Penn Libraries
University of Pennsylvania

University of Pennsylvania
ScholarlyCommons

Publicly Accessible Penn Dissertations

1-1-2012

Evolution and Dynamics of the Human Gut Virome

Samuel Schwartz Minot
University of Pennsylvania, sminot@mail.med.upenn.edu

Follow this and additional works at: http://repository.upenn.edu/edissertations

Part of the Biology Commons, and the Virology Commons

Recommended Citation

Minot, Samuel Schwartz, "Evolution and Dynamics of the Human Gut Virome" (2012). *Publicly Accessible Penn Dissertations*. 552.
http://repository.upenn.edu/edissertations/552

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/edissertations/552
For more information, please contact libraryrepository@pobox.upenn.edu.

Evolution and Dynamics of the Human Gut Virome

Abstract
Advisor: Frederic D. Bushman, PhD.

The human body contains large numbers of viral particles (over 1012 per person), largely bacteriophage, but little is known of how these viral communities influence human health and disease. To study the viruses of the human gut (the so-called gut `virome') during a known environmental perturbation we collected stool samples from healthy individuals participating in a controlled diet study. Viral DNA was purified and deep-sequenced using 454 and Illumina technologies, yielding over 48 billion bases of viral sequence spread across 28 samples from 12 healthy individuals. Computational analysis of this unprecedentedly large database of viral sequences allowed us to characterize these communities on a genomic level. We found that the vast majority of viruses from the human gut were novel species of bacteriophage, and that only 1 of these 12 individuals contained a known eukaryotic DNA virus. Temporal changes in these viral communities were correlated with experimental manipulation of diet, and parallel deep sequencing of gut bacteria revealed co-variation between bacterial and viral communities, supporting the hypothesis of linked reproduction between these two groups. A large proportion of viral contigs have markers of temperate lifestyle, indicating that there is a significant role of lysogeny in the gut microbiome. Analysis of genetically variable elements within these viral genomes revealed novel classes of diversity-generating retroelements targeting immunoglobulin-superfamily proteins, suggesting a surprising example of convergent evolution with the vertebrate immune system. Optimization of assembly algorithms for these samples improved the recovery of complete and partial genome sequences. While the assembled genomes were highly dissimilar on the nucleotide level, analysis of syntenic protein-coding sequences revealed conserved gene cassettes that display an inferred structural and functional conservation despite a high degree of nucleotide substitution. Through high-throughput shotgun sequencing of viral DNA, we found that the healthy human gut contains a wide variety of extremely diverse bacteriophages encoding novel and unexpected functions. This work sets the stage for thorough genomic analysis of complex viral communities, and presents the intriguing problem of how this immense pool of genetic diversity has evolved and persisted.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Cell & Molecular Biology

First Advisor
Frederic D. Bushman

Second Advisor
Ronald Collman

This dissertation is available at ScholarlyCommons: http://repository.upenn.edu/edissertations/552

Keywords
Bacteriophage, Evolution, High-Throughput Sequencing, Microbiome

Subject Categories
Biology | Microbiology | Virology

This dissertation is available at ScholarlyCommons: http://repository.upenn.edu/edissertations/552

EVOLUTION AND DYNAMICS OF THE HUMAN GUT VIROME

Samuel Schwartz Minot

A DISSERTATION

in

Cell and Molecular Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2012

Supervisor of Dissertation

____

Frederic D. Bushman PhD.
Professor of Microbiology

Dissertation Committee

Ronald Collman, M.D.
Professor of Medicine
Jeffery Weiser, M.D.,
Professor of Pediatrics and Microbiology
Paul Bates, PhD.
Professor of Microbiology
Mark Goulian, PhD.
Professor of Biology

Graduate Group Chairperson

____

Daniel S. Kessler, PhD.
Graduate Group Chair
Associate Professor of Cell and Developmental Biology

DEDICATION

To my family, for raising me to read, cook, and find joy in the esoteric.

ACKNOWLEDGMENT

I am grateful to my advisor, Rick Bushman, for his guidance and support as I pursued an unconventional project in his lab. Whenever the path forward seemed unclear, he had the insight that helped me forge ahead. I also thank my thesis committee for their advice and support, Jeff Weiser, Paul Bates, Ron Collman, and Mark Goulian. In particular I owe this work to the knowledge and support of the whole Bushman lab, who taught me the diverse set of skills I needed to work independently. Christian Hoffman for helping me develop extractions and the ecological foundation for the project; Kyle Bittinger, Rohini Sinha, and Nirav Malani for teaching me what little programming I know; Greg Peterfreund, Rithun Mukherjee, and Emily Charlson for stimulating and lengthy discussions; Young Hwang for hours spent together troubleshooting cloning or purifying proteins in the cold room; Stephanie Grunberg for her patience and for always demonstrating a superior standard of labwork; Aubrey Bailey for keeping Mumak running under the heaviest loads; Scott Sherril-Mix for his analytical insight; and the whole lab for providing support and assistance whenever it was needed.

For his help as I navigated my graduate career, I thank Bob Doms, who provided advice and assistance that was always focused on what would help me grow and be successful throughout my scientific life.

For her support in life outside of lab I thank above all others Kristen van der Veen, as well as Marian Schwartz, Reid Minot, John Minot, and the larger Minot confederation, Dianne van der Veen, Steve van der Veen, and the entire van der Veen clan, who even honored my research at our wedding. For keeping Johnson Pavilion connected to the outside world, I thank Hilary DeBardeleben, Jason Diaz, Nathaniel Snyder, Rob Plasschaert, and Sudil Mahendra,

because friends that can talk science but don't have to are invaluable. The entire Bushman lab was a constant source of support and encouragement, and I want to thank Karen Ocweija, Caitlin Greig, Mali Skotheim, Troy Brady, Frances Male, Rebecca Custers-Allen, Serena Dollive, and Shannah Roth.

ABSTRACT

EVOLUTION AND DYNAMICS OF THE HUMAN GUT VIROME

Samuel Schwartz Minot

Frederic D. Bushman, PhD.

The human body contains large numbers of viral particles (over 10^{12} per person), largely bacteriophage, but little is known of how these viral communities influence human health and disease. To study the viruses of the human gut (the so-called gut ‘virome’) during a known environmental perturbation we collected stool samples from healthy individuals participating in a controlled diet study. Viral DNA was purified and deep-sequenced using 454 and Illumina technologies, yielding over 48 billion bases of viral sequence spread across 28 samples from 12 healthy individuals. Computational analysis of this unprecedentedly large database of viral sequences allowed us to characterize these communities on a genomic level. We found that the vast majority of viruses from the human gut were novel species of bacteriophage, and that only 1 of these 12 individuals contained a known eukaryotic DNA virus. Temporal changes in these viral communities were correlated with experimental manipulation of diet, and parallel deep sequencing of gut bacteria revealed co-variation between bacterial and viral communities, supporting the hypothesis of linked reproduction between these two groups. A large proportion of viral contigs have markers of temperate lifestyle, indicating that there is a significant role of lysogeny in the gut microbiome. Analysis of genetically variable elements within these viral genomes revealed novel classes of diversity-generating retroelements targeting immunoglobulin-superfamily proteins, suggesting a surprising example of convergent evolution with the vertebrate immune system. Optimization of assembly algorithms for these samples improved the recovery of complete and partial genome sequences. While the assembled genomes were highly dissimilar on the nucleotide level, analysis of syntenic protein-coding sequences revealed conserved gene cassettes that display an inferred structural and functional conservation despite a high degree of nucleotide substitution. Through high-throughput shotgun sequencing of viral DNA, we found that the healthy human gut contains a wide variety of extremely diverse bacteriophages encoding novel and unexpected functions. This work sets the stage for thorough genomic analysis of complex viral communities, and presents the intriguing problem of how this immense pool of genetic diversity has evolved and persisted.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 – Introduction

In microbial communities from the oceans to the human gut, there are at least as many virus-like particles as microbial cells [1,2]. However, little is known of how these viruses impact microbial communities and human health. This dissertation interrogates the composition, dynamics, and evolution of the viral communities of the human gut – the human gut 'virome' – in order to gain insight into how they affect human health and disease. First the salient technological and biological systems are described in the introductory chapter, and then later chapters explore these issues in greater detail.

The microbiota of the human gut are of particular interest to human health, not only due to their role in harvesting energy from food [3], but also through their strong interaction with the human immune system [4]. While human disease states such as obesity and inflammatory bowel disease are accompanied by changes in the gut microbiome [3,5,6], the field is still developing its knowledge of how those microbial changes are initiated, are perpetuated, and lead to disease. For example, studies have shown that gut bacterial composition is influenced by diet [7], geography [8], and method of delivery at birth [9], and there are likely many more important factors that are as yet unknown. Given the large role of the gut microbiome in human diseases of metabolism and inflammation, a goal of this field is to eventually design therapeutics that treat diseases of metabolism and inflammation through direct manipulation of the human microbiome.

1.1 Sampling the human gut virome

Prior to the advent of high-throughput sequencing, the composition of viral communities of the human body could only be surveyed by physical methods (such as electron microscopy), which suggested that the major types resembled tailed bacteriophages [2]. This was confirmed by sequencing approaches, which found that DNA viruses were generally novel bacteriophage species and that RNA viruses were predominantly plant viruses [10,11].

Many analytical techniques take the approach of purifying viral nucleic acids away from human and bacterial DNA and RNA [12]. The methods used to purify viral particles vary according to the characteristics of viral particles in the sample. In seawater, with a relatively low density of particulates, tangential flow filtration has been used to concentrate viral particles, following an initial filtering step at 0.2µm or 0.4µm to remove cells. The samples used in this work are human stool, which has a high concentration of both viral particles and other contaminating particulates. For this application, isopycnic cesium chloride density ultra-centrifugation is used on samples that have been suspended in buffer and filtered at 0.2µm to remove cells. Such a purification technique has been shown previously to isolate viral particles and exclude bacterial cells and free DNA [10,12,13]. These methods have added the power of metagenomic analysis to the study of viral ecology and pathogenesis.

1.2 Eukaryotic viruses in the human microbiome

One of the primary questions about the composition of the human gut virome is the presence and abundance of eukaryotic viruses. The first sequence-based studies of DNA viruses purified from healthy human stool found no convincing evidence of eukaryotic viruses, with a small proportion of reads containing a limited similarity to known eukaryotic viruses [10,14]. Moreover, while studies that focused on people with idiopathic diarrheal symptoms have found a wide variety of pathogenic RNA viruses, they only found a single (adenoviral) representative of eukaryotic DNA viruses [15,16]. One of the challenges of these studies is that a small number of sequence reads with only weak database hits that do not reach statistical significance may truly result from a novel pathogenic virus at low abundance [17]. Moreover, it is possible that a novel virus may not be captured by physical sampling techniques or comparison of short sequence reads to reference databases. If the goal of viral sequencing is to diagnose idiopathic disease and influence medical treatment, such a conclusion is unsatisfying. With the continued development of sequencing technology, an acceptable diagnostic standard for identification of human viruses may become the complete assembly of a recognizable pathogen, which would help to differentiate between spurious hits and novel viruses [17]. Complete assembly of possible novel DNA viruses from stool samples has not been achieved in studies prior to this dissertation, which have been limited to pyrosequencing technology. One important development in the identification of novel DNA viruses has been use of Rolling-Circle Amplification (RCA), which enriches small circular genomes and has resulted in the finding that multiple novel polyomaviruses are chronically shed from healthy human skin [18,19]. The instances where shotgun sequencing has yielded novel eukaryotic viruses appear to currently be limited to viruses isolated from animals in disease states, including human children [17,20], salmon [21], bats [22], harbor seals [23], dogs [24], and possum [25]. In contrast, the question of whether healthy humans harbor potentially pathogenic viruses is of considerable interest to the study of human disease, and will be addressed by this thesis by using ultra-deep Illumina sequencing technology.

1.3 Kill-the-winner models of microbial dynamics

There are two major mechanisms by which bacteriophages are thought to influence bacterial communities. The first is through predatory pressure. There is a body of speculation that bacteriophages control bacterial abundance through so-called “kill-the-winner” dynamics, where the growth of any single bacterial species results in the subsequent growth of its phage (following Lotka-Volterra dynamics), which then lowers the bacterial abundance to its original level [26-28]. This hypothesis is based in part on a theoretical model of microbial community dynamics [26,29].

One of the basic assumptions of this model is that most bacteriophages are obligately lytic – they reproduce through replication inside host cells that quickly results in lysis. Another central assumption is that bacteriophages are highly host specific, such that a given phage or phage strain can only infect a subset of the bacteria that belong to a single species. The predicted community dynamics that result from this type of predation are different from those of a community that lacks bacteriophages in a few key ways: a greater number of bacterial genotypes or ‘species’ are able to coexist, and those species exist in a more even set of proportions, such that small numbers of strains cannot dominate [28]. One argument supporting “kill-the-winner” dynamics in the environment is that the level of bacterial diversity observed in nature is greater than could be expected in its absence [30]. It should be noted that this type of diversity is both at the level of the number of distinct bacterial genotypes as well as the even distribution of those genotypes in an environment. However, there are a number of complexities inherent to microbial communities that have been demonstrated to be important to this system, including lysogeny [31], spatial structure [32], and the fitness cost of resistance [33], and have yet to be incorporated into these models. Another argument for “kill-the-winner” has to do with the composition of bacterial genomes. It has been observed that regions of bacterial genomes that are strain-specific (sometimes called the ‘dispensable’ component of the pan-genome[34]) are disproportionately enriched in functions of resistance to phage infection [30]. One difficulty of this analysis is that many bacterial genes are unannotated, and that existing annotations are not always accurate. However the conclusions are consistent with the hypothesis that bacteria experience considerable selective pressure from a set of strain-specific bacteriophages. In addition to explaining an unexpectedly even distribution of bacterial genotypes, this model also explains the high level of genomic diversity within bacterial species, proposing that the bacterial pan-genome exists to combat phage predation [30]. While both of these lines of argument are suggestive, they fall short of directly testing the “kill-the-winner” model. In order to rigorously test this hypothesis, both bacterial and bacteriophage communities would need to be characterized with a temporal resolution close to that of the generation time, which may be as short at 30 minutes [35]. Moreover, these strains would need to be monitored at the level of genomic composition as well as absolute abundance. As challenging as it may appear to test the “kill-the-winner”, the needed techniques may soon be within reach, due to rapid advances in high-throughput sequencing, computational analysis, and mathematical modeling that will be developed in part by this dissertation.

1.4 Bacteriophage-mediated horizontal gene transfer

The second mechanism by which bacteriophages may influence bacterial communities is the horizontal transfer of genetic material. The lifecycle of lysogenic bacteriophages can include periods where they exist as prophages, replicating as either an integrated part of the host chromosome [31] or as a plasmid [36]. For these lysogens the phage genome itself can be classified as horizontally transferred DNA. Moreover, genomic diversity within bacterial species often includes integrated prophages that are specific to one strain or another [37-39]. This type of horizontal transfer has been demonstrated to impact human health, as there are multiple instances of bacterial toxins causing human disease that are carried on and mobilized by lysogenic bacteriophages, including cholera and diphtheria toxin [40]. Even in the absence of known resistance genes, various prophages of E. coli have been demonstrated to confer resistance to antibiotics [41]. A study of viruses from the human gut found an enrichment of genes encoding glycan metabolism on well-assembled bacteriophage contigs, which may be used in host bacterial carbohydrate metabolism [13]. In the ocean, bacteriophages are observed to contain a variety of metabolic genes that may carry out carbon and phosphate metabolism as well as photosynthesis

[42,43]. Moreover, generalized transduction – the non-specific packaging of host DNA into phage particles and subsequent transfer between cells – has been demonstrated to mobilize antibiotic resistance in gut bacteria [44]. Because bacteria are observed to have undergone a high degree of horizontal gene transfer that is apparently independent of known plasmids, transposons, or lysogenic phages [45], the role of novel phage in horizontal gene transfer is of great interest. It is worth mentioning that the evolutionary pressures and models of population dynamics that are implied by the "kill-the-winner" model of phage predation are significantly different from those implied by a temperate lifestyle, and little is known of how those conflicting pressures are balanced in the environment of the microbiome. This thesis will specifically address the degree of lysogeny within bacteriophages of the human gut, adding to our understanding of the role of horizontal gene transfer within the human microbiome.

1.5 Diversity-generating retroelements

Many studies of the human microbiome focus on the representation of bacterial taxa or gene families across different environments and samples. However, the decreasing cost of high-throughput sequencing will enable researchers to additionally characterize the nucleotide polymorphisms present within metagenomic samples, which can yield insight into functional changes and selective pressures acting on genomes. While investigating highly variable elements of viral genomes, we explore a bacteriophage-encoded system of directed mutagenesis carried out by a diversity-generating retroelement (DGR). This element was discovered in the *Bordetella* phage BPP-1 as the determinant of rapid host-switching [46]. Species of *Bordetella* undergo phase variation, where the proteins on the outer cell surface are rapidly exchanged. *Bordetella* exists in either a positive phase or minus phase, corresponding to which set of proteins is displayed on its surface [47]. The phage BPP-1 is able to infect the positive phase variant of

Bordetella, but not the minus phase variant [48]. However, BPP-1 has a high rate of switching between variants that can infect either the negative phase, the positive phase, or both phases. It was discovered that this host-tropism switching depends on what has been termed a “diversity-generating retroelement,” consisting of an error-prone reverse transcriptase, template repeat, and variable repeat (each repeat is ~100bp in length). The variable repeat is located within a gene that encodes the major tropism determinant (mtd) protein, which forms part of the phage tail. Through the action of this reverse transcriptase (which is not the replicative polymerase), the template repeat is transcribed and reverse-transcribed such that each adenine is mutated to a random base. The mutagenized copy is then used to copy over the variable repeat locus. In this manner, a small number of codons are mutagenized while the rest of the protein remains constant. This model of copy-and-paste DNA mutagenesis through an RNA intermediate is supported by experiments in which a self-splicing intron was inserted into the template repeat, but was not found in the resulting variable repeat sequence [49]. The mutagenized amino acids are found in the binding pocket of the mtd protein, and are responsible for the binding of the phage to its host [50]. Because this system directly modifies the bacteriophage genome, the nucleotide changes in the mtd gene are inherited normally by its progeny. It is not known what the rate of variation is *in vivo*, or how many genomes are altered per generation. Through the activity of its DGR, the phage BPP-1 is able to quickly change the host surface molecule that it binds to, and thereby rapidly switch tropism. DGRs have not been mechanistically described in any organism other than BPP-1--in this thesis we report that they are present and active in the human microbiome.

1.6 Clustered, regularly interspaced, short palindromic repeats (CRISPRs)

The strong selective pressure of phage predation in the environment can be seen in the wide array of genetic systems that bacteria have evolved to resist infection. One of the resistance mechanisms described in this thesis is that of clustered, regularly interspaced, short palindromic repeats (CRISPRs). CRISPRs are adaptive and sequence-specific, and their activity leaves a heritable genomic record of past infections. CRISPRs are widespread throughout both Bacteria and Archaea and can be encoded by different guilds of effector proteins (the CRISPR associated, or ‘cas’ proteins), which each carry out the same series of actions: 1) Non-self DNA (from an invading bacteriophage or plasmid) is recognized and cleaved into short (26-72bp) segments. 2) Those segments are incorporated into the CRISPR locus, which consists of alternating direct repeats (21-48bp) and spacers. Spacers correspond to the short, cleaved segments of foreign DNA, and a single CRISPR locus may contain up to hundreds of spacers [51]. 3) The CRISPR array is transcribed and processed into short crRNAs that are made up of a single spacer and a small amount of adjacent repeat sequence. 4) Those crRNAs form a complex with cas proteins such that complimentary DNA sequences (other than the source CRISPR array) are recognized and subsequently targeted for degradation. In this manner foreign DNAs with homology to CRISPR spacers are recognized and degraded. CRISPR arrays evolve rapidly in a wide variety of organisms, apparently in response to infection by bacteriophages and plasmids. While CRISPRs have only been identified within a single bacteriophage genome (a prophage of Clostridium [52]), it remains to be seen whether the system has been effectively adopted by free-living bacteriophages, possibly as a means of superinfection immunity. Bacteriophages have clearly evolved multiple mechanisms to restrict superinfection by other phage species [31,53,54], and the CRISPR system would yield all the same benefits of adaptive immunity to a bacteriophage as to its host. However, the fitness cost of a CRISPR locus may be higher for bacteriophages (compared to bacteria) due to either a constraint on genome size, or a low rate of superinfection. Nevertheless, by interrogating viral communities of the human gut virome, this work reports that

CRISPR loci are found in free-living bacteriophages, providing important insight into the evolutionary range and novelty of this new and fascinating form of prokaryotic immunity.

1.7 Summary

In order to characterize the human gut virome broadly, this work will focus on three different scales of evolutionary time. Chapter 2 investigates the differences between viral communities within different individuals, how these communities change over days, and how they respond to environmental perturbation. As an early survey of viral communities using high-throughput sequencing, this section also explores the role of CRISPRs and lysogeny. Chapter 3 focuses on polymorphisms that can be found within genes of individual viral species, using hypervariable regions to infer biological function. This section uncovers an unexpectedly rich collection of diversity-generating retroelements in the human gut virome that target a set of novel and intriguing protein folds. Chapter 4 develops a computational pipeline of *de novo* sequence assembly designed for viral communities and uses those long, high-quality genomes to explore long-term evolution in bacteriophage genomes. While these genomes have little similarity on the nucleotide level, they contain open reading frames that are conserved at the level of encoded protein sequence, gene order, and gene orientation. This thesis will broadly explore the composition, dynamics, and evolution of the human gut virome, providing novel insight into diverse genetic systems and selective pressures in this important environment.

1.8 References

1. Suttle CA (2005) Viruses in the sea. Nature 437: 356-361.
2. Weinbauer MG (2004) Ecology of prokaryotic viruses. FEMS Microbiol Rev 28: 127-181.
3. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027-1031.
4. Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI (2011) Human nutrition, the gut microbiome and the immune system. Nature 474: 327-336.

5. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65.
6. Greenblum S, Turnbaugh PJ, Borenstein E (2012) Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc Natl Acad Sci U S A 109: 594-599.
7. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334: 105-108.
8. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, et al. (2012) Human gut microbiome viewed across age and geography. Nature 486: 222-227.
9. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, et al. (2010) Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci U S A 107: 11971-11975.
10. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185: 6220-6223.
11. Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, et al. (2006) RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol 4: e3.
12. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. Nat Protoc 4: 470-483.
13. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, et al. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466: 334-338.
14. Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, et al. (2008) Viral diversity and dynamics in an infant gut. Res Microbiol 159: 367-373.
15. Jones MS, Kapoor A, Lukashov VV, Simmonds P, Hecht F, et al. (2005) New DNA viruses identified in patients with acute viral infection syndrome. J Virol 79: 8230-8236.
16. Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, et al. (2008) Metagenomic analysis of human diarrhea: viral detection and discovery. PLoS Pathog 4: e1000011.
17. Greninger AL, Runckel C, Chiu CY, Haggerty T, Parsonnet J, et al. (2009) The complete genome of klassevirus - a novel picornavirus in pediatric stool. Virol J 6: 82.
18. Schowalter RM, Pastrana DV, Pumphrey KA, Moyer AL, Buck CB (2010) Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. Cell Host Microbe 7: 509-515.
19. van der Meijden E, Janssens RW, Lauber C, Bouwes Bavinck JN, Gorbalenya AE, et al. (2010) Discovery of a new human polyomavirus associated with trichodysplasia spinulosa in an immunocompromized patient. PLoS Pathog 6: e1001024.
20. Finkbeiner SR, Li Y, Ruone S, Conrardy C, Gregoricus N, et al. (2009) Identification of a novel astrovirus (astrovirus VA1) associated with an outbreak of acute gastroenteritis. J Virol 83: 10836-10839.
21. Palacios G, Lovoll M, Tengs T, Hornig M, Hutchison S, et al. (2010) Heart and skeletal muscle inflammation of farmed salmon is associated with infection with a novel reovirus. PLoS One 5: e11487.
22. Quan PL, Firth C, Street C, Henriquez JA, Petrosov A, et al. (2010) Identification of a severe acute respiratory syndrome coronavirus-like virus in a leaf-nosed bat in Nigeria. MBio 1.
23. Ng TF, Wheeler E, Greig D, Waltzek TB, Gulland F, et al. (2011) Metagenomic identification of a novel anellovirus in Pacific harbor seal (Phoca vitulina richardsii) lung samples and its detection in samples from multiple years. J Gen Virol 92: 1318-1323.
24. Kapoor A, Simmonds P, Gerold G, Qaisar N, Jain K, et al. (2011) Characterization of a canine homolog of hepatitis C virus. Proc Natl Acad Sci U S A 108: 11608-11613.

