2016

Covalent DNA Modifications in Phage and Bacterial Dynamics

Alexandra Bryson
University of Pennsylvania, bryson.alexandra@gmail.com

Covalent DNA Modifications in Phage and Bacterial Dynamics

Abstract
The microorganisms on and in the human body play a significant role in health and disease; however, little is known about how the interactions between these complex communities affect our wellbeing. This study examines how bacteria and phage interact through bacterial nucleases that restrict infection, such as restriction enzymes and CRISPR systems, and the covalent DNA modifications that neutralize them. Multiple targeted nucleases equip bacteria with an innate immune response against phage, and CRISPR systems provide an adaptive immune response. I report three main studies. 1) To study the human gut microbiome and virome (comprised predominately of phage), we collected fecal samples from a healthy individual over four years. From the fecal samples, total bacterial DNA and DNA from purified virus like particles (VLPs) were sequenced using Illumina and Pacific Bioscience single-molecule real-time (SMRT) sequencing to yield information about genome sequences and covalent modifications. Using computational methods we identified seven bacterial contigs and one phage contig with CRISPR arrays targeting phage contigs. This suggests that both bacteria and phage use CRISPR systems to compete with other phage. 2) Covalent DNA modifications are known to block the nuclease activity of restriction enzymes, however it was unknown if they can block the nuclease activity of CRISPR systems. To address this, we test if the CRISPR-Cas9 system could target wild type T4 phage and two T4 mutants. Wild type T4 modifies all its cytosines to glycosylated hydroxymethylcytosine (glc-HMC), and the two mutant T4 phage contain either hydroxymethylcytosine (HMC) or unmodified cystosines (C). These tests confirmed that glc-HMC and HMC in high concentrations can block CRISPR-Cas9. 3) To explore interactions between bacteria and phage further, we used covalent DNA modification data to link bacteria and phage pairs from the human gut microbiome, based on the idea that phage and bacterial DNAs in the same cell have been exposed to the same DNA modifying enzymes and thus share modification patterns. Overall, 443 modified motifs were shared between phage and bacteria, suggesting many possible phage-host pairs. In our data, 73% of phage genomes and 56% of bacterial genomes contained motifs that were completely modified, highlighting how ubiquitous and important the roll of DNA modifications are. These data allowed us to begin to specify the extent and types of interactions between phage and bacteria in longitudinal data. This work explores the complex interactions between bacteria and phage, a crucial step in understanding how these organisms contribute to human health and disease.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Cell & Molecular Biology

First Advisor
Frederic D. Bushman

Keywords
CRISPR, DNA modifications, Human Microbiome, Human Virome, Phage Dynamics, T4 Phage

COVALENT DNA MODIFICATIONS IN PHAGE AND BACTERIAL DYNAMICS

Alexandra Lynn Bryson

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

2016

Supervisor of Dissertation

_____

Frederic Bushman

W. M. Measey Professor

Chair, Department of Microbiology

Graduate Group Chairperson

_____

Dan Kessler, Associate Professor of Cell and Developmental Biology

Dissertation Committee

Mark Goulian, Professor of Biology

Mecky Pohlschröder, Professor of Biology

Rahul Kohli, Assistant Professor of Medicine

Elizabeth Grice, Assistant Professor of Dermatology

Dedication

To Rick Bushman, the Bushman Lab members, and the MVP matriculating class of 2011.

ABSTRACT

COVALENT DNA MODIFICATIONS IN PHAGE AND BACTERIAL DYNAMICS

Alexandra Lynn Bryson

Dr. Frederic Bushman's Laboratory

The microorganisms on and in the human body play a significant role in health and disease; however, little is known about how the interactions between these complex communities affect our wellbeing. This study examines how bacteria and phage interact through bacterial nucleases that restrict infection, such as restriction enzymes and CRISPR systems, and the covalent DNA modifications that neutralize them. Multiple targeted nucleases equip bacteria with an innate immune response against phage, and CRISPR systems provide an adaptive immune response. I report three main studies. 1) To study the human gut microbiome and virome (comprised predominately of phage), we collected fecal samples from a healthy individual over four years. From the fecal samples, total bacterial DNA and DNA from purified virus like particles (VLPs) were sequenced using Illumina and Pacific Bioscience single-molecule real-time (SMRT) sequencing to yield information about genome sequences and covalent modifications. Using computational methods we identified seven bacterial contigs and one phage contig with CRISPR arrays targeting phage contigs. This suggests that both bacteria and phage use CRISPR systems to compete with other phage. 2) Covalent DNA modifications are known to block the nuclease activity of restriction enzymes, however it was unknown if they can block the

nuclease activity of CRISPR systems. To address this, we test if the CRISPR-Cas9 system could target wild type T4 phage and two T4 mutants. Wild type T4 modifies all its cytosines to glycosylated hydroxymethylcytosine (glc-HMC), and the two mutant T4 phage contain either hydroxymethylcytosine (HMC) or unmodified cystosines (C). These tests confirmed that glc-HMC and HMC in high concentrations can block CRISPR-Cas9. 3) To explore interactions between bacteria and phage further, we used covalent DNA modification data to link bacteria and phage pairs from the human gut microbiome, based on the idea that phage and bacterial DNAs in the same cell have been exposed to the same DNA modifying enzymes and thus share modification patterns. Overall, 443 modified motifs were shared between phage and bacteria, suggesting many possible phage-host pairs. In our data, 73% of phage genomes and 56% of bacterial genomes contained motifs that were completely modified, highlighting how ubiquitous and important the roll of DNA modifications are. These data allowed us to begin to specify the extent and types of interactions between phage and bacteria in longitudinal data. This work explores the complex interactions between bacteria and phage, a crucial step in understanding how these organisms contribute to human health and disease.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1: Introduction

1.1 The human gut virome and microbiome

The microorganisms on and in the human body play a significant role in health and disease. Commensal microbes help our immune systems develop properly, aid us in the breakdown and digestion of nutrients, and guard against pathogen invasion(1-5). However, dysbiosis (disease associated with microbial ecosystem shifts) among the microbiota has been linked to heart disease, inflammatory bowel disease, depression, autoimmune disease, and obesity(6-17).

The human gastrointestinal tract contains one of the most densely populated microbial communities in the human body, and recent advances in deep sequencing technologies reveal complex communities of bacteria and viruses living and interacting together. Phage, in addition to being the vast majority of viruses detected within the human gut, are also the most prevalent biological entities on Earth (estimated at 10^{31} virions) and known to dynamically regulate bacterial populations(2-5, 18-20). Studies of phage, since their independent discovery in the early 1910s by Twort and d'Herelle, helped establish fundamental principals in molecular biology and genetics(21, 22); however, once the groundwork was laid, those fields quickly transitioned to studying higher order model organisms. The recognition that phage play an integral role in the healthy

human microbiome has brought about a recent resurgence of phage biology in the context of the human virome.

The human gut virome can be studied by isolating virus like particles (VLPs) from fecal samples. Fecal samples offer a non-invasive way to study microbiota of the lower gastrointestinal tract. Fecal samples are homogenized, passed though a 0.2μm filter (to remove human and bacterial cells) then treated with DNAse and RNAse to remove DNA that is not contained in a VLP. The VLPs are then broken down with proteinase K and their nucleic acids are extracted for sequencing. The sequenced viral genomes are assembled *de novo* using computational methods. These proposed genomes are referred to as contigs. RNA and DNA viruses known to infect human cells have been found using these methods, however the majority of isolated viruses are DNA phage. RNA viruses found in healthy, fecal samples are typically plant pathogens, which are thought to have been ingested with food and are passing though transiently(23).

To study bacteria of the human gut, total DNA is extracted from fecal samples. The DNA can then be prepared and sequenced using two different methods. The first is shotgun metagenomics, where the total, extracted DNA is sequenced using high-throughput technology. The resulting sequencing reads are filtered to remove those mapping to the human genome. The remaining sequence reads are assembled *de novo* using computational methods. Roughly

96% of resulting contigs belong to bacteria, with the remaining 4% belonging to phage. The second method is 16S ribosomal DNA (16S rDNA) amplicon sequencing. The gene encoding 16S ribosomal RNA is ubiquitous and highly conserved among bacteria. The 16S rDNA also contains nine hyper-variable regions that can be used to distinguish between bacterial species. Primers targeting the conserved segments of the 16s rDNA are used to amplify the gene. These amplicons are sequenced and the variable regions they contain are used to determine which bacteria are present within a fecal sample.

Understanding the dynamics between bacteria and phage in healthy individuals is a crucial step in learning to treat disorders involving a dysbiotic microbial communities. However determining which phage infect which bacteria is a difficult challenge. The human gut provides a unique niche for bacteria and phage to coexist that has yet to be replicated in a laboratory setting. Previous efforts have failed to grow gut phage *in vitro* (outside of gut) despite being able to observe the phage propagating within a gnotobiotic mouse gut*(24)*. Matching phage-host pairs is further complicated because phage can sometimes infect multiple bacterial hosts. Traditionally, there have been many phage known to bind a specific receptor, and they do not bind other structures that vary only slightly from their host receptor. Research done by Jensen et al. suggests that current methods of isolating phage artificially select for phage with a specific host

and that phage with a broad-range of hosts are more prevalent than previously thought(25). Jansen et al. even identified phage that plaque on both *Escherichia coli* and *Sphaerotilus natans*, which belong to different classes, *Gammaproteobaceria* and *Betaproteobacteria* respectively(25).

In this body of work we seek to identity phage/ bacterial-hosts pairs and understand how they interact, particularly through host-parasite competition based on nucleases and protection of DNA with covalent modifications.

1.2 Phage influence on bacterial communities

There are two major mechanisms by which phage influence the survival of bacterial communities. Phage can exert a predatory pressure killing their host or providing advantageous genetic information though lateral gene transfer. Most of what is known about the control of bacterial populations by phage comes from environmental studies, particularly in lakes and seawater(26-31). The literature indicates that dynamics between phage and bacteria including predation rates can vary significantly. Work done in Germany's Lake Plussee demonstrated that bacterial mortality from viral lysis varies within the same body of water based on the steep temperature and oxygen gradients of the lake. In the warm and oxic epilimnion, 8% to 42% of bacterial deaths are attributed to viral lysis, whereas in

the cooler, anoxic hypolimnion 88% to 94% of bacterial mortality is attributed to viral lysis(28).

Phage can provide survival advantages to their bacterial hosts by existing in a lysogenic state as a prophage or episome. Research conducted on bacteria living in the gulf of Mexico suggests that 0.07% to 4.4% of the bacteria harbor a prophage(27). Other studies suggest higher rates of lysogenized bacteria in different environments (32). A filamentous phage, f327, is though to help *Pseudoalteromonas* survive in the Artic Sea by enhancing motility and chemotaxis of their host(32). Other marine phage provide their hosts with genes to carry out carbon and phosphate metabolism as well as photosynthesis(33, 34).

Phage are also known to transfer antibiotic resistance genes between bacteria(35). The mobility of antibiotic resistance is of particular concern to heath care providers as multidrug resistant pathogens are become more prevalent.

1.3 Restriction-Modification Systems

Bacteria use restriction-modification systems to protect themselves from foreign DNA, such as the DNA injected by a phage. To distinguish self from non-self DNA, bacteria use enzymes to add covalent modifications to their own genomes at specific nucleotide motifs known as recognition sites. The number of

nucleotides in a recognition site is typically 4 to 8 bases, and the exact motif sequences vary among restriction/modification systems. Restriction enzymes scan DNA for recognition sites. If a recognition site is unmodified the restriction enzyme binds to the motif and cleaves the DNA. If the recognition site is modified, the restriction enzyme will not be able to bind and cleave the DNA due to steric hindrance. The first evidence for restriction-modification systems arose in the 1950s from work done by Salvador Luria and Giuseppe Bertani when they found phage λ could grow on some *E. coli* strains but not others(36, 37). In the 1960s Werner Arber and Matthew Meselson demonstrated this restriction of phage growth was caused by enzymatic cleavage (38-40).

Restriction enzymes are traditionally classified into four types. Type I enzymes (originally discovered to target phage λ) contain multiple subunits that preform restriction-and-modification functions within in one enzyme. Their restriction cut sites are random and distant from the recognition motif. Type II restriction enzymes cut at defined positions close to or within a specific DNA sequence motif. HindII was the first Type II restriction enzyme to be characterized, when in 1970, Hamilton Smith observed phage P22 DNA degrading in the presence of Haemophilus influenza cell extract (41, 42). Type III enzymes are combination restriction-and-modification complexes that cleave outside of their recognition sequence and required two motifs in opposite

orientations. In contrast to Types I-III, Type IV enzymes recognize and cleave modified (typically methylated) DNA such as McrBC in *E. coli*.

