can be fully metabolized for ATP production are instead allocated as carbon skeletons to the production/maintenance of a particular cellular structure, this diversion eliminates their availability for other purposes, a consequence that we refer to as the **opportunity cost.** We use P_O to symbolize the unit of opportunity cost.

Opportunity costs can also be calculated from basic cell-biological knowledge though a rigorous definition such as the one provided in the Supplementary Information does not seem to have been previously provided (4, 9, 14). Specifically, we estimate the opportunity cost of a precursor metabolite as the number of ATPs (or ATP equivalents) that could have been generated had the precursor metabolite not been diverted towards the synthesis of molecular building blocks (Figure 2A, SI). Figure 2B demonstrates the placement of metabolic precursors that are implicated in the synthesis of molecular building blocks across metabolic pathways. In arriving at cost estimates for molecular building blocks, we assumed glucose as the primary carbon source, so these cost estimates may need to be modified when considering another carbon source as input. The estimated opportunity costs for each precursor metabolite in the context of bacterial and eukaryotic metabolism are shown in Figure 2C. The difference in these costs between eukaryotic and bacterial metabolism stems from the higher efficiency of eukaryotic metabolism in producing ≈6 more ATPs per glucose than bacterial metabolism,

which generates about 26 ATPs per glucose (15, 16) (see SI). We hope that future studies will also reveal these costs in the context of the archaeal metabolism.

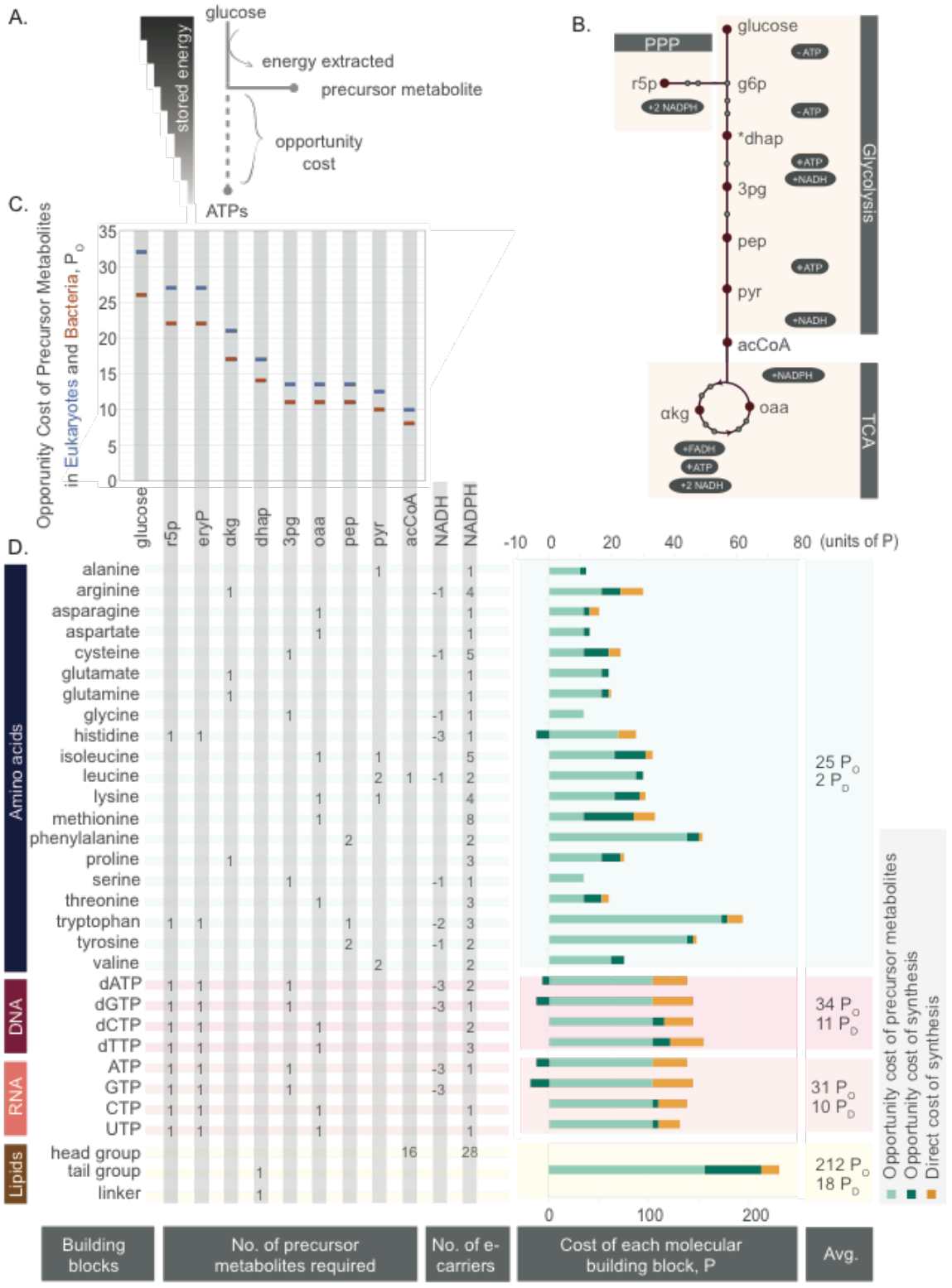


Figure 2. The energetic cost of molecular building blocks. A) The concept of opportunity cost is shown schematically for the situation in which glucose is

the sole carbon source. B) The metabolic pathways from which the opportunity cost of each precursor metabolite can be estimated. From dihydroxyacetone phosphate (dhap) onwards, there are two molecules of each precursor metabolite generated. Precursor metabolites that are not implicated in the synthesis of building blocks are shown as grey circles. Positive and negative signs indicate gains and losses in ATPs or electron carrier molecules from the conversion of one precursor metabolite to another. The names for each precursor metabolite denoted are as follows: ribose-5-phosphate (r5p), erythrose-4-phosphate (e4p, or eryP), alpha-ketoglutarate (αkg), dihydroxyacetone phosphate (dhap), 3-phosphoglycerate (3pg), oxaloacetate (oaa), phosphoenolpyruvate (pep), pyruvate (pyr), acetyl-CoA (acCoA). C) The opportunity cost of each precursor metabolite estimated in the context of both heterotrophic bacteria and eukaryotic cells as detailed in the SI. D) The opportunity cost of a molecular building block is equivalent to the number of precursor metabolites and electron carrier molecules used during its synthesis times their respective opportunity costs. The direct cost of synthesizing a molecular building block is the number of ATP (or ATP-equivalent) hydrolysis events required during the synthesis of each building block (orange). All costs shown are estimated in the context of a heterotrophic bacterial metabolism. The average direct cost provided for building blocks does not include post-synthesis costs such as polymerization or maintenance. See SI Dataset 1 for further details.

Another source of opportunity cost is the pool of electron carrier molecules used in the synthesis of molecular building blocks. If electron carriers such as NAD(P)H were to be preserved rather than used in the synthesis of molecular building blocks, they would result in 2 and 2.5 ATP molecules within the bacterial and eukaryotic electron transport chain, respectively (16). To distinguish between these two sources of opportunity cost, we have denoted them in two shades of green in both Figures 1 and 2. However, unless denoted

otherwise, we will refer to the opportunity cost of a molecular building block as the sum of all opportunity costs.

Moreover, the **total cost** of producing a trait and diverting structures from alternative usage to do so is then the sum of the direct and opportunity costs,

$$c_T = c_D + c_O, \quad [1]$$

where all costs represent the cumulative expenditures over the entire lifespan of the cell. Whereas c_D is expected to reflect ATP (and ATP-equivalent) hydrolysis reactions resulting in heat dissipation in the cell, c_O will not be manifested in heat production, given that the ATP is not actually produced or consumed. We have used the symbol P_T to denote the number of ATP formation/hydrolysis reactions associated with the total cost definition (Figure 2).

2.3 Which is more biologically relevant, the direct or the total cost?

Depending on the experimental context or question at hand, one of the two definitions can be more appropriate than the other. Under the direct cost definition, we simply account for the number of ATP hydrolysis events, and ignore the effects of diverting molecular building blocks from energy-producing pathways. This cost definition is useful when we are comparing direct cost estimates to calorimetric studies or attempting to estimate heat production and power generation of a cellular process. It is also of interest in scenarios where there is a constrained rate of ATP production, e.g. because of a limited amount of membrane real-estate for the respiration machinery (17) or a limitation on the amount of glycolytic enzymes. Most of the pioneering studies in bioenergetics have reported the costs of molecular building blocks in terms of their direct costs (11). More recent works

on cellular energetics, however, have implicitly adhered to the total cost definition, thereby including the opportunity costs of molecular building blocks (4, 9, 10, 14, 18).

The total cost definition is useful when the overall carbon source availability is limiting. It also offers the advantage of comparison with results from chemostat experiments as a way of determining the cost of a particular biological structure relative to the collective costs of a cell at a given growth rate. During a chemostat experiment, the number of glucose molecules required per unit time to grow cells at a set growth rate is measured. This number can then be converted to an energetic value by assuming that every glucose molecule consumed by the cell is fully metabolized. However, not all glucose molecules are used for energy production. In fact, as we will briefly discuss in the following section, depending on the growth rate, the majority of glucose molecules consumed may be converted to biomass. As a result, the cost of a cell as measured through chemostat experiments includes both the direct and the opportunity costs of cellular processes, and is inherently a total cost estimate. Thus, when considering the energetic burden of a given process on a cell's energy budget as measured by chemostat experiments, the total cost definition provides a more meaningful approach.

2.4 What fraction of a cell's total cost is direct cost?

One approach to estimating the total cost of a bacterium is to consider its mass in carbon. With a doubling time of 30 minutes, an *E. coli* cell growing on glucose as its sole carbon source will be ≈50% carbon in dry weight which amounts to 0.1 pg (5×10^9 carbon molecules) for a cell with an approximate volume of 1 μm^3 (19). Considering that every glucose molecule contributes 6 carbons, we can deduce that this bacterium's energy and carbon demand is met by $\approx 10^9$ glucose molecules. Because every glucose molecule can generate about 26 ATPs during aerobic respiration in *E. coli*, we can convert the number of

glucose molecules to a total energetic cost of ≈3 x 10^{10} P_T. This estimate for the total cost of *E. coli* is similar to those obtained from growth experiments in chemostats (10).

To obtain the direct cost estimate of a bacterial cell, we can make use of the fact that translation incurs the greatest cost for cells (5, 9, 10, 14). An *E. coli* cell with a volume of 1 μm^3 (the size can range between ≈0.5 and 2 μm^3 depending on growth rate) contains on the order of three million proteins (20). With an average protein length of 300 amino acids (21), an *E. coli* cell of this size will be comprised of 10^9 amino acids. The direct costs of polymerization (4 P_D) and synthesis from precursor metabolites (2 P_D) (Figure 2) are 6 ATPs per amino acid, and thus the direct cost of translation is about ≈6 x 10^9 P_D. We can compare this cost to the direct cost of genome replication, ≈ 10^8 P_D, for an *E. coli* genome that is comprised of 5 x 10^6 base pairs. This is because the direct cost of polymerization (2 P_D) and synthesis of nucleotides from precursor metabolites (11 P_D) amounts to 13 P_D per nucleotide. As such, the direct cost of translation far outweighs the direct cost of genome replication. The cost of transcription in both cellular and viral contexts has also been shown to be lower than the cost of translation (8, 10). Thus, we will take the direct cost of translation as a proxy for the sum of all direct costs. With that assumption, the ratio of direct cost of an *E. coli* cell (≈6 x 10^9 P_D) to its total cost (≈3 x 10^{10} P_T) is 0.2. This estimate suggests that the majority of glucose molecules consumed by an *E. coli* cell during rapid growth are not fully metabolized and instead are used to synthesize biomass.

2.5 Energy as one of several possible limiting factors to growth

Cellular fitness need not always be strongly limited by energy. For example, photosynthetic plankton populations often experience an overabundance of energy relative to some nutrient such as nitrogen or phosphorus. For microbes growing in laboratory

conditions on a defined medium with a single compound providing carbon and energy, the growth yield per carbon consumed increases linearly with the substrate heat of combustion (which is inversely related to the degree of oxidation) up until a threshold value, thereafter leveling off (Figure 3). This suggests that below a critical substrate value of ≈10 kcal/g carbon, growth in such conditions is limited by energy, whereas above this threshold the food supply contains excess energy relative to carbon content required for growth. Notably, the most common substrate used in growth experiments with microbes, glucose, has a heat of combustion of 9.3 kcal/g carbon, close to the threshold at which growth is equally limited by carbon and energy. Very few commonly used substrates have heats of combustion much beyond the apparent threshold (values being 11.0, 13.6, and 14.8 kcal/g carbon for glycerol, ethanol, and methanol, respectively).

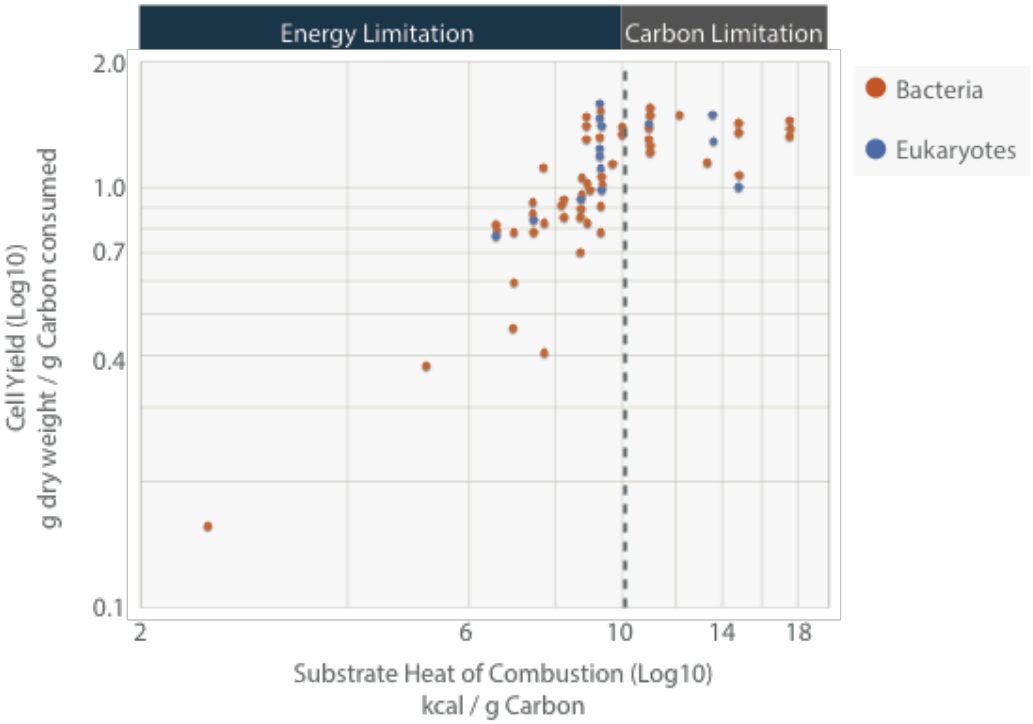


Figure 3. Growth yields per unit carbon consumed as a function of the heats of combustion of the carbon substrate (which also serves as the source of energy). Results are taken from various sources in the literature; all results for eukaryotes involve fungi (mostly yeasts).

Although analyses involving alternative limiting factors will not be pursued here, where deemed necessary, they could be implemented in a parallel fashion by altering the budgetary currency (e.g., to carbon, nitrogen or some other limiting nutrient). Consider, for example, an organism living in an environment plentiful in a carbon/energy source but limited by a micronutrient such as iron. In this case, depending on the internal regulatory structures of the cell, energy extracted from the food source may be in excess supply, resulting in under-utilization of ingested carbon and energy. In such a case one can choose to adopt a framework as described here but define costs in terms of the iron atoms needed.

The energetic costs can still be evaluated as given here but their effect on growth will be diminished due to the greater impact of iron shortage on growth rate.

Regardless of the substrates being consumed, all aspects of cellular maintenance and growth require energy. Although the yields of microbes (per unit carbon consumed) grown on alternative substrates vary substantially with the nature of the substrate, the direct costs necessary to build an offspring cell are relatively constant. This is consistent with the positive scaling in Figure 3 (except in the above mentioned case of highly reduced substrates such as some alcohols and methane), and provides at least partial justification for using ATP as a universal currency as opposed to carbon or some other micronutrient.

2.6 The evolutionary implications of the energetic cost of a cellular structure

Given a value for the total energetic cost of a cellular structure (or modification thereof), what are the overall evolutionary implications? To answer that question we need to consider two central issues: 1) what is the scaling of the cost of a cellular structure relative to the total cost of building and maintaining the cell? and 2) how do we convert the appropriately scaled measure of this energetic cost to a corresponding change in fitness. Supposing the cell has a baseline total energy budget per cell cycle of C_T (which includes the costs of both growth and maintenance, with a capital C denoting a whole-cell cost), the addition of an energetic burden to the trait under consideration alters the lifetime energy budget to $C_T' = C_T + c_T$. Under the assumption that energy availability influences fitness, if the cell-division time is τ in the absence of the trait, then we qualitatively expect this additional energetic investment in the trait to alter the cell-division time to $\tau' > \tau$.

To understand the total energetic investment in a cellular structure from a fitness perspective, we need to define its effects on the cell's reproductive rate relative to that for a cell

without such an additional energetic investment. From standard haploid selection theory, the selective disadvantage of building a cellular structure in an energy-limited environment is defined as

$$s = \frac{\mu' - \mu}{\mu} \ln 2, \qquad [2a]$$

where $\mu = \ln(2)/\tau$ denotes an exponential rate of growth. Assuming that $\Delta\tau = (\tau' - \tau) \ll \tau$ which we suspect will be valid for most single-gene modifications, we can simplify the selective disadvantage to

$$s \simeq \frac{\Delta\tau}{\tau} \ln 2. \qquad [2b]$$

If we assume that energy is the only limiting factor and further assume that $c_T \ll C_T$ so that the increment in cell-division time scales proportionally with the increased investment, we can write

$$\tau' \simeq \tau \left(1 + \frac{c_T}{C_T}\right). \qquad [3]$$

This then leads to the simple result that

$$\mathrm{s} \simeq \frac{c_T}{C_T} \ln 2, \qquad [4]$$

showing that the intrinsic selective disadvantage associated with the energetic cost of a trait under energy limited conditions scales directly with the proportional increase in the total energy demand per cell cycle.

It is important to note that there are some caveats with respect to the preceding derivation. First, it is assumed that the addition of the trait does not somehow alter the cell's basic metabolic makeup in ways that would modify the total baseline energy budget C_T. Even if this does occur, the result given in Equation 4 will be only slightly modified if the fractional

alteration to C_T is small, which seems likely for cellular modifications involving just one or two genes. Second, we have focused entirely on the bioenergetic costs of producing a trait. There may be additional costly side effects, if for example a novel protein promiscuously interacts with inappropriate substrates, aggregates with other cellular structures, and/or excessively occupies cellular volume or membrane real estate.

2.7 Supplementary Information

2.7.A Opportunity cost of precursor metabolites in heterotrophic bacteria

In this section we will use the units of P, without any subscripts, to refer to ATP (and ATP-equivalent) hydrolysis events (using the subscripts without having first derived the opportunity cost of a metabolite is meaningless). Once we have derived the opportunity cost of a precursor metabolite here in this section, we will accompany it with the symbol P_O in later figures and tables to clearly mark these costs as opportunity costs. We will estimate the opportunity cost of precursor metabolites, $C_{opportunity}$, by

$$C_{opportunity} = E_{glu} - E_{partial}, \quad [1]$$

where E_{glu} represents the net energetic gain from the complete metabolism of a glucose molecule into water and carbon dioxide, and $E_{partial}$ is the net energetic gain from the partial metabolism of a glucose molecule into a precursor metabolite. Under aerobic respiration, E_{glu} is ≈ 26 P in *E. coli* (15). In the event that there is a net energetic *cost* from the conversion of glucose into a precursor metabolite, $E_{partial}$ will be a negative value.

In the synthesis of lipids, dihydroxyacetone-phosphate (dhap) is used as a precursor. This precursor is generated during glycolysis (Figure 2). Each glucose molecule results in the production of two dihydroxyacetone-phosphate molecules, with this process having a net

energetic cost of 2 P ($E_{partial} = -2$). The opportunity cost of two dihydroxyacetone-phosphate molecules is 2 P greater than that of glucose, or 28 P (SI Eq. 1). The opportunity cost of one dihydroxyacetone-phosphate molecule is therefore ≈ 14 P (Figure 2).

To simplify the opportunity cost estimates further, we could employ a shortcut. Rather than estimating the opportunity cost of each precursor metabolite by calculating the $E_{partial}$ from glucose as the starting point, we could obtain the opportunity cost of metabolite *j*, $C^{j}_{opportunity}$, from the opportunity cost of metabolite *i*, $C^{i}_{opportunity}$, by

$$C^{j}_{opportunity} = C^{i}_{opportunity} - E^{i \to j}_{partial}, \quad [2]$$

where $E^{i \to j}_{partial}$ represents the net energetic gain from the conversion of metabolite *i* to metabolite *j*.

For example, during glycolysis each molecule of dihydroxyacetone-phosphate is converted to a molecule of 3-phosphoglycerate (3pg), resulting in the production of 1 NADH molecule and 1 ATP. In *E. coli*, each NADH molecule results in the production of ≈ 2 ATP molecules under aerobic conditions (11); therefore the net energetic gain from this conversion is ≈ 3 P. Hence, the opportunity cost of each 3-phosphoglycerate molecule would be ≈ 11 P, which is 3 P less than the opportunity cost of a dihydroxyacetone-phosphate molecule (SI Eq. 2).

If not used as a precursor, 3-phosphoglycerate is converted to phosphoenolpyruvate (pep) in glycolysis (Figure 2). In this process, however, there is zero energy expenditure or gain. As such, the opportunity cost of a phosphoenolpyruvate molecule is the same as that of a 3-phosphoglycerate molecule's, which is ≈ 11 P. Both of these precursors come before pyruvate (pyr) in glycolysis. In converting a phosphoenolpyruvate molecule into a pyruvate molecule, there is a net energy gain of 1 P. The opportunity cost of a pyruvate molecule is

therefore ≈ 10 P (1 P less than a phosphoenolpyruvate opportunity cost) (Figure 2). Pyruvate can be converted to oxaloacetate (oaa) with the expenditure of 1 ATP. The opportunity cost of oxaloacetate is ≈ 11 P (10 P + 1 P) (Figure 2).

Pyruvate is further converted to acetyl-CoA (acCoA), and in the process one molecule of NADH is generated, which is equivalent to 2 P (Figure 2). The opportunity cost of acetyl-CoA is therefore ≈ 8 P (10 P - 2 P) (Figure 2). One molecule of acetyl-CoA and one molecule of oxaloacetate are then eventually converted to alpha-ketoglutarate (αkg) (Figure 2). The sum of the opportunity costs of acetyl-CoA (8 P) and oxaloacetate (11 P) is 19 P, and because 1 molecule of NADH is generated in their conversion to alpha-ketoglutarate, the opportunity cost of alpha-ketoglutarate is ≈ 17 P (or, 19 P – 2 P) (Figure 2). Similarly, alpha-ketoglutarate is eventually converted to oxaloacetate, and 2 NADH, 1 GTP, and 1 $FADH_2$ molecules are generated (Figure 2). This is a net gain of ≈ 6 (assuming 1 P from each $FADH_2$), reducing the opportunity cost of oxaloacetate to ≈ 11 P (17 P – 6 P). Note, this is consistent with the opportunity cost of oxaloacetate derived from the anaplerotic pathway described earlier (the conversion of pyruvate to oxaloacetate via the pyruvate decarboxylase enzyme).

Glucose can also be converted to ribose-5-phosphate (r5p) in the pentose phosphate pathway, and in the process 2 NADPH molecules are generated, which is equivalent to 4 P (Figure 2). We subtract 4 P from the possible 26 P that glucose would be converted to under respiratory conditions, and we arrive at 22 P as the opportunity cost of ribose-5-phosphate (SI Eq. 1) (Figure 2). The same calculation can be used for erythrose-4-phosphate (e4p), resulting in 22 P as its opportunity cost (Figure 2).

2.7.B Opportunity cost of precursor metabolites in heterotrophic eukaryotes

To estimate the opportunity costs of precursor metabolites in heterotrophic eukaryotes, we can carry out very similar calculations to those performed for heterotrophic bacteria. For eukaryotes, E_{glu}, or the total energetic gain from the complete metabolism of a glucose molecule into carbon dioxide and water is higher. This is because each NAD(P)H and $FADH_2$ molecule results in a higher number of ATPs within the mitochondrial electron transport chain compared to the bacterial electron transport chain. Specifically, each NAD(P)H molecule is equivalent to ≈ 2.5 P and each $FADH_2$ molecule corresponds to ≈ 1.5 P ((16),(22) pages 517-518), resulting in 30-32 P per glucose molecule. Note, the theoretical yield of 38 P per glucose molecule has been shown to be an overestimate due to the outdated assumptions that each NAD(P)H molecule is equivalent to 3 P and that each $FADH_2$ molecule generates 2 P (16). We will therefore use 32 P as E_{glu}.

Each glucose molecule results in the production of two dihydroxyacetone-phosphate molecules, with this process having a net energetic cost of 2 P. The opportunity cost of two dihydroxyacetone-phosphate molecules is 2 P greater than that of glucose, or 34 P (SI Eq. 1). The opportunity cost of one dihydroxyacetone-phosphate molecule is therefore 17 P (Figure 2). As described earlier, in the conversion of dihydroxyacetone-phosphate molecule into a 3-phosphoglycerate molecule, 1 NADH and 1 ATP molecules are produced. This is equivalent to a net energetic gain of ≈ 3.5 P. The opportunity cost of a 3-phosphoglycerate molecule is therefore ≈ 13.5 P (17 P – 3.5 P) (SI Eq. 2) (Figure 2).

The opportunity cost of a phosphoenolpyruvate is the same as that of a 3-phosphoglycerate, which is ≈ 13.5 P. In converting phosphoenolpyruvate into pyruvate, there is a net energy gain of 1 P (1 ATP molecule is formed). The opportunity cost of pyruvate is

therefore ≈ 12.5 P. Pyruvate can be converted to oxaloacetate with the expenditure of 1 ATP. Hence, the opportunity cost of oxaloacetate will be ≈ 13.5 P.

Pyruvate is converted to acetyl-CoA, and in the process one molecule of NADH is generated. As a result, the opportunity cost of acetyl-CoA is ≈ 10 P (12.5 P – 2.5 P). One molecule of acetyl-CoA and one molecule of oxaloacetate are converted to alpha-ketoglutarate in the TCA cycle. The sum of the opportunity costs of acetyl-CoA (10 P) and oxaloacetate (13.5 P) is 23.5 P, and because 1 molecule of NADH is generated in their conversion to alpha-ketoglutarate, the opportunity cost of alpha-ketoglutarate is ≈ 21 P (23.5 P – 2.5 P). Alpha-ketoglutarate is eventually converted to oxaloacetate, and 2 NADH, 1 GTP, and 1 $FADH_2$ molecules are generated. This is a net gain of ≈ 7.5 P (assuming 1.5 P from each $FADH_2$), reducing the opportunity cost of oxaloacetate to ≈ 13.5 P (21 P – 7.5 P). Note, this is again consistent with the opportunity cost of oxaloacetate derived from the anaplerotic pathway.

Glucose can also be converted to ribose-5-phosphate in the pentose phosphate pathway, and in the process 2 NADPH molecules are generated, which is equivalent to 5 P. The opportunity cost of ribose-5-phosphate is thus ≈ 27 P. The same calculation can be used for erythrose-4-phosphate, resulting in ≈ 27 P as its opportunity cost.

SI Dataset 1. A detailed breakdown of the energetic cost of molecular building blocks depicted in Figure 2. Provided in a GitHub repository: https://github.com/gitamahm/thesis

2.8 References

1. Berg JM, Tymoczko JL, & Stryer L (2012) *Biochemistry* 7th Ed.
2. Voet D & Voet JG (2011) Biochemistry, 4-th Edition. *New York: John Wiley& Sons Inc*:492-496.
3. Bauchop T & Elsden SR (1960) The growth of micro-organisms in relation to their energy supply. *Microbiology* 23(3):457-469.
4. Atkinson DE (1971) Adenine nucleotides as universal stoichiometric metabolic coupling agents. *Advances in enzyme regulation* 9:207-219.
5. Stouthamer AH (1973) A theoretical study on the amount of ATP required for synthesis of microbial cell material. *Antonie van Leeuwenhoek* 39(1):545-565.
6. Tempest DW & Neijssel OM (1984) The status of YATP and maintenance energy as biologically interpretable phenomena. *Annual Reviews in Microbiology* 38(1):459-513.
7. Russell JB & Cook GM (1995) Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiological reviews* 59(1):48-62.
8. Mahmoudabadi G, Milo R, & Phillips R (2017) Energetic cost of building a virus. *Proceedings of the National Academy of Sciences* 114(22):E4324-E4333.
9. Akashi H & Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proceedings of the National Academy of Sciences* 99(6):3695-3700.
10. Lynch M & Marinov GK (2015) The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences* 112(51):15690-15695.
11. Neidhardt FC, Ingraham JL, & Schaechter M (1990) Physiology of the bacterial cell: a molecular approach.

12. Geiler-Samerotte KA*, et al.* (2013) Quantifying condition-dependent intracellular protein levels enables high-precision fitness estimates. *PloS one* 8(9):e75320.

13. Dekel E & Alon U (2005) Optimality and evolutionary tuning of the expression level of a protein. *Nature* 436(7050):588-592.

14. Wagner A (2005) Energy constraints on the evolution of gene expression. *Molecular biology and evolution* 22(6):1365-1374.

15. Kaleta C, Schäuble S, Rinas U, & Schuster S (2013) Metabolic costs of amino acid and protein production in Escherichia coli. *Biotechnology journal* 8(9):1105-1114.

16. Silverstein T (2005) The mitochondrial phosphate-to-oxygen ratio is not an integer. *Biochemistry and Molecular Biology Education* 33(6):416-417.

17. Andersen KB & von Meyenburg K (1980) Are growth rates of Escherichia coli in batch cultures limited by respiration? *Journal of Bacteriology* 144(1):114-123.

18. Craig CL & Weber RS (1998) Selection costs of amino acid substitutions in ColE 1 and colIa gene clusters harbored by Escherichia coli. *Molecular biology and evolution* 15(6):774-776.

19. Milo R, Jorgensen P, Moran U, Weber G, & Springer M (2010) BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic acids research* 38(suppl 1):D750-D753.

20. Milo R (2013) What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays* 35(12):1050-1055.

21. Milo R & Phillips R (2015) *Cell biology by the numbers* (Garland Science).

22. Berg JM, Tymoczko JL, & Stryer L (2002) Biochemistry. 5th. (New York: WH Freeman).

Chapter III

Energetic Cost of Building a Virus

3.1 Introduction

Viruses are biological 'entities' at the boundary of life. Without cells to infect, viruses as we know them would cease to function, as they rely on their hosts to replicate. Though the extent of this reliance varies for different viruses, all viruses consume from the host's energy budget in creating the next generation of viruses. There are many examples of viruses that actively subvert the host transcriptional and translational processes in favor of their own replication (1). This viral takeover of the host metabolism manifests itself in a variety of forms such as in the degradation of the host's genome or the inhibition of the host's mRNA translation (1). There are many other experimental studies (discussed in the SI) (2-6) that demonstrate viruses to be capable of rewiring the host metabolism. These examples also suggest that a viral infection requires a considerable amount of the host's energetic supply. In support of this view are experiments on T4 (7), T7 (8), Pseudoalteromonas phage (9), and Paramecium bursaria chlorella virus-1 or PBCV-1 (10), demonstrating that the viral burst size correlates positively with the host growth rate. In the case of PBCV-1, the burst size is reduced by 50% when its photosynthetic host, a freshwater algae, is grown in the dark (10). Similarly, slow growing E. coli with a doubling time of 21 hours affords a T4 burst size of just one phage (11), as opposed to a burst size of 100-200 phages during optimal growth conditions.

These fascinating observations led us to ask the following questions: what is the energetic cost of a viral infection, and what is the energetic burden of a viral infection on the

host cell? To our knowledge, the first attempt to address these problems is provided through a kinetic model of the growth of Qß phage (12). A more recent study performed numerical simulations of the impact of a phage T7 infection on its *E. coli* host, yielding important insights into the time course of the metabolic demands of a viral infection (13).

To further explore the energetic requirements of viral synthesis, we made careful estimates of the energetic costs for two viruses with very different characteristics, namely the T4 phage and the influenza A virus. T4 phage is a double-stranded DNA (dsDNA) virus with a 169 kb genome that infects *E. coli*. The influenza virus is a negative-sense, single-stranded RNA virus (-ssRNA) with a segmented genome that is 10.6 kb in total length. The influenza virus is a eukaryotic virus infecting various animals, with an average burst size of 6000, though note that the burst size depends upon growth conditions (14). Similar to many other dsDNA viruses, T4 phage infections yield a relatively modest burst size, with the majority of T4 phages resulting in a burst size of approximately 200 during optimal host growth conditions (15). To determine the energetic demand of viruses on their hosts, the cost estimate for building a single virus has to be multiplied by the viral burst size and placed in the context of the host's energy budget during the viral infection.

Concretely, the costs associated with building a virus can be broken down into the following processes that are common to the life-cycles of many viruses: 1) viral entry 2) intracellular transport, 3) genome replication, 4) transcription, 5) translation, 6) assembly and genome packaging, and 7) exit. Our strategy was to examine each of these processes for both viruses in parallel, comparing and contrasting the energetic burdens of each of the steps in the viral life-cycle.

3.2 Energetic cost units and definitions

Given that the energetic processes of the cell take place in many different energy currencies ranging from ATP and GTP hydrolysis to the energy stored in membrane potentials, it is important to have a consistent scheme for reporting those energies. Adenosine triphosphate (ATP) serves as the most common energy currency of the cell, a function that is universally conserved across all known cellular life-forms (16, 17). Under physiological conditions, the hydrolysis of ATP usually releases about -50 kJ/mol (16, 18). In addition to ATP, which most commonly serves as the energy currency of the cell, there are other nucleoside triphosphates such as GTP that are approximately energetically equivalent to ATP. We will refer to these molecules as ATP-equivalent.

We follow others in their reporting of cellular costs by using the number of ATP (and ATP-equivalent) hydrolysis events as a proxy for energetic cost. Similar to Lynch and Marinov (19), we will use the symbol P as a shorthand notation to represent an ATP (or an ATP-equivalent) hydrolysis event. We will additionally employ subscripts to clearly label the results obtained under different energetic cost definitions, which have already been introduced in Chapter II (e.g. opportunity cost denoted by P_O, direct cost denoted by P_D, and total cost denoted by P_T). Lastly, in reporting some of our final cost estimates we will convert the number of ATP (or ATP-equivalent) hydrolysis events to units of Joules and k_BT by assuming 50kJ of negative free energy change per mole of P at physiological conditions.

As a reminder of topics discussed in Chapter 2, the distinction between total and direct cost definitions is that under the total cost definition, in addition to accounting for direct costs, we attribute an energetic cost to the building blocks that are usurped from the host during viral synthesis. Both energetic cost definitions have physical significance. For

example, the direct cost definition is the more appropriate choice when estimating heat production and power consumption of a viral infection (SI sections B, K). The total cost definition, on the other hand, is aligned with traditional energetic cost estimates made from growth experiments in chemostats, where substrate consumption and cell yield is monitored, and allows for a clear comparison between the cost of an infection and the cost of a cell. This is because the cost of a cell, derived through chemostat experiments, implicitly includes the opportunity cost component, which is the cost of diverting precursor metabolites from energy-producing pathways towards the synthesis of molecular building blocks. And while our approach would certainly benefit from detailed experimental studies that reveal the fluxes in the host metabolome during an infection, the assignment of an energetic value to each metabolite allows us to simplify the problem from reporting changes in the concentration of hundreds of metabolites to reporting a single energetic value associated with the viral infection. This value can then be compared to the cellular energy budget.

We will generally estimate the cost of a certain viral process for a single virus, and then multiply this cost by the viral burst size to determine the infection cost of a given process. Subscript v will denote the cost estimates made for a single virus, and the subscript i will refer to a cost estimate made for an infection. We relegate the energetic cost estimates for all viral processes to the SI sections C through I.

3.3 The energetic costs of T4 and Influenza

By estimating the energetic costs of influenza and T4 life-cycles, we show that surprisingly the cost of synthesizing an influenza virus and a T4 phage are nearly the same (Table 1). The outcome of the analysis to be discussed in the remainder of the chapter is summarized pictorially in Figure 1 for bacteriophage T4 and Figure 2 for influenza. For both

viruses, the energetic cost of translation outweighs other costs (Table 1, Figures 1, 2, 3), though as we will show at the end of the chapter, since translation scales with the surface area of the viral capsid and replication scales as the volume of the virus, for double-stranded DNA phages larger than a critical size, the replication cost outpaces the translation cost.

To get a sense for the numbers, here we provide order-of-magnitude estimates of both the costs of translation and replication and refer the interested reader to the SI sections C through I for full details. As detailed in the SI Tables 1 and 2, both T4 and influenza are comprised of about 10^6 amino acids. We can estimate the total cost of translation by appealing to a few simple facts. First, the average opportunity cost per amino acid is about 30 P_O. Second, the average direct cost to produce amino acids from precursor metabolites is 2 P_D per amino acid (SI Figure 1). Finally, each polypeptide bond incurs a direct cost of 4 P_D. We can see that the total cost of an amino acid is approximately 36 P_T (30 P_O + 6 P_D). As a result, the translational cost of an influenza virus and a T4 phage both fall between 10^7 to 10^8 P_T (Table 1).

Table 1. The direct, opportunity and total energetic costs of viral processes for T4 and influenza. The T4 infection costs are estimated based on an average burst size of 200, and the influenza infection costs are based on an average burst size of 6000. Direct costs shown represent the number of phosphate bonds directly hydrolyzed during the viral lifecycle (P_D), whereas the total costs (P_T) include both direct costs (P_D) as well as opportunity costs (P_O) incurred during the viral life-cycle (See SI sections C through I). Empty entries correspond to viral processes that did not result in an energetic cost or were not applicable to the given virus. Note, to obtain the total cost estimates, the sum of opportunity and direct costs used exact numbers and was *then* rounded (this is why the sum of the rounded versions of direct and

opportunity costs do not exactly match up to the total costs presented in this table).

			Replication	Transcription	Viral Entry	Packaging	Intracellular Transport	Viral Exit	Translation	Sum
Direct Cost (P_D)	Per Virion	T4	4×10^6	7×10^5	-	3×10^5	-	-	7×10^6	10^7
		Flu	3×10^5	7×10^4	-	-	10^3	2×10^6	10^7	10^7
	Per Infection	T4	9×10^8	10^8	-	7×10^7	-	-	10^9	3×10^9
		Flu	2×10^9	4×10^8	10^3	-	6×10^6	10^{10}	6×10^{10}	8×10^{10}
Opportunity Cost (P_O)	Per Virion	T4	10^7	7×10^5	-	-	-	-	3×10^7	4×10^7
		flu	8×10^5	2×10^5	-	-	-	3×10^7	5×10^7	9×10^7
	Per Infection	T4	2×10^9	10^8	-	-	-	-	6×10^9	8×10^9
		flu	5×10^9	10^9	-	-	-	2×10^{11}	3×10^{11}	5×10^{11}
Total Cost (P_T)	Per Virion	T4	2×10^7	10^6	-	3×10^5	-	-	4×10^7	6×10^7
		flu	10^6	3×10^5	-	-	10^3	4×10^7	6×10^7	10^8
	Per Infection	T4	3×10^9	3×10^8	-	7×10^7	-	-	8×10^9	10^{10}
		flu	6×10^9	2×10^9	10^3	-	6×10^6	2×10^{11}	4×10^{11}	6×10^{11}

The cost of viral replication can be approximated in a similar fashion: we have to consider that the T4 genome is comprised of roughly 4×10^5 DNA bases and that the influenza genome is composed of an order of magnitude fewer RNA bases ($\approx 10^4$). The total costs of a DNA nucleotide and an RNA nucleotide, including the opportunity costs as well

as the direct costs of synthesis and polymerization, are approximately 50 P_T (SI Figure 1, SI Dataset 3). As a result of T4's longer genome length, its total cost of replication ($\approx 10^7$ P_T) is about an order of magnitude higher than that of an influenza genome (Table 1, Figure 1, Figure 2, SI section E).

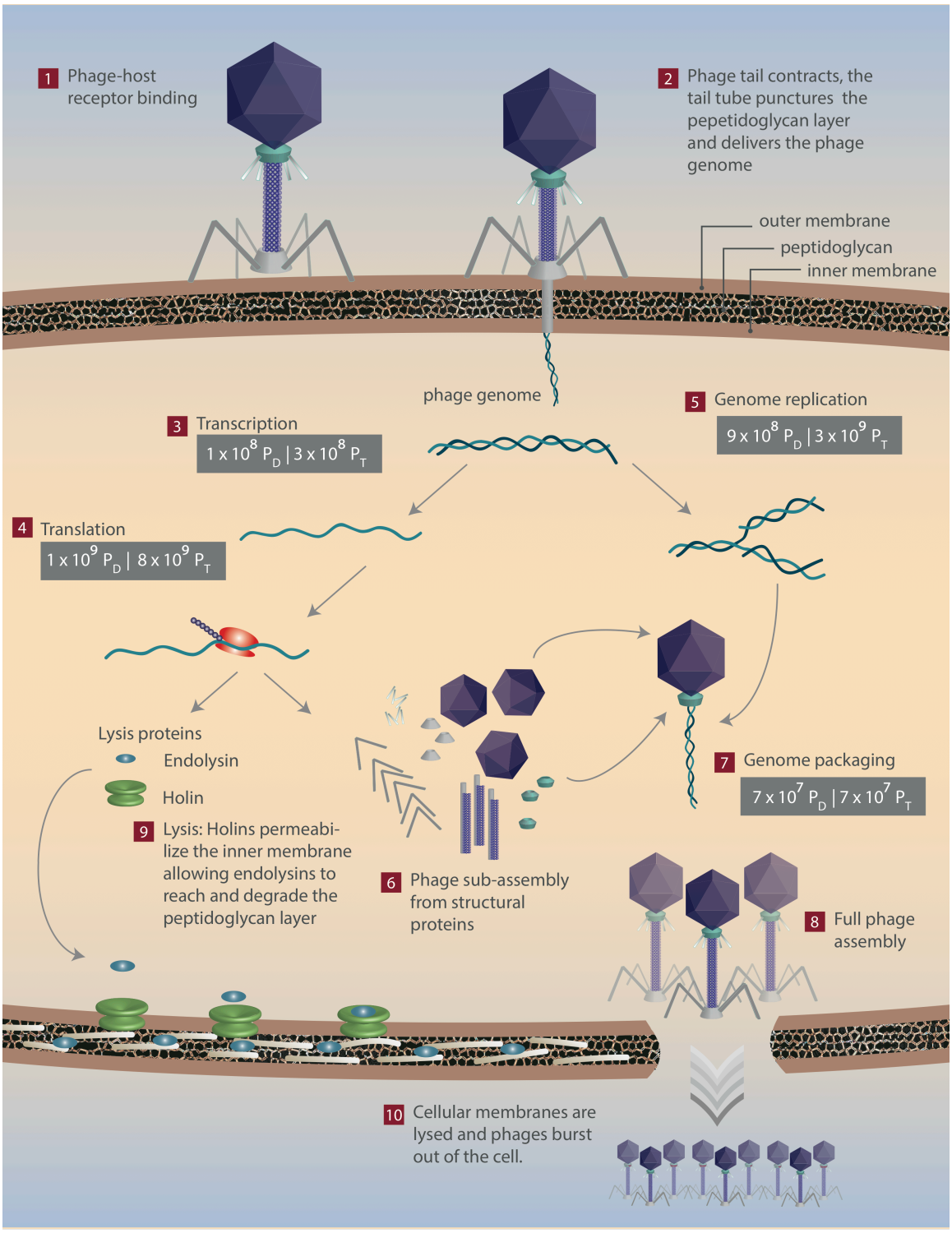


Figure 1. The energetics of a T4 phage infection. The direct and total costs of viral processes are denoted and can be distinguished by their units (P_D and P_T, respectively). The energetic requirements of transcription (step 3),

translation (step 4), genome replication (step 5), and genome packaging (step 7) are shown. See SI sections C through I and Table 1.

The cost estimates of different viral processes during T4 and influenza infections are summarized in Figures 1-3 and Table 1. The overall cost of a T4 infection is obtained by the sum of replication ($E_{REP/i}$), transcription ($E_{TX/i}$), translation ($E_{TL/i}$), and genome packaging ($E_{Pack/i}$) costs required during the infection (SI sections C-I, Table 1, Figure 1, Figure 3). These costs together amount to ≈3 x 10^9 P_D in direct cost and 1 x 10^{10} P_T in total cost (SI sections C-I, Table 1, Figure 1, Figure 3, assuming a burst size of 6000). The total cost of a T4 infection is also equivalent to the aerobic respiration of ≈4 x 10^8 glucose molecules by *E. coli* (26 ATP per glucose (20)). Alternatively, it is equivalent to ≈2 x 10^{11} k_BT (assuming 1 ATP ≈ 20 k_BT (21)).

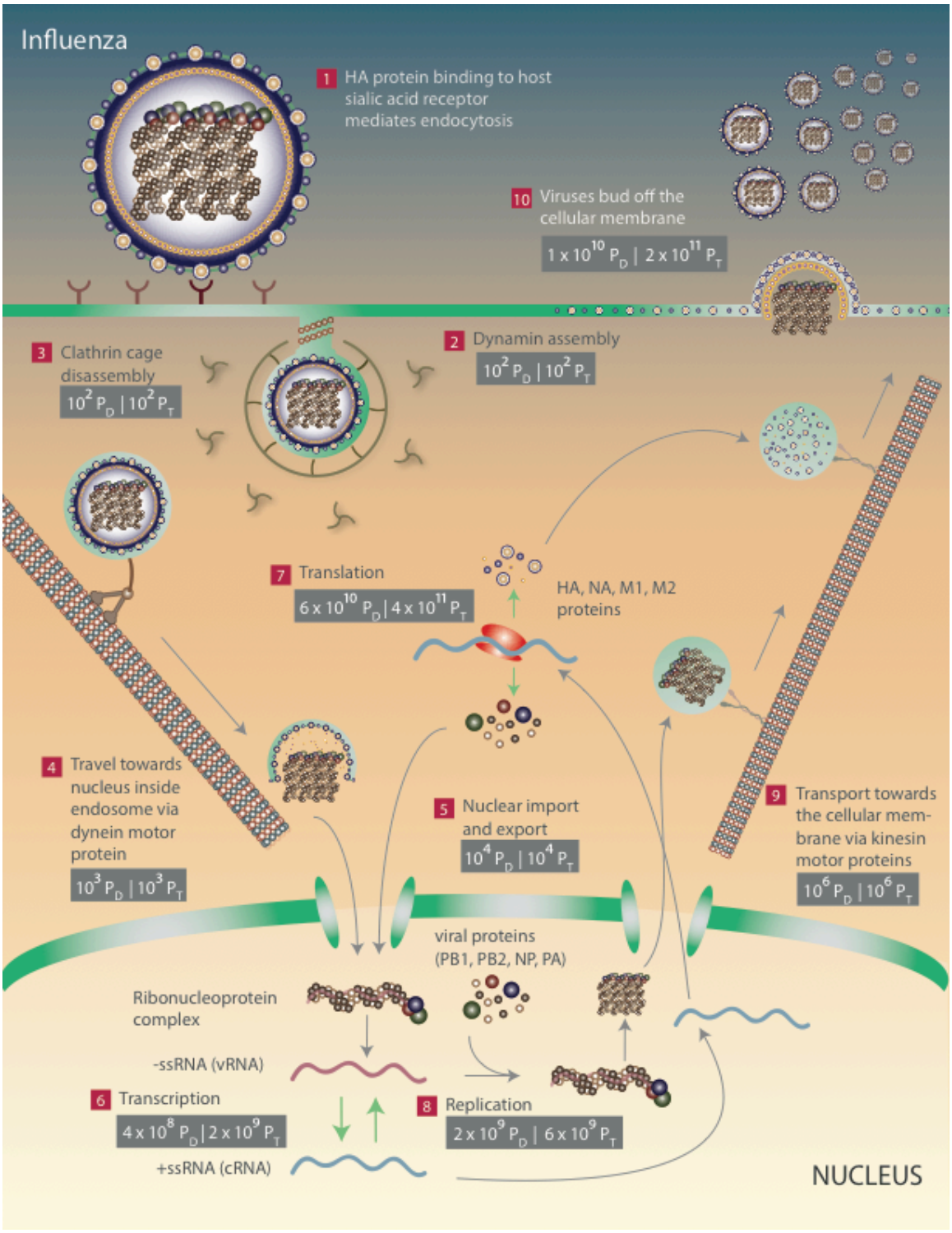


Figure 2. The energetics of an influenza infection. The direct and total costs of viral processes are denoted and can be distinguished by their units (P_D and P_T, respectively). The energetic requirements of viral entry (steps 2,3), intracellular transport (steps 4,5,9), transcription (step 6), translation (step 7),

genome replication (step 8) and viral exit (step 10) are shown. See SI sections C through I and Table 1.

Similarly, the cost of an influenza infection is obtained by adding up the costs of entry (E_{Entry}), intracellular transport ($E_{Transit/i}$), replication ($E_{REP/i}$), transcription ($E_{TX/i}$), translation ($E_{TL/i}$), and exit ($E_{Exit/i}$) required during the infection (SI sections C-I, Table 1, Figure 2, Figure 3). These processes have a cumulative cost of ≈8 x 10^{10} P_D and 6 x 10^{11} P_T, for the assumed burst size of 200. The sum of costs in an influenza infection (6 x 10^{11} P_T) is equivalent to the aerobic respiration of ≈2 x 10^{10} glucose molecules by a eukaryotic cell (32 ATP per glucose). It is also equivalent to ≈10^{13} k_BT. It is interesting to note that for both viral infections the opportunity cost components are the dominant component of the total costs (Table 1).