25. Dunowska M, Biggs PJ, Zheng T, Perrott MR (2012) Identification of a novel nidovirus associated with a neurological disease of the Australian brushtail possum (Trichosurus vulpecula). Vet Microbiol 156: 418-424.
26. Thingstad EV (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnology and Oceanography 45: 1320-1328.
27. Fuhrman JA, Schwalbach M (2003) Viral influence on aquatic bacterial communities. Biol Bull 204: 192-195.
28. Winter C, Bouvier T, Weinbauer MG, Thingstad TF (2010) Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "killing the winner" hypothesis revisited. Microbiol Mol Biol Rev 74: 42-57.
29. Thingstad EV, Lignell R (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquatic Microbial Ecology 13: 19-27.
30. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, et al. (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7: 828-836.
31. Hendrix RW (1983) Lambda II. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory. 694 p.
32. Forde SE, Thompson JN, Bohannan BJ (2004) Adaptation varies through space and time in a coevolving host-parasitoid interaction. Nature 431: 841-844.
33. Gomez P, Buckling A (2011) Bacteria-phage antagonistic coevolution in soil. Science 332: 106-109.
34. Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11: 472-477.
35. Abedon ST, Hyman P, Thomas C (2003) Experimental examination of bacteriophage latent-period evolution as a response to bacterial availability. Appl Environ Microbiol 69: 7499-7506.
36. Lobocka MB, Rose DJ, Plunkett G, 3rd, Rusin M, Samojedny A, et al. (2004) Genome of bacteriophage P1. J Bacteriol 186: 7032-7068.
37. Jeong H, Barbe V, Lee CH, Vallenet D, Yu DS, et al. (2009) Genome sequences of Escherichia coli B strains REL606 and BL21(DE3). J Mol Biol 394: 644-652.
38. Betancor L, Yim L, Fookes M, Martinez A, Thomson NR, et al. (2009) Genomic and phenotypic variation in epidemic-spanning Salmonella enterica serovar Enteritidis isolates. BMC Microbiol 9: 237.
39. Lehours P, Vale FF, Bjursell MK, Melefors O, Advani R, et al. (2011) Genome sequencing reveals a phage in Helicobacter pylori. MBio 2.
40. Boyd EF (2012) Bacteriophage-encoded bacterial virulence factors and phage-pathogenicity island interactions. Adv Virus Res 82: 91-118.
41. Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, et al. (2010) Cryptic prophages help bacteria cope with adverse environments. Nat Commun 1: 147.
42. Sharon I, Alperovitch A, Rohwer F, Haynes M, Glaser F, et al. (2009) Photosystem I gene cassettes are present in marine virus genomes. Nature 461: 258-262.
43. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, et al. (2011) Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci U S A 108: E757-764.
44. Mazaheri Nezhad Fard R, Barton MD, Heuzenroeder MW (2011) Bacteriophage-mediated transduction of antibiotic resistance in enterococci. Lett Appl Microbiol 52: 559-564.

45. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, et al. (2011) Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480: 241-244.

46. Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, et al. (2002) Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. Science 295: 2091-2094.

47. Cummings CA, Bootsma HJ, Relman DA, Miller JF (2006) Species- and strain-specific control of a complex, flexible regulon by Bordetella BvgAS. J Bacteriol 188: 1775-1785.

48. Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, et al. (2004) Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature 431: 476-481.

49. Guo H, Tse LV, Barbalat R, Sivaamnuaiphorn S, Xu M, et al. (2008) Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification. Mol Cell 31: 813-823.

50. Dai W, Hodes A, Hui WH, Gingery M, Miller JF, et al. (2010) Three-dimensional structure of tropism-switching Bordetella bacteriophage. Proc Natl Acad Sci U S A 107: 4347-4352.

51. Bhaya D, Davison M, Barrangou R (2011) CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation. Annual Review Genetics, Vol 45 45: 273-297.

52. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, et al. (2006) The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. Nat Genet 38: 779-786.

53. Kliem M, Dreiseikelmann B (1989) The superimmunity gene sim of bacteriophage P1 causes superinfection exclusion. Virology 171: 350-355.

54. Hutchison CA, 3rd, Sinsheimer RL (1971) Requirement of protein synthesis for bacteriophage phi X174 superinfection exclusion. J Virol 8: 121-124.

CHAPTER 2 –INTER-INDIVIDUAL VARIATION AND DYNAMIC RESPONSE TO DIET

The contents of this chapter have been published as:

> Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu G, Lewis J, Bushman FD. (2011) The human gut virome: inter-individual variation and dynamic response to diet. Genome Research. 21: 1616-1625.

2.1 Abstract

Immense populations of viruses are present in the human gut and other body sites. Understanding the role of these populations (the human "virome") in health and disease requires much deeper understanding of their composition and dynamics in the face of environmental perturbation. Here we investigate viromes from human subjects on a controlled feeding regimen. Longitudinal fecal samples were analyzed by metagenomic sequencing of DNA from virus-like particles (VLP) and total microbial communities. Assembly of 336 Mb of VLP sequence yielded 7,175 contigs, many identifiable as complete or partial bacteriophage genomes. Contigs were rich in viral functions required in lytic and lysogenic growth, as well as unexpected functions such as viral CRISPR arrays and genes for antibiotic resistance. The largest source of variance among virome samples was inter-personal variation. Parallel deep sequencing analysis of bacterial populations showed covaration of the virome with the larger microbiome. The dietary intervention was associated with a change in the virome community to a new state, in which individuals on the same diet converged. Thus these data provide an overview of the composition of the human gut virome and associate virome structure with diet.

All sequence reads have been deposited in NCBI's Sequence Read Archive with the following accession numbers: SRX020379, SRX020378, SRX020505, SRX020504, and SRX020587.

2.2 Introduction

Bacteriophages are the most abundant biological entities on Earth, with an estimated population of ~10^{31} total particles [1,2], but their roles in human health are only beginning to be studied [3-5]. Phage model systems were pivotal in the early development of molecular biology [6,7]. Today much of phage research is focused on phage in their natural environments, including the viral component of the human microbiome [5,8-10]. The new emphasis on studies of whole populations has been made possible in part by the development of "next generation" sequencing methods, which allow quantification of the types and proportions of phages in complex mixtures by deep sequencing of environmental samples [11].

Lysogenic or temperate phages are able to integrate their chromosomes into the bacterial genome [12,13], and so can alter the phenotype of the host bacterium by lysogenic conversion [14]. Transduction of genes for toxins by phage is well known, as in the case of cholera [15] and Shiga toxin [14]. Additional functionality, identified more recently, may promote bacterial adaptation to the host environment--genes for functions involved in energy harvest [5] and platelet adhesion [10] have been identified in viral metagenomic data, and cryptic prophages of E. coli have been shown to encode genes for resistance to antibiotics and other environmental stresses [16]. The contributions of these and other phage genes to microbiome function are just beginning to be studied.

Diet is expected to alter the composition of the human microbiome, and specific microbiome assemblages in turn are expected to affect the welfare of the human host, but interactions between phage and diet in the human microbiome are mostly unexplored. One recent study used next generation sequencing to characterize human gut viruses from four twin pairs and their mothers [5] and found similarity of communities between twins and their mothers, and stability of viral communities over time. Dynamics in this study did not show cyclic changes in phage and bacterial abundance as would be expected for Lotka-Volterra predator-prey relationships [17], or episodes of outgrowth of particular bacterial species followed by blooms of their phage as in "kill-the-winner" dynamics [18]. The factors responsible for the observed longitudinal stability have not been fully clarified.

Here we present a study of the dynamics of the human gut virome during a deliberate perturbation by a dietary intervention. We compared shotgun metagenomic sequences from virome samples, as well as metagenomic sequences from bacterial populations. We found that the predominant source of variation was differences among individuals, but that significant changes in viral populations were detectable associated with switching to a defined diet and that convergence of viral populations was seen for individuals on similar diets.

2.3 Results and Discussion

Sampling and sequencing

We purified virus-like particles (VLPs) from stool samples collected longitudinally from 6 healthy volunteers between the ages of 18 and 40 years who had normal bowel frequency, normal body mass index, no history of chronic intestinal disease, diabetes, or immune deficiency, and who had not been treated with antibiotics for a minimum of 6 months prior to entering the study. Two individuals were fed a high-fat/low-fiber diet, three were fed a low-fat/high-fiber diet, and one was on an *ad-lib* diet. Samples were collected at up to four time points (days 1, 2, 7, and 8), with the controlled diet starting after sample collection on day 1. VLPs were purified (Fig. 2-1) by filtration and CsCl density gradient fractionation. In what follows, we use "VLP" to refer to these preparations. Although we are able to isolate multiple phage types from these preparations, and EM analysis confirms the presence of virus-like particles, the fraction of particles that are replication-competent virions is unknown, so we avoid referring to the full population as "viruses". VLPs were treated exhaustively with DNase, then deproteinized and total VLP DNA was purified [11]. The VLP-associated DNA was randomly amplified by Phi29 polymerase and shotgun sequenced using the Roche/454 GS FLX Titanium platform. Amplification with Phi29 polymerase can distort the ratios of different members of the community [19] but all samples studied here were processed similarly, allowing consistent comparisons between samples. After filtering, the VLP data set yielded 936,213 high-quality sequences with a mean length of 359nt (336Mb total). Initial analysis of individual reads showed that 98% of these sequences had no significant match to an identified sequence in the non-redundant database (E-value $< 10^{-5}$) when analyzed individually, consistent with previous studies of similar preparations [4,5,20].

To track bacterial populations, total DNA was isolated from the same stool samples and analyzed using deep sequencing of 16S rDNA amplicons and shotgun sequencing of total DNA [21]. The 16S rDNA sequence tag data set contained 63,405 reads, with a mean length of 268nt. Sequences were filtered using QIIME [22] and assigned to bacterial lineages using RDP [23]. Bacterial communities were compared using UniFrac [24]. Shotgun sequencing of total stool DNA (mostly from bacteria) yielded 1,007,534 reads with a mean length of 344nt.

To quantify the purity of our VLP preparations, we checked bacterial 16S sequences in the VLP DNA. VLP DNA preps were confirmed to be at least 10,000X reduced in bacterial 16S DNA by Q-PCR (data not shown), and VLP DNA samples contained only 21 reads with similarity to bacterial 16S, a 35-fold reduction compared to bacterial shotgun sequencing ($p<10^{-15}$, chi-square test). Bacterial sequences could be present in the VLP preparations as contamination, or as a result of generalized transduction, specialized transduction, or incorporation in Gene Transfer Agents (GTAs) [25,26]. Thus the origin of the low level bacterial sequences in our data set is uncertain.

Assembly and initial analysis

The average read length in our VLP data set was longer than in most previous metagenomic studies of viral DNA, allowing extensive assembly of individual reads into contigs. We assembled VLP sequences using the Newbler assembler [27] (40bp overlap, 90% identity), which yielded 7,175 contigs at least 500bp in length. Fully 86.6% of the sequence reads were recruited into these contigs (Fig. 2-2A). The longest contig was 46kb, 73 contigs were longer than 10kb, 279 contigs were longer than 5kb, and 3,028 contigs were longer than 1kb.

The approximate size of the VLP community can be estimated by PHACCS (PHAge Communities from Contig Spectrum) [28], which calculates the degree to which a group of sequences co-assemble into contigs, and compares them to simulated communities of different sizes. Similar to what has been seen recently in the human gut virome [5], the median richness (number of species) of the 16 samples was 44 (range = 19-785), and the average Shannon Diversity was 3.46 (s.d. = 0.59). The most abundant genotype was predicted to account for

16.20% of the total (s.d. = 2.09%), predicting the complete assembly of the most abundant VLP genomes in our study.

Analysis of gene content

To characterize these VLP contigs, we compared both nucleotide sequences and open reading frames (ORFs) to 1) the NCBI non-redundant database, 2) the Pfam database of conserved amino acid motifs [29], 3) the Clusters of Orthologous Groups (COG) database of annotated bacterial protein families [30], 4) A CLAssification of Mobile genetic Elements (ACLAME) [31], 5) the Antibiotic Resistance Genes Database (ARDB) [32], and 6) the Virulence Factors Database (VFDB) [33]. All VLP contigs and associated annotation can be viewed using a web-based interface at http://microb215.med.upenn.edu/cgi-bin/gbrowse/phage_metagenomics/. A total of 22% of ORFs contained recognizable Pfam motifs. Essential bacteriophage functions were well represented, including functions required for both lytic and lysogenic growth (Fig. 2-2B).

VLP contigs were classified according to their similarity to ICTV-defined bacteriophage families (Fig. 2-2C). There were 1,268 contigs with amino acid similarity to members of the Siphoviridae family (18% of the total), 686 (10%) to Myoviridae, 344 (4.8%) to Podoviridae, 68 (0.9%) to Microviridae, and 0.4% to other families. Of the remaining contigs, 813 (11%) had amino acid similarity to multiple bacteriophage families, and 3,969 (55%) did not have significant similarity to any bacteriophage families. Membership in these families was similar among the subjects studied (Fig. 2-2C). No strong candidates for eukaryotic viruses were detected either through nucleotide comparison to known eukaryotic viral genome sequences, or through similarity to conserved eukaryotic protein domains. In a few cases a gene was annotated as similar to a eukaryotic virus, but for each of these genes contigs could be identified that also encoded multiple phage genes, suggesting that these proteins contain motifs common to both bacterial and eukaryotic viruses. Thus this classification indicates that the viruses studied here were comprised primarily of tailed bacteriophage, with some representation of additional DNA phage families.

Analysis showed that nine VLP contigs formed closed circles, 4.5 – 6kb in length, suggestive of completion of these genome sequences. None of the nine had significant nucleotide similarity to any sequence in the NCBI non-redundant database (E-value < 10^{-3}). However, eight of the nine contigs contain an ORF that aligns to the capsid F protein sequence from Microphages [34], the family of 4.5 – 6kb circular ssDNA phage containing the proto-type bacteriophage ϕX174. Three of the eight contigs were closely related to the chp1-like Microphage. Those eight genomes also contain proteins similar to phage proteins involved in replication (n=4), proteolysis (n=2), and scaffolding assembly (n=2). Only two of these genomes have significant sequence similarity to each other--contigs c03390 and c04421 are 93% identical, but were too diverse to co-assemble. The ninth genome (contig c04570) contains proteins that resemble those found on plasmids from a wide range of bacteria, primarily *Firmicutes*, and may represent a temperate phage that maintains itself as a plasmid [35].

Comparison of total community and VLP metagenomes

We next investigated the proportion of phage DNA in our total (bacterial) metagenomic data set. We calculated the proportion of VLP sequences that have a significant match in the total shotgun data set (E-value < 10^{-5}), and vice versa (Fig. 2-3A). As expected [5], the VLP sequences represent a minority of the total DNA from stool, in the range of 4-17%, although this value is dependent on sampling depth, because the VLP community is a subset of the total community. We also assembled and annotated the total community shotgun sequence data set (347Mb) using COG [30], then compared it to the VLP dataset. The proportions of COG classes present in the total community (mostly bacteria) and VLPs were quite different, as expected (Fig. 2-3B). Bacteria were significantly enriched in genes for synthesis of amino acid and carbohydrate precursors, ion transport and metabolism, translational machinery, and cell wall/membrane biogenesis, while VLP contigs were enriched in genes for replication, recombination and repair, and unknown functions ($p<0.05$, t-test). Thus the profile indicates that viruses recruit host cell machinery for translation, energy production and synthesis of macromolecular precursors, while using their own coding capacity to encode functions for replication.

Prophage abundance

A key question in investigating the gut virome centers on what fraction of phages are temperate, since this group can install new genes in bacteria and alter phenotype via lysogenic conversion [5,14]. We used three indications of a temperate lifestyle to annotate the VLP contigs: 1) nucleotide identity to sequenced bacterial genomes, which is indicative of prophage formation (90% of bases aligned at 90% identity), 2) presence of integrase genes (according to Pfam [29] and COG [30] annotations), and 3) significant similarity of multiple proteins to prophages annotated in the ACLAME [31] database of mobile genetic elements (E-value < 10^{-5}). This strategy provides a minimal estimate of the number of temperate phages in our dataset, since authentic temperate phages may not be positive by any of these criteria. Of the 3,029 VLP contigs of at least 1kb in length, 428 (14%) had multiple proteins that significantly resembled an annotated prophage, 73 (2.4%) contained an integrase gene, and 37 (1.2%) aligned to a sequenced bacterial genome (Fig. 2-4A). Of the 505 contigs that fit at least one of these three criteria, 442

(88%) also contained an annotated bacteriophage gene. When individual VLP reads were mapped to known bacterial genomes, only 1.5% (13,808) aligned at 90% identity or higher. Of those reads, 74% map to just 5 genomes, *Bacteroides vulgatus* ATCC 8482 (n=4,555), *Eubacterium eligens* ATCC 27750 (n=2,147), *Faecalibacterium prausnitzii* L2/6 (n=2,081), *Parabacteroides distasonis* ATCC 8503 (n=785), and *Bacteroides thetaiotaomicron* VPI-5482 (n=583). These reads aligned to short, mostly contiguous regions of each genome and assembled into contigs that encode bacteriophage proteins (Fig. 2-4B,C), as expected for prophages. All of the reads matching *Eubacterium eligens* (n=2,147) mapped to its plasmid, which also contains annotated phage proteins, suggesting that this plasmid may contain an integrated prophage or itself be a non-integrating prophage, potentially analogous to phage P1 [35]. Of our 3,029 VLP contigs over 1kb, a minimum of 505 (or 17%), can be tentatively classified as temperate phages by at least one of the above criteria.

Variation among animals in prophage induction has been previously shown in mouse models [5], leading us to investigate induction here. We probed the level of induction of the above five prophages indirectly by comparing shotgun sequences from total community DNA (mostly bacterial), with VLP sequences (Fig. 2-S2). In only one case (*F. prausnitzii*) did the abundance of the bacterial host correlate with the abundance of its corresponding prophage. The lack of correlation or the other four prophages suggests the possibility that the degree of induction varied among the individuals studied.

We were able to infer the groups of bacteria infected by these putative temperate phage by comparing the VLP contigs to prophage sequences in the ACLAME database. Of the 2,814 contigs that had significant amino acid similarity (E-value < 10^{-5}) to an ACLAME prophage sequence, the contigs that were assigned exclusively to the *Firmicutes* accounted for fully 41% of the total (n=1,148) (Fig. 2-4D). Surprisingly, similarity to prophage from the other common gut resident, *Bacteroidetes*, accounted for less than 0.9% of the assigned contigs (n=24). While *Proteobacteria* account for only a modest fraction of the total gut bacteria, contigs with similarity to *Proteobacteria* accounted for 16% (n=447) of the total. This pattern also extended to contigs that were similar to multiple bacterial phyla ('Multiple' in Fig. 2-4D). Of the 1,176 VLP contigs that were similar to more than one bacterial Phylum, 34% had a match to *Firmicutes*, 0.8% had a match to *Bacteroides*, and 13% had a match to *Proteobacteria*. This suggests that the prevalence of temperate bacteriophage in the gut may differ among bacterial phyla, though the conclusions depend on representation of phage sequences in the databases used for analysis, and so conclusions may evolve as further sequence data accumulates.

Functionality in the gut virome: CRISPRs

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) are mediators of a recently described form of bacterial adaptive immunity. Short DNA sequences (26-72nt in length) are captured from invading DNA elements and installed as CRISPR spacer sequences in bacterial chromosomes. Multiple spacer sequences are separated by partially-palindromic repeats within CRISPR arrays. After transcription of CRISPR arrays, RNA copies of the CRISPR spacer sequences act as recognition modules in nucleoprotein complexes, which bind to targeted nucleic acid from genomic invaders such as phage or plasmids and program their degradation [36,37]. Thus CRISPR spacers provide a record of the genomic parasites that bacteria have encountered.

Analysis of CRISPRs in our assembled data provided a detailed record of phage-host and phage-phage competition. A total of 38 CRISPR arrays were found in shotgun metagenomic assemblies for our total community data set. The 38 contigs containing these CRISPRs aligned primarily to genomes of gut *Bacteriodetes* (n=16, E-value < 10^{-5}). Of the 45 spacer sequences found within those 38 CRISPR arrays, fully 80% (n=36) showed no matches to any known sequence, but one spacer perfectly matched a VLP contig sequence (contig c05189; 100% identity over 43 bases). Both the bacterially-encoded CRISPR spacer and VLP target were found within a single individual, suggesting that the CRISPR spacer sequence may have restricted phage replication within that subject.

Unexpectedly, we also found 22 CRISPR arrays in our VLP contigs. CRISPRs have not previously been reported from free phage particles, but one report did identify CRISPR sequences in potential prophages of *Clostridium difficile* [38]. A related CRISPR array was found in our virome samples (with >95% identity in repeat regions to the annotated *C. difficile* CRISPR5). Analysis of novel CRISPR spacer sequences showed one high stringency match between a VLP spacer (on contig c05834) and a separate VLP contig (c02690) (95% identity over 39 bases). The CRISPR spacer-target pair were isolated from the same individual and both were detected at the same two time points. Phage are well known to encode an extensive set of functions for competing with other phage [39,40]--these data indicate that CRISPRs may mediate phage-phage competition as well. It will be valuable to assess CRISPR evolution and targeting in further phage metagenomic data sets.

<u>Functionality in the gut virome: antibiotic resistance</u>

To interrogate antibiotic resistance in the VLP metagenomic data set, the ORFs found on contigs and unassembled reads were compared to known antibiotic resistance genes (BLASTp against the ARDB database [32]; E-value < 10^{-5}), yielding 614 matches to antibiotic resistance genes. These included multidrug efflux transporters (n=355), vancomycin resistance genes (n=129), tetracycline resistance genes (n=18), and beta-lactamases (n=16). The 33 highest quality matches to assembled contigs were examined further. Some of these encoded clear antibiotic resistance genes (e.g. beta-lactamase, drug efflux pumps, streptogramin acetyltransferase), while others encoded relatives of antibiotic resistance proteins that are of uncertain importance to resistance (e. g. ABC transporters and the VanRS two-component signaling system). Five of the contigs also contained identifiable phage genes. Transmissible antibiotic resistance is believed to involve primarily mobilization by plasmids and transposons [41,42]--our results indicate that the role of mobilization by phage may deserve further investigation.

Variation of the gut virome among individuals and during dietary intervention

We next asked how viromes varied among individuals, and how the dietary intervention affected virome structure. We characterized each virome sample by enumerating the proportion of sequence reads recruited into each VLP contig over the full contig data set. In Fig. 2-5A the proportional abundance of each VLP contig is shown by the color code. The Euclidean distance was then computed between all pairs of virome samples and used for statistical analysis. Comparison of the collection of within-subject distances among time points to between-subject distances showed that between-subject distances were significantly greater (Fig. 2-5B; p<0.001, significance assessed by permutation of sample labels). Thus each individual contained a unique virome that was globally stable over the 8 day time course. A previous study showed individual distinctiveness and stability of the human virome over a year [5].

In order to test the effect of diet on the composition of the gut virome, we compared the distance between VLP communities in individuals both before and after they started their controlled diet (Fig. 2-5C). The distance between the gut viromes of individuals on the same diet (Fig. 2-5C, left side) was significantly smaller at the end of their dietary treatment than it was at the start (p=0.05, significance assessed by permutation of diet labels). There was no increase over time in virome similarity for individuals on different diets (Fig. 2-5C, right side). A group of 39 VLP contigs were identified that changed in association with the dietary intervention. In comparison to the full data set of 7,175 contigs, this subset showed a trend toward enrichment in Siphoviridae (p<0.06) and depletion of Myoviridae (p<0.08), which is suggestive of a phylogenetically distinct diet-responsive bacteriophage population.