Phage evolved to avoid restriction modification systems by commandeering their host's modifying enzymes or alternatively by encoding their own. Replicating phage are thereby exposed to the same modifying enzymes as their host and are likely to share modification patterns. In this work, we hypothesize that phage and their hosts can be linked matching DNA modification patterns between the two.

1.4 CRISPR

The Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) proteins form an adaptive immune system used by roughly 50% of bacteria to confer resistance against phage, plasmid-mediated lateral gene transfer, and other mobile genetic elements (MGE)(43, 44). The CRISPR-Cas system allows for acquisition and storage/memory of phage and plasmid-derived sequences that can be used to identify future infections based on sequence homology. The CRISPR locus is comprised of an AT-rich leader sequence followed by an array of uniform repeat sequences alternating with unique segments of viral or MGE sequences (spacers). The spacers, averaging 32bp in length, are acquired from an invading virus or plasmid and integrated into the CRISPR locus near the 5' leader end creating a linear history of invading

DNA. The region in the viral or plasmid genome that matches the spacer is referred to as the protospacer(45).

Upstream of the CRISPR array lies the *Cas* genes. Different organisms contain distinctive clusters of *Cas* genes, which fall into two major classes that can be further broken down into five types and sixteen subtypes(46, 47). CRISPR systems requiring multiple proteins for interference belong to class I, while CRISPR systems requiring only one protein for interference are a part of class II. Cas proteins process transcribed CRISPR arrays into smaller fragments called CRISPR RNAs (crRNA). Cas proteins in complex with a crRNA form a functional unit that base pairs with and cleaves DNA containing the homologous protospacer sequence and a specific protospacer-adjacent motif (PAM) of ~3-7 bases(45, 48, 49). The PAM sequence plays an important role in allowing the CRISPR system to distinguish between self and non-self DNA. The absence of PAMs in CRISPR arrays prevents the CRISPR system from degrading its own DNA.

Two modes of spacer acquisition have been reported for class I CRISPR systems: native and primed(50-57). So far, only native spacer acquisition has been reported in class II CRISPR systems(58-61). In native spacer acquisition, proteins Cas1 and Cas2 are necessary and sufficient to obtain new spacers from foreign DNA. Primed spacer acquisition requires Cas1, Cas2, Cas3, Cascade (a

complex of Cas proteins) and a “priming” spacer targeting an existing protospacer. Priming enhances spacer acquisition 10 to 20 fold compared to native acquisition(62). Spacer acquisition through priming increases when the priming spacer is an imperfect match to the protospacer or when the protospacer does not have the correct PAM(62). This is thought to counteract against phage acquiring point mutations to escape the sequence homology requirements of CRISPR targeting.

The class II, type II systems are among the best-characterized CRISPR systems largely because of their simple four *Cas* gene structure (*Cas9, Cas1, Cas2, and Csn2 or Cas4*). *Cas9* is the signature gene within these systems. The Cas9 protein aids in crRNA biogenesis and cleavage of target DNA. These systems also require a trans-activating crRNA (tracrRNA)(63). In crRNA biogenesis, the tracrRNA hybridizes with the transcribed CRISPR array forming dsRNA that is then cleaved by RNase III (a bacterial host enzyme) to release the individual crRNAs. Cas9, the crRNA, and the tracrRNA then together form a functional unit to cleave target DNA. Cas9 has an HNH domain and a RuvC-like domain, which cleave protospacer targets that match the crRNA(63).

Phage use several mechanisms to evade CRISPR systems. Point mutations acquired in the PAM or protospacer sequences allow phage to escape the sequence homology requirement of CRISPR systems. Exact homology between

the first eight bases of the spacer and protospacer (known as the seed sequence) are crucial for the CRISPR system to recognize and degrade its target. Outside of the seed sequence, multiple point mutations can accumulate before inhibiting the CRISPR system(64). To date, three phage encoded anti-CRISPR (Acr) proteins have been reported: AcrF1 (from phage JBD30), AcrF2 (from phage D3112), and AcrF3 (from phage JBD5)(64, 65). These proteins block the DNA-binding activity of the CRISPR-Cas complex and bind the Cas3 helicase-nuclease so it cannot be recruited to target DNA bound by the CRISPR-Cas complex(65). One study evaluated if a single adenine-N6-methyl DNA modification within a phage protospacer could block CRISPR activity, since methyl groups are known to block other nucleases(66). The single adenine-N6-methyl group was not sufficient to block CRISPR activity; however, phage are known to have multiple unique and unusual DNA modifications (many of which are significantly larger than methyl groups). Thus, here we seek to determine if larger DNA modifications can block the CRISPR-Cas9 system. We also search for CRISPR systems targeting phage within the human gut microbiome for an improved understanding of how the phage and bacteria communities are intertwined.

1.5 Covalent DNA modifications in Phage

At least ten covalent DNA modifications have been reported in phage such as α-gluThy in SP10, 5-dhpUra in SP15, 5-mCyt in χP12, and 2-n-Ade in S2L(67). Despite the remarkable bases seen in phage, very few phage have been analyzed for covalent modifications, and no studies have looked at DNA modifications of phage in the human microbiome. DNA modifications of bacteria within the human microbiome were previously evaluated in one paper yielding information about two *Bacteroides* genomes(68). In this body of work, we evaluate the modification profiles of the complete microbiome, including both phage and bacteria, providing insight into the frequency and breadth of DNA modifications of the human gut microbiome. Additionally, we demonstrate a new function for several different phage DNA modifications.

1.6 References

1. **Rakoff-Nahoum S, Paglino J, Eslami-Varzaneh F, Edberg S, Medzhitov R.** 2004. Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. Cell **118:**229-241.
2. **Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI.** 2005. Host-bacterial mutualism in the human intestine. Science **307:**1915-1920.
3. **Ley RE, Peterson DA, Gordon JI.** 2006. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell **124:**837-848.

4. **O'Hara AM, Shanahan F.** 2006. The gut flora as a forgotten organ. EMBO Rep **7:**688-693.

5. **Zoetendal EG, Vaughan EE, de Vos WM.** 2006. A microbial world within us. Mol Microbiol **59:**1639-1650.

6. **Emoto T, Yamashita T, Kobayashi T, Sasaki N, Hirota Y, Hayashi T, So A, Kasahara K, Yodoi K, Matsumoto T, Mizoguchi T, Ogawa W, Hirata KI.** 2016. Characterization of gut microbiota profiles in coronary artery disease patients using data mining analysis of terminal restriction fragment length polymorphism: gut microbiota could be a diagnostic marker of coronary artery disease. Heart Vessels doi:10.1007/s00380-016-0841-y.

7. **Patel T, Bhattacharya P, Das S.** 2016. Gut microbiota: an Indicator to Gastrointestinal Tract Diseases. J Gastrointest Cancer doi:10.1007/s12029-016-9820-x.

8. **Gensollen T, Iyer SS, Kasper DL, Blumberg RS.** 2016. How colonization by microbiota in early life shapes the immune system. Science **352:**539-544.

9. **Rao K, Higgins PD.** 2016. Epidemiology, Diagnosis, and Management of Clostridium difficile Infection in Patients with Inflammatory Bowel Disease. Inflamm Bowel Dis doi:10.1097/MIB.0000000000000793.

10. **Bultman SJ.** 2016. Interplay between diet, gut microbiota, epigenetic events, and colorectal cancer. Mol Nutr Food Res doi:10.1002/mnfr.201500902.

11. **Backhed F, Ding H, Wang T, Hooper LV, Koh GY, Nagy A, Semenkovich CF, Gordon JI.** 2004. The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A **101:**15718-15723.

12. **Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI.** 2005. Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A **102:**11070-11075.

13. **Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI.** 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444:**1027-1031.

14. **Ordovas JM, Mooser V.** 2006. Metagenomics: the role of the microbiome in cardiovascular diseases. Curr Opin Lipidol **17:**157-161.

15. **Zheng P, Zeng B, Zhou C, Liu M, Fang Z, Xu X, Zeng L, Chen J, Fan S, Du X, Zhang X, Yang D, Yang Y, Meng H, Li W, Melgiri ND, Licinio J, Wei H, Xie P.** 2016. Gut microbiome remodeling induces depressive-like behaviors through a pathway mediated by the host's metabolism. Mol Psychiatry doi:10.1038/mp.2016.44.

16. **Emge JR, Huynh K, Miller EN, Kaur M, Reardon C, Barrett KE, Gareau MG.** 2016. Modulation of the Microbiota-Gut-Brain Axis by Probiotics in a Murine Model of Inflammatory Bowel Disease. Am J Physiol Gastrointest Liver Physiol doi:10.1152/ajpgi.00086.2016**:**ajpgi 00086 02016.

17. **Moloney RD, Johnson AC, O'Mahony SM, Dinan TG, Greenwood-Van Meerveld B, Cryan JF.** 2016. Stress and the Microbiota-Gut-Brain Axis in Visceral Pain: Relevance to Irritable Bowel Syndrome. CNS Neurosci Ther **22:**102-117.

18. **Whitman WB, Coleman DC, Wiebe WJ.** 1998. Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A **95:**6578-6583.

19. **Gorski A, Dabrowska K, Switala-Jelen K, Nowaczyk M, Weber-Dabrowska B, Boratynski J, Wietrzyk J, Opolski A.** 2003. New insights into the possible role of bacteriophages in host defense and disease. Med Immunol **2:**2.

20. **Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermudez-Humaran LG, Gratadoux JJ, Blugeon S, Bridonneau C, Furet JP, Corthier G, Grangette C, Vasquez N, Pochart P, Trugnan G, Thomas G, Blottiere HM, Dore J, Marteau P, Seksik P, Langella P.** 2008. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proc Natl Acad Sci U S A **105:**16731-16736.

21. **D'Herelle F.** 2007. On an invisible microbe antagonistic toward dysenteric bacilli: brief note by Mr. F. D'Herelle, presented by Mr. Roux. 1917. Res Microbiol **158:**553-554.

22. **Twort FW.** 1936. Further Investigations on the Nature of Ultra-Microscopic Viruses and their Cultivation. J Hyg (Lond) **36:**204-235.

23. **Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW, Hibberd ML, Liu ET, Rohwer F, Ruan Y.** 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol **4:**e3.

24. **Reyes A, Wu M, McNulty NP, Rohwer FL, Gordon JI.** 2013. Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. Proc Natl Acad Sci U S A **110:**20236-20241.

25. **Jensen EC, Schrader HS, Rieland B, Thompson TL, Lee KW, Nickerson KW, Kokjohn TA.** 1998. Prevalence of broad-host-range lytic bacteriophages of Sphaerotilus natans, Escherichia coli, and Pseudomonas aeruginosa. Appl Environ Microbiol **64:**575-580.

26. **Bongiorni L, Magagnini M, Armeni M, Noble R, Danovaro R.** 2005. Viral production, decay rates, and life strategies along a trophic gradient in the North Adriatic Sea. Appl Environ Microbiol **71:**6644-6650.

27. **Weinbauer MG, Suttle CA.** 1996. Potential significance of lysogeny to bacteriophage production and bacterial mortality in coastal waters of the gulf of Mexico. Appl Environ Microbiol **62:**4374-4380.

28. **Weinbauer MG, Hofle MG.** 1998. Significance of viral lysis and flagellate grazing as factors controlling bacterioplankton production in a eutrophic lake. Appl Environ Microbiol **64:**431-438.

29. **Wieltschnig C, Fischer UR, Kirschner AK, Velimirov B.** 2003. Benthic bacterial production and protozoan predation in a silty freshwater environment. Microb Ecol **46:**62-72.

30. **Simek K, Pernthaler J, Weinbauer MG, Hornak K, Dolan JR, Nedoma J, Masin M, Amann R.** 2001. Changes in bacterial community composition and dynamics and viral mortality rates associated with enhanced flagellate grazing in a mesoeutrophic reservoir. Appl Environ Microbiol **67:**2723-2733.

31. **Pradeep Ram AS, Sime-Ngando T.** 2010. Resources drive trade-off between viral lifestyles in the plankton: evidence from freshwater microbial microcosms. Environ Microbiol **12:**467-479.

32. **Yu ZC, Chen XL, Shen QT, Zhao DL, Tang BL, Su HN, Wu ZY, Qin QL, Xie BB, Zhang XY, Yu Y, Zhou BC, Chen B, Zhang YZ.** 2015. Filamentous phages prevalent in Pseudoalteromonas spp. confer

properties advantageous to host survival in Arctic sea ice. ISME J **9:**871-881.