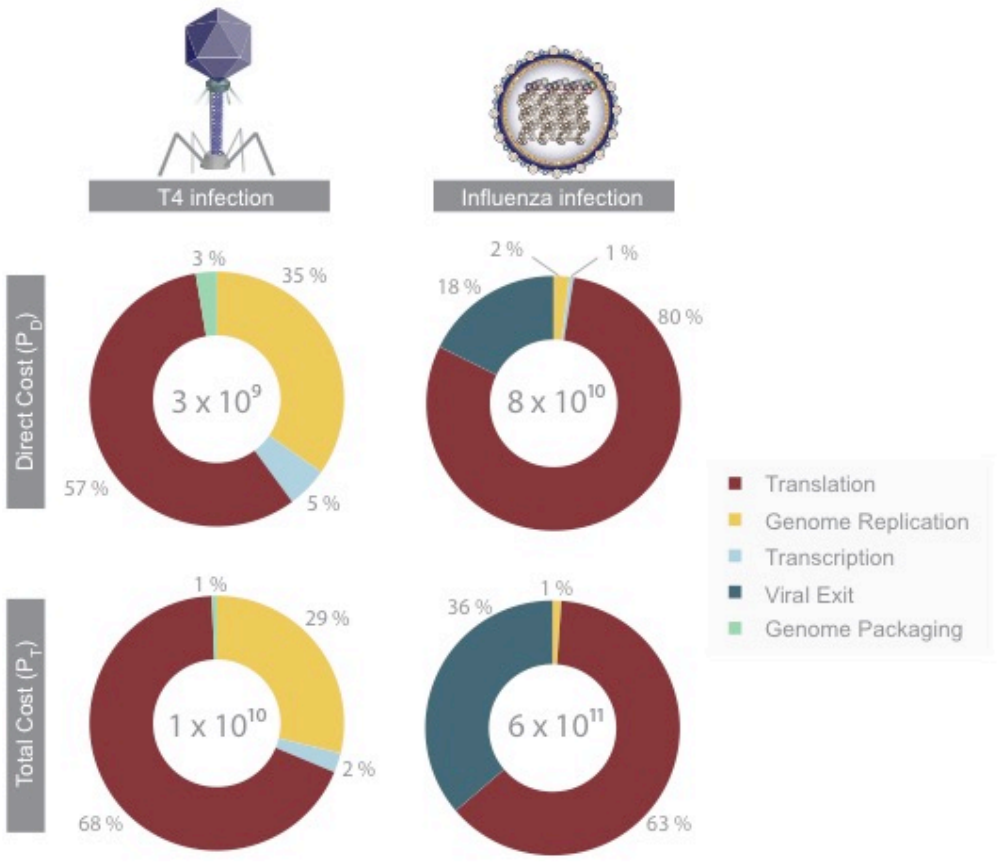


Figure 3. A breakdown of the direct cost (top) and the total cost (bottom) of various viral processes during T4 (left) and influenza (right) viral infections (normalized to the sum of all costs during an infection, as shown in the center of each pie chart). The direct cost of a T4 phage infection is approximately 3 x 10^9 P_D (top) while the total cost is 10^{10} P_T (bottom). The direct and total costs of an influenza infection are approximately ~8 x 10^{10} P_D and 6 x 10^{11} P_T, respectively. Numbers are rounded to the nearest percent, and viral processes costing below 0.5% of the infection's cost are not shown. See SI sections C through I for energetic cost estimates for viral entry, intracellular transport, transcription, viral assembly, and viral exit.

Even though individually a T4 phage and an influenza virus have comparable energetic costs, because of their different burst sizes, the direct cost of a T4 phage infection

is only ≈3% of the direct cost of an influenza infection. Similarly, the total cost of a T4 phage infection is ≈2% of the total cost of an influenza infection. To contextualize these numbers, the host energy budget (or the host energetic cost, depending on the viewpoint of a virus versus a cell) during the infection has to be taken into account. The total cost of a cell is experimentally tractable through growth experiments in chemostats, in which cultures are maintained at a constant growth rate. The number of glucose molecules taken up per cell per unit time can be determined. The number of glucose molecules can then be converted to an energetic supply by assuming typical conversion ratios of 26 or 32 ATPs per glucose molecule depending on the organism (20). This energetic cost estimate will be a total cost estimate because not all glucose molecules taken up by the cell are fully metabolized to carbon dioxide and water to generate ATPs and are used as building blocks for biomass components instead. During the cellular life-cycle, the cell has to double its number of building blocks prior to division, and to do so, a fraction of glucose molecules taken up is diverted away from energy production towards biosynthesis pathways. As such, cellular energetic cost estimates that are derived from chemostat experiments are total cost estimates because they report on the combined opportunity and direct costs of a cell.

Based on chemostat growth experiments (19), the total energy used by a bacterium and a mammalian cell with volumes of 1 μm^3 and 2000 μm^3, respectively, are ≈3 x 10^{10} P_T and ≈5 x 10^{13} P_T, during the course of their viral infections (SI section J). A simpler estimate for arriving at the total cost of *E. coli* with a 30-minute doubling time is by considering the dry weight of *E. coli* (≈0.6 pg at this growth rate) (22). Given that about half of the cell's dry weight is composed of carbon (22), an *E. coli* is composed of ≈2 x 10^{10} carbons, supplied from ≈3 x 10^9 glucose molecules, since each glucose contributes 6 carbons. With the 26 ATP per glucose conversion for *E. coli*, this is equivalent to a total cost of ≈7 x 10^{10} P_T,

which is similar to the number obtained from chemostat growth experiments (19)(SI section J).

Moreover, we estimate the fractional cost of a viral infection as the ratio of total cost of an infection, $E_{T/i}$, to the total cost of the host during the infection, $E_{T/h}$. For the T4 infection with a burst size of 200 virions, $E_{T/i}$ ≈1 x 10^{10} P_T (Table 1) and $E_{T/h}$ ≈3 x 10^{10} P_T, therefore the fractional cost of the T4 infection is ≈0.3. Interestingly, a calorimetric study of a marine microbial community demonstrated that 25% of the heat released by microbes is due to phage activity. (23) If we assume that the majority of the direct cost of a cell is associated with translation (19, 21), these calorimetric studies square well with our estimate for the ratio of direct costs. In contrast, the influenza infection despite its larger burst size (6000 virions) leading to a higher $E_{T/i}$ (≈6 x 10^{11} P_T) has a fractional cost of just 0.01 as the host cell is much bigger. Finally, we estimate that the heat release due to the T4 and influenza viral infections are approximately 0.2 nJ and 2 nJ, respectively (SI section K). While influenza infection results in an order of magnitude more heat, the average power of T4 and influenza infections are surprisingly very similar, on the order of 200 fW (SI section K).

3.4 Scaling of viral energetics with size for phages

While we have concluded that for the influenza virus and the T4 phage the translational cost outweighs the replication cost, the ratio of these two costs varies according to the dimensions of a virus. In the case of T4 and influenza, these two viruses have comparable dimensions and consequently were built from a similar number of amino acids (SI Tables 1 and 2). However, because for double-stranded DNA phages, the capsid is mostly composed of proteins whereas the virion interior is mostly dedicated to the genetic material (24), it follows that with the diminishing surface area to volume ratio of a spherical

object as it grows in size, the ratio of translational cost to replication cost also diminishes with increasing radius of a spherical capsid. This simple rule governs not just the nucleotide or amino acid composition of a virus, but more fundamentally, it governs the elemental composition of viruses with spherical-like geometries (24).

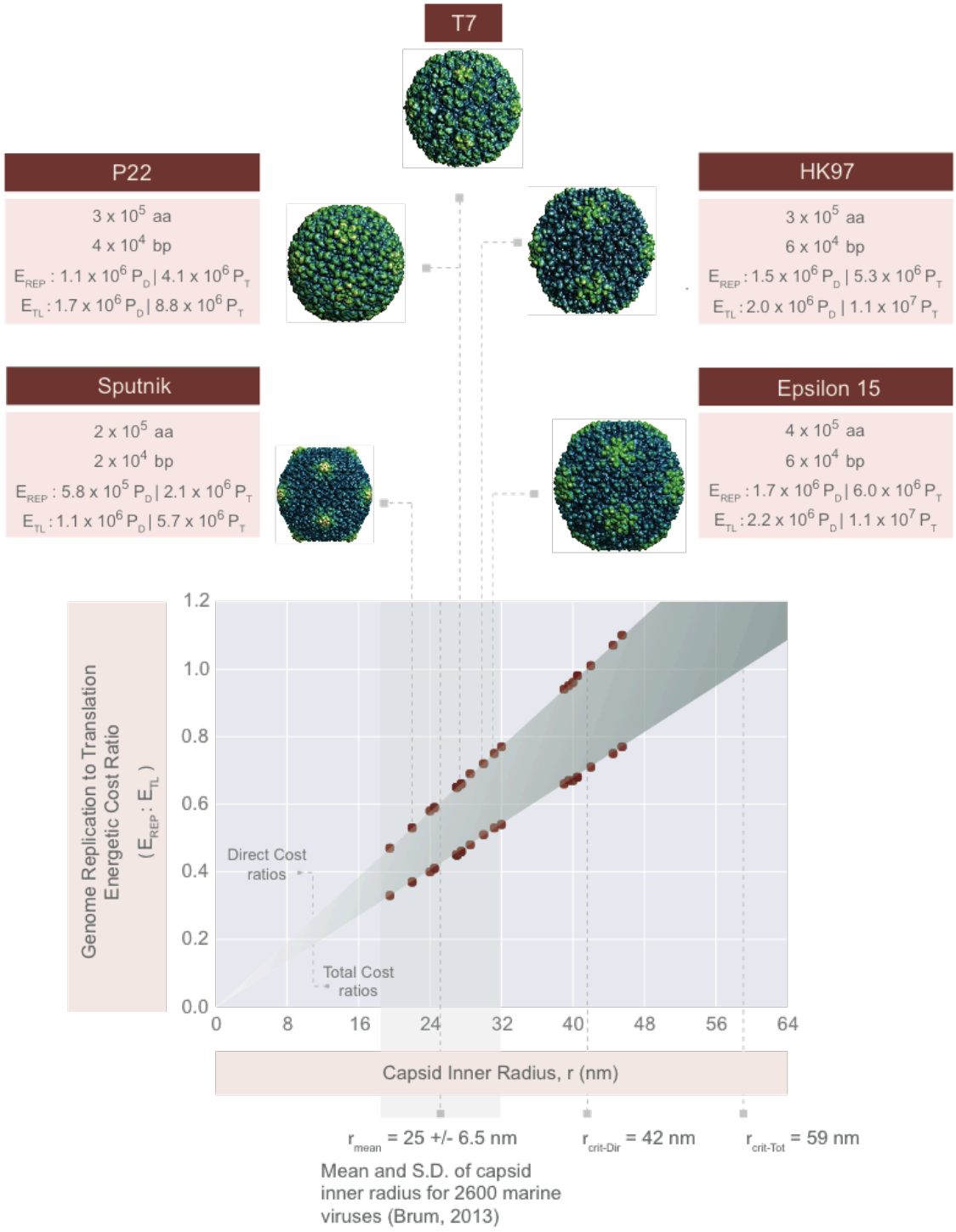


Figure 4. Generalizing viral energetics. A plot of the genome replication cost (E_{REP}) to translational cost (E_{TL}) ratio as a function of the virus inner radius, r. The plot uses the geometric parameters of double-stranded DNA viruses

with icosahedral geometries (SI Dataset 1, SI section L). The predicted numbers of amino acids and nucleotides are derived in SI Dataset 1. Cost ratios are shown for both direct and total cost estimates. All viruses shown infect bacteria except Sputnik, which is a satellite virus of the giant Mimivirus. We have zoomed in on viruses Sputnik (r = 22 nm), P22 (r = 27.5 nm), T7 (r = 27.5 nm), HK97 (r = 30 nm), and Epsilon15 (r = 31.2 nm). The capsid structures for these representative viruses were obtained from the VIPERdb (25) and image sizes were scaled based on radii shown in SI Dataset 1 to accurately represent the relative sizes of each capsid. The critical radii for the total cost ($r_{crit-Tot}$) and the direct cost ($r_{crit-Dir}$) estimates are shown. We have also included the mean (r_{mean} = 25 nm) and standard deviation (gray vertical box, $\pm$6.5 nm) of viral capsid inner radii from 2,600 viruses collected by the Tara Oceans Expeditions (26). Note, here we have subtracted the mean capsid thickness (3 nm) from the mean capsid radius reported by Brum *et al.* to arrive at the mean *inner* capsid radius.

The full derivation of replication and translational cost estimates as a function of viral capsid inner radius, r, can be found in the SI section L. From these expressions, it is clear that the translational cost of a virus scales with r^2, whereas the replication cost scales with r^3 (Figure 4). The critical radius at which replication will outweigh translation in cost is ≈60 nm for total cost estimates, $r_{crit-Tot}$ (Figure 4, SI section L). For the direct cost estimates, the critical radius, $r_{crit-Dir}$, is ≈40 nm. Interestingly, a survey of structural diversity encompassing 2,600 viruses inhabiting the world's oceans reveals that the average outer capsid radius is 28 nm (26) (25 nm inner radius), which is much smaller than the critical radii at which replication becomes the dominant cost (Figure 4). As such, for the majority of viruses, we predict translation is the dominant cost of a viral infection.

Furthermore, we provide genome replication to translation cost ratios for about 30 different double-stranded viruses, primarily phages (SI Dataset 1, Figure 4). While we have omitted calculations for the virus tails, they can be simply treated as hollow cylinders and will further decrease the expected replication to translation cost ratio for the tailed viruses. Although we have calculated these ratios for double-stranded DNA phages, similar principles can be applied to modeling the energetics of other viral groups.

3.5 Forces of evolution operating on viral genomes.

Inspired by efforts to consider the evolutionary implications of the cost of a gene to cells of different sizes (19, 27), we were curious whether similar considerations might be in play in the context of viruses. For example, we asked which evolutionary forces are prominently operating on neutral genetic elements that are incorporated into viral genomes, either by horizontal gene transfer, gene duplication or other similar types of events. We further asked whether the viral size is a parameter of interest in the tug of war between different forces of evolution. We will address these topics by assuming that the viral infection, consistent with our findings for T4, consumes a substantial portion of the host energy budget. We further assume that the energetic cost of a genetic element translates to a proportional fitness cost. We believe this assumption to be relevant when the host growth condition is energy or carbon substrate limited.

For a genetic element to remain in the population, regardless of whether it is beneficial or not, it must face the consequences of genetic drift which scales with the viral effective population size, N_e, as ${N_e}^{-1}$. We follow the treatment of Lynch and Marinov who argue that the net selective advantage of a genetic element is $s_n = s_a - s_c$, where s_a and s_c denote the selective advantage and disadvantage, respectively (Figure 5B). For a genetic

element within a viral genome that is non-transcribed and non-translated (Figure 5C), only the energetic cost of its replication poses a selective disadvantage. Assuming the genetic element provides no benefit to the virus ($s_a = 0$), the net selective advantage can be stated as $s_n = -s_c$, the absolute value of which must be much greater than N_e^{-1} for selection to operate effectively. Following Lynch and Marinov and others (27, 28), we make the simplifying assumption that a neutral genetic element's selection coefficient, s_c, is proportional to its fractional energetic cost, E_g (Figure 5C). This means that the viral infection is energy (or carbon source) limited. Because we assumed that the energetic cost of a viral infection is comparable to the total energy budget of a cell, any increase to the cost of a virus would necessitate a smaller burst size. Using the viral burst size as a proxy for the viral growth rate, we are then able to relate the additional fractional energetic cost of a neutral genetic element to a fitness cost.

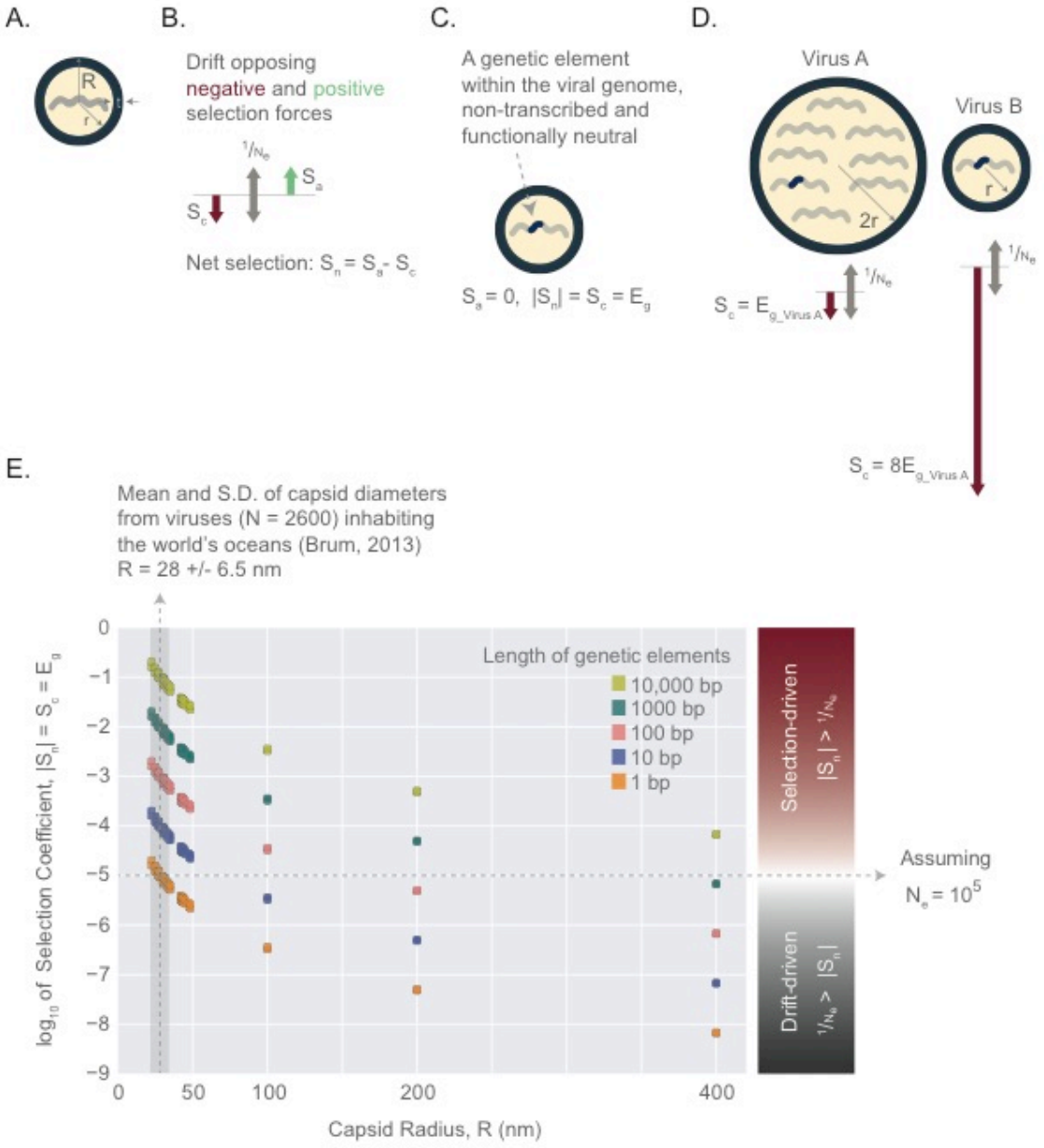


Figure 5. Evolutionary forces acting on genetic elements within viral genomes. A) Schematic of a virus as a spherical object, with an inner radius, r, an outer radius, R, and a capsid thickness, t. The capsid is composed of viral proteins, while the inner volume holds the viral genome. B) Positive and negative selective forces (s_a and s_c) at a tug of war with the force of genetic drift, which scales as N_e^{-1}, where N_e is the viral effective population size. C) A schematic of a genetic element within a viral genome. It is assumed to be non-functional ($s_a = 0$) and non-transcribed, resulting in $|s_a| = s_c = E_g$,

where s_a corresponds to the net selection coefficient and E_g corresponds to the fractional cost of a genetic element. D) The evolutionary forces acting on a genetic element within Virus A and Virus B genomes. The fractional cost of a genetic element in Virus B, E_{g_VirusB}, is 8 times higher than the fractional cost of the same element in Virus A, E_{g_VirusA}. Note, Virus A has twice the radius of Virus B, and therefore its genome is 8 times longer than that of Virus B (schematically represented by the number of genetic segments). Both viruses are assumed to have radii greater than critical radii, $r_{crit-Tot}$ and $r_{crit-Dir}$. E) Log_{10} E_g estimates for non-transcribed and neutral genetic elements of different lengths (1 – 10,000 base pairs) within the context of 30 dsDNA viruses ranging from ~20 nm to 400 nm in radius (SI Dataset 2; viruses with R > 50 nm are hypothetical dsDNA viruses). Log_{10} E_g estimates derived from both direct and total cost estimates are included (there is minimal difference between these estimates, which is not visible in this figure, see SI Dataset 2). Assuming $N_e = 10^5$, the region above the horizontal dashed line represents a selection-dominated regime, and the region below it represents a drift-dominated regime. For comparison, we have included the mean (vertical dashed line, 28 nm) and standard deviation (gray vertical box, $\pm$6.5 nm) of viral capsid outer radii obtained from 2600 viruses collected during the Tara Oceans Expeditions (26).

In the case of a non-transcribed genetic element, $E_g = \frac{E_{REP/v}}{E_v}$, where $E_{REP/v}$ corresponds to its replication cost and E_v is the sum of all costs of a virus (Figure 5C). Given that replication cost scales as r^3 the effects of selection relative to genetic drift could be different for viruses of different sizes. Consider two viruses with the same burst size, with Virus A, having a radius that is two times larger than that of Virus B (Figure 5D). Because both viruses are assumed to have radii larger than the critical radius, we imagine the scenario in which the cost of genome replication is the dominant cost of synthesizing these viruses.

The fractional cost of a genetic element in the smaller virus, $E_{g\_Virus\ B}$ is then equal to $8E_{g\_Virus\ A}$, where $E_{g\_Virus\ A}$ is the fractional cost of the genetic element in the larger virus. This is because the length of the genome is proportional to r^3, and consequently, E_g is inversely proportional to r^3 (Figure 5D).

Figure 5E and SI Dataset 2 provide E_g estimates for genetic elements of different lengths (1 – 10,000 base pairs) within 30 dsDNA viruses. To illustrate the effect of scaling in the example provided above, we made the simplifying assumption that the viruses are large enough that their E_v are approximately equal to their replication costs. However, for E_v values in Figure 5E and SI Dataset 2, we provide more precise estimates, treating E_v as the sum of both the replication cost and the translational cost of a virus. The cost of replicating a double-stranded genetic element can be obtained from SI Eq. 3. For a 1 kb element, which is about the average length of a bacterial gene, the direct and total costs of its replication per virus, $E_{REP/v}$, are 3 x 10^4 P_D and 9 x 10^4 P_T, respectively. Both direct and total cost estimates indicate that the strength of selection acting on a 1 kb, non-transcribed element ranges from 2 x 10^{-2} - 7 x 10^{-6} (SI Dataset 2, Figure 5E) when considering viruses with radii ranging from ≈20 nm to 400 nm. The difference between direct and total estimates of selection strength is minimal within this range of capsid radii and continues to diminish as the capsids grow in size.

To examine whether selection or genetic drift will decide the fate of a genetic element we need to assess each virus's effective population size. This is difficult because the effective population size of most viruses is unknown and subject to great variability due to several environmental factors (29). The current effective population size estimates regarding HIV, influenza, dengue, and measles fall within 10^1 to 10^5 (29-31). Based on the wide range

of variation in these effective population sizes, it is difficult to make conclusive statements. It is, however, apparent that the strength of selection on neutral genetic elements is a non-linear function of the viral capsid radius and becomes much weaker as viruses get larger (Figure 5E). In fact, for giant viruses (with outer radius, R > 200 nm), assuming an $N_e^{-1} = 10^{-5}$, genetic drift could overpower selection, allowing for the persistence of neutral elements of lengths 100 bp or shorter in the population. For the majority of viruses ($R = 28 \pm 6.5$ nm) (26), however, selection is likely to be the dominant force and drift may only play a role for genetic elements that are just a few base pairs long (Figure 5E, SI Dataset 2).

3.6 Discussion

There have been several experiments that imply that a viral infection requires a significant portion of the host energy budget (3, 5, 10, 11, 32-34). Following these experimental hints, we enumerated the energetic requirements of two very different viruses on the basis of their life-cycles, and thereby estimated the energetic burdens of these viral infections on the host cells. According to our total cost estimates, a T4 infection with a burst size of 200 will consume a significant portion (about 30%) of the host energy supply. This result, demonstrating a significant fraction of the host energy used by an infection, supports the experimental findings that the T4 burst size is correlated positively with the host growth rate (7, 11). It also lends further credence to the hypothesis that auxiliary metabolic genes within phage genomes are not just evolutionary accidents; rather, they have come to serve a functional role in boosting the host's metabolic capacity, which translates into larger viral burst sizes (3, 4, 34, 35). These calculations make it all the more interesting to develop high-precision, single-cell calorimetric techniques to monitor energy usage during viral infections.

Perhaps the most promising support for T4's cost estimate is the observation that

the maximum T4 burst size is 1,000 virions (15). Using the total cost to make new viruses, at a burst size of 1,000, the viral infection would consume 170% of the host normal energy supply at a 30 min growth rate, consistent with the observed apparent upper limits on burst size.

It is however important to note that in all of our estimates, we make the assumption that the sources of nitrogen, sulfur, phosphorus, and other trace elements are in excess, which is typical of culture conditions in the laboratory and from which most burst measurements are obtained (SI Section B), but this assumption may not be valid in certain natural environments as demonstrated by the phosphorus-limited environments of marine ecosystems (36). In such limited environments, phages are shown to carry auxiliary genes and to actively rewire the host metabolism (full discussion can be found in the SI Section A). It would be interesting to have additional experimental studies that go beyond the ideal conditions of a laboratory experiment to fully explore the range of possible limiting factors in a viral infection.

While there are several fascinating studies that explore the link between the host metabolism and phage infections (3, 6, 12, 13), similar studies focusing on viruses of multicellular eukaryotes are largely lacking. To that end, we chose to estimate the energetic cost of a representative virus for this category, namely, the influenza virus. The influenza virus and T4 phage are functionally and evolutionarily very different viruses. Yet, they have a very similar per-virus cost, regardless of whether the total or the direct cost estimates are being considered. This is primarily due to the fact that they have a similar translational cost, which dominates all other costs. Their comparable cost of translation is due to the fact that these viruses have similar dimensions and are both composed of about a million amino acids.

Perhaps even more surprising is that both viral infections have very similar average power consumptions, on the order of 200 fW (SI Section K).

Even with its higher burst size, an influenza infection has a total cost that is just 1% of the total energetic budget of a eukaryotic cell over the characteristic time of the viral infection. This is because a typical eukaryotic cell is estimated to have much higher energy supply than a typical bacterium under the same growth conditions. So far in our estimates, we do not account for the possible inefficiencies at various stages of the viral infection, which may drain more of the host energy than we estimated. Specifically, burst sizes are typically reported from plaque assays, which count the number of infectious virions that create plaques. However, we don't have a good estimate for the number of non-infectious viruses that arise from faulty genome replication, transcription, or viral assembly, for example. This point is especially important when considering RNA-based viruses such as influenza or HIV, which have higher mutation rates (10^{-4}-10^{-6} mutations per base pair per generation; (37)) compared to dsDNA viruses such as T4 (10^{-6}-10^{-8} mutations per base pair per generation; (37)). As a result of these higher error rates, RNA-based viruses may have greater hidden costs associated with aborted viral synthesis or a greater fraction of faulty and non-infectious virions.

Even infectious viruses cannot all be guaranteed to enter the lytic cycle upon infecting a host cell. For example, only 10% of influenza-infected host cells have been shown to generate infectious virions (38), demonstrating the cumulative inefficiency of an influenza infection. Counting plaques to measure viral burst sizes likely underestimates the true burst size, and result in an underestimation of the infection cost. As such, single-cell studies of viral infection could provide a detailed breakdown of inefficiencies at various steps of the viral life-cycle and enable more exact cost estimates.

In addition to the energetics of viral synthesis, another burst-limiting metric to consider is a volumetric one, namely the fraction of the host volume occupied by viruses during an infection. Taking influenza and T4 as our representative viruses, it is clear that they both occupy a relatively small percentage of the total host volume (39). A T4 infection takes up less than 5% of its host's total volume. An influenza infection takes up even less space (≈0.2% of its host volume). These estimates suggest that 1) for T4, the energy requirement is more likely a burst-limiting factor than the volumetric requirement, and 2) for influenza, neither energetic nor volumetric factors seem to be limiting the burst size.

We have already considered several possible causes for the inefficiencies of an influenza infection, which have experimentally been shown to result in only one infectious virus out of every ten produced (38). Accounting for this inefficiency, boosts the total cost of an influenza infection to ≈10% of its host's. A second consideration that could explain the relatively low energy uptake of an influenza infection is the growth state of its host cell. Our current estimate assumes that the host cell is under maximal growth conditions. When the host cell is not dividing, however, its energy supply could be as low as 10^{12} P_T (estimated for a 12-hour infection period, SI section J). In considering both the 10% infection efficiency and assuming a slow-growing host cell, the influenza infection's total cost could also be a significant fraction of its host's total cost.

Another consideration is that implicit in our original question is the assumption that all viruses must conform to producing the maximum burst size allowed by their host energetic supply. This assumption, while it appears compatible with our findings for phages, may simply not be true for viruses such as influenza. It may be that a eukaryotic virus within a multicellular setting has an entirely different growth strategy than a prokaryotic virus infecting a single-celled organism. The influenza virus burst size is not only under selection

pressure within its host cell, but also within the multicellular organism that serves as its secondary host.

Finally, we will need future experimental studies to test the assumptions underlying the relationship between the fractional cost of a neutral genetic element and the strength of negative selection acting on the viral population. There is also a great need for estimates of the effective population sizes of different viruses within their natural environments. With current effective population size estimates for viruses it appears that selection likely determines the fate of genetic elements for the majority of viruses, which have on average 28 nm radii (26) (Figure 5E, SI Dataset 2). However, for larger viruses (R > 200 nm), the diminishing, fractional cost of a gene may enable the interference of genetic drift to the extent that neutral genetic elements could persist in the viral population. The result of such a phenomenon could be genome expansions in the form of gene duplication events, cooption of previously noncoding, horizontally transferred elements into functional genes and regulatory domains, and perhaps even a trend towards greater autonomy over large evolutionary time-scales. This effect may explain the unusual number of duplication events in the genomes of giant viruses such as that of the Mimivirus (40, 41).

3.7 Supplementary Information

3.7.A Viral rewiring of the host metabolism

Viruses rely entirely on their host as an energy source. Instead of passively exploiting the host's metabolic energy, some viruses appear to augment it (2). A particularly compelling example is demonstrated by phages that infect cyanobacteria. Cyanophages carry genes for photosystem II, high-light inducible protein, transaldolase, and ribonucleotide reductase, which are all transcribed during an infection (3). Given the unprecedented presence of

photosynthetic genes in viral genomes and the active expression of these genes during an infection, it is proposed that cyanophages carry these genes to increase the host energy supply and deoxynucleotide production for their own replication (3). An analogous finding is the presence of sulfite reductase genes in genomes of phages that infect deep-sea bacteria, which use sulfur as their energy source (34). Here too, these phages are hypothesized to add to the host's metabolic output. It was also shown that large DNA algal viruses encode deoxynucleotide synthesis enzymes. For example, PBCV-1 encodes 13 nucleotide metabolism enzymes and EsV-1 encodes an ATPase and both subunits of ribonucleotide reductase (32). The most recent study on this topic identified more than 200 virus-encoded auxiliary metabolic genes such as those used in nitrogen and sulfur metabolism in marine viral metagenomes (6).

Adenoviruses have been shown to reprogram the host's glutamine metabolism by up-regulating glutamine transporters and glutamine catabolism enzymes. Glutamine is a critical amino acid used in the synthesis and import of other amino acids. Interestingly, this viral rewiring of glutamine metabolism is shown to boost the concentration of certain amino acids as well as increase glutamine reductive carboxylation. Together these effects are required for optimal viral production not only during an adenovirus infection, but also during herpes and influenza viral infections (4). In addition to virus-infected cells, which have high demand for molecular building blocks and energy, cancer cells, immune cells, and other proliferating cells similarly rewire glutamine metabolism for energy production and biosynthesis (42).

Moreover, several examples from the early years of virology have shown that viruses such as Rous sarcoma virus and Feline Leukemia Virus increase their hosts' glycolytic rate upon infection (33). Similarly, the Vaccinia Virus was shown to upregulate mitochondrial

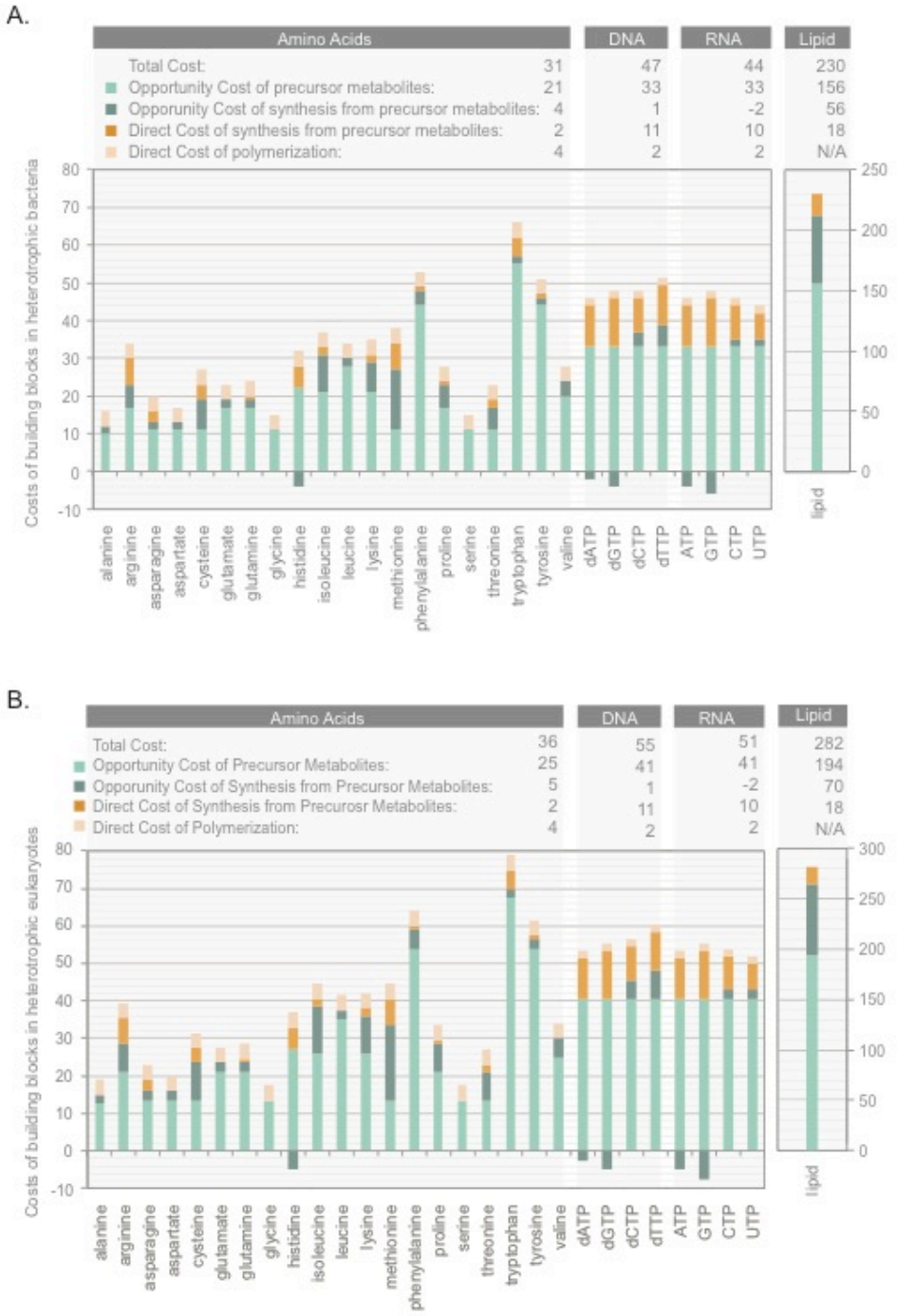
genes involved in the electron transport chain, thereby increasing the ATP production within its host (5).

3.7.B Energetic cost definitions and assumptions

Viral synthesis requires the expenditure of the host's ATP-equivalent molecules as well as the usurpation of the host's monomeric building blocks such as nucleotides, amino acids, and in the case of some viruses, lipids. To synthesize these monomeric building blocks and generate ATP-equivalent molecules, heterotrophic cells, such as the hosts of T4 or influenza, rely on reduced carbon sources from their environment. In calculating the energetic cost of a viral infection and its impact on the host energy supply, it is critical that we state several assumptions about the host growth conditions. First, we assume that the host is growing aerobically at 37°C, with glucose as the sole carbon source. Second, we assume that sources of nitrogen, sulfur, phosphorus, and other trace elements are in excess, which is typical of culture conditions in the laboratory, and from which burst size measurements are commonly obtained. Third, we assume that the growth media contains only inorganic sources of nitrogen, requiring cells to synthesize amino acids rather than salvaging them from a growth medium supplemented with peptides (although this assumption should be modified in the case of a mammalian host cell which cannot synthesize all amino acids).

As the sole carbon source in the growth media, glucose will serve both as a source of energy and biomass. SI Figure 1 provides detailed estimates of the opportunity and direct costs of precursor metabolites in heterotrophic bacteria and eukaryotes, assuming glucose as the sole carbon source.

As discussed in Chapter II, in the synthesis of building blocks from precursor metabolites there is an additional source of opportunity cost, namely the oxidation of electron carrier molecules. Had these electron carriers been preserved for energy production rather than being oxidized during biosynthesis pathways, 2 ATP molecules could have been generated from each NAD(P)H molecule using the bacterial ETC (43) and 2.5 ATP molecules from the eukaryotic ETC (44). The combination of the opportunity cost of precursor metabolites and the opportunity cost of building block synthesis from precursor metabolites will be referred to as the opportunity cost of a building block, (SI Figure 1). To simplify, we will generally refer to the opportunity cost of a building block in our calculations.



SI Figure 1. A breakdown of direct and opportunity costs associated with amino acids, DNA, RNA, and lipids in the context of A) bacterial and B) eukaryotic metabolism. Average cost values are reported in the table above the chart. See SI Dataset 3 for a detailed derivation of these costs

3.7.C Viral entry cost

The process of viral entry varies extensively across different viral groups. While many animal viruses enter their host cell through clathrin-mediated endocytosis or fusion with the cell membrane (45), most phages inject their genetic material with the capsid remaining outside of the cell. In both cases, however, entry is mediated by the interaction between viral entry proteins and host receptors (45).

For T4, it is the interactions between a minimum of 3 long tail fibers and cellular receptors that initiates a cascade of conformational changes (46) (Figure 1, step 1). After this preliminary interaction, the base plate is subsequently brought closer to the cell membrane, allowing the short tail fibers to interact with their host receptors. The tail sheath contracts, resulting in the tail tube puncturing the outer cell membrane (Figure 1, step 2). Then, conformational changes in gp5 phage protein activate its lysozyme domains, resulting in the digestion of the peptidoglycan layer (Figure 1, step 2). Effectively, the tail tube passes through the periplasm. At this point, the phage DNA is passed through the inner membrane via the tail tube and is then exposed to the intracellular environment (Figure 1, step 2). In general, viral entry proceeds through protein conformational changes and doesn't rely on ATP expenditure (47). In the case of the T5 phage, this point has been explicitly demonstrated (48).

The influenza virus is composed of a capsid that is enveloped by a lipid membrane (Figure 2). Inside the capsid reside the ribonucleoprotein complexes, which are composed of the segmented viral genome encapsidated by proteins. The viral membrane is decorated with hemagglutinin (HA) proteins, which bind to the host sialic acid receptors, thereby initiating clathrin-mediated endocytosis (Figure 2, step 1). During endocytosis, a self-assembled protein cage composed of clathrin triskelions forms around the inward budding vesicle.

Once the clathrin cage has formed, dynamins perform the last stage of endocytosis (Figure 2, step 2).

Dynamin is a mechanochemical GTPase that self-assembles into multimeric spirals at the necks of clathrin-coated endocytic vesicles to catalyze membrane fission. The dynamin helix wrapped around the neck of an endocytic vesicle forms a protein-lipid tube with an inner diameter of 20 nm. In the presence of GTP, the dynamin helix undergoes structural changes that result in the reduction of the inner diameter (49). Because the dynamin helix is composed of at least 2 turns (35), each turn is composed of 13 dynamins, and each dynamin consumes 1 GTP, the energy requirement for vesicle fission can be approximated as 30-100 P_D (Figure 2, step 2).

Once the vesicle is released from the cell membrane, the clathrin cage has to be disassembled. This process requires the expenditure of 3 ATPs per triskelion (50). For a clathrin cage composed of 36 triskelions, this is equivalent to about 100 P_D (Figure 2, step 3). As the endosome matures, the endosomal lumen becomes more acidic. The influenza virus has exploited this feature for the uncoating of its lipid membrane as well as the disassembly of its capsid. The endosomal pH drop activates the viral transmembrane proton channel, M2. The influx of H^+ ions from the endosome through M2 leads to the dissociation of the viral capsid proteins. The acidic environment also triggers conformational changes that expose the viral HA2 fusion peptide, which subsequently fuses the viral and endosomal membranes (51). These two events together enable the release of the viral ribonucleoprotein complexes into the host's cytoplasm (52). To summarize, the cost of entry for influenza per virion, E_{Entry}, falls within the range of 10^2 to 10^3 P_D. It's important to note that in our estimates we will generally not include the cost of host protein production, unless a protein is exclusively produced for viral synthesis. For example, the costs of producing the clathrin

cage and dynamin proteins are not included due to the fact that these proteins are recycled and produced for the host's own functions.

T4 efficiently exposes its genomic material to the host cytoplasm where it can be readily transcribed and translated. The influenza virus, however, due to the much larger volume and extensive compartmentalization of its eukaryotic host, faces additional obstacles in the way of reaching sites of transcription and translation. We will describe these obstacles in the next section.

3.7.D Viral intracellular transport cost

Replication and transcription of the influenza virus occur inside the nucleus. Like other cargo destined towards the nucleus from the plasma membrane, the endosome carrying the influenza virus is transported via the action of the dynein motor proteins along microtubule tracks (Figure 2, step 4). Unlike the kinesin motor protein, dynein is one that takes variable step sizes along the microtubule. As a result of having a hexameric ring of AAA+ ATPases (similar to Vps4, described later SI section I), dynein has multiple ATP binding sites. For the purposes of our estimates, we assume a step size of 8 nm and the expenditure of 1 ATP per step (53). If we assume that the nucleus resides roughly in the center of a cell $\approx 10\ \mu m$ in diameter (18), this will require approximately 10^3 dynein steps in carrying the endosome. As a result, the cost of transport for the viral genome from the vicinity of the cell membrane to the nucleus is approximately $10^3\, P_D$ (Figure 2, step 4).

The ribonucleoproteins are imported to and exported from the nucleus via nuclear localization signals (Figure 2, step 5). The nuclear import of influenza ribonucleoproteins has been experimentally demonstrated to be an energy-consuming process (54). To get access to the nucleus, these ribonucleoproteins have to pass through the nuclear pore complex. They

do so by binding to importins using nuclear localization signals. Each cargo imported into the nucleus will require the hydrolysis of one GTP. Whether ribonucleoproteins travel in and out of the nucleus separately or together is not definitely known, but there is recent evidence suggesting that the eight ribonucleoproteins travel as one cargo (55). As such, we estimate that the import of the viral genome will cost 1 P_D per virus.

Once inside the nucleus, the influenza genome is transcribed and replicated, the costs of which will be discussed in later sections. The production of the full ribonucleoprotein complex is a convoluted process, requiring the transcription of the viral genome (Figure 2, step 6), the nuclear export of the resulting transcripts (Figure 2, step 5), as well as the translation of viral mRNA transcripts (Figure 2, steps 7). Once the viral proteins are generated, some proteins travel back into the nucleus to encapsidate the viral -ssRNA genome and form the next generation of ribonucleoprotein complexes (PB1, PB2, NP, and PA proteins) (Figure 2, steps 5, 8). Other viral proteins are transported to the cell membrane, and together with ribonucleoprotein complexes, which are destined to the same site, bud off.

Similar to the nuclear entrance, cargo exiting the nucleus must also pay a price. The nuclear export cost through the CRM1 export pathway, which is the one used by influenza, incurs 1 P_D per cargo, similar to the import pathway (55). Considering that on average 6000 influenza genomes have to leave the nucleus to eventually give rise to 6000 influenza virions, we estimate the export of ribonucleoproteins from the nucleus will cost $\approx 6 \times 10^3$ P_D. The mechanism and energetics of viral mRNA export from the nucleus is not well understood (56) though there's a growing body of evidence implicating the involvement of the nuclear RNA export factor 1. We suspect the cost of mRNA nuclear export is on par with the cost of ribonucleoprotein nuclear export and estimate the cost of nuclear import and export to be $\approx 10^4$ P_D.

Upon exiting the nucleus, ribonucleoproteins have to be transported towards the cell membrane where they can co-assemble with other viral proteins and bud off. Their path to the cytoplasm starts at the Microtubule Organizing Center (MTOC) and proceeds via Recycling Endosomes bound to the RAB-11 GTPase (55). This endocytic transport is powered by kinesin motor proteins trekking along microtubules with a step size of 8 nm and the hydrolysis of 1 ATP per step (57) (Figure 2, step 9). Assuming the eight ribonucleoproteins travel as one cargo – consistent with the reported typical size of an endosome (200 nm in diameter) (58) – a 5 μm transport of 6000 ribonucleoproteins from the nucleus to the cytoplasm costs ≈3 x 10^6 P_D. Under a similar assumption (i.e. each endosome carries the proteins required to build a single virus), the cost of protein transport to the apical cell surface is similar to the cost of ribonucleoprotein transport. Considering the sum of the costs from nuclear import and export as well as travel along the microtubules, the cost of intracellular transport for an influenza infection, $E_{Transit/i}$, can be approximated as 6 x 10^6 P_D.

The cost of GTPases mediating endosome docking and fusion is likely negligible compared to the heavy cost of motor proteins mobilizing the endosomes. This is because during endosome trafficking, there are far more steps taken by motor proteins than there are endosome fusion or docking events. Yet, each fusion or docking event incurs a similar cost to a step taken by a motor protein. More detailed estimates of endosomal trafficking are exceedingly difficult due to the unavailability of studies that shed light on 1) the average number of proteins carried per endosome, 2) the average number of GTPases needed per endosome, and 3) the average length that an endosome travels within the cell, among other topics.

3.7.E Viral genome replication cost

The direct cost. Genome replication is a complex phenomenon requiring many different steps, such as the unwinding of the parent helix, RNA primer synthesis, Okazaki fragment ligation, and proofreading. Though there are many facets of genome replication in cells that require energy expenditure, the direct cost of replication lies primarily in the direct synthesis cost of nucleotides from precursor metabolites as well as the polymerization of individual nucleotides. The energetic cost of genome replication for a virus with a double-stranded DNA genome can be approximated as

$$E_{REP(dsDNA)/v} \approx 2L_g(e_d + e_p). \quad [3]$$

Here, L_g corresponds to the genome length and is multiplied by two to account for T4's double-stranded genome. The cost of each DNA nucleotide can be stated as the sum of e_d, which represents the average direct cost of DNA synthesis from precursor metabolites, and e_p, which denotes the cost of chain elongation per base (equivalent to 2 P_D, (19)) (Figure 1, step 5). The energetic cost of replicating a -ssRNA genome is similarly

$$E_{REP(-ssRNA)/v} \approx 2L_g(e_r + e_p), \quad [4]$$

where e_r represents the average direct cost of RNA synthesis from precursor metabolites. The factor of two stems from the fact that a -ssRNA has to first be converted to +ssRNA before a second copy of -ssRNA can be synthesized (Figure 2, step 8). e_d in the context of bacterial metabolism is equivalent to 11 P_D (SI Figure 1, SI Dataset 3). The average synthesis cost of an RNA base from a precursor metabolite in the context of eukaryotic metabolism is 10 P_D (SI Figure 1, SI Dataset 3). With these values in hand, we estimate the direct cost of T4

($L_g \approx 2 \times 10^5$) and influenza ($L_g \approx 1 \times 10^5$) genome replication to be ≈4 x 10^6 P_D and ≈3 x 10^5 P_D, respectively.

The opportunity cost. The opportunity cost of T4 phage replication can be estimated as $2L_g e_{od}$, where e_{od} corresponds to the average opportunity cost of a DNA base synthesized in bacteria and is approximately 34 P_O (SI Figure 1, SI Dataset 3). Note, e_{od} represents the sum of the average opportunity cost of precursor metabolites required for DNA synthesis (33 P_O) and the average opportunity cost of DNA synthesis from precursor metabolites (1 P_O). Under this estimate, the opportunity cost of genome replication for a T4 phage is ≈1 x 10^7 P_O. Similarly, the opportunity cost of genome replication for an influenza virus can be estimated as $2L_g e_{or}$, where e_{or} refers to the average opportunity cost of an RNA nucleotide synthesized in a eukaryotic organism, which is ≈ 39 P_O (SI Figure 1, SI Dataset 3). The opportunity cost of influenza genome replication is therefore ≈8 x 10^5 P_O.

The total cost. The total cost of T4 phage replication is the sum of opportunity and direct costs: $2L_g(e_d + e_p + e_{od})$. Under this estimate, the cost of genome replication for a T4 phage is ≈2 x 10^7 P_T. The same logic follows for influenza where the total cost of its genome replication can be estimated as $2L_g(e_d + e_p + e_{or})$. The total cost of influenza genome replication is therefore ≈1 x 10^6 P_T.

The replication cost component of an infection. Further, the replication cost per infection is the cost of replication for a virus multiplied by its burst size, B, resulting in

$$E_{REP/i} \approx BE_{REP/v}. \quad [5]$$

The direct, opportunity and total costs of genome replication during a T4 infection with an average burst size of 200 are ≈9 x 10^8 P_D, ≈ 2 x 10^9 P_O and ≈ 3 x 10^9 P_T (Figure 1, step 5).

The direct, opportunity and total costs of genome replication during an influenza infection with an average burst size of 6000 are ≈ 2 x 10^9 P_D, ≈ 5 x 10^9 P_O and ≈ 6 x 10^9 P_T (Figure 2, step 8).

3.7.F Viral transcriptional cost

As in the case of replication, transcription involves various energy-consuming processes, such as transcriptional activation, initiation and termination, as well as proofreading and mRNA splicing. However, the dominant cost in transcription is similar to that in replication, namely the investment in the synthesis of the nucleotides themselves (19). See Figure 1 (step 3) and Figure 2 (step 6).

The direct cost. The direct cost of transcription is approximately

$$E_{TX/v} \approx (e_{er} + e_r) \sum_{i=1}^{m} L_{ri} N_{ri}, \quad [6]$$

where L_{ri} and N_{ri} represent the length and the copy number of each viral mRNA transcript, respectively. The average direct cost of synthesizing an RNA base from precursor metabolites is denoted as e_r, which is 10 P_D (SI Figure 1, SI Dataset 3). The direct cost of polymerizing an RNA base is symbolized by e_{er}, which is equal to 2 P_D (19). Because the mRNA copy number for each viral gene is largely an unknown parameter, we approximated the viral genome as one long gene. This allowed us to eliminate the L_r and N_r dependence, and replace them with constants, l_r and n_r, such that

$$E_{TX/v} \approx (e_{er} + e_r) l_r n_r. \quad [7]$$

The constant l_r, corresponds to the length of the mRNA transcript, and thus is equal to the length of the genome. The average copy number of this transcript, n_r, can be approximated by the observed ratio of 1 mRNA transcript per 1,000 resulting proteins both in prokaryotes

(18) as well as in mammalian cells (59). The average protein copy number of a virus, n_p, can be related to its average transcript number according to $n_r \approx \frac{n_p}{1000}$.