This dietary effect was also seen in a subject-by-subject analysis. For each subject, the distances between samples taken before the defined diet were compared, and distances between samples taken after defined diet were compared (Fig. 2-5D). As a contrast, distances between samples where one was before-diet and the other was after-diet were also compared. The distances were greatest for comparisons between before-diet and after-diet samples (Fig. 2-5D; p<0.05, significance assessed by permutation of day labels). Thus over the period of the dietary intervention, the viral community changed detectably to a new state.

The gut virome is dominated by prokaryotic viruses which prey on gut bacteria, raising the question of how changes in bacterial and viral communities are linked. We tested covariation by comparing 16S rDNA amplicon data from the gut DNA samples, which measured bacterial diversity, with the VLP shotgun metagenomic data, which measured viral diversity (Fig. 2-5E). Variation was significantly correlated between bacterial and VLP communities (Mantel test, R=0.40, p<0.001). This correlation was significant when tested by subject label permutation (p<0.001), indicating that inter-personal variation is correlated between these communities.

Conclusions

In summary, the VLP sequence data presented here could be assembled into contigs that recruited fully 86.6% of sequence reads, providing a resource of over 7,000 complete and partial phage genomes and making possible extensive interpretation of ORF function. A substantial portion of the phages in our samples are likely temperate, so that genes contained within phage may alter phenotype of the bacterial host by lysogenic conversion. Novel functionalities emphasized here include phage-encoded CRISPR arrays and antibiotic resistance genes. Also seen were the expected diversity of lysins [43], holins [44], bacteriocins [45], restriction/modification systems [46] and virulence factors [47-49].

The major determinants of microbiome community composition and dynamics remain to be fully clarified. Bacteriophage and bacterial abundances did not oscillate detectably over the time period studied, as would be predicted by Lotka-Volterra predator-prey dynamics, nor did we detect boom and bust outgrowth of phage-host pairs, indicative of kill-the-winner dynamics, consistent with previous work [5,50]. A number of factors may confound the detection of competition in natural populations. The rate at which ‘winning’ bacterial clones [18] arise and are preyed upon may be too fast or too slow to be apparent in this set of samples. The sample set may not be dense enough to detect rapid changes. In addition, phage types that infect different bacterial hosts may be indistinguishable at the current depth of sequencing and with available annotation [50]. Viromes were relatively stable within each individual, and interpersonal variation was the largest source of variance observed even when individuals were on the same diet. This allows us to rule out the idea that phage populations are predominantly acquired on a daily time scale as transients in food, because individuals eating the same food did not come to harbor identical viromes. However, the gut virome changed significantly during the change in diet by alteration of the proportions of pre-existing populations, so that subjects on the same diet showed more similar, though not identical, virome composition. Whether the changes in phage abundance are simply a result of changes in abundance of their hosts, or whether additional mechanisms are involved will require further work to clarify, though initial data suggests a possible contribution of lysogenic induction. It will be valuable going forward to develop improved methods for quantifying phage induction in vivo. Considerable further study will be required to understand the acquisition of gut viral communities and the factors mediating the balance between long term stability and dynamic response to the environment.

2.4 Methods

Sample collection

Six healthy adult volunteers (at least 18 years old) were recruited to provide stool samples within the Center for Clinical and Translational Research at the Hospital of the University of Pennsylvania. Exclusion criteria included having had diarrhea within one week prior to the sample collection, abnormal bowel movement frequency (at least once every 2 days and no more than 3 times a day), consumption of any antibiotics within six months prior to sample collection, or any prior diagnosis with inflammatory bowel disease, irritable bowel syndrome, celiac sprue, or other chronic inflammatory diseases of the intestines, or BMI outside the range of 18.5 and 35. All collection was carried out after subjects provided informed consent under an approved IRB protocol.

DNA isolation, PCR amplification, and purification

VLP DNA was isolated in the following manner (Fig. 2-1). Approximately 0.5g of stool was homogenized in 40mL SM buffer [51]. Particulate matter was spun down at 4700g for 30min and supernatant was filtered at 0.22μm (PES filter, Nalgene, Rochester, NY). The filtrate was centrifuged on a CsCl step gradient (described in detail [11,51]), and the 1.35-1.5g/mL fraction was extracted from the column. We note that some viruses are known to have densities outside this range (e. g. Tectiviridae, Poxviridae, Herpesviridae), and so would be lost during purification [11]; no attempt was made to isolate RNA viruses. Samples were treated with chloroform for 10min, treated with DNase (Invitrogen, Calsebad, CA) for 10min at 37C, and then extracted with the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA). VLP DNA was amplified with Genomiphi V2 polymerase (GE Healthcare, Piscataway, NJ).

Total stool DNA was isolated using the QIAamp DNA Stool Mini Kit (Qiagen, Valencia, CA), as described previously [52].

454/Roche sequencing methods

Both total DNA and amplified VLP DNA were randomly sheared and ligated to adapters using the 454 GS Titanium Rapid Library Preparation Kit with MID adaptors (454, Branford, CT). Bacterial 16S amplicons were amplified with primers BSR 357-A (barcoded) and BSF8-B (not barcoded), which anneal to the V1-V2 region of the bacterial 16S rRNA gene, and include 454 sequencing adaptors [52]. Samples were pooled and sequenced using GS FLX Titanium chemistry on a 454 Genome Sequencer [53].

16S sequence analysis

Bacterial 16S sequences were analyzed using Pyronoise, assigned by comparison to the RDP database [23], and communities compared to each other using UniFrac [24] scores, all within the Qiime software package [22]. Sequences were filtered according to size (between 200nt and 800nt), ambiguous bases were removed (max=2), homopolymer runs were truncated (max=20), and primer mismatches removed (max=0).

VLP contig sequence analysis

Quality Control: 454 Pyrosequencing yielded a total of 1,052,246 VLP sequences. These reads were filtered for quality using an in-house pipeline. This pipeline removed sequences sequentially due to 1) incorrect barcodes (n=23,290), 2) length less than 50nt (n=16,939), and 3) multiple ambiguous bases (n=19,708). For removing the human reads from the data, the sequences were aligned to the human genome using BLAT [54]. A total of 476 reads (approximately .05% of the data) had greater than 70% similarity to the human genome and were removed. QIIME [22] was used to remove duplicate reads from the data set in the following manner. First, 5 bases were trimmed off from 3' end of each of the sequences. The trimmed reads were then used to form OTUs using the TRIE algorithm at 97% identity. A total of 55,620 duplicate reads were discarded and the final data set of 936,213 sequences (with the final 5 bases preserved) was used for further analysis. The size distribution and quality scores of these high-quality reads are shown in Figure 2-S1.

Assembly: The 454 Newbler Assembler [27] was used to *de novo* assemble VLP sequences. Default parameters were used to obtain contigs, with minimum overlap length of 40bp and minimum overlap identity of 90%. One sequence was only allowed to be assigned to one contig. The Newbler assembler uses overlap-layout-consensus method to assemble sequences where overlaps are computed by pair-wise sequence alignments and then multiple sequence alignment is used to derive the consensus sequence. Similarity between VLP contigs was assessed using YASS [55], a DNA local alignment tool.

Analysis of VLP genomes and visualization: The contigs generated from assembly were compared (using BLASTn) to the NCBI non-redundant nucleotide database, ACLAME (A CLAssification of Mobile genetic Elements) [31] and VFDB (Virulence Factors Database) [33]. In-house scripts were used to derive ORFs from contigs by translating the reads in all 6 reading frames using NCBI's Bacterial, Archaeal and Plant Plastid Code. Only ORFs at least 100aa were considered in further analysis. The ORFs were compared (BLASTp) to the NCBI non-redundant amino acid sequence database, Pfam [29], COG [30], Antibiotic Resistance Genes Database (ARDB) [32], ACLAME protein and VFDB protein databases.

The Generic Genome Browser (GBrowse) [56] was configured on the local server (http://microb215.med.upenn.edu/cgi-bin/gbrowse/phage_metagenomics/) to visualize assembled VLP contigs and singletons. MySql adaptor was used to set up the databases. The "Phage metagenomics" database contains the contigs, ORFs and hits obtained by querying the reads using BLAST. Another database, "Phage Circular Genomes" was created to view the 9 complete circular phage genomes, the ORFs from the genomes and BLASTn/BLASTp/BLAST-rps hits from the databases listed above. The contigs, ORFs and the BLAST results information were converted to the (GBrowse-compatible) gff3 file format using in-house scripts which were then configured as MySql tables to view in GBrowse.

Taxonomic classification: Contigs were assigned to different bacteriophage families according to their amino acid similarity to the genomes of ICTV-defined groups. To be placed in a certain group, a contig must have amino acid similarity (E-value < 10^{-3}) to a protein from the genome of a member of that group. Contigs were excluded that had a single ORF that was similar to members of multiple families (cutoff at 90% of the top score), or had multiple ORFs that were similar to different families. Samples were characterized according to their membership in these taxonomic families by counting the number of reads from each sample that were assembled into a contig with a given classification.

CRISPR identification and spacer similarity: CRISPR arrays were identified in both the VLP contig and total DNA contig data sets using the CRISPRfinder utility [57]. Both direct repeat and spacer sequences were compared to previously sequenced CRISPR spacers and direct repeats (from CRISPRdb [58]) using blat (90% aligned at 90% identity), and to the non-redundant database (from NCBI) using BLASTn (90% aligned at 90% identity). Spacers from CRISPRdb [58] were also compared to the VLP contig data set using blat (90% aligned at 90% identity).

Functional comparison of VLP DNA and total DNA

Biochemical functionality was predicted according to similarity of ORFs to proteins within the COG database [30]. Membership in a given COG category was determined by the proportion of reads (out of a given data set) that fell within ORFs that had significant similarity (E-value < 10^{-5}) to a protein in the COG database in that category. Differences between VLP and total DNA were calculated via two-sample t test.

PHACCS Analysis

Community structure was estimated by using PHACCS (v1.1.2) [28], implemented with the Octave (v3.2) environment, using contig spectra as input that were generated by Circonspect (v0.2.4). The contig spectra were generated by randomly sampling 10,000 sequences truncated to 100bp in length, assembling at 98% identity and 35bp overlap, and counting the number of contigs with 1 member, 2 members, etc. PHACCS was run on those spectra using a genome size of 50kb, under a power law scenario.

Statistical methods

Euclidean distances were calculated by the R function *dist*. This function takes vectors as inputs, where each element was the number of sequences from a given sample that assembled into a given contig.

To assess the significance of diet effect and association between two distance matrices, day labels were permuted for samples from the same subject, the test statistic (t-statistic and Pearson correlation, respectively) was recalculated using the permuted distance matrix and compared to the observed test statistic to yield P-values. To visualize this covariation, we used Procrustes (part of the Qiime software package [22]), which rotates two distance matrices to maximize superimposition, and then plotted using KiNG 2.16. Significance was called at a p value of 0.05 or below, though we note it would be helpful to have more samples and use a lower cut value. To assess the difference in taxonomic categorization within the contigs that were accountable for the effect of diet (above), we took a permutation approach. The complete set of contigs was randomly sampled 10,000 times, and that was compared to the observed distribution of the diet-associated contigs. The probability of obtaining the observed proportions was estimated by the number of permutations that had a proportion that was no less extreme.

2.5 Figures

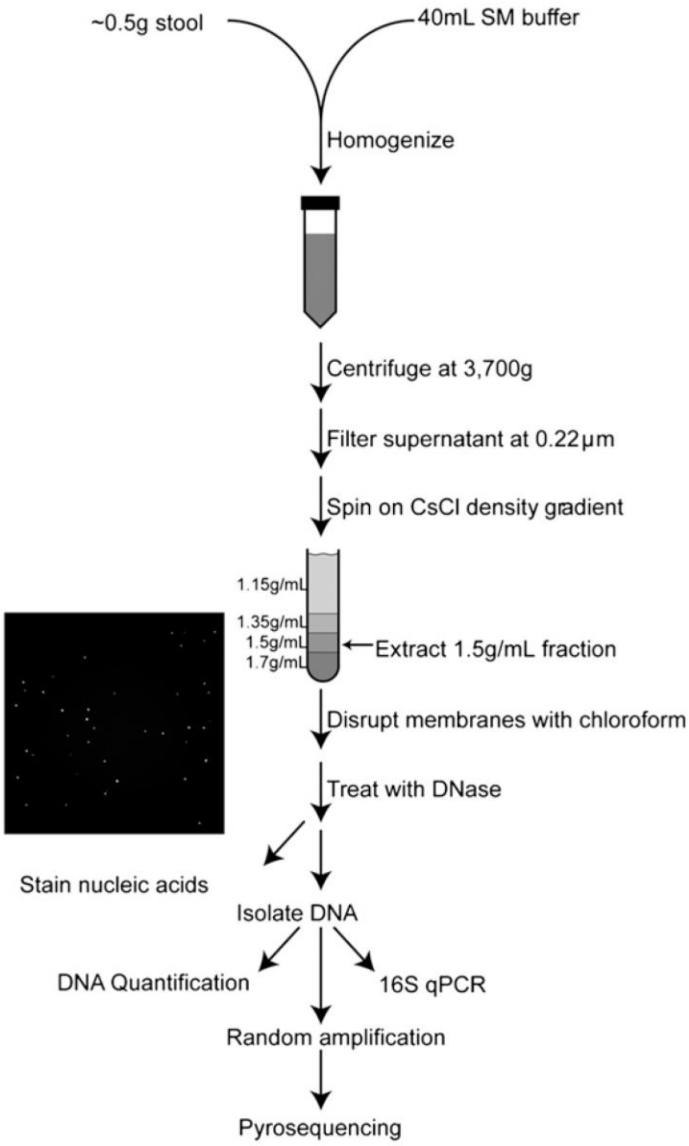

Figure 2-1. Purification of VLP DNA. Purification scheme was adapted from [11]. Stool was homogenized in SM Buffer; particulate matter was spun down; supernatant was filtered at 0.22µm to remove cells; VLPs were purified on a CsCl density gradient and treated with nuclease to eliminate unprotected DNA. The absence of bacterial cells was confirmed by staining VLP preparations for nucleic acids. VLP DNA was quantified, amplified, and pyrosequenced.

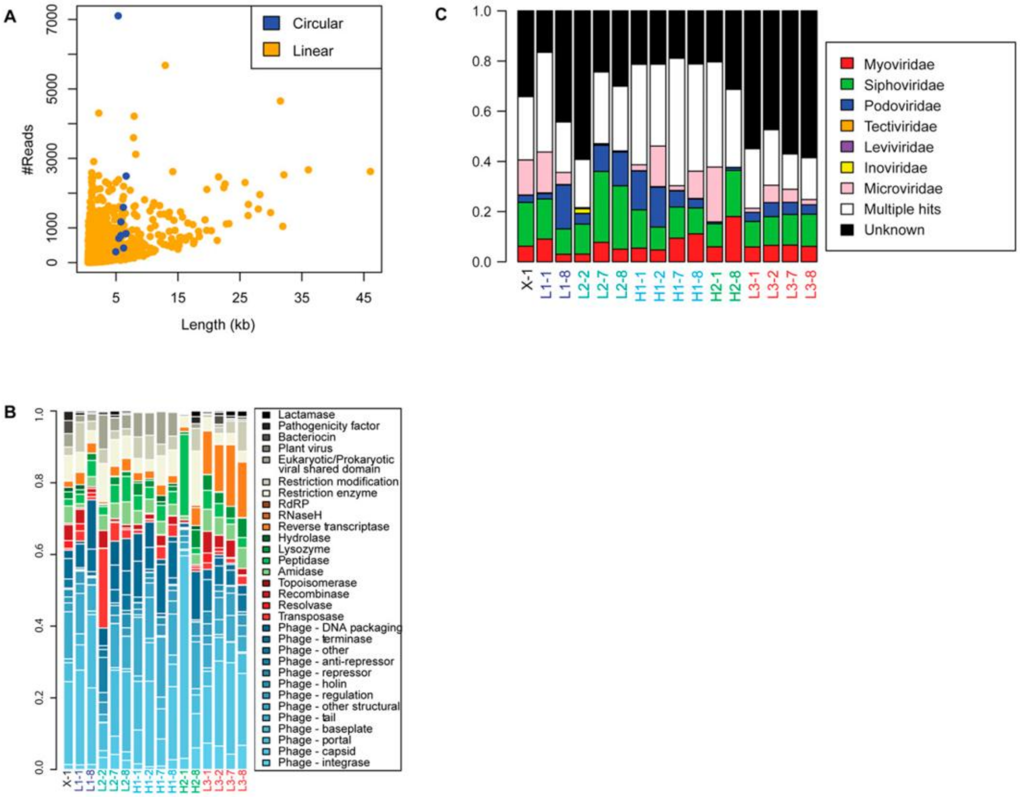

Figure 2-2. Assembly and functional annotation of shotgun metagenomic sequences from the human gut virome. (A) Analysis of recruitment of VLP sequence reads into contigs. The y-axis shows the number of sequence reads, the x-axis shows contig length. Pyrosequencing data from the human virome were assembled into 7,147 contigs up to 47.8kb in length. Linear contigs are shown in yellow, circular contigs are shown in blue. (B) Analysis of protein functions in VLP contigs. The functions encoded in VLP contigs were predicted using the Pfam database, then grouped using a custom database relating Pfam domain identifiers to phage functions (Table S4). The relative proportions of pyrosequencing reads falling within ORFs of different annotations were plotted according to their sample of origin (y-axis). Each bar is indicated by the sample code, where L or H indicates low or high fat diet, the adjacent number indicates the subject number, and the number after the hyphen indicates the day of the study. "X-1" indicates *ad-lib* diet. C) Taxonomic classification of VLP communities is consistent across samples. Samples (in columns, labeled as in Figs 1 and 5) are characterized according to the number of sequences from each sample that are assembled into a contig that is classified by taxonomic family. Phage families are indicated by the color code to the right. 'Unknown' (black) indicates contigs that cannot be classified in any way. 'Multiple hits' (white) indicates contigs that have proteins that are similar to multiple families.

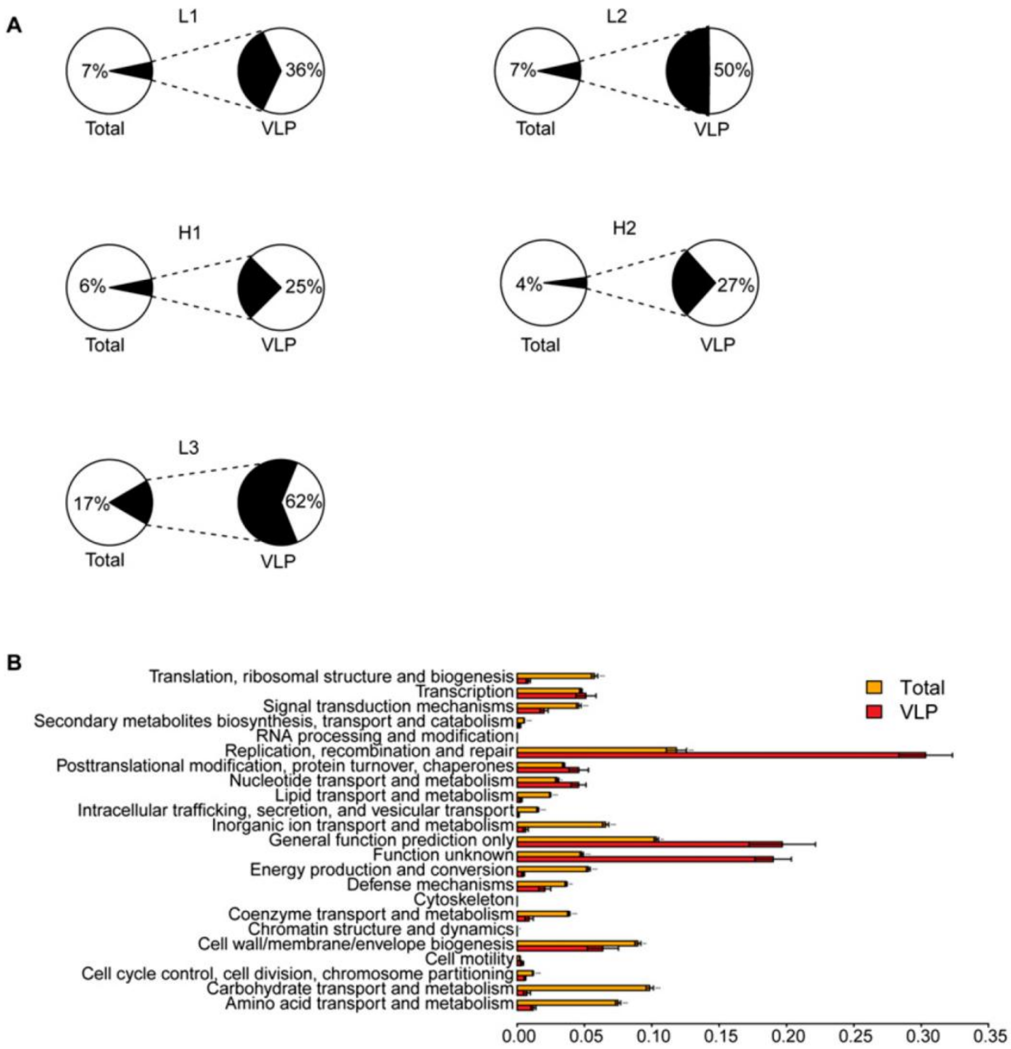

Figure 2-3. Comparison of gene content in total microbial communities and VLP communities. (A) The proportions of total genes identified in shotgun metagenomic analysis of total stool DNA (left) is compared to genes in VLP communities (right). (B) VLP DNA encodes biochemical functionality that is markedly different from the total microbiome. Function annotation was performed according to comparison to the COG database [30]. The proportion of reads from each assembled data set that fall within an ORF of the indicated annotation are plotted on the x-axis (mean±standard error). Asterisks indicate significant differences between VLP and microbial communities by two-sample t-test (one for p<0.05, two for p<0.01, and three for p<0.001).

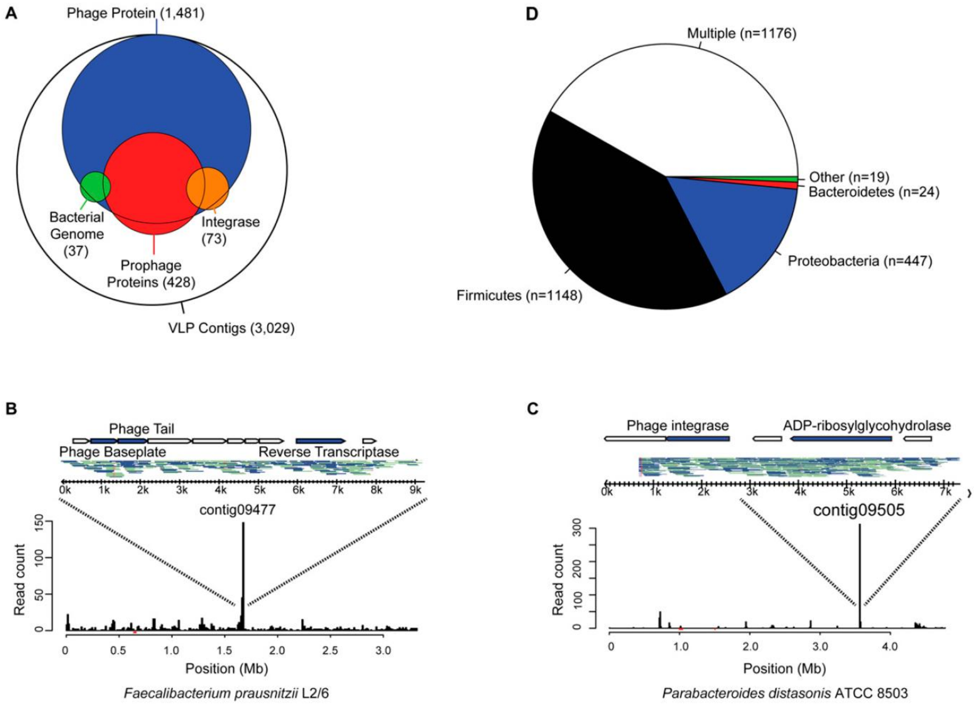

Figure 2-4. Analysis of temperate bacteriophages in the human gut virome. (A) Venn diagram indicating frequency of functions associated with temperate phages. VLP contigs were annotated according to the presence of integrase-like sequences (orange), BLAST alignment to sequenced bacterial genomes, suggestive of prophage formation (90% length at 90% identity; green), and presence of multiple genes with significant similarity to a prophage element within the ACLAME database (red). Map of VLP pyrosequencing reads aligning to the genomes of (B) *Faecalibacterium prausnitzii* L2/6 and (C) *Parabacteroides distasonis* ATCC 8503 (90% length at 90% identity). Inset are the contigs that correspond to each peak, the reads that make up each contig, and their annotated genes. Top strand reads are indicated by blue, bottom strand reads by grey, and mismatches are indicated by vertical red lines. (D) Prevalence of prophage sequences according to bacterial phylum of origin. Contigs with amino acid similarity (E-value < 10^{-5}) to prophage sequences within the ACLAME databases are shown according to which bacterial phylum they match. Any contig with similarity to more than one phylum is classified as 'Multiple.' Roughly 39% of VLP contigs (n=2,814) did not have significant similarity to any prophage sequence in this database.