33. **Sharon I, Alperovitch A, Rohwer F, Haynes M, Glaser F, Atamna-Ismaeel N, Pinter RY, Partensky F, Koonin EV, Wolf YI, Nelson N, Beja O.** 2009. Photosystem I gene cassettes are present in marine virus genomes. Nature **461:**258-262.

34. **Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW.** 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci U S A **108:**E757-764.

35. **Boyd EF.** 2012. Bacteriophage-encoded bacterial virulence factors and phage-pathogenicity island interactions. Adv Virus Res **82:**91-118.

36. **Bertani G, Weigle JJ.** 1953. Host controlled variation in bacterial viruses. J Bacteriol **65:**113-121.

37. **Luria SE, Human ML.** 1952. A nonhereditary, host-induced variation of bacterial viruses. J Bacteriol **64:**557-569.

38. **Dussoix D, Arber W.** 1962. Host specificity of DNA produced by Escherichia coli. II. Control over acceptance of DNA from infecting phage lambda. J Mol Biol **5:**37-49.

39. **Lederberg S, Meselson M.** 1964. Degradation of Non-Replicating Bacteriophage DNA in Non-Accepting Cells. J Mol Biol **8:**623-628.

40. **Meselson M, Yuan R.** 1968. DNA restriction enzyme from E. coli. Nature **217:**1110-1114.

41. **Kelly TJ, Jr., Smith HO.** 1970. A restriction enzyme from Hemophilus influenzae. II. J Mol Biol **51:**393-409.

42. **Smith HO, Wilcox KW.** 1970. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. J Mol Biol **51:**379-391.

43. **Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P.** 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science **315:**1709-1712.

44. **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.

45. **Yosef I, Goren MG, Qimron U.** 2012. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res **40:**5569-5576.

46. **Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV.** 2015. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol **13:**722-736.

47. **Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV.** 2011. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol **9:**467-477.

48. **Shah SA, Erdmann S, Mojica FJ, Garrett RA.** 2013. Protospacer recognition motifs: mixed identities and functional diversity. RNA Biol **10:**891-899.

49. **Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C.** 2009. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology **155:**733-740.

50. **Erdmann S, Garrett RA.** 2012. Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. Mol Microbiol **85:**1044-1056.

51. **Erdmann S, Le Moine Bauer S, Garrett RA.** 2014. Inter-viral conflicts that exploit host CRISPR immune systems of Sulfolobus. Mol Microbiol **91:**900-917.

52. **Fineran PC, Gerritzen MJ, Suarez-Diez M, Kunne T, Boekhorst J, van Hijum SA, Staals RH, Brouns SJ.** 2014. Degenerate target sites mediate rapid primed CRISPR adaptation. Proc Natl Acad Sci U S A **111:**E1629-1638.

53. **Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, Edgar R, Qimron U, Sorek R.** 2015. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature **520:**505-510.

54. **Richter C, Dy RL, McKenzie RE, Watson BN, Taylor C, Chang JT, McNeil MB, Staals RH, Fineran PC.** 2014. Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Res **42:**8516-8526.

55. **Savitskaya E, Semenova E, Dedkov V, Metlitskaya A, Severinov K.** 2013. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. RNA Biol **10:**716-725.

56. **Shmakov S, Savitskaya E, Semenova E, Logacheva MD, Datsenko KA, Severinov K.** 2014. Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic Acids Res **42:**5907-5916.

57. **Swarts DC, Mosterd C, van Passel MW, Brouns SJ.** 2012. CRISPR interference directs strand specific spacer acquisition. PLoS One **7:**e35888.

58. **Westra ER, Brouns SJ.** 2012. The rise and fall of CRISPRs--dynamics of spacer acquisition and loss. Mol Microbiol **85:**1021-1025.

59. **Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, Marraffini LA.** 2015. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature **519:**199-202.

60. **Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S.** 2008. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J Bacteriol **190:**1390-1400.

61. **Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadan AH, Moineau S.** 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature **468:**67-71.

62. **Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E.** 2012. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun **3:**945.

63. **Sorek R, Lawrence CM, Wiedenheft B.** 2013. CRISPR-mediated adaptive immune systems in bacteria and archaea. Annu Rev Biochem **82:**237-266.

64. **Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, van der Oost J, Brouns SJ, Severinov K.** 2011. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U S A **108:**10098-10103.

65. **Bondy-Denomy J, Garcia B, Strum S, Du M, Rollins MF, Hidalgo-Reyes Y, Wiedenheft B, Maxwell KL, Davidson AR.** 2015. Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. Nature **526:**136-139.

66. **Dupuis ME, Villion M, Magadan AH, Moineau S.** 2013. CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. Nat Commun **4:**2087.

67. **Warren RA.** 1980. Modified bases in bacteriophage DNAs. Annu Rev Microbiol **34:**137-158.

68. **Leonard MT, Davis-Richardson AG, Ardissone AN, Kemppainen KM, Drew JC, Ilonen J, Knip M, Simell O, Toppari J, Veijola R, Hyoty H, Triplett EW.** 2014. The methylome of the gut microbiome: disparate Dam methylation patterns in intestinal Bacteroides dorei. Front Microbiol **5:**361.

CHAPTER 2: Rapid evolution of the human gut virome

Samuel Minot, Alexandra Bryson, Christel Chehoud, Gary Wu, James Lewis, and Frederic Bushman

This paper is published: Proc Natl Acad Sci USA. 2013 Jul 23;110(30):12450-5. doi: 10.1073/pnas.1300833110. Epub 2013 Jul 8.
PMID:23836644

2.1 Contributions

My contributions to this paper are the CRISPR analysis and MetaPhlan analysis. I generated Table 1, Figure 5, Figure S2, and Figure S3.

2.2 Abstract

Humans are colonized by immense populations of viruses, which metagenomic analysis shows are mostly unique to each individual. To investigate the origin and evolution of the human gut virome, we analyzed the viral community of one adult individual over 2.5 y by extremely deep metagenomic sequencing (56 billion bases of purified viral sequence from 24 longitudinal fecal samples). After assembly, 478 well-determined contigs could be identified, which are inferred to correspond mostly to previously unstudied bacteriophage genomes. Fully 80% of these types persisted throughout the duration of the 2.5-y study, indicating long-term global stability. Mechanisms of base substitution, rates of accumulation, and the amount of variation varied among viral types.

Temperate phages showed relatively lower mutation rates, consistent with replication by accurate bacterial DNA polymerases in the integrated prophage state. In contrast, Microviridae, which are lytic bacteriophages with single-stranded circular DNA genomes, showed high substitution rates (>10(-5) per nucleotide each day), so that sequence divergence over the 2.5-y period studied approached values sufficient to distinguish new viral species. Longitudinal changes also were associated with diversity-generating retroelements and virus-encoded Clustered Regularly Interspaced Short Palindromic Repeats arrays. We infer that the extreme interpersonal diversity of human gut viruses derives from two sources, persistence of a small portion of the global virome within the gut of each individual and rapid evolution of some long-term virome members.

2.3 Introduction

There are an estimated 1031 viral particles on earth, and human feces contain at least 109 virus-like particles per gram (1–3). Many of these are identifiable as viruses that infect bacteria (bacteriophages), but the great majority remains unidentified. Even today, gut virome samples taken from different human individuals still yield mostly novel viruses (4–8), and only a small minority of viral ORFs resembles previously studied genes (7). Bacteriophages are of biomedical importance because of their ability to transmit genes to their bacterial hosts, thereby conferring increased pathogenicity, antibiotic resistance, and perhaps

new metabolic capacity (4, 5, 9, 10). Despite their importance, the forces diversifying bacteriophage genomes in human hosts have not been studied in detail. Humans show considerable individual variation in the bacterial lineages present in their guts (11–13); this variation likely is one reason for the differences in their phage predators (5–8, 14). The large differences in phage populations among individuals also may be influenced by within individual viral evolution. To investigate the origin and nature of human viral populations, we carried out a detailed study of a single human gut viral community. Ultra-deep longitudinal analysis of DNA sequences from the viral community, combined with characterization of the host bacteria, revealed rapid change over time and begins to specify some of the mechanisms involved.

2.4 Results

2.4.1 Sample Collection, Viral Purification, and DNA Sequencing.

Stool samples (n = 24) were collected from a healthy male at 16 time points spread over 884 days (Fig. 1A). For eight of the time points, two separate samples taken 1 cm apart were purified and sequenced independently to allow estimation of within-time point sample variation. Virus-like particles were extracted by sequential filtration, Centricon ultrafiltration, nuclease treatment, and solvent extraction. Purified viral DNA was subjected to linear amplification using

Φ29 DNA polymerase, after which quantitative PCR showed that bacterial 16S sequences were reduced to less than 10 copies per nanogram of DNA, and human sequences were reduced to below 0.1 copies per nanogram, the limit of detection. Paired-end reads then were acquired using Illumina HiSeq sequencing, yielding more than 573 million reads (Q ≥ 35; mean read length, 97.5 bp), with 15–39 million reads per sample (Table S1). No attempt was made to study gut RNA viruses, which also are known to exist, although some samples were dominated by abundant plant RNA viruses ingested with food (15). Sequence reads from each sample were first assembled individually using MetaIDBA (16). When reads were aligned back onto contigs generated within each sample, only 71% of reads could be aligned. Improved contigs then were generated using a hybrid assembly method combining all samples, taking advantage of the fact that viruses that are rare at one time point may be abundant at another. After this step, 97.6% of the reads could be aligned to contigs, allowing assessment of within-contig diversity. Rarefaction (collector's curve) analysis showed that the detection of these contigs was saturated at 20-fold coverage (median, 82-fold); from the purification results, we infer these contigs to be mostly or entirely DNA viruses (Fig. 1C). Sixty contigs assembled as closed circles (ranging in size from 4–167 kb), an indication of probable completion of these genome sequences, providing an estimate of the viral population size and composition in unprecedented detail. One circular genome

was sequenced independently using the Sanger method and was confirmed to have the structure predicted from the Solexa/Illumina data (SI Methods). The abundance of each contig at each time point was measured by the proportion of reads that aligned to it, normalized to the length of each contig. The correlation coefficient between replicate samples from the same time point was at least 0.99, indicating a high degree of reproducibility (Fig. S1).

2.4.2 Viral Groups Detected

Taxonomic analysis of these contigs indicated recovery of Microviridae, Podoviridae, Myoviridae, and Siphoviridae, but contigs with taxonomic attributions were a minority, only 13%, emphasizing the enormous sequence variation present in bacteriophages. Microviridae (the group including ΦX174) predominated, but this predominance could be a consequence of favored amplification by Φ29 polymerase of the small circular genomes that characterize this group. The most abundant contigs were mostly retained over the duration of the experiment. Because there are many possible pairwise comparisons between time points, distances between time points analyzed (Fig. 2A, x-axis) were compared with Jaccard index values (Fig. 2A, y-axis), which score shared membership, over all of the possible pairwise comparisons of time points. On average, more than 80% of contigs were found in common between the time

points separated by 850 d (points at the right side of the plot), the longest time intervals compared. No contigs corresponded to known viruses infecting eukaryotic cells. To investigate the possible presence of eukaryotic cell viruses further, we aligned the raw sequence reads to the National Center for Biotechnology Information viral genome database. Thirty-two percent coverage was seen for Gyrovirus in one time point, and pooling reads over all time points yielded 42% coverage. Gyrovirus is a Circovirus genus with very small genome sizes (~2.3 kb) recently reported to infect humans (17). However, the number of reads aligning was modest (10 total), and in no case did both reads of the paired end reads align. Because of these results, and in addition to the small target size, we believe that the detection of Gyrovirus is uncertain. All other animal cell virus genomes showed <10% coverage, so detection is questionable. The rarity of eukaryotic virus sequences is typical of gut virome samples from healthy individuals (4–6, 18, 19), emphasizing the tremendous size of the bacteriophage populations of the gut.

2.4.3 Host Bacteria

To allow tracking of the bacterial hosts, for three of the time points we also sequenced a total of 5.2 Gb of DNA purified from unfractionated stool, which yields predominantly bacterial DNA. Attribution of bacterial lineages using MetaPhlAn (20) showed members of the Bacteroides and Firmicutes phyla to be

the most abundant community members (Fig. S2). Bacterial community membership and taxonomic proportions showed only modest variation over time.