To obtain n_p we have used available data on viral protein copy numbers. For a T4 phage, with a total estimated protein count of 5,000 representing 35 structural and 3 lysis genes (SI Table 1), the average protein copy number per virus is ≈130. Influenza's total protein count is also approximately 5,000, representing the products of 9 proteins (SI Table 2). This results in an estimate for the average protein copy number for an influenza virus of ≈550. With these numbers in hand, we estimate n_r to be approximately 0.13 for a single T4 phage and 0.55 for an influenza virus. We note that generally mRNAs are relatively short-lived and each individual mRNA on average produces between 10 and 100 proteins. However, the pool of such proteins is then used to synthesize all the virus particles that make up a given burst.

While it is sufficient to consider only the synthesis and polymerization costs of nucleotides in the direct cost estimates for genome replication, the transcriptional cost should additionally encompass the cost of mRNA re-polymerization (19). This is because mRNA transcripts have a short lifespan and must be regenerated throughout the duration of the infection. This cost is perhaps more prominent in bacteria, where the lifetime of a transcript is about 3 minutes (18), whereas the lifetime of a mammalian cellular transcript is approximately 10 hours (18), and therefore comparable to the lifetime of the influenza infection itself. The first step in calculating the mRNA re-polymerization cost is to multiply the lifetime of the infection, t, by the mRNA degradation rate, δ_r.

The second step is to take into account the average transcript copy number and its length to determine the number of RNA bases that have to be re-polymerized during the

course of an infection. Effectively, the assumption is that the RNA nucleotides are only being re-polymerized, not re-synthesized. The re-polymerization cost of transcription can be stated as the number of nucleotides to be re-polymerized, $l_r n_r \delta_r t$, multiplied by e_{er}. The direct cost of transcription per virus can then be revised as such:

$$E_{TX/v} \approx l_r n_r (e_r + e_{er} \delta_r t). \quad [8]$$

The lifetime of a T4 infection is ≈30 minutes (60), and the lifetime of an influenza infection is roughly 12 hours (14). With these parameters in hand, the direct cost of transcription (per virus) for T4 and influenza are ≈7 x 10^5 P_D and ≈7 x 10^4 P_D, respectively (Table 1).

The opportunity cost. The opportunity cost of transcription can be obtained by $l_r n_r e_{or}$, where e_{or} represents the opportunity cost of an RNA nucleotide. e_{or} is approximately 31 P_O in bacteria and 39 P_O in eukaryotes (SI Dataset 3, SI Figure 1). Note, that in this expression we don't account for re-polymerization events as RNA nucleotides are assumed to be recycled rather than resynthesized. The opportunity costs of transcription for a single T4 phage and an influenza virus are thus ≈7 x 10^5 P_O and ≈2 x 10^5 P_O, respectively.

The total cost. The total cost of transcription can be obtained by $l_r n_r (e_{or} + e_r + e_{er} \delta_r t)$, which represents the sum of opportunity and direct costs of transcription. According to this estimate, the total costs of transcription for a single T4 phage and an influenza virus are ≈1 x 10^6 P_T and ≈3 x 10^5 P_T, respectively (Table 1).

The transcriptional cost of an infection. The transcriptional cost of an infection is the transcriptional cost of a virus multiplied by its burst size, namely,

$$E_{TX/i} \approx B E_{TX/v}. \quad [9]$$

For T4 with an average burst size of 200 and for influenza with an average burst size of 6000, the direct cost of transcription at the level of an infection is ≈1 x 10^8 P_D (Figure 1, step 3) and ≈4 x 10^8 P_D, respectively (Figure 2, step 6). The opportunity cost of transcription for these two infections are ≈1 x 10^8 P_O (T4, Figure 1, step 3) and ≈4 x 10^8 P_O (influenza, Figure 2, step 6). Their total costs are ≈3 x 10^8 P_T (T4, Figure 1, step 3) and ≈2 x 10^9 P_T (influenza, Figure 2, step 6) (Table 1).

3.7.G Viral translational cost

There are important biosynthetic costs associated with proteins just as there are with nucleic acids. Here, we attempt to capture the most significant costs in the protein synthesis pathway while making some simplifying assumptions that neglect substantially smaller cost components such as the costs of translational initiation and termination and post-translational modifications (19). The estimate for translational cost follows the same rationale as the cost calculation for transcription.

The direct cost. The direct cost of translating the viral proteome can be estimated as

$$E_{TL/v} \approx (e_a + e_{ea}) \sum_{j=1}^{k} L_{pj} N_{pj}. \quad [10]$$

Here we have multiplied the total number of amino acids by the per-amino acid costs of synthesis from precursor metabolites, e_a, and polymerization, e_{ea}. The arrays L_p and N_p hold values for the length of each protein and its copy number per virus, respectively (SI table 1, SI table 2). We show that $\sum_{j=1}^{k} L_{pj} N_{pj}$ for the influenza virus and the T4 phage are about 1.7 and 1.2 million amino acids, respectively (SI table 1, SI table 2). In both bacteria and eukaryotes, e_a, is on average equal to 2 P_D (SI Figure 1, SI Dataset 3) and e_{ea} is 4 P_D

(21). Due to the relatively slow protein degradation rates for bacteria ($\frac{0.4}{hr}$) and human cells ($\frac{0.08}{hr}$) (19, 22) compared to infection lifetimes, we have neglected costs associated with this process. Using this information, the direct cost of translation for a T4 phage and an influenza virus are ≈7 x 10^6 P_D and ≈1 x 10^7 P_D, respectively (Table 1). This finding stems mainly from the fact that both viruses are based on ≈10^6 amino acids.

SI Table 1. T4 bacteriophage structural proteins and their average copy numbers per virion. The number of amino acids comprising each virion is calculated by the product of the average protein copy number and the length of the corresponding protein.

Protein	Protein length (L_{pj})	Average protein copy number (N_{pj})	Total no. of amino acids in protein j ($L_{pj}N_{pj}$)
23*	521	930	484,530
20*	524	12	6,288
24*	427	55	23,485
soc*	80	840	67,200
hoc*	376	160	60,160
22*	269	576	154,944
21*	212	72	15,264
IPIII*	194	370	71,780
IPI*	95	360	34,200
IPII*	100	360	36,000
alt*	682	40	27,280
68*	141	240	33,840
67*	80	341	27,280
3†	176	6	1,056
53†	196	6	1,176
5†	575	3	1,725
6†	660	12	7,920
7†	1,032	6	6,192
8†	334	12	4,008
9†	288	18	5,184
10†	602	18	10,836
11†	219	18	3,942
12†	527	18	9,486
15†	272	6	1,632
18†	659	144	94,896
19†	163	144	23,472
25†	132	6	792
26†	208	Assumed 1	208
27†	391	3	1,173
28†	177	Assumed 1	177
29†	590	6	3,540
48†	364	6	2,184
54†	320	6	1,920
td†	286	3	858
frd†	193	6	1,158
holin‡	218	20	4,360
endolysin‡	164	20	3,280
spanin‡	216	20	4,320
Totals		$\sum_{j=1}^{k} N_{pj}$ = 4,805 proteins	$\sum_{j=1}^{k} L_{pj}N_{pj}$ = 1,225,786 aa

The number of amino acids comprising each virion is calculated by the product of the average protein copy number and the length of the corresponding protein.

*These genes together compose the phage head.

†These genes are those that make up the tail tube, the tail sheath, and the base plate (table modified from ref. 85).

‡These genes are those that are involved in lysis (76).

SI Table 2. Influenza A virus proteins and their average copy numbers per virion. The number of amino acids comprising each virion is calculated by the product of the average protein copy number and the length of the corresponding protein (Table modified from (61)).

RNA segment lengths (no. of nucleotides)	Protein product	Protein length (L_{pj})	Average protein copy number (N_{pj})	Total no. of amino acids in protein j ($L_{pj}N_{pj}$)
1 (2,341)	Polymerase PB2	759	45	34,155
2 (2,341)	Polymerase PB1	757	45	34,065
3 (2,233)	Polymerase PA	716	45	32,220
4 (1,778)	Hemagglutinin	566	500	283,000
5 (1,565)	Nucleoprotein	498	1,000	498,000
6 (1,413)	Neuraminidase	454	100	45,400
7 (1,027)	Matrix protein M1	252	3,000	756,000
	Matrix protein M2	97	40	3,880
8 (890)	NS1	230	0	0
	NS2	121	165	19,965
Totals			$\sum_{j=1}^{k} N_{pj}$ = 4,940 proteins	$\sum_{j=1}^{k} L_{pj}N_{pj}$ = 1,706,685 aa

The opportunity cost. The opportunity cost of viral translation is approximately $e_{oa} \sum_{j=1}^{k} L_{pj}N_{pj}$, where e_{oa} denotes the average opportunity cost of an amino acid, and corresponds to 25 P_O in bacteria and 30 P_O in eukaryotes (SI Figure 1, SI Dataset 3). The opportunity cost of viral translation for a T4 phage and an influenza virus are therefore ≈3 x 10^7 P_O and ≈5 x 10^7 P_O, respectively.

The total cost. The total cost of viral translation is the sum of direct and opportunity costs of translation. The total cost of viral translation for a T4 phage and an influenza virus are therefore ≈4 x 10^7 P_T and ≈6 x 10^7 P_T, respectively (Table 1).

The translational cost component of an infection. The translational cost of an infection is simply the cost of translation per virion multiplied by its burst size, namely,

$$E_{TL/i} \approx BE_{TL/v}. \quad [11]$$

The direct, opportunity and total translational costs for a T4 infection with a burst size of 200 are ≈1 x 10^9 P_D, 6 x 10^9 P_O and 8 x 10^9 P_T (Figure 1, step 4). For an influenza infection with a burst size of 6000, these costs are ≈ 6 x 10^{10} P_D, 3 x 10^{11} P_O and 4 x 10^{11} P_T (Figure 2, step 7)(Table 1).

Protein folding and quality control. Just as in any other biological process, protein folding is subject to errors. To correct for such errors and prevent the aggregation of misfolded proteins, cells from all three domains of life have evolved elaborate mechanisms for the detection of misfolded proteins. Through various ATP-dependent (e.g. Hsp90, Hsp70, and Hsp60) and ATP-independent processes, a triage is carried out in which some proteins are re-folded and others are degraded (62-64). From an energetic standpoint, protein quality control mechanisms are likely to cost substantially less than the cost of translation. As shown above, a protein with an average length of 300 amino acids (22) will have a direct translational cost of 1,800 P_D and a total cost of 9,300 P_T in bacteria and 10,800 P_T in eukaryotes, respectively. On the other hand, the energetic cost of various protein quality control pathways can range from a few ATPs (65) to a few hundred ATPs per protein (66), which is required for protein degradation. Thus, it is likely that protein quality control will be a fractional cost compared to the translational cost of a protein. A similar conclusion was drawn in the context of cellular protein cost (19). Because we were unable to ascertain the fraction of viral proteins that may be degraded, and because different proteins require different quality control pathways (67), any more detailed estimates are difficult to make. Future experimental studies would be needed to determine any substantial costs incurred by protein quality control or translation at large that may be missing from our estimates.

Another very interesting experimental avenue would be to explore the consequences of the quality control mechanisms as they are being partially recruited towards maintaining the viral proteome. Particularly in the context of a host cell that survives the infection, how the cell responds to the additional burden from viral protein production and maintenance would be a fascinating topic of study. Constructed from roughly 5000 proteins (SI Table 2), and with an average burst size of about 6000, the influenza infection will produce about 3 x 10^7 viral proteins. Considering that a human cell will harbor more than 10^9 proteins (68), we would expect the extra load from viral proteins on the quality control machinery to be minimal. For a T4 phage, composed of roughly 5000 proteins (SI Table 1) and with a burst size of roughly 200, the total number of viral proteins during an infection would be approximately 10^6. This is comparable to the number of bacterial proteins, which is estimated to be 10^6 per cubic micron (69), or the approximate volume of an *E. coli* cell. As such, in the case of a T4 infection, the viral quality control pathways are likely to more heavily affected than in the case of an influenza infection.

3.7.H Assembly and genome packaging cost

Upon translation, the influenza viral proteins and ribonucleoproteins travel towards the cell membrane via Rab-bound endosomes that are carried by kinesins on microtubules. Interactions of the matrix protein M1 with ribonucleoproteins and the viral transmembrane proteins, namely HA, NA, and M2, result in the assembly of the influenza virus (70). Though the kinetics of the assembly steps remain to be delineated, influenza virus assembly and genome packaging are not regarded as energy-consuming processes. In general, virus assembly is described as an energetically favorable process, typically driven by the burial of hydrophobic surfaces (71, 72), and therefore independent of host energy expenditure. As an

example, the assembly of hepatitis B virus is shown to occur spontaneously through weak protein-protein interactions (73).

While the assembly of the T4 capsid is spontaneous, the packaging of the genome inside the capsid is not (Figure 1, step 7). The cost of genome packaging for a T4 phage is

$$E_{Pack/v} = e_p L_g, \quad [12]$$

where the cost to package a base pair, e_p, is 2 P_D (74). For the 169 kb genome of T4, this cost is ≈3 x 10^5 P_D.

Genome packaging cost of an infection. The packaging cost of a T4 infection is simply the cost of packaging for a single T4 phage multiplied by the T4 burst size:

$$E_{Pack/i} = BE_{Pack/v}. \quad [13]$$

For T4, with a burst size of 200, the contribution of packaging to the total cost of infection is ≈7 x 10^7 P_D (Figure 1, step 7) (Table 1).

3.7.I Viral exit cost

Viruses use two primary exit strategies. Generally, enveloped viruses such as influenza and HIV bud off from the host membrane. Phages, on the other hand, generally lyse their host cells. For T4, the cost of exit is primarily the production cost of proteins that together break down the cell wall. We have already included the cost of lysis proteins in our translational cost estimates. The lysis proteins include holin, endolysin, and spanin proteins. The holins create holes in the host inner membrane, enabling the endolysins to reach the peptidoglycan layer. The spanins fuse the inner and the outer membrane as a requirement for lysis of gram-negative bacteria (Figure 1, step 9, only holins and endolysins are shown). Considering that T4 holin, endolysin, and spanin proteins are 218, 164, and 216 amino acids

in length, respectively, and each have a copy number of about 4,000 per infection (or approximately 20 per virus considering a burst size of 200) (75), the contribution of lysis proteins to the translational cost of an infection is negligible.

In the case of influenza, the exocytosis of virions is energy-consuming. However, the exact mechanism remains a mystery, with influenza M2 protein so far serving as the most likely agent for mediating exocytosis (76). Three separate costs exist: 1) the cost to locally bend the membrane outward, 2) the cost to scisse the budding virion from the cell membrane, and the cost of the cellular membrane that becomes part of the viral membrane (Figure 2, step 10). The cost to bend the membrane into the shape of a sphere of any size, is equivalent to 25 P_D (21). One way to estimate the cost of scission is to assume that it incurs a comparable cost to the HIV scission process, which, similar to several other enveloped viruses, is mediated through the ESCRT-III (endosomal sorting complexes required for transport) assembly. ESCRT-III complex self-assembles into filaments around the neck of a budding vesicle (similar to dynamin), and its disassembly requires the expenditure of 6 P_D via the Vps4 ATPase (77). The sum of these two costs results in the usage of 31 P_D as one influenza virus leaves the cell. Because 6000 virions exit the cell on average, exocytosis costs approximately 2×10^5 P_D.

An alternative, order-of-magnitude estimate could be made with the assumption that the cost of membrane scission during endocytosis equals the cost of membrane scission during exocytosis. In estimating the influenza entry cost, we showed that the cost of membrane scission during influenza endocytosis via the dynamin polymer is $\approx$30 P_D. Together with the cost of membrane bending, $\approx$25 P_D, this exit estimate is slightly higher (55 P_D per virus) than the previous (31 P_D per virus). Under this estimate, the cost of exocytosis for all 6000 influenza virions amounts to $\approx 3 \times 10^5$ P_D.

The primary cost of viral exit for influenza, however, is the cost of lipids that are taken from the host cell to form the viral membrane. The cost of lipids per virus can be estimated by the number of lipid molecules needed per virion multiplied by the cost of a lipid molecule, e_l:

$$E_{Exit/v} = \frac{8\pi r^2}{s}\, e_l. \quad [14]$$

In SI Eq. 14 the numerator represents twice the viral surface area (accounting for the bilayer) and the denominator, *s*, denotes the surface area of a lipid head group, which is approximately 0.5 nm^2 (22, 78). With an average radius of 50 nm, the influenza virus is comprised of $\approx 10^5$ lipid molecules. The direct, opportunity and total cost of a lipid molecule in a eukaryotic cell are 18 P_D, 264 P_O and 282 P_T, respectively (SI Dataset 3, SI Figure 1). As a result, the direct, opportunity and total costs of lipids per influenza virus are 2 x 10^6 P_D, 3 x 10^7 P_O and 4 x 10^7 P_T, respectively. The exit costs of an influenza infection can be derived by

$$E_{Exit/i} = BE_{Exit/v}, \quad [15]$$

and are approximately 1 x 10^{10} P_D, 2 x 10^{11} P_O and 2 x 10^{11} P_T for an infection with a burst size of 6000 (Table 1).

3.7.J Estimating the total host energy budget

The total basal and growth metabolic requirements of various organisms have been shown to correlate with the cellular volume (19). We have used SI Eqs. 16-18 presented by Lynch and Marinov (19) to estimate the energetic budget of hosts considered in this study. Basal metabolic requirement of a cell scales with the cell volume according to

$$E_M = 0.39V^{0.88}, \quad [16]$$

where V, represents the cell volume in units of μm^3, and E_M is in the units of 10^9 ATP per cell per hour. The growth metabolic requirement of a cell similarly scales with the cell volume according to

$$E_G = 27V^{0.97}, \quad [17]$$

where E_G is in the units of 10^9 ATP per cell. The total energy requirement of a cell is simply the sum of the basal and growth energy requirements,

$$E_T = E_G + tE_M, \quad [18]$$

where t is the cell-division time in the units of hours.

For a bacterial cell with a volume of 1 μm^3, the maintenance metabolic cost is $\approx 10^8$ P_T in the duration of a T4 phage infection which lasts about 30 minutes. A mammalian cell, on the other hand, with a characteristic volume of 2000 μm^3, has a basal metabolic cost of $\approx 10^{12}$ P_T over the course of a 12-hour influenza infection. The total energetic cost of a cell should also encompass the cost of cellular growth. The total energetic cost of a bacterium and a mammalian cell with the dimensions highlighted above are therefore $\approx 3 \times 10^{10}$ P_T and $\approx 5 \times 10^{13}$ P_T, respectively, during the course of their viral infection. The correlation between cellular volume and metabolic capacity is supported by the observation that larger *E. coli* cells produce higher T4 burst sizes (79).

3.7.K Heat production and power consumption of a viral infection

In our estimates for heat production and power consumption of a viral infection, we will not consider the total cost of an infection as it contains the opportunity costs; by definition, these opportunity costs do not represent direct usage of ATP (and ATP-

equivalent) molecules; rather, they represent the ATP (and ATP-equivalent) molecules that could have been generated in the absence of a viral infection. For these estimates we will rely on the direct costs.

To estimate the amount of heat that is generated due to a viral infection, we have to consider the inefficiency of aerobic metabolism. The burning of glucose results in the production of approximately 2800 kJ/mol of heat (80). The same reaction takes place inside our cells, with the difference being that cells are capable of harnessing a fraction of the free energy that would otherwise be liberated as heat. When glucose is aerobically metabolized into water and carbon dioxide, a fraction of the free energy is used to convert ADP into ATP, while the remaining free energy is dissipated as heat. By assuming physiological conditions, the free energy change of ATP hydrolysis can be approximated as -50kJ/mol (16). In bacterial metabolism, 26 ATPs are generated from each glucose molecule; hence the free energy captured by the conversion of ADP into ATP is approximately -1300 kJ/mol of glucose. As a result, in this simple estimate, we consider that about 50% of the energy from the aerobic metabolism of glucose is dissipated as heat: $\left(1 - \frac{1300\frac{kJ}{mol}}{2800\frac{kJ}{mol}}\right) \times 100\%$. For eukaryotic cells, with 32 ATPs generated per glucose molecule, about 40% of the energy stored in glucose is dissipated as heat.

The T4 infection has a direct cost of 3 x 10^9 P_D (Table 1). Because each glucose molecule results in 26 ATPs in the bacterial metabolism, T4 infection's direct cost would require the complete metabolism of 10^8 glucose molecules. The influenza infection's direct cost is about an order of magnitude higher than that of T4 (Table 1), and is equivalent to the aerobic metabolism of approximately 10^9 glucose molecules. Based on the number of glucose molecules required to cover the direct costs of each infection, the free energy stored in

glucose (-2800 kJ/mol), and the percentage of the energy released as heat during the aerobic metabolism of glucose (≈ 40-50%), we can conclude that the heat generated during T4 and influenza infections are approximately 0.2 nJ and 2 nJ, respectively.

In half an hour, the T4 infection results in the hydrolysis of ATP-equivalent molecules at an average rate of 2 x 10^6 P_D per second. In half a day, an influenza infection also has an average ATP-hydrolysis rate of 2 x 10^6 P_D per second. Put in terms of the more familiar units of Watts (by assuming -50kJ/mol of free energy change per P_D), the power of both viral infections is on the order of 200 fW.

3.7.L Generalizing viral energetics for double-stranded DNA phages

Figure 5A.1 shows how we can generalize the estimates presented here by thinking of dsDNA phages as approximately spherical objects with an outer layer of thickness, *t*. In this model, the inner volume of a virus containing the viral genome is given by $\frac{4\pi r^3}{3}$, which can be used to estimate the viral genome length (24). The total cost of genome replication for a double-stranded DNA genome can be obtained from SI Eq. 3 (Figure 4). However, instead of using the viral genome length directly, we can divide the capsid inner volume by the volume of a base pair, v_d (approximately 1 nm^3) (22):

$$E_{REP(dsDNA)/v} \approx \frac{4\pi r^3}{3v_d}\left(e_d + e_p + e_{od}\right). \quad [19]$$

Because for many dsDNA phages only about half of the capsid is filled with the viral genome (24), the cost of a DNA base was not multiplied by a factor of 2 (even though the genome is double stranded) as the two multipliers cancel each other out. The direct cost of replication can be obtained similarly by the exclusion of the opportunity cost of a DNA base or e_{od} (in bacteria) from SI Eq. 19. Moreover, the translational cost of a virus can be

obtained from a modification of $(e_a + e_{ea} + e_{oa}) \sum_{j=1}^{k} L_{pj} N_{pj}$, derived previously (SI section G). The total cost of viral translation,

$$E_{TL/v} \approx \frac{4\pi(R^3 - r^3)}{3v_a} (e_a + e_{ea} + e_{oa}), \quad [20]$$

can be obtained from multiplying the total number of amino acids by the total cost of an amino acid. The number of amino acids is estimated by dividing the outer capsid volume (denoted by the shaded blue region in Figure 5A), $\frac{4\pi(R^3 - r^3)}{3}$, by the volume of an amino acid, v_a, which can be approximate as 0.1 nm^3 (81). This expression can be further simplified by replacing the outer radius, R, with $r + t$ (Figure 4):

$$E_{TL/v} \approx \frac{4\pi((r+t)^3 - r^3)}{3v_a} (e_a + e_{ea} + e_{oa}) \approx \frac{4\pi r^2 t}{v_a} (e_a + e_{ea} + e_{oa}). \quad [21]$$

The direct translation cost of a virus can be similarly obtained from Eq. 21 by excluding the average opportunity cost of an amino acid, e_{oa}. The critical radius, r_{crit}, at which translation and replication will have equal cost can be obtained by setting Eqs. 19 and 21 equal and solving for r (Figure 4). Because capsid shell thickness is relatively conserved across icosahedral viruses studied to date, it can be treated as a constant equal to 3 nm (82). The critical radius for total cost estimates, $r_{crit-Tot}$, is 59 nm. For the direct cost estimates, the critical radius, $r_{crit-Dir}$, is 42 nm (Figure 4).

SI Dataset 1. A list of viruses and their associated costs used to estimate replication to translation cost ratios shown in Figure 4. Provided in a GitHub repository: https://github.com/gitamahm/thesis

SI Dataset 2. A list of direct and total fractional cost estimates, E_g, for genetic elements of lengths 1, 10, 100, 1000, and 10,000 base pairs across 30 dsDNA viruses (Figure 5). Genetic elements are assumed to have no functional benefit to the virus and to be non-transcribed. Virus A, B, and C correspond to hypothetical viruses. Provided in a GitHub repository: https://github.com/gitamahm/thesis

SI Dataset 3. A detailed breakdown of opportunity and direct costs of building blocks across heterotrophic bacterial and eukaryotic metabolisms (using glucose as the sole carbon source). Provided in a GitHub repository: https://github.com/gitamahm/thesis

3.8 References

1. Kutter E & Sulakvelidze A (2004) *Bacteriophages: biology and applications* (CRC Press).
2. Rosenwasser S, Ziv C, van Creveld SG, & Vardi A (2016) Virocell Metabolism: Metabolic Innovations During Host–Virus Interactions in the Ocean. *Trends in Microbiology* 24(10):821-832.
3. Lindell D*, et al.* (2007) Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* 449(7158):83-86.
4. Thai M*, et al.* (2015) MYC-induced reprogramming of glutamine catabolism supports optimal virus replication. *Nature communications* 6.
5. Chang C-W, Li H-C, Hsu C-F, Chang C-Y, & Lo S-Y (2009) Increased ATP generation in the host cell is required for efficient vaccinia virus production. *Journal of biomedical science* 16(1):1.
6. Roux S*, et al.* (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*.
7. Hadas H, Einav M, Fishov I, & Zaritsky A (1997) Bacteriophage T4 development depends on the physiology of its host Escherichia coli. *Microbiology* 143(1):179-185.
8. Ahuka-Mundeke S*, et al.* (2010) Full-length genome sequence of a simian immunodeficiency virus (SIV) infecting a captive agile mangabey (Cercocebus agilis) is closely related to SIVrcm infecting wild red-capped mangabeys (Cercocebus torquatus) in Cameroon. *Journal of General Virology* 91(12):2959-2964.
9. Middelboe M (2000) Bacterial growth rate and marine virus–host dynamics. *Microbial Ecology* 40(2):114-124.
10. Van Etten JL, Burbank DE, Xia Y, & Meints RH (1983) Growth cycle of a virus, PBCV-1, that infects Chlorella-like algae. *Virology* 126(1):117-125.

11. Golec P, Karczewska-Golec J, Łoś M, & Węgrzyn G (2014) Bacteriophage T4 can produce progeny virions in extremely slowly growing Escherichia coli host: comparison of a mathematical model with the experimental data. *FEMS microbiology letters* 351(2):156-161.

12. Kim H & Yin J (2004) Energy-efficient growth of phage Qβ in Escherichia coli. *Biotechnology and bioengineering* 88(2):148-156.

13. Birch EW, Ruggero NA, & Covert MW (2012) Determining host metabolic limitations on viral replication via integrated modeling and experimental perturbation. *PLoS Comput Biol* 8(10):e1002746.

14. Stray SJ & Air GM (2001) Apoptosis by influenza viruses correlates with efficiency of viral mRNA synthesis. *Virus research* 77(1):3-17.

15. Delbrück M (1945) The burst size distribution in the growth of bacterial viruses (bacteriophages). *Journal of bacteriology* 50(2):131.

16. Berg JM, Tymoczko JL, & Stryer L (2012) *Biochemistry* 7th Ed.

17. Voet D & Voet JG (2011) Biochemistry, 4-th Edition. *New York: John Wiley& Sons Inc*:492-496.

18. Milo R & Phillips R (2015) *Cell biology by the numbers* (Garland Science).

19. Lynch M & Marinov GK (2015) The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences* 112(51):15690-15695.

20. Kaleta C, Schäuble S, Rinas U, & Schuster S (2013) Metabolic costs of amino acid and protein production in Escherichia coli. *Biotechnology journal* 8(9):1105-1114.

21. Phillips R, Kondev J, Theriot J, & Garcia H (2012) *Physical biology of the cell* (Garland Science).

22. Milo R, Jorgensen P, Moran U, Weber G, & Springer M (2010) BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic acids research* 38(suppl 1):D750-D753.

23. Djamali E, Nulton JD, Turner PJ, Rohwer F, & Salamon P (2012) Heat output by marine microbial and viral communities.

24. Jover LF, Effler TC, Buchan A, Wilhelm SW, & Weitz JS (2014) The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nature Reviews Microbiology* 12(7):519-528.

25. Shepherd CM*, et al.* (2006) VIPERdb: a relational database for structural virology. *Nucleic acids research* 34(suppl 1):D386-D389.

26. Brum JR, Schenck RO, & Sullivan MB (2013) Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *The ISME journal* 7(9):1738-1751.

27. Wagner A (2005) Energy constraints on the evolution of gene expression. *Molecular biology and evolution* 22(6):1365-1374.

28. Koonin EV (2015) Energetics and population genetics at the root of eukaryotic cellular and genomic complexity. *Proceedings of the National Academy of Sciences* 112(52):15777-15778.

29. Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* 10(3):195-205.

30. Novitsky V, Wang R, Lagakos S, & Essex M (2010) HIV-1 subtype C phylodynamics in the global epidemic. *Viruses* 2(1):33-54.

31. Bedford T, Cobey S, & Pascual M (2011) Strength and tempo of selection revealed in viral gene genealogies. *BMC Evolutionary Biology* 11(1):1.

32. Van Etten JL, Graves MV, Müller DG, Boland W, & Delaroque N (2002) Phycodnaviridae–large DNA algal viruses. *Archives of virology* 147(8):1479-1516.

33. Maynard ND, Gutschow MV, Birch EW, & Covert MW (2010) The virus as metabolic engineer. *Biotechnology journal* 5(7):686-694.

34. Anantharaman K*, et al.* (2014) Sulfur oxidation genes in diverse deep-sea viruses. *Science* 344(6185):757-760.

35. Roux A (2014) Reaching a consensus on the mechanism of dynamin? *F1000prime reports* 6.

36. Zeng Q & Chisholm SW (2012) Marine viruses exploit their host's two-component regulatory system in response to resource limitation. *Current Biology* 22(2):124-128.

37. Lauring AS, Frydman J, & Andino R (2013) The role of mutational robustness in RNA virus evolution. *Nature Reviews Microbiology* 11(5):327-336.

38. Brooke CB*, et al.* (2013) Most influenza a virions fail to express at least one essential viral protein. *Journal of virology* 87(6):3155-3162.

39. Weitz JS (2016) *Quantitative Viral Ecology: Dynamics of Viruses and Their Microbial Hosts* (Princeton University Press).

40. Raoult D*, et al.* (2004) The 1.2-megabase genome sequence of Mimivirus. *Science* 306(5700):1344-1350.

41. Suhre K (2005) Gene and genome duplication in Acanthamoeba polyphaga Mimivirus. *Journal of virology* 79(22):14095-14101.

42. Mullen AR*, et al.* (2012) Reductive carboxylation supports growth in tumour cells with defective mitochondria. *Nature* 481(7381):385-388.

43. Neidhardt FC, Ingraham JL, & Schaechter M (1990) Physiology of the bacterial cell: a molecular approach.

44. Silverstein T (2005) The mitochondrial phosphate-to-oxygen ratio is not an integer. *Biochemistry and Molecular Biology Education* 33(6):416-417.

45. Dimitrov DS (2004) Virus entry: molecular mechanisms and biomedical applications. *Nature Reviews Microbiology* 2(2):109-122.

46. Rossmann MG, Mesyanzhinov VV, Arisaka F, & Leiman PG (2004) The bacteriophage T4 DNA injection machine. *Current opinion in structural biology* 14(2):171-180.

47. Chen BJ & Lamb RA (2008) Mechanisms for enveloped virus budding: can some viruses do without an ESCRT? *Virology* 372(2):221-232.

48. Maltouf AF & Labedan B (1983) Host cell metabolic energy is not required for injection of bacteriophage T5 DNA. *Journal of bacteriology* 153(1):124-133.

49. Sundborger AC*, et al.* (2014) A dynamin mutant defines a superconstricted prefission state. *Cell reports* 8(3):734-742.

50. Rothnie A, Clarke AR, Kuzmic P, Cameron A, & Smith CJ (2011) A sequential mechanism for clathrin cage disassembly by 70-kDa heat-shock cognate protein (Hsc70) and auxilin. *Proceedings of the National Academy of Sciences* 108(17):6927-6932.

51. Samji T (2009) Influenza A: understanding the viral life cycle. *Yale J Biol Med* 82(4):153-159.

52. Pinto LH & Lamb RA (2006) The M2 proton channels of influenza A and B viruses. *Journal of Biological Chemistry* 281(14):8997-9000.

53. Reck-Peterson SL*, et al.* (2006) Single-molecule analysis of dynein processivity and stepping behavior. *Cell* 126(2):335-348.

54. Kemler I, Whittaker G, & Helenius A (1994) Nuclear import of microinjected influenza virus ribonucleoproteins. *Virology* 202(2):1028-1033.

55. Eisfeld AJ, Neumann G, & Kawaoka Y (2015) At the centre: influenza A virus ribonucleoproteins. *Nature Reviews Microbiology* 13(1):28-41.

56. Chen Z & Krug RM (2000) Selective nuclear export of viral mRNAs in influenza-virus-infected cells. *Trends in microbiology* 8(8):376-383.

57. Coy DL, Wagenbach M, & Howard J (1999) Kinesin takes one 8-nm step for each ATP that it hydrolyzes. *Journal of Biological Chemistry* 274(6):3667-3671.

58. Barysch SV, Aggarwal S, Jahn R, & Rizzoli SO (2009) Sorting in early endosomes reveals connections to docking-and fusion-associated factors. *Proceedings of the National Academy of Sciences* 106(24):9697-9702.

59. Schwanhäusser B*, et al.* (2011) Global quantification of mammalian gene expression control. *Nature* 473(7347):337-342.

60. Vafabakhsh R*, et al.* (2014) Single-molecule packaging initiation in real time by a viral DNA packaging machine from bacteriophage T4. *Proceedings of the National Academy of Sciences* 111(42):15096-15101.

61. Lamb RA & Krug RM (2001) Orthomyxoviridae: the viruses and their replication, p 1487–1531. *Fields virology* 1.

62. Hartl FU & Hayer-Hartl M (2002) Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* 295(5561):1852-1858.

63. Yerbury JJ, Stewart EM, Wyatt AR, & Wilson MR (2005) Quality control of protein folding in extracellular space. *EMBO reports* 6(12):1131-1136.

64. Wickner S, Maurizi MR, & Gottesman S (1999) Posttranslational quality control: folding, refolding, and degrading proteins. *Science* 286(5446):1888-1893.

65. Sharma SK, De Los Rios P, Christen P, Lustig A, & Goloubinoff P (2010) The kinetic parameters and energy cost of the Hsp70 chaperone as a polypeptide unfoldase. *Nature chemical biology* 6(12):914-920.

66. Peth A, Nathan JA, & Goldberg AL (2013) The ATP costs and time required to degrade ubiquitinated proteins by the 26 S proteasome. *Journal of Biological Chemistry* 288(40):29215-29222.

67. Chaudhry C*, et al.* (2003) Role of the γ-phosphate of ATP in triggering protein folding by GroEL–GroES: function, structure and energetics. *The EMBO Journal* 22(19):4877-4887.

68. Wolff S, Weissman JS, & Dillin A (2014) Differential scales of protein quality control. *Cell* 157(1):52-64.

69. Milo R (2013) What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays* 35(12):1050-1055.

70. Hutchinson EC & Fodor E (2013) Transport of the influenza virus genome from nucleus to nucleus. *Viruses* 5(10):2424-2446.

71. Bahadur RP, Rodier F, & Janin J (2007) A dissection of the protein–protein interfaces in icosahedral virus capsids. *Journal of molecular biology* 367(2):574-590.

72. Katen S & Zlotnick A (2009) The thermodynamics of virus capsid assembly. *Methods in enzymology* 455:395-417.

73. Ceres P & Zlotnick A (2002) Weak protein-protein interactions are sufficient to drive assembly of hepatitis B virus capsids. *Biochemistry* 41(39):11525-11531.

74. Roos WH, Ivanovska IL, Evilevitch A, & Wuite GJL (2007) Viral capsids: mechanical characteristics, genome packaging and delivery mechanisms. *Cellular and molecular life sciences* 64(12):1484-1497.

75. Moussa SH, Kuznetsov V, Tran TAT, Sacchettini JC, & Young R (2012) Protein determinants of phage T4 lysis inhibition. *Protein Science* 21(4):571-582.

76. Rossman JS, Jing X, Leser GP, & Lamb RA (2010) Influenza virus M2 protein mediates ESCRT-independent membrane scission. *Cell* 142(6):902-913.

77. Caillat C*, et al.* (2015) Asymmetric ring structure of Vps4 required for ESCRT-III disassembly. *Nature communications* 6.

78. Brügger B*, et al.* (2006) The HIV lipidome: a raft with an unusual composition. *Proceedings of the National Academy of Sciences of the United States of America* 103(8):2641-2646.

79. Choi C, Kuatsjah E, Wu E, & Yuan S (2010) The effect of cell size on the burst size of T4 bacteriophage infections of Escherichia coli B23. *Journal of Experimental Microbiology and Immunology* 14:85-91.

80. Lodish H & Zipursky SL (2001) Molecular cell biology. *Biochemistry and Molecular Biology Education* 29:126-133.

81. Counterman AE & Clemmer DE (1999) Volumes of individual amino acid residues in gas-phase peptide ions. *Journal of the American Chemical Society* 121(16):4031-4039.

82. Božič AL, Šiber A, & Podgornik R (2013) Statistical analysis of sizes and shapes of virus capsids and their resulting elastic properties. *Journal of biological physics* 39(2):215-228.

Chapter IV

A Comprehensive and Quantitative Exploration of Thousands of Viral Genomes

4.1 Introduction

There are an estimated 10^{31} virus-like particles inhabiting our planet, outnumbering all cellular life forms (1, 2). Despite their presence in astonishing numbers and their impact on the population dynamics and evolutionary trajectories of their hosts, our quantitative knowledge of trends in the genomic properties of viruses remains largely limited with many of the key quantities used to characterize these genomes either scattered across the literature or unavailable altogether. This is in contrast to the growing ability exhibited in resources such as the BioNumbers database (3) to assemble in one curated collection the key numbers that characterize cellular life forms. Our goal has been to complement these databases of key numbers of cell biology (3-6) with corresponding data from viruses. With the advent of high-throughput sequencing technologies, recent studies have enabled genomic and metagenomic surveys of numerous natural habitats, untethering us from the organisms we know and love and giving us access to a sea of genomic data from novel organisms (7). Such advances allow us to appreciate the genomic diversity that is a hallmark of viral genomes (7-12) and now make it possible to assemble some of the key numbers of virology.

In contrast to cellular genomes, which are universally coded in the language of double-stranded DNA (dsDNA), genomes of viruses are remarkably versatile. Viral genomes can be found as single or double-stranded versions of DNA and RNA, packaged in segments or as one piece, and present in both linear and circular forms. Additionally, based

on their rapid infectious cycles, large burst sizes, and often highly error-prone replication, viruses collectively survey a large genomic sequence space, and comprise a great portion of the total genomic diversity hosted by our planet (13, 14). Recently, through a large study of metagenomic sequences, the known viral sequence space was increased by an order of magnitude (7), and much more of the viral "dark matter" likely remains unexplored (15).

In analyzing an increasing spectrum of sequence data, we are faced with a considerable challenge that is unique to viruses, namely, how to find those features within viral genomes that might reveal hidden aspects of their evolutionary history. To put this challenge in perspective, when analyzing non-viral data, universal markers from the ribosomal RNA such as 16S sequences are used to classify newly discovered organisms and to locate them on the evolutionary tree of life (16). Virus genomes on the other hand are highly divergent and possess no such universally shared sequences (17).

In the absence of universal genomic markers, viruses have historically been classified based on a variety of attributes, perhaps most notably morphological characteristics, proposed in 1962 by the International Committee on Taxonomy of Viruses or ICTV (18), or based on the different ways by which they produce mRNA, proposed by David Baltimore in 1971 (19) (see SI and Figure 1 for a detailed description of the ICTV and the Baltimore classification categories). Given the prevalence of these viral classification systems in the categorization of viruses today, it is worth remembering that their inception predates the sequencing of the first genome in 1976. With the fastest and cheapest rates of sequencing available to date, we live at an opportune moment to explore viral genomic properties and evaluate these existing classification systems in light of the growing body of sequence information.

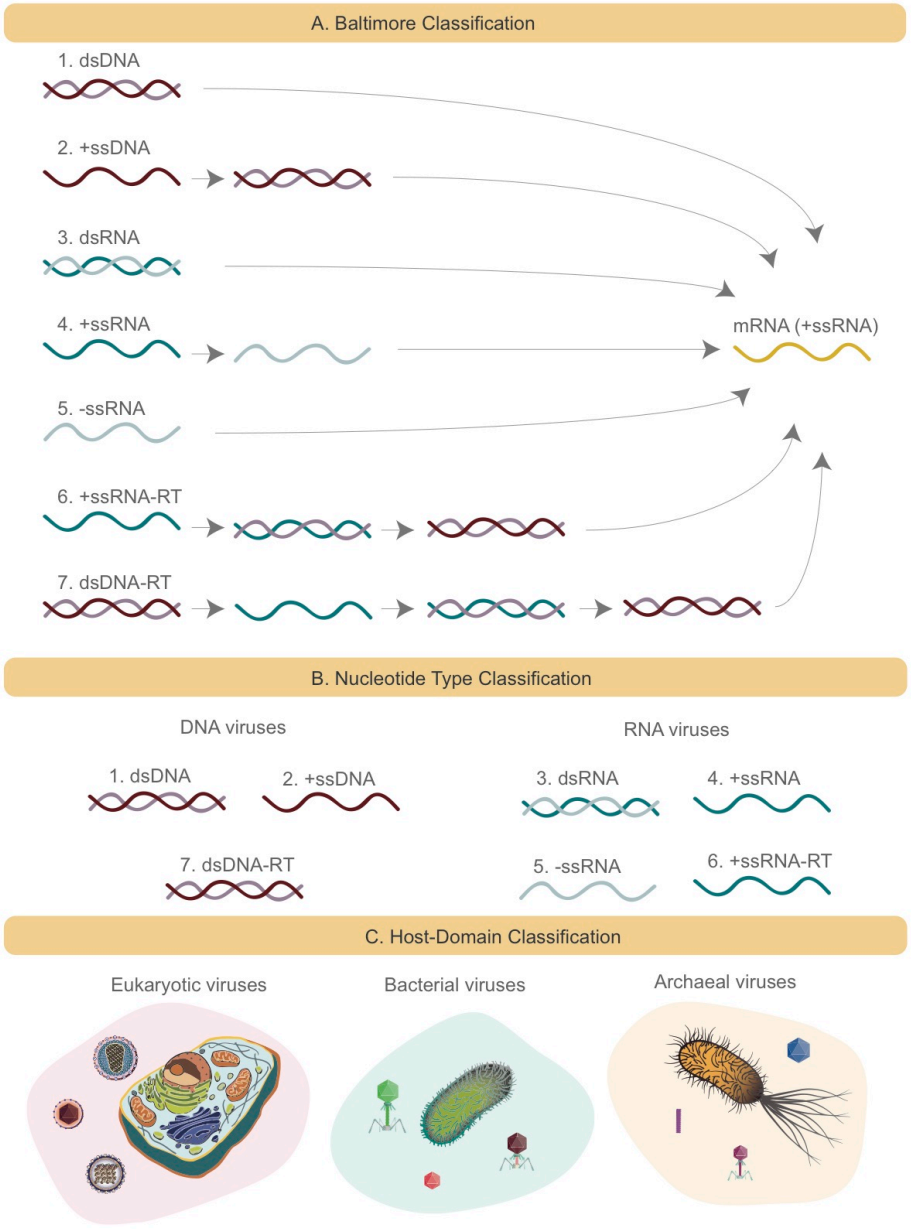


Figure 1. Schematics of several viral classification systems explored in this study. A) The Baltimore classification divides all viruses into seven groups based on how the viral mRNA is produced. DNA strands are denoted in red (+ssDNA in darker shade of red than -ssDNA). Similarly RNA strands are denoted in green (+ssRNA in darker shade of green than -ssRNA). In the case of Baltimore groups 1, 2, 6, and 7, the genome either is or is converted to dsDNA, which is then converted to mRNA through the action of DNA-dependent RNA polymerase. In the case of Baltimore groups 3, 4, and 5, the

genome is or is converted to +ssRNA, which is mRNA, through the action of RNA-dependent RNA polymerase. B) Nucleotide type classification divides viruses based on their genomic material into DNA and RNA viruses. Baltimore viral groups 1, 2, and 7 are all considered DNA viruses, and the remaining viral groups are considered RNA viruses. C) Host Domain classification groups viruses based on the host domain that they infect. Three groups are formed: eukaryotic, bacterial and archaeal viruses.

In addition to the ICTV and the Baltimore classifications (see SI) we used a simple classification system based on the host domain information, and divided viruses into bacterial, archaeal and eukaryotic viruses (Figure 1). The underpinning motivation behind this kind of classification is the Coevolution Hypothesis (20, 21). Viruses are obligate organisms unable to survive without their host, and as a corollary it is hypothesized that they have coevolved with their hosts as the hosts diverged over billions of years to form the three domains of life (20, 21). A possible piece of supporting evidence for this hypothesis is that there are to date no reported infections of hosts from one domain by viruses of another observed. We also explored a minimal classification system that divides the virus world into two groups based on their nucleotide type (RNA and DNA), here termed "Nucleotide Type" classification (Figure 1). This classification is introduced as a simplified version of the Baltimore classification system. In practice, we have assigned Baltimore groups 1, 2, and 7 to the DNA viral category, and the remaining Baltimore groups to the RNA viral category.

Although many viruses are uncharacterized, at the time of the analysis of the data presented here, there were 4,378 completed genomes available from the NCBI viral genomes resource (22) (data acquired in August, 2015). However, large-scale analyses of genomic properties for these viruses are generally unavailable. This stands in stark contrast to the in-depth analyses performed on partially assembled viral genomes or viral contigs derived from

metagenomic studies (7, 23). Although these studies have uncovered many important aspects of viral ecology with relatively little bias in sampling, they are limited by the fact that metagenomic studies typically do not result in the full assembly of genomes. An interesting example that illustrates the difficulty of complete genome assembly from metagenomic studies is the crAssphage genome, which despite taking prominent fractions of reads across various metagenomic datasets, had gone undetected and remained unassembled (24).

Without complete viral genomes, it is difficult to develop systematic understanding of key aspects of viral genomic architecture. To address this problem at least in part, we set out to provide a large-scale analysis of various genomic metrics measured from complete viral genomes. To perform a comprehensive analysis on complete viral genomes, we first explored the diversity of known viruses and their hosts within the NCBI database (see Methods). We then created distributions on a number of metrics, namely genome length, gene length, gene density, percentage of noncoding DNA (or RNA), functional gene category abundances, and gene order. We have provided brief introductions to these metrics in the following subsections.

Viral genome length, gene length and gene density. Genomes are replete with information about an organism's past and present. A central and revealing piece of information is the genome length. As more and more complete genomes have become available, we have learned that genome lengths of cellular organisms vary quite extensively, specifically by six orders of magnitude (4, 25). Because these studies focused on cellular organisms, and because genome length information is generally inaccessible through metagenomic studies, large-scale analyses that systematically capture viral genome length distributions in light of different classification systems and in relation to other genomic parameters are lacking. One such genomic parameter is the number of genes that are

encoded per genome, also referred to as gene density (26, 27). Another set of missing distributions involves gene lengths, and here too, it is important to see how they vary across different viral classification categories.

The noncoding percentages of viral genomes. One of the most surprising discoveries of the past several decades was the rich and enormous diversity of noncoding DNA in the human genome (28). Though originally thought of as "junk DNA", the noncoding regions of our genomes were later shown to be of great functional importance. Noncoding DNA is an umbrella term for very different elements, for example functional RNAs such as micro RNAs (miRNA), regulatory elements such as promoters and enhancers, and also transposons and pseudogenes.

Moreover, genomes vary widely in their noncoding percentages. While multicellular eukaryotic genomes such as plants and vertebrates have 50% or more of their genomes filled with noncoding regions, single-cell eukaryotic genomes have 25-50% of their genomes present as noncoding regions and prokaryotic genomes have even lower percentages of noncoding DNA, generally 15 to 20% (29-31). As such, the noncoding percentage of the genome is thought to correlate with the phenotypic complexity of the organism, and consequently, much of the investigation into noncoding fractions of genomes has been focused on higher eukaryotes. However, the discovery of the bacterial immunity against phages and other sources of foreign DNA, otherwise known as CRISPR/Cas system (Clustered Regularly Interspaced Short Palindromic Repeats), as well as the discovery of a new class of antibiotics targeting bacterial noncoding DNA (32) demonstrate the level of biotechnological impact and scientific insight that the study of noncoding elements in bacteria can provide. Even less is known about the noncoding fraction of viral genomes.

The literature on viral noncoding DNA or RNA is relatively sparse but highly intriguing. The first viral noncoding RNAs were discovered in adenoviruses, dsDNA viruses that infect humans, and were ~160 base pairs long (33-35). These sequences were shown responsible for viral evasion of host immunity by inhibition of protein kinase R- a cellular protein responsible for the inactivation of viral protein synthesis (36). In ovine herpesvirus, miRNAs have been shown to maintain viral latency (37). These are just several examples in which viral noncoding elements have been shown to enable viral escape from host immunity, as well as regulate viral life-cycle and viral persistence (34). Despite many interesting studies exploring the topic of cellular noncoding DNA (29-31), there are no studies, to our knowledge, that reveal the statistics of noncoding percentage of viral genomes.

Viral functional gene categories. There are detailed studies on the counts of cellular genes belonging to each broad functional category (38, 39). These studies have helped us better understand the scaling of functional categories across different clades of organisms. In fact there was an intriguing conclusion that for prokaryotic genomes, there exists a universal organization which governs the relative number of genes in each category (38). Such depictions of viral genomes, however, are largely lacking. Thus, we set out to better understand how viral genes are distributed across different functional categories and how these distributions might differ across various viral groups.

Viral genome organization. Viral genome organization is a topic that has great depth but limited breadth. There exist highly detailed genome-wide diagrams that illustrate the location, direction, and predicted function of viral genes, which are then compared to similar illustrations from a small number of viral genomes (40-43). While this highly detailed approach is indispensible for studying individual viruses, a simplified illustration of genome

organization is a requirement of any high-throughput visualization and comparison of genomes. The latter approach could help us uncover general rules governing genomic organization, in the same way that synteny, or conserved gene order, has been used to compare animal genomes (44, 45).

4.2 Exploring the NCBI viral database

We used the largest available dataset of completed viral genomes available from the National Center for Biotechnology Information (NCBI) viral genomes resource (22), containing a total of 4,378 complete viral genomes at the time of data acquisition (August, 2015). After implementing several manual and programmed steps towards curating the data, a total of 2,399 viruses (excluding satellite viruses) could be associated with a host using NCBI's documentation (see Methods). These viruses were included for further analysis, and unless noted otherwise, will constitute our dataset in this study. By examining these viruses through different classifications (Figure 2), it is clear that they are largely DNA viruses (Figure 2.B4), and more specifically, they are primarily double-stranded DNA (dsDNA) viruses (Figure 2.C4). This is in contrast to the RNA viruses, which are mostly single-stranded (Figure 2.B4 and Figure 2.C4).