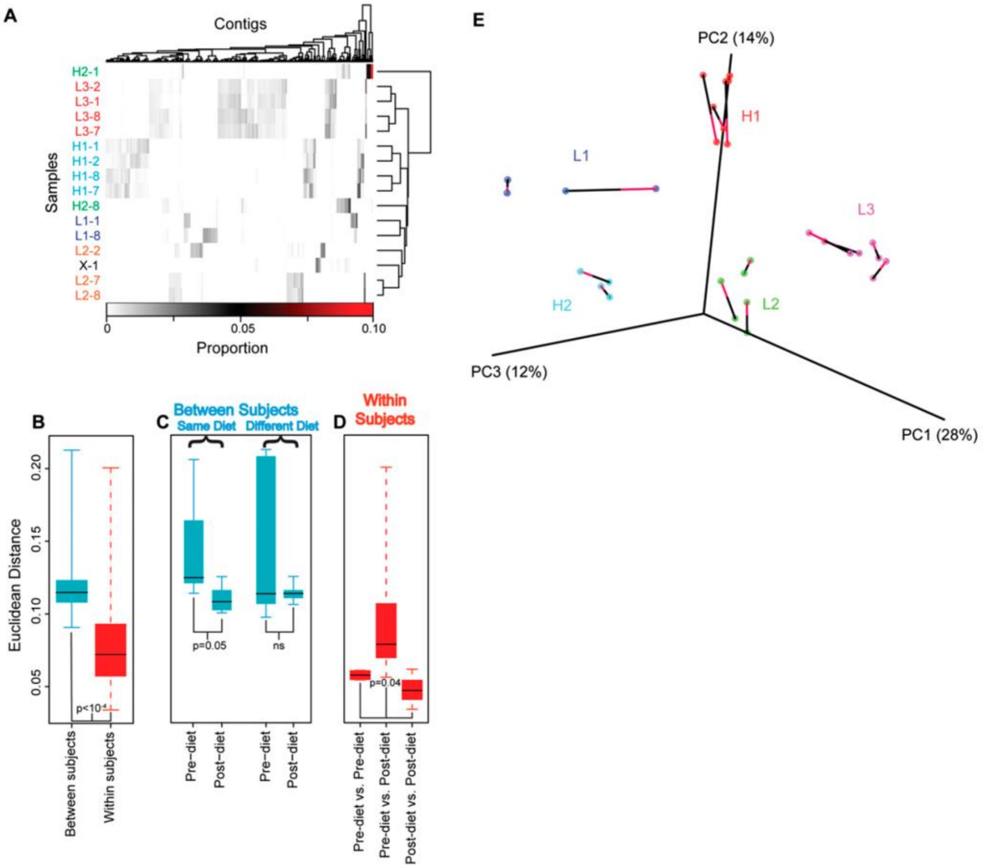

Figure 2-5: Alterations in VLP contig abundance associated with diet. (A) Proportions of VLP sequence reads in contigs from different subjects and time points. Vertical bars indicate the proportion of reads within each contig. The contigs are shown in columns, and subject/time point combinations in rows. Hierarchical clustering was performed on both rows and columns according to Euclidean distance and complete distance agglomeration. Samples are labeled by subject according to diet: high-fat (H1, H2), low-fat (L1, L2, L3), and ad-lib (X), as well as day of dietary intervention (days 1, 2, 7, or 8). The proportion of all VLP reads in each contig are shown by the scale at the bottom. (B, C, D) The Euclidean distance between samples is shown according to median (line), quartile (box), and range (whisker). (B) Between-subject variation is significantly greater than within-subject variation ($p<10^{-4}$, subject label permutation). (C) The distance between the gut viromes of individuals on the same diet (left) was significantly smaller at the end of their dietary treatment than it was at the start (p=0.05, permutation of diet labels), while there was no increase in similarity for individuals on different diets (right). (D) The distances within subjects that was measured between samples taken on the first two days of the timecourse (left) and the last two days of the timecourse (right) were significantly lower than the distances between those two sets of days (middle; p=0.04, permutation of day labels). (E) Covariation of bacterial and VLP community diversity. Distances between pairs of bacterial and VLP communities were calculated as described in Materials and Methods. The similarity of bacterial and VLP communities is shown through PCoA analysis, where each data set was rotated and scaled for maximum superimposition. Each circle represents a sample, either bacterial or

VLP. The bacterial and VLP communities from the same sample are connected by a line, where the red half of the line touches the VLP community, and the black half touches the bacterial community. The percent of total variation accounted for by each axis is shown in the axis label. The alignment of bacterial and VLP communities was highly significant (p = 0.004).

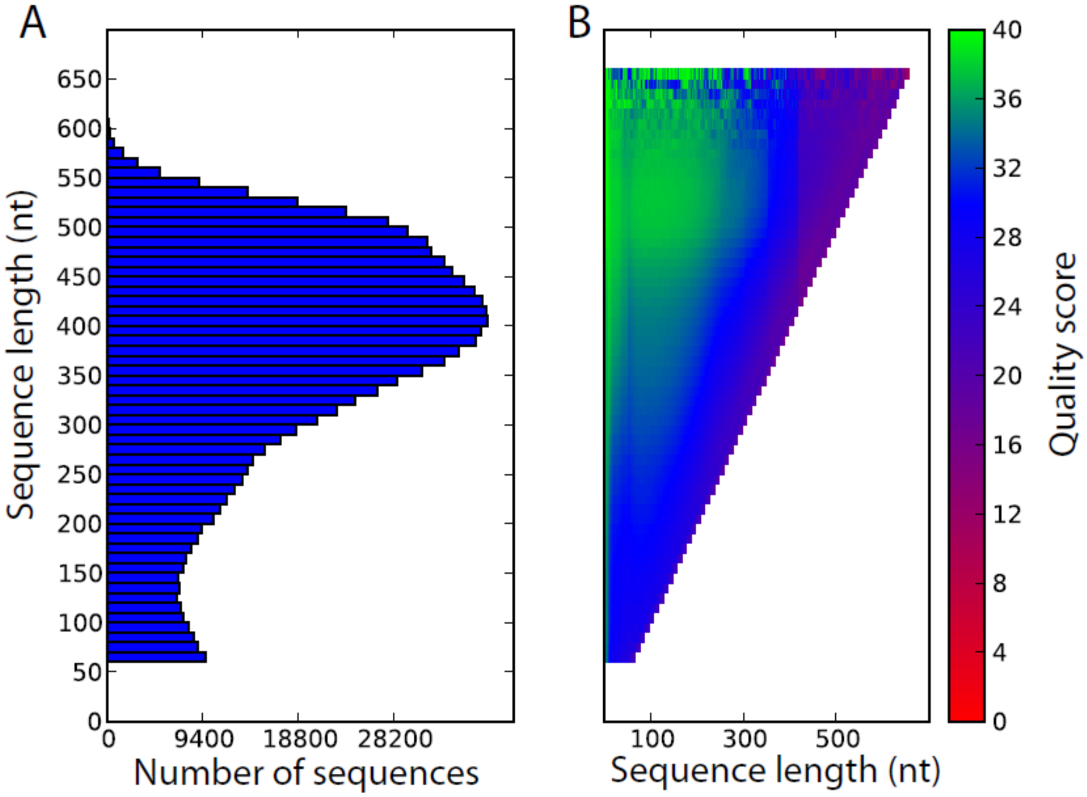

Figure 2-S1: Quality scores of VLP DNA sequence reads used in shotgun metagenomic analysis. (A) Histogram of read length by 10nt bin. The majority of sequences fall within the 300-500nt range. (B) Heatmap of sequence quality scores binned by sequence length (y-axis) and sequence position (x-axis). The average quality score (454 Q20 score) at a certain position is given by color (indicated in legend). The bins on the vertical axis from (A) are the same as in (B).

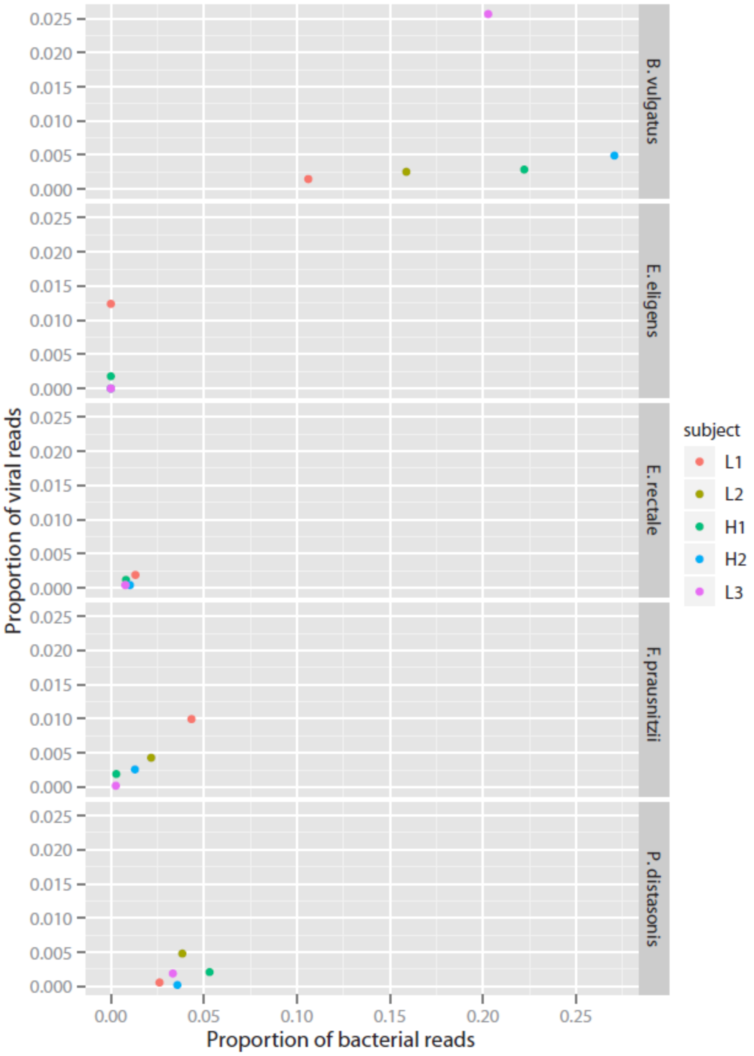

Figure 2-S2: Comparison of reads from five common bacterial genomes among complete metagenomic and VLP data sets. The proportion of reads are shown from total DNA shotgun sequences (x-axis) and VLP sequences (y-axis) that have significant similarity (E-value $< 10^{-5}$) to the indicated genomes. Reads are pooled by subject (as in Figure 3), shown in color.

2.6 References

1. Suttle CA (2005) Viruses in the sea. Nature 437: 356-361.
2. Weinbauer MG (2004) Ecology of prokaryotic viruses. FEMS Microbiol Rev 28: 127-181.
3. Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3: 504-510.
4. Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, et al. (2008) Viral diversity and dynamics in an infant gut. Res Microbiol 159: 367-373.
5. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, et al. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466: 334-338.
6. Cairns J, Stent GS, Watson JD (1966) Phage and the Origins of Molecular Biology. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
7. Judson OP, Normark BB (2000) Evolutionary genetics. Sinless originals. Science 288: 1185-1186.
8. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. (2008) Functional metagenomic profiling of nine biomes. Nature 452: 629-632.
9. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, et al. (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS One 4: e7370.
10. Willner D, Furlan M, Schmieder R, Grasis JA, Pride DT, et al. (2010) Microbes and Health Sackler Colloquium: Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. Proc Natl Acad Sci U S A.
11. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. Nat Protoc 4: 470-483.
12. Hendrix RW, Roberts JW, Stahl FW, Weisberg RA (1983) Lambda II. Cold Spring Harbor: Cold Spring Harbor Laboratory.
13. Ptashne M (1992) A Genetic Switch. Cambridge, MA: Cell Press and Blackwell Scientific Publications.
14. Brussow H, Canchaya C, Hardt WD (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 68: 560-602.
15. Waldor MK, Mekalanos JJ (1996) Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. Science 272: 1910.
16. Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, et al. (2010) Cryptic prophages help bacteria cope with adverse environments. Nat Commun 1: 147.
17. Bohannan BJM, Lenski RE (1997) Effect of resource enrichment on a chemostat community of bacteria and bacteriophage. Ecology 78: 2303-2315.
18. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, et al. (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7: 828-836.
19. Yilmaz S, Allgaier M, Hugenholtz P (2010) Multiple displacement amplification compromises quantitative analysis of metagenomes. Nat Methods 7: 943-944.
20. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185: 6220-6223.
21. Hoffmann C, Hill DA, Minkah N, Kirn T, Troy A, et al. (2009) Community-wide response of gut microbiota to enteropathogenic Citrobacter infection revealed by deep sequencing. Infection and immunity.

22. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7: 335-336.
23. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73: 5261-5267.
24. Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71: 8228-8235.
25. Bushman FD (2001) Lateral DNA Transfer: Mechanisms and Consequences. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
26. McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, et al. (2010) High frequency of horizontal gene transfer in the oceans. Science 330: 50.
27. Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95: 315-327.
28. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, et al. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6: 41.
29. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic acids research 26: 320-322.
30. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.
31. Leplae R, Lima-Mendez G, Toussaint A (2009) ACLAME: a CLAssification of Mobile genetic Elements, update 2010. Nucleic Acids Res 38: D57-61.
32. Liu B, Pop M (2009) ARDB--Antibiotic Resistance Genes Database. Nucleic Acids Res 37: D443-447.
33. Yang J, Chen L, Sun L, Yu J, Jin Q (2008) VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Res 36: D539-542.
34. Rohwer F, Edwards R (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol 184: 4529-4535.
35. Sternberg N, Austin S (1981) The maintenance of the P1 plasmid prophage. Plasmid 5: 20-31.
36. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, et al. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321: 960-964.
37. Sorek R, Kunin V, Hugenholtz P (2008) CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol 6: 181-186.
38. Sebaihia M (2006) The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. Nature Genetics 38: 779-786.
39. Calendar R (1988) The Bacteriophages; Fraenkel-Conrat F, Wagner RR, editors. New York: Plenum Press.
40. Refardt D (2011) Within-host competition determines reproductive success of temperate bacteriophages. ISME J.
41. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, et al. (2009) Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev 33: 376-393.
42. Barlow M (2009) What antimicrobial resistance has taught us about horizontal gene transfer. Methods Mol Biol 532: 397-411.

43. Seo HS, Xiong YQ, Mitchell J, Seepersaud R, Bayer AS, et al. (2010) Bacteriophage lysin mediates the binding of streptococcus mitis to human platelets through interaction with fibrinogen. PLoS Pathog 6.
44. Wang IN, Smith DL, Young R (2000) Holins: the protein clocks of bacteriophage infections. Annu Rev Microbiol 54: 799-825.
45. Dawid S, Roche AM, Weiser JN (2007) The blp bacteriocins of Streptococcus pneumoniae mediate intraspecies competition both in vitro and in vivo. Infection and immunity 75: 443-451.
46. Kobayashi I (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. Nucleic Acids Res 29: 3742-3756.
47. Campos J, Martinez E, Izquierdo Y, Fando R (2010) VEJ{phi}, a novel filamentous phage of Vibrio cholerae able to transduce the cholera toxin genes. Microbiology 156: 108-115.
48. O'Brien AD, Newland JW, Miller SF, Holmes RK, Smith HW, et al. (1984) Shiga-like toxin-converting phages from Escherichia coli strains that cause hemorrhagic colitis or infantile diarrhea. Science 226: 694-696.
49. Lainhart W, Stolfa G, Koudelka GB (2009) Shiga toxin as a bacterial defense against a eukaryotic predator, Tetrahymena thermophila. J Bacteriol 191: 5116-5122.
50. Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, et al. (2010) Viral and microbial community dynamics in four aquatic environments. ISME J 4: 739-751.
51. Sambrook J, Russell DW (2001) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
52. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, et al. (2010) Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. BMC Microbiol 10: 206.
53. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.
54. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664.
55. Noe L, Kucherov G (2005) YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res 33: W540-543.
56. Podicheti R, Gollapudi R, Dong Q (2009) WebGBrowse--a web server for GBrowse. Bioinformatics 25: 1550-1551.
57. Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35: W52-57.
58. Grissa I, Vergnaud G, Pourcel C (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 8: 172.

CHAPTER 3 – HYPERVARIABLE LOCI IN THE HUMAN GUT VIROME

The contents of this chapter have been published as:

> Minot S, Grunberg S, Wu G, Lewis J, Bushman FD. Hypervariable loci in the human gut virome. PNAS (2012) 109 (10): 3962-3966

3.1 Abstract

Genetic variation is critical in microbial immune evasion and drug resistance, but variation has rarely been studied in complex heterogeneous communities such as the human microbiome. To begin to study natural variation, we analyzed DNA viruses present in the lower gastrointestinal tract of 12 human volunteers by determining 48 billion bases of viral DNA sequence. Viral genomes mostly showed low variation, but 51 loci of ~100 bp showed extremely high variation, so that up to 96% of the viral genomes encoded unique amino acid sequences. Some hotspots of hypervariation were in genes homologous to the bacteriophage BPP-1 viral tail fiber gene, which is known to be hypermutagenized by a unique reverse-transcriptase (RT)-based mechanism. Unexpectedly, other hypervariable loci in our data were in novel gene types, including genes encoding predicted Ig-superfamily proteins. Most of the hypervariable loci were linked to genes encoding RTs of a single clade, which we find is the most abundant clade among gut viruses but only a minor component of bacterial RT populations. Hypervariation was targeted to 5'-AAY-3' asparagine codons, which allows maximal chemical diversification of the encoded amino acids while avoiding formation of stop codons. These findings document widespread targeted hypervariation in the human gut virome, identify new types of genes targeted for hypervariation, clarify association with RT gene clades, and motivate studies of hypervariation in the full human microbiome.

3.2 Introduction

Key aspects of host-parasite interactions are mediated by targeted changes in DNA. The vertebrate adaptive immune system is based on covalent DNA rearrangements that diversify genes encoding Ig-domain antigen binding proteins. In response, viral and cellular pathogens encode genetic systems that vary antigens bound by host antigen receptors [1,2].

In this study we begin to characterize patterns of sequence variation in heterogeneous natural communities, using the human microbiome as a model. We chose to study viral samples because they represent a medically important microbiome component, but contain a smaller aggregate genome size than the full microbiome, allowing sequencing to a depth that permits empirical assessment of variation.

A newly discovered mechanism of targeted hypermutation, particularly pertinent here, involves the *Bordetella* bacteriophage BPP-1, which has been shown to vary the sequence of the gene encoding its phage tail fiber to bind divergent cell surface receptors [3-6]. The phage-encoded major tropism determinant (MTD) gene, which encodes the tip of the tail fiber, is subjected to targeted hypermutation by a reverse transcriptase (RT)-dependent mechanism [7,8]. The 3' part of the tail fiber gene is duplicated in the phage genome, and the duplicated template repeat (TR) is transcribed and reverse transcribed in an error-prone fashion. The mutated copy is then incorporated into the MTD gene variable repeat (VR), leading to very high mutation rates. Diversity generating systems involving related RTs and genes encoding C-type lectin folds have been inferred from prokaryotic genome sequences [3,4], but only the BPP-1 system has been characterized functionally.

Here we have used the Solexa/Illumina HiSeq method to interrogate 48 billion bases of DNA sequence from populations of gut DNA viruses, which allowed us to identify regions of targeted hypervariation in the primary sequence data. We found that RT-associated hypervariation systems were present in 11 out of 12 subjects examined, and act on a much wider range of gene types than was known previously. Analysis of the sequence information further specifies the chemical logic of the mutational targeting, and suggests that the most common role of RTs in the gut virome is targeted hypervariation.

3.3 Results

Sequence and assembly of 48Gb of gut viral DNA

To study diversity in natural populations of the human virome, we collected stool samples from 12 healthy individuals (3 per subject) over up to 2 months, then purified viral particles by sequential filtration, banding in CsCl density gradients, and treatment with nuclease as described [9]. DNA was isolated from viral particles, amplified, and then sequenced using the Solexa/Illumina HiSeq paired-end sequencing platform. A total of 495,053,311 reads were generated, averaging 97.2bp in length. As an empirical error control, 153 million reads were determined for DNA from phage ΦX174, showing an accuracy of 99.94%. A total of 48Gb of data were collected, the largest survey of viral sequences yet reported.

The raw sequences were assembled into contigs using the deBruijn graph-based assembler SOAPdenovo [10]. The depth of sequencing for the gut viral contigs averaged 49X and ranged up to 3,000X (Fig 3-1A). There were 78 contigs longer than 1kb that assembled as complete circles, indicating probable completion of the viral genome sequence. Circular assemblies could arise either by completing the sequence of a circular genome or by sequencing concatemers, which are intermediates in the replication of many DNA viruses. The mean number of contigs per subject longer than 1kb was 1390, ranging from 573 to 3390.

Protein functions were inferred by comparing the conceptual translation of predicted open reading frames to a curated database of protein families. A broad range of viral functions were identified in the encoded proteins (Fig. 3-1B), as observed previously [9,11]. On average, 72% of the ORFs did not resemble any recognizable protein family, emphasizing the immense diversity of novel genes in gut viral populations.

To assess the relationship to known viral genome sequences, contigs were compared to the NCBI RefSeq collection of viral genomes. The five database sequences with the most extensive similarity are shown in Figure 3-1C. Most of the recognizable viral sequences showed matches to prokaryotic viruses (DNA bacteriophages). Regions of similarity were typically short patches (median alignment length 202bp at E-value <= 10^{-5}), supporting the idea that bacteriophage functions are commonly organized in genetic cassettes [12,13]. Only one well characterized virus known to replicate on human cells was detected--Human Papillomavirus type 6b (Fig. 3-1C), which was only found within a single subject and sequenced to a depth of 23-fold. The next best hit to a eukaryotic virus was unconvincing, indicating that this viral fraction in healthy subjects is overwhelmingly composed of bacterial viruses.

Hypervariable loci in gut DNA viruses

To investigate sequence variation within each contig, we aligned the raw reads back to contigs and quantified variation at each base. Initial analysis showed that multiple small regions showed extremely high variation against a background of low variation. Comparison of filtering criteria led us to focus on regions of at least 90bp that were sequenced to a depth of at least 5X, contained a proportion of unique sequences of at least 40%, and contained a proportion of polymorphic bases of at least 5%, yielding 36 regions of the highest variability.

Analysis of these regions revealed that 12 resembled the diversity generating retroelement of phage BPP-1 described above [6,8]. This system is comprised of ~100bp repeat regions--the donor template repeat (TR), the targeted variable repeat (VR), and an RT, which is required to mutagenize the VR at positions where the TR contains an adenine [6]. Because of the central role of RT in this process, we identified all of the genes encoding RT-like sequences within the full collection of contigs, revealing 185 genes. Duplicated sequences can potentially break contig assemblies, so we manually inspected all of the RT-containing contigs to identify broken assemblies near the RT sequences suggestive of TR/VR pairs. We repaired 33 contigs through a combination of directed resequencing and manual re-alignment of shotgun Solexa/Illumina reads. This increased the number of variable regions to 51, those variable regions falling within a TR/VR pair to 36, and those with both an RT and a TR/VR pair to 29. In every case where a variable region was near an RT, it also contained a TR/VR pair. Such elements were found in 11 out of 12 subjects studied. Based on the resemblance to BPP-1, we refer to these systems as diversity generating retroelements below. Eighteen out of twenty-nine were found in contigs that could be tentatively assigned to a specific bacteriophage family.