Longitudinal Base Substitution in Viral Contigs. The depth of sequence information available and the quality of the viral contigs allowed a detailed assessment of the rates of accumulation of base substitutions. For each viral contig at each time point, the extent of nucleotide polymorphism was determined by aligning reads within each sample. The extent of nucleotide substitution then was compared for each contig between time points, and substitution frequencies were correlated with biological features. Substitution rates varied with viral family and replication style (Fig. 2B). The Microviridae showed the highest substitution rate ($P < 0.004$). Microviridae package ssDNA genomes, which have been reported to show higher mutation rates than dsDNA genomes in vitro (21, 22), and this study confirms this result in a human host. The Podo-, Myo-, and Siphoviridae all package dsDNA genomes and showed lower substitution frequencies. The lowest substitution rates were seen for temperate bacteriophage ($P = 0.015$, Kruskal–Wallace test), which can integrate into the host bacterial genome. Temperate phages were identified as contigs satisfying at least one of three criteria: (i) encoding integrase genes, (ii) homologs present as prophage in sequenced bacterial genomes, or (iii) annotated as resembling previously studied temperate phage (5). When integrated, temperate phage DNA

is replicated by high-fidelity bacterially encoded machinery, and temperate phage also may undergo fewer lytic replication cycles; both result in lower substitution rates. Temperate bacteriophage showed significantly lower substitution rates even when Microviridae were excluded from the comparison ($P = 0.044$). There was no significant difference in rates among the families of large dsDNA viruses. The four contigs with the highest rate of nucleotide substitution were all members of the Microviridae (Fig. 3A). The main variant for each lineage showed 1–4% nucleotide substitutions over the course of the experiment (more than one substitution per 105 nt per day). An alternative explanation for these high substitution rates could be the immigration of new closely related Microviridae into the community. To investigate this possibility, we reconstructed the consensus genome for the four contigs at multiple time points and aligned them against a large collection of Microviridae genomes. In every case the contig consensus sequences for all time points clustered closely together (Fig. 3B), arguing against immigration of related Microviridae and supporting the model of continuous substitution in long-term viral residents. A detailed analysis of the longitudinal change of each SNP detected (Fig. 4) showed that a complex community of variants was present at most time points and that new SNPs accumulated on this background. Substitutions could accumulate either at a steady rate or in an episodic fashion, for example in response to a change in selective pressure. Linear modeling of substitution rates versus time showed

correlation coefficients of 0.91–0.99, consistent with generally steady substitution rates, although with considerable sample-specific fluctuations. Longitudinal sequence divergence in major variants predicted from the Illumina data were confirmed using Sanger sequencing for two of the Microviridae (described in SI Methods).

2.4.4 Clustered Regularly Interspaced Short Palindromic Repeats Targeting Phage Genomes

One force driving phage sequence variation is the bacterial Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) system (23–26). DNA sequences from invaders such as bacteriophage or plasmids are incorporated as spacers into arrays in the bacterial genome. Transcription of such arrays allows the CRISPR spacer RNAs to be incorporated into nucleoprotein effector complexes that target the destruction of sequence-complementary invaders. Thus, bacteriophages are under pressure to mutate to evade degradation by the CRISPR system, as has been documented in model systems (23–25, 27). The deep analysis of viral sequences presented here, together with the shotgun metagenomic analysis of host bacterial sequences, allowed the influence of the CRISPR system in vivo to be studied in detail. A total of 34 types of CRISPR repeat sequences and their associated spacers were identified in the bacterial metagenomic sequence. Table 1 shows that several of

these spacers targeted contigs from the virome sequence data. Up to 28 spacers could be identified targeting a single viral contig. The CRISPR-targeted viral contigs were analyzed for their relative abundance over time. No simple pattern was seen relating the presence of CRISPR spacers to the relative abundance over all of the targeted viruses. In one case, a viral contig accumulated a base substitution in a CRISPR target site, and the mutant contig increased in abundance while the original contig declined, suggestive of CRISPR evasion by mutation (Fig. S3). Of the CRISPR arrays identified, four appeared to be encoded by temperate phage. Several previous reports also have documented phage-encoded CRISPR arrays (5, 28, 29). An analysis of longitudinal variation in phage CRISPR arrays would be useful, but uncertainties in reconstructing arrays from short read data precluded a detailed analysis. For the CRISPR array with the most sequence coverage (contig 117), we found that the entire collection of spacers was replaced over the time series studied. The phage-encoded CRISPR array on phage contig 117 encoded spacers that targeted four different phage contigs from our study (Fig. 5 shows one example). We previously reported another example from a different subject of a phage-encoded CRISPR spacer targeting a different phage in the same virome sample (5). Evidently phages commonly use CRISPR systems to compete with one another.

2.4.5 Identifying Phage Hosts

Characterization of bacteriophage populations by sequencing typically does not specify the host bacterial species, leaving important gaps in our understanding of phage–host interactions. Analysis of CRISPRs, however, provides a means of connecting phage–host pairs (Table 1). Three previously sequenced bacterial genomes, from Ruminococcus bromii, Eubacterium siraeum, and Bacteriodes fragilis, contain CRISPR repeats that were found here linked to spacers matching virome contigs from this study (contig 232_308, contig 132_57, and contig 111_52, respectively), allowing us to infer that these phages infect these three bacteria in the subject studied. In another approach to associating phage–host pairs, phage sequences annotated as integrated prophages in sequenced bacterial genomes could be recognized that resembled our newly sequenced phage contigs, thereby also specifying potential hosts (4–6). Bacterial lineages identified as harboring phage from the virome analysis included Bacteroides fragilis, Eubacterium siraeum, Ruminococcus bromii, Blautia hansenii, and Lachnospiraceae, all of which were found to be present in metagenomic sequence analysis of total stool DNA (Fig. S2). Overall, 19 of the phage contigs sequenced here could be associated with bacterial hosts by at least one of the two approaches (Table S2), although for the great majority the hosts remain unknown.

2.4.6 Longitudinal Sequence Variation Driven by Diversity-Generating Retroelements

Another force diversifying bacteriophage genomes are diversity-generating retroelements (DGRs), which are reverse transcriptase-based systems that introduce mutations at adenines in specific repeated sequences using a copy–paste targeting mechanism (6, 30–33). We analyzed the viral contigs described here to investigate whether DGRs were detectably active within the human gut. DGRs were identified by searching contigs for regions that matched three criteria: (i) they contained proteincoding regions resembling reverse transcriptases, (ii) they encompassed short repeat regions containing mismatches in adenine positions, and (iii) they contained hypervariable regions. Of the 20 contigs with both a reverse transcriptase and an adenine mismatched repeat, six were associated with hypervariable regions (located no more than 100 bp away; Table S3) and were selected for further study. As was found previously, hypervariation was directed toward asparagine AAY codons in genes encoding either predicted C-type lectin or Ig-superfamily proteins (6, 30–33). We next asked whether any of the DGRs were detectably active over the time series studied. The longest gap between sample collections was 22 mo, so to maximize sensitivity we asked whether the hypervariable regions had evolved to become clearly different over this time interval. Of the two hypervariable regions with sufficient longitudinal coverage for analysis, one (contig 42) showed change over

the 22-mo time period, and change was greater than for samples closer together in time ($P < 0.0001$) or for pairs of samples from the same time point ($P < 0.0001$). For the second (contig d03-2), we did not obtain evidence for longitudinal variation. We conclude that one of our DGRs was active in the human gut. For the others, it is unclear whether they were inactive or whether we did not have enough sequence coverage to detect activity. Analysis showed that DGR containing contigs were not among the most variable, highlighting the local nature of DGR variation and emphasizing the contributions of other mechanisms. The possibility that some of the DGRs were inactive raises the question of whether the mutagenic activity might be regulated in the human host.

2.5 Discussion

Here we report a study of longitudinal variation in the human gut virome and some of the mechanisms responsible for change over time. Loss and acquisition of viral types was uncommon: Fully ~80% of viral forms persisted over the 2.5-y time course studied, as is consistent with previous studies of shorter duration (4–6). Most viral contigs showed diversity within each time point and accumulated variation over time. Temperate DNA phages showed relatively modest rates of variation compared with lytic phage, as is consistent with temperate phage DNA replication by accurate bacterial polymerases in the prophage state, and potentially fewer total rounds of replication. In contrast, the

strictly lytic ssDNA Microviridae showed up to 4% substitutions in the major variants present over the time period studied. DGRs showed high diversity in variable repeat regions, and one was detectably active over the time series studied. CRISPR arrays encoded in viral genomes also were associated with longitudinal variation. Thus, multiple mechanisms contributed to viral sequence variation, and our data provide a detailed picture of their relative contributions. This study did not yield any clear examples of known DNA viruses infecting animal cells. Rare reads did align to genomes of animal cell viruses, but it is uncertain whether these alignments represent true detection of these viruses or rare regions of homology between animal cell viruses and phages. In contrast, several studies have reported frequent detection of animal cell viruses in metagenomic analysis of stool DNA from humans and other primates, raising the question of how these studies differed. One observation is that samples from sick individuals (34, 35) or SIV-infected macaques (36) have yielded animal cell viruses more frequently than samples from healthy controls. Some of these studies did not attempt to analyze bacterial viruses, instead using bioinformatic filters to extract animal cell viruses from complex sequence mixtures, potentially leading to an under-appreciation of the size of the phage populations. Thus, our data emphasize that in the healthy human gut bacterial viruses are much more numerous than animal cell viruses, although it remains possible that some of our contigs with no database matches correspond to previously unknown viruses

infecting human cells. Given the findings reported here, we can return to the question of why human gut viromes differ so greatly among human individuals. One factor must be the differences in bacterial populations in the guts of different humans. Many metagenomic studies emphasize that, although the human gut typically contains bacteria from only a few phyla, the bacterial strains are mostly different between individuals (11–13). Phages can be highly selective for different bacterial lineages—indeed, phage sensitivity is used clinically to distinguish some bacterial strains (e.g., refs. 37 and 38)—likely explaining some of the differences in phage populations in different individuals. However, a second basis for the differences among individuals, highlighted in data reported here, is rapid within-host viral evolution. Microviridae lineages showed up to 4% substitution in the main variant over the 2.5-y period studied, consistent with laboratory experiments also showing high mutation rates for Microviridae (39). There is no single threshold of sequence identity accepted for splitting related viruses into separate species (40), but different Microviridae species specified by the International Committee on Taxonomy of Viruses show as little as 3.1% divergence (Table S4). Evidently the divergence seen here for Microviridae contigs 122_321 and 001_39 approaches the level sufficient for designation as speciation events. Extrapolating from these rates, our data suggest that multiple new viral species commonly will arise in the gut of a typical human over the course of a human life. Thus, part of the explanation for the extremely large

populations of gut viruses inferred from sequence information and for the extreme differences among individual humans appears to be rapid within-individual evolution of long-term viral residents.

2.6 Methods

Longitudinal stool samples were collected from a single healthy male adult under a protocol approved by the Internal Review Board of the Perelman School of Medicine at the University of Pennsylvania. Samples of viral particles were purified by filtration, Centricon ultrafiltration, and nuclease treatment, and then total DNA was extracted using the QIAamp DNA Stool kit. Sequence information was acquired using Illumina paired-end technology. Sequences were assembled by iterative deBruijn graph assembly using MetaIDBA, and contigs were combined using Minimo. Taxonomy was assigned using Blastp, ORFs were predicted using Glimmer, and bacterial taxa were called using Metaphlan. Oligonucleotides used in this study are presented in Table S5. All sequence information has been deposited at the National Center for Biotechnology Information. For further details see SI Methods.