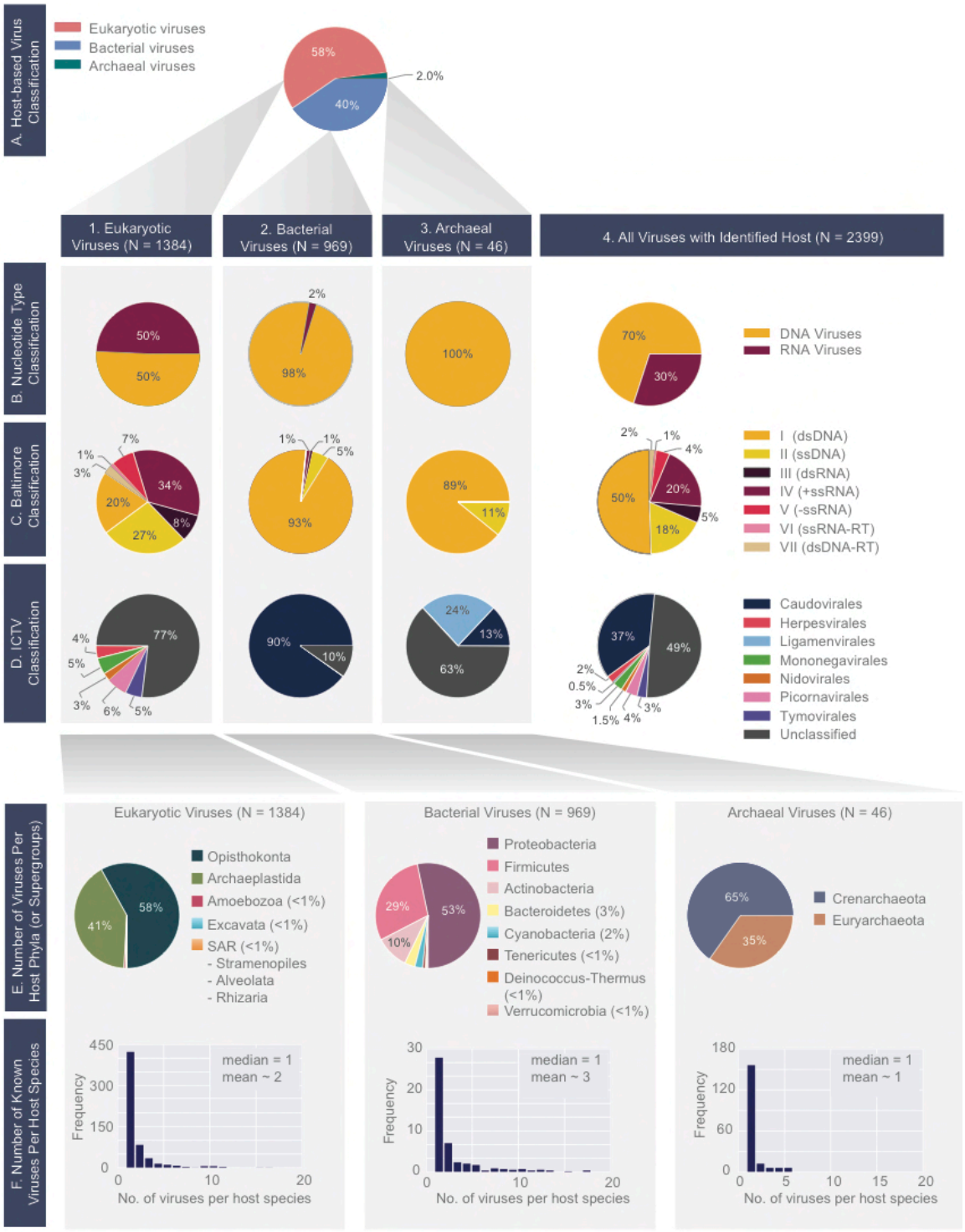


Figure 2. A census of all viruses with complete genomes reported to NCBI that were matched to a host (N= 2399). A) Percentage of viruses infecting hosts from the three domains of life. 1) Eukaryotic, 2) bacterial, and 3) archaeal viromes are further classified according to the B) Nucleotide Type,

C) Baltimore, and D) ICTV classification systems. E) Distributions of host phyla (or supergroups) infected by the 1) eukaryotic, 2) bacterial, and 3) archaeal viruses is shown. As in the case panel F, the host taxonomic identification is derived from the NCBI Taxonomy database (see Methods). F) Histograms of the number of known viruses infecting host species. Median and mean number of viruses infecting a host species is provided in each plot. The full-range of x-values for the bacterial and eukaryotic histograms extends beyond n=20 (see virusHostHistograms.ipynb).

We further observed that eukaryotes host nearly an equal number of DNA and RNA viruses (Figure 2.B1). In contrast to prokaryotes, which are predominantly host to viruses with double-stranded genomes, eukaryotes are host to a higher number of viruses with single-stranded genomes. Why are double-stranded DNA viruses, despite their high prevalence in the bacterial and archaeal world, only the third largest group of viruses in eukaryotes? One explanation proposed is the physical separation of transcriptional processes from the cytoplasm by way of the eukaryotic nucleus (46). This physical separation is thought to impose an additional barrier for DNA viruses in gaining access to the host's transcriptional environment.

More than half of viruses with complete genomes have not been assigned to any viral orders under the ICTV classification (Figure 2.D4). About one third of all known viruses are assigned to the *Caudovirales* order, while the other orders are in the minority. The vast majority of the bacterial viruses are categorized as part of the *Cauodvirales* order (Figure 2.D2), but the majority of archaeal and eukaryotic viruses remain unassigned to any order.

Before any further exploration of this dataset, we aimed to assess its diversity and possible sources of bias (Figure 2.E-F, SI Figure 1). It was immediately clear, for example, that archaeal viruses were heavily undersampled. In contrast, bacterial viruses infect hosts

from a diverse array of bacterial phyla (Figure 2.E2). However, even for bacterial viruses, there are host phyla whose viruses are entirely missing from the database, for example *Synergistes* and *Acidobacteria*, whose members are typically unculturable soil bacteria. Given that the isolation and characterization of archaeal and bacterial viruses has traditionally been dependent on the culturing of their hosts, the majority of viruses with unculturable hosts remain unexplored. Moreover, the eukaryotic viruses in the database infect hosts primarily from the *Viridiplantae* or the *Opisthokonta* supergroups (Figure 2.E1). Among *Viridiplantae*, the majority of hosts belong to the *Streptophytina* group (land plants), and within the *Opisthokonta* supergroup, the majority of viruses are metazoan. We further examine the distribution of viruses from the *Opisthokonta* supergroup in SI Figure 1.

We continued to explore host diversity at a finer resolution and mapped out the number of viruses that infect each host species (Figure 2.F). As expected, organisms such as *Staphylococcus aureus*, *Escherichia coli*, and *Solanum lycopersicum*, which are host species with either medical, research or agricultural relevance, have many known viruses and are outliers in the skewed distributions shown in Figure 2.F. However, the median number of viruses known to infect a eukaryotic or a prokaryotic host species is approximately 1 (Figure 2.F). This signifies that even for host species that are already represented in our collection, the number of known viruses is likely an underestimate considering the larger numbers of viruses known to infect the more heavily studied host species.

4.3 Viral genome lengths, gene lengths and gene densities

Genome lengths for all fully sequenced viral genomes varied widely by three orders of magnitude (Figure 3.A, Table 1). According to the Host Domain classification, prokaryotic viruses tend to have longer genomes than eukaryotic viruses (SI Table 1, SI

Figure 2). However, this difference can be better explained by the Nucleotide Type classification, as the median RNA virus genome length is four times shorter than the median DNA virus genome length. Thus, the comparison between prokaryotic and eukaryotic viral genome lengths is confounded by the fact that the prokaryotic virome is primarily composed of DNA viruses, whereas the eukaryotic virome is only half composed of DNA viruses (Figure 2.C4).

With respect to viral genome lengths, the Baltimore classification seems to offer the most explanatory power. Knowing whether a viral genome is DNA- or RNA-based already provides a strong indication about viral genome length, especially for RNA viruses where the standard deviation is just a few kilobases (SI Table 1). However, by distinguishing between ssDNA, dsDNA and dsDNA-RT viruses, the Baltimore classification offers a more complete view of genome length distributions compared to the binary Nucleotide Type classification (Figure 3.A). Across all Baltimore groups, dsDNA viruses have genome lengths that have the largest standard deviation, however considering the limited range of genome lengths associated with other Baltimore groups, it is very likely that a larger viral genome will be composed of dsDNA (Figure 3.A). We provide a more detailed view of genome length distributions by layering different classification systems, first applying the Baltimore classification, followed by the Host Domain and the ICTV family classifications (Figure 3.B, SI Table 1).

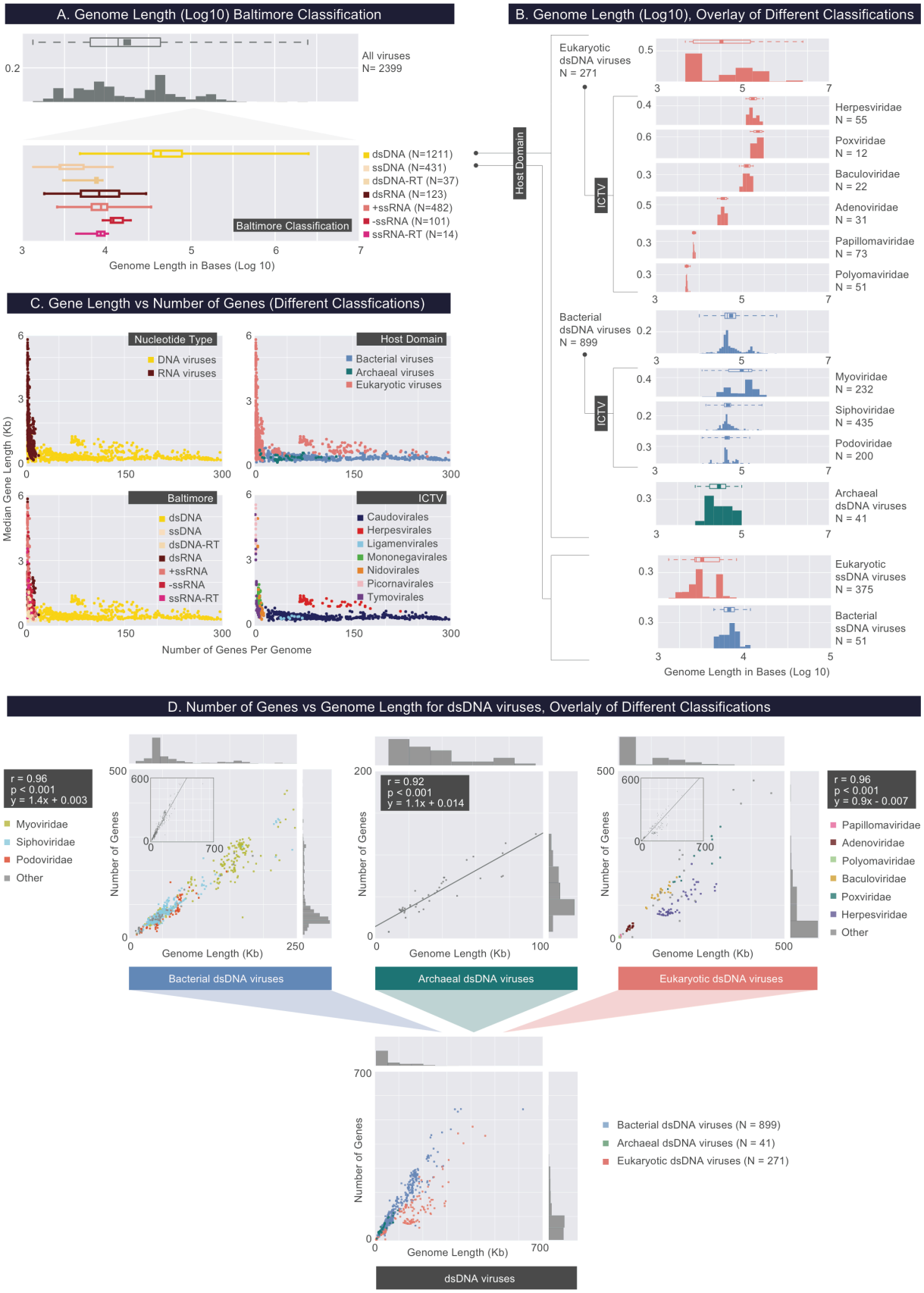

A. Genome Length (Log10) Baltimore Classification
All viruses
N= 2399
dsDNA (N=1211)
ssDNA (N=431)
dsDNA-RT (N=37)
dsRNA (N=123)
+ssRNA (N=482)
-ssRNA (N=101)
ssRNA-RT (N=14)
Baltimore Classification
Genome Length in Bases (Log 10)
B. Genome Length (Log10), Overlay of Different Classifications
Host Domain
ICTV
Eukaryotic dsDNA viruses N = 271
Herpesviridae N = 55
Poxviridae N = 12
Baculoviridae N = 22
Adenoviridae N = 31
Papillomaviridae N = 73
Polyomaviridae N = 51
Bacterial dsDNA viruses N = 899
Myoviridae N = 232
Siphoviridae N = 435
Podoviridae N = 200
Archaeal dsDNA viruses N = 41
Eukaryotic ssDNA viruses N = 375
Bacterial ssDNA viruses N = 51
Genome Length in Bases (Log 10)
C. Gene Length vs Number of Genes (Different Classfications)
Nucleotide Type
DNA viruses
RNA viruses
Host Domain
Bacterial viruses
Archaeal viruses
Eukaryotic viruses
Baltimore
dsDNA
ssDNA
dsDNA-RT
dsRNA
+ssRNA
-ssRNA
ssRNA-RT
ICTV
Caudovirales
Herpesvirales
Ligamenvirales
Mononegavirales
Nidovirales
Picornavirales
Tymovirales
Median Gene Length (Kb)
Number of Genes Per Genome
D. Number of Genes vs Genome Length for dsDNA viruses, Overlaly of Different Classifications
r = 0.96
p < 0.001
y = 1.4x + 0.003
Myoviridae
Siphoviridae
Podoviridae
Other
r = 0.92
p < 0.001
y = 1.1x + 0.014
r = 0.96
p < 0.001
y = 0.9x - 0.007
Papillomaviridae
Adenoviridae
Polyomaviridae
Baculoviridae
Poxviridae
Herpesviridae
Other
Number of Genes
Genome Length (Kb)
Bacterial dsDNA viruses
Archaeal dsDNA viruses
Eukaryotic dsDNA viruses
Bacterial dsDNA viruses (N = 899)
Archaeal dsDNA viruses (N = 41)
Eukaryotic dsDNA viruses (N = 271)
dsDNA viruses

Figure 3. Describing viral genomes through distributions of genome length, gene length and gene density. A) Box plots of genome lengths (Log10) across all viruses included in our dataset (top), further partitioned based on the Baltimore classification categories (bottom). The number of viruses included in each group is denoted by N. B) A closer examination of dsDNA and ssDNA viral genome lengths through the overlay of Host Domain and ICTV classification systems. Distributions of genome lengths associated with eukaryotic, bacterial and archaeal viruses are shown in salmon, blue, and teal, respectively. ICTV viral families with only a few members are omitted. Distributions of genome lengths across different classification systems along with various statistics are shown in SI Figure 2 and SI Table 1. Note that the bimodal distribution of eukaryotic ssDNA viruses, which also appears in the next figure, arises from the Begomoviruses, which are plant viruses with circularized monopartite and bipartite genomes (47). C) Median gene length is plotted against the number of genes for each genome for all genomes in our dataset, color-coded according to different classification systems. D) Number of genes per genome length (gene density) for dsDNA viruses based on the overlay of Host Domain (bottom) and ICTV family classification categories (top) (Pearson correlations and their statistical significance are denoted).

Table 1. Viral genomic statistics based upon different classification systems. Only median values are reported in this table. A more complete version of these statistics can be found in SI Table 1- SI Table 3. Genome length data is rounded to the nearest kilobase. N corresponds to the number of viruses from which data is obtained.

Classification		N	Genome length (kb)	Percent noncoding (DNA/RNA)	Median gene length (bases)
Host Domain	Eukaryota Viruses	1384	8	10	1055
	Bacteria Viruses	969	43	9	408
	Archaea Viruses	46	24	10	400
Baltimore	Group I (dsDNA)	1211	44	9	429
	Group II (ssDNA)	431	3	14	588
	Group III (dsRNA)	123	8	8	2291
	Group IV (+ssRNA)	482	9	5	2366
	Group V (-ssRNA)	101	12	7	1353
	Group VI (ssRNA-RT)	14	8	16	1799
	Group VII (dsDNA-RT)	37	8	11	558
Nucleotide Type	DNA Viruses	1679	38	10	444
	RNA Viruses	720	9	6	2072
ICTV (orders)	Caudovirales	879	44	9	408
	Herpesvirales	55	159	19	1107
	Ligamenvirales	11	37	12	372
	Mononegavirales	71	12	8	1266
	Nidovirales	35	27	3	672
	Picornavirales	89	8	11	7056
	Tymovirales	73	8	4	693
Combinations of different classifications	All Eukaryotic dsDNA viruses	271	33	11	990
	All Bacterial dsDNA viruses	899	44	9	408
	All Archaeal dsDNA viruses	41	28	10	396
	All Eukaryotic ssDNA viruses	375	3	14	732
	All Bacterial ssDNA viruses	51	7	14	348

In viewing the relationship between median gene length and number of genes per viral genome (Figure 3.C), two different coding strategies become apparent. Namely,

compared to DNA viruses, RNA viruses exhibit a large range of gene lengths. This trend is at least in part reflective of the challenges faced by RNA viruses when encountering the requirements of their host's translational machinery (48). For example, many of the RNA genomes we examined closely contained genes that encode polyproteins, ribosomal slippage (frame-shifting) or codon read-through events, among other non-canonical translational mechanisms.

As in the case of genome lengths, by examining only the ICTV or the Host Domain classifications it would be difficult to draw meaningful conclusions about the observed patterns, and in the case of the Host Domain classification, our conclusions would be confounded by the disproportionate ratio of RNA to DNA viruses that are known to infect each host domain. However, the layering of these classification systems offers new insights, which we will discuss in the following paragraphs.

We follow others (26, 27) in defining the gene density of a genome as the number of genes divided by the genome length (Figure 3.D). We further partitioned dsDNA viruses according to the Host Domain and subsequently the ICTV (family) classifications. We observed a strong linear correlation between dsDNA viral genome lengths and the number of genes encoded by these genomes (Figure 3.D). The mean (and median) gene densities for bacterial, archaeal and eukaryotic dsDNA viral genomes are approximately 1.4, 1.6 and 0.9 genes per kilo basepairs. As illustrated by the slopes of the regression lines, as well as through a nonparametric statistical test performed on eukaryotic and bacterial dsDNA viral gene densities (one-sided Mann-Whitney U test, $P < 0.001$), bacterial dsDNA viruses have significantly higher gene densities than their eukaryotic counterparts.

A closer examination of median gene lengths more clearly reveals the significantly longer gene lengths of RNA viruses compared to DNA viruses (one-sided Mann-Whitney U

test, P < 0.001) (Figure 4, Table 1). By focusing on DNA viruses, and further dividing these viruses based on Baltimore, Host Domain and ICTV (family) classifications, we arrive at an interesting trend. Namely, eukaryotic viruses, whether dsDNA or ssDNA, have significantly longer gene lengths compared to bacterial viruses from the same Baltimore classification category (Figure 4, SI Table 2) (one-sided Mann-Whitney U test, P < 0.001). This trend follows what we see across cellular genomes, since prokaryotic genes and proteins are shown to be significantly shorter than eukaryotic ones (5, 49).

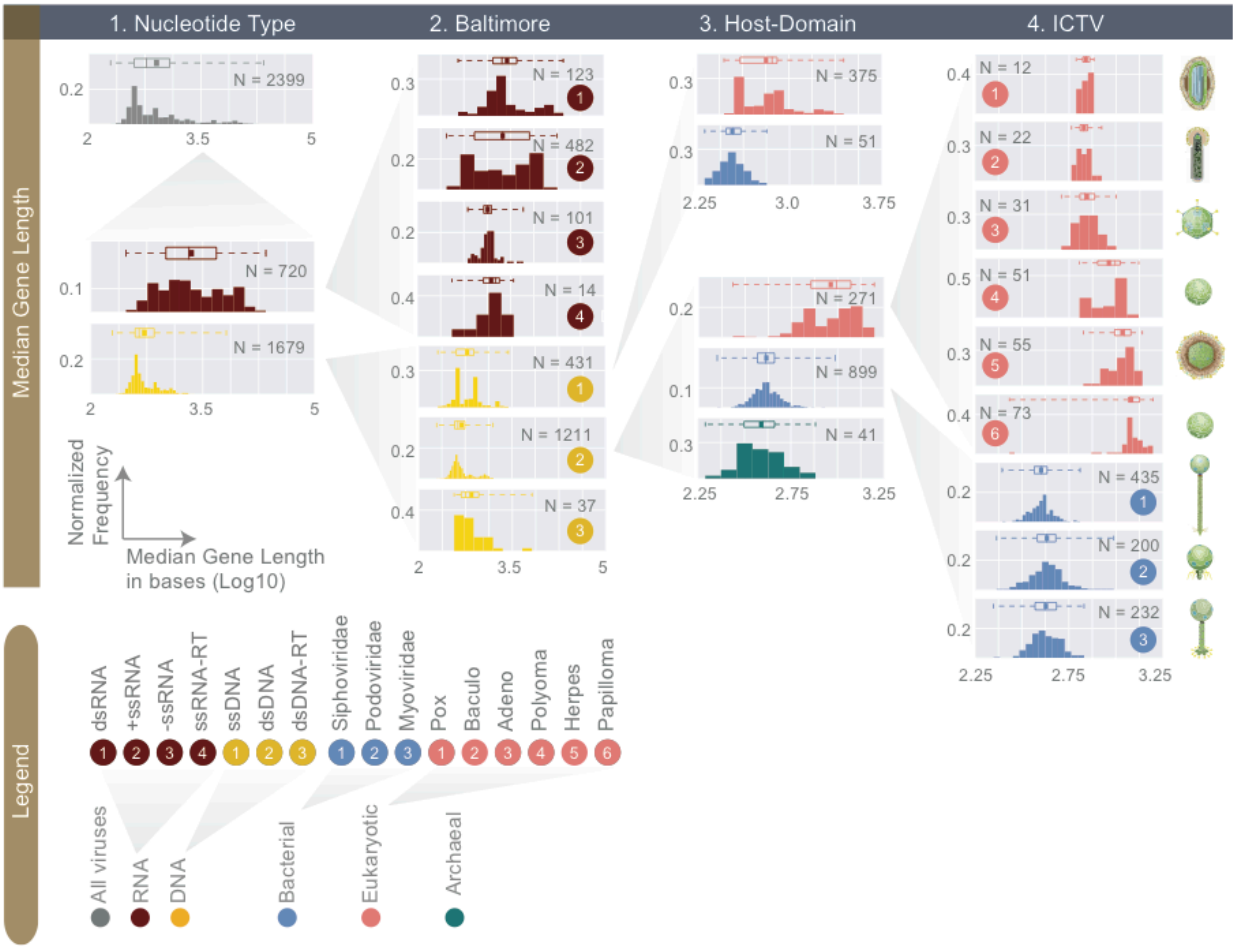


Figure 4. Normalized histograms of median gene lengths (log10) across all complete viral genomes associated with a host. Instead of showing absolute viral counts on y-axes, the counts are normalized by the total number of

viruses in each viral category (denoted as N inside each plot). The mean of each distribution is denoted as a dot on the boxplot. For all histograms, bin numbers and bin widths are systematically decided by the Freedman-Diaconis rule (35). Viral schematics are modified from ViralZone (38). Key statistics describing these distributions can be found in Table 1 and SI Table 2.

4.4 Noncoding percentages of viral genomes

So far we have primarily focused on the coding fractions of viral genomes. Thus, we created distributions of noncoding percentage of viral genomes (see Methods, Figure 5, Table 1, SI Table 3). In general, DNA viral genomes contain about 10% noncoding regions which is even lower than the noncoding percentage of bacterial genomes (29, 30). With a median noncoding percentage of just 6%, RNA viral genomes have significantly lower noncoding percentage compared to DNA viruses (one-sided Mann-Whitney U Test, $P < 0.001$). A notable exception to the RNA viral group is the ssRNA-RT with a median noncoding percentage of 16%. Interestingly, both retroviral groups had relatively high noncoding DNA percentages. This is likely due to the presence of defunct retroviral genes. For example, the *Xenopus laevis* endogenous retrovirus (NCBI taxon ID 204873) belonging to the ssRNA-RT group has a noncoding percentage of 93%. This high noncoding percentage can be explained by the fact that this virus genome contains three pseudogenes previously coding for *env*, *pol,* and *gag* proteins.

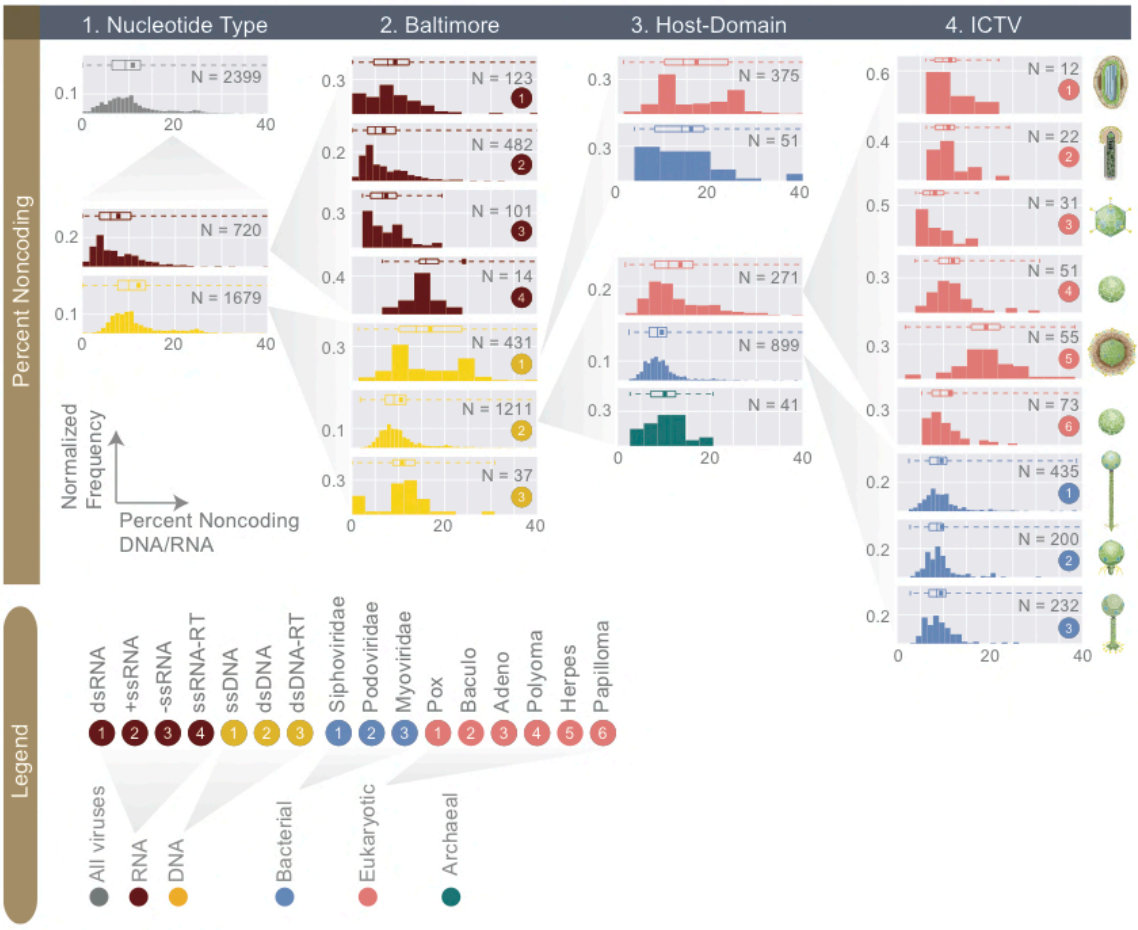


Figure 5. Normalized histograms of noncoding DNA/RNA percentage across all complete viral genomes associated with a host. The counts of viruses are normalized by the total number of viruses in each viral category (denoted as N inside each plot). The mean of each distribution is denoted as a dot on the boxplot. For all histograms, bin numbers and bin widths are systematically decided by the Freedman-Diaconis rule (35). Viral schematics are modified from ViralZone (38). Key statistics describing these distributions can be found in Table 1 and SI Table 3.

4.5 Viral functional gene categories

We categorized viral genes according to several major functional categories, including structural genes such as capsid and tail genes, metabolic genes, informational

genes, which we define as those involved in replication, transcription or translation of the viral genetic code, among other categories (Figure 6, see Methods). In addition to the fraction of viral genes that we were able to assign to these functional categories, there still remains what we will refer to as an "unlabeled" fraction that contains hypothetical genes or genes with poor annotation (see Methods). When reporting the relative abundance of different functional gene categories, we will normalize the number of genes belonging to each functional category by the total number of labeled genes.

RNA, dsDNA, and ssDNA viruses, despite differences in the detailed categorization of their genes (Figure 6.B) share similar general features (Figure 6.A). For example, across all three viral groups, roughly half of all genes are structural. Similarly, dsDNA viruses of eukaryotes and bacteria, in contrast to having different genomic properties and morphologies surprisingly have very similar distribution of gene functional category and subcategory abundances. The major difference between these two viral groups, as expected from our knowledge of viral morphologies, is that a larger portion of eukaryotic dsDNA viral genes are envelope and matrix genes, whereas a greater portion of bacterial dsDNA genes are portal and tail-associated genes. By further zooming in on bacterial dsDNA viruses, it is again interesting to see that *Myoviridae*, *Siphoviridae*, and *Podoviridae* viral groups, with their different morphologies and wide range of hosts, having very similar functional gene category abundances even at the level of subcategories.

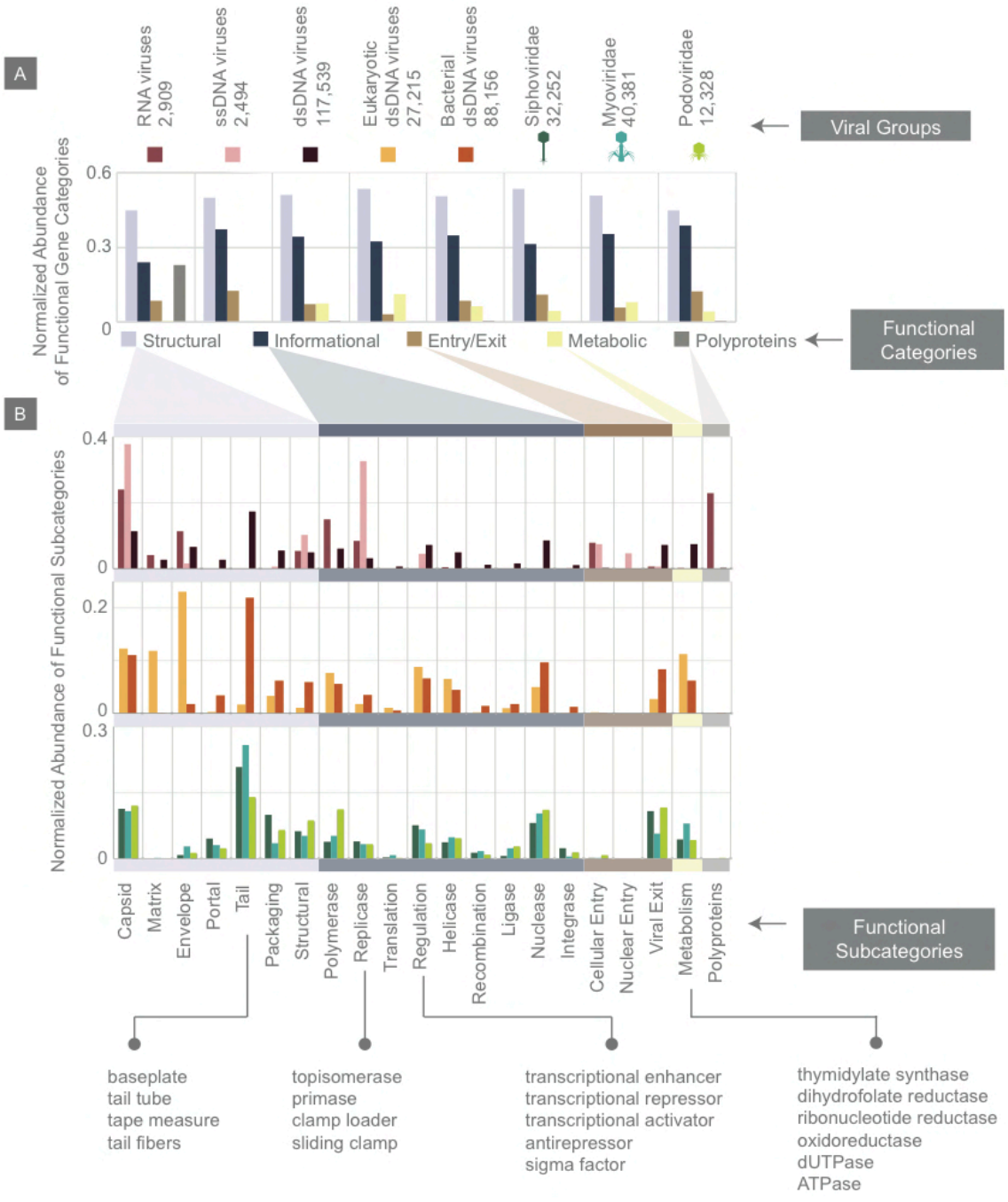


Figure 6. Normalized abundance of functional gene categories across different viral groups. A) Abundances of functional gene categories across 8 viral groups normalized to the number of labeled genes in each viral group (the number of labeled genes in each viral group is shown above the panel). B) Abundances of functional gene subcategories across 8 viral groups: RNA,

ssDNA, and dsDNA viral groups (top plot); eukaryotic and bacterial dsDNA viral groups (middle); *Siphoviridae*, *Myoviridae*, and *Podoviridae* viral groups (bottom). A few examples of the types of genes contained as part of each functional subcategory are provided.

4.6 Viral genome organization

To explore viral genome organization we developed a coarse-grained method for visualizing a large number of genomes in one snapshot. We first defined genome organization as the order in which genes appear across a genome. We then symbolized each gene by a letter, indifferent to the gene's length or its orientation on the genome. Genes with similar functions are grouped and are represented by the same letter (Figure 7). Therefore each viral genome, analogous to a nucleotide sequence, is compactly described by a sequence of letters that represent its gene order (Figure 7), which we will refer to as the gene order sequence. Because we aimed to study gene order sequences across different viral groups, we focused on genes whose functions are universally required, namely structural genes. SI text file 1 provides the structural gene order sequences for all viruses (see Methods for filters applied), though the script developed can be easily modified to visualize the placement of any number of genes or user-defined gene groups.

Furthermore, by focusing on bacterial dsDNA viruses, we were able to identify the most common gene order patterns across this virome (see Methods). One particular gene order pattern and its variations exist across various types of dsDNA bacterial viruses. We will refer to it as gene order pattern A, or pattern A for short (Figure 7.A). In pattern A, gene packaging, portal and capsid-related genes are mostly tightly clustered and are followed by tail-associated genes. Interestingly, this pattern occurs at the beginning of the genome for some viruses, and for others it seems to have been shifted further down on the genome.

Pattern A occurs across viruses from five different host phyla. The other two most common gene order patterns (patterns B and C) occur across viruses with more limited host range and morphologies.

To further examine the extent to which gene order sequences in a given pattern may be related at the sequence level, we used BLASTN to identify genomes in pattern C that share any regions of homology. We have shown examples of genomes (see SI) with very little sequence similarity, which due to having similar gene order sequence were grouped into the same gene order pattern, suggesting that at least for some viral genomes the gene order can remain conserved in time.

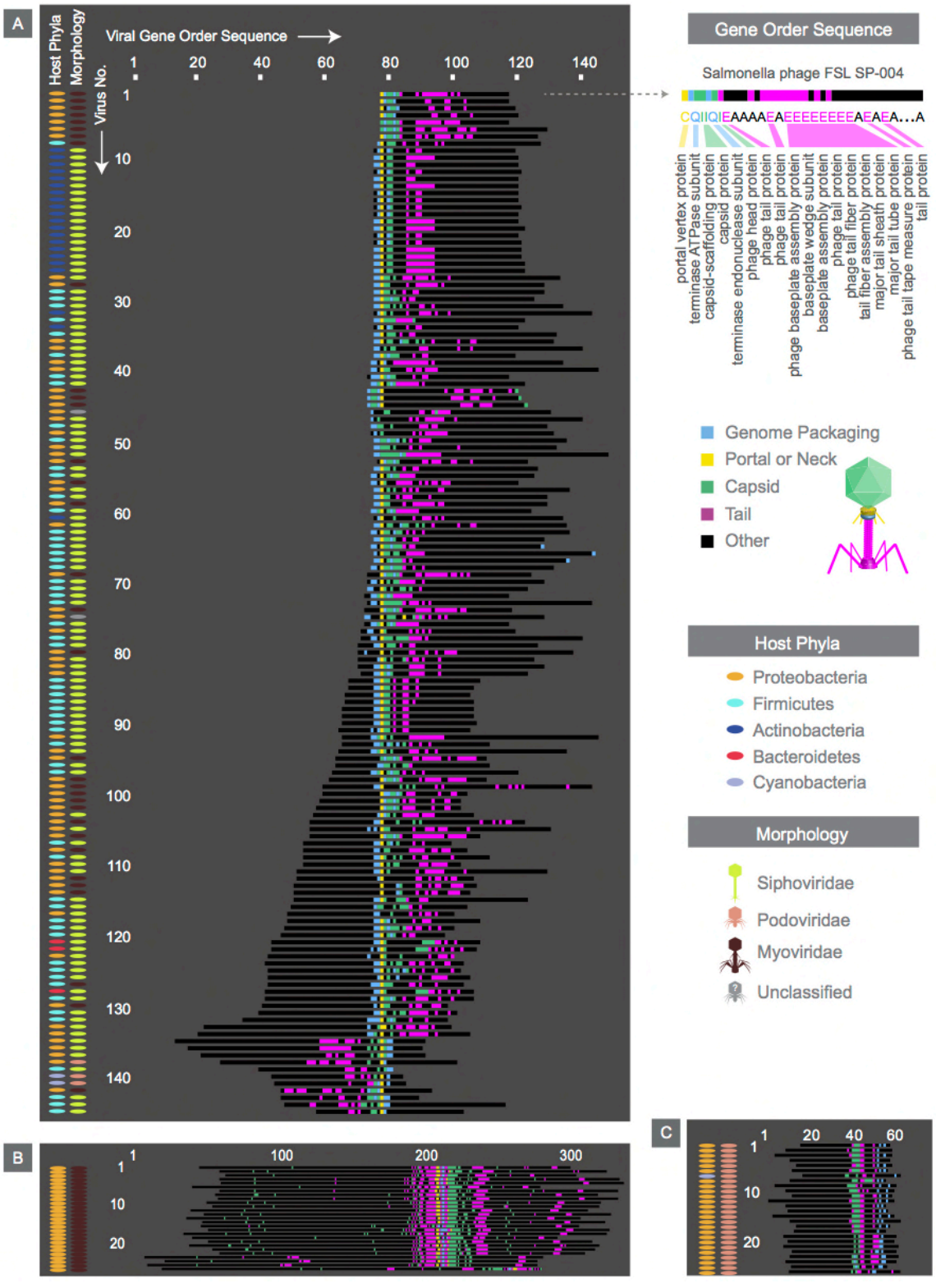

A
Host Phyla
Morphology
Virus No.
Viral Gene Order Sequence
1
20
40
60
80
100
120
140
Gene Order Sequence
Salmonella phage FSL SP-004
CQIIQIEAAAAEAEEEEEEEEAEAEA...A
portal vertex protein
terminase ATPase subunit
capsid-scaffolding protein
capsid protein
terminase endonuclease subunit
phage head protein
phage tail protein
phage tail protein
phage baseplate assembly protein
baseplate wedge subunit
baseplate assembly protein
phage tail protein
phage tail fiber protein
tail fiber assembly protein
major tail sheath protein
major tail tube protein
phage tail tape measure protein
tail protein
Genome Packaging
Portal or Neck
Capsid
Tail
Other
Host Phyla
Proteobacteria
Firmicutes
Actinobacteria
Bacteroidetes
Cyanobacteria
Morphology
Siphoviridae
Podoviridae
Myoviridae
Unclassified
B
C

Figure 7. Alignment of the most common gene order patterns for dsDNA bacterial viruses. Each genome is summarized by a sequence of letters, with each letter corresponding to a gene, positioned in the order that it appears on the genome. As an example, the gene order sequence *for Salmonella phage FSL SP-004* is shown. Note the letters shown serve to only denote genes with similar functions. Structural genes are assigned colors, whereas other genes are denoted in black. Across all three panels, each row corresponds to the gene order sequence for a given virus, and thus, the length of the sequence denotes the number of genes within a given genome. The left two columns accompanying each panel provide further information on hosts and viral morphologies. Panel A, B, and C, represent gene order patterns A, B, and C, respectively. Geneious global alignment (37) was used to align gene order sequences (see Methods).

4.7 Discussion

The motivation for conducting a large-scale study of viral genomes was above all to provide the distributions of key numbers that characterize viral genomes. Furthermore, implicit in this aim was to compare different viral classification systems. Because viral classification systems were constructed prior to the emergence of sequencing, we were interested to see how well they can describe genomic trends. Based on a comparison of classification systems across various genomic metrics, the Baltimore classification and in some cases its more minimal form (Nucleotide Type classification) seem to provide the clearest explanation for the observed trends. We suspect that this is due to the Baltimore classification's discernment of RNA, ssDNA, and dsDNA genomes, which have striking physical differences.

The greater stability of dsDNA compared to RNA (50) and ssDNA is thought to be an important factor in the observed variations in genome lengths. The 2'-hydroxyl group in

RNA makes it more susceptible to hydrolysis events and cleavage of the backbone compared to DNA. It has been shown that for bacteria and viruses, the mutation rate and the genome length are inversely correlated (51, 52), and it is therefore hypothesized that the lack of proofreading mechanisms in RNA replication and the resulting higher mutation rates compared to DNA replication (52) imposes length limits on RNA viral genomes. In support of the suspected link between mutation rates and genome length, it has been shown that long RNA viruses (above 20 kb) contain 3'-5' exonuclease, which is a homolog of the DNA-proofreading enzymes (53).

Similarly, the hydrolysis of cytosine into uracil occurs two orders of magnitude faster in ssDNA genomes than in dsDNA genomes (54). This may explain the high mutation rates of ssDNA viruses, which is within the range of RNA viral mutation rates, despite using error-correcting host polymerases to replicate. In contrast to genome length in which ssDNA and RNA viruses have similar distributions, it was interesting to see that ssDNA viruses are actually more similar to dsDNA viruses in terms of their gene lengths and noncoding percentages.

While the Baltimore classification serves as a meaningful coarse-grained classification system, it can be expanded by the addition of subcategories. As is shown by gene length distributions (Figure 4), the additional layer of categorization provided by the Host Domain classification offers new insight. For example, dsDNA and ssDNA viruses of eukaryotes have much longer gene lengths compared to their prokaryotic counterparts, an observation that may be hinting at the coevolution of host and viral genomes and proteomes since the eukaryotic genes and proteins are also shown to be significantly longer than prokaryotic ones (49, 55, 56). It is well known that certain eukaryotic viral genomes, similar to their hosts' genomes, contain genes with introns (57-59), which may explain the longer median gene

length for eukaryotic viruses. In fact mRNA splicing was discovered for the first time in a study of adenovirus mRNA expression (60). Virus proteomes are also shown to be tuned to their hosts' proteomes by having similar codon usage and amino acid preferences (61). However, future studies are needed to further ascertain the mechanisms responsible for the differences in eukaryotic and prokaryotic viral gene lengths.

The ICTV classification also offers some supporting data (e.g. viral morphology or in some cases host information), perhaps as the final layer of classification, but it is limited by the fact that it leaves many viruses unclassified and, more importantly, that it lacks truly systematic classification criteria. As our exploration of viruses shifts its basis from culturing of viruses to sequencing of viruses from their natural habitats, morphological data is likely to become more and more scarce. As a result, ICTV will need to adapt its classification system to operate exclusively on genomic data, a viewpoint that is broadly shared by many experts in the field (10).

In this work, we have described our attempt at providing a comprehensive and quantitative view of fully sequenced viral genomes. Similar to earlier work on biological numeracy, as exemplified by the BioNumbers database (3), we have identified a number of interesting trends associated with viral genomes that will be helpful in gaining a broad overview of vastly different viral groups.

4.8 Materials and Methods

Data acquisition and curation. All genomic data was retrieved from the NCBI Genome FTP server (retrieved July 2015) (Brister, 2015). Matching viruses to their hosts was done by parsing ASN files from the NCBI Genome FTP server while searching for the term "nat-host". All other taxonomic data, including host and viral lineages, was retrieved from

NCBI's Taxonomy database using the NCBI Taxa class of the ETE Toolkit (62). Once we had the "nat-host" name of organisms in English, we retrieved their taxids using ETE Toolkit. These were in turn used to identify the host's taxonomic lineage. Hosts with complete genomes were identified by searching the assembly reports of the NCBI Genome FTP server for assemblies labeled "Complete Genome", then using the associated FTP address to download the _assembly_stats.txt files and _protein.faa files. Only viruses that could be matched to a host were included for further analysis. Additionally, various quality checks were manually performed to ensure that viruses with improper annotations were excluded from further analysis. For example, we found viruses and hosts with incomplete or incorrect taxonomic information which we excluded from further analysis. The list of excluded viruses can be found in our code (see next section).

Data and code availability. We have compiled all scripts (in the format of Jupyter Notebooks) used to write this manuscript in a GitHub repository (https://github.com/gitamahm/VirologyByTheNumbers), and invite others to explore our methods. viromePieChartsVF.ipynb and virusHostHistogramsVF.ipynb were used to create Figure 2. The code for Figure 3 through Figure 5 can be found in genomeLengthsVF.ipynb, geneLengthsVF.ipynb, and percentNoncodingVF.ipynb, respectively. Finally, the code for Figure 6 and Figure 7 is provided in geneOrderAndGeneAbundanceVF.ipynb. SI text files 1 through 6 can also be found in this repository.

Genome length and gene densities. Genome lengths were extracted from .ptt files and _assembly_stats.txt files for viruses. The .ptt files were parsed to find "complete genome - 1.." which is followed by the length of the genome. The _assembly_stats.txt files were parsed to find the first instance of "total-length", which is followed by the length of the genome. For segmented genomes, the total length of the segments is reported as the genome

length. The number of protein-coding genes, which was used in calculating gene densities, was found by parsing .faa files using the BioPython Seq class. For gene length histograms, we first obtained the gene lengths for each virus, and then create a histogram based on the median gene length associated with each virus. To have a systematic scheme for determining the number of bins needed for each histogram, we employ the Freedman-Diaconis' rule (63) for all histograms shown in this paper.

Noncoding DNA/RNA percentages. To extract the percent of the genome that is noncoding, we could not merely subtract the lengths of the genes from the length of the genome, as this would not take overlapping genes into account. Instead, we used the .ptt files to identify where each gene began and ended in the genome, then added all indices between protein-coding genes to a set. We then could subtract the size of this set from the genome length to arrive at the number of noncoding bases, which is then turned into a percentage.

Decomposition of Viral Genes into Functional Categories. To obtain the abundance of various gene functional categories, we collected the COG product annotations (64) accompanying each gene from .ptt file(s) provided for each virus. Based on the most frequent COG product names, we constructed a dictionary of search terms to query viral genes and measure the abundance of various functional categories (by measuring abundance, we are referring to the number of genes that belong to a given functional category). To determine the most common search terms, we derived the unique set of COG product annotations for different viromes. We used the annotations shared between viromes to exclude problematic search terms with multiple meanings. As a result we avoided search terms with multiple functional associations such as “gp41”, which in the context of HIV

signifies a transmembrane glycoprotein, and in the context of *Mycobacterium phage Bxb1* denotes a 3'-5' exonuclease involved in DNA replication.

While the dictionary constructed contains many key words that capture essential gene functional categories common to many viruses, it does not account for COG annotations that are non-descriptive (e.g. "phage protein" or "Z protein"). Additionally, there is typically a large number of genes that code for "hypothetical proteins". Together, these two fractions make up the unlabeled component, which we do not include for further analysis. Despite the limitations introduced by these unlabeled genes, there are still a large number of genes ($\sim 10^5$) that are included in our analysis. In constructing the relative abundances of different gene functional categories (Figure 6), we divide the abundance of a gene functional category by the total number of labeled genes (denoted at the top of Figure 6.A for each viral group).

Gene Order. In visualizing gene order we employed a similar search strategy to the one explained in the previous section. To detect potentially conserved patterns in gene order across vastly different viral genomes, we searched only structural genes as they are essential to any virus. We used .ptt files to determine gene order since they contain the beginning and end indices of genes. The code developed uses .ptt files as input, and outputs a string of characters per viral genome, which we have referred to as the gene order sequence. Each character represents a viral gene in the order that it appears on the genome (without distinguishing between the strand of DNA on which the gene is located). All genes belonging to the same functional category, for example all tail-related genes, are represented by the same character. All unlabeled genes (i.e. non-structural, hypothetical, or poorly annotated genes) are also represented by the same character. Each gene order sequence,

analogous to a nucleotide sequence, can be aligned against other gene order sequences by existing alignment software.

Though it would be ideal to calculate a pairwise distance matrix between gene order sequences and to quantitatively define a gene order pattern based on gene order sequence similarity (akin to defining an Operational Taxonomic Unit), this effort would require the development of appropriate alignment algorithms and inference methods fit to process gene order sequences. In the meantime, we used existing alignment software as a guide and grouped gene order sequences based on generally shared features

We used Geneious software (65) to align gene order sequences using global alignment with free end gaps and identity cost matrix (with default gap open and extension penalties). Using Geneious global alignment as a guide, we further manually improved the alignment by aligning similar characters, without introducing any gaps. This step was necessary because any alignment algorithm will aim to maximize the alignment between unlabeled genes, unable to distinguish between these characters and the more meaningful characters corresponding to labeled structural genes. Moreover, because of the high fraction of genes that have "hypothetical protein" COG annotation, we had to impose filters to extract gene order sequences that are not entirely composed of unlabeled genes. To generate the alignments shown in Figure 7, we imposed that at least 15% of characters in a gene order sequence have to correspond to labeled genes, and that the gene order sequence has to be at least 40 characters long. For the gene order sequences shown in SI text file 1 the sequence order length limit was not imposed.