Short hairpins near the VR are essential for activity in the BPP-1 system [14]. Similar hairpin sequences were found in only 13 of the above 29 elements. Of those 13 hairpins, 6 were found in ORFs, raising the interesting question of how this structure may constrain amino acid evolution in the host gene. This also suggests that some of the novel diversity generating retroelements described here may use initiation mechanisms that do not involve hairpin structures.

All of the 29 variable regions adjoining both an RT and a TR/VR pair were within an intact ORF longer than 500bp, allowing the targeted genes to be analyzed. We used BLASTp alignments and the homology-based structural prediction pipeline Phyre2 to analyze each ORF [15]. Fourteen ORFs had the hypervariable region near the 3' end of the coding region and showed a predicted C-type lectin fold (Phyre2 confidence score 90-100%), resembling the well-studied MTD of BPP-1 (% identity 15-33%; Fig 3-2A; the arrow indicates the direction of information transfer inferred from the BPP-1 model).

Novel gene types at hypervariable loci

Surprisingly, six hypervariable ORFs encoded proteins that aligned to cadherins, invasins, and fibronectins, which contain Ig-superfamily beta-sandwich domains. Proteins aligned with modest percent identity (11% to 15%), but structure predictions suggested multiple beta-sandwich domains with high confidence (Phyre2 confidence scores of 97-99%). Several of these proteins were also homologous to each other (80-90% identity), despite being isolated from different individuals. Ig-superfamily proteins are common in bacteriophage [16-18], but genes encoding Ig-superfamily proteins were not previously known to be subject to hypervariation by an RT-associated mechanism. The VRs for Ig-superfamily proteins were in the middle of the target ORF, so that the TR and VR were separated by an average of 1,938bp. In comparison, the MTD-like TR and VR are separated by an average of only 356bp. The greater distance between repeats in Ig-superfamily genes may have hindered their detection in previous studies based on DNA sequence comparisons [3,6].

Three hypervariable ORFs also encoded predicted leucine-rich repeat (LRR) proteins N-terminal to the hypervariable region (Phyre2 confidence score of 95-97%). These were all embedded in large ORFs, ranging from 1716-2181 predicted amino acids. One of these also had a C-type lectin fold in the hypervariable region. The remaining eight ORFs containing hypervariable regions did not have convincing similarity to known proteins, and may represent still further types of proteins subject to targeted hypermutagenesis.

Hypervariation at 5’-AAY-3’ adenine residues

Adenine residues were targeted in all the hypervariable regions identified here (Fig. 3-3A). 5'-AAY-3' sequences were particularly strongly affected (e. g. Fig. 3-2, bottom). This substitution pattern in 5'-AAY-3', which encodes asparagine, allows access to many different chemistries in the encoded amino acid side chains while suppressing creation of stop codons, as was originally pointed out for the MTD system [3]. The size of the dataset reported here allowed us to carry out statistical analysis of the placement of the 5'-AAY-3' relative to the three possible reading frames, which showed that the 5'-AAY-3' sequences were overwhelmingly in the asparagine-encoding frame (Fig. 3-3B, $P<10^{-163}$). Thus, variable region sequences have evolved to take advantage of asparagine-codon diversification while suppressing other types of changes.

RT gene populations in gut DNA viruses and bacteria

We next took advantage of the above data to annotate functions of gut virome RT genes. All of the RT sequences previously associated with hypervariable regions [3,6,19,20] cluster in a monophyletic clade containing the BPP-1 RT (Fig. 3-4A, cluster marked "DGR" for diversity generating retroelements). In our data set, we found that most of the new RTs clustered in this group (n=99), including all the RTs found to be associated with hypervariable regions (Figure 3-4A, green symbols). There was no obvious correlation between RT phylogeny and targeted gene type. Far fewer gut virome RTs clustered with group II intron RTs (n=8), and retron RTs (n=6).

We observed two novel groups of RT sequences. Five sequences fall into “Novel 1” (Fig. 3-4A), which is most similar to the Unk2 family [19]. Seven sequences fall into “Novel 2” (Fig. 3-4A), which is a sister clade to the retron RTs. The average pairwise distance of the pooled RTs associated with diversity generating systems described here is 1.14, which greatly increases the diversity of this group (previously 0.90), and rivals the diversity of the large retroviral/LTR retrotransposon RT clade (1.20).

We compared the distribution of RT clades in gut DNA viruses described here to that of their bacterial hosts. The bacterial genomes were dominated by the RT clades associated with group II introns and retrons, and thus differed from the DNA viruses, where RTs associated with diversity generating retroelements dominated (Fig. 3-4B).

3.4 Discussion

We report that DNA viruses of the human gut are rich in hypervariable regions, and that these are associated with template repeat/variable repeat pairs and characteristic RTs. The frequency of substitutions was so high that up to 96% of alleles in hypervariable regions encoded unique protein sequences. Most of the RT genes in the virome data set were in the clade linked to diversity generating retroelements--thus targeted hypervariation appears to be the major role of RTs in DNA viruses of the human gut.

Surprisingly, several of the genes subject to hypervariation were predicted to encode Ig-superfamily proteins--thus both gut viruses and vertebrate antigen receptors have evolved to use Ig domains as scaffolds for displaying highly diversified polypeptides. Evolution may have converged on these beta-sheet-rich domains because they are relatively rigid and so can maintain their folds despite primary sequence diversification, as has also been suggested for the C-type lectin fold [3]. The placement of diversified regions on Ig domains appears to differ between vertebrates and phage. Although more complete structural characterization is needed, modeling suggests that the phage Ig-superfamily domains may be diversified along one surface and into the adjoining linker between domains, while the vertebrate antigen receptors are diversified in loops between beta sheets within an Ig domain. The mechanism of diversification in phage clearly differs from that in the vertebrate immune system--the phage genes are diversified by error-prone reverse transcription [8], while the Gnathostomata immune system is diversified by V(D)J recombination, which involves DNA double strand breaks [21], and targeted deamination by AID [22].

The functions of the new viral hypervariable genes found here are not fully clarified. Hypervariable genes may encode viral structural proteins targeted by human IgA, which is secreted into the gut in large amounts, so that diversification of viral structural proteins may allow immune evasion. A role in ligand binding may be more likely, however, because only specific short regions are targeted for hypermutagenesis, so nonvariable regions would still be exposed to the immune system, possibly making immune evasion ineffective. Some of the hypervariable Ig proteins may be homologs of T4 hoc (highly immunogenic outer capsid) protein, which encodes an Ig protein related in sequence to those studied here. Hoc decorates T4 heads by binding to six-fold symmetric vertices in hexameric capsomeres, thereby providing a polyvalent binding moiety on the outside of phage heads. Hoc is proposed to mediate binding of T4 to surfaces such as the *E. coli* host cell [23], and has also been used for phage display to create new binding specificities for biotechnology applications [24]. If the Ig-superfamily proteins studied here are also accessory head proteins, they may mediate binding of viral particles to candidate host cells or environmental materials, allowing selection to enrich for those binding specificities that optimize reproductive success. The most useful binding specificities may differ widely during replication in the human gut or after shedding in feces, but hypervariation allows optimization in each new environment. Further research will be needed to clarify the full biological roles of these viral diversity generating systems.

3.4 Methods

Sample collection

Stool samples were collected from twelve healthy adult volunteers that were enrolled in a controlled feeding study, as described in [25] and [9]. Subjects were at least 18 years old, had a BMI between 18.5 and 35, were free of gastrointestinal disorders, and had not consumed antibiotics within six months prior to sample collection. Collections proceeded according to protocols approved by the institutional IRB. Six of the twelve subjects were the same as sampled previously [9] and all twelve were studied in [25]. In [9], multiple separate samples were sequenced for each subject, while in this study we pooled DNA from three samples per subject for sequencing.

Isolation and sequencing of viral DNA

Viral DNA was isolated from these stool samples as previously described [9]. Approximately 0.5g of stool was resuspended through homogenization in 40mL SM buffer [26]. Centrifugation at 4700g was carried out for 30min to remove large solids, and the resulting supernatant was passed through a 0.22µm PES filter (Nalgene, Rochester, NY). The 0.22µm filtrate was loaded onto a CsCl density step gradient (described further [26,27]), finally extracting the middle (1.35-1.5g/mL) fraction from the column using a sterile syringe. This study made no attempt to isolate RNA viruses or DNA viruses with densities outside this range. Chloroform was incubated with these samples for 10min prior to DNase (Invitrogen, Carlsbad, CA) treatment for 10min at 37C. Finally, viral DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA). This viral DNA was amplified with phi29 polymerase using random hexamer primers prior to pooling and sequencing (Genomiphi V2, GE Healthcare, Piscataway, NJ).

Viral DNA purity was assessed by quantifying the abundance of bacterial 16S rDNA, which is never encoded in viral genomes, via qPCR. Ten separate viral extractions were quantified in triplicate, using a plasmid standard curve to determine the number of 16S copies/ng DNA. First, the number of 16S copies/ng in bacterial DNA was estimated using an average genome size of 5Mb and an average of five 16S copies per genome. Analysis of total stool DNA shows, ~10^6 16S copies/ng, consistent with the majority of DNA originating from bacterial genomes [25]. The number of 16S copies/ng in the isolated viral DNA was 825 (standard deviation=805). The relative proportion of bacterial DNA in these viral preparations (compared to the total) was therefore estimated to be $9.0*10^{-4}$ (standard deviation=$8.7*10^{-4}$).

Three separate samples were pooled for each subject in this study, following isolation and amplification. Those pooled samples were randomly sheared by sonication (Covaris, Woburn, MA), and barcoded sequencing adapters were ligated using the Illumina TruSeq DNA Sample Prep Kit (Illumina, San Diego, CA). The same set of pooled samples was also prepared for Illumina sequencing using the Nextera DNA Sample Prep Kit (Epicentre, Madison, WI). The Illumina-prepared and Nextera-prepared samples were each pooled independently and sequenced on their own single lane of a HiSeq 2000 flow cell (Illumina, San Diego, CA). One lane of that same flow cell was set aside for Illumina control DNA isolated from the bacteriophage ΦX174.

Assembly and mapping of viral sequences

Viral sequences were trimmed by quality score (cutoff at Q35 using FASTX v0.0.13), and then assembled using SOAPdenovo (v1.05) [10] (kmer size = 63, additional flags "-p 20 –M 3 –u –G 200 –R"). We found that optimal assembly occurred when the reads from each sample were assembled separately, and when the largest possible kmer size was used. An insert size of 300bp was chosen based on the fragment size that was selected for sequencing. Reads were mapped back to those contigs using the Burrows-Wheeler Aligner (BWA v0.5.9-r16) [28] and the resulting alignments were visualized using the Integrative Genomics Viewer (IGV v2.0) [29]. Open reading frames were predicted using Glimmer (v3.02) [30], and functions were predicted using RPSBLAST [31] (v2.2.20) against the Pfam and NCBI Conserved Domain Database [32] (accessed 3/28/2011).

The contigs generated above were compared to the RefSeq collection of viral sequences using BLASTn [31]. The five genomes with the largest amount of sequence that was similar to at least one contig were selected, and raw reads were mapped to those contigs using BWA, as above. The pileup figures were generated using IGV.

Identification of variable regions

Variable regions were identified using a custom R script (available upon request) that uses Rsamtools to parse the BAM alignment files output by BWA (above). We estimated the basal error rate of Illumina sequencing by mapping control reads to the ΦX174 genome (gi 9626372). After excluding positions with >0.1 polymorphism (suggesting heterogeneity in the starting population), the error rate was calculated as the proportion of bases that did not match the reference. The script scans along every contig in a 50bp window (step size = 5bp) and extracts the sequences that cover that region completely. For each window we calculated the number of sequences, the proportion of those sequences that were unique (complementary to the proportion of sequences that were a duplicate of another), and the proportion of bases that did not match the consensus sequence. The criteria we chose to identify the most variable elements in this dataset were a minimum of 5 sequences, 0.4 unique alleles, and 0.05 polymorphic bases. Importantly, we required that 9 adjacent windows (a total of 90 contiguous basepairs) fulfilled these criteria.

Manual re-sequencing of contigs

Primers were selected that flanked the target gap using Primer3 (v0.4.0) [33]. The region of interest was amplified using AccuPrime Taq (Invitrogen, Carlsbad, CA), and the following thermocycler program: 94C for 15sec, 30 cycles of 94C for 15s and 68C for 3min, and finally 68C for 10min and cool to 4C. The resulting PCR products were purified using a QIAquick PCR Purification Kit (QIAGEN, Valencia, CA) and either sequenced directly, or cloned into a TOPO-TA vector (Invitrogen, Carlsbad, CA) for Sanger sequencing. The resulting Sanger reads were used in combination with shotgun reads to manually repair contigs, closing gaps with the new sequences. Those repaired contigs were then put through the analysis pipeline above, including read mapping, functional prediction, polymorphism scanning, etc.

Taxonomic classification of variable contigs

Each variable contig was compared to the viral proteins in RefSeq using BLASTx with a cutoff of $e<=10^{-40}$. The taxonomic classification of each RefSeq genome was retrieved from the NCBI website. When hits overlapped, the hit with the lowest evalue was retained. When one contig resembled reference genomes from multiple viral families, all of the matching families were recorded (e.g. “S/M”).

Sequence structure adjacent to VR

Recent work has identified short hairpins (8bp stem, 4bp loop, 20bp total) located in the IMH (‘initiation of mutagenic homing’ region) at the 3’ end of the VR as essential for DGR function [14]. We identified short hairpins in this dataset using a custom R script which scanned in 26bp windows looking for hairpins with either even or odd numbers of bases in the loop, and at least 7bp in the stem. An additional characteristic of the IMH is a 14bp GC-only sequence. We identified GC-only sequences using a custom R script that scanned each contig in 12bp windows, identifying each GC-only window, and then merging overlapping windows. The R script correctly called the experimentally verified signals in BPP-1 [14].

Structural prediction of variable ORFs

The amino acid sequence of ORFs covering hypervariable regions was generated using custom scripts. The structure of those amino acid sequences was predicted using Phyre2 [15] which uses homology of the input to sequences with known structures in order to generate the output. A threshold of 95% confidence was used to evaluate the output models.

Phylogenetic analysis of reverse transcriptase sequences

Reverse transcriptase (RT) sequences were identified using homology to the Pfam PF00078 (RPSBLAST; evalue < 0.00001), and the amino acid sequences were found using a custom script. Reference RT sequences were selected from two previous analyses of RTs associated with diversity generating systems [6,19], as well as representatives from each family of retroviruses (LTR-group RTs). The RT sequences from this dataset were aligned along with the reference sequences by individual alignment against the HMM contained in PF00078. A master alignment that preserved the position of each sequence relative to that HMM was generated by hmmalign (HMMER v3.0) [34]. This method does not involve any comparisons between the selected sequences, but rather relies on their similarity to conserved elements within the curated position-specific scoring matrix that constitutes the Pfam PF00078. An approximately maximum-likelihood tree was generated by FastTree [35]. The figure was generated using FigTree (http://tree.bio.ed.ac.uk/software/figtree/), coloring internal branches according to the confidence estimates generated by FastTree. Branch tips were adjusted and circles were added to indicate tips corresponding to novel RT sequences from this dataset. Overlapping circles were merged to avoid overplotting. A distance matrix was also calculated by MEGA using the Poisson-corrected distance, using only the subset of sequences described in [6].

In order to compare the relative abundance of different clades of RT sequences, we used the sequences (above) that fell into the Hypervariation, Group II Intron, Retron, and putatively Novel clades to compare against a variety of nucleotide databases using BLASTx [31]. The three genome databases that we used were 1) the collection of complete and partial viral genomes generated in this study, 2) all of the phage RefSeq genomes, and 3) all of the bacterial RefSeq genomes. For each of the genomes in those databases we recorded the clade that most closely resembled their RT sequences.

3.5 Tables and Figures

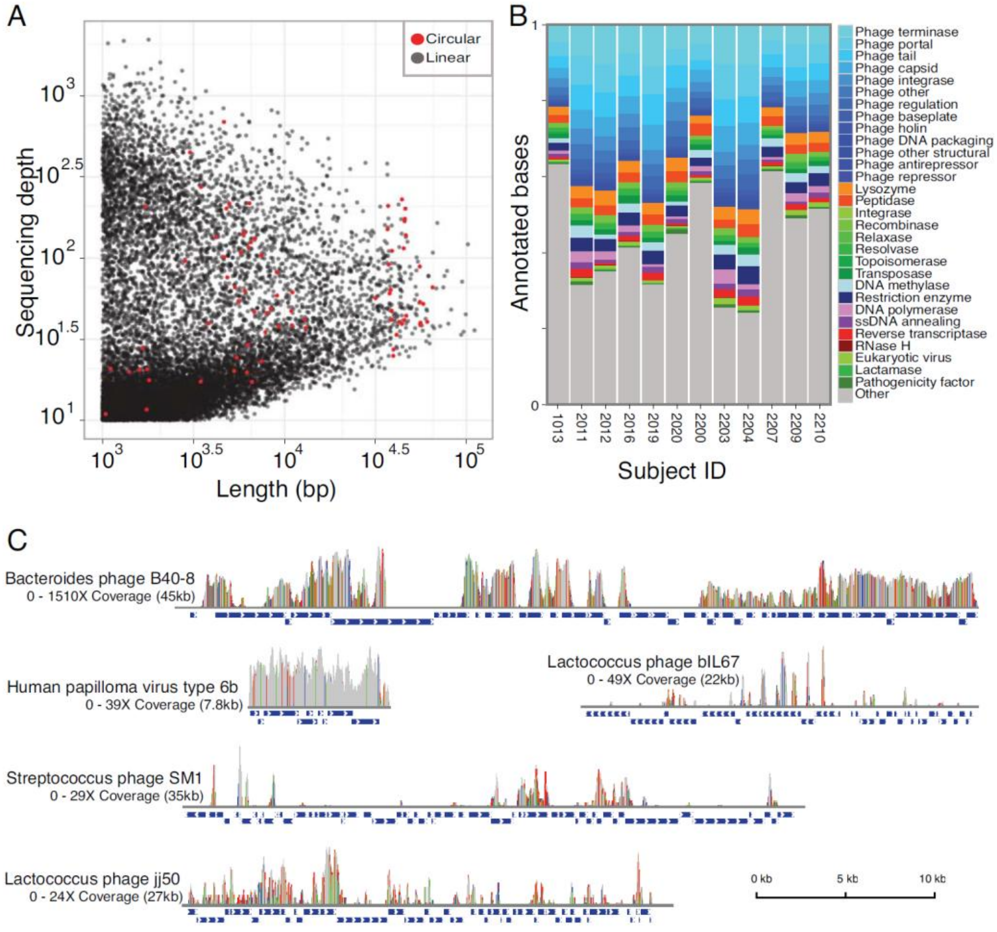

Figure 3-1. Assembly, functional assessment, and identification of viral sequences. (A) Summary of contigs assembled from viral sequences. Each contig is shown as a point, the length is shown on the x-axis, and the depth of reads mapped to each contig on the y-axis. Circular contigs are shown in red. B) Assignment of gene functions from viral contigs using the Pfam database of protein families; assignment of Pfam domains to viral functions is described in [9]. The proportion of sequences that were assigned for each function is indicated on the y-axis. On average, 21% of each sample was assigned to a Pfam protein family. C) The five viral RefSeq genomes with the most similarity to sequences generated in this study. Vertical lines indicate the depth of sequencing at that position, and colored lines indicate mismatches with the reference sequence. The range of coverage is noted to the left of each plot. Blue boxes below each genome indicate annotated genes.

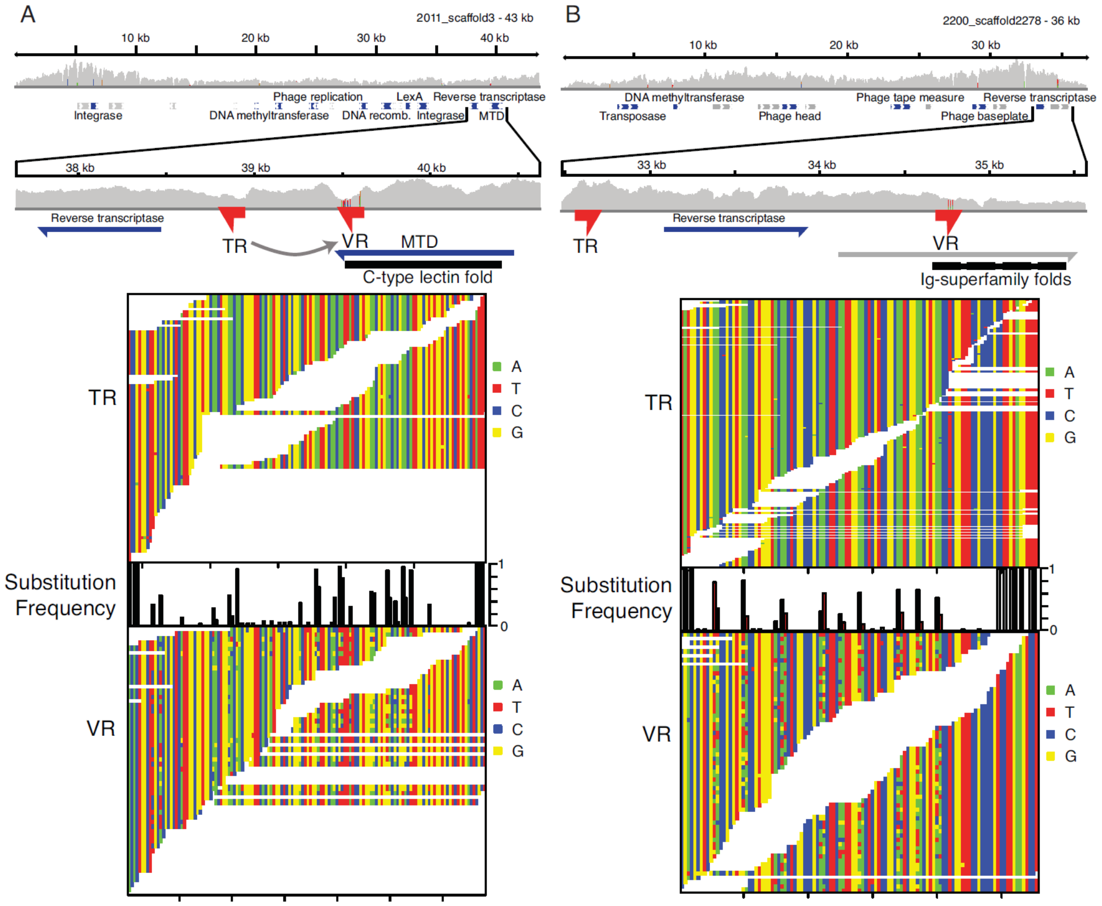

Figure 3-2. RT-associated hypervariable regions from the human gut virome. A) Hypervariation in a gene predicted to encode a protein with an MTD-like C-type lectin fold. B) Hypervariation in a gene predicted to encode an Ig-superfamily fold. In each panel, the top shows the contig of origin, with grey vertical lines showing sequencing depth and boxes showing annotated proteins. The indicated area is expanded below to show the template repeat (TR), the corresponding variable repeat (VR), reverse transcriptase, and the ORF that contains the targeted VR. The inferred direction of information transfer between the TR and VR is shown with arrow. The bottom of each plot shows an alignment of the sequences spanning the TR and VR for each element (white space indicates gaps between reads). Above the VR sequence is a barplot indicating the proportion of bases in the VR that differ from the consensus base in the TR. DNA bases are indicated by colors as indicated on the sides of the panels.