2.8 Tables

Table 1

Table 1. CRISPR arrays from bacterial metagenomic sequence targeting viral contigs detected in this study

CRISPR	Organism hosting CRISPR	No. of spacers associated with repeat	Median spacer length (bp)	Viral contig targeted	No. of spacers matching viral contig
CRISPR-2	*Ruminococcus bromii* L2-63 (temperate phage)	64	30 (29–31)	232_308	1
CRISPR-3	Unknown	38	30 (21–33)	112_6	2
CRISPR-7	Unknown	64	36 (22–40)	051_116	1
				75	1
CRISPR-21	Unknown	59	35 (30–38)	111_52	4
CRISPR-31	*Eubacterium siraeum* V10Sc8a	110	37 (25–40)	132_57	1
CRISPR-32	*Eubacterium siraeum* V10Sc8a	230	37 (22–46)	132_57	27
CRISPR-37	*Bacteroides fragilis* NCTC 9343	32	30 (29–30)	111_52	1

Table S1

Table S1. Virus like particle sequence sample characteristics

Sampling day	Replicate	Read	Aligned reads	Percent aligned
0	1	20,312,322	18,331,267	90.25
0	2	24,435,114	22,289,142	91.22
180	1	24,875,744	24,691,054	99.26
181	1	23,959,436	23,841,743	99.51
182	1	15,505,820	15,208,373	98.08
182	2	16,812,218	16,706,701	99.37
183	1	17,774,454	17,264,914	97.13
184	1	19,893,608	19,702,285	99.04
851	1	25,101,704	25,084,614	99.93
852	2	27,714,314	27,697,853	99.94
852	1	28,434,716	28,387,692	99.83
852	2	29,751,810	29,692,193	99.80
853	1	15,903,684	15,896,899	99.96
853	2	25,144,920	25,123,536	99.92
854	1	29,634,128	29,623,353	99.96
855	1	22,257,952	22,255,753	99.99
879	1	16,071,666	16,070,799	99.99
879	2	16,405,346	16,402,036	99.98
880	1	21,052,128	21,046,503	99.97
880	2	19,138,984	19,136,500	99.99
881	1	29,389,988	29,372,515	99.94
881	2	35,751,748	35,724,342	99.92
882	1	29,211,340	29,205,227	99.98
883	1	39,238,778	39,234,718	99.99

Table S2

Table S2. Assignment of phage contigs to bacterial hosts

Contig	Length (bp)	Bacterial (GI)	Bacterial species	Connection
111_52	36,084	60491031	*Bacteroides fragilis* NCTC 9343	CRISPR
132_57	7,156	291556121	*Eubacterium siraeum* V10Sc8a	CRISPR
232_308	5,336	291541372	*Ruminococcus bromii* L2-63	CRISPR
111_107	5,222	224485451	*Bacteroides* sp. 2_1_7	Prophage
221_131	4,177	224485451	*Bacteroides* sp. 2_1_7	Prophage
231_217	5,118	224485451	*Bacteroides* sp. 2_1_7	Prophage
021_4	37,938	319644666	*Bacteroides* sp. 3_1_40A	Prophage
031_147	4,924	319644666	*Bacteroides* sp. 3_1_40A	Prophage
231_103	11,455	319644666	*Bacteroides* sp. 3_1_40A	Prophage
231_91	13,236	319644666	*Bacteroides* sp. 3_1_40A	Prophage
38	20,452	256402715	*Blautia hansenii* DSM 20583	Prophage
44	5,669	256402715	*Blautia hansenii* DSM 20583	Prophage
231_106	26,844	256402715	*Blautia hansenii* DSM 20583	Prophage
74	10,031	331640228	*Lachnospiraceae bacterium* 3_1_46FAA	Prophage
232_270	6,065	331640228	*Lachnospiraceae bacterium* 3_1_46FAA	Prophage
232_349	4,323	331640228	*Lachnospiraceae bacterium* 3_1_46FAA	Prophage
011_27	27,157	291541372	*Ruminococcus bromii* L2-63	Prophage
117	15,472	291541372	*Ruminococcus bromii* L2-63	Prophage
107	36,432	291541372	*Ruminococcus bromii* L2-63	Prophage

Shown are viral contigs assigned to bacterial hosts by both CRISPR spacer matches and annotation as prophage in sequenced genomes.

Table S3

Table S3. Variation in DGR contigs over time

Contig	Significant change over time	ORF length	CDD (hit - bit score - evalue)	Phyre2
231_106	No	381	164824 - MTD - 104–5e-25	Clec (MTD)
38	No	381	164824 - MTD - 108–2e-26	Clec (MTD)
42	Yes	351	32846 - FxsA - 28.2–3.7	Clec
032_43621	No	603	48198 - GlucD - 34.5–0.15	Clec (MTD)
166	No	592	145488 - Big_2 - 37.3–0.001	Ig superfamily (α-amylase)
90	No	365	164824 - MTD - 80.2–1e-16	Clec (MTD)

Contigs queried for significant variation and gene types affected. CDD, conserved domains database; MTD, major tropism determinant.

Table S4

Table S4. Nucleotide divergence among Microviridae from the International Committee on Taxonomy of Viruses

	Chp2	Alpha3	St-1	ID18	WA13	phiX174	G4	ID2 Moscow/ID/2001	Chp4	PhiCPG1	Chp3
Chlamydia phage Chp2		0.0	0.0	0.0	0.0	0.0	0.0	0.0	90.8	91.5	96.9
Enterobacteria phage alpha3	0.0		90.1	0.0	63.6	5.0	0.0	0.0	0.0	0.0	0.0
Enterobacteria phage St-1	0.0	90.1		0.0	62.2	5.0	0.0	0.0	0.0	0.0	0.0
Enterobacteria phage ID18	0.0	0.0	0.0		4.1	5.6	44.3	70.7	0.0	0.0	0.0
Enterobacteria phage WA13	0.0	63.6	62.2	4.1		5.3	0.0	3.9	0.0	0.0	0.0
Enterobacteria phage phiX174	0.0	5.0	5.0	5.6	5.3		0.0	6.4	0.0	0.0	0.0
Enterobacteria phage G4	0.0	0.0	0.0	44.3	0.0	0.0		47.8	0.0	0.0	0.0
Enterobacteria phage ID2 Moscow/ID/2001	0.0	0.0	0.0	70.7	3.9	6.4	47.8		0.0	0.0	0.0
Chlamydia phage 4	90.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0		94.8	90.6
Chlamydia phage PhiCPG1	91.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	94.8		91.0
Chlamydia phage 3	96.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	90.6	91.0	

Entries in the matrix show the identity between the isolates compared.

Table S5

Table S5. Oligonucleotides used in this study

Position relative to 122_321	Orientation	Name	Sequence	Comments
408	F	122_321_408_F	TTCGCTAGCCAACAGTCCTT	Sequencing
427	R	122_321_427_R	AAGGACTGTTGGCTAGCGAA	Sequencing/ amplify 122_321
496	F	122_321_496_F	TGTACTTCGGCAGCATTGAG	Sequencing/ amplify 122_321
918	F	122_321_918_F	CGCCGTTTGTCCGTAAGTAT	Sequencing
937	R	122_321_937_R	ATACTTACGGACAAACGGCG	Sequencing
987	F	122_321_987_F	AGGAGCAGTTGCGTTTCCTA	Sequencing
1,299	F	122_321_1299_F	AGAAGCAGCACCTTTTCCAA	Sequencing
1,318	R	122_321_1318_R	TTGGAAAAGGTGCTGCTTCT	Sequencing
2,154	F	122_321_2154_F	AGACCGGAGAATGTTCGATG	Sequencing
2,173	R	122_321_2173_R	CATCGAACATTCTCCGGTCT	Sequencing
3,238	F	122_321_3238_F	ATTTGGGGCGTGTATTACCA	Sequencing
3,257	R	122_321_3257_R	TGGTAATACACGCCCCAAAT	Sequencing
4,072	F	122_321_4072_F	CGGGGTTAATGCGTAAAGAA	Sequencing
4,091	R	122_321_4091_R	TTCTTTACGCATTAACCCCG	Sequencing
4,653	F	122_321_4653_F	GACGAGCATAAACACGAGCA	Sequencing
4,672	R	122_321_4672_R	TGCTCGTGTTTATGCTCGTC	Sequencing
6,121	F	122_321_6121_F	GGCACGAAAAGACCATTGTT	Sequencing
6,140	R	122_321_6140_R	AACAATGGTCTTTTCGTGCC	Sequencing

F, forward; R, reverse.

2.7 Figures

Figure 1

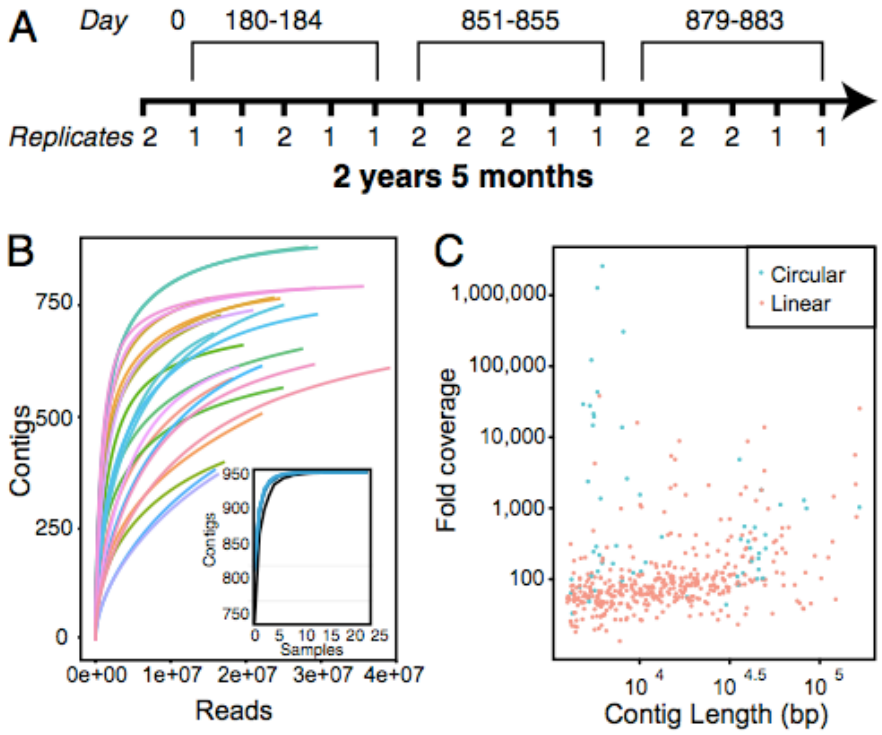


Fig. 1. Longitudinal analysis of the human gut virome from a single individual. (A) Timeline of sample collection. Note that at some time points, two separate portions of the stool sample, taken approximately 1 cm apart, were processed and sequenced independently to assess reproducibility. (B) Rarefaction analysis of sampling depth by number of reads; detection of each contig is scored as positive if 50% of the contig is covered by sequence reads. (Inset) Contig recovery. The x-axis is the number of samples included (black line: 2 million reads; blue line: 15 million reads). (C) Contig spectrum, relating the lengths of the contigs assembled in bas pairs (x-axis) to the depth of coverage (y-axis). Circular contigs are shown as blue and linear contigs as red.

Figure 2

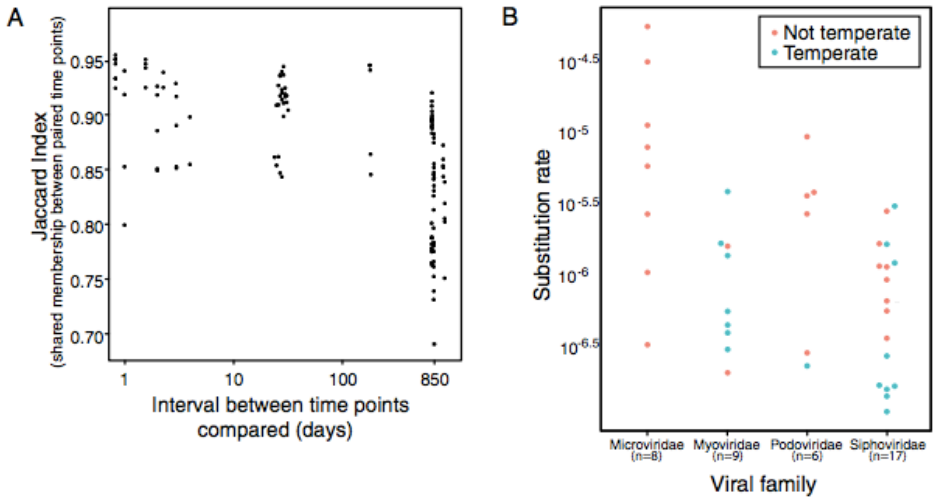


Fig. 2. Stability and change in the gut virome of the individual studied. (A) Conserved membership in the viral community over time intervals analyzed using the Jaccard index. Because many pairwise comparisons are possible between the 24 time points, we plotted shared membership for all pairs of time points as a function of the length of time between each pair. The x-axis shows the time interval between time points, and the y-axis shows shared membership in the two communities compared summarized using the Jaccard index. Perfect identity yields a value of 1, and complete divergence yields a value of 0. (B) Comparison of substitution rates among viral families. Temperate phages are shown in blue, and lytic phages are in red. The viral families studied are shown at the bottom; substitution rates on the y-axis are substitutions per base, per day. Only contigs with clear taxonomic attributions were analyzed; such contigs comprise a minority of all contigs.