To further explore gene order pattern C, we used BLASTN and accession numbers shown in SI Figure 3 to BLAST all genomes in pattern C against each other. We have summarized these results in SI Figure 3.A, wherein orange boxes correspond to homologous

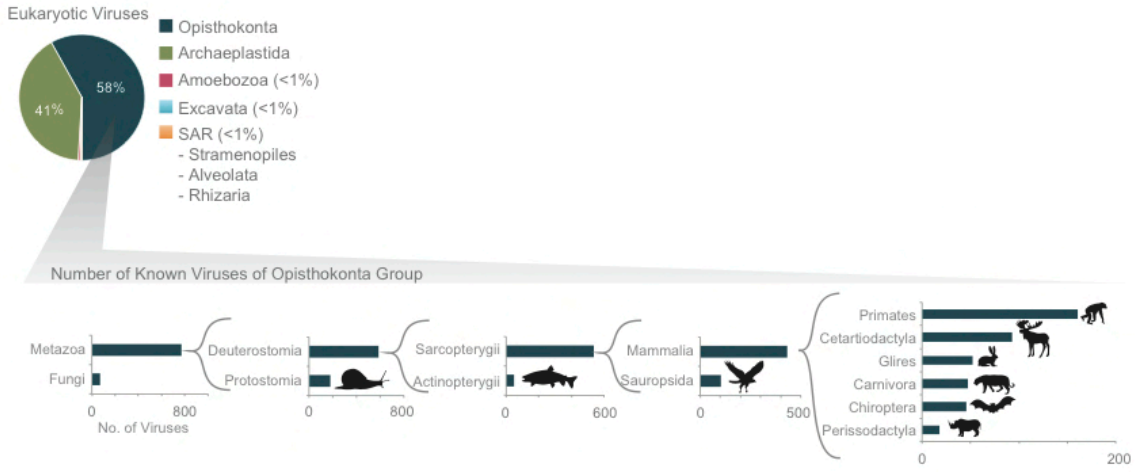
relationships (E values < 10^{-5}). All non-homologous relationships are shown in blue. We then used protein-protein BLAST to identify relationships between genomes that had no homology to any other genomes in pattern C. We targeted tail tube A, tail tube B, capsid, tail fiber, and large terminase proteins (SI Figure 3.B). The BLASTP reports are provided as SI text file 2- SI text file 6.

4.9 Supplementary Information

Virus classifications. The ICTV classifies viruses into seven orders: *Herpesvirales*, large eukaryotic double-stranded DNA viruses; *Caudovirales*, tailed double-stranded DNA viruses typically infecting bacteria; *Ligamenvirales*, linear double-stranded viruses infecting archaea; *Mononegavirales*, nonsegmented negative (or antisense) strand single-stranded RNA viruses of plants and animals; *Nidovirales*, positive (or sense) strand single-stranded RNA viruses of vertebrates; *Picornavirales*, small positive strand single-stranded RNA viruses infecting plants, insects, and animals; and finally, the *Tymovirales*, monopartite positive single-stranded RNA viruses of plants. In addition to these orders, there are ICTV families, some of which have not been assigned to an ICTV order. Only those ICTV viral families with more than a few members present in our dataset are explored.

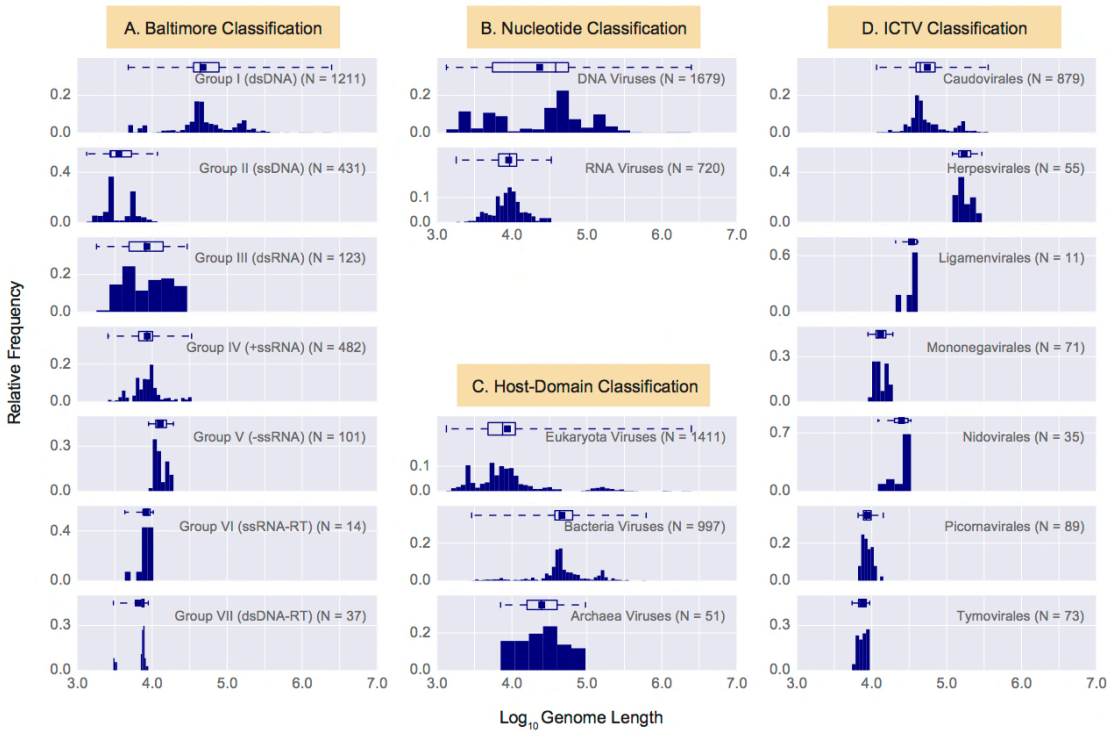
The Baltimore classification groups viruses into seven categories (Figure 1): double-stranded DNA viruses (Group I); single-stranded DNA viruses (Group II); double-stranded RNA viruses (Group III); positive single-stranded RNA viruses (Group IV); negative single-stranded RNA viruses (Group V); positive single-stranded RNA viruses with DNA intermediates (Group VI), commonly known as retroviruses; and the double-stranded DNA retroviruses (Group VII).

Comparing the sequence similarity of genomes found in Gene Order Pattern C. We wanted to examine the extent to which gene order sequences in a given pattern may be related at the sequence level, so we used BLASTN to identify genomes in pattern C that share any regions of homology. While the majority of sequences share at least a small degree of homology across their genomes (see Methods, SI Figure 3), genomes of *Acinetobacter phage Petty*, *Lelliottia phage phD2B*, and *Synechococcus phage S-CBP4* do not share any significant homology at the nucleotide level with any other genome in pattern C (SI Figure 3). When we examined several structural proteins from these viruses using protein-protein BLAST, we found that they have homologous large terminase and tail tube B proteins. However, they have weak to nonexistent homology across their capsid, tail fiber, and tail tube A proteins (see Methods, SI Figure 3, SI text file 2- SI text file 6). This finding demonstrates that at least for some viruses despite limited homology across their nucleotide sequences, genome organization could still be detectably similar.

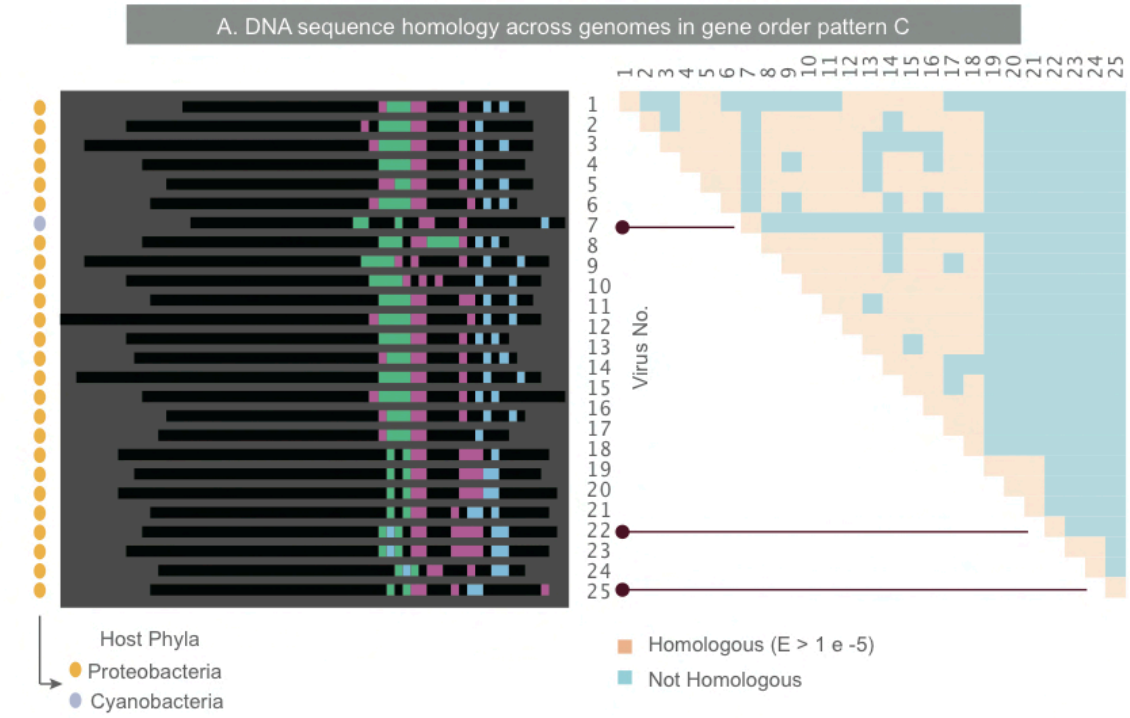


SI Figure 1. Further exploration of the largest fraction of the eukaryotic virome: viruses of Opisthokonta supergroup (animals). The x-axis corresponds to the number of viruses infecting each host group. In a

recursive fashion, the host group with the largest number of known viruses is further zoomed in on (host groups infected by only a few known viruses are not shown). The host classification was obtained from the NCBI taxonomic database.

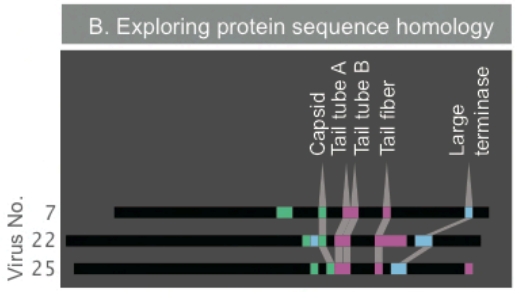


SI Figure 2. Histograms of genome length (Log10) across all complete viral genomes associated with a host. Histograms are grouped according to four viral classification systems: A) Baltimore classification, B) Nucleotide type classification, C) Host Domain Classification, and D) ICTV classification. Instead of showing absolute viral counts on the y-axis, the counts are normalized by the total number of viruses in each viral category (the total counts of viruses in each category is denoted as N inside the plots). The mean of each distribution is denoted as a dot on the boxplots. The relevant statistics for each distribution is provided in SI Table 1. In each histogram the number of bins and their width is set by Freedman-Diaconis rule (35).



Virus name and NCBI genome accesion no.

Virus No.		
1.	Pseudomonas phage gh 1	NC_004665.1
2.	Yersinia pestis phage phiA1122	NC_004777.1
3.	Enterobacteria phage T3	NC_003298.1
4.	Salmonella phage Vi06	NC_015271.1
5.	Yersinia phage Berlin	NC_008694.1
6.	Yersinia phage vB YenP AP5	NC_025451.1
7.	Synechococcus phage S CBP4	NC_025464.1
8.	Cronobacter phage Dev2	NC_023558.1
9.	Citrobacter phage CR44b	NC_023576.1
10.	Erwinia phage vB EamP L1	NC_019510.1
11.	Kluyvera phage Kvp1	NC_011534.1
12.	Yersinia phage phiYeO3 12	NC_001271.1
13.	Pseudomonas phage Phi S1	NC_021062.1
14.	Pseudomonas phage phiPSA2	NC_024362.1
15.	Citrobacter phage CR8	NC_023548.1
16.	Pseudomonas phage philBB PF7A	NC_015264.1
17.	Klebsiella phage KP32	NC_013647.1
18.	Enterobacteria phage K1F	NC_007456.1
19.	Pseudomonas phage phikF77	NC_012418.1
20.	Pseudomonas phage MPK6	NC_022746.1
21.	Pseudomonas phage LUZ19	NC_010326.1
22.	Lelliottia phage phD2B	NC_025450.1
23.	Xylella phage Paz	NC_022982.1
24.	Xylella phage Prado	NC_022987.1
25.	Acinetobacter phage Petty	NC_023570.1



SI Figure 3. Exploring homology across genomes and proteins in gene order pattern C. A) Qualitative depiction of BLASTN results, wherein orange boxes correspond to homologous genomes (E value < 10^{-5}), and blue denote non-homologous genomes. The genome number can be used to identify the genome name and its NCBI accession number. B) Gene order sequences for *Acinetobacter phage Petty*, *Lelliottia phage phD2B*, and *Synechococcus phage S-CBP4*, which are viruses with genomes that have no homology to any other genomes in pattern C. We compared their tail tube A, tail tube B, capsid, tail fiber, and large terminase proteins using protein-protein BLAST. SI text file 2- SI text file 6 provide the BLASTP results.

SI Table 1. Genome length statistics for viral groups across different classification systems (rounded to the nearest kilobase).

		Genome Length Statistics (kb)						
Classification	**Classification Categories**	**Min**	**Max**	**25th Percentile**	**Median**	**75th Percentile**	**Mean**	**Stdev.**
Host Domain	Eukaryota Viruses (N = 1384)	1.3	2473.9	4.9	7.5	11.1	27.7	110.5
	Bacteria Viruses (N = 969)	2.9	617.5	37.3	43.4	65.8	63.4	55.7
	Archaea Viruses (N = 46)	7.0	95.7	15.7	24.4	40.9	31.7	21.9
Baltimore	Group I (dsDNA) (N = 1211)	4.8	2473.9	35.5	43.6	77.7	75.9	121.4
	Group II (ssDNA) (N = 431)	1.3	11.7	2.7	2.8	5.3	4.0	1.8
	Group III (dsRNA) (N = 123)	1.8	29.1	4.8	8.1	14.0	10.6	7.0
	Group IV (+ssRNA) (N = 482)	2.6	33.5	6.6	8.5	10.1	9.6	5.5
	Group V (-ssRNA) (N = 101)	8.9	19.2	11.1	11.9	15.4	13.0	2.6
	Group VI (ssRNA-RT) (N = 14)	4.3	10.3	7.5	8.4	9.5	8.3	1.5
	Group VII (dsDNA-RT) (N = 37)	3.0	8.8	7.3	7.5	7.8	6.8	1.8
Nucleotide Type	DNA Viruses (N = 1679)	1.3	2473.9	5.5	38.3	56.6	55.9	108.0
	RNA Viruses (N = 720)	1.8	33.5	6.6	9.3	11.5	10.2	5.5
ICTV (orders)	Caudovirales (N = 879)	11.6	358.7	39.1	44.5	70.2	67.9	52.5
	Herpesvirales (N = 55)	119.5	295.1	144.9	159.2	211.5	177.0	45.1
	Ligamenvirales (N = 11)	20.9	41.2	31.8	36.9	40.4	34.7	6.5
	Mononegavirales (N = 71)	8.9	19.2	11.4	12.0	15.5	13.4	2.5
	Nidovirales (N = 35)	12.1	33.5	20.1	26.7	31.0	25.8	5.6
	Picornavirales (N = 89)	6.6	14.3	7.8	8.4	9.8	8.9	1.4
	Tymovirales (N = 73)	5.5	9.4	6.7	7.9	8.5	7.6	1.0
Combinations of different classifications	All Eukaryotic dsDNA viruses (N = 271)	4.8	2473.9	7.3	33.0	152.7	109.0	232.5
	Baculoviridae (N = 22)	84.3	176.7	108.6	127.6	151.1	127.2	23.9
	Poxviridae (N = 12)	150.0	307.7	170.6	237.2	282.9	233.1	56.2
	Herpesvirales (N = 55)	119.5	295.1	144.9	159.2	211.5	177.0	45.1
	Papillomaviridae (N = 73)	7.0	8.3	7.3	7.6	7.7	7.6	0.3
	Adenoviridae (N = 31)	26.3	45.8	31.6	35.1	43.4	36.1	6.0
	Polyomaviridae (N = 51)	4.8	6.2	5.0	5.1	5.3	5.1	0.2
	All Bacterial dsDNA viruses (N = 899)	10.1	617.5	39.0	44.4	69.8	67.8	55.5
	Siphoviridae (N = 435)	14.3	280.0	38.0	43.1	53.1	50.5	30.5
	Podoviridae (N = 200)	11.7	145.9	39.2	42.5	50.4	47.2	19.5
	Myoviridae (N = 232)	11.6	358.7	47.4	136.4	164.0	118.1	69.2
	All Archaeal dsDNA viruses (N = 41)	8.1	95.7	17.4	28.3	41.2	34.5	21.6
	All Eukaryotic ssDNA viruses (N = 375)	1.3	8.1	2.7	2.8	5.2	3.5	1.4
	All Bacterial ssDNA viruses (N = 51)	4.4	11.7	5.8	6.8	7.8	6.8	1.5

SI Table 2. Median gene length statistics for viral groups across different classification systems (rounded to the nearest base). It is important to clarify that the median values of this table represents the median of median gene lengths.

		Median Gene Length Statistics (bases)						
Classification	**Classification Categories**	**Min**	**Max**	**25th Percentile**	**Median**	**75th Percentile**	**Mean**	**Stdev.**
Host Domain	Eukaryota Viruses (N = 1384)	272	22173	702	1055	2129	2192	2770
	Bacteria Viruses (N = 969)	204	5100	366	408	456	429	192
	Archaea Viruses (N = 46)	195	762	324	400	462	412	119
Baltimore	Group I (dsDNA) (N = 1211)	195	1577	380	429	555	539	269
	Group II (ssDNA) (N = 431)	204	2777	404	588	774	692	419
	Group III (dsRNA) (N = 123)	453	22173	1638	2291	3978	4148	4409
	Group IV (+ssRNA) (N = 482)	297	17715	828	2366	6372	3742	3266
	Group V (-ssRNA) (N = 101)	648	5052	1167	1353	1568	1448	633
	Group VI (ssRNA-RT) (N = 14)	362	3530	1154	1799	2103	1805	921
	Group VII (dsDNA-RT) (N = 37)	368	6537	477	558	915	873	998
Nucleotide Type	DNA Viruses (N = 1679)	195	6537	393	444	708	586	354
	RNA Viruses (N = 720)	297	22173	1014	2072	4812	3452	3360
ICTV (orders)	Caudovirales (N = 879)	224	972	369	408	456	419	76
	Herpesvirales (N = 55)	669	1382	978	1107	1200	1092	151
	Ligamenvirales (N = 11)	315	462	342	372	429	384	45
	Mononegavirales (N = 71)	648	1896	1055	1266	1367	1218	275
	Nidovirales (N = 35)	297	4920	537	672	1056	1045	1007
	Picornavirales (N = 89)	3375	10041	6372	7056	8232	6963	1580
	Tymovirales (N = 73)	402	5103	554	693	1014	1183	1138
Combinations of different classifications	All Eukaryotic dsDNA viruses (N = 271)	272	1577	714	990	1179	958	271
	Baculoviridae (N = 22)	582	843	647	672	711	680	57
	Poxviridae (N = 12)	614	762	650	695	729	691	45
	Herpesvirales (N = 55)	669	1382	978	1107	1200	1092	151
	Papillomaviridae (N = 73)	272	1577	1170	1209	1338	1231	187
	Adenoviridae (N = 31)	510	999	636	681	771	706	104
	Polyomaviridae (N = 51)	639	1320	797	990	1055	930	158
	All Bacterial dsDNA viruses (N = 899)	224	972	369	408	456	419	78
	Siphoviridae (N = 435)	248	644	366	401	429	402	57
	Podoviridae (N = 200)	231	972	378	426	480	438	100
	Myoviridae (N = 232)	224	678	372	419	483	431	76
	All Archaeal dsDNA viruses (N = 41)	195	762	315	396	459	405	120
	All Eukaryotic ssDNA viruses (N = 375)	300	2777	404	732	806	741	426
	All Bacterial ssDNA viruses (N = 51)	204	653	303	348	404	352	84

SI Table 3. Percent noncoding DNA (or RNA) for viral groups across different classification systems (rounded to the nearest percentage).

		Percent Noncoding (DNA/RNA)						
Classification	**Classification Categories**	**Min**	**Max**	**25th Percentile**	**Median**	**75th Percentile**	**Mean**	**Stdev.**
Host Domain	Eukaryota Viruses (N = 1384)	0	93	5	10	15	11	9
	Bacteria Viruses (N = 969)	3	92	7	9	11	10	6
	Archaea Viruses (N = 46)	3	21	7	10	13	10	4
Baltimore	Group I (dsDNA) (N = 1211)	2	92	7	9	12	11	7
	Group II (ssDNA) (N = 431)	1	91	10	14	24	17	10
	Group III (dsRNA) (N = 123)	0	47	5	8	12	9	8
	Group IV (+ssRNA) (N = 482)	0	43	3	5	9	7	5
	Group V (-ssRNA) (N = 101)	2	20	4	7	10	8	4
	Group VI (ssRNA-RT) (N = 14)	7	93	15	16	19	24	23
	Group VII (dsDNA-RT) (N = 37)	0	31	9	11	14	11	6
Nucleotide Type	DNA Viruses (N = 1679)	0	92	8	10	14	12	8
	RNA Viruses (N = 720)	0	93	4	6	10	8	7
ICTV (orders)	Caudovirales (N = 879)	3	92	7	9	11	10	5
	Herpesvirales (N = 55)	2	38	16	19	22	19	6
	Ligamenvirales (N = 11)	8	21	9	12	17	13	4
	Mononegavirales (N = 71)	2	20	4	8	10	8	4
	Nidovirales (N = 35)	1	8	2	3	5	4	2
	Picornavirales (N = 89)	2	23	9	11	12	11	4
	Tymovirales (N = 73)	2	13	3	4	4	4	2
Combinations of different classifications	All Eukaryotic dsDNA viruses (N = 271)	2	86	8	11	16	14	9
	Baculoviridae (N = 22)	6	24	8	10	12	11	4
	Poxviridae (N = 12)	6	22	7	10	13	11	5
	Herpesvirales (N = 55)	2	38	16	19	22	19	6
	Papillomaviridae (N = 73)	5	51	8	9	12	11	8
	Adenoviridae (N = 31)	4	18	6	7	11	8	3
	Polyomaviridae (N = 51)	4	31	9	11	14	12	5
	All Bacterial dsDNA viruses (N = 899)	3	92	7	9	11	10	5
	Siphoviridae (N = 435)	3	39	7	9	11	10	5
	Podoviridae (N = 200)	3	55	7	9	10	10	6
	Myoviridae (N = 232)	3	92	7	9	11	10	6
	All Archaeal dsDNA viruses (N = 41)	3	21	7	10	13	10	4
	All Eukaryotic ssDNA viruses (N = 375)	1	80	10	14	24	17	10
	All Bacterial ssDNA viruses (N = 51)	4	91	8	14	19	16	14

SI text files

SI text files are included as part the SI Text Files folder at the following GitHub repository: https://github.com/gitamahm/VirologyByTheNumbers

SI text file 1. Gene order sequences for all viruses whose genomes contained at least 15% labeled genes. Letters I, C, E, and Q correspond to capsid-related, portal-related, tail-related, and genome packaging-related genes, respectively. All other genes are denoted by the letter A.

SI text file 2. BLASTP report for the tail tube A protein found in *Acinetobacter phage Petty*, *Lelliottia phage phD2B*, and *Synechococcus phage S-CBP4* genomes.

SI text file 3. BLASTP report for the tail tube B protein found in *Acinetobacter phage Petty*, *Lelliottia phage phD2B*, and *Synechococcus phage S-CBP4* genomes.

SI text file 4. BLASTP report for the tail fiber protein found in *Acinetobacter phage Petty*, *Lelliottia phage phD2B*, and *Synechococcus phage S-CBP4* genomes.

SI text file 5. BLASTP report for the capsid protein found in *Acinetobacter phage Petty*, *Lelliottia phage phD2B*, and *Synechococcus phage S-CBP4* genomes.

SI text file 6. BLASTP report for the large terminase protein found in *Acinetobacter phage Petty*, *Lelliottia phage phD2B*, and *Synechococcus phage S-CBP4* genomes.

4.10 References

1. Suttle CA (2005) Viruses in the sea. *Nature* 437(7057):356-361.
2. Wigington CH*, et al.* (2016) Re-examination of the relationship between marine virus and microbial cell abundances. *Nature microbiology* 1:15024.
3. Milo R, Jorgensen P, Moran U, Weber G, & Springer M (2010) BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic acids research* 38(suppl 1):D750-D753.
4. Phillips R, Kondev J, Theriot J, & Garcia H (2012) *Physical biology of the cell* (Garland Science).
5. Milo R & Phillips R (2015) *Cell biology by the numbers* (Garland Science).
6. Phillips R & Milo R (2009) A feeling for the numbers in biology. *Proceedings of the National Academy of Sciences* 106(51):21465-21471.
7. Paez-Espino D*, et al.* (2016) Uncovering Earth's virome. *Nature* 536(7617):425-430.
8. Edwards RA & Rohwer F (2005) Viral metagenomics. *Nature Reviews Microbiology* 3(6):504-510.
9. Rohwer F & Thurber RV (2009) Viruses manipulate the marine environment. *Nature* 459(7244):207-212.
10. Simmonds P*, et al.* (2017) Consensus statement: Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*.
11. Simmonds P (2015) Methods for virus classification and the challenge of incorporating metagenomic sequence data. *Journal of General Virology* 96(6):1193-1206.
12. Mokili JL, Rohwer F, & Dutilh BE (2012) Metagenomics and future perspectives in virus discovery. *Current opinion in virology* 2(1):63-77.

13. Kristensen DM, Mushegian AR, Dolja VV, & Koonin EV (2010) New dimensions of the virus world discovered through metagenomics. *Trends in microbiology* 18(1):11-19.

14. Hendrix RW (2003) Bacteriophage genomics. *Current opinion in microbiology* 6(5):506-511.

15. Youle M, Haynes M, & Rohwer F (2012) Scratching the surface of biology's dark matter. *Viruses: Essential agents of life*, (Springer), pp 61-81.

16. Hug LA*, et al.* (2016) A new view of the tree of life. *Nature Microbiology* 1:16048.

17. Kristensen DM, Cai X, & Mushegian A (2011) Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. *Journal of bacteriology* 193(8):1806-1814.

18. King AMQ (2011) Virus taxonomy: classification and nomenclature of viruses: Ninth Report of the International Committee on Taxonomy of Viruses. (Elsevier).

19. Baltimore D (1971) Expression of animal virus genomes. *Bacteriological reviews* 35(3):235.

20. Mahy BWJ & Van Regenmortel MHV (2010) *Desk encyclopedia of human and medical virology* (Academic Press).

21. Forterre P (2010) Defining life: the virus viewpoint. *Origins of Life and Evolution of Biospheres* 40(2):151-160.

22. Brister JR, Ako-Adjei D, Bao Y, & Blinkova O (2015) NCBI viral genomes resource. *Nucleic acids research* 43(D1):D571-D577.

23. Roux S*, et al.* (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*.

24. Dutilh BE*, et al.* (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature communications* 5.

25. Alberts B*, et al.* (2002) Molecular Biology of the Cell 4th Edition: International Student Edition. (Routledge).

26. Keller B & Feuillet C (2000) Colinearity and gene density in grass genomes. *Trends in plant science* 5(6):246-251.

27. Hou C, Li L, Qin ZS, & Corces VG (2012) Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Molecular cell* 48(3):471-484.

28. Elgar G & Vavouri T (2008) Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in Genetics* 24(7):344-352.

29. Mattick JS & Makunin IV (2006) Non-coding RNA. *Human molecular genetics* 15(suppl 1):R17-R29.

30. Morris KV (2012) *Non-coding RNAs and epigenetic regulation of gene expression: Drivers of natural selection* (Horizon Scientific Press).

31. Mattick JS (2004) RNA regulation: a new genetics? *Nature Reviews Genetics* 5(4):316-323.

32. Howe JA*, et al.* (2015) Selective small-molecule inhibition of an RNA structural element. *Nature*.

33. Reich PR, Forget BG, Weissman SM, & Rose JA (1966) RNA of low molecular weight in KB cells infected with adenovirus type 2. *Journal of molecular biology* 17(2):428-439.

34. Tycowski KT*, et al.* (2015) Viral noncoding RNAs: more surprises. *Genes & development* 29(6):567-584.

35. Steitz J, *et al.* (2011) Noncoding RNPs of viral origin. *Cold Spring Harbor perspectives in biology* 3(3):a005165.

36. Mathews MB & Shenk T (1991) Adenovirus virus-associated RNA and translation control. *Journal of virology* 65(11):5657.

37. Riaz A, *et al.* (2014) Ovine herpesvirus-2-encoded microRNAs target virus genes involved in virus latency. *Journal of General Virology* 95(2):472-480.

38. Molina N & van Nimwegen E (2009) Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends in genetics* 25(6):243-247.

39. Grilli J, Bassetti B, Maslov S, & Lagomarsino MC (2012) Joint scaling laws in functional and evolutionary categories in prokaryotic genomes. *Nucleic acids research* 40(2):530-540.

40. Labonté JM, *et al.* (2015) Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *The ISME journal* 9(11):2386-2399.

41. Casjens SR, *et al.* (2005) The generalized transducing Salmonella bacteriophage ES18: complete genome sequence and DNA packaging strategy. *Journal of bacteriology* 187(3):1091-1104.

42. Marinelli LJ, *et al.* (2012) Propionibacterium acnes bacteriophages display limited genetic diversity and broad killing activity against bacterial skin isolates. *MBio* 3(5):e00279-00212.

43. Brüssow H & Hendrix RW (2002) Phage genomics: small is beautiful. *Cell* 108(1):13-16.

44. Telford MJ & Copley RR (2011) Improving animal phylogenies with genomic data. *Trends in Genetics* 27(5):186-195.

45. Jaillon O*, et al.* (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* 431(7011):946-957.

46. Koonin EV, Dolja VV, & Krupovic M (2015) Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* 479:2-25.

47. Melgarejo TA*, et al.* (2013) Characterization of a New World monopartite begomovirus causing leaf curl disease of tomato in Ecuador and Peru reveals a new direction in geminivirus evolution. *Journal of virology* 87(10):5397-5413.

48. Firth AE & Brierley I (2012) Non-canonical translation in RNA viruses. *Journal of General Virology* 93(7):1385-1409.

49. Brocchieri L & Karlin S (2005) Protein length in eukaryotic and prokaryotic proteomes. *Nucleic acids research* 33(10):3390-3400.

50. Lindahl T (1993) Instability and decay of the primary structure of DNA. *nature* 362(6422):709-715.

51. Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences* 88(16):7160-7164.

52. Sanjuan R, Nebot MR, Chirico N, Mansky LM, & Belshaw R (2010) Viral mutation rates. *Journal of virology* 84(19):9733-9748.

53. Lauber C*, et al.* (2013) The footprint of genome architecture in the largest genome expansion in RNA viruses. *PLoS Pathog* 9(7):e1003500.

54. Frederico LA, Kunkel TA, & Shaw BR (1990) A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* 29(10):2532-2537.

55. Zhang J (2000) Protein-length distributions for the three domains of life. *Trends in Genetics* 16(3):107-109.

56. Tiessen A, Pérez-Rodríguez P, & Delaye-Arredondo LJ (2012) Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC research notes* 5(1):85.

57. Himmelspach M, Cavaloc Y, Chebli K, Stevenin J, & Gattoni R (1995) Titration of serine/arginine (SR) splicing factors during adenoviral infection modulates E1A pre-mRNA alternative splicing. *Rna* 1(8):794-806.

58. Barksdale S & Baker CC (1995) Differentiation-specific alternative splicing of bovine papillomavirus late mRNAs. *Journal of virology* 69(10):6553-6556.

59. Ge H & Manley JL (1990) A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA in vitro. *Cell* 62(1):25-34.

60. Flint SJ*, et al.* (2000) *Principles of Virology: Molecular Biology, Patogenesis and* (ASM Press. Washington DC USA).

61. Bahir I, Fromer M, Prat Y, & Linial M (2009) Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Molecular systems biology* 5(1):311.

62. Huerta-Cepas J, Dopazo J, & Gabaldón T (2010) ETE: a python Environment for Tree Exploration. *BMC bioinformatics* 11(1):24.

63. Freedman D & Diaconis P (1981) On the histogram as a density estimator: L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 57(4):453-476.

64. Tatusov RL, Galperin MY, Natale DA, & Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research* 28(1):33-36.

65. Kearse M, *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12):1647-1649.

Chapter V

Human Phageprints: Commensal phage communities reveal individual-specific and temporally-stable signatures

5.1 Introduction

The study of bacteriophages (phages) often relies on culturing the bacterial hosts. Because the vast majority of microbes are currently unculturable, we have only recently become aware of the overwhelming presence of phages in natural environments through metagenomic and imaging studies (1, 2). These recent studies collectively paint phages not only as the most numerous and diverse biological entities on our planet, but also as regulators of microbial ecosystems through rapid infection cycles and horizontal gene transfer events (3-7). Yet, compared to their bacterial hosts, and despite their medical, biogeochemical and agricultural importance, environmental phages remain largely under-sampled and poorly characterized (8-12).

One of the defining features of phage genomes (and viruses in general) is the lack of the ribosomal RNA sequence that has remained highly conserved across cellular genomes. The 16S ribosomal RNA sequence (18S in eukaryotic genomes) is used as a universal marker, and its sequence variation across genomes can be used to draw conclusions about cellular classification and evolution (13-15), as well as geographical distribution and community composition (16, 17). This marker-based approach to microbiology is indispensible, as it enables a high coverage depth of the 16S region via PCR-amplification and high-throughput sequencing. This depth of coverage provides a precise and reproducible depiction of bacterial community composition across space and time (18-20).

Given current sequencing technology, the trade-off for coverage depth is the coverage breadth, as only a small region in the genome can be sequenced (Figure 1).

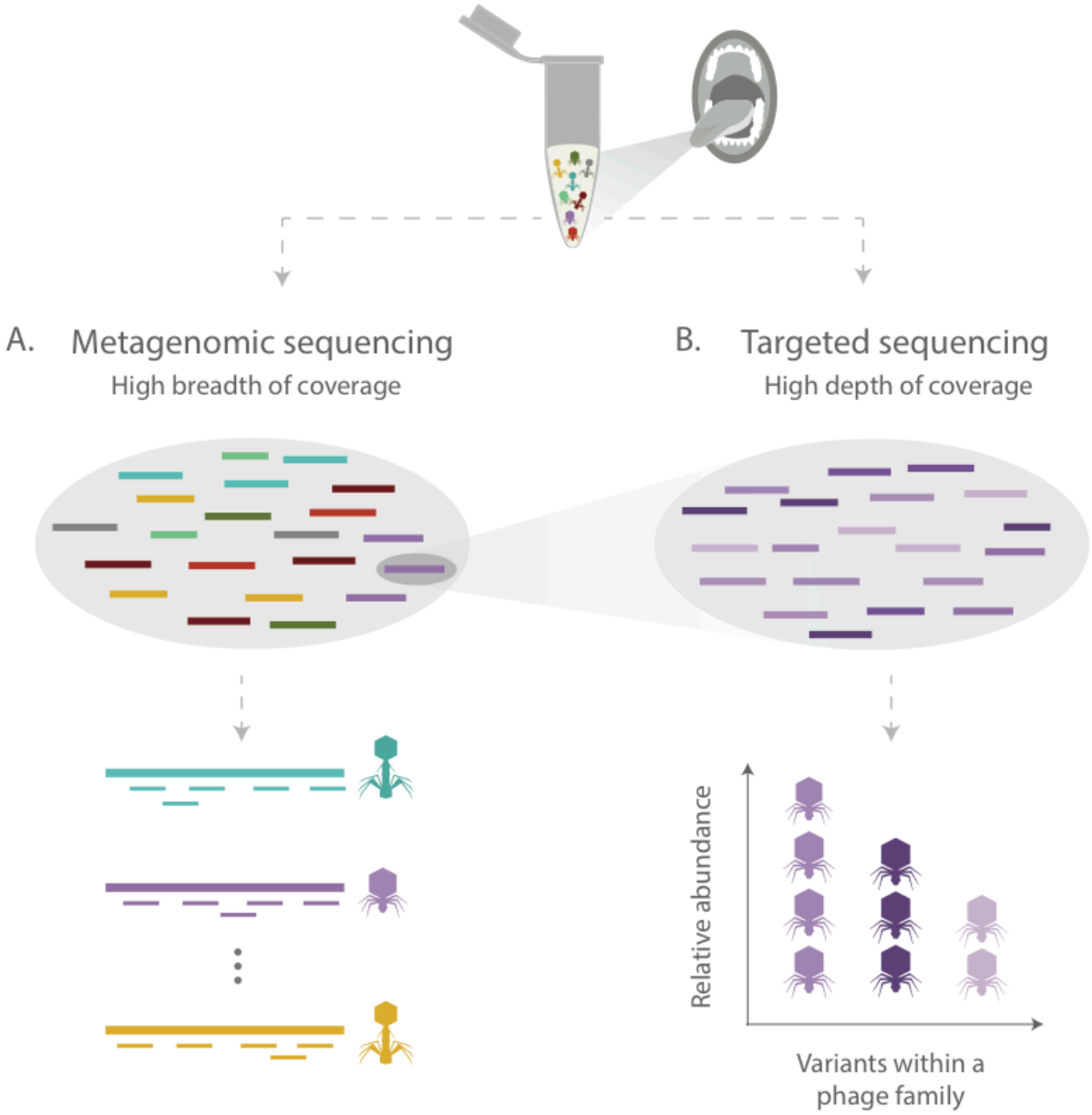


Figure 1. Comparison of A) metagenomic sequencing and B) targeted sequencing approaches. A) Metagenomic sequencing offers high breadth of coverage, spanning genomes from many different organisms, however it suffers from low depth of coverage (shown here by the incomplete assembly of phage genomes). B) Targeted sequencing approaches, such as 16S

sequencing, which use PCR to amplify a specific genomic region, exchange breadth of coverage for depth. Targeted sequencing studies, due to their greater depth of coverage, provide much higher resolution for constructing the community composition by equating coverage depth with relative abundance of species or strains.

In comparison to the targeted marker-based approach, metagenomic studies provide much greater breadth in coverage (Figure 1). Typically, these studies obtain genomic sequences from many different organisms, but the coverage depth remains comparatively low. Assuming an equal distribution of species or variants within a sample, coverage depth can be estimated as $C = \frac{LN}{G}$, where L represents sequencing read length, N represents the total number of reads, and G represents the total length of genome(s) or genetic segment(s) of interest (21). For example, using 100 million, 150-bp Illumina HiSeq reads to sequence a marker of length 150 bp, would provide a coverage depth of $\frac{150 \times 10^8}{150}$ or 10^8x. This would mean that each base pair along the makrer's length would be sequenced 10^8 times. On the other hand, if the same sequencing conditions are used for metagenomic sequencing of a sample hypothetically composed of 100 different bacterial genomes (assuming an average bacterial genome length of 10^6 bp (22)) the coverage depth would drop to just $\frac{150 \times 10^8}{100 \times 10^6} =$ 150x. Natural environments are often far more complex, containing many different species and strains of organisms with varying abundances (23). One of the manifestations of this problem is that *de novo* assembly of genomes from natural environments through current metagenomic techniques remains a significant challenge (24), even for the most abundant members of an environment with relatively short genome lengths (25).

Moreover, the coverage depth of a genome through metagenomics varies greatly across the length of the genome. Even for complete genome assemblies from metagenomic samples, the assemblies represent consensus genomes (average representation of highly similar genomes within an environment). In these studies, it is typical to see genomic segments with ~100x coverage depth that are islands in the sea of low coverage depth regions (< 10x) (26-28). However, when a variant appears 1 in 1000 sequences, 100x coverage does not provide adequate resolution needed to assess the sequence diversity of this genomic region. Because of this limitation of metagenomics, the marker-based exploration of microbial community composition often provides a more reliable and accurate approach (29-31).

Due to the lack of a 16S-analog within viral genomes, the study of environmental phages has typically relied on metagenomic studies (5, 32-34). While these studies have unveiled intriguing facets of environmental phage biology, insufficient coverage depth fails to provide a detailed view of phage community composition. Thus, to explore environmental phage communities at a higher resolution than is typically offered through metagenomics, we decided to take a targeted, marker-based approach. We hypothesized that due to the immense sequence diversity of phage genomes (35-37), a high-resolution view of their communities could provide novel insights.

Considering the lack of universal markers for phages, we aimed to discover environment-specific phage markers so that we could study previously unexplored phage families (we will provide a more precise description of this term later in this section). As we will demonstrate, this marker-based approach reveals phage community composition with a resolution that cannot be achieved through typical metagenomic approaches. Specifically, we will follow others (38, 39) in using the term "community composition" to refer to the

relative abundance profile of members within a phage family. We will further demonstrate that at this resolution, phage community composition can serve as a fingerprint, or a "phageprint" – highly unique to an environment and apparently stable over time.

In our search for candidate phage markers, we combined the advantages of metagenomic and marker-based approaches. Namely, to study previously unexplored phage families, we first used existing metagenomic datasets to identify candidate phage families, and then by targeting these families using degenerate primers and PCR, we were able to explore them with high coverage depth (see Methods). We limited our bioinformatic and later experimental search for ubiquitous phage families to those inhabiting the human oral cavity. Our motivation to study phages within the human oral environment was rooted in this environment's diversity of microbial inhabitants, key role in human health and disease, feasibility of sample collection, and the wealth of existing oral metagenomic datasets (40-43).

Briefly, to arrive at markers representing ubiquitous oral phage families, we imposed several guidelines: 1) candidate marker sequences should be unique to phages; 2) candidate phage markers should be present across different metagenomic datasets so to increase the likelihood that the markers will represent core (though not necessarily abundant) members of the human oral environment; and 3) candidate markers representing different phage families should not share any significant sequence similarity.

To meet the first criterion, we focused our search on the terminase large subunit (TerL) sequences, which are shown to be unique to phages (44, 45). To meet the second criterion, we used two metagenomic datasets, the Xie (46) and the Mira (47) datasets, to identify potentially ubiquitous phage makers based on their presence in both datasets. Using two larger metagenomic projects, namely the HOMD (43) and the HMP (48), we confirmed the presence of candidate phage markers across a greater number of human oral samples. In

meeting the third criterion, pairwise sequence similarity analysis was performed to exclude markers that share any sequence similarity to each other (see Methods). Finally, by designing degenerate primers for the candidate phage markers, we experimentally confirmed marker presence in human oral samples, and developed a marker-based survey of phage communities with high resolution. In the absence of a taxonomic convention for viral genomic data, we use the term "phage family" to refer to phages that share homologous TerL sequences.

For the bulk of our sample collection, we took a citizen science approach. By creating instructional videos and sample collection kits (see Methods), volunteers collected ~700 oral samples (representing ~100 individuals). Using this large collection of samples, we were able to explore phage communities as a function of space (Figure 3, Figure 5, Figure 6, Figure 12) and time (Figure 9, Figure 10, Figure 11). By examining phage communities at 6 different oral sites, and by comparing phage communities of individuals living across the globe, we were able to study the effect of spatial separation, ranging from several millimeters to thousands of kilometers. We found that the spatial separation of just a few centimeters (the distance between the gingival sites and the hard palate, for example) can already result in highly distinct phage communities. For larger distances, spanning the phage communities of different individuals, we did not observe any correlation between spatial distance and phage community composition. In other words, individuals residing in the same city did not have any more similar phage communities than individuals living on different continents.

Additionally, we found that neither genetics nor cohabitation played a role in the relatedness of phage communities across individuals. Cohabitating siblings and even identical twins did not have phage communities that were any more similar than those of unrelated individuals (Figure 12). The only factor we observed that contributes to phage community

relatedness is direct exchange of saliva between individuals, as is demonstrated by the similarity between phage communities of couples (Figure 11, Figure 12).

By exploring phage communities across the span of a month, we observed highly stable communities (Figure 8, Figure 9, Figure 10, Figure 11). These studies consistently point to the existence of remarkably diverse and personal phage families that are stable in time.

5.2 Results

5.2.A Discovery of phage families ubiquitous across humans

To test our bioinformatics predictions regarding the presence of ubiquitous phage families, we designed degenerate primers (see Methods). We then tested to see if the primers would be able to amplify the candidate phage marker sequences from oral samples previously collected from 10 individuals and 6 oral sites (40). We found that many samples were positive for the phage families that were targeted via four of the primer sets (Figure 2). To further explore these phage families, we developed instructional videos and collection kits to recruit volunteers for our study (see Methods). 700 additional oral samples from 100 different individuals were collected, the results of which will be discussed in later sections. We will focus our discussions on the most ubiquitous phage families, namely HA, HB1 and PCA2.

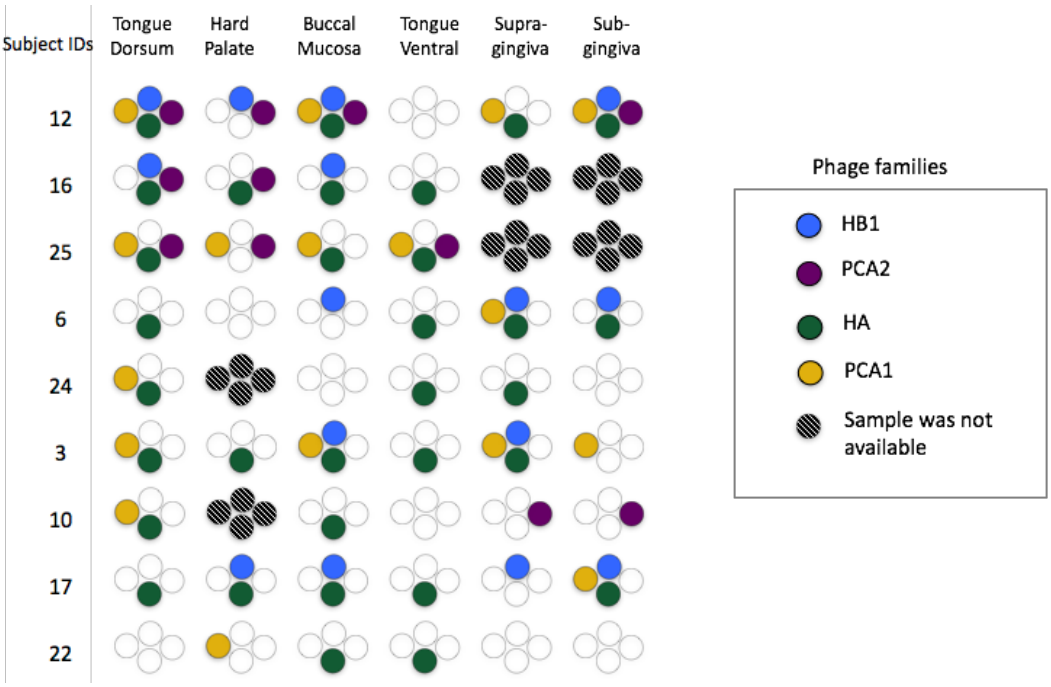


Figure 2. A qualitative depiction of phage family presence in oral samples collected from 10 healthy individuals and 6 different oral sites. Filled-in circles indicate presence of a phage family, and blank circles correspond to absence of that phage family.

5.2.B An exploration of three phage families reveal the presence of highly personal phage communities with varying degrees of conservation across different oral sites

As previously defined in the introduction, due to the high depth of coverage afforded through targeted sequencing, the relative abundance of OTUs provides a highly reproducible snapshot of the phage community composition (see SI). As shown in Figure 3 and Figure 4, phage community composition is highly skewed towards one or two dominant OTUs. However, there are also many other OTUs with abundance values that are consistent across space (different oral sites) and time. Generally, the dominant OTUs are not the same across different individuals, and the presence of numerous other OTUs with stable relative

abundances, gives rise to phage community compositions that appear highly personal. Therefore, we sometimes use "phageprint" as shorthand to refer to a community composition plot.

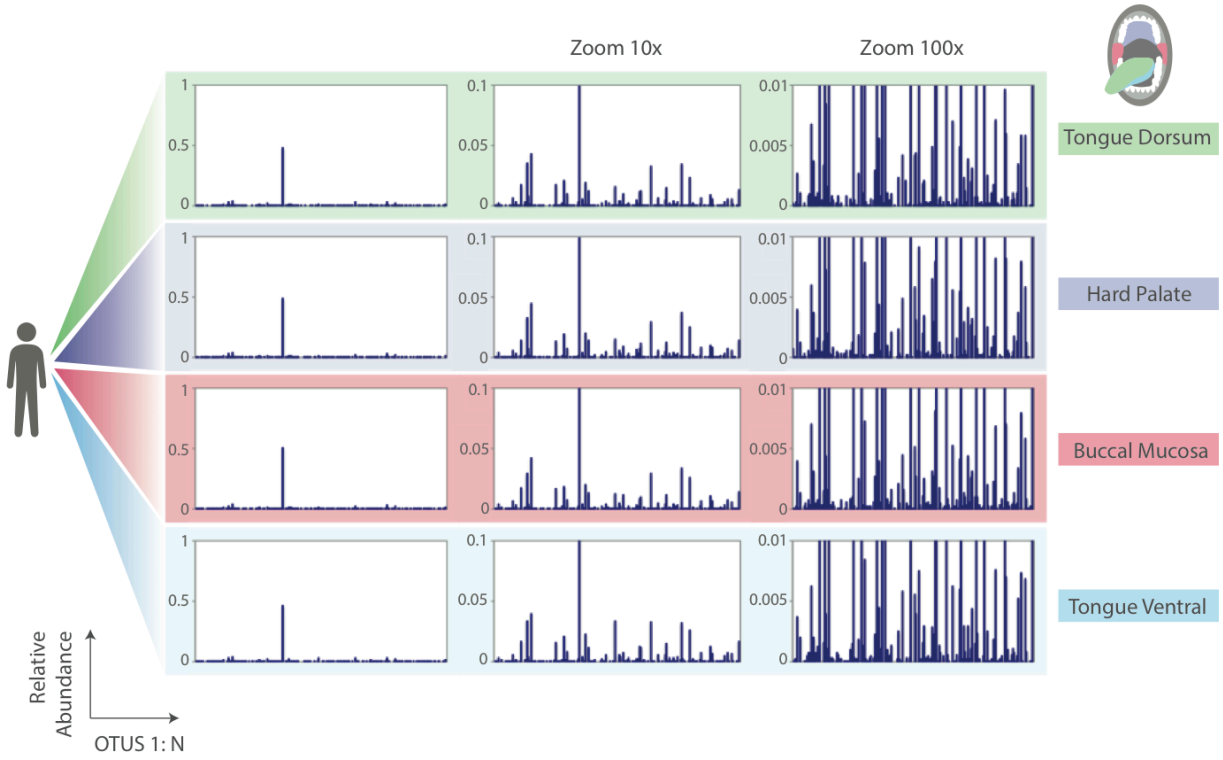


Figure 3. HA phage community compositions (phageprints) across 4 different oral sites in subject 16. Each phageprint is derived from the analysis of 4000 sequences (see SI). OTUs are defined at 98% sequence similarity and OTUs with less than or equal to 0.1% relative abundance across all phageprints were filtered out (see SI).