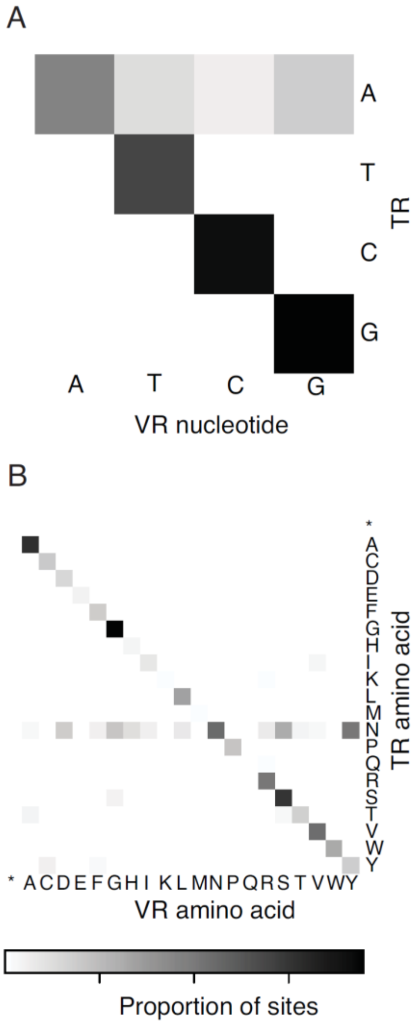

Figure 3-3. Characteristics of RT-associated hypervariation in the gut virome. A) Heatmap showing the relationship of positions in the template repeat (TR) (y-axis) to the resulting nucleotides in the variable repeat (VR) (x-axis). Out of 15,447 mutated bases, 14,930 (97%) are located at adenine-positions relative to the TR. B) Amino acid substitution heatmap showing the relationship of codons in the TR (y-axis) to the resulting codons in the VR (x-axis). Out of 11,462 mutated codons, 9,212 (80%) are located at asparagine (N) codons in the TR.

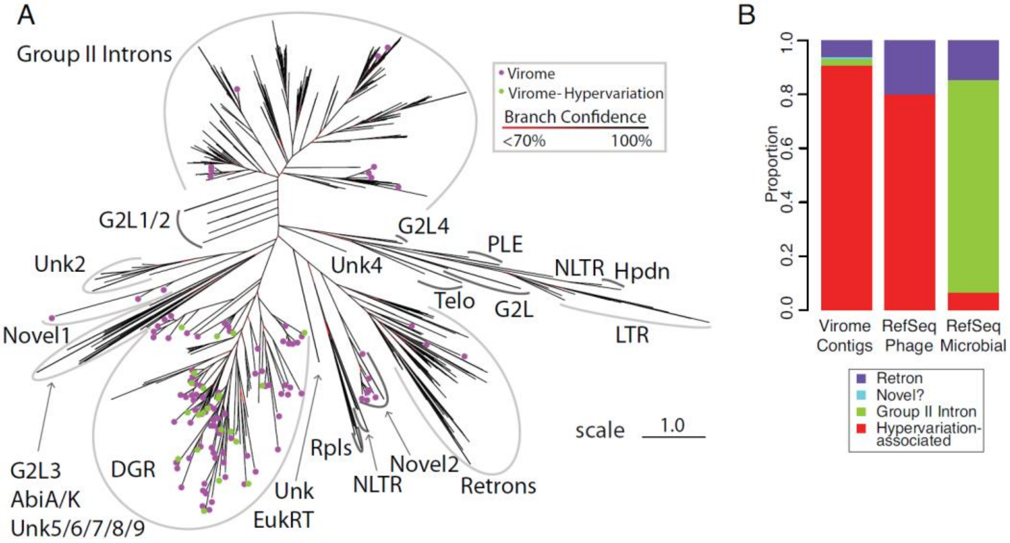

Figure 3-4. Reverse transcriptase (RT) sequences found in DNA viruses of the human gut. A) Phylogentic tree of RT sequences. Each sequence was aligned to a position-specific scoring matrix to construct a multiple sequence alignment. The tree was constructed using the maximum likelihood method. Green circles indicate RT sequences on viral contigs from this dataset that contain hypervariable regions and TR/VR pairs. Purple circles indicate other RT sequences from this data set, the remaining leaves indicate reference sequences from NCBI. RT clades were adapted from [6,19], and are indicated by grey lines. The bootstrap support of internal nodes is indicated by the color of internal branches as described in the key. Clades are marked according to [6,19]: "DGR"--diversity generating retroelements, "G2L" -group II intron-like families, "PLE"-Penelope-like elements, "NLTR"-non-LTR retrotransposons, “Hpdn”-hepadnaviruses, “LTR”-LTR retrotransposons and retroviruses, "Telo"-telomerase, “Unk”-unknown families [19], "Rpls"-retroplasmid, and “Abi” -abortive-phage-infection. Scale bar indicates the log-corrected distance metric used by FastTree, adapted from BLOSUM45. Distances range from 0, indicating a perfect match, to 3, indicating no overlap (scale bar indicates distance 1.0). B) Relative proportions of RTs in viruses studied here, the RefSeq phage genome database, and the RefSeq bacterial genome database.

3.6 References

1. Craig NL, Craigie R, Gellert M, Lambowitz AM (2002) Mobile DNA II: ASM Press.
2. Bushman FD (2001) Lateral DNA Transfer: Mechanisms and Consequences. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
3. McMahon SA, Miller JL, Lawton JA, Kerkow DE, Hodes A, et al. (2005) The C-type lectin fold as an evolutionary solution for massive sequence variation. Nat Struct Mol Biol 12: 886-892.
4. Miller JL, Le Coq J, Hodes A, Barbalat R, Miller JF, et al. (2008) Selective ligand recognition by a diversity-generating retroelement variable protein. PLoS Biol 6: e131.
5. Dai W, Hodes A, Hui WH, Gingery M, Miller JF, et al. (2010) Three-dimensional structure of tropism-switching Bordetella bacteriophage. Proc Natl Acad Sci U S A 107: 4347-4352.
6. Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, et al. (2004) Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature 431: 476-481.
7. Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, et al. (2002) Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. Science 295: 2091-2094.
8. Guo H, Tse LV, Barbalat R, Sivaamnuaiphorn S, Xu M, et al. (2008) Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification. Mol Cell 31: 813-823.
9. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, et al. (2011) The human gut virome: Inter-individual variation and dynamic response to diet. Genome Res.
10. Li R, Zhu H, Ruan J, Qian W, Fang X, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20: 265-272.
11. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, et al. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466: 334-338.
12. Hatfull GF (2008) Bacteriophage genomics. Curr Opin Microbiol 11: 447-453.
13. Veesler D, Cambillau C (2011) A common evolutionary origin for tailed-bacteriophage functional modules and bacterial machineries. Microbiol Mol Biol Rev 75: 423-433.
14. Guo H, Tse LV, Nieh AW, Czornyj E, Williams S, et al. (2011) Target site recognition by a diversity-generating retroelement. PLoS Genet 7: e1002414.
15. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc: 363-371.
16. Fraser JS, Yu Z, Maxwell KL, Davidson AR (2006) Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J Mol Biol 359: 496-507.
17. Fraser JS, Maxwell KL, Davidson AR (2007) Immunoglobulin-like domains on bacteriophage: weapons of modest damage? Curr Opin Microbiol 10: 382-387.
18. Pell LG, Gasmi-Seabrook GM, Morais M, Neudecker P, Kanelis V, et al. (2010) The solution structure of the C-terminal Ig-like domain of the bacteriophage lambda tail tube protein. J Mol Biol 403: 468-479.
19. Simon DM, Zimmerly S (2008) A diversity of uncharacterized reverse transcriptases in bacteria. Nucleic Acids Res 36: 7219-7229.
20. Kojima KK, Kanehisa M (2008) Systematic survey for novel types of prokaryotic retroelements based on gene neighborhood and protein architecture. Mol Biol Evol 25: 1395-1404.
21. Schatz D, Oettinger M, Baltimore D (1989) The V(D)J Recombination Activating Gene, RAG-1. Cell 59: 1035-1048.

22. Pavri R, Nussenzweig MC (2011) AID targeting in antibody diversity. Adv Immunol 110: 1-26.
23. Fokine A, Islam MZ, Zhang Z, Bowman VD, Rao VB, et al. (2011) Structure of the three N-terminal immunoglobulin domains of the highly immunogenic outer capsid protein from a T4-like bacteriophage. J Virol 85: 8141-8148.
24. Oslizlo A, Miernikiewicz P, Piotrowicz A, Owczarek B, Kopciuch A, et al. (2011) Purification of phage display-modified bacteriophage T4 by affinity chromatography. BMC Biotechnol 11: 59.
25. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334: 105-108.
26. Sambrook J, Russell DW (2001) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
27. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. Nat Protoc 4: 470-483.
28. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
29. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative genomics viewer. Nat Biotechnol 29: 24-26.
30. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23: 673-679.
31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: 421.
32. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, et al. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39: D225-229.
33. Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S, editors. Bioinformatics Methods and Protocols: Methods in Molecular Biology. Totowa, NJ: Humana Press. pp. 365-386.
34. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29-37.
35. Price MN, Dehal PS, Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5: e9490.

CHAPTER 4 – CONSERVATION OF GENE CASSETTES AMONG DIVERSE VIRUSES OF THE HUMAN GUT

The contents of this chapter have been submitted as:

> Minot S, Wu G, Lewis J, Bushman FD. Conservation of gene cassettes among diverse viruses of the human gut. *PLoS One*. Submitted

4.1 Abstract

Viruses are a crucial component of the human microbiome, but large population sizes, high sequence diversity, and high frequencies of novel genes has hindered genomic analysis by high-throughput sequencing. Here we investigate approaches to metagenomic assembly to probe genome structure in a sample of 5.6Gb of gut viral DNA sequence from six individuals. Tests showed that a new pipeline based on DeBruijn graph assembly yielded longer contigs that were able to recruit more reads than the equivalent non-optimized, single-pass approach. To characterize gene content, the database of viral RefSeq proteins was compared to the assembled viral contigs, generating a bipartite graph with functional cassettes linking together viral contigs, which revealed a high degree of connectivity between diverse genomes involving multiple genes of the same functional class. In a second step, open reading frames were grouped by their co-occurrence on contigs in a database-independent manner, revealing conserved cassettes of co-oriented ORFs. These methods reveal that free-living bacteriophages, while usually dissimilar at the nucleotide level, often have significant similarity at the level of encoded amino acid motifs, gene order, and gene orientation. These findings thus connect contemporary metagenomic analysis with classical studies of bacteriophage genomic cassettes. Software is available at https://sourceforge.net/projects/optitdba/.

3.2 Introduction

Advances in DNA sequencing technology have made it possible to characterize microbial communities using extremely large numbers of short sequence reads [1-3]. This offers a powerful tool for interrogating complex communities of uncultured organisms, but analyzing the shotgun sequence data from mixtures of organisms poses considerable computational challenges [4-7]. Here we address the problem of assembling genomes from complex viral communities to investigate conserved features of gene content and order.

Viral communities critically influence environmental microbial populations and human health, but their study is hampered by a large degree of uncharacterized sequence diversity. It has been estimated that 0.0002% of the global viral gene pool has been sequenced [8] and deep sequencing of viruses purified from the environment typically yields a large majority of unidentified sequences [2,3,9-11]. Therefore the effective study of these viruses depends on the efficient computational assembly of individual reads into large genome fragments without reference to known genomes.

The assembly of mixed viral reads presents a number of challenges. Viral genomes are small but range widely in size, from 5kb to >1Mb [12-14], so size cannot easily be used to assess genome completion. Different viral genomes can be present in widely differing proportions [15,16], complicating the use of coverage to judge correct assembly. Viral genomes can also evolve quickly, including frequent recombinational exchange of protein-coding cassettes [17-19] and high rates of nucleotide substitution [11], further confusing assembly.

However, many viral genomes are either circular, such as φX174 [20], or are circularly permuted, such as T4, which includes 1.02 genome copies in each viral head [21]. Thus assembly into circles indicates probable completion of the sequence, but circularity not yet been used widely to improve viral sequence assembly.

The problem of *de novo* assembly of high-throughput sequencing datasets has been greatly aided by the development of de Bruijn graph assemblers [22-27]. In the de Bruijn graph method, extremely large sets of short sequences (such as those generated by Illumina HiSeq technology) can be assembled into complete and partial genomes by mapping them onto a de Bruijn graph (Fig 4-1A) [24,28]. Each read is computationally fragmented into sequences of length k (the so-called 'kmer'), then each kmer is used to form an edge between nodes corresponding to sequences of length k-1. By drawing such edges for every read in the dataset, one constructs a de Bruijn graph, which contains the information necessary to reconstruct the genome sequences that gave rise to the graph. In a subsequent step, a consensus contig sequence is constructed from the de Bruijn graph, which involves popping bubbles, trimming branches, resolving repeats, and more complex operations to generate a linear graph [4]. Assembly by this method scales linearly with increasing sequence number, while the more familiar overlap method of assembly scales exponentially, explaining why the de Bruijn graph method is used for very large data sets [22].

However, complexities within the sequence population, such as nucleotide polymorphisms or short sequence repeats, can introduce misleading connections in the de Bruijn graph [5,29]. In Figure 4-1 we demonstrate how the optimal kmer -- i. e. one that minimizes misleading connections -- depends on the nature of the underlying sequence. A set of genomes with three independent SNPs separated by 25bp (Figure 4-1B) will produce a de Bruijn graph with three isolated bubbles at a kmer of 23, while it will produce a much more complex structure at a kmer of 27. In contrast, two unrelated genomes with 25bp of identical sequence (Figure 4-

1C) will be joined together at a kmer of 23bp, but not at one of 27bp. These examples demonstrate how the difficulty of parsing a de Bruijn graph depends both on the nature of the underlying polymorphism and the kmer value used.

Thus in a mixture of multiple microbial genomes, it is likely that the optimal kmer value for assembly will vary [5,30]. One group [5] found that combining the assemblies constructed across a range of kmer values yielded a large number of long contigs, but that these contigs did not faithfully represent the underlying genomes. Another group developed IDBA, which performs sequential assemblies while stepping through kmers of increasing lengths [30]. At each kmer value, IDBA removes the best contigs and the reads used to make those contigs. A metagenomic version of this program, MetaIDBA, has been developed [29].

In this paper our goal is to find patterns of genome conservation in the highly diverse collection of viruses found in the human gut [2,3,11,31,32]. We implement an optimized iterative de Bruijn graph assembly approach, significantly increasing the length and depth of the assembled contigs compared with previous virome studies. We present results for 5.6 Gb of Illumina paired-end sequence data from six human gut virome samples (a subset of samples reported initially in [11]). While only a minority of the assembled ORFs in the sample could be annotated – emphasizing the vast diversity of gut viral populations – the annotated ORFs tended to group by predicted function. Moreover, a large fraction of the complete collection of ORFs could be clustered into inferred cassettes with conserved gene order and orientation. Analysis of the contigs produced emphasizes the extreme variation in gut bacteriophage populations across individuals, while at the same time suggesting that viral genomes are organized in conserved functional gene cassettes.

3.3 Results

Assembly of viral contigs

In order to analyze protein conservation among viruses derived from mixed environmental samples, it is necessary to generate contigs that most closely approximate complete viral genomes. To generate long contigs that faithfully represent the underlying genome structure, we developed and employed an optimized iterative de Bruijn graph assembly approach (OPTITDBA). We compared this assembly method to two previously published methods (SOAPdenovo and MetaIDBA) using 5.6 Gb of Illumina HiSeq data (100 bp paired end reads) derived from stool virome samples from six healthy human subjects [11]. An example of assembly for samples from one of the six subjects is shown in Fig. 4-5. We found that assembly using our interative method (OPTITDBA) resulted in fewer reads mapped to contigs less than 1kb in length, and more reads mapped to contigs in each of the three longer size classes (1-3kb, 3-10kb, and >10kb) (Fig 4-2), performing better than either SOAPdenovo or MetaIDBA ($p<0.05$; Wilcox signed-rank test).

In order to measure the accuracy of this method, we used synthetic viral communities composed from previously sequenced communities. Of the 6 subjects in this dataset, one contained sequences that align closely to Human Papilloma Virus type 6b (as reported previously in [11]). These reads were added in varying amounts to a collection of reads from a subject completely lacking HPV reads, and the resulting synthetic datasets were used to assess quality of assembly. We assembled these synthetic viral communities using OPTITDBA, MetaIDBA, or SOAPdenovo and compared the efficiency of HPV recovery (measured in this case as the length of the longest HPV contig as a proportion of the whole HPV genome). For every level of sequencing (6, 13, 19, and 23X coverage), the HPV genome was better assembled using this pipeline than using the single pass SOAPdenovo assembly ($p<0.0005$; Wilcox signed-rank test) (Fig. 4-6). On average, this pipeline performed 61% better than the corresponding single assembly using SOAPdenovo. There was no significant difference in HPV genome recovery between MetaIDBA and OPTITDBA, though our pipeline was better than MetaIDBA in producing contigs that recruited the maximum number of reads. In summary, OPTITDBA assembled viral reads into longer contigs at no cost to accuracy in the reconstruction of control genomes.

Network analysis of bacteriophage proteins

In order to characterize the assembled viral genomes, we predicted open reading frames (ORFs) using Glimmer, yielding 29,017 ORFs longer than 100bp from the 6 datasets. Of these, only 3,066 had similarity at a cutoff of $E<10^{-10}$ to the RefSeq collection of viral proteins (10.6%). At a more stringent cutoff of $E<10^{-50}$, only 690 ORFs were similar (2.4%). Searching for conserved amino acid motifs contained within the Conserved Domain Databases (CDD) yielded 3,374 ORFs with a match in the CDD at $E<10^{-10}$ (11.6%), but only 777 with a match at $E<10^{-50}$(2.7%).

In order to investigate which of these RefSeq annotations were shared among contig-encoded ORFs, we carried out a network analysis (Fig. 4-3). The nodes in this network represent either contigs (orange circles) or RefSeq viral proteins (smaller black circles). Edges (connections) are drawn between contigs and RefSeq proteins when an ORF (encoded by a contig) is highly similar to a RefSeq protein ($E<10^{-50}$). Groups of RefSeq proteins that are similar to multiple contigs are highlighted by light blue ovals. While in some cases these groups of reference proteins encode only a single function (in which case they are likely all similar to a single ORF on each of the indicated contigs), in others there are multiple predicted functions (in which case there is a similar collection of genes found on all of the indicated contigs). For example, multiple contigs are linked by genes encoding both capsid and terminase proteins, while others are linked by genes encoding transcription and DNA packaging functions. These examples parallel previous work which showed that bacteriophage genomes are often organized into cassettes of functionally related genes [33-35]

Bacteriophage genomes contain conserved cassettes encoded by divergent nucleic acid sequences

After finding only a low frequency of similarities between ORFs in viral contigs and database sequences, we searched for conserved gene cassettes in a database-independent manner. ORFs were compared within the assembled sequences to find encoded amino acids sequences that were repeated among multiple contigs, which we refer to as protein-coding families. Of the 29,007 predicted ORFs, 16,944 (58%) were found to be members of families, that is, ORFs on different contigs showed alignments with at least 30% identity. A total of 2,961 families contained 2 ORFs each. The largest family contained 25 ORFs. Of these 5,135 protein-coding families, only 1,287 (25%) had any similarity to the Conserved Domain Database, emphasizing the amount of unexplored diversity in genes of the gut virome.

Relationships among these protein-encoding families were interrogated by grouping families into cassettes, consisting of different families that were found on the same group of contigs. We found 28 types of cassettes that contained from 2 to 8 protein-coding families. On average, the amount of each contig that was covered by each cassette was 1.6kb, ranging from 105bp to 11.5kb. The mean proportion of each contig that was covered by a cassette was 27%, ranging from 1% to 90%. The most common cassette was found on 20 contigs generated from all six subjects studied. Of the 16,944 ORFs found in families, 651 (4%) were found in cassettes (Table 4-1). While a small proportion of the total number of predicted ORFs were grouped into cassettes, this accounted for a disproportionately large amount of the input sequence reads. The contigs containing at least one ORF accounted for $3.1*10^7$ reads. The contigs containing a cassette accounted for $5.9*10^6$ reads, or 18% of all contigs (Table 4-2). Therefore while the proportion of contigs that harbor cassettes is relatively small, contigs with cassettes represent highly abundant lineages.

Bacteriophage cassettes commonly show conserved gene orientation as well as conserved gene type, so we investigated orientation as well. The degree of co-orientation among protein-coding regions in cassettes was found to be high, with an average co-orientation score of 99% (compared to 25% co-orientation expected by chance), providing strong support for cassette structure.

In a few cases, the proteins encoded within a cassette showed potentially related annotations, such as N-6 DNA methylase and DEAD-like helicase (Fig 4-7) or phage portal and terminase (Fig. 4-4A). In many cases, specific unannotated ORFs were repeatedly found near ORFs annotated as phage proteins. In one case, proteins with less than 30% amino acid identity between them (resulting in their being grouped in different families) were assigned the same functional annotation (Phage Mu F: morphogenesis-related protein) and located in the same functional cassette (Fig. 4-4B), suggesting preservation of protein function and genetic organization despite nucleotide and amino acid divergence (Fig. 4-4B).

3.4 Discussion

One difficulty of studying viruses in the environment is that high-throughput sequencing data is difficult to interpret when high proportions of reads are unknown or unrecognizable. One way to address this problem is through *de novo* assembly, generating complete and partial genome sequences. For environmental bacteriophage this is usually necessary because previously sequenced and closely related genomes are usually not available. We demonstrated that optimizing the assembly process according to the characteristics of viral genomes dramatically improves the degree of assembly at no cost to accuracy. Using our assemblies, we found that the viral open reading frames often cluster in related cassettes, but that the cassettes show considerable sequence divergence among genomes.

In our assembly pipeline we have improved on iterative kmer based de Bruijn graph assembly for use with viral samples in three ways. 1) We picked the optimal kmer value to use at each iteration, rather than cycling once through kmers of increasing length. 2) We removed at each iteration the set of reads that aligned to the best contigs, not those reads that were used to construct those contigs, because due to the nature of the de Bruijn graph assembly process, the set of reads used to construct a contig may not fully contain the set of reads that align well to that contig. 3) We reasoned that circular sequences would represent complete viral genomes, either as circular genomes or circularly permuted genomes, and so used this also as a criterion for calling finished contigs.

As a measure of the quality of assembly, we monitored correct assembly of Human Papillomavirus Type 6b (HPV), the one known virus in our data set. We found that our pipeline was better able to assemble a single contig matching HPV across a range of sequencing depths than was SOAPdenovo (the underlying assembly algorithm used in our pipeline). Both our pipeline and MetaIDBA reconstructed HPV about equally well (however, as described below, our method yielded contigs explaining a larger proportion of the reads). The ability to reconstruct viral genomes present at low abundance is particularly important when trying to detect pathogens in sequence mixtures, such as in efforts to identify novel pathogens in samples from outbreaks of infectious diseases.

A more complicated challenge is assessing the quality of assembly of unknown viral genomes. One common metric for assessing assembly quality is the length of contigs produced (N50). However, a recent study [5] found that for one implementation of De Bruijn graph assembly of short sequences from known genomes, the method that yielded the highest N50 yielded the lowest similarity to the known genomes. Therefore we chose to measure how well the contigs explain the input data by mapping reads back to contigs, and counting the number of reads that mapped to contigs of different size classes, thereby generating an estimate of how well the assembly process reconstructed the primary data. We found that our pipeline performed better than MetaIDBA or SOAPdenovo.

Analysis of protein conservation emphasized the cassette structure of the viral genomes in our samples. We annotate the viral contigs by aligning new ORFs to available databases, and by identifying ORFs of unknown function that aligned with other ORFs in our data set. We found that viral ORF families often clustered in cassettes, where genes with similar sequences were almost always in the same orientations. Cassette structure has been well documented in many bacteriophage families [13,33-37]--here we show that these structural patterns are accessible after assembly of metagenomic data.