Figure 3

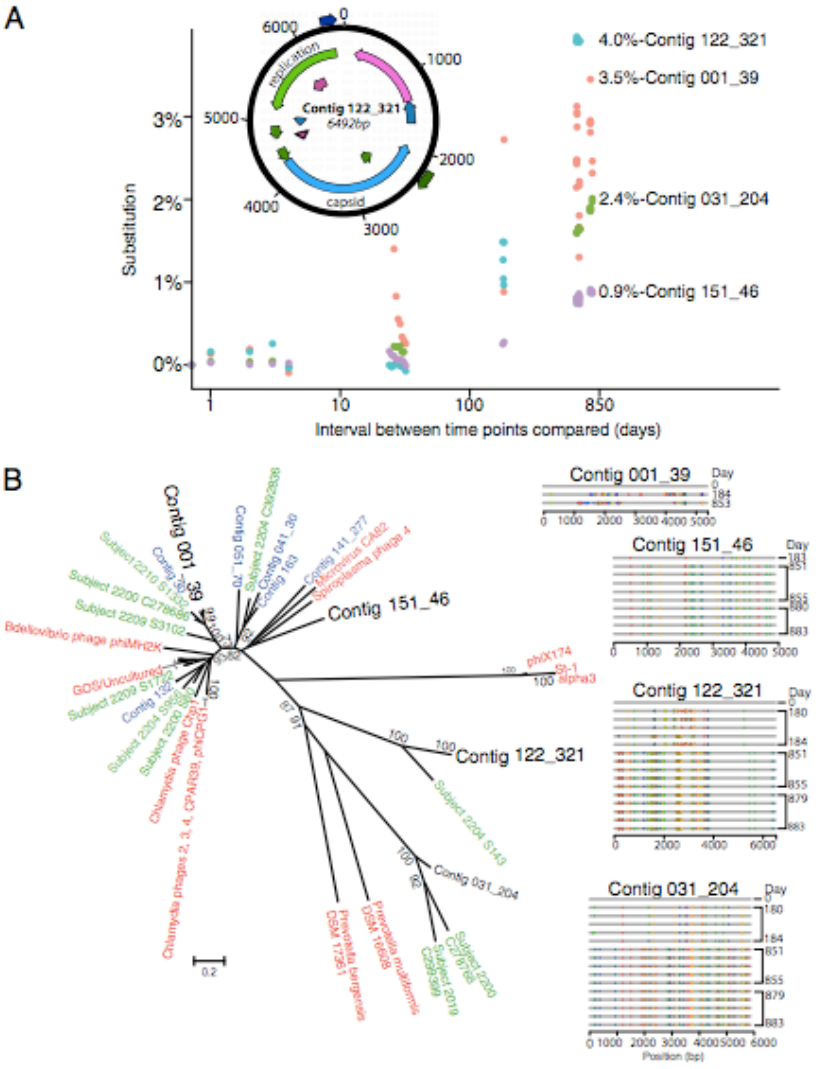


Fig. 3. Longitudinal DNA substitution in Microviridae. (A) Substitution rates in the four Microviridae genomes with the highest values measured. Because many pairwise comparisons are possible between the time points at which each virus was detected, the plot shows distances between time points on the x-axis and the percent substitution on the y-axis. The percent substitution values within each time point were subtracted from the between-time point values before the plot was constructed. Colors differentiate the four viruses studied. (Inset) The genome with the highest substitution rate (contig 122_321). (B) Phylogenetic tree of microphages detected in this and other studies. The four microphage contigs with the highest substitution rates observed in this study are shown in large black lettering. Database microphages are shown in red, microphages from ref. 6 are shown in green, and additional microphages identified in this study are shown in

blue. (Scale bar: the proportion of amino acid substitutions within the 919-aa major coat protein, which was aligned to make the tree.) Longitudinal maps of substitution accumulation are shown to the right. Note that all of the variations shown in the sequences to the right are plotted in the phylogenetic tree but are not visible because of the comparatively low divergence. Only time points with high-quality complete-genome assemblies are shown.

Figure 4

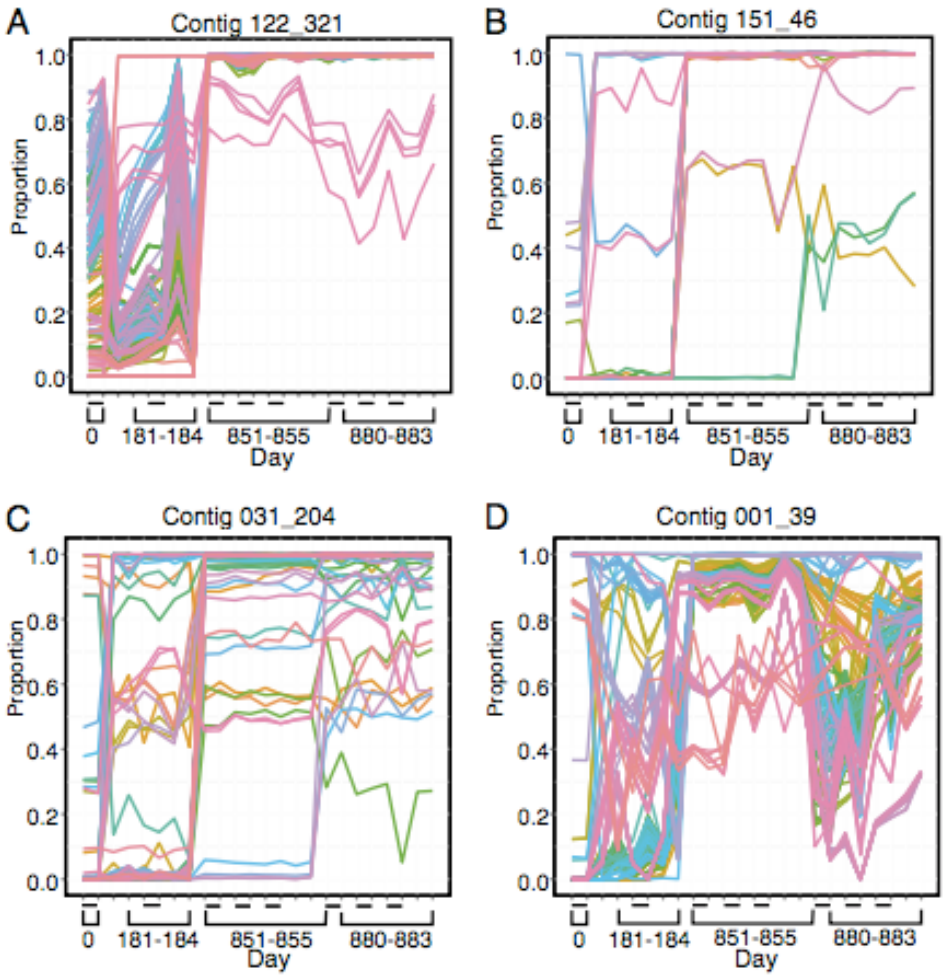


Fig. 4. Relative abundance of SNPs in four Microviridae genomes analyzed longitudinally. Contigs studied are marked above each figure panel. The x-axis shows elapsed time since the start of the study. The y-axis shows the relative proportion of each variant in the population. The dashes on the x-axis show replicate analysis of single time points, allowing assessment of within-time point variability. Only positions with SNPs that transitioned from minor (0.5) are plotted. The

colors are used to make the different positions easier to visualize. Panel labels A–D show data for the contigs indicated at the top of each panel.

Figure 5

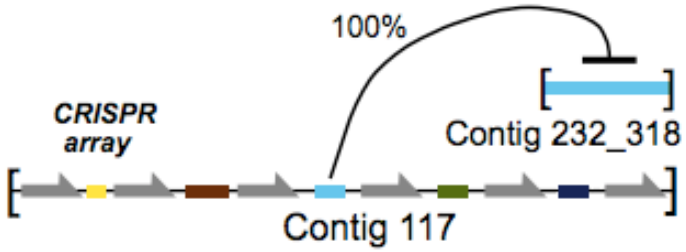


Fig. 5. A phage-encoded CRISPR array targeting another phage. The array shown (contig 117) was detected in the viral contig collection. Gray indicates CRISPR repeats, and colors indicate CRISPR spacers. The target contig (contig 102) also was identified and observed to be present at some of the same time points; three other contigs also were targeted by the CRISPR array in contig 117. The CRISPR array in viral contig 117 is closely similar to CRISPR-2 detected in the total stool metagenomic sequencing.

Figure S1

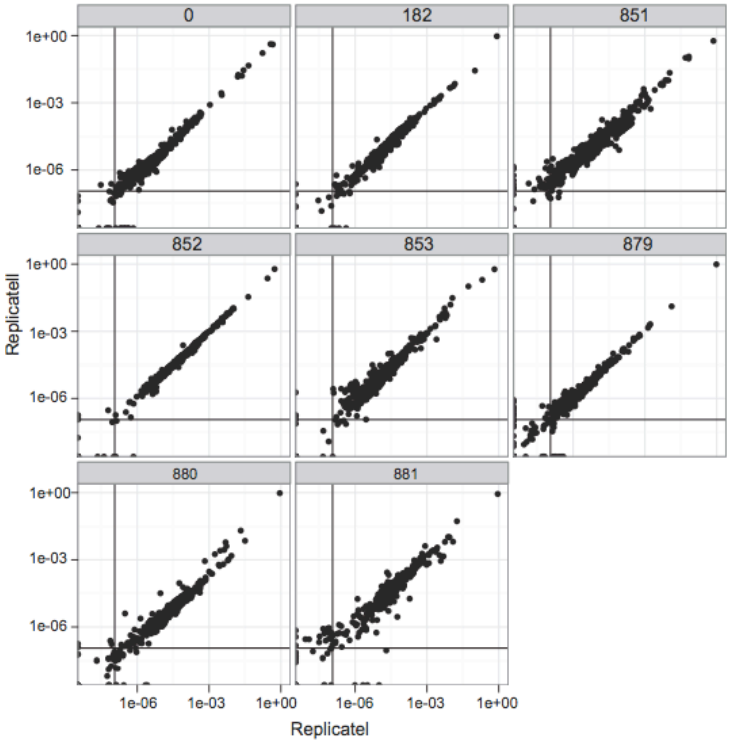


Fig. S1. Reproducibility between replicates. Each point represents the normalized abundance of a contig in a pair of replicate virome samples from the same time point. All contigs and pairs of technical replicates are represented in the figure.

Figure S2

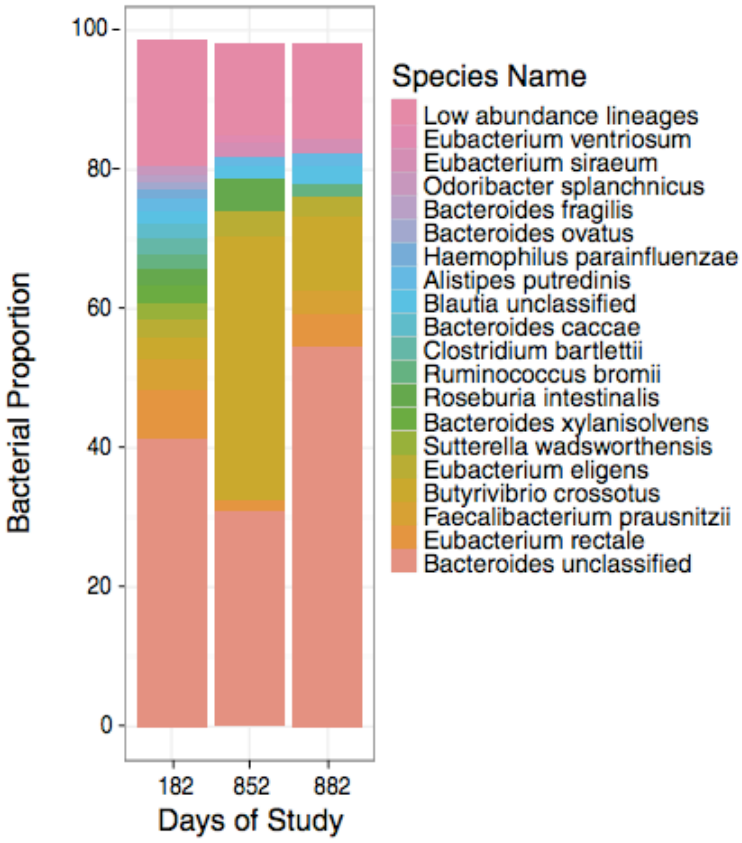


Fig. S2. Bacterial species detected in Illumina sequencing of unfractionated stool DNA. Bacterial lineages were identified using MetaPhlan.