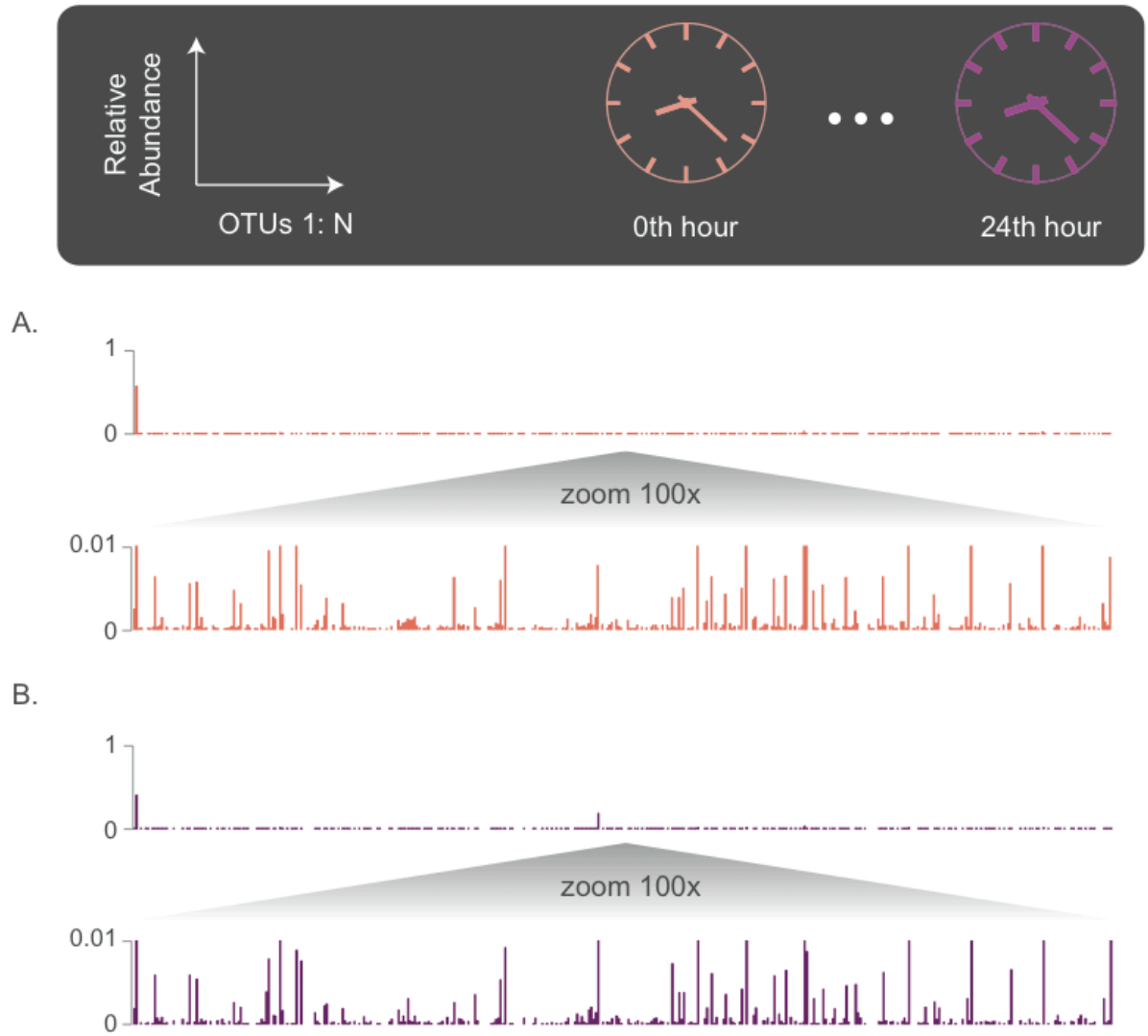


Figure 4. HA Phage community compositions (phageprints) from subject 37 at two different time points. Samples were collected from the tongue dorsum. A) Subject 37's phageprint at 0^{th} time point, collected right after brushing tongue dorsal and teeth surfaces. B) Subject 37's phageprint 24 hours after the initial time point (no brushing in between time points). Each phageprint is derived from the analysis of 4000 sequences (see SI). OTUs are defined at 98% sequence similarity.

To further quantify the differences between phage community compositions, we depict each pairwise comparison of community compositions by their Pearson correlation

coefficient (r-value). In comparing any pair of communities, there were m OTUs. Pearson correlation coefficient of community compositions A and B, was calculated according to $\frac{\sum_{i=1}^{m}(A_i-\bar{A})(B_i-\bar{B})}{\sqrt{\sum_{i=1}^{m}(A_i-\bar{A})^2}\sqrt{\sum_{i=1}^{m}(B_i-\bar{B})^2}}$. Here, A_i and B_i are relative abundance values of OTU_i in community compositions A and B, respectively, and $\bar{A}$ and $\bar{B}$ are average OTU relative abundances for community compositions A and B. A k by k matrix of Pearson correlation coefficients was created and shown as a heatmap for each phage family, with k representing the number of phage community compositions.

Figure 5 summarizes the Pearson correlation matrix for the HB1 phage family across 9 individuals and 4 different oral sites. An immediately recognizable pattern is that the phage community compositions of an individual are highly correlated. In stark contrast are the correlations between the phage community compositions of different individuals.

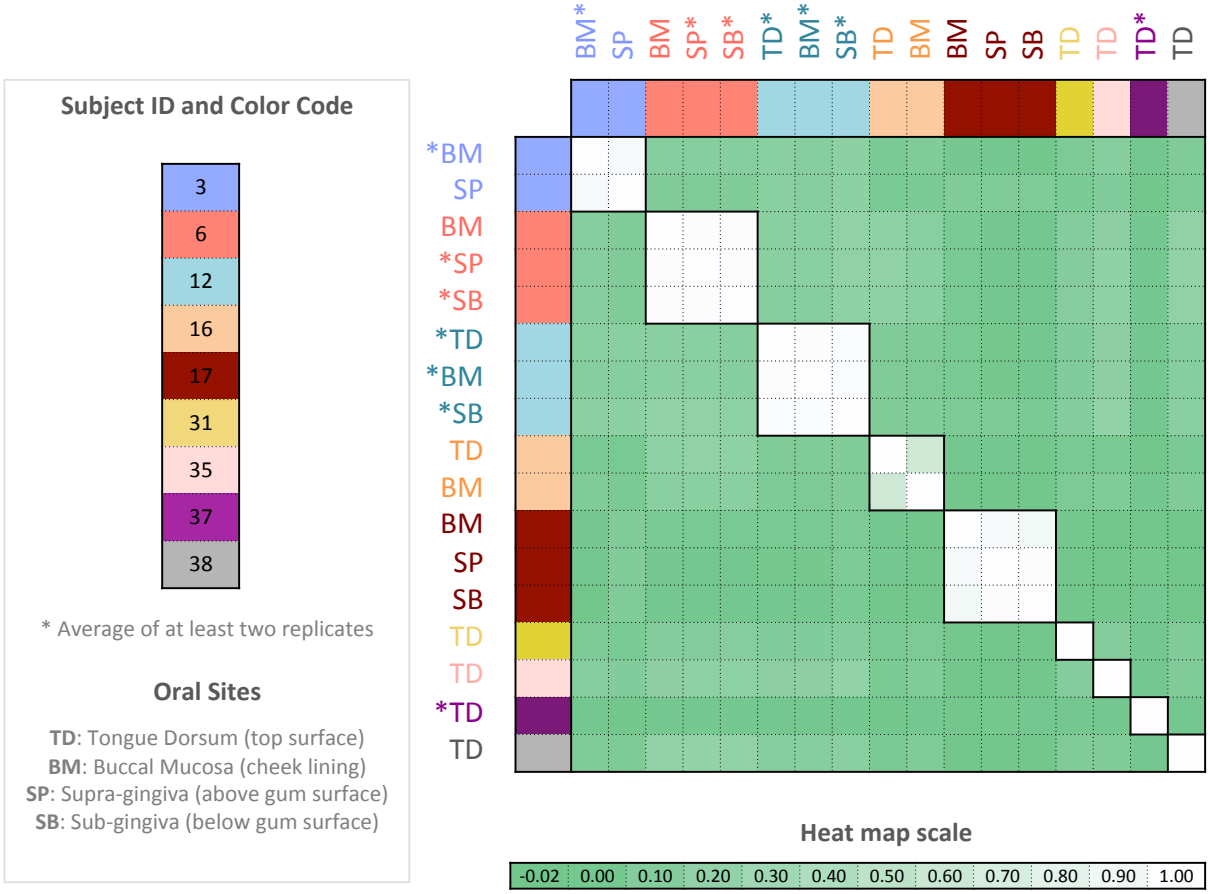


Figure 5. Pearson correlation coefficient matrix of HB1 phage community compositions spanning 9 subjects and four oral sites. Each community composition is derived from the analysis of 4000 sequences associated with an individual and a particular oral site. OTUs are defined at 98% sequence similarity and OTUs with less than or equal to 0.1% relative abundance across all phageprints were filtered out (see SI). Phageprints are color-coded based on the individual they originate from. Community compositions that have been replicated at least twice and averaged have an asterisk next to them (see SI).

As in the case of the HB1 phage family, there is low to non-existing correlation between the HA phage community compositions of different individuals (Figure 6), reinforcing the notion of highly personal phage communities. However,

unlike HB1, not all oral sites within the same subject are highly or even moderately correlated (see subjects 3, 12, and 17). In subject 12 for example, the tongue dorsum has a correlation close to zero with supra-gingiva and sub-gingiva sites, which are nearly perfectly correlated. Similarly, in subject 3, the hard palate and the tongue ventral surface have nearly identical phage community compositions while they have a very low correlation with the community at the tongue dorsum. However, unlike subject 12, the tongue dorsum in subject 3 seems to be an intermediate community, having a moderate correlation with all other sites that are distinct from each other. In subject 17 as well, buccal mucosa serves as the intermediate community, having a moderate correlation with the disparate communities of sub-gingiva and the hard palate.

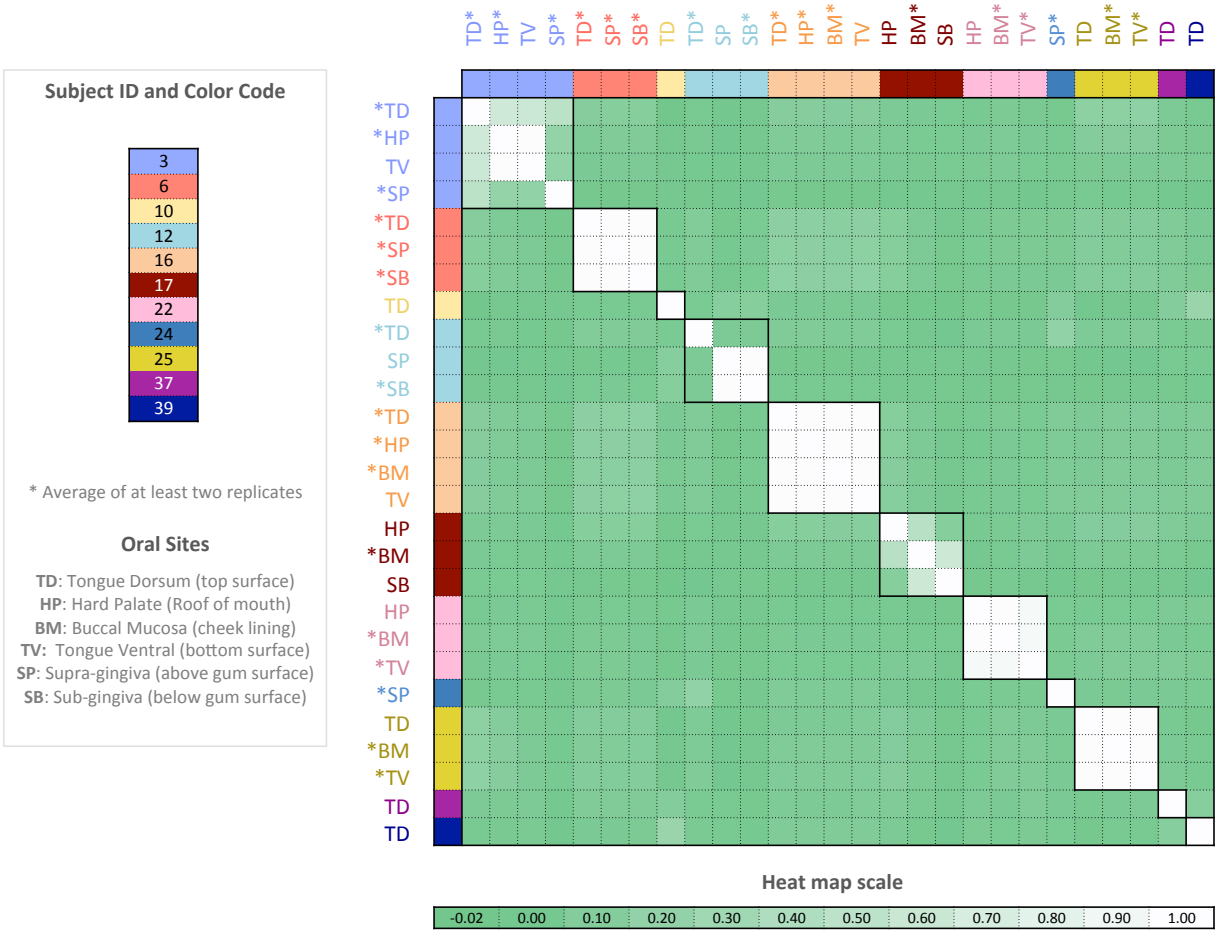


Figure 6. Pearson correlation coefficient matrix of HA community compositions encompassing 11 subjects and six oral sites. Each community composition is derived from the analysis of 4000 sequences associated with an individual and a particular oral site. Samples are color-coded based on the individual they originate from. Oral sites shown are the tongue dorsum (TD), buccal mucosa (BM), supra-gingiva (SP), sub-gingiva (SB), hard palate (HP), and ventral surface of the tongue (TV). Samples whose community composition has been replicated at least twice and averaged have an asterisk next to them (see SI).

SI Figure 1 represents yet another phage family's Pearson correlation coefficient matrix. As with the HA and HB1 phage families, PCA2 phage community compositions appear specific to individuals, and in some cases they are even specific to an individual's oral

site. While subject 16 and subject 25 have highly correlated phage community compositions across different oral sites, subjects 10 and 12 have uncorrelated phage communities separated by intermediate communities similar to subjects 3 and 17's HA phage communities (Figure 6).

In contrast to the other phage families, which were found across all oral sites, the HB1 phage family is absent from the hard palate and the tongue ventral surfaces. The HB1 communities at the gingival sites and the buccal mucosa tend to be highly correlated, and the buccal mucosa community is moderately correlated with the tongue dorsum community (SI Figure 2). Similarly, the HA phage communities at the gingival sites are highly correlated (SI Figure 2).

Using bipartite network diagrams (Figure 7, SI Figure 3, SI Figure 4, SI Figure 5) we can further examine phage communities. For example, the HB1 phage family network degree distribution appears to fit a power-law model, with the majority of OTUs occurring only in one subject. Using these networks we can begin to see the differences between the community compositions at different oral sites that were identified as only moderately correlated using the Pearson correlation matrices. In viewing the HB1 phage family network (SI Figure 5), we can see that in subject 16, in addition to OTU 5, which is abundant on the tongue dorsum and the buccal mucosa, OTU 8 has exclusively thrived on the buccal mucosa, while OTU 11 is only pervasive on the tongue dorsum. Similarly, subjects 17 and 3 each have a second-abundant OTU that is exclusive to the buccal mucosa surface, distinguishing this site from the phage communities at gingival sites, which they are moderately correlated to.

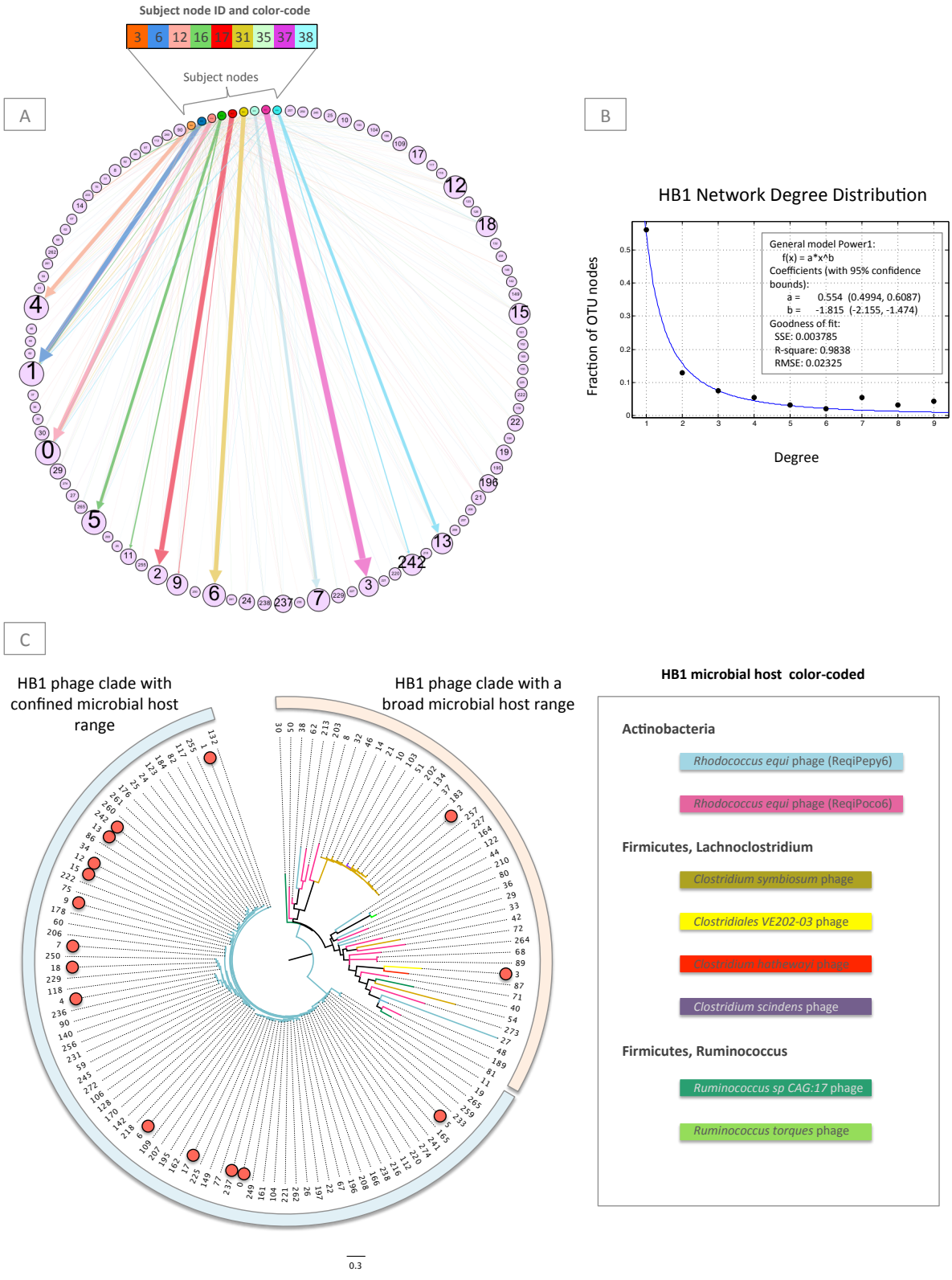


Figure 7. HB1 phage-host network and degree distribution. A) Two types of nodes exist: OTU nodes (purple), and subject nodes. Subject nodes and

edges are color-coded based on the individual they represent. Each directed edge connects an individual to a phage OTU that he/she harbors, and the edge weight is proportional to the relative abundance of the OTU in that individual's oral community. OTU node sizes and labels are proportional to the number of individuals the OTU is present in. For OTU nodes, the node ID is the OTU ID, which can be matched to IDs in SI Table 1 and SI Table 2 for obtaining taxonomic information regarding each OTU's representative sequence. Refer to SI Figure 5 to see a similar network that shows OTUs present at different oral sites. B) The degree distribution of HB1 phage network. This distribution was obtained by making a histogram of OTU in-degrees, or the number of individuals that each OTU is present in. C) Maximum Likelihood phylogenetic tree of OTU representative sequences (see SI Figure 7 for their nucleotide alignment) for the HB1 phage family. Red circles are placed next to OTUs that are prevalent (appear in multiple samples). Branches are color-coded based on the organism in whose genome the closest homolog of the OTU representative sequence was found (see SI Table 1, SI Table 2).

In all of our analysis so far, we have used the Pearson correlation because it allows for a coarse-grained depiction of differences between phage communities, as it is dominated by the most abundant OTUs. We will also use other distance metrics to explore different facets of phage communities in later sections. Moreover, we have carefully explored the effects of sequence similarity thresholds when defining OTUs, and we discovered that the phage community compositions, as represented by pairwise Pearson correlation coefficients, are highly robust to sequence similarity thresholds used for defining OTUs (see SI).

5.2.C A bioinformatic search for the bacterial hosts

Because we aimed to study previously uncharacterized phages, the bacterial hosts for the phage families in this study have not yet been cultured or identified. However, using homology search we can identify candidate host species. Figure 7 (right panel) demonstrates organisms that were found to have terminase sequences homologous to HB1 sequences found in the human mouth. Each OTU's most abundant sequence served as its representative sequence and was used as a query for BLASTx homology search. With the exception of a few sequences tagged as "putative proteins", all resulting homologs were terminase sequences with very low E-values (less than 10^{-19}) (SI Table 1). This was true for the HA and the PCA2 marker sequences as well (SI Table 3 and SI Table 5

SI Table 5).

The majority of HB1 homologs belonged to ReqiPepy6 phage isolated from *Rhodococcus equi,* a member of the Actinobacteria phylum. Recognized as a major pathogen in foals and an emerging pathogen in immunocompromised humans, *R. equi* inhabits soil as well as a diverse range of organisms such as cattle and pigs (49). Other OTU homologs were matched to ReqiPoco6, another *R. equi* phage, and six species spread across two different families of the Firmicutes phylum.

In the phylogenetic tree composed of HB1 OTU representative sequences (Figure 7), two major clades have formed. The first clade is entirely composed of sequences that are ReqiPepy6 phage terminase homologs. The second clade contains a mixture of sequences with homologs captured from phages of two different host phyla and three different families. Despite the great diversity and host range observed in sequences from the second clade, red circles placed next to prevalent OTUs (those that appear in multiple individuals) (Figure 7) reveal that the majority of these OTUs belong to the first clade. Moreover, the

HA phage family infects only a single genus of Firmicutes (*Streptococcus)*, but appears embedded in the genomes of many different species within this genus (SI Table 4).

5.2.D Oral phage community temporal dynamics in the span of 30 days

We have so far demonstrated the highly personal nature of phage communities residing in the human mouth. To better understand the temporal stability of these phage communities, 10 subjects collected one sample per day (from tongue dorsum) for 30 days. The HB1 community composition as it evolved over 30 days inside subject 1 is depicted in Figure 8. Here, to provide a more detailed view of this community, we cluster the reads into OTUs based on 100% similarity (hence the 10,000 OTUs that are listed). These OTUs are shown in Figure 8 ordered based on their phylogenetic distance.

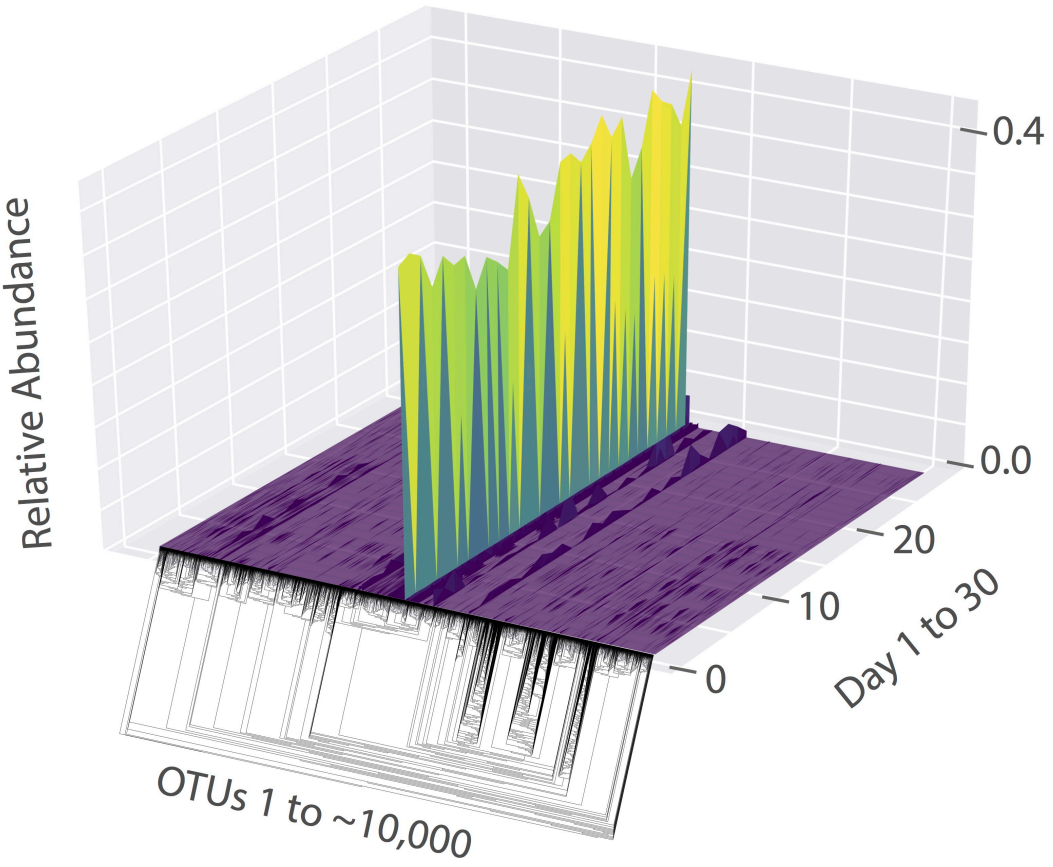


Figure 8. A 3D surface plot depicting the HB1 phage community composition as it evolves over 30 days on subject 1's tongue dorsum. The x-axis contains ~10,000 OTUs ordered according to the depicted phylogenetic tree of the OTU sequences (the phylogenetic tree is provided largely to serve as a schematic since it is hard to visualize the details of this tree). Each OTU is composed of identical sequences (i.e. 100% sequence similarity threshold). The y-axis depicts the relative abundance of each OTU, and the z-axis shows the fluctuations in relative abundance of each OTU in time. (Note, the colors shown correspond to an abundance-based heatmap, which provides redundant information since the y-axis provides this information, however the colors are kept because they allow for a better visualization of fluctuations in low-abundant OTUs.)

Surprisingly, over 30 days, the main features of the HB1 community composition is preserved, though there are also interesting fluctuations that are well above the experimental error and detection threshold (see SI). Figure 9 demonstrates different degrees of temporal stability and phylogenetic diversity across individuals. However, a global trend is that the dominant OTUs remain dominant over the span of 30 days in all subjects.

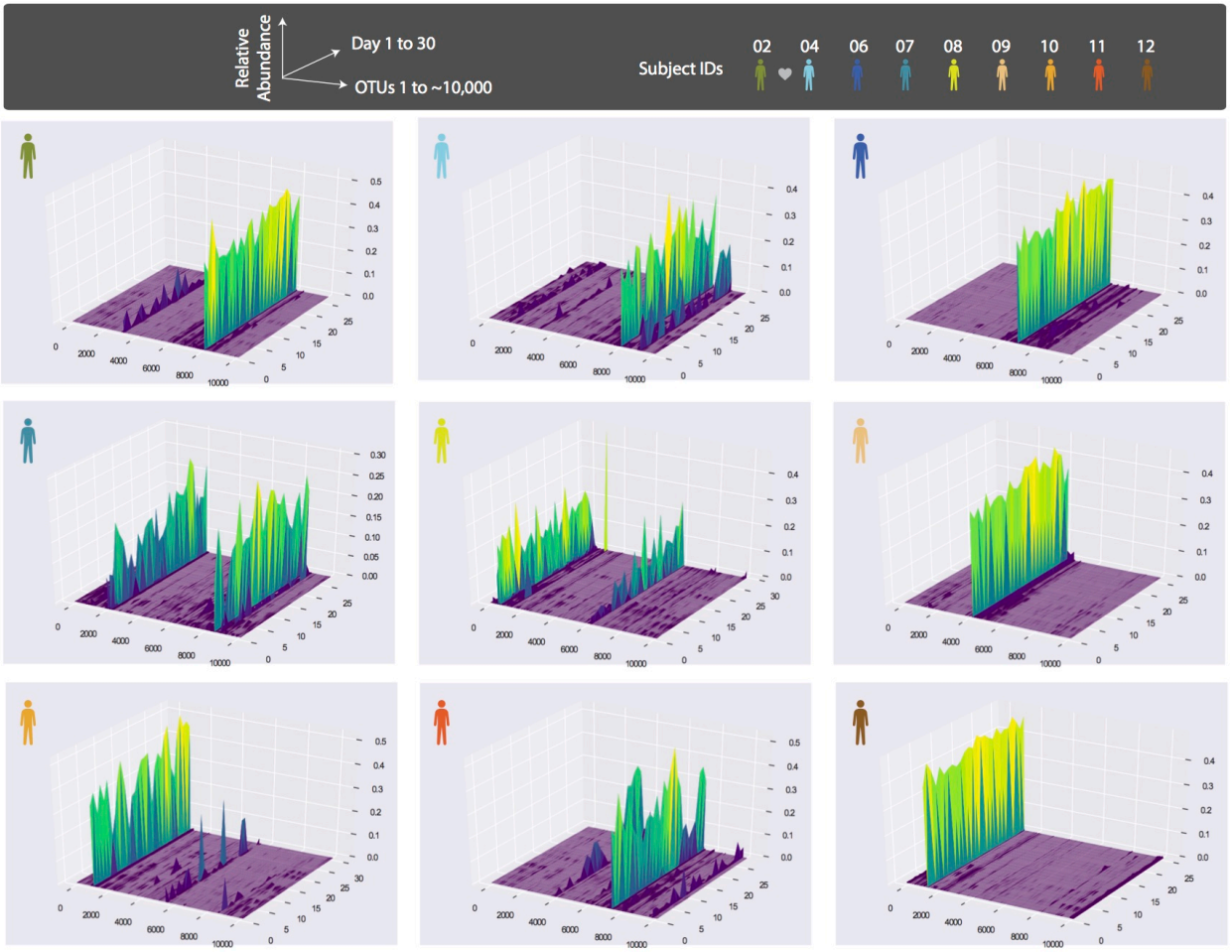


Figure 9. Depictions of HB1 phage community evolution in different subjects over 30 days. The format of the plots is the same as that of Figure 8, and the order of OTUs is based on their phylogenetic distance and identical across all plots. All samples are collected from the tongue dorsum. Note that subject 2 and 4 are a couple, and their phage community compositions share some main features.

To further examine the phage community dynamics, we used several distance metrics to create pairwise comparisons within and between phage community compositions (Figure 10). The first metric explored is binary Jaccard distance, which is equal to one minus the ratio of the intersection to the union of two samples' OTUs: $1 - \frac{|A \cap B|}{|A|+|B|-|A \cap B|}$. Here, A and B represent the OTUs that are present in sample 1 and 2, respectively. This is a binary method of comparing samples simply based on the presence/absence of the OTUs. In addition to the Pearson distance (1- Pearson correlation), we chose two other abundance-based distance metrics, namely abundance-weighted Jaccard and Bray-Curtis. Abundance-weighted Jaccard, which is equal to $1 - \frac{UV}{U+V-UV}$ (50), is similar to Jaccard but here U and V represent the sum of relative abundances of OTUs shared between samples 1 and 2, respectively. Bray-Curtis dissimilarity (51) is defined as $\frac{\sum |x_{ik} - x_{jk}|}{\sum x_{ik} + x_{jk}}$, where x_{ik} and x_{jk} correspond to the relative abundance of OTU k in samples i and j.

Lastly, we explored unweighted Unifrac, a phylogenetic distance metric (52). The Unifrac algorithm operates on a phylogenetic tree containing sequences from all samples. It proceeds to create pairwise comparisons between samples by identifying the branch lengths that are shared between two samples, as well as the branch lengths that are unique to each sample. The Unifrac distance is then defined as the unshared branch lengths divided by the total branch lengths, where total branch lengths is the sum of shared and unshared branch lengths. If two samples are identical, the fraction of the tree's branch lengths that is unique to one sample or the other will be zero, and thus, the Unifrac distance will be zero.

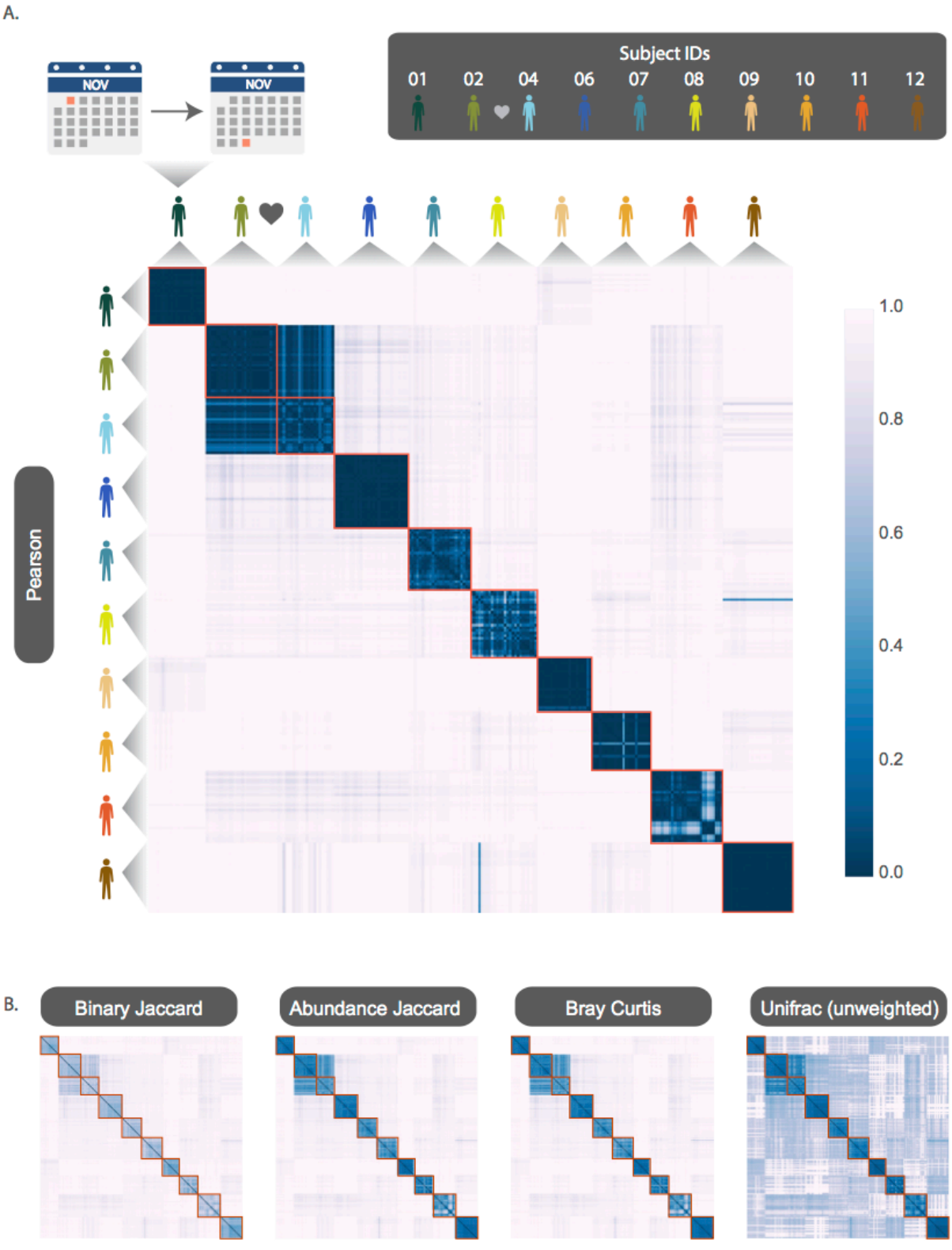

A.
NOV
NOV
Subject IDs
01
02
04
06
07
08
09
10
11
12
Pearson
1.0
0.8
0.6
0.4
0.2
0.0
B.
Binary Jaccard
Abundance Jaccard
Bray Curtis
Unifrac (unweighted)

Figure 10. HB1 phage community temporal dynamics (previously shown graphically in Figure 9) depicted here by pairwise distance metrics: A) Peason, B) Binary Jaccard, Abundance Jaccard, Bray Curtis and unweighted Unifrac. The heatmap scale applies to all heatmaps shown. Subjects 02 and 04 are a couple. Samples from each subject are chronologically ordered.

All distance metrics explored paint similar pictures of the HB1 phage communities, depicting them as highly personal and stable over time. Because phage communities in different individuals have such distinct composition (Figure 9) abundance-based metrics are especially suitable for describing them. However, even the binary Jaccard and weighted Unifrac distance metrics demonstrate a similar message, which is consistent with HB1 network's degree distribution (Figure 7). Figure 11 further demonstrates the intra-and - interpersonal distances as measured through these various distance metrics. As is expected from the heatmaps shown in Figure 10, the intra-personal distances are markedly different from the inter-personal, with the notable exception being subject 2 and 4, who are the couple in the study.

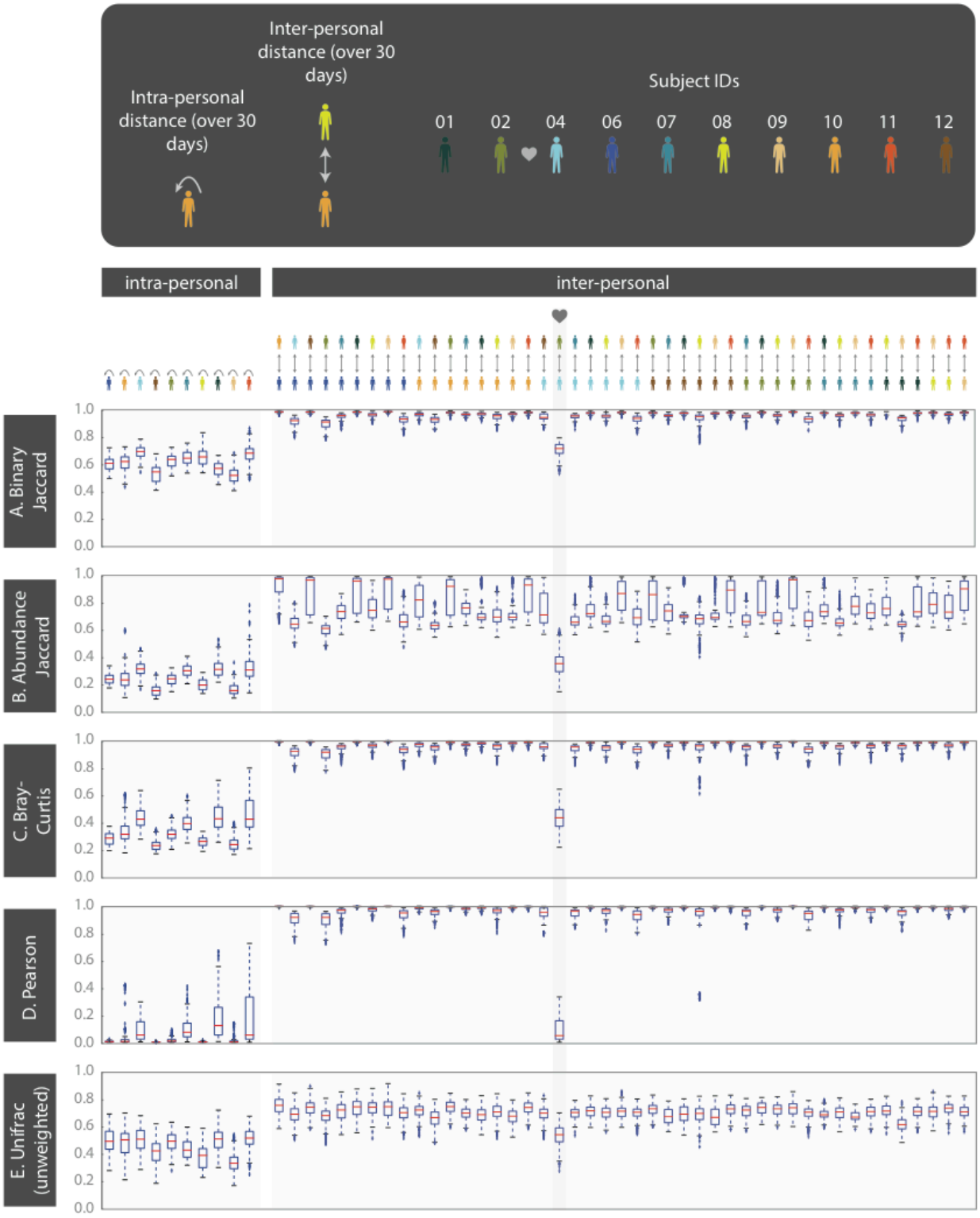

Inter-personal distance (over 30 days)
Intra-personal distance (over 30 days)
Subject IDs
01
02
04
06
07
08
09
10
11
12
intra-personal
inter-personal
A. Binary Jaccard
B. Abundance Jaccard
C. Bray-Curtis
D. Pearson
E. Unifrac (unweighted)
1.0
0.8
0.6
0.4
0.2
0.0

Figure 11. Intra-and inter-personal distances between HB1 phage communities from 10 subjects, over the span of 30 days (further quantifying the heatmaps from Figure 10). Box-plots depict distances from pairwise comparisons made using the following metrics: A) Binary Jaccard, B) Abundance Jaccard, C) Bray-Curtis, D) Pearson, and E) unweighted Unifrac. The outliers defined as those outside of the 1.5 x IQR (inter-quartile range) are denoted by "+". The box-plots corresponding to the comparisons between the couple in this study are highlighted.

5.2.E Phage community comparisons across siblings, couples, and non-related individuals residing across the globe

Given the ubiquitous presence of the phage families across subjects residing in the U.S. we wondered whether phage families (HA and HB1, specifically) are globally distributed, and whether subjects residing in the same country would have more similar phage communities. We discovered that phage families were in fact found in individuals from various ethnicities, nationalities, and ages. Surprisingly, neither from abundance-based nor phylogenetic distance comparisons did we find an indication that people residing in the same country share more similar phage communities (Figure 12). We found individuals typically have highly unique phage communities.

Even siblings who were either living in the same household or had previously, do not have any more similar phage communities than unrelated individuals. In fact, one of the four sibling groups with uncorrelated phageprints are identical twins (Figure 12). However, 3 out of 4 couples in this study exhibited highly similar phage communities. These results suggest that genetics and cohabitation do not significantly impact a person's oral phage community. The more impactful factor appears to be direct oral contact with another person. To further

test these trends, larger studies encompassing a greater number of individuals and regions in the world are required.

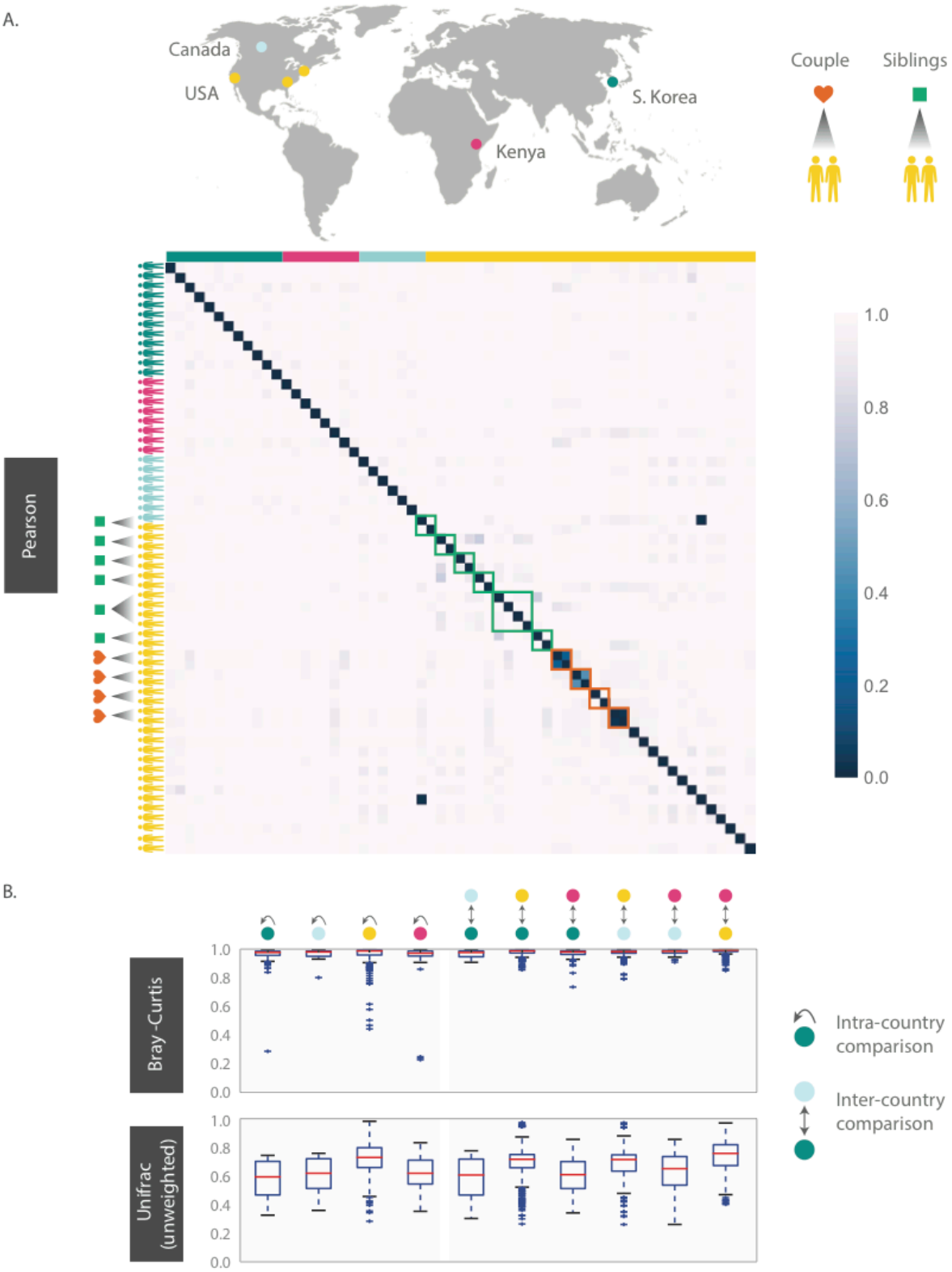

A.
Canada
USA
S. Korea
Kenya
Couple
Siblings
Pearson
1.0
0.8
0.6
0.4
0.2
0.0
B.
Bray-Curtis
Unifrac (unweighted)
1.0
0.8
0.6
0.4
0.2
0.0
Intra-country comparison
Inter-country comparison

Figure 12. HB1 phage community across 61 individuals residing across different parts of the globe. Samples are obtained from the tongue dorsum. A) Pearson distance (1 – Pearson correlation) is shown as a heatmap. A subset of individuals residing in the U.S. are either couples or siblings. Green and red boxes are drawn around samples from each sibling group and couple, respectively. B) Intra- and inter-country distances from pairwise comparisons made using Bray-Curtis and unweighted Unifrac distance metrics. The outliers are denoted as points outside of the 1.5 x IQR (inter-quartile range). Siblings and couples are excluded from this analysis.

5.3 Discussion

Our method for finding ubiquitous human oral phages relied on a relatively small metagenomic dataset, which contained sequences from 6 individuals residing in Spain (47). Yet, on the basis of markers designed from this small dataset we were able to identify the same phage families in at least 10 times as many individuals from across the globe. This finding seems to suggest that despite the great sequence diversity that has been revealed through viral metagenomic surveys, there exist certain phage families that are a stable feature of the human oral microbiome. Studies of phages from various natural environments (e.g. marine, soil, lakes) also report the finding of phage families that are distributed across similar types of habitats despite vast geographical distances and barriers that exist between these habitats (9, 53, 54). The discovery of core bacterial members within the human microbiome (40, 55, 56) that are present globally (42) further support our discovery of globally distributed phage families. Similar to our findings for phages, the oral bacteria of individuals from the same part of the world was as different from each other as they were to individuals from other parts of the world (42).

The ubiquitous presence of the identified phage families in individuals, together with their temporal stability, seems to suggest that they likely play important roles in this environment. The observed temporal stability of these phage families over the span of a month is supported through metagenomic studies of oral phages (57, 58) as well as 16S sequencing of bacterial communities inhabiting various sites in the human body (55, 59, 60). Our study represents one of the largest studies of human oral phages. As a comparison, the most recent version of the Human Microbiome Project, contains samples from 265 individuals (61). However, future studies are required to more systematically account for diet, ethnicity, and other parameters across tens of thousands of individuals across the globe.

A particularly important aspect of our study is that it combined the advantages of metagenomics with targeted sequencing to not only identify core phage families inhabiting the human oral cavity, but to also characterize their communities with a resolution that is unavailable through metagenomic studies of phages. This detailed view allowed us to clearly observe the highly complex, and personal nature of phage community compositions. Moreover, the emergence of phageprints is directly the result of the remarkable phage sequence diversity that we were able to capture via targeted sequencing. For example, we observed a few hundred HB1 OTUs (defined at 97% sequence similarity). This is a staggering number when considering that this is about the same number as the total number of bacterial OTUs (defined at 97% sequence similarity) in the human mouth (47, 55). Even though the HB1 phage family is only one of many oral phage families, it by itself contains the same level of sequence diversity as the entire bacterial population in the human mouth. This perhaps explains the need to use highly elaborate algorithms applied onto 16S and metagenomic sequences from all bacterial strains to be able to identify a person based on

his/her microbiome with a similar level of accuracy as our method, which only employs sequences from one phage family (62).

Because of the great diversity of sequences associated with just one phage family, in our study of about 60 individuals we found only one case where unrelated individuals whose phageprints had similar correlation coefficients. This may be due to experimental error or due to the course-graining associated with Pearson correlation matrices. However, if we conservatively assume that each phage family can provide only 50 unique patterns, then the combination of phageprints from just 6 phage families would already provide a greater number of possible patterns than the size of the current human population. This is assuming that these phage families will be physiologically independent of each other. Future studies are needed to test the long-term stability of the human phageprints especially with regard to perturbations such as exposure to antibiotics. To our knowledge, this is the first study that demonstrates the potential application of phage sequences for human identification.