A conjecture to explain the observed phage genome structure invokes pressure for sequence diversification from the CRISPR system. Many bacterial genomes harbor a series of repeated sequences spaced by short sequences derived from phage or plasmids, called CRISPR arrays. The CRISPR arrays are transcribed, then the spacer sequence RNAs are used as recognition elements to program degradation of incoming sequence-complimentary DNA. Thus bacteriophages that infect CRISPR-containing hosts are regularly under pressure to alter their DNA sequences to evade attack. Assisting this, bacteriophage replication cycles can be as short as 20 minutes and burst sizes large, allowing rapid evolution. Constraining the allowable DNA substitutions, of course, is the requirement for proper function of the encoded proteins. In a few cases three dimensional structures have been determined for multiple phage proteins encoded in syntenic regions from functionally interchangeable cassettes, and the structures can be surprisingly similar given the low DNA and protein similarity. For example, the repressor and Cro proteins of Lambda, 434, and P22 show little similarity at the nucleic acid level (median identity 34%) or amino acid level (median identity 17%) [38], but share common alpha-helical structures and helix-turn-helix motifs [39,40]. Thus the modules emerging from the metagenomic assembly may represent functionally similar gene sets that have diversified to elude attack by the CRISPR system, perhaps helping to explain why bacteriophage populations show such extreme variation.

3.5 Methods

Iterative assembly pipeline

Here we first describe the basic steps of the optimized iterative de Bruijn graph assembly pipeline (available at https://sourceforge.net/projects/optitdba/), and then describe the implementation of each step in more detail. For each iteration, OPTITDBA 1) selects the optimal kmer, 2) generates a de Bruijn graph for the optimal kmer length, 3) removes the reads that map to the most highly abundant contigs from the dataset or reads that map to circular contigs, and 4) starts another iteration using all of the reads that do not map to those contigs. The loop ends when there are no highly abundant contigs meeting the criteria outlined below. At that point, all of the remaining reads will be assembled and mapped using the optimal values from the final iteration.

Selecting the optimal kmer

OPTITDBA assembles over a range of kmer values (63, 59, 55, 51, 47, 43, 39, 35, 31, 27, 23, and 19) using SOAPdenovo v1.05 [25] (flags: -p 10 -d 1 -M 3 -u -G 200 -R). All the steps taken to simplify the de Bruin graph, clipping tips, removing low-coverage links, resolving tiny repeats, and merging bubbles, were implemented as described in [25]. Each assembly is queried for whether any circular contigs longer than 2kb were generated, suggesting complete assembly. If multiple kmer values resulted in circular genomes, than the largest such kmer value is selected.

If no such circles are generated, then the kmer values are scored by the length and depth of sequencing of its most abundant members. For each assembly, the contigs are sorted by the number of reads used to construct them and the cumulative length of the top 20 contigs is recorded. All contigs <1kb in length are excluded. The kmer value with the longest cumulative length of its 20 most abundant is selected as the most optimal for that loop. The number of contigs selected (20) is arbitrary, can be specified by the user, and is used to balance computational resources against thorough assembly.

If more than 10^6 reads are used as input, OPTITDBA randomly selects 10^6 reads to use for the assembly trials. Preliminary tests indicated that this strategy reduced computation time at no detriment to optimal kmer determination.

Removing reads that map to the most highly abundant contigs

If no circular contigs are found, then the top 20 contigs from the assembly with the optimal kmer value are retained. If circular contig(s) is/are found, then the circular contig(s) and the top 20 contigs are retained. In pilot tests it was found that reducing the number of contigs that are retained at each step increases the final number of iterations as well as computational time, while the proportion of reads that were mapped was not impacted significantly.

The reads are then mapped to those retained contigs using BWA v0.5.8c [41]. The full set of reads is used to map, not the random subset described above (if used). All reads not mapping to these contigs are then used to start another iteration. The training set consisted entirely of paired reads, and both members of the pair were removed, even if only one mapped to a contig. We observed that when only one read in a pair mapped, the other often covered the junction of circular contigs, or gaps in the assembly. The cycle ends when either zero reads map, or there are zero contigs >=1kb in length.

While the iterative assembly pipeline developed here implements SOAPdenovo to perform assembly and BWA to perform mapping, the concept is independent of both programs.

Benchmark sequences

The data used to benchmark this pipeline are those described in [3,11]. Viral DNA was isolated from human fecal samples using sequential filtration and CsCl density ultra-centrifugation, then unprotected DNA was digested using DNaseI [42]. Viral DNA was subsequently recovered from particles, yielding a sample that was depleted in bacterial DNA by >100-fold (as measured by 16S rDNA qPCR [11]. Three samples each from six human subjects were extracted, pooled, and submitted for sequencing using the Illumina HiSeq 2000 platform

(100bp paired-end sequencing). Ten million reads were randomly selected from each dataset (except for Subject 6, which only had 5,754,268 reads) while preserving all read pairings, and assembled using either OPTITDBA, MetaIDBA v0.19 [29], or using SOAPdenovo with a kmer value of 63 and all of the same flags as in the iterative assembly pipeline. The kmer value of 63 for SOAPdenovo was found in previous tests to produce the highest N50 score using this dataset. MetaIDBA was run using default settings. See Table 4-3 for a summary of each dataset.

Detection of Human Papillomavirus Virus

A previous analysis of these sequences found evidence of a single eukaryotic virus: Human Papilloma Virus type 6b [11]. Reads mapping to the HPV genome (NCBI gi: 9626053) were extracted and used to mix back in various quantities to a dataset that did not previously have any detected HPV sequences. The mixing was done by randomly selecting a total of $4*10^6$ reads for each test. The number of HPV reads varied across a range (500, 1000, 1500, or 1798 reads, corresponding to 6X, 13X, 19X, or 23X coverage), with three replicates of each. Each set of mixed reads was assembled using OPTITDBA, MetaIDBA, or SOAPdenovo as described above.

Network analysis of viral proteins

In order to classify the assembled viral contigs according to their similarity with known proteins, we compared the predicted open reading frames (ORFs) on these contigs with 1) the RefSeq [43] collection of viral proteins, or 2) the Conserved Domain Database (CDD) [44] of conserved amino acid motifs. ORFs were predicted using Glimmer v3.02 [45], compared to Viral RefSeq (downloaded on 12/16/11) using blastp [46] (v2.2.25+, build 1/3/12), and compared to CDD [44] (downloaded on 10/18/11) using rpsblast (v2.2.25) [46]. Because of the difficulty of manually identifying patterns of similarity among contigs, we converted the protein similarity data into a format that could be viewed in the interactive network visualization tool Cytoscape [47]. In this bipartite network scheme, there are two classes of nodes: contigs and RefSeq proteins. When a RefSeq viral protein has a highly significant match ($E<10^{-50}$) to an ORF encoded by a contig, a connection is made between those two nodes. For ease of visualization, we excluded all contigs that were either shorter than 3kb, or had fewer than 5 hits to RefSeq proteins. The nodes and connections for all six datasets were combined and loaded into Cytoscape. The network was arranged using the spring-embedded layout (data available upon request).

Protein family organization

In order to search for conserved protein families in a database-independent manner, we clustered the ORFs described above using UCLUST v1.2.22q [48]. Each of those protein families was compared to the Conserved Domain Database using rpsblast. Those protein families were next grouped into cassettes, meaning multiple protein families that can be found together on our contigs. Cassette discovery proceeded in the following manner. Each protein family was classified according to the list of contigs that encoded it. Next, all of the protein families were compared, seeing how many of those occurred on common contigs. A given pair of protein coding families was grouped into a cassette when the smaller of the two families was found on a shared contig at least 80% of the time. This process was performed iteratively, recalculating the overlap scores after each pair of protein families was merged together. In subsequent iterations, protein families could also merge in the same way with cassettes that formed earlier.

If a pair of proteins formed a cassette found on multiple contigs, we expect shared ORFs to be in the same relative orientations. To calculate the consistency of orientation across contigs, we used a simple co-orientation score, calculated in the following way. Any two genes have four possible relative orientations. For every pair of protein clusters in a module, we calculate the proportion of contigs that contain the orientation found most commonly. Discovery and analysis of protein modules was implemented in an R script that is available along with the iterative assembly pipeline, at https://sourceforge.net/projects/optitdba/.

Computation

Computation was carried out on a home-built computer with 192 Gb of RAM and 12 cores (24 hyperthreaded). The computer was assembled from parts costing $16,060 (USD) (a full parts list is available at http://microb230.med.upenn.edu/protocols/comput_resources.html). Computation times for assembly of single viral communities ($5.7*10^6$ – 10^7 reads) using this pipeline were 20.7 to 132.1 wall clock hours, with a median of 39.0 hours. The computation time may vary with the community complexity and number of reads, as the dataset with the longest compute time (#1: 132.1 hours), also had the largest number of predicted species by PHACCS (data not shown), the dataset with the shortest compute time (#6: 20.7 hours) was the only one to have less than 10^7 sequences, and all of the other datasets ranged between 38 and 56 hours.

3.6 Figures

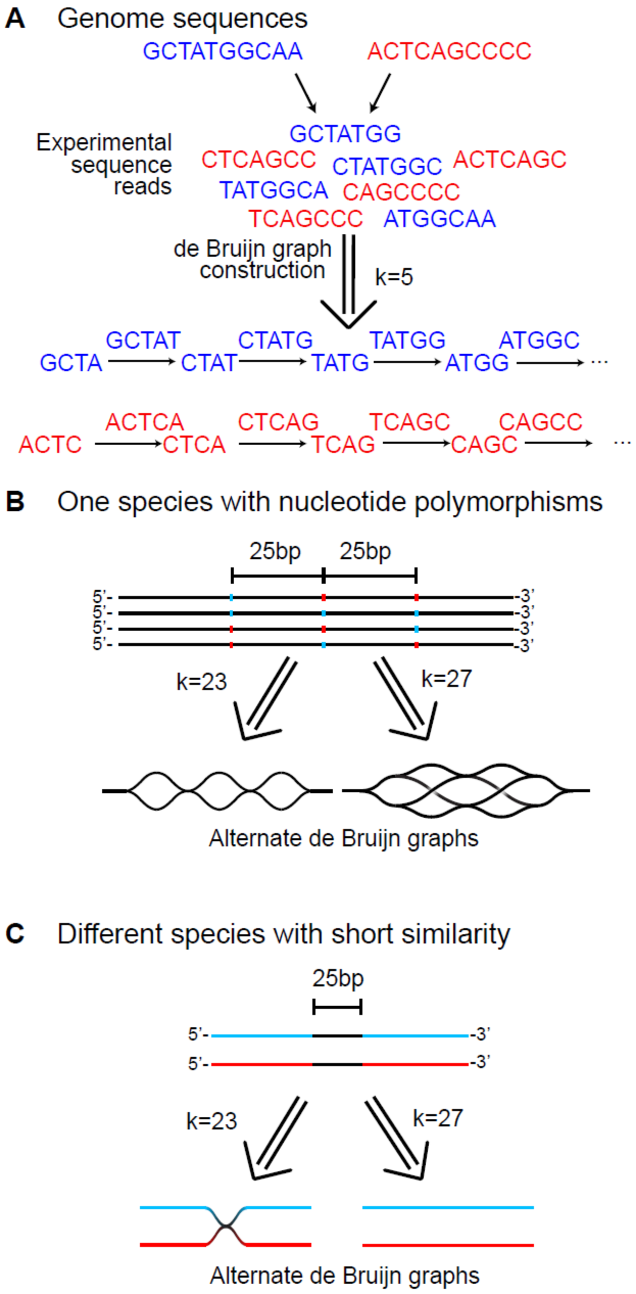

Figure 4-1. The de Bruijn graph assembly method and the influence of genomic variation on de Bruijn graph complexity. A) Shotgun sequences are produced from two different genomes (shown in blue and red at the top). Those sequences are used to construct a de Bruijn graph, where nodes are formed by all possible sequences of length k-1 (in this case 4 bases), which are connected by edges of length k (5 bases). Since there are no 4mers shared between these two example genomes, the resulting de Bruijn subgraphs are separate. B) Nucleotide polymorphisms are better resolved by short kmers. We consider a mixture of four genomes, each with three polymorphic positions separated by 25bp. The identity at each polymorphic position is represented by either blue or red to indicate different nucleotides. At all other positions the genomes are identical. The de Bruijn graph that is constructed from this mixture of genomes using a kmer of 23 is shown on the left, where three independent bubbles form around each polymorphic position. The equivalent graph at k=27 is shown on the right, where three independent sets of bubbles overlap, forming a more complex and suboptimal graph structure. C) Short regions of similarity are better resolved by long kmers. We consider a mixture of two genomes which are entirely different except for a 25bp region of sequence identity (shown in black). The de Bruijn graph that is constructed from this mixture at k=23 is shown on the left, where the two resulting subgraphs intersect at the 23mer of similarity. The de Bruijn graph at k=27 is shown on the right, where the two resulting subgraphs (corresponding to the two genomes) do not intersect, since they have no 26mer in common. The examples in B and C together illustrate how different kmers can be optimal for assembling graphs with different types of polymorphisms.

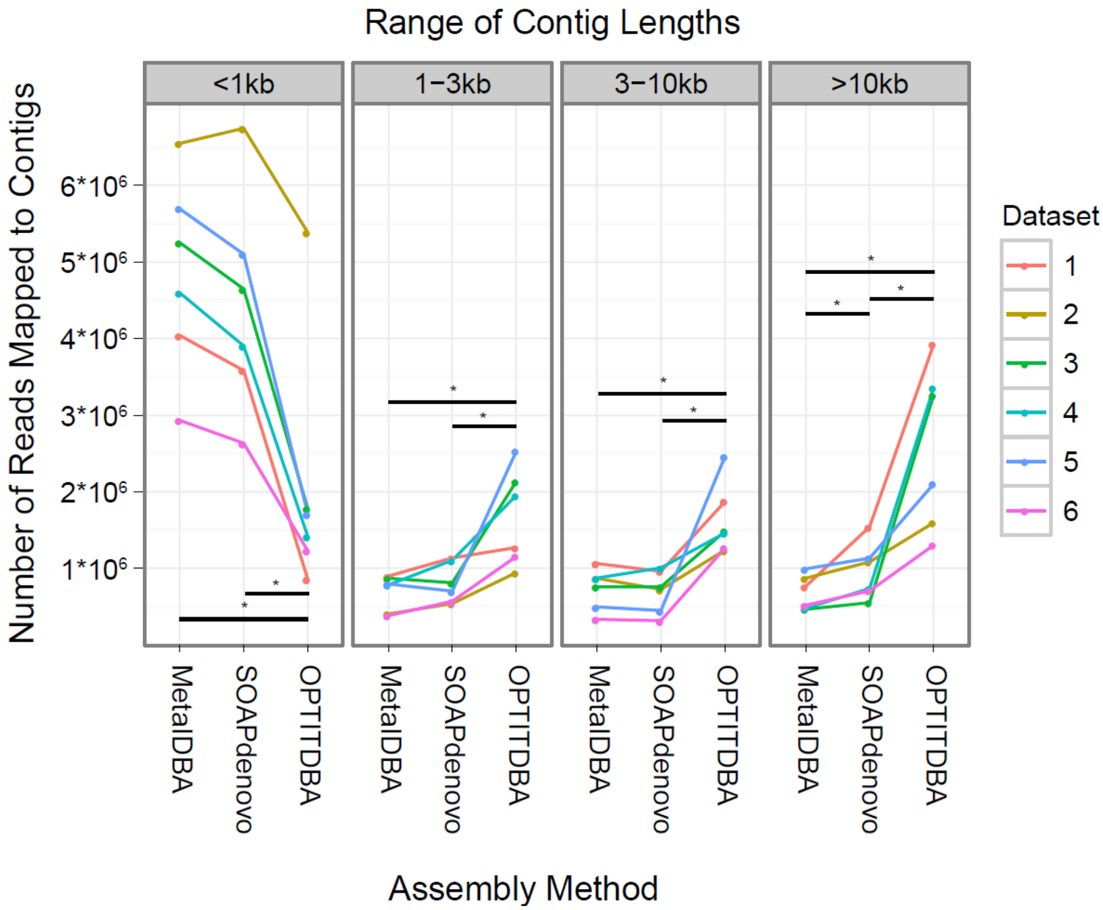

Figure 4-2. Comparison of assembly methods by read alignment. The vertical axis indicates the number of reads from each dataset that align to contigs of different size classes (either less than 1kb, between 1kb and 3kb, between 3 and 10kb, or longer than 10kb). The horizontal axis separates assembly method. Each dataset is indicated by color (see key on right; numbers indicate gut virome communities from different human subjects). * indicates $p<0.05$ by Wilcoxon signed-rank test for the indicated pair of assembly methods.

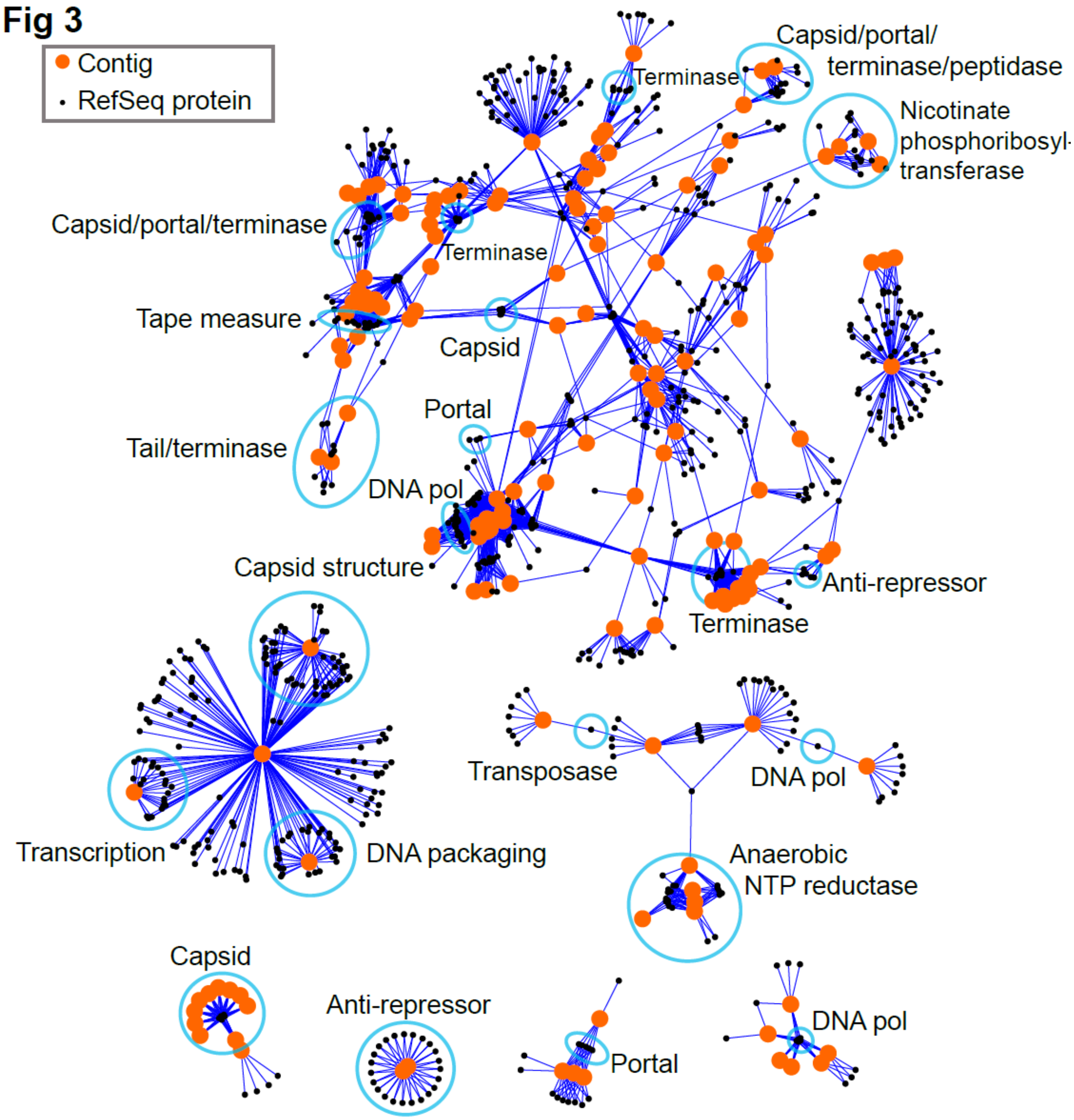

Figure 4-3. Network based annotation of viral contigs. Orange circles represent viral contigs no shorter than 3kb. Black circles represent proteins in the RefSeq viral database. RefSeq proteins are connected to viral contigs when an ORF encoded by that contig resembles that protein at $E<10^{-50}$ (blastp). Blue outlines indicate groups of RefSeq proteins and ORFs from contigs that share the function indicated by the adjacent label.

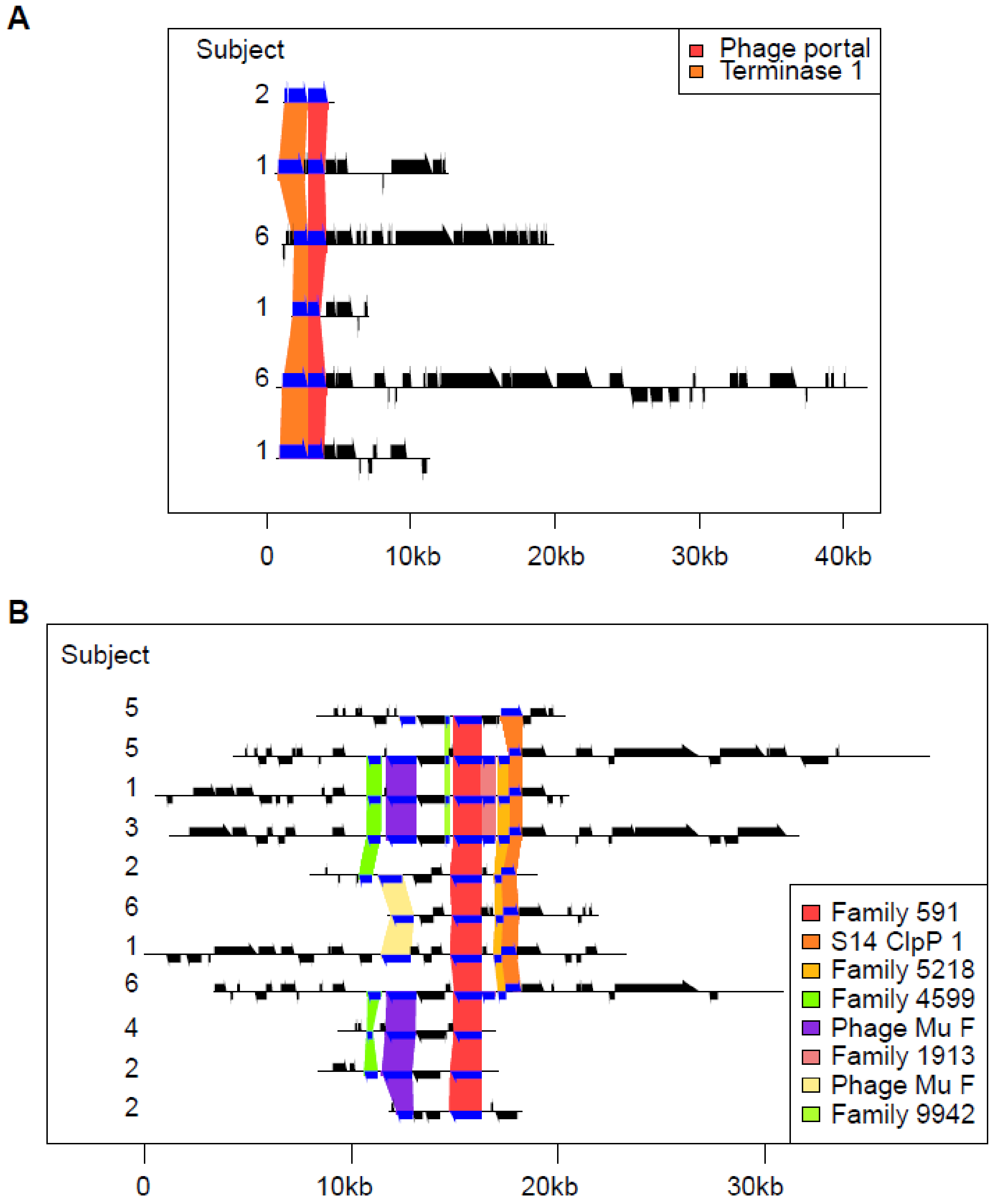

Figure 4-4. Two examples of phage cassettes (A and B). Contigs are shown as horizontal black lines, ORFs on those contigs are shown by black arrows above and below those lines, and the organization of those ORFs into protein-coding families is shown with colored boxes. The subject that each contig was assembled from is shown on the left of each panel. When a protein-coding family was functionally annotated according to its similarity with the CDD, that annotation is listed in the legend. Otherwise a unique identification number is shown (e. g. Family 591).