Figure S3

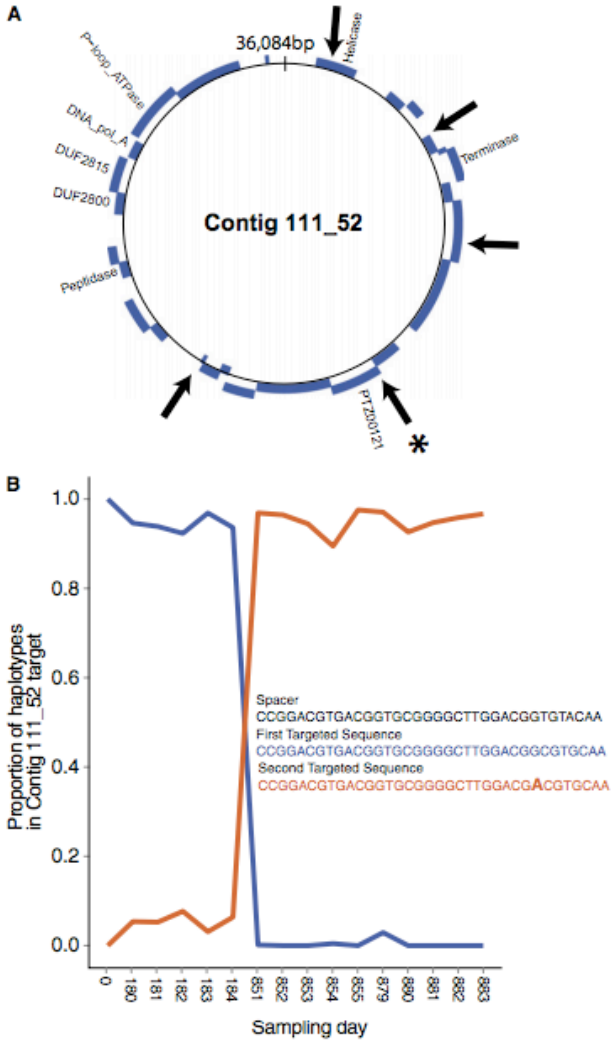


Fig. S3. A possible case of an escape mutation allowing evasion of CRISPR pressure. An example of bacterial CRISPRs targeting viral contig 111_52 and a possible example of an escape mutation. (A) Mapping of bacterial CRISPR target sites on the phage genome. The CRISPR spacer targets are shown by the arrows, and the spacer described in B is indicated by the asterisk. (B) Longitudinal abundance of a phage genome with an additional mismatch at a CRISPR homologous site. The genome containing an additional mismatch in the CRISPR recognition site (red) versus the original sequence (blue) increased in abundance over time.

2.7 Acknowledgments

We thank members of the F.D.B. laboratory for help and suggestions. This work was supported by Human Microbiome Roadmap Demonstration Project Grant UH2DK083981 (to G.D.W., F.D.B., and J.D.L.), the Penn Genome Frontiers Institute, and the University of Pennsylvania Center for AIDS Research Grant P30 AI 045008. S.M. was supported by National Institutes of Health Training Grant T32AI060516.

2.8 References

1. Rohwer F (2003) Global phage diversity. Cell 113(2):141.

2. Schoenfeld T, et al. (2010) Functional viral metagenomics and the next generation of molecular tools. Trends Microbiol 18(1):20–29.

3. Suttle CA (2005) Viruses in the sea. Nature 437(7057):356–361.

4. Reyes A, et al. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466(7304):334–338.

5. Minot S, et al. (2011) The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21(10):1616–1625.

6. Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD (2012) Hypervariable loci in the human gut virome. Proc Natl Acad Sci USA 109(10):3962–3966.

7. Minot S, Wu GD, Lewis JD, Bushman FD (2012) Conservation of gene cassettes among diverse viruses of the human gut. PLoS ONE 7(8):e42342.

8. Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI (2012) Going viral: nextgeneration

sequencing applied to phage populations in the human gut. Nat Rev Microbiol 10(9):607–617.

9. O'Brien AD, et al. (1984) Shiga-like toxin-converting phages from Escherichia coli strains that cause hemorrhagic colitis or infantile diarrhea. Science 226(4675): 694–696.

10. Waldor MK, Mekalanos JJ (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. Science 272(5270):1910–1914.

11. Turnbaugh PJ, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457(7228):480–484.

12. Yatsunenko T, et al. (2012) Human gut microbiome viewed across age and geography. Nature 486(7402):222–227.

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

14. Rodriguez-Valera F, et al. (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7(11):828–836.

15. Zhang T, et al. (2006) RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol 4(1):e3.

16. Peng Y, Leung HC, Yiu SM, Chin FY (2012) IDBA-UD: a de novo assembler for singlecell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28(11):1420–1428.

17. Sauvage V, et al. (2011) Identification of the first human gyrovirus, a virus related to chicken anemia virus. J Virol 85(15):7948–7950.

18. Breitbart M, et al. (2008) Viral diversity and dynamics in an infant gut. Res Microbiol 159(5):367–373.

19. Breitbart M, et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185(20):6220–6223.

20. Segata N, et al. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9(8):811–814.

21. Domingo-Calap P, Sanjuán R (2011) Experimental evolution of RNA versus DNA viruses. Evolution 65(10):2987–2994.

22. Berman L, et al. (2008) Defining surgical therapy for pseudomembranous colitis with toxic megacolon. J Clin Gastroenterol 42(5):476–480.

23. Brouns SJ, et al. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321(5891):960–964.

24. 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(3):181–186.

25. Karginov FV, Hannon GJ (2010) The CRISPR system: small RNA-guided defense in bacteria and archaea. Mol Cell 37(1):7–19.

26. Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea.

Science 327(5962):167–170.

27. Semenova E, et al. (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci USA 108(25):10098–10103.

28. Sebaihia M, et al. (2007) Genome sequence of a proteolytic (Group I) Clostridium botulinum strain Hall A and comparative analysis of the clostridial genomes. Genome Res 17(7):1082–1092.

29. Seed KD, Lazinski DW, Calderwood SB, Camilli A (2013) A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature 494(7438): 489–491.

30. McMahon SA, et al. (2005) The C-type lectin fold as an evolutionary solution for massive sequence variation. Nat Struct Mol Biol 12(10):886–892.

31. Miller JL, et al. (2008) Selective ligand recognition by a diversity-generating retroelement variable protein. PLoS Biol 6(6):e131.

32. Dai W, et al. (2010) Three-dimensional structure of tropism-switching Bordetella bacteriophage. Proc Natl Acad Sci USA 107(9):4347–4352.

33. Doulatov S, et al. (2004) Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature 431(7007):476–481.

34. Gevers D, et al. (2012) The Human Microbiome Project: a community resource for the healthy human microbiome. PLoS Biol 10(8):e1001377.

35. Ursell LK, et al. (2012) The interpersonal and intrapersonal diversity of human-associated microbiota in key body sites. J Allergy Clin Immunol 129(5):1204–1208.

36. Handley SA, et al. (2012) Pathogenic simian immunodeficiency virus infection is associated with expansion of the enteric virome. Cell 151(2):253–266.

37. Sell TL, Schaberg DR, Fekety FR (1983) Bacteriophage and bacteriocin typing scheme for Clostridium difficile. J Clin Microbiol 17(6):1148–1152.

38. Mahony DE, Clow J, Atkinson L, Vakharia N, Schlech WF (1991) Development and application of a multiple typing system for Clostridium difficile. Appl Environ Microbiol 57(7):1873–1879.

39. Cuevas JM, Domingo-Calap P, Sanjuán R (2012) The fitness effects of synonymous mutations in DNA and RNA viruses. Mol Biol Evol 29(1):17–20.

40. Cantalupo PG, et al. (2011) Raw sewage harbors diverse viral populations. MBio 2(5):1–11.

CHAPTER 3: Covalent modification of bacteriophage T4 DNA inhibits CRISPR-Cas9

Alexandra L. Bryson, Young Hwang, Scott Sherrill-Mix, Gary D. Wu, James D. Lewis, Lindsay Black, Tyson A. Clark, and Frederic D. Bushman

This paper is published: MBio. 2015 Jun 16;6(3):e00648. doi:10.1128/mBio.00648-15.

PMID: 26081634

3.1 Abstract

The genomic DNAs of tailed bacteriophages are commonly modified by attachment of chemical groups. Some forms of DNA modification are known to protect phage DNA from cleavage by restriction enzymes, but others are of unknown function. Recently the CRISPR-Cas nuclease complexes were shown to mediate bacterial adaptive immunity using RNA-guided target recognition, raising the question of whether phage DNA modifications may also block attack by CRISPR-Cas9. We investigated phage T4 as a model system, where cytosine is substituted with glucosyl-hydroxymethylcytosine (glc-HMC). We first quantified the extent and distribution of covalent modifications in T4 DNA using single molecule DNA sequencing and enzymatic probing. We then designed CRISPR spacer sequences targeting T4, and found that wild-type T4 containing

glc-HMC was insensitive to attack by CRISPR-Cas9, but mutants with unmodified cytosine were sensitive. Phage with HMC only showed intermediate sensitivity. While this work was in progress, another group reported examples of heavily engineered CRISRP-Cas9 complexes that could in fact overcome the effects of T4 DNA modification, indicating that modifications can inhibit but do not always fully block attack.

3.2 Importance

Bacteria were recently found to have a form of adaptive immunity, the CRISPR-Cas systems, which use nucleic acid paring to recognize and cleave genomic DNA of invaders such as bacteriophage. Historic work of tailed phages has shown that phage DNA is often modified by covalent attachment of large chemical groups. Here we demonstrate that DNA modification in phage T4 inhibits attack by the CRISPR-Cas9 System. This finding provides insight into mechanisms of host-virus competition, and also a new set of tools that may be useful in modulating the activity of CRISPR-Cas9 in genome engineering applications.

3.3 Introduction

The functional importance of covalent DNA modification was first demonstrated in 1952 in studies of bacteriophage (1, 2). In bacteriophage T4, genomic DNA contains 5-hydroxymethylcytosine (HMC), which is further modified by attachment of glucose to yield glucosyl HMC (glc-HMC) (Fig. 1) (3-6). Incorporation of HMC blocks DNA cleavage by many restriction endonucleases, and the glc-HMC modification further blocks attack by the HMC-specific McrABC (Rgl/MspJI) nuclease. Today more than ten different types of covalent modification are known in bacteriophage DNA, many of which are of unknown function (7, 8). Eukaryotic DNA also can be modified to methyl-cytosine, hydroxymethyl-cytosine and glucosylated hydroxymethyldeoxyuridine(9, 10).

Recently striking studies have revealed a new class of nucleases in bacteria, the CRISPR-Cas systems, which provide bacteria with a form of adaptive immunity against infection by genomic parasites such as phages or plasmids(11-17). Short sequences from genomic parasites are incorporated into CRISPR arrays in the bacterial chromosome, which consist of repeated sequences and unique spacers (typically ~30 nt) that are derived from invaders(18-20). Transcription of the arrays and RNA processing produces

spacer RNA sequences (crRNAs), which are bound by a nuclease (Cas9 for the type II CRISPR systems)(21). The crRNA is then used to recognize DNA of invaders by base pairing, allowing subsequent nucleic acid cleavage by Cas9(22) (reviewed in (23-25)). These programmable nuclease systems are now used widely in biotechnology applications (26-31).

The discovery of the CRISPR-Cas9 system raised the question of whether T4 DNA modification might have an additional function—protecting phage DNA from cleavage by CRISPR-Cas9. One previous paper analyzed the effects of a smaller DNA modification--adenine N6 methylation--and showed that phage DNA with this modification was still sensitive (32). Another study demonstrated that genome engineering by Cas9 in eukaryotes is also unaffected by 5-methyl cytosine (33). Here we investigate the effects of the larger DNA modifications found in T4. We first characterize T4 DNA modification in detail using single molecule sequencing and nuclease digestion. We go on to show that the bulkier HMC and glc-HMC modifications can in fact inhibit CRISPR-Cas9 attack. While this work was in progress, another group reported examples were T4 with glc-HMC modification could be in fact be sensitive to attack by CRISPR-Cas9, which we further analyzed and attribute to use of a heavily engineered and optimized CRISPR-Cas9 system (34).