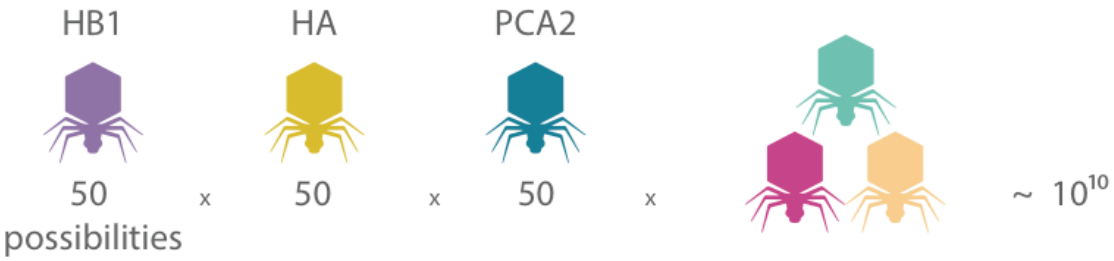


Figure 13. An estimate for the number of additional globally-distributed phage families needed to achieve the number of possible phageprint patterns that surpass the current human population. Assuming that phageprints from each phage family can provide 50 unique patterns, there would only be 3 additional globally distributed phage families needed.

5.4 Materials and Methods

Materials and Methods (an overview). This section will comprise a condensed description of phage marker discovery, experimental methods such as sample collection, barcoded PCR, sequencing, and experimental reproducibility analysis, as well as bioinformatic methods such as sequence quality control measures, sequence demultiplexing and clustering. Figure 14 provides a schematic summary of some of the main experimental and bioinformatic methods employed.

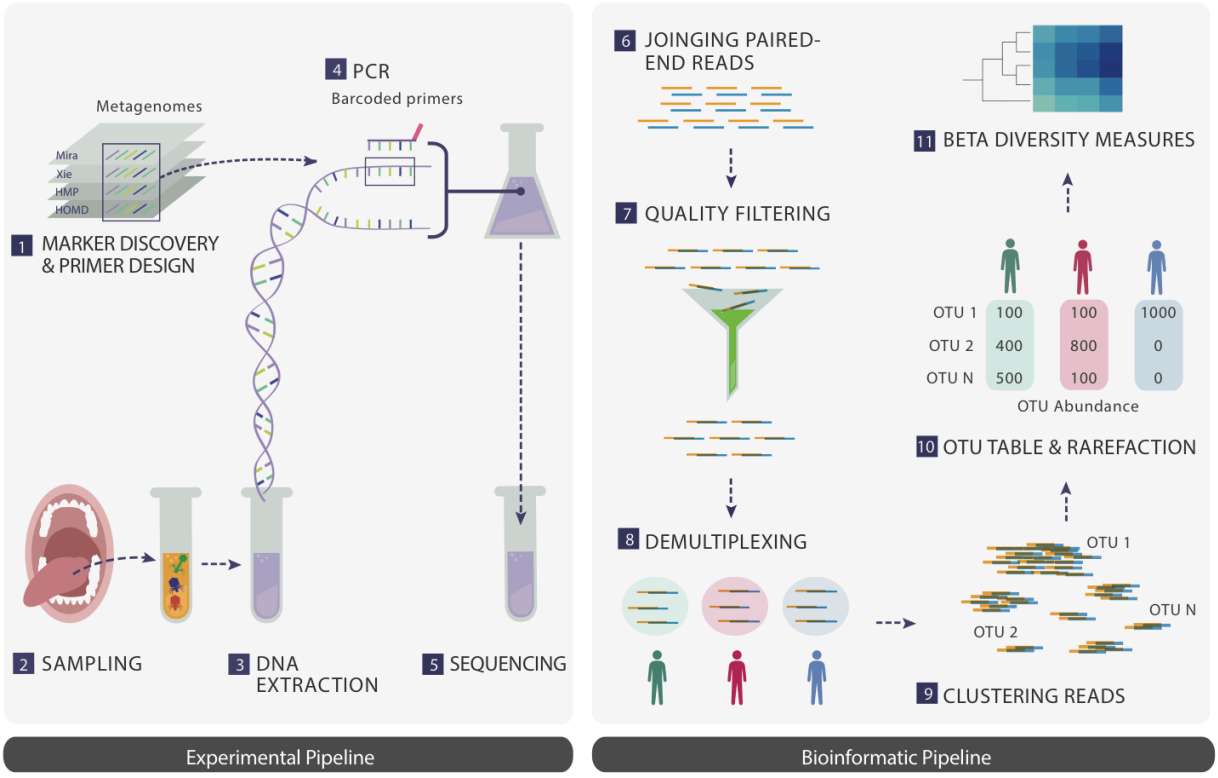


Figure 14. A schematic summary of the main experimental and bioinformatic methodologies presented in this chapter: 1) Discovery of ubiquitous phage families by examining large terminase sequences that occur across different metagenomic datasets, 2) sampling, 3) DNA extraction from oral biofilm samples, 4) PCR using barcoded primers followed by PCR clean-up, 5) paired-end Next Generation Sequencing, 6) joining paired-end reads to

eliminate sequencing errors, 7) various additional quality control steps to further eliminate errors, 8) demultiplexing of reads based on their barcode sequence and tagging sequence names by the sample they originate from, 9) gathering reads from all samples and clustering them based on sequence similarity (OTUs), 10) counting the number of sequences belonging to each OTU from each sample (i.e. constructing an OTU table), and rarefying the table so that each sample is represented by the same total number of sequences, and 11) performing various down-stream diversity analysis (e.g. community composition plots or phageprints) using the constructed OTU table as the basis.

Phage marker discovery. Focusing on the oral cavity, we applied a metagenome clustering program called MetaCAT (63) to eight oral metagenomes obtained from six human subjects with varying degrees of oral hygiene (47). MetaCAT reports the most abundant genes in a metagenome based on an annotated reference library provided by the user. As a reference library we used the viral RefSeq database (v48) (64) containing 2727 distinct viral taxonomic IDs, spanning 96713 viral gene records. Applying MetaCAT to the eight Mira metagenomes yielded 8757 known RefSeq viral genes. This list was manually screened for RefSeq genes containing the case insensitive keywords "terL", "terminase", or "packa" in the definition of the known reference gene or in the features field of the GenPept file provided by MetaCAT. From the 1.9×10^6 contigs, 119 contained TerL genes, which when cross-BLASTed against a second oral metagenomic dataset obtained from a healthy individual (46), reduced to eleven candidate markers.

After removing redundant markers using additional alleles recovered from public genomic oral datasets we found seven full length TerL genes representing unrelated lineages (maximum percent identity < ~30%), labeled HA, HB1, HB2, PCA1, PCA2, AB1, and AB2,

all present in a context of phage-related genes. The labels were made according to the database from which each marker was identified (H=healthy, PC=past caries, A=active caries, C=cavities), followed by the patient index (patient A or B for the given dental category), and a counter of the TerL solution found in the given metagenome. Moreover, in order to derive markers that span the full length of the TerL gene we applied a bootstrapping approach in which the seven TerL+ contigs were used as hooks to recover additional alleles from several datasets. This was achieved by BLASTing the amino acid sequence of the seven representative contigs against the Xie, Mira, HOMD (Human Oral Microbiome Database)(43) and the HMP (Human Microbiome Project) (48) datasets. After identifying markers and designing degenerate primers against them, we began to test for their presence in a variety of samples.

Measures against contamination. A common source of contamination in PCR originates from previously amplified template sequences that enter new PCR reactions. To prevent contamination this type of contamination, four physically separated workstations were developed for DNA extraction (station A1), PCR preparation (station A2), PCR and gel electrophoresis (station B1), and PCR cleanup (station B2). A and B specify two different buildings at Caltech while 1 and 2 refer to two different rooms within the same building. The flow of materials was from building A to B and never the vice-versa. Every station had its own set of lab equipment, materials, and storage space. Disposable lab coats (Sigma-Aldrich®) were worn and disposed of at the end of every procedure to ensure that DNA was not carried between stations via clothing. Facemasks (Fisher Scientific) were also worn at all times to prevent any oral or nasal droplets from entering reactions. Prior to the start of every DNA extraction, lab equipment and bench tops were cleaned using sterile wipes and DNA AWAY™ (Thermo Scientific), a surface decontaminant that eliminates

DNA and DNAses. PCR preparations and aliquoting of reagents were carried out in a PCR flowhood (AirClean® Systems) equipped with a UV light and laminar airflow capabilities. Lab equipment required for PCR preparation was designated to the PCR preparation flowhood. At the end of every experimental session and when introducing new equipment into the flowhood, all surfaces were first wiped with DNA AWAY™ solution and then exposed to UV radiation for 60 minutes. Prepackaged, sterile gloves were used for PCR preparation. To prevent sample-to-sample contamination during DNA extraction, PCR preparation, and PCR cleanup, gloves were frequently exchanged. Most importantly, 5 No Template Control (NTC) reactions accompanied every PCR run. Similarly, to test the presence of contaminants in extraction reagents, for every extraction experiment, 3 reactions were carried out without the addition of any sample. PCR using phage primers was performed on these extraction control reactions.

DNA Extraction (Station A1). DNA extraction of human oral samples was done according to the manual from MoBio PowerBiofilm® DNA Isolation Kit. The advantage to using this kit for DNA extraction and purification is that it combines the use of chemical and mechanical (bead-beating) treatments for an increased efficiency in biofilm disruption, lysis, and removal of inhibitors such as humic acid. The final concentrations of DNA were measured using Nanodrop. The concentration range of the total extracted genomic DNA was typically between 5 to 50 ng/μL.

PCR preparation (Station A2) and PCR (Station B1). Each PCR reaction contained 12.5 μL of PerfeCTa® qPCR SuperMix, ROX™ (Quanta Biosciences), a premix containing AccuStart™ Taq DNA polymerase, $MgCl_2$, dNTPs, and ROX reference dye for qPCR applications. Additionally, each reaction contained 10.5 μL of RT-PCR Grade Water (Ambion®) which is free of nucleic acids and nucleases, 1 μL of extracted DNA at 1 ng/μL,

0.5 μL of forward and 0.5 μL of reverse primers, each at 50 ng/μL (synthesized by IDT). A higher than recommended primer concentration was used because the phage primers used are 32-64 fold degenerate. The thermocycling protocol was made according to PerfeCTa qPCR SuperMix recommendations: 1) a 10-minute activation of AccuStart™ Taq DNA polymerase at 95°C, 2) 10 seconds of DNA denaturation at 95°C, 3) 20 seconds of annealing at 60°C, and 4) 30 seconds of extension at 72°C, 40 cycles repeating steps 2 to 4, followed by 5 minutes of final extension at 72°C.

Gel electrophoresis (Station B1) and PCR cleanup (Station B2). Phage PCR products were visualized using 2% agarose in TAE buffer. After gels were cast, 5 μL of each PCR product was mixed with 1 μL of 6X loading dye and loaded into a well. 5 μL of 100 base-pair ladder was used, and the gel electrophoresis instrument was set to run for 30 minutes at 100V. Phage PCR positive hits were purified using the QIAquick PCR Purification Kit (QIAGEN). 20μL of PCR products were used and purified according to the QIAquick PCR Purification manual.

Illumina sequencing. Upon PCR cleanup, double stranded DNA concentration in each sample was measured using Qubit instrument. Qubit measurements were performed in Building C due to practical considerations rather than a necessary treatment for preventing contamination. Samples were combined into one reaction (~2 μg dsDNA) and submitted to GENEWIZ, Inc for library preparation and MiSeq 2x300bp Paired-End sequencing.

DNA barcodes for multiplexed sequencing. To enable multiplexing, unique DNA barcodes (Table 1) were appended onto the forward primer sequences (Table 3) used to amplify each phage marker. These barcoded primer sequences were synthesized by IDT. Using this scheme, ~100 samples were submitted per MiSeq sequencing run (Table 1) and

by matching the barcode sequence to the sample ID, information about who and where the sample came from was accessible. More specifically, Hamady error-correcting 8-letter barcodes (65) were used. Hamady DNA barcodes are an example of Hamming code wherein the addition of parity bits allow for detection and correction of errors within the barcode sequence. In the case of Hamady barcodes, up to 2 errors in the barcode sequence can be detected and one error can be corrected.

Error-detecting barcode design. Because we were unable to use Hamady barcodes to amplify PCA2 marker sequences under various PCR conditions, 4-letter error-detecting DNA barcodes were designed. The general scheme used is that the last letter of the barcode sequence (Table 4) serves as a parity base and contains information about the preceding letters, enabling a check on single-base sequencing errors occurring in the barcode sequence. More specifically, in designing four-letter barcodes the first three letters in the barcode sequence together are set to represent a number between 1 and 64 written in base 4, and converted to DNA bases using the following assignments: 0 = A, 1 = C, 2 = G, and 3 = T. For example, to generate a barcode representing the number 60, we re-write 60 as $3 \times 4^2 + 3 \times 4^1 + 0 \times 4^0$, and convert the coefficients to DNA bases. As such, 60 can be coded as TTA. To enable error detection, the fourth letter in a barcode represents $\sum coefficients \ mod \ 4$, which is the remainder of the sum of the coefficients divided by 4. The full sequence of barcode #60 is TTAG. Although using 4 letters and 3 positions 64 barcode sequences are possible, certain barcode sequences were excluded due to repetition of bases and/or high GC content, for example barcode #42 (GGGG) and barcode #41 (GGCC). Note that barcode numbers 1 through 21, 24, 32, 33, 36, and 48, $\sum coefficients \ mod \ 4$ result in fractions, and had more barcodes been necessary for our

experiments it would be simple to include additional assignments for conversion of fractions into DNA bases.

Barcode verification. To verify the efficacy of all barcodes used in the experiments, PCR was once performed on each sample using non-barcoded primers (Table 1), and once again using barcoded primers. We were unable to use Hamady barcoded primers to amplify the PCA2 marker; therefore shorter, 4-letter error-detecting barcodes were designed and experimentally verified (Table 2).

The sample collection kit and measures against sampling contamination. To obtain samples, we developed a sample collection kit and prepared kit contents within the PCR flowhood. Before and after every kit preparation session, the flowhood surfaces and pipettes were wiped using sterile wipes, DNA AWAY™, and 95% ethanol. At the end of each session the surfaces were also UV-sterilized (60 minutes). Each kit contains plastic tongue scrapers (Yellow CeraSpoon Safe Ear Curettes, Bionix) that were first autoclaved and then UV-sterilized for 60 minutes, 1.5 mL gamma-sterilized and pre-packaged collection tubes certified as pyrogen- RNase- DNA- and ATP-free (VWR), each containing 200 uL sterile 1X PBS buffer (VWR), along with pre-packaged sterile gloves (VWR). Each collection tube and tongue scraper pair was placed inside a sterile bag and the bags were placed in another bag. The next steps were performed outside of the flowhood. Each collection bag was put inside a Styrofoam box along with ice gel packs. Ice gel packs and Styrofoam boxes were not reused to prevent cross contamination between individuals in case of a spill, which would already be highly unlikely due to multiple layers of packaging. Upon arrival of samples, collection tubes were taken out of their original bags, wiped with 95% ethanol and DNA AWAY™ using sterile wipes and placed into a new sterile bag. Gloves were frequently exchanged both during this step and before proceeding to the next collection tube to

prevent cross contamination. In addition to standard lab attire such as gloves and lab coat, a facemask was worn to prevent contamination during kit preparation and sample storage.

Subject recruitment and sample collection. We made an educational video to introduce a diverse audience to the fascinating world of phages, explain our study and to recruit volunteers. We also created an instructional video for prospective volunteers on subject disqualifying criteria and subject rights, and to provide a step-by-step demonstration of sample collection, storage, and shipment. Among other exclusion criteria, subjects could not have taken antibiotics for the preceding 3 months and subjects could not have active cavities or gum disease. Qualified subjects were sent a kit and were asked not to brush their teeth or tongue for a minimum of 8 hours prior to sample collection to allow for a substantial build up of plaque on the tongue dorsum. Put simply, subjects were instructed to 1) wear gloves, 2) scrape their tongue (dorsal surface) several times using the tongue scraper, 3) deposit their sample into the collection tube, 4) place the tube back into the bag, and 5) store the bag in their freezer along with ice gel packs prior to an over-night shipment of their samples. They were also instructed to report any sources of error that occurred at any step, and to send their samples along with their signed consent form and questionnaire. Our sample collection and processing protocols were approved by Caltech Institutional Review Board (IRB protocol 14-0430) and Institutional Biosafety Committee (IBC protocol 13-198).

Subject recruitment and sample collection (Bik *et al.* samples). 10 subjects included in this study are those included in a previous study of oral microbial diversity by Bik *et al.* (40). Briefly, samples were collected from 10 individuals by a dentist who examined subjects for their oral health, thereby excluding subjects with active cavities, gingivitis, or periodontal disease. For each of the 10 individuals, samples from 6 different oral sites were collected using sterile curettes and deposited separately in 1.5 mL collection tubes containing

PBS buffer. The 6 oral sites sampled include plaque from tongue dorsum, tongue ventral, buccal mucosa, hard palate, supra-gingiva, and sub-gingiva.

Table 1. List of non-barcoded phage primer sequences used to amplify markers HB1, PCA2, HA, and PCA1.

Markers	Forward Primer Sequence (5' to 3')	Reverse Primer Sequence (5' to 3')
HB1	CCGATCTGTCICARGGIGAYGA	GTTACGAACTCTTTGGCRTTRTAIGGRTC
PCA2	GTGCGGCAACWAARCARGAICA	CTGATTATTTGGGTGTGCRTGWARYTCRTC
HA	CGTGATGGCTGYCTWGARTTYGAYGA	CGTAAGGAGTGCTYTCRTCCARCATIGG
PCA1	CCTTTGYTTGGCITGGTTYGARGA	CAGCRACICCCCAYTCRCC

Table 2. A list of 4-letter error-detecting barcodes designed for multiplexed sequencing of PCA2 marker sequences from various samples. The barcode number dictates the first three letters in the barcode sequence according to base 4 arithmetic. The last letter in the barcode sequence is a parity letter and allows for detection of errors within the barcode sequence (See Error-detecting Barcode Design).

Barcode Number	Barcode Sequence (5' to 3')
63	TTTC
62	TTGA
61	TTCT
60	TTAG
59	TGTA
58	TGGT
57	TGCG
56	TGAC
55	TCTT
54	TCGG
53	TCCC
52	TCAA
51	TATG
50	TAGC

49	TACA
47	GTTA
45	GTCG
44	GTAC
43	GGTT
40	GGAA
39	GCTG
37	GCCA
35	GATC
31	CTTT
30	CTGG
29	CTCC
28	CTAA
27	CGTG
25	CGCA
23	CCTC
22	CCGA

Quality control steps to eliminate sequencing errors. We used Illumina MiSeq's 2x300bp paired-end configuration (GENEWIZ, Inc). Each sequencing run produced about 20-25 Million paired-end reads. Paired-end reads were joined using *join_paired_ends.py* script from QIIME (Quantitative Insights Into Microbial Ecology) package, and unless noted otherwise scripts used in this chapter are part of QIIME (66). When a base is confirmed by both reads, higher Phred score is increased by up to 3 points. If paired reads had any mismatches across their overlapping bases, the paired reads was eliminated from any further analysis (QC step #1). Taking one of the sequencing runs as an example, this step reduced the number of paired reads from ~21Million to ~6.3Million joined reads (Figure 1), but it also likely reduced the probability of sequencing error in the surviving reads. For markers

HB1, PCA2, and HA the overlap between the paired reads entirely covers the marker sequence, hence eliminating many sequencing errors.

Upon joining reads and eliminating those with mismatches in the region of overlap *seqQualityFilters.py*, an in-house script, was used to preform QC step #2: taking joined reads from QC step #1, and eliminating any sequences that have one or more bases marked by a Phred score below 30. At this step the number of sequences was reduced from ~6.3Million to ~6.0 Million, reflecting the fact that QC step #1 already ensures that the majority of surviving reads have high Phred scores across their bases (Figure 1). Excluded from QC step #2 were the first two bases in the beginning and end of each sequence, which for majority of reads have much lower quality scores.

Using *seqQualityFilters.py,* sequences were placed in 4 different bins according to their primer sequences, and any sequence that did not have the correct barcode length, or the correct primer sequences at the expected positions, was eliminated (QC step #3). Additionally, nearly half of remaining sequences had to be reverse complemented so that all sequences were oriented in the 5' to 3' direction. Using the same script, primer and barcode sequences were removed, and barcode sequences were written to a separate file (to be used as input to *split_libraries_fastq.py*). At this point sequences that did not have the correct length were filtered out (QC step #3). Sequences were demultiplexd using *split_libraries_fastq.py* and reads with errors in the barcode sequence were eliminated (QC step #4).

Phage community composition plots ("Phageprints"). After demultiplexing quality-controlled reads, sequences were clustered according to a specified sequence similarity threshold using UCLUST *de novo* clustering algorithm (67) used in *pick_otus.py* script. Using *make_otu_table.py*, OTU tables were generated. An OTU table summarizes counts of sequences assigned to each OTU across each sample. We refer to this per-sample

sequence count as the OTU size. As long as an OTU of size 1 or greater exists in at least one sample, it is included in the OTU table. In this way, the counts of OTUs for samples containing the same marker remains the same, though their size could vary widely across different samples. Later we will demonstrate the effects of noise filters applied to the OTU table. The relative abundance of each OTU within each sample was calculated via *processOtuTable.py,* another in-house script. In plotting the relative OTU abundance values for different samples, we arrived at complex, individual-specific patterns. We dubbed these phage community composition plots as "phageprints".

Examining the effect of OTU sequence similarity threshold on Pearson correlation coefficient matrices. In analyzing 16S sequences, clusters or Operational Taxonomic Units (OTUs) are conventionally defined at 97% sequence similarity threshold. To examine the effect of sequence similarity threshold for phage OTU formation, we tested OTU sequence similarity thresholds of 98%, 97%, 95%, 90%, and 80%. Figure 15 is a matrix of Pearson correlation coefficients calculated during the pairwise comparison of HB1 community compositions using different sequence similarity thresholds for defining OTUs. Very similar Pearson correlation matrices are obtained as the sequence similarity threshold is lowered from 98% to 80%. However, because the number of cluster is reduced as we reduce the sequence similarity threshold, with lower sequence similarity thresholds, the chance that individual-specific variations are lumped into the same cluster is increased. If noise-induced sequence variations are effectively accounted for, higher sequence similarity thresholds for defining OTUs can enable a more accurate and detailed depiction of a person's phage community composition. For this reason, we used a sequence similarity threshold of 98% for the study of different oral sites, and later we used a 100% sequence similarity threshold for the temporal and the global study.

A

	3.3	3.5	6.3	6.5	6.6	10.1	10.3	12.1	12.2	12.3	12.6	16.1	16.3	17.2	17.3	17.5	17.6
3.3	1.00	0.95	0.12	0.12	0.12	0.02	0.04	0.12	0.11	0.11	0.12	0.05	0.02	0.00	0.02	0.00	0.00
3.5	0.95	1.00	0.08	0.08	0.08	0.01	0.03	0.10	0.10	0.10	0.11	0.04	0.01	0.07	0.08	0.07	0.07
6.3	0.12	0.08	1.00	1.00	1.00	0.02	0.05	0.12	0.12	0.14	0.16	0.18	0.06	0.00	0.04	0.01	0.00
6.5	0.12	0.08	1.00	1.00	1.00	0.02	0.06	0.13	0.14	0.15	0.17	0.18	0.06	0.00	0.04	0.01	0.00
6.6	0.12	0.08	1.00	1.00	1.00	0.02	0.05	0.12	0.12	0.14	0.16	0.19	0.06	0.00	0.04	0.01	0.00
10.1	0.02	0.01	0.02	0.02	0.02	1.00	0.47	0.02	0.02	0.02	0.02	0.00	-0.01	-0.01	0.00	0.00	-0.01
10.3	0.04	0.03	0.05	0.06	0.05	0.47	1.00	0.04	0.04	0.04	0.04	0.01	0.00	0.00	0.00	0.00	0.00
12.1	0.12	0.10	0.12	0.13	0.12	0.02	0.04	1.00	1.00	1.00	0.95	0.06	0.04	0.00	0.03	0.01	0.00
12.2	0.11	0.10	0.12	0.14	0.12	0.02	0.04	1.00	1.00	1.00	0.95	0.06	0.04	0.00	0.03	0.01	0.00
12.3	0.11	0.10	0.14	0.15	0.14	0.02	0.04	1.00	1.00	1.00	0.95	0.06	0.04	0.01	0.04	0.02	0.01
12.6	0.12	0.11	0.16	0.17	0.16	0.02	0.04	0.95	0.95	0.95	1.00	0.06	0.03	0.10	0.12	0.11	0.10
16.1	0.05	0.04	0.18	0.18	0.19	0.00	0.01	0.06	0.06	0.06	0.06	1.00	0.26	0.00	0.01	0.00	0.00
16.3	0.02	0.01	0.06	0.06	0.06	-0.01	0.00	0.04	0.04	0.04	0.03	0.26	1.00	0.00	0.00	0.00	0.00
17.2	0.00	0.07	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.01	0.10	0.00	0.00	1.00	0.92	1.00	1.00
17.3	0.02	0.08	0.04	0.04	0.04	0.00	0.00	0.03	0.03	0.04	0.12	0.01	0.00	0.92	1.00	0.95	0.92
17.5	0.00	0.07	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.02	0.11	0.00	0.00	1.00	0.95	1.00	1.00
17.6	0.00	0.07	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.01	0.10	0.00	0.00	1.00	0.92	1.00	1.00

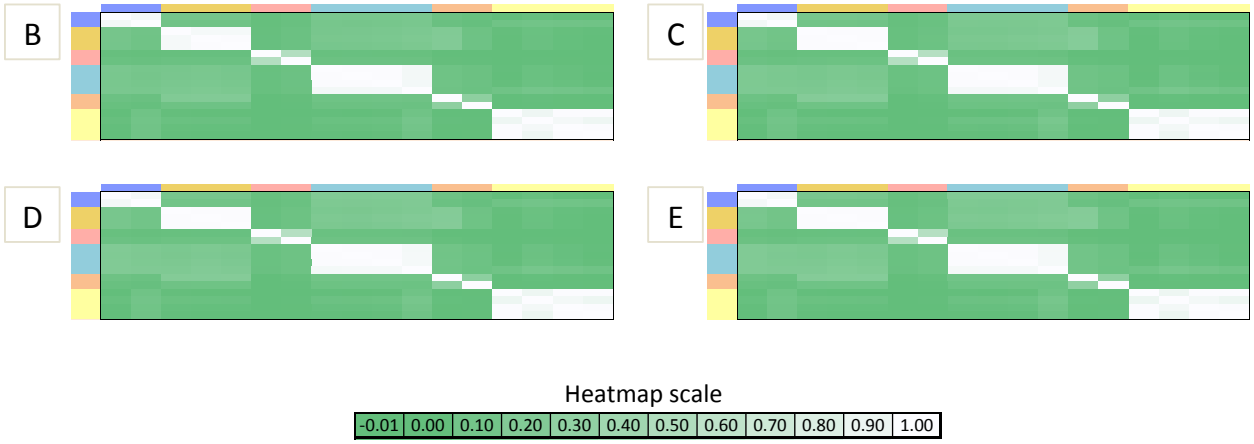


Figure 15. Pairwise Pearson correlation coefficient values calculated for HB1 phage community compositions as a function of A) 98%, B) 97%, C) 95%, D) 90%, and E) 80% sequence similarity thresholds for OTU formation. Sample IDs can be decoded as before: subject ID precedes oral site ID. Oral sites 1-6 correspond to tongue dorsum, hard palate, buccal mucosa, ventral tongue, supra-gingiva, and sub-gingiva respectively (e.g. 3.3 corresponds to subject 3 community composition derived from the buccal mucosa, and 3.5 is subject 3 supra-gingiva community composition). The number of OTUs generated at 98%, 97%, 95%, 90%, and 80% sequence similarity thresholds are 210, 181, 172, 170, and 80, respectively.

Rarefaction. One common observation when analyzing sequence diversity is that it is dependent on the number of sequences analyzed. Rarefaction is typically performed to

illustrate the minimum number of sequences needed to entirely capture sequence diversity in a given environment. Because both PCR and sequencing are error-prone processes, they contribute noise-induced sequence variation and result in an over-estimation of natural sequence diversity. To resolve this issue, we present a comparison of rarefaction plots in Figure 16 and Figure 17.

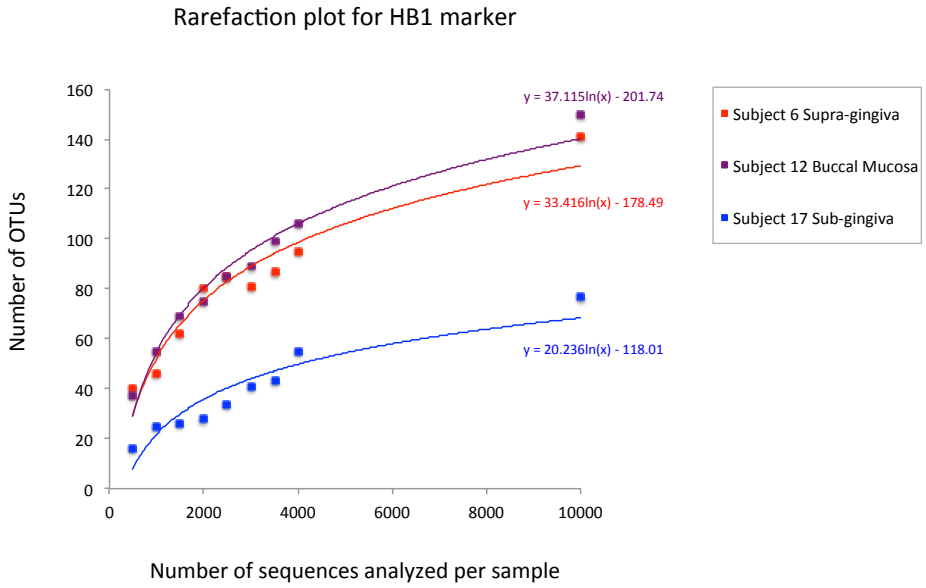


Figure 16. Rarefaction plot for HB1 marker using data from 3 samples, belonging to subject 6 supra-gingiva (red), subject 12 buccal mucosa (purple), and subject 17 sub-gingiva (blue). The y-axis contains the number of OTUs (defined at 98% sequence similarity) that contain one or more sequences. The x-axis demonstrates the number of sequences analyzed per sample. Trend lines demonstrate a logarithmic relationship between number of OTUs and the number of sequences analyzed per sample.

Figure 16 is a rarefaction plot of HB1 marker present in three different environments: subject 6 supra-gingiva (red), subject 12 buccal mucosa (purple), and subject 17 sub-gingiva (blue). By randomly selecting *x* number of sequences from each environment and plotting the number of observed OTUs, *y,* at each value of *x,* the relationship can be

modeled as $y=ln(x)$. From this plot, one might conclude that even 10,000 sequences are insufficient to entirely capture sequence diversity within each environments. This result would place significant limits on the number of samples that reads from a sequencing run can be divided into. It also provides no insight for identifying OTUs that are either noise-induced or so naturally low in abundance that practically they would not be reproducible using current sequencing technologies. However, if we stratify OTUs based on their relative abundance, we arrive at a more meaningful picture.

Figure 17 shows HB1 rarefaction plots of the same subjects as in Figure 16: subject 6 supra-gingiva (panel A), subject 12 buccal mucosa (panel B), and subject 17 sub-gingiva (panel C). As in Figure 16, OTUs are defined at 98% sequence similarity threshold. In Figure 17.A OTUs have been binned into 5 categories: Empty OTUs, which are present in other samples or are non-empty at higher values of x, but have zero sequences associated with them in this sample; below-threshold OTUs that are present in less than 0.1% of population x; rare OTUs which are present in 0.1-1%; common OTUs with 1%-10% presence; and finally, abundant OTUs which take up anywhere from 10% to 100% of reads.

As we increase x from 500 to 10,000, the number of abundant OTUs remains constant. The number of common OTUs fluctuates between 6 and 7, and the number of rare OTUs decreases from 32 to 25. At x =1500, OTUs below 0.1% of population emerge and they are OTUs defined by a single sequence. As x increases previously empty OTUs transform into below-threshold OTUs defined by 1 to 10 sequences. However, as the number of below-threshold OTUs continues to increase the rare, common, and abundant OTU counts begin to stabilize and have nearly exact values at x=4000 as they do at x=10,000.

In Figure 17.B and C, we have combined counts of rare, common, and abundant OTUs and together call them above-threshold OTUs. Below-threshold OTU counts continue to increase as a function of x while above-threshold OTU counts have the same value at x=4000 as they have at x=10,000. We have limited our analysis to a maximum of 10,000 sequences because only a few samples are represented by more than 10,000 sequences. By rarifying the OTU table multiple times and creating an average OTU table at different values of x, the count of above-threshold OTUs could stabilize at even a smaller value of x. While it's possible that some below-threshold OTUs are highly rare species in the population, they are more likely OTUs that arose as a result of sequencing and/or PCR errors. From this exercise we have taken away two useful parameters: the minimum number of sequences to analyze per sample, and the relative abundance threshold for detection, which we will further demonstrate in the following subsections.

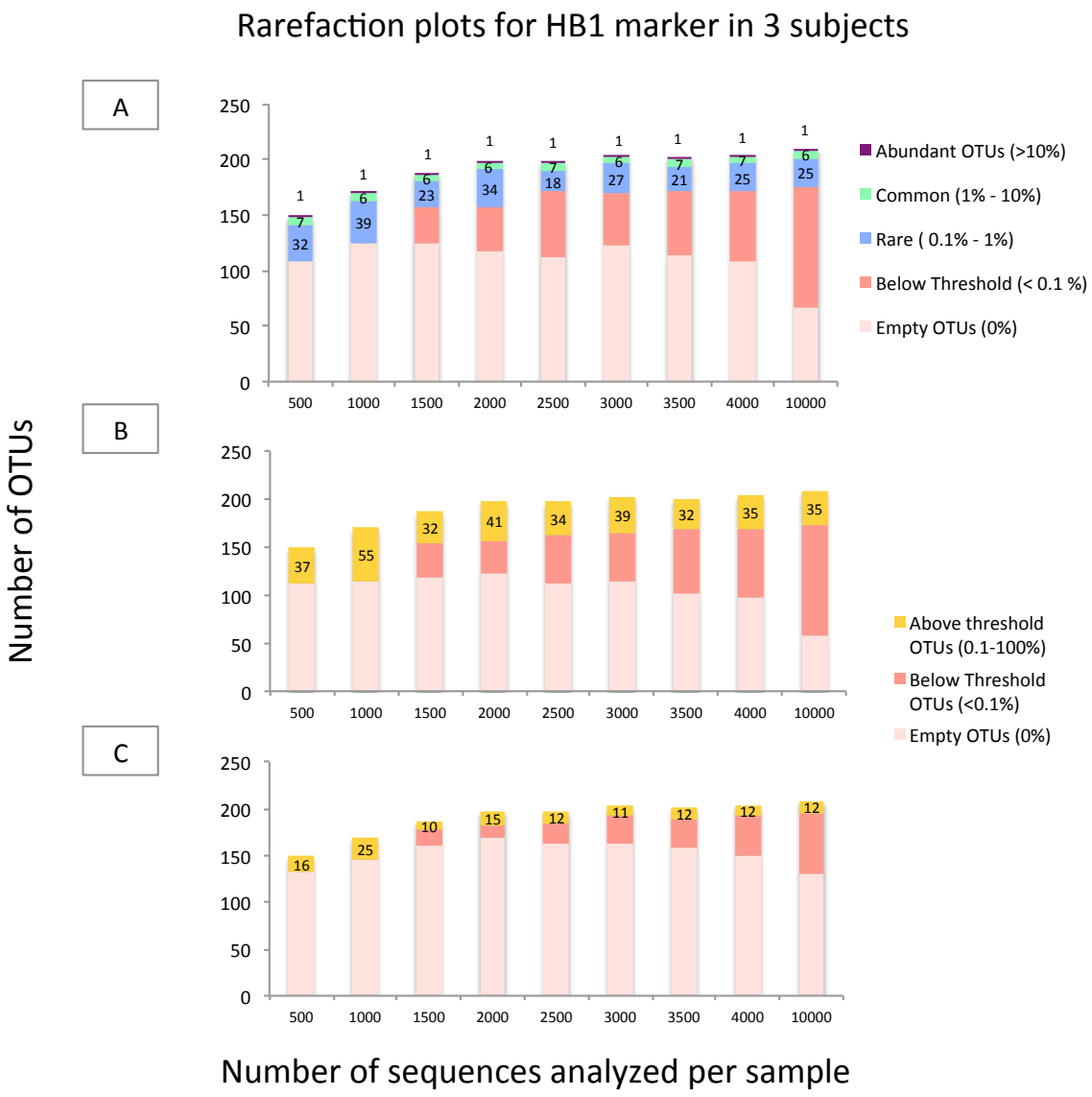


Figure 17. Rarefaction plots of HB1 marker segmented based on OTU relative abundance. Samples are analyzed at rarefaction intervals ranging from 500 to 10,000 sequences. A) Subject 6 supra-gingiva rarefaction plot. OTUs that are present in other samples but have zero sequences associated with them in this sample are shown in light pink, below threshold OTUs with a relative abundance of less than 0.1% but greater than zero are in darker pink, Rare OTUs (0.1%-1%) are in blue, followed by Common OTUs (1%-10%) in green, and Abundant OTUs (>10%) in purple. The count of OTUs for rare, common, and abundant OTUs are depicted. Rarefaction plots for B) subject 12 buccal mucosa and C) subject 17 sub-gingiva. The

count of Rare, Common and Abundant OTUs are summed and shown in yellow as above-threshold OTUs (0.1%-100% relative abundance). The counts of above-threshold OTUs are shown.

Quantifying experimental noise. How reproducible is a phage community composition plot? Figure 18 summarizes the sources of noise from all experimental processes performed during this study. First, it's important to capture sampling variation. How consistently can we capture a phage community from an individual's oral site given that we are sampling different clumps of biofilm each time? Another factor that could contribute to sampling variation are the personal differences in the rate of biofilm mass accumulation on the tongue dorsum. Secondly, we need to ask whether processes of lysis and DNA extraction allow for the availability of the same template DNA sequences in the same relative abundances across different extraction runs.

Third, we need to evaluate the OTU abundance variations that could result in PCR due to both errors as well as other stochastic events. For example, it's possible that very rare template sequences are left out of the initial cycles of PCR and their relative abundance at the end of PCR is lower than their relative abundance prior to PCR. In this hypothetical scenario PCR could serve as a biased amplifier. PCR purification is similar to extraction and sampling in that it does not introduce sequence errors; however it is unlike these processes because after PCR billions of template copies are created and it's unlikely that the loss of a fraction of templates during PCR purification will dramatically change OTU relative abundances. Finally, Illumina MiSeq sequencing is another error-prone process not only at the level of base-calling, but at the level of bridge amplification which like PCR could introduce errors that propagate exponentially. Refer to Figure 18 for a summary of processes that could result in irreproducible OTUs or variation in OTU relative abundances.

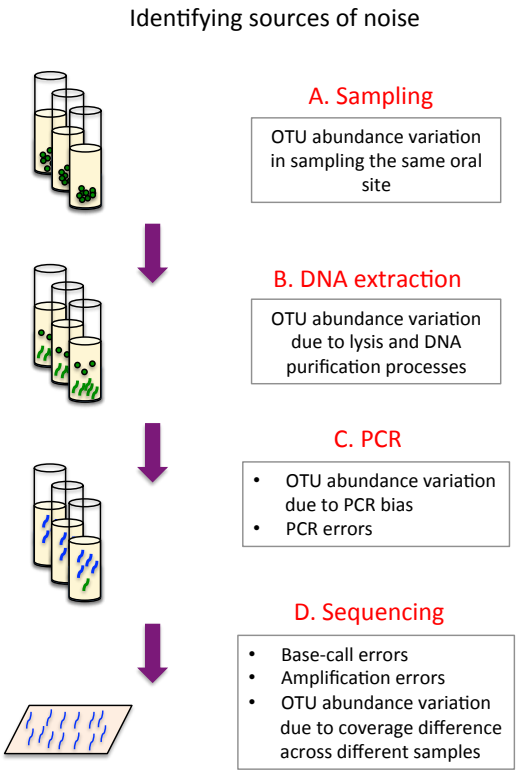


Figure 18. Sources of error and variation in experimental processes used in this study. A) Sampling of the same oral site in the same individual could result in collection of different microbial communities, which could introduce new OTUs or change relative abundance of existing OTUs. B) DNA extraction is not 100% efficient and the fraction of DNA extracted from an environment could serve as a source of variation across different samples. C) PCR introduces errors that could present themselves as novel OTUs or cause variation in abundance of genuine OTUs. D) Sequencing also introduces errors both at the level of base-calling and bridge amplification.

To quantify how reproducible a given phage community composition is, we obtained 3 different samples from subject 37 tongue dorsum. We then performed DNA extraction and PCR separately on each sample and sent samples for sequencing (sequencing run #2). The logic behind this experiment was to capture a lumped measure of noise arising from various processes depicted in Figure 18. After performing quality control steps 1-4,

demultiplexing reads based on their barcode sequences, clustering reads based on 98% sequence similarity threshold for OTU formation, rarefying the OTU table to 4000 reads per sample, and calculating the relative abundances of OTUs, we measured the standard deviation in the relative abundance of each OTU across these three samples (Figure 19). Remarkably, relative abundance values across these three samples were highly consistent, with the majority of OTUs having standard deviations below 0.2% and the maximum standard deviation observed was less than 0.7% relative abundance.

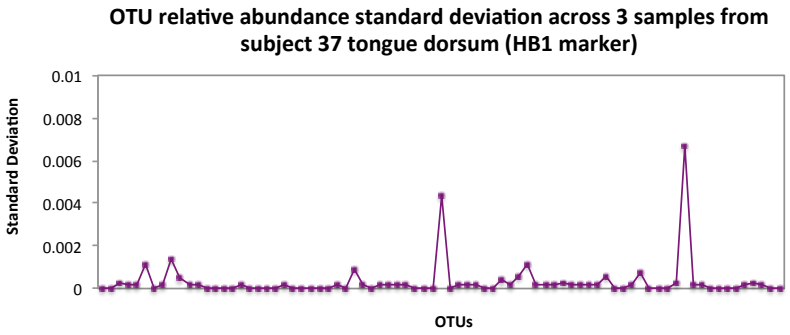


Figure 19. Standard deviations of OTU relative abundances calculated for all experimental processes. Three data points per OTU are used for standard deviation calculations. These three data points correspond to measurements of OTU relative abundances obtained for three different samples obtained from subject 37 tongue dorsum (HB1 marker) which underwent separate sampling, DNA extraction, PCR and PCR cleanup procedures. The maximum standard deviation observed is less than 0.007 relative abundance, and majority are close to 0.

Identifying non-reproducible OTUs. To identify OTUs that were non-reproducible across the three samples from subject 37's tongue dorsum (HB1 marker), we flagged OTUs that had appeared in only one or two samples out of three. We then plotted the histogram of non-reproducible OTUs as a function of their relative abundance (for those

OTUs appearing in 2 out 3 samples, the higher relative abundance value was used). The thresholds defining each bin, *b*, were selected to be the following: $0>b_1\geq0.00025$ (OTU of size 1 sequence since the total number of sequences per sample is 4000), $0.00025>b_2\geq0.0005$ (2 sequences), $0.0005>b_3\geq0.00075$ (3 sequences), $0.00075>b_4\geq0.001$ (4 sequences), and $0.001<b_5$ (5 or more sequences).

Figure 20 demonstrates the number of non-reproducible OTUs drops as a function of OTU relative abundance, and all OTUs with more than 4 sequences (0.001 relative abundance) are reproducible. This result is consistent with previously established OTU noise threshold of 0.001 relative abundance, obtained during the rarefaction study (see Rarefaction). To conclude, using two different approaches and across two different sequencing runs, we arrived at 0.001 relative abundance as the detection threshold for OTUs.

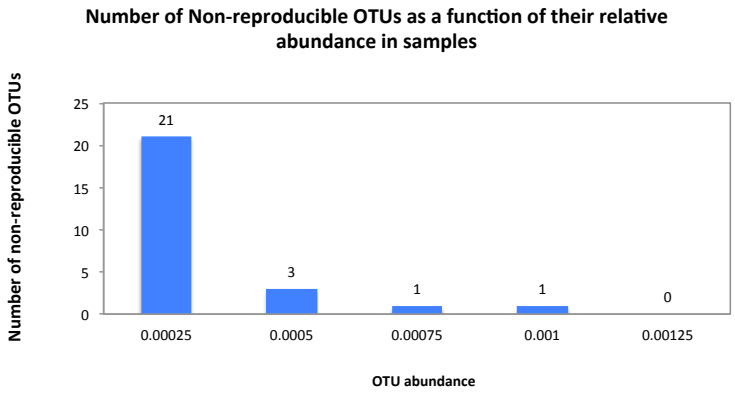


Figure 20. Number of non-reproducible OTUs across three samples obtained from subject 37 tongue dorsum (HB1 marker), presented as a function of OTU relative abundance. A total of 30 OTUs appear in one or two samples out of three, and therefore are considered non-reproducible. 21 out of 30 OTUs are defined by a single sequence which translates into 0.00025 relative abundance since samples are rarefied to 4000 sequences. The

number of non-reproducible OTUs drops as a function of OTU relative abundance, and all OTUs with more than 4 sequences (0.001 relative abundance) are reproducible across three samples. This relative abundance threshold for noise-induced OTUs was previously obtained using a different approach, different set of samples, and a different sequencing run (Figure 17).

In addition to capturing a lumped sum of noise across all experimental processes for subject 37 tongue dorsum sample (Figure 19,Figure 20), for samples from subjects 3, 6, 10, 16, and 17, we performed a second set of PCR on previously extracted DNA samples, and submitted those samples for sequencing (Figure 21). In addition to these replicates, we acquired new samples from the tongue dorsum for subjects 31, 35, 37, and 38, and submitted these samples for the second sequencing run. In obtaining replicate phageprints, we were able to demonstrate that with proper quality filtration steps phageprints are highly reproducible even when they are generated from two separate PCR and sequencing steps (Figure 21).

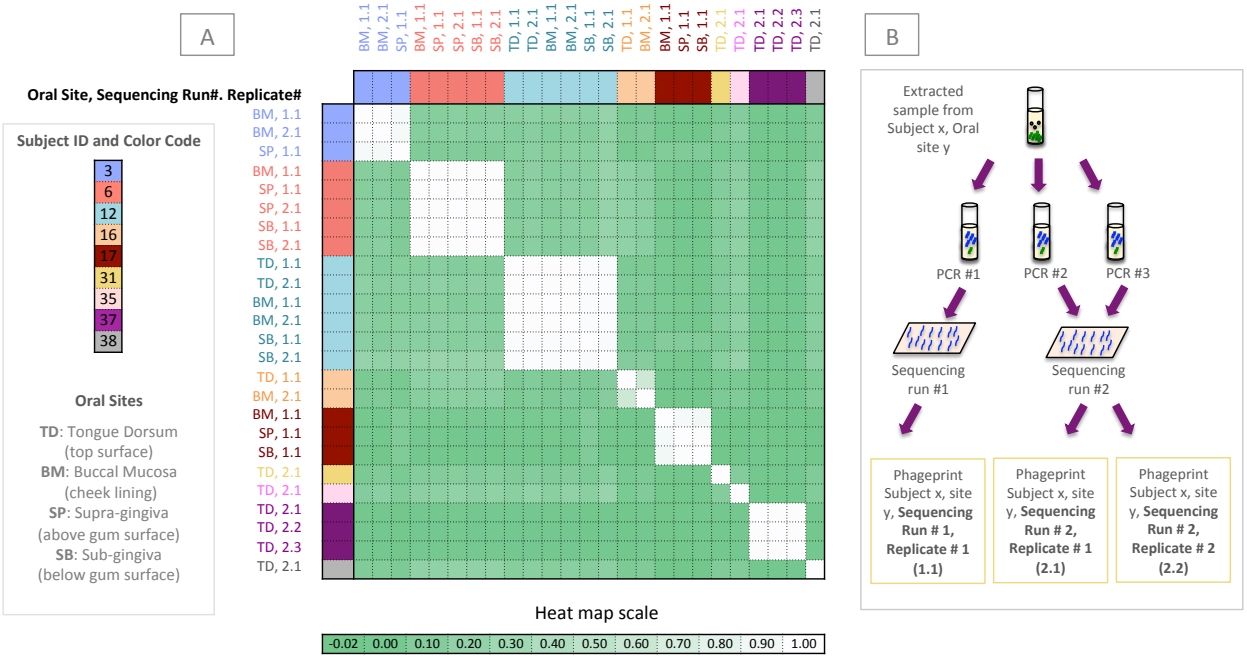


Figure 21. Panel A is the Pearson correlation matrix of all HB1 phageprints. Each phageprint is derived from the analysis of 4000 sequences associated with an individual and a particular oral site. OTUs are defined at 98% sequence similarity and OTUs with less than or equal to 0.1% relative abundance across all phageprints were filtered out. Phageprints are color-coded based on the individual they originate from. Oral sites shown to be positive for the HB1 marker are the tongue dorsum (TD), buccal mucosa (BM), supra-gingiva (SP), and sub-gingiva (SB). Phageprints that were acquired from sequencing run #1, are those marked as replicate #1. Panel B shows that to confirm reproducibility of phageprints, a second set of PCR was performed on previously extracted DNA from all samples included in sequencing run #1 and those PCR products were included in sequencing run #2. Phageprints derived from the second sequencing run are marked as replicate #2.

Identifying phage marker homologs and phylogenetic tree construction. The most abundant sequence from each OTU was retrieved using *pick_rep_set.py* to serve as a

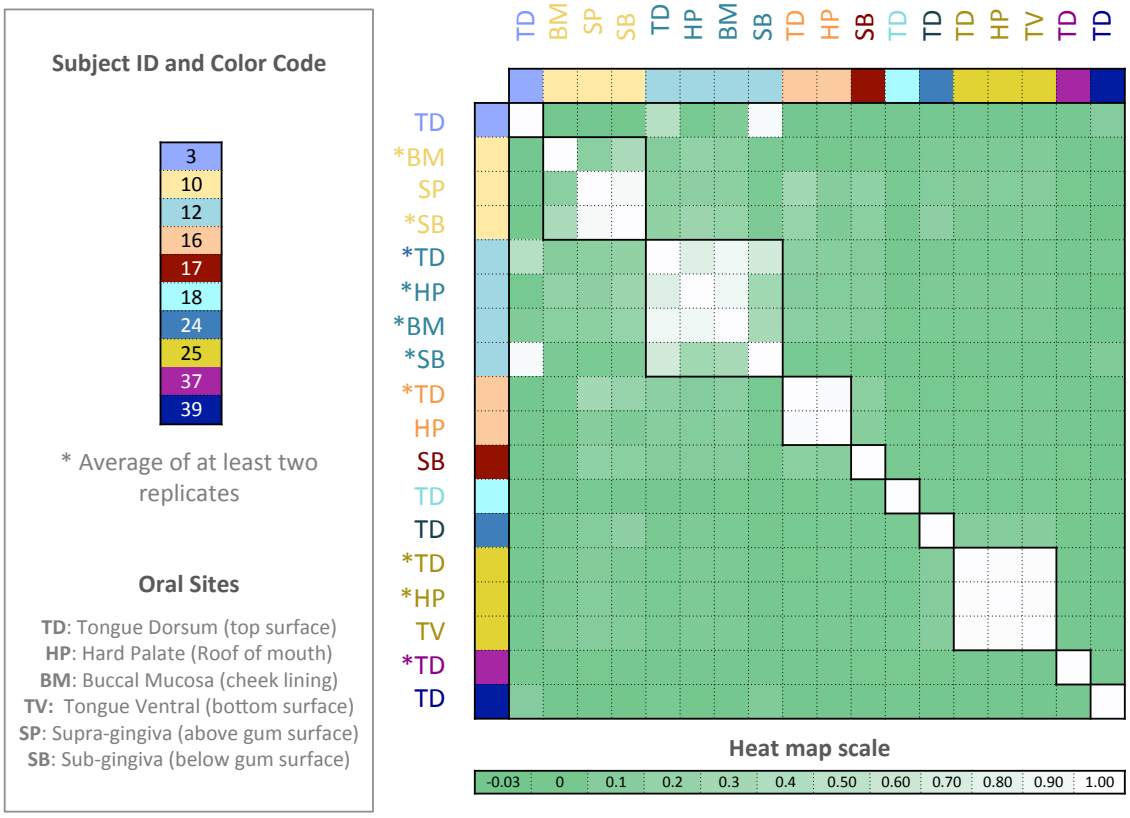
representative sequence. BLASTx function was used to detect the closest homolog to each OTU's representative sequence from within the NCBI's non-redundant protein database. The output of BLASTx for each marker is summarized in SI Table 1 and SI Table 3. These tables summarize the amino acid percent identity between each OTU representative and its closest homolog, as well as the E-value associated with the match.