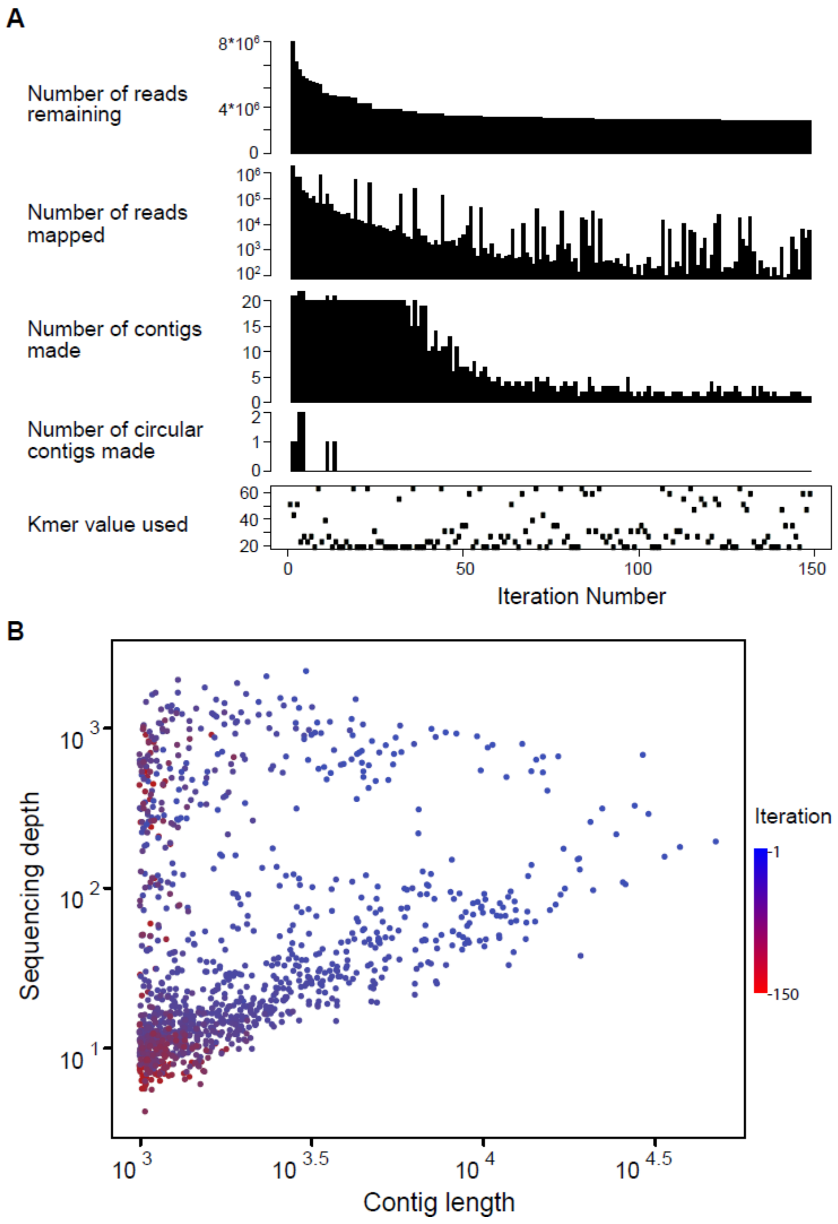

Figure 4-5. Optimized iterative de Bruijn graph assembly of 107 viral metagenomic sequences. A) Summary of run statistics for each iteration of the assembly, in which reads mapping to newly assembled contigs were removed at each iteration. The horizontal axis indicates the iteration number. For each of those iterations, the vertical axes indicate the number of reads remaining at the end of the iteration, the number of reads mapped during that iteration, the number of contigs made, the number of circular contigs made, and the optimal kmer chosen for that iteration. B) Characteristics of contigs by iteration of assembly. Each point is a contig with a length shown on the horizontal axis, depth of the assembly is shown on the vertical axis, and the iteration at which it was assembled indicated by color. The contigs that were assembled at earlier cycles (shown with bluer points) are generally longer and more deeply sequenced.

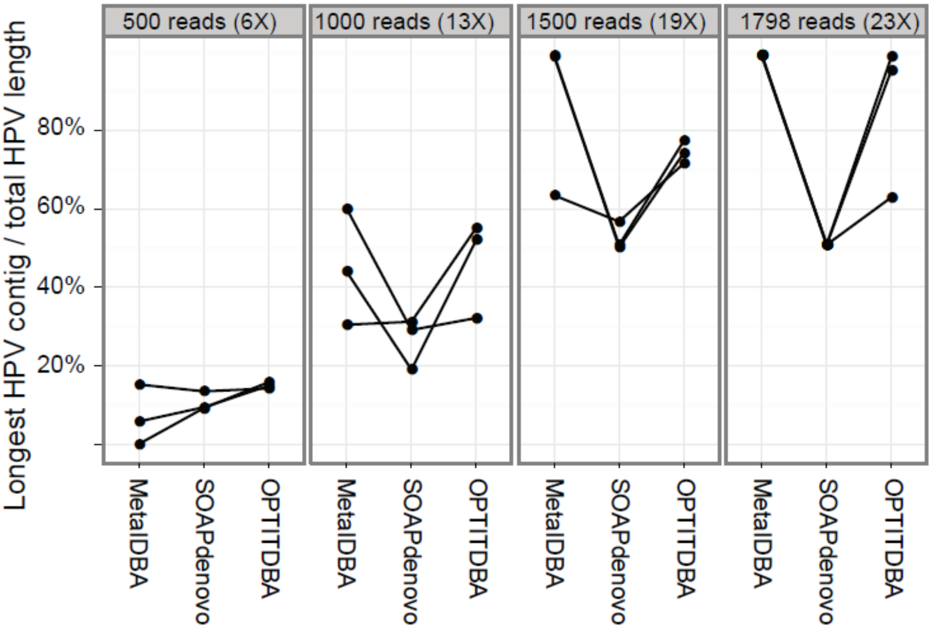

Figure 4-6 Comparison of assembly methods by known genome reconstruction. Shotgun sequences from HPV Type 6b were extracted from one dataset and added back to another dataset lacking HPV in varying amounts, as indicated in the grey boxes above each plot. The success of HPV reconstruction was measured as the length of the longest HPV-matching contig as a proportion of the total HPV length (vertical axis). The horizontal axis indicates the three assembly methods used. Three independent random samples were created for each level of coverage, and the assemblies using the same dataset are connected with a line.

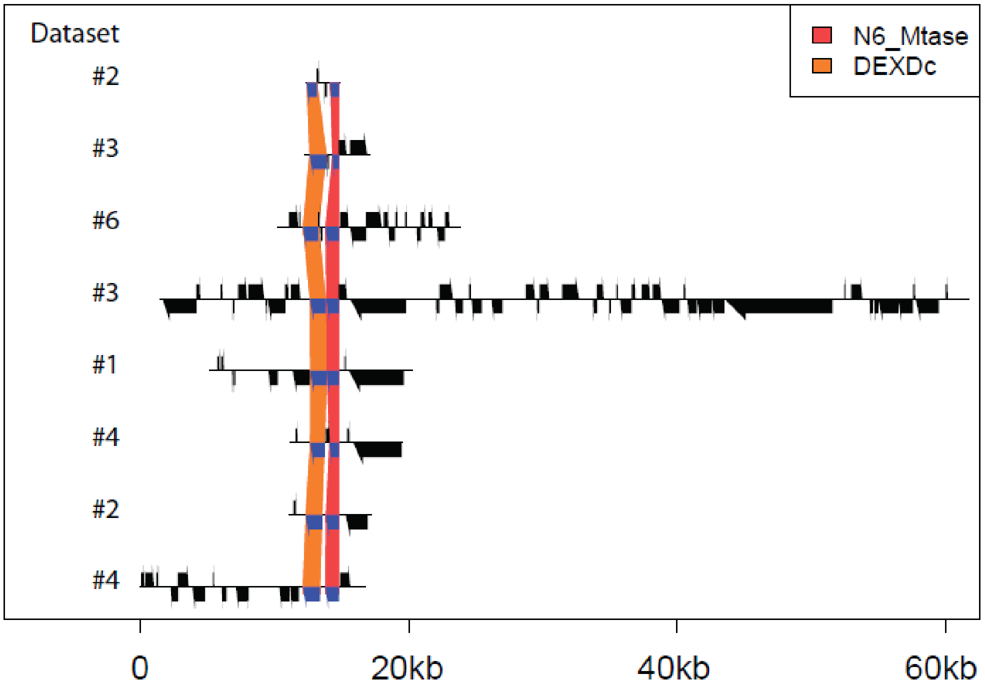

Figure 4-7. One additional example of phage cassette. Contigs are shown as horizontal black lines, ORFs on those contigs are shown by black arrows above and below those lines, and the organization of those ORFs into protein-coding families is shown with colored boxes. The subject that each contig was assembled from is shown on the left of each panel. When a protein-coding family was functionally annotated according to its similarity with the CDD, that annotation is listed in the legend.

4.6 Tables

Table 4-1. ORFs in families and cassettes.

	Dataset						
	1	**2**	**3**	**4**	**5**	**6**	**Total**
Contigs	8403	13258	4755	6067	4375	2415	39273
ORFs	9507	4508	3143	5618	4009	2232	29007
ORFs in families	5980	3056	2139	3648	2825	1527	16944
ORFs in cassettes	118	135	106	116	107	74	651

The number and proportion of ORFs predicted in each dataset that belong to protein-coding families (i.e. are not unique), and/or belong to cassettes (groups of protein-coding families that are found on the same set of contigs.

Table 4-2. Contigs and reads that form cassettes.

Contig Criteria	Contigs	Reads that align to those contigs
With at least 1 ORF	10032	31883951
With at least 1 ORF family	7024	29697888
With at least 1 cassette	326	5886117

The number of contigs, and the number of reads that align to those contigs, that contain at least 1 ORF, more than 1 ORF, at least 1 ORF family, and/or at least 1 cassette. The percentage of the total number of reads that align to contigs with at least 1 ORF is shown in parentheses.

Table 4-3. Assembly statistics.

		Optimized iterative assembly pipeline (OPTITDBA)				SOAPdenovo			
Subject	Number of reads	Number of contigs	Longest Contig (bp)	N50 (bp)	Circular	Number of contigs	Longest Contig (bp)	N50 (bp)	Circular
1	10000000	3516	58746	2981	22	1565	108497	5649	3
2	10000000	1400	42939	3415	10	960	93814	7605	2
3	10000000	977	60257	3986	8	487	45986	5289	1
4	10000000	1617	44975	3661	16	805	40190	6165	2
5	10000000	1150	47449	4561	16	463	117255	13073	2
6	5754268	588	77186	5772	10	340	47138	9053	2

Note: All contigs shorted than 1,000bp are omitted from the above table.

4.7 References

1. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65.
2. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, et al. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466: 334-338.
3. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, et al. (2011) The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21: 1616-1625.
4. Kingsford C, Schatz MC, Pop M (2010) Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics 11: 21.
5. Charuvaka A, Rangwala H (2011) Evaluation of short read metagenomic assembly. BMC Genomics 12 Suppl 2: S8.
6. Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT (2011) Individual genome assembly from complex community short-read metagenomic datasets. ISME J.
7. Pignatelli M, Moya A (2011) Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. PLoS One 6: e19984.
8. Rohwer F (2003) Global phage diversity. Cell 113: 141.
9. Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, et al. (2008) Viral diversity and dynamics in an infant gut. Res Microbiol 159: 367-373.
10. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185: 6220-6223.
11. Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD (2012) Hypervariable loci in the human gut virome. Proc Natl Acad Sci U S A.
12. Petrov VM, Ratnayaka S, Nolan JM, Miller ES, Karam JD (2010) Genomes of the T4-related bacteriophages as windows on microbial genome evolution. Virol J 7: 292.
13. Rohwer F, Edwards R (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol 184: 4529-4535.
14. Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci U S A 108: 17486-17491.
15. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, et al. (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7: 828-836.
16. Hoffmann KH, Rodriguez-Brito B, Breitbart M, Bangor D, Angly F, et al. (2007) Power law rank-abundance models for marine phage communities. FEMS Microbiol Lett 273: 224-228.
17. Lima-Mendez G, Toussaint A, Leplae R (2011) A modular view of the bacteriophage genomic space: identification of host and lifestyle marker modules. Res Microbiol 162: 737-746.
18. Lucchini S, Desiere F, Brussow H (1999) Comparative genomics of Streptococcus thermophilus phage species supports a modular evolution theory. J Virol 73: 8647-8656.
19. Botstein D (1980) A theory of modular evolution for bacteriophages. Ann N Y Acad Sci 354: 484-490.
20. Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, et al. (1978) The nucleotide sequence of bacteriophage phiX174. J Mol Biol 125: 225-246.
21. Rao VB, Black LW (2005) DNA packaging in bacteriophage T4. In: Catalano C, editor. Viral Genome Packaging Machines. Georgetown, TX: Landes Bioscience. pp. 40-58.

22. Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95: 315-327.
23. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, et al. (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19: 1117-1123.
24. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-829.
25. Li R, Zhu H, Ruan J, Qian W, Fang X, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20: 265-272.
26. Zhang W, Chen J, Yang Y, Tang Y, Shang J, et al. (2011) A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PLoS One 6: e17915.
27. Lin Y, Li J, Shen H, Zhang L, Papasian CJ, et al. (2011) Comparative studies of de novo assembly tools for next-generation sequencing technologies. Bioinformatics 27: 2031-2037.
28. Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A 98: 9748-9753.
29. Peng Y, Leung HC, Yiu SM, Chin FY (2011) Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics 27: i94-101.
30. Peng Y, Leung HC, Yiu SM, Chin FY (2010) IDBA - A Practical Iterative de Bruijn Graph De Novo Assembler. Research in Computational Molecular Biology 6044: 426-440.
31. Ng TF, Willner DL, Lim YW, Schmieder R, Chau B, et al. (2011) Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. PLoS One 6: e20579.
32. Willner D, Furlan M, Schmieder R, Grasis JA, Pride DT, et al. (2011) Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. Proc Natl Acad Sci U S A 108 Suppl 1: 4547-4553.
33. Botstein D, Herskowitz I (1974) Properties of hybrids between Salmonella phage P22 and coliphage lambda. Nature 251: 584-589.
34. Hendrix RW, Roberts JW, Stahl FW, Weisberg RA (1983) Lambda II. Cold Spring Harbor: Cold Spring Harbor Laboratory.
35. Hershey AD (1971) The Bacteriophage Lambda. Cold Spring Harbor Laboratory, New York: Cold Spring Harbor Press.
36. Wilgus GS, Mural RJ, Friedman DI, Fiandt M, Szybalski W (1973) Lambda imm lambda-434: a phage with a hybrid immunity region. Virology 56: 46-53.
37. Leiman PG, Shneider MM (2012) Contractile tail machines of bacteriophages. Adv Exp Med Biol 726: 93-114.
38. Sauer RT, Yocum RR, Doolittle RF, Lewis M, Pabo CO (1982) Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. Nature 298: 447-451.
39. Ohlendorf DH, Anderson WF, Lewis M, Pabo CO, Matthews BW (1983) Comparison of the structures of cro and lambda repressor proteins from bacteriophage lambda. J Mol Biol 169: 757-769.
40. Ptashne M (1986) A Genetic Switch. Cambridge, Massachusetts: Cell and Blackwell Scientific Press.
41. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
42. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. Nat Protoc 4: 470-483.

43. Pruitt K, Brown G, Tatusova T, Maglott D (2011) The Reference Sequence (RefSeq) Database. In: McEntyre J, Ostell J, editors. The NCBI Handbook. Bethesda, MD: National Center for Biotechnology Information.
44. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, et al. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39: D225-229.
45. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23: 673-679.
46. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: 421.
47. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431-432.
48. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460-2461.

CHAPTER 5 – CONCLUSION AND FUTURE STUDIES

One of the most striking findings of this dissertation work is that the viruses of the human gut are extremely diverse at the nucleotide level. In Chapter 3 [1] twelve humans' viromes were sequenced deeply, and the only virus that had been previously identified was the Human Papillomavirus. While the remaining bacteriophage genomes were novel, I show in Chapter 4 [2] that they contain recognizable homology at the levels of protein sequence and gene organization. Moreover, when viromes are compared between humans, there is little overlap of these novel species; however, temporal sampling reveals that an individual's gut virome is relatively stable, such that a majority of the genotypes are predicted to be shared by samples collected days, weeks, or months apart [3,4]. However, these quantitative predictions of overlap between viral communities have only been achieved through a simulation approach [5] in which shotgun reads are compared to each other, but not a common reference genome, so the identity of the constituent viruses is unknown. The observation that each human's gut virome is composed of previously unsequenced phages is consistent with the high rate at which bacteriophages isolated from the environment appear to be novel [6]. This remarkably large degree of genomic novelty could be explained by two non-exclusive models: a large global gene pool and/or a high rate of nucleotide evolution. It may be that the total number of viral species worldwide vastly exceeds the current extent of sequencing efforts, and it could also be that those viral species change rapidly, such that they are not recognizable at the nucleotide level after a relatively short period of time. Either of these explanations, or some combination of the two, would result in the observed high rate of genomic novelty among newly isolated bacteriophages.

It is not known whether the co-evolution of bacteria and phage in the environment can be expected *a priori* to result in a high rate of nucleotide substitution. While co-evolutionary "arms race" dynamics have been observed in culture [7], they may be attenuated by complex population structure [8] and spatial heterogeneity [9] in the environment. One possible mechanism of nucleotide substitution is selection by CRISPR arrays, which acquire novel spacers quickly but can be evaded by single nucleotide substitutions [10]. It is not known how any of these proposed mechanisms impact the evolution of bacteriophages. In order to address the question of how bacteriophages evolve with their hosts, it is essential to monitor natural communities as they change over time. I demonstrate in Chapter 4 that Illumina sequencing can be optimized for viral communities, yielding large contigs that explain a large fraction of the data (the raw reads). Using this approach, it would be possible to condense sequences generated from a single environment over multiple timepoints into a single set of consensus genomes. By comparing the similarity of reads at different timepoints to those consensus genomes, one could directly measure the substitution rate of viruses in the environment, gaining insight into the nature of bacteriophage evolution.

The question that drives our study of the human virome is that of how these viruses contribute to the overall function of the human microbiome and ultimately impact human health. The field has generated a number of intriguing hypotheses that deserve to be tested: that phage alter bacterial functionality by horizontal transfer of DNA, that phage limit bacterial abundance and promote bacterial diversity through "kill-the-winner" dynamics, and that phage drive the evolution of the bacterial 'accessory' pan-genome as a reservoir for resistance genes. Due to the complex nature of this community, a wide variety of experimental techniques may be required to address these questions, some of them only recently developed. For example, the colonization of germ-free animals with known collections of bacteria has been a useful way to test specific predictions regarding the mammalian microbiome [11]. Reyes, Rohwer, Gordon, and colleagues used the same technique to show that lysogens are induced by a variety of environmental signals in the mouse intestine [4]. By colonizing mice with pure and genetically characterized collections of bacteria and phage, one could study *in vivo* dynamics with the knowledge of which specific pairs were interacting, with the added ability to test mechanistic hypotheses via genetic manipulation of those strains. Another technique that has been used to link phage to their host is single-cell isolation [12,13], which can be combined with high-throughput sequencing to yield complete genome sequences [14]. Not surprisingly, this method can also capture viral genomes that are also associated with those cells [15]. One caveat of this technique is that physical separation of single cells may yield viral particles that are not capable of carrying out a complete cycle of infection, but when carried out in a large scale a consistent pattern of association may provide sufficient evidence for an infective relationship. Moreover, these data may provide evidence to test whether infection by phage is relatively common or rare in the human microbiome, which has strong implications for all of the models of population dynamics and evolution developed thus far.

Even with a complete genome sequence in hand, a great deal about can remain unknown about the genetic functionality it encodes. This fact was demonstrated by an expression cloning screen for antibiotic resistance genes using DNA generated from the human microbiome, which identified a large proportion of genes that had not been previously identified as such [16]. A similar approach could help to identify a large number of bacteriophage functions whose presence is inferred, but not known, by cloning and expressing viral DNA. For example, bacteriophage must encode a mechanism for cell lysis and exit, and the genes responsible could be identified through a screen for clones that cause cell death. In addition, it would be reasonable to propose that bacteriophage encode a wide variety of genes that exist to counteract bacterial immunity (such as restriction-modification systems or CRISPRs). For any immunity system that can be stably reconstructed in a culture setting, phage resistance genes could be found in the same manner. In addition to helping us better annotate phage genomes, identifying these negative regulators of immunity may provide useful reagents for the mechanistic study of such systems. In every case, by screening for genes among DNA from the human virome, we would gain a better understanding of the selective determinants of that system.

In this dissertation I examined a broad range of questions concerning the human gut virome. This study took advantage of high-throughput sequencing technology, which has allowed unprecedentedly large communities of microbes to be characterized on a genomic level. However, it is clear that increasingly powerful sequencing techniques will only do so much to help us understand this community. Further exploration will require a wide variety of experimental techniques, including (and likely not limited to) controlled colonization of mammals, single-cell sequencing, functional genetic screens, and genetic manipulation in culture. We now have some idea of mechanisms that could be important about the viruses in the gut: horizontal gene transfer, “kill-the-winner,” lysogeny, a molecular “arms race,” rapid nucleotide evolution, etc. The next task is to determine which of those are most important, playing a significant role in establishing and maintaining the human microbiome. It is my hope that this dissertation has laid a foundation for the study of the human gut virome and will enable further discoveries that advance our knowledge of human health and microbial ecology.

References

1. Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD (2012) Hypervariable loci in the human gut virome. Proc Natl Acad Sci U S A 109: 3962-3966.
2. Minot S, Wu GD, Lewis JD, Bushman FD Conservation of gene cassettes among diverse viruses of the human gut. Submitted.

3. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, et al. (2011) The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21: 1616-1625.
4. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, et al. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466: 334-338.
5. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. (2006) The marine viromes of four oceanic regions. PLoS Biol 4: e368.
6. Hatfull GF, Hendrix RW (2011) Bacteriophages and their genomes. Curr Opin Virol 1: 298-303.
7. Wichman HA, Millstein J, Bull JJ (2005) Adaptive molecular evolution for 13,000 phage generations: a possible arms race. Genetics 170: 19-31.
8. Hall AR, Scanlan PD, Morgan AD, Buckling A (2011) Host-parasite coevolutionary arms races give way to fluctuating selection. Ecol Lett 14: 635-642.
9. Forde SE, Thompson JN, Bohannan BJ (2004) Adaptation varies through space and time in a coevolving host-parasitoid interaction. Nature 431: 841-844.
10. Bhaya D, Davison M, Barrangou R (2011) CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation. Annual Review Genetics, Vol 45 45: 273-297.
11. Faith JJ, McNulty NP, Rey FE, Gordon JI (2011) Predicting a human gut microbiota's response to diet in gnotobiotic mice. Science 333: 101-104.
12. Lasken RS (2007) Single-cell genomic sequencing using Multiple Displacement Amplification. Curr Opin Microbiol 10: 510-516.
13. Kvist T, Ahring BK, Lasken RS, Westermann P (2007) Specific single-cell isolation and genomic amplification of uncultured microorganisms. Appl Microbiol Biotechnol 74: 926-935.
14. Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo MJ, Dupont CL, et al. (2011) Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. Nat Biotechnol 29: 915-921.
15. Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, et al. (2011) Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science 332: 714-717.
16. Sommer MO, Dantas G, Church GM (2009) Functional characterization of the antibiotic resistance reservoir in the human microflora. Science 325: 1128-1131.