3.4 Results

3.4.1 The extent of modification in T4 and mutant derivatives

Prior to testing sensitivity to CRISPR-Cas9, wild-type bacteriophage T4 (termed “T4(glc-HMC)” in the following) and mutants altered in DNA modification were analyzed to characterize the nature and extent of genomic DNA modification. T4(147) is mutant in genes encoding the alpha and beta glucosyl transferases, which attach glucose to HMC DNA, so that genomes contain HMC only (termed “T4(HMC)” in the following). T4(GT7) is mutant in genes required to substitute HMC for dCTP in nucleotide pools, so genomes contain only unmodified cytosines (termed “T4(C)” below). Complete T4 genotypes are in Table S1.

Several studies were carried out to verify the presence of the expected DNA modifications in each phage and evaluate the extent of base substitution. First, genomic DNAs were purified from phage stocks and probed by exposure to DNA modifying enzymes of known specificities. DNA from T4(C), but not T4(HMC) or T4(glc-HMC), was sensitive to digestion by the restriction enzyme AluI (Fig. 2A, top), as expected for DNA containing unmodified cytosines. T4 genomic DNAs were next incubated with MspJI, which cleaves HMC-containing DNA selectively. T4(HMC) DNA was digested, but not T4(glc-HMC) or T4(C)(Fig. 2A, middle). The T4 DNAs were also exposed to a glucosyl-

transferase and a glucose donor, which resulted in reduced mobility of T4(HMC) DNA, consistent with glucose attachment to HMC, but no changes were observed for the faster migrating T4(C) or slower migrating T4(glc-HMC) (Fig. 2A, bottom). T4 DNA modification was further probed by infection of *E. coli* strains expressing the Rgl nuclease, which cleaves HMC-containing DNA selectively. Infection by T4(HMC) was undetectable in Rgl-containing strains, but infection by T4(glc-HMC) or T4(C) was not restricted. These data confirm the expected modification patterns in the T4 DNA stocks studied, and indicate that the extents of HMC incorporation and subsequent glucosyl conjugation are high, consistent with an analysis of nucleotides generated after enzymatic degradation of T4(glc-HMC) DNA (35).

3.4.2 Single molecule sequencing to characterize T4 DNA modification

To characterize the extent and distributions of DNA modifications in more detail, we subjected each T4 DNA sample to analysis by single molecule real-time (SMRT) sequencing using Pacific Biosciences technology (36, 37). In this method, single DNA molecules are sequenced by synthesis on immobilized DNA polymerase enzymes. Sequence information is acquired by detection of fluorescently labeled nucleotides during each incorporation step. The presence of DNA modifications in the template can slow the kinetics of incorporation, allowing DNA modification to be quantified as an increase in interpulse duration

(IPD). IPD values are calculated for each position in the template sequence and are compared to an *in silico* model of IPD values for an unmodified sequence (IPD ratio) (38). In favorable cases, different forms of DNA modification show distinguishable kinetic profiles (38-41).

Figure 2B and Fig. S1A and B summarize the SMRT sequencing results for T4(glc-HMC), T4(HMC), T4(C), and a T4 genome lacking all forms of modification made by copying wild-type T4 DNA with a DNA polymerase *in vitro* (whole genome amplification, "WGA"). Kinetic profiles of T4(glc-HMC) showed many increased IPD ratios associated with expected positions of glc-HMC (Fig. 2B, top). Particularly increased IPDs were seen for potential glc-HMC sites 3' of a G residue, or 5' of a pyrimidine (Fig. S2). The mechanism of these sequence context effects is unknown--they could either reflect different extents of glucose attachment dependent on local sequence, or differential effects of sequence on polymerase kinetics. Kinetic perturbations were also seen at additional base positions, commonly those near C residues in the sequence, suggesting that modifications may contact polymerase from nearby positions in the DNA chain. For T4(HMC), kinetic lags were also associated with sites of potential C modification (Fig. 2B, middle), but the magnitude of the effects were typically less than for T4(glc-HMC). For T4(HMC) modification, IPD ratios were particularly high for pairs of expected HMC residues (Fig. S2). T4(C) and the WGA DNA

showed no notable alterations in kinetics at C residues (Fig. 2B, bottom two panels).

T4 also encodes an (N6-adenine)-methyltransferase that methylates A residues at 5'-GATC-3' sequences. The role of this modification is unknown, but it may help protect T4 DNA from the cellular methyl-directed mismatch repair system, which carries out double strand cleavage when mismatches are detected near unmodified 5'-GATC-3' sequences (42-44). Methylation was evident as increased IPD values at A in 5'-GATC-3' in all three T4 genomic DNAs but not in the WGA control. However, the extents of modification differed (Fig. 2C). The extent of 5'-GATC-3' adenine methylation was higher in both T4(C) and T4(HMC) than in T4(glc-HMC) DNA, paralleling a previous report (45). This is consistent with a steric interference model, which posits that glucosylation of HMC obstructs access of the adenine methyltransferase to T4 DNA and thereby reduces the extent of 5'-GATC-3' adenine methylation.

Sequencing data for T4(C) showed a high proportion of reads that mapped to the *E. coli* genome (56.7%; Fig. S3). Far fewer *E. coli* reads were detected in the T4(glc-HMC) and T4(HMC) samples (1.5% and 0.8% respectively). The T4(C) strain has been widely used in generalized transduction for genetic mapping in *E. coli* (46-48). T4(C) contains a mutation that inactivates the gene encoding the denB nuclease, which normally degrades host cell DNA,

thus allowing *E. coli* DNA to compete for packaging in T4 particles, as well as mutations inactivating the T4 encoded dCTPase and dCMP hydroxymethylase enzymes. In T4(C), different segments of the *E. coli* DNA was not packaged with uniform frequency (Fig. S3), suggesting possible involvement of sequence-specific recognition during T4 packaging(49).

3.4.3 Inhibition of CRISPR-Cas9 by T4 DNA modification

Given that the densities of DNA modification in T4(glc-HMC) and T4(HMC) are high, we sought to investigate whether the CRISPR-Cas system was blocked by T4 DNA modification. As a representative CRISPR-Cas system, we chose the type II system of *Streptococcus pyogenes*, because it has been widely used in biotechnology applications and functions well in *E. coli* (27). CRISPR spacers (targeting sequences) were designed to target four regions of the T4 genome (termed “protospacers”), each proximal to the required downstream 5’-NGG-3’ protospacer adjacent motif (PAM) in the T4 target.

Protospacer sequences and T4 DNA modification densities in these regions are shown in Fig. 3. The spacers were designed to target regions of the T4 genome with varying cytosine arrangement and density. None of the spacers contain the adenine methylation target sequence, GATC. Spacer T4 CRISPR 1 was designed to target a region of the T4 genome that maximizes the number of cytosines on the target strand and in the “seed sequence”, which is the inferred

3’ region of the CRISPR RNA that is reported to be most important for recognition (50). Spacer T4 CRISPR 2 maximizes the number of cytosines or modified derivatives on both the target and complementary strands in the T4 protospacer. Spacer T4 CRISPR 3 minimizes the number of cytosines in the target strand, but maximizes the number of cytosines on the complementary strand of the T4 protospacer. Spacer T4 CRISPR 4 minimized the numbers of cytosine residues on both the target and complementary strands of the T4 protospacer. IPD ratio analysis of T4(glc-HMC) and T4(HMC) DNA showed slowed kinetics at C-residues in the complement of the 5’-NGG-3’ PAM sequence and at internal cytosines, indicative of DNA modification.

The efficiency of phage infection was then tested on the CRISPR-Cas9-containing strains. To confirm the CRISPR system was active using our engineered spacers, we transformed the CRISPR-Cas9 containing bacteria with pUC19-derived plasmids encoding either the corresponding T4 protospacers and PAM sequences or a nonspecific control sequence (Fig. 4A). Using spectinomycin antibiotic selection for the target plasmid, we quantified efficiency of transformation, comparing the number of bacteria containing the incoming plasmid with a target matching T4 protospacer versus a control plasmid lacking the target (Fig. 4B). All of the CRISPR-Cas9 containing bacteria showed

reduced acquisition of the target-containing plasmid, indicating that the CRISPR systems are functional and reduce transformation by at least two logs (Fig. 4B).

The ability of T4 and mutant derivatives to infect CRISPR-Cas9 containing bacteria was then scored in plaque assays (Fig. 4C). Figure 4D-F shows illustrative experiments in which T4 phage were plated on CRISPR-Cas9-containing strains or controls and the efficiency of plaquing quantified. Infection with the cytosine-only strain T4(C) resulted in reduced or undetectable plaque formation in the presence of the T4-targeting spacers (Fig. 4D, right-most four spacers). Plaque formation was not affected by the presence of Cas9 and a nonspecific spacer, or in *E. coli* with no CRISPR-Cas9 system (Fig. 4D, left two samples marked “None” and “non-sp”). T4 CRISPR 1 showed the weakest activity, possibly due to high G/C content or the presence of homopolymeric sequences in the crRNA, which were previously reported to inhibit function (51). Titration studies on strains containing Cas9 and CRISPR 2, 3, and 4 showed the efficiency of plating to be reduced by >10,000-fold. CRISPR1 was weaker, showing only about 3-fold reduction, paralleling many studies showing variation in the efficiency of CRISPR targeting.

T4(glc-HMC), in contrast, formed plaques on strains expressing T4 CRISPR 1-4 and Cas9 efficiently (Fig. 4E). Infection of three of the four CRISPR-containing strains was as efficient as for the strains with control

nonspecific spacers. The fourth (T4 CRISPR 4) contained the spacer with the fewest modified C resides on both DNA strands, and so the lowest modification density. Infection by T4(glc-HMC) was reduced 2-10 fold in repeated assays, and plaque size was reduced by about two-thirds, indicating some sensitivity. Note that two modified cytosines are present in T4 CRISPR 4 target in the DNA complementary to the 5'NGG'3 PAM, and glucosylation of these likely exerted some inhibition. A previous study showed both DNA strands of the PAM are important for target recognition in other CRISPR systems(52). T4(glc-HMC) infection was not inhibited in strains expressing T4 CRISPR 3, which contains no cytosines on the target DNA strand but seven on the complementary strand, indicating that modifications on either strand can interfere with CRISPR attack. Thus glucosylation of HMC mostly protects T4 from attack by the CRISPR-Cas9 system, but a region with few glc-HMC residues showed modest but detectable sensitivity.

For T4(HMC)(Fig. 4F), the T4 CRISPR 1-3 targeting constructs did not inhibit infection, indicating that substitution of cytosine with HMC was also sufficient to block CRISPR-Cas9 attack. However, for T4 CRISPR 4, which has the fewest C residues on both the target and complementary strands, T4(HMC) was highly sensitive—efficiency of plating was reduced by at least 10,000-fold (Fig 4F). This indicates that the HMC modification alone on the cytosines on the

complementary strand of the 5'-NGG-3' PAM is not enough to inhibit CRISPR-Cas9 attack. Evidently the HMC modification is a less effective blocker than the glc-HMC modification, though both suffice at a high enough density.

3.4.4 Comparison to results of Yaung et al.

While this work was in progress, Yaung et al. reported three spacers in an engineered type II CRISPR system that were functional against wild-type T4(glc-HMC) phage and a T4 mutant containing HMC DNA (34). We obtained their spacer plasmids, and confirmed that they were able to restrict growth of T4(HMC) and T4(glc-HMC) phage in plaque assays as reported (Fig. S4A&B). These spacers differed from ours in that the crRNAs were engineered so that they were fused to tracrRNAs also know as a single-guide RNAs(26). The tracrRNA is a small RNA bound by Cas9 that is required for crRNA processing and as a cofactor for Cas9 nuclease activity(53). Fusion of the two RNAs is convenient in some genome engineering applications (26).

We cloned the spacers of Yaung et al. into the type II CRISPR system used in our studies, where crRNAs are not fused to tracrRNA, as in the natural *S. pyogenes* CRISPR-Cas9 system. We found that two of the three spacers were ineffective against the modified DNA of phage T4(glc-HMC) (Fig. S4D), but all three spacers were effective against unmodified DNA (Fig S5). Two of the spacers restricted growth of T4(HMC), while the third showed partial activity (Fig.

S4C). We confirmed that the Cas9 nucleases used here and by Yaung et al. functioned similarly in side-by-side tests, indicating that the CRISPR RNAs and not the Cas9 nuclease were responsible for the functional differences (Fig. S5E&F). Evidently the spacers of Yaung et al., with the synthetic single-guide RNA fusion, shows higher activity against modified T4 DNA (Fig. S4). These data indicate that DNA modifications can inhibit a biologically-occurring type II CRISPR system, but that particularly potent crRNAs can overcome