HB1 representative sequences were aligned using Geneious (68), using a gap open penalty of 30 and gap extension penalty of 15 and a 65% similarity cost matrix. No gaps were introduced. The alignment is shown in SI Figure 7. SeaView was used to create a PhyML maximum likelihood phylogenetic tree from the alignment. General Time Reversible model was used with empirical nucleotide equilibrium frequencies.

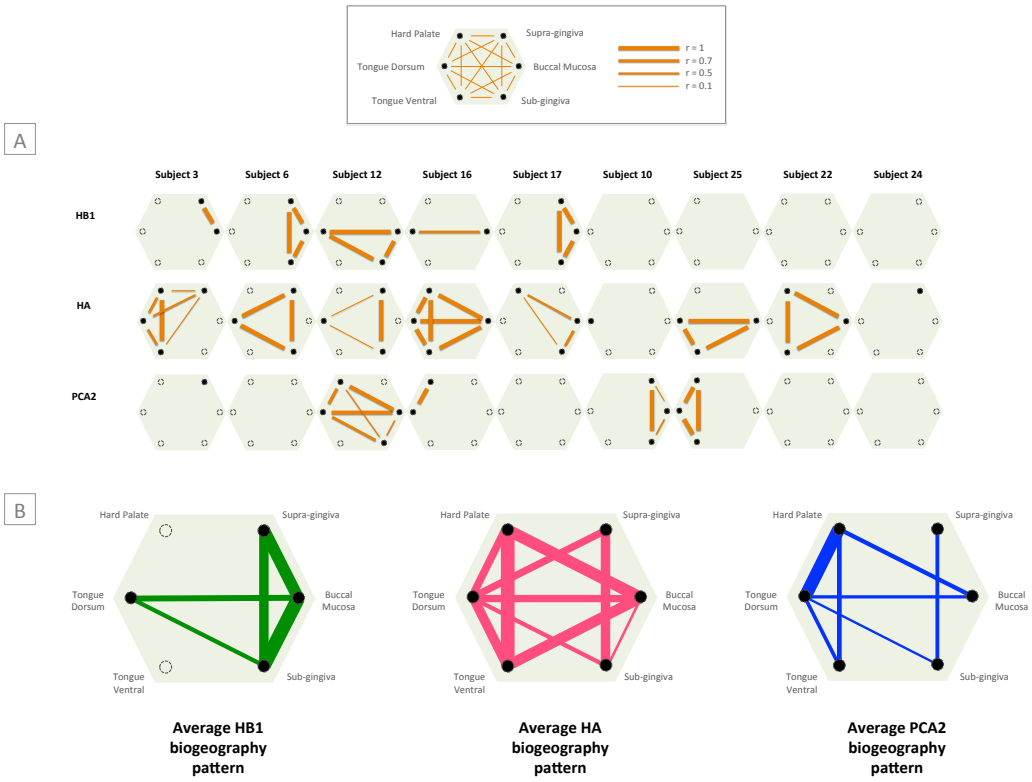
Phage networks. OTU tables were input to *createNetwork.py,* an in-house script that creates node and edge tables. The nodes represent samples and phage OTUs, and a directed edge connects samples to the OTUs that they host. The weight of this connection is based on the relative abundance of the OTU in that sample. Gephi software (69) was used to visualize the resulting networks, and to obtain the degree distribution.

Averaged phage biogeography patterns. Given 6 oral sites, there are up to 15 pairwise interactions possible. For each possible interaction, *i*, the number of subjects positive for the interaction was divided by the total number of subjects, while still accounting for the missing oral samples. The shorthand for this fraction is E_i. Moreover, the Pearson correlation coefficients associated with each interaction were averaged across the subjects positive for the interaction (r_i). The edges in SI Figure 2.B are weighted based on the product of E_i and r_i; however, edges with low r_i (<0.35) are not shown.

5.5 SI

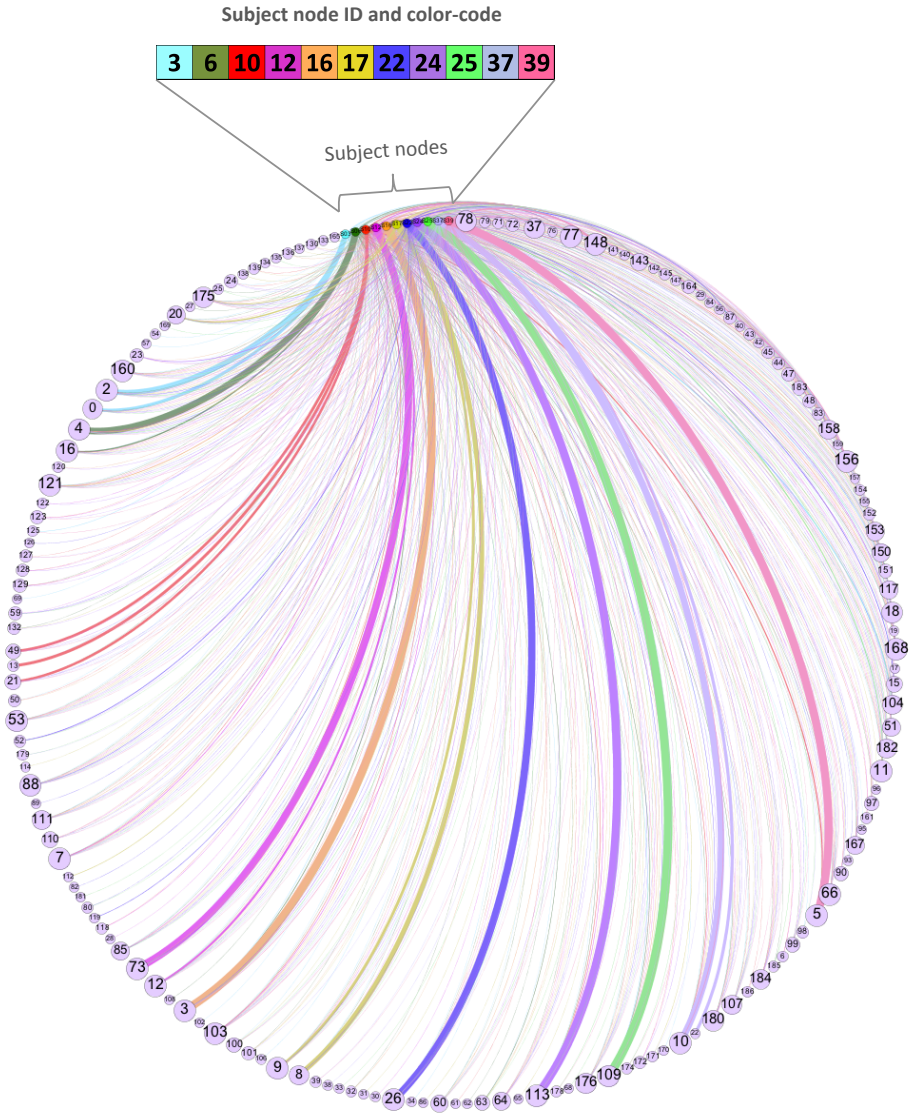


SI Figure 1. Pearson correlation matrix of PCA2 phage family. Samples that have an asterisk are those that are used in error analysis (see Methods) and have been replicated experimentally at least twice.



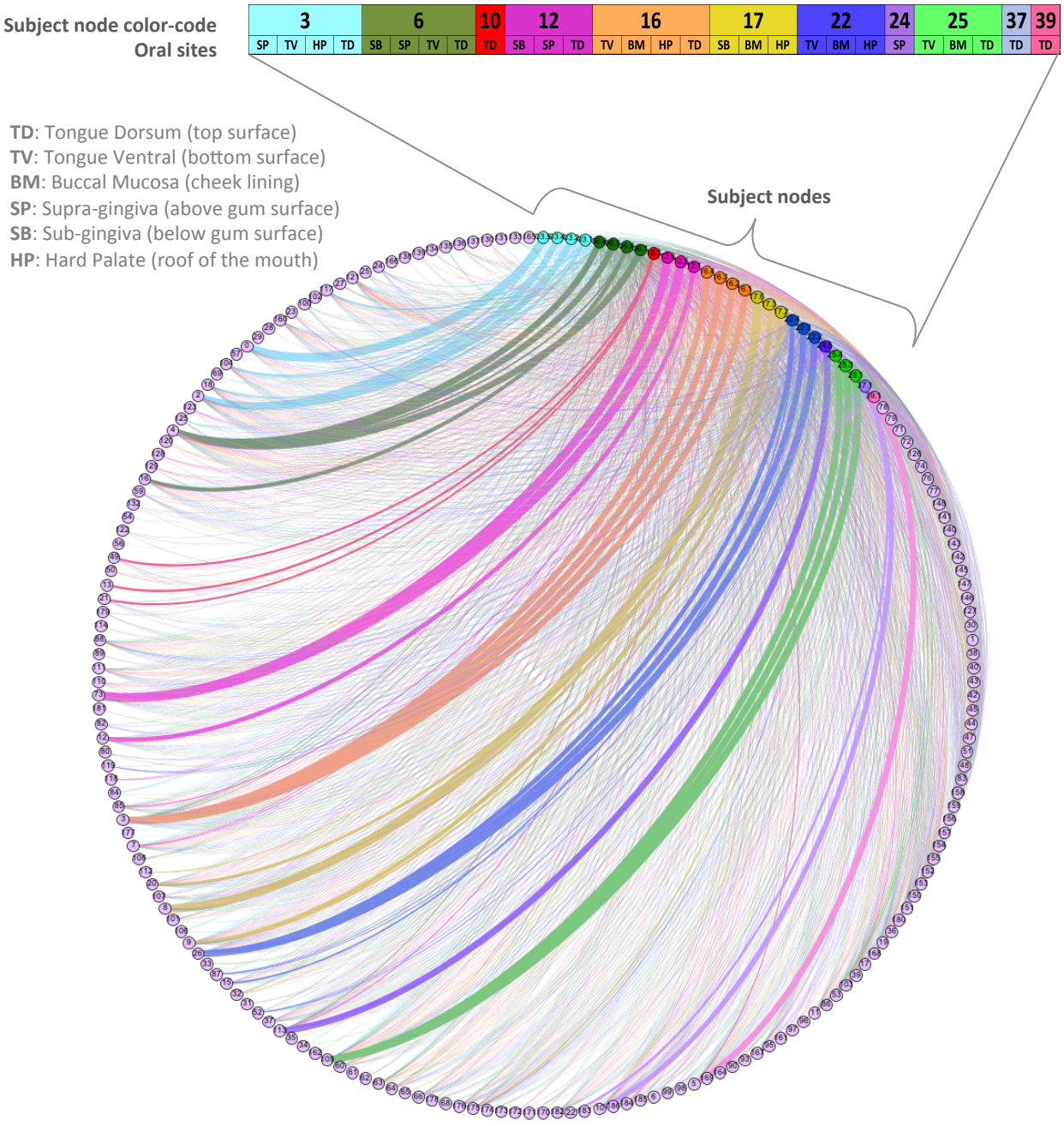
SI Figure 2. Phage biogeography patterns in the human mouth. Six different oral sites were tested for presence of phage families HA, HB1, and PCA2. A) Each hexagon summarizes the community composition of a particular phage family across different oral sites within a given individual. Each vertex of a hexagon symbolizes an oral site, and the presence of a marker is signified by a filled-in vertex. The marker is considered present at an oral site if at least 4000 high-quality sequences (see Methods) belonging to a terminase family are found at that oral site. The unfilled vertices are those samples that did not harbor phage markers. The missing vertices, correspond to samples that were unavailable and consequently were not tested for presence/absence of phage makers. The edges connecting different vertices represent the similarity of the phage community at one oral site to another. The thicker the edge, the more similar are the phageprints of the two connected oral sites. Here, the measure of similarity is the Pearson correlation value obtained by the pair-wise comparison of phageprints from different oral sites. B) The average biogeography patter for each phage family is denoted. The edge weight is a product of f and r_{avg}, where f represents the fraction of individuals that

harbor the connection and r_{avg} is the average of Pearson correlation value for that connection across all individuals who harbor the connection.



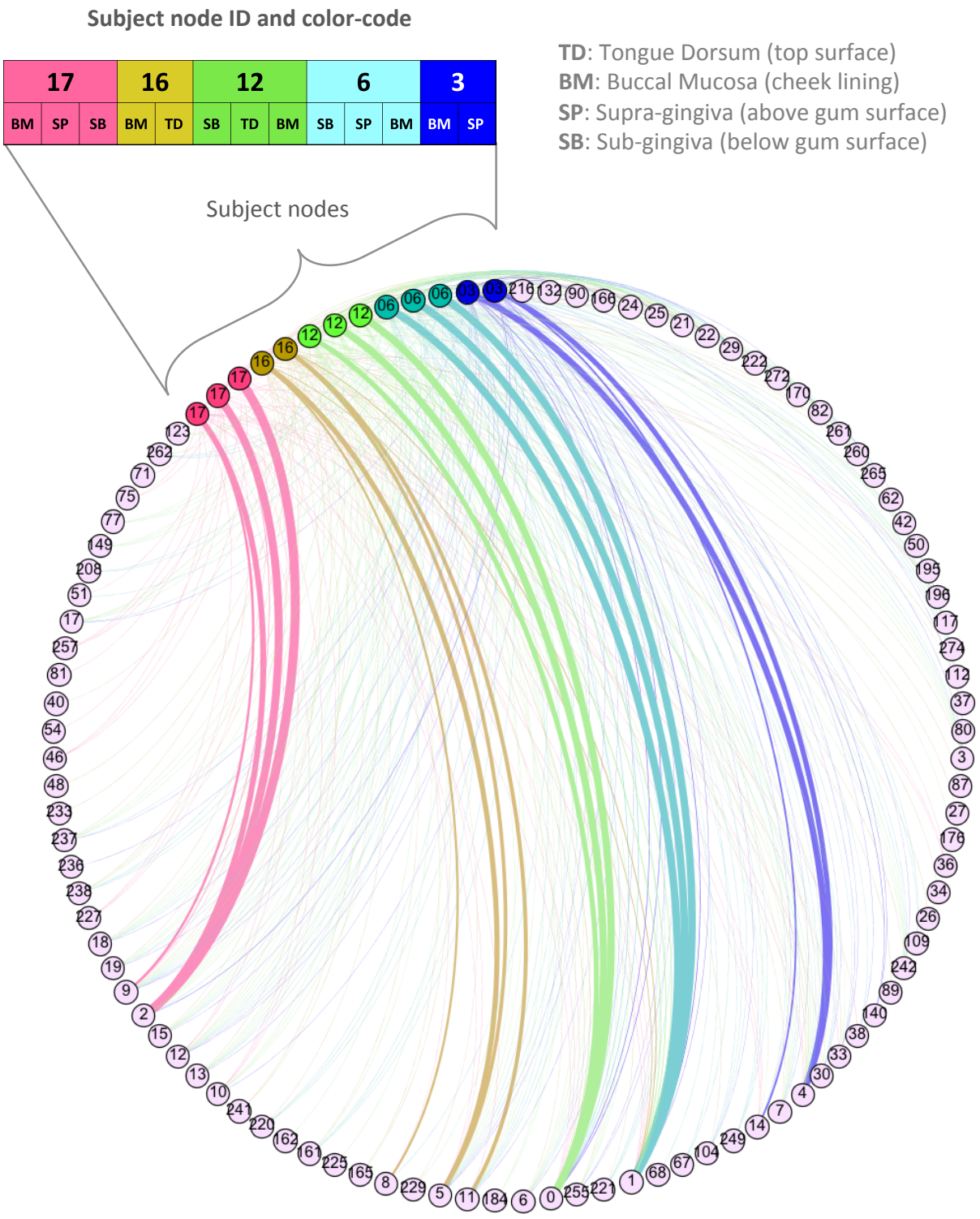
SI Figure 3. HA phage family network. Two types of nodes exist: OTU nodes (purple), and subject nodes. Subject nodes and edges are color-coded based on the individual they represent. Each directed edge connects an individual to a phage OTU that he/she harbors, and the edge weight is proportional to the relative abundance of the OTU in that individual's oral community. OTU node sizes and labels are proportional to the number of individuals the OTU is present in. For OTU nodes, the node ID is the OTU

ID, which can be matched to IDs in SI Table 3 for obtaining taxonomic information regarding each OTU's representative sequence. See SI Figure 4 to see a similar network that shows OTUs present at different oral sites.



SI Figure 4. HA phage-host network (expanded version of SI Figure 3, showing each oral site). Purple nodes are the OTU nodes and all other nodes represent samples. Sample nodes and edges are color-coded based on the

individual they originate from. Subject node color code, ID, and the oral sites are displayed above sample nodes. Each edge connects an OTU a sample it exists in, and the edge weight is proportional to the relative abundance of the OTU in that sample. Node IDs are displayed. For OTU nodes, the node ID is the OTU ID which can be matched to IDs in SI Table 3 for identifying taxonomic information regarding each OTU. For sample nodes, the nodes IDs are simply the subjects' IDs.



SI Figure 5. HB1 phage family network (expanded version of Figure 7). Purple nodes are the OTU nodes and all other nodes represent samples. Sample nodes and edges are color-coded based on the individual they

originate from. The oral site associated with each sample is abbreviated next to the sample's node. Each edge connects an OTU a sample it exists in, and the edge weight is proportional to the relative abundance of the OTU in that sample. Node IDs are displayed. For OTU nodes, the node ID is the OTU ID which can be matched to IDs in SI Table 2 for identifying taxonomic information regarding each OTU. For sample nodes, the nodes IDs are simply the subjects' IDs.

SI Table 1. Closest homolog to each OTU's representative sequence (HB1 phage family). Each OTU's representative sequence was used as a query for NCBI's BLASTx homology search against the non-redundant protein database. The table summarizes the E-value and the percent amino acid identity across the query sequence and the closest homolog, as well as the closest homolog's name, sequence ID, and taxon ID. The taxon ID is color coded, and the taxonomic classification corresponding to each taxon ID can be retrieved from the following table. Note with the exception of a few "putative uncharacterized" homolog names that most are identified as terminases or TerLs (terminase large subunits).

Query Sequence ID (OTU ID)	Percent Identity	E value	Closest Homolog	Closest Homolog Sequence ID	Closest Homolog Taxon ID
0	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
1	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
10	67.9	3.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
103	69.14	5.00E-30	terminase [[Clostridium] scindens]	gi|639772655|ref|WP_024738760.1|	29347
104	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
106	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
109	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
11	72.84	1.00E-34	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
112	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
117	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
118	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
12	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
122	70.37	2.00E-19	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
123	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
128	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
13	70.37	6.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
132	70.37	4.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
134	67.9	3.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
14	67.9	3.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
140	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
142	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
149	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
15	71.6	5.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
161	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
162	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
164	72.84	2.00E-32	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
165	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
166	71.6	5.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
17	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
170	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
176	71.6	3.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
178	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
18	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
183	66.67	7.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
184	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
189	71.6	4.00E-33	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
19	72.84	1.00E-34	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
195	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
196	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
197	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
2	66.67	7.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
202	67.9	3.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
203	65.43	2.00E-28	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
206	70.37	2.00E-32	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
207	71.6	5.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
208	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
21	67.9	3.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
210	71.6	2.00E-31	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
213	69.14	2.00E-29	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
216	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
218	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
22	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
220	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
221	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
222	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
225	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
227	66.67	7.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
229	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
231	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
233	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965

236	70.37	2.00E-32	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
237	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
238	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
24	71.6	5.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
241	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
242	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
245	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
249	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
25	71.6	3.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
250	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
255	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
256	72.84	4.00E-34	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
257	66.67	7.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
259	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
26	71.6	5.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
260	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
261	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
262	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
264	71.6	3.00E-32	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
265	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
27	62.96	2.00E-26	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
272	70.37	3.00E-32	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
273	67.9	2.00E-30	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
274	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
29	74.07	2.00E-35	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
3	67.9	3.00E-29	terminase [Clostridiales bacterium VE202-03]	gi|639707411|ref|WP_024723669.1|	1232439
30	64.2	2.00E-28	putative uncharacterized protein [Ruminococcus sp. CAG:17]	gi|547240587|ref|WP_021976510.1|	1262951
32	67.9	3.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
33	67.9	3.00E-31	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
34	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
36	76.54	2.00E-35	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
37	67.9	3.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
38	66.67	2.00E-29	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
4	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
40	66.67	3.00E-32	putative uncharacterized protein [Ruminococcus sp. CAG:17]	gi|547240587|ref|WP_021976510.1|	1262951
42	64.2	3.00E-26	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
44	66.67	5.00E-19	terminase [[Ruminococcus] torques]	gi|490985259|ref|WP_004846995.1|	33039
46	67.9	3.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
48	71.6	9.00E-34	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
5	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
50	86.42	3.00E-44	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
51	67.9	3.00E-30	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
54	65.43	6.00E-31	terminase [[Clostridium] symbiosum]	gi|489596073|ref|WP_003500516.1|	1512
59	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
6	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
60	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
62	69.14	2.00E-30	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
67	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
68	74.07	1.00E-34	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
7	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
71	70.37	2.00E-31	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
72	62.96	4.00E-27	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
75	70.37	8.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
77	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
8	56.79	9.00E-26	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
80	69.14	3.00E-31	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
81	74.07	7.00E-33	putative uncharacterized protein [Ruminococcus sp. CAG:17]	gi|547240587|ref|WP_021976510.1|	1262951
82	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
86	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
87	67.9	1.00E-27	terminase [[Clostridium] hathewayi]	gi|493833739|ref|WP_006781000.1|	154046
89	74.07	1.00E-34	TerL [Rhodococcus phage ReqiPoco6]	gi|593774729|ref|YP_009012597.1|	691964
9	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965
90	71.6	2.00E-33	TerL [Rhodococcus phage ReqiPepy6]	gi|593779801|ref|YP_009017628.1|	691965

SI Table 2. Taxonomic classification of closest homologs (HB1 phage family). Majority of OTUs (86 out of 123) have the closest match to ReqiPoco6 terminase large subunit, whereas 15 OTUs have closest homologs belonging to ReqiPepy6.

Closest Homolog Taxon ID	Kingdom	Phylum	Class	Order	Family	Genus	Species
1262951	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus	Ruminococcus sp. CAG:17
691964	Viruses	dsDNA viruses, no RNA stage	Caudovirales	Siphoviridae	unclassified Siphoviridae	Rhodococcus phage	ReqiPoco6
691965	Viruses	dsDNA viruses, no RNA stage	Caudovirales	Siphoviridae	unclassified Siphoviridae	Rhodococcus phage	ReqiPepy6
1512	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnoclostridium	Clostridium symbiosum
29347	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnoclostridium	Clostridium scindens
33039	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia	Ruminococcus torques
1232439	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified Clostridiales	unclassified Clostridiales	Clostridiales bacterium VE202-03
154046	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnoclostridium	Clostridium hathewayi

SI Table 3. Closest homolog to each OTU's representative sequence (HA phage family). Each OTU's representative sequence was used as a query for NCBI's BLASTx homology search against the non-redundant protein database. The table summarizes the E-value and the percent amino acid identity across the query sequence and the closest homolog, as well as the closest homolog's name, sequence ID, and taxon ID. The taxon ID is color coded, and the taxonomic classification corresponding to each taxon ID can be retrieved from the following table. Note with the exception of a few "putative uncharacterized" homolog names that most are identified as terminases or TerLs (terminase large subunits).

Query Sequence ID (OTU ID)	Percent Identity	E value	Closest Homolog	Closest Homolog Sequence ID	Closest Homolog Taxon ID
0	97.56	4.00E-49	hypothetical protein [Streptococcus sp. F0442]	gi|497418421|ref|WP_009732619.1|	999425
1	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
10	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
100	98.78	4.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
101	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
102	100	2.00E-49	terminase [Streptococcus infantis]	gi|493136448|ref|WP_006154887.1|	68892
103	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
104	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
106	100	2.00E-49	terminase [Streptococcus infantis]	gi|493136448|ref|WP_006154887.1|	68892
107	100	2.00E-49	terminase [Streptococcus infantis]	gi|493136448|ref|WP_006154887.1|	68892
108	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
109	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
11	98.78	1.00E-48	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
110	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
111	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
112	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
113	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
114	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
117	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
118	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
119	98.78	3.00E-49	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
12	100	2.00E-49	terminase [Streptococcus infantis]	gi|493136448|ref|WP_006154887.1|	68892
120	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
121	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
122	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
123	98.78	3.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
125	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
126	98.78	3.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
127	98.78	3.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
128	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
129	98.78	1.00E-48	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
13	98.78	1.00E-49	hypothetical protein [Streptococcus sp. F0442]	gi|497418421|ref|WP_009732619.1|	999425
130	98.78	2.00E-49	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
131	98.78	6.00E-48	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
132	98.78	1.00E-48	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
133	98.78	7.00E-49	terminase [Streptococcus infantis]	gi|493136448|ref|WP_006154887.1|	68892
134	98.78	6.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
135	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
136	98.78	5.00E-49	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
137	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
138	98.78	3.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
139	98.78	1.00E-48	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
140	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
141	98.78	1.00E-48	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
142	97.56	4.00E-48	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
143	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
145	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
146	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
147	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
148	98.78	2.00E-48	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
15	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
150	98.78	9.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
151	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
152	98.78	3.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
153	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
154	97.56	1.00E-48	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
155	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
156	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
157	98.78	3.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
158	97.56	1.00E-48	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
159	98.78	9.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
16	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
160	98.78	4.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
161	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
162	98.78	3.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
164	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
165	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
166	97.56	1.00E-48	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
167	100	2.00E-49	terminase [Streptococcus infantis]	gi|493136448|ref|WP_006154887.1|	68892
168	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
169	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
17	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
170	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
171	98.78	3.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
172	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
173	100	2.00E-49	terminase [Streptococcus infantis]	gi|493136448|ref|WP_006154887.1|	68892
174	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
175	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
176	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
177	97.56	1.00E-48	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
178	98.78	4.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
179	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
18	100	2.00E-49	terminase [Streptococcus infantis]	gi|493136448|ref|WP_006154887.1|	68892
180	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
181	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
182	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416

183	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
184	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
185	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
186	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
19	98.78	3.00E-49	terminase [Streptococcus parasanguinis]	gi|671602035|ref|WP_031575397.1|	1318
2	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
20	96.34	2.00E-48	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
21	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
22	91.46	8.00E-46	hypothetical protein [Streptococcus sp. F0442]	gi|497418421|ref|WP_009732619.1|	999425
23	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
24	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
25	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
26	98.78	3.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
27	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
28	98.78	7.00E-50	terminase [Streptococcus parasanguinis]	gi|671602035|ref|WP_031575397.1|	1318
29	98.78	7.00E-49	terminase [Streptococcus infantis]	gi|493136448|ref|WP_006154887.1|	68892
3	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
30	100	2.00E-49	terminase [Streptococcus infantis]	gi|493136448|ref|WP_006154887.1|	68892
31	98.78	6.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
32	98.78	4.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
33	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
34	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
35	98.78	1.00E-49	hypothetical protein [Streptococcus sp. F0442]	gi|497418421|ref|WP_009732619.1|	999425
36	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|565851306|ref|WP_023933954.1|	257758
37	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
38	98.78	7.00E-50	terminase [Streptococcus parasanguinis]	gi|671602035|ref|WP_031575397.1|	1318
39	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
4	98.78	4.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
40	98.78	2.00E-49	hypothetical protein [Streptococcus sp. F0442]	gi|497418421|ref|WP_009732619.1|	999425
42	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
43	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
44	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
45	98.78	9.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
47	98.78	7.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
48	98.78	3.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
49	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
5	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
50	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
51	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
52	98.78	3.00E-49	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
53	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
54	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
56	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
57	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
59	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
6	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
60	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
61	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
62	98.78	4.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
63	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
64	98.78	3.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
65	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
66	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
68	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
69	98.78	4.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
7	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
71	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
72	100	1.00E-49	ative phage terminase, large subunit [Streptococcus tigurir	gi|494783687|ref|WP_007519095.1|	1077464
73	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
74	98.78	3.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
76	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
77	98.78	3.00E-49	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
78	98.78	3.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
79	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
8	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
80	98.78	3.00E-49	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
82	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
83	97.56	1.00E-48	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
84	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
85	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
86	95.12	3.00E-48	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
87	98.78	3.00E-49	terminase [Streptococcus oralis]	gi|446545997|ref|WP_000623343.1|	1303
88	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
89	98.78	2.00E-48	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
9	98.78	3.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
90	98.78	4.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
93	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
95	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
96	98.78	4.00E-49	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
97	100	1.00E-49	terminase [Streptococcus pseudopneumoniae]	gi|446545996|ref|WP_000623342.1|	257758
98	100	9.00E-50	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416
99	97.56	1.00E-48	terminase [Streptococcus sp. SR1]	gi|726981126|ref|WP_033585808.1|	1161416

SI Table 4. Taxonomic classification of closest homologs to each OTU's representative sequence (HA phage family).

Closest Homolog Taxon ID	Kingdom	Phylum	Class	Order	Family	Genus	Species
1318	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	Streptococcus parasanguinis
1077464	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	Streptococcus tigurinus
257758	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	Streptococcus pseudopneumoniae
999425	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	Streptococcus sp. F0442
1161416	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	Streptococcus sp. SR1
1303	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	Streptococcus oralis
68892	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	Streptococcus infantis

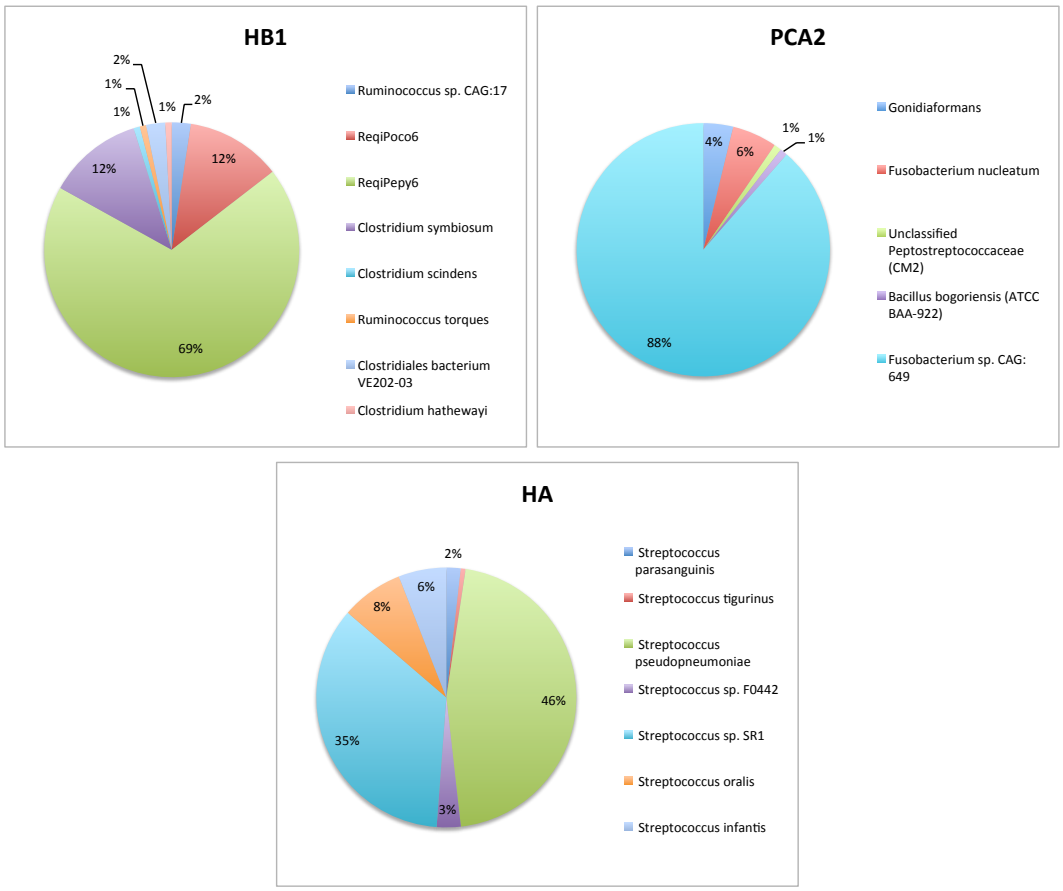
SI Table 5. Closest homolog to each OTU's representative sequence (PCA2 phage family). Each OTU's representative sequence was used as a query for NCBI's BLASTx homology search against the non-redundant protein database. The table summarizes the E-value and the percent amino acid identity across the query sequence and the closest homolog, as well as the closest homolog's name, sequence ID, and taxon ID. The taxon ID is color coded, and the taxonomic classification corresponding to each taxon ID can be retrieved from the following table. Note with the exception of a few "putative uncharacterized" homolog names, most are identified as terminases or TerLs (terminase large subunits).

Query Sequence ID (OTU ID)	Percent Identity	E value	Closest Homolog	Closest Homolog Sequence ID	Closest Homolog Taxon ID
0	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
10	90	2.00E-28	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
100	76.27	4.00E-21	terminase [Peptostreptococcaceae bacterium CM2]	gi|497213446|ref|WP_009527708.1|	796939
101	98.33	1.00E-30	terminase [Fusobacterium nucleatum]	gi|495968206|ref|WP_008692785.1|	851
103	98.33	2.00E-30	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
104	75	1.00E-21	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
107	98.33	1.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
108	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
109	75	1.00E-21	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
11	93.33	3.00E-30	terminase [Fusobacterium periodonticum]	gi|496096975|ref|WP_008821482.1|	860
110	75	1.00E-21	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
111	98.33	3.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
112	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
113	98.33	6.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
115	73.33	4.00E-21	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
116	73.33	5.00E-21	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
117	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
12	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
120	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
123	83.33	6.00E-25	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
124	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
126	98.33	1.00E-30	terminase [Fusobacterium nucleatum]	gi|495968206|ref|WP_008692785.1|	851
127	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
128	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
129	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
13	35.59	4.00E-04	terminase [Bacillus bogoriensis]	gi|651939129|ref|WP_026673624.1|	246272
132	98.33	1.00E-30	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
135	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
136	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
137	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
138	75	1.00E-21	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
139	96.67	4.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
14	96.67	5.00E-30	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
140	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
141	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
142	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
143	91.67	1.00E-29	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
144	98.33	4.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
146	98.33	2.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
147	98.33	6.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
149	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
15	73.33	2.00E-21	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
150	98.33	2.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
151	98.33	3.00E-31	terminase [Fusobacterium periodonticum]	gi|496096975|ref|WP_008821482.1|	860
152	98.33	1.00E-30	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
153	98.33	2.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
154	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
155	98.33	6.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
156	98.33	5.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
16	98.33	1.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
2	91.67	1.00E-29	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
21	73.33	7.00E-21	terminase [Fusobacterium nucleatum]	gi|495968206|ref|WP_008692785.1|	851
22	75	1.00E-21	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
23	73.33	5.00E-21	terminase [Fusobacterium nucleatum]	gi|495968206|ref|WP_008692785.1|	851
24	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
25	98.33	4.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
26	98.33	1.00E-30	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
27	75	1.00E-21	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
29	98.33	2.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
3	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
32	98.33	3.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
33	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
34	98.33	4.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
36	96.67	1.00E-29	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
4	96.67	1.00E-30	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
40	75	1.00E-21	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
41	98.33	3.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
45	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
47	73.33	1.00E-21	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
5	75	1.00E-21	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
50	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
51	98.33	2.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
54	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
56	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
60	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
62	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
63	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
64	93.33	3.00E-29	terminase [Fusobacterium periodonticum]	gi|496096975|ref|WP_008821482.1|	860
65	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
66	98.33	3.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
67	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900

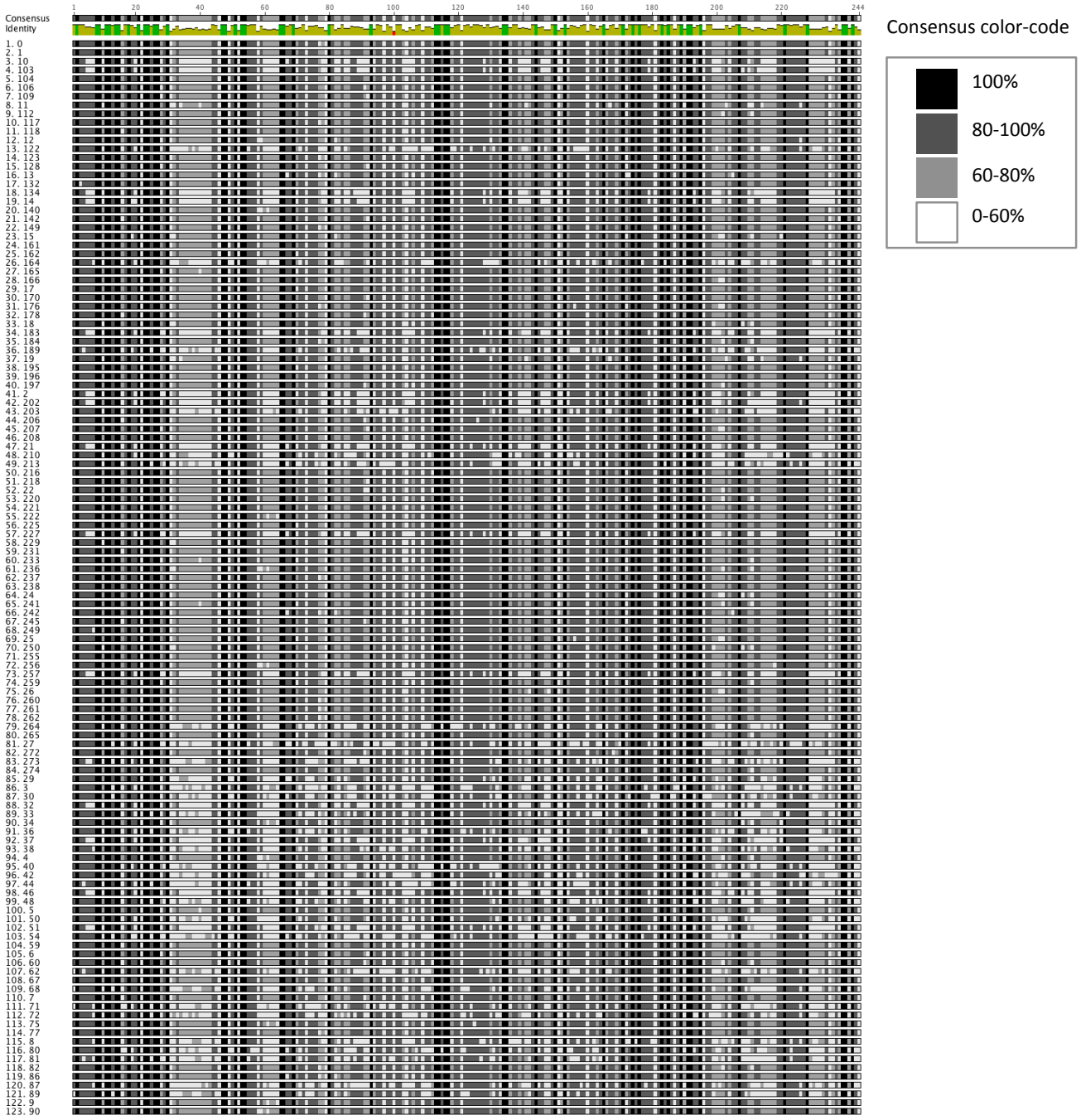
70	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
72	98.33	3.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
73	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
75	96.67	2.00E-30	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
77	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
78	90	5.00E-29	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
79	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
8	95	5.00E-30	terminase [Fusobacterium periodonticum]	gi|496096975|ref|WP_008821482.1|	860
80	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
81	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
82	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
83	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
84	98.33	1.00E-30	terminase [Fusobacterium nucleatum]	gi|495968206|ref|WP_008692785.1|	851
86	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
89	90	5.00E-29	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
90	98.33	3.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
91	98.33	3.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
92	98.33	8.00E-31	terminase [Fusobacterium nucleatum]	gi|495968206|ref|WP_008692785.1|	851
93	96.67	9.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
94	96.67	4.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
95	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
96	100	7.00E-32	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
97	98.33	3.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900
99	98.33	3.00E-31	putative phage terminase large subunit [Fusobacterium sp. CAG:649]	gi|547450305|ref|WP_022069933.1|	1262900

SI Table 6. Taxonomic classification of closest homologs to each OTU's representative sequence (PCA2 phage family).

Closest Homolog Taxon ID	Kingdom	Phylum	Class	Order	Family	Genus	Species
860	Bacteria	Fusobacteria	Fusobacteria	Fusobacteriales	Fusobacteriaceae	Fusobacterium	gonidiaformans
851	Bacteria	Fusobacteria	Fusobacteria	Fusobacteriales	Fusobacteriaceae	Fusobacterium	Fusobacterium nucleatum
796939	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	unclassified Peptostreptococcaceae	unclassified Peptostreptococcaceae (CM2)
246272	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus	Bacillus bogoriensis (ATCC BAA-922)
1262900	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium	Fusobacterium sp. CAG:649



SI Figure 6. Percentage of HB1, PCA2 and HA phage family OTUs belonging to each taxonomic group identified in SI Figure 4, SI Figure 5, and SI Table 3.



SI Figure 7. The nucleotide alignment of HB1 phage family OTU representative sequences. Sequences were aligned using Geneious (68). No gaps were introduced. Each base is color-coded according to its relative abundance within a column in the alignment. Conserved bases are black and highly variable sites are denoted in white.

5.6 References

1. Suttle CA, Chan AM, & Cottrell MT (1990) Infection of phytoplankton by viruses and reduction of primary productivity. *Nature* 347(6292):467-469.
2. Bergh Ø, BØrsheim KY, Bratbak G, & Heldal M (1989) High abundance of viruses found in aquatic environments. *Nature* 340(6233):467-468.
3. Suttle CA (2007) Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology* 5(10):801-812.
4. Rohwer F & Thurber RV (2009) Viruses manipulate the marine environment. *Nature* 459(7244):207-212.
5. Roux S*, et al.* (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*.
6. Mokili JL, Rohwer F, & Dutilh BE (2012) Metagenomics and future perspectives in virus discovery. *Current opinion in virology* 2(1):63-77.
7. Navarro F & Muniesa M (2017) Phages in the Human Body. *Frontiers in microbiology* 8.
8. Youle M, Haynes M, & Rohwer F (2012) Scratching the surface of biology's dark matter. *Viruses: Essential agents of life*, (Springer), pp 61-81.
9. Paez-Espino D*, et al.* (2016) Uncovering Earth's virome. *Nature* 536(7617):425-430.
10. Reyes A, Semenkovich NP, Whiteson K, Rohwer F, & Gordon JI (2012) Going viral: next generation sequencing applied to human gut phage populations. *Nature Reviews. Microbiology* 10(9):607.
11. Penadés JR, Chen J, Quiles-Puchalt N, Carpena N, & Novick RP (2015) Bacteriophage-mediated spread of bacterial virulence genes. *Current opinion in microbiology* 23:171-178.

12. Stone R (2002) Food and agriculture: testing grounds for phage therapy. *Science* 298(5594):730-730.

13. Hug LA*, et al.* (2016) A new view of the tree of life. *Nature Microbiology* 1:16048.

14. Yarza P*, et al.* (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews. Microbiology* 12(9):635.

15. Woese CR, Kandler O, & Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences* 87(12):4576-4579.

16. Barberán A*, et al.* (2015) Continental-scale distributions of dust-associated bacteria and fungi. *Proceedings of the National Academy of Sciences* 112(18):5756-5761.

17. Shanks OC*, et al.* (2013) Comparison of the microbial community structures of untreated wastewaters from different geographic locales. *Applied and environmental microbiology* 79(9):2906-2913.

18. Wu GD*, et al.* (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334(6052):105-108.

19. Caporaso JG*, et al.* (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences* 108(Supplement 1):4516-4522.

20. Yatsunenko T*, et al.* (2012) Human gut microbiome viewed across age and geography. *nature* 486(7402):222.

21. Lander ES & Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2(3):231-239.

22. Milo R, Jorgensen P, Moran U, Weber G, & Springer M (2010) BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic acids research* 38(suppl 1):D750-D753.

23. Escobar-Zepeda A, de León AV-P, & Sanchez-Flores A (2015) The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in genetics* 6.

24. Yu F*, et al.* (2017) Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *bioRxiv*:114496.

25. Dutilh BE*, et al.* (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature communications* 5.

26. Li Y*, et al.* (2016) VIP: an integrated pipeline for metagenomics of virus identification and discovery. *Scientific reports* 6.

27. Wüthrich D*, et al.* (2016) Exploring the virome of cattle with non-suppurative encephalitis of unknown etiology by metagenomics. *Virology* 493:22-30.

28. Manso CF, Bibby DF, & Mbisa JL (2017) Efficient and unbiased metagenomic recovery of RNA virus genomes from human plasma samples. *Scientific Reports* 7.

29. Sanschagrin S & Yergeau E (2014) Next-generation sequencing of 16S ribosomal RNA gene amplicons. *Journal of visualized experiments: JoVE* (90).

30. Steven B, Gallegos-Graves LV, Starkenburg SR, Chain PS, & Kuske CR (2012) Targeted and shotgun metagenomic approaches provide different descriptions of dryland soil microbial communities in a manipulated field study. *Environmental microbiology reports* 4(2):248-256.

31. Tessler M*, et al.* (2017) Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports* 7.

32. Modi SR, Lee HH, Spina CS, & Collins JJ (2013) Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* 499(7457):219-222.

33. Duerkop BA & Hooper LV (2013) Resident viruses and their interactions with the immune system. *Nature immunology* 14(7):654-659.

34. Anantharaman K*, et al.* (2014) Sulfur oxidation genes in diverse deep-sea viruses. *Science* 344(6185):757-760.

35. Ng TFF*, et al.* (2011) Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PloS one* 6(6):e20579.

36. Grose JH & Casjens SR (2014) Understanding the enormous diversity of bacteriophages: the tailed phages that infect the bacterial family Enterobacteriaceae. *Virology* 468:421-443.

37. Suttle CA (2005) Viruses in the sea. *Nature* 437(7057):356-361.

38. Langfeldt D*, et al.* (2014) Composition of microbial oral biofilms during maturation in young healthy adults. *PloS one* 9(2):e87449.

39. Szafranski SP*, et al.* (2015) High-resolution taxonomic profiling of the subgingival microbiome for biomarker discovery and periodontitis diagnosis. *Applied and environmental microbiology* 81(3):1047-1058.

40. Bik EM*, et al.* (2010) Bacterial diversity in the oral cavity of 10 healthy individuals. *The ISME journal* 4(8):962-974.

41. Edlund A, Santiago-Rodriguez TM, Boehm TK, & Pride DT (2015) Bacteriophage and their potential roles in the human oral cavity. *Journal of oral microbiology* 7.

42. Nasidze I, Li J, Quinque D, Tang K, & Stoneking M (2009) Global diversity in the human salivary microbiome. *Genome research* 19(4):636-643.

43. Chen T*, et al.* (2010) The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* 2010.

44. Casjens S (2003) Prophages and bacterial genomics: what have we learned so far? *Molecular microbiology* 49(2):277-300.

45. Tadmor AD & Phillips R (2015) Host-Virus Interaction: From Metagenomics to Single-Cell Genomics. *Encyclopedia of Metagenomics*, (Springer), pp 257-265.

46. Xie G*, et al.* (2010) Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Molecular oral microbiology* 25(6):391-405.

47. Belda-Ferre P*, et al.* (2012) The oral metagenome in health and disease. *The ISME journal* 6(1):46.

48. Turnbaugh PJ*, et al.* (2007) The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* 449(7164):804.

49. Prescott JF (1991) Rhodococcus equi: an animal and human pathogen. *Clinical microbiology reviews* 4(1):20-34.

50. Chao A, Chazdon RL, Colwell RK, & Shen TJ (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology letters* 8(2):148-159.

51. Bray JR & Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs* 27(4):325-349.

52. Lozupone C & Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology* 71(12):8228-8235.

53. Breitbart M & Rohwer F (2005) Here a virus, there a virus, everywhere the same virus? *Trends in microbiology* 13(6):278-284.

54. Holmfeldt K*, et al.* (2013) Twelve previously unknown phage genera are ubiquitous in global oceans. *Proceedings of the National Academy of Sciences* 110(31):12798-12803.

55. Hall MW*, et al.* (2017) Inter-personal diversity and temporal dynamics of dental, tongue, and salivary microbiota in the healthy oral cavity. *npj Biofilms and Microbiomes* 3(1):2.

56. Costello EK*, et al.* (2009) Bacterial community variation in human body habitats across space and time. *Science* 326(5960):1694-1697.

57. Abeles SR*, et al.* (2014) Human oral viruses are personal, persistent and gender-consistent. *The ISME journal* 8(9):1753-1767.

58. Pride DT*, et al.* (2012) Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *The ISME journal* 6(5):915-926.

59. Oh J*, et al.* (2016) Temporal stability of the human skin microbiome. *Cell* 165(4):854-866.

60. Belstrøm D*, et al.* (2016) Temporal stability of the salivary microbiota in oral health. *PLoS One* 11(1):e0147472.

61. Lloyd-Price J*, et al.* (2017) Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* 550(7674):61-66.

62. Franzosa EA*, et al.* (2015) Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences* 112(22):E2930-E2938.

63. Tadmor AD, Ottesen EA, Leadbetter JR, & Phillips R (2011) Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science* 333(6038):58-62.

64. Pruitt KD, Tatusova T, & Maglott DR (2006) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* 35(suppl_1):D61-D65.

65. Hamady M, Walker JJ, Harris JK, Gold NJ, & Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature methods* 5(3):235-237.

66. Caporaso JG*, et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature methods* 7(5):335-336.

67. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461.

68. Kearse M*, et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12):1647-1649.

69. Bastian M, Heymann S, & Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. *Icwsm* 8:361-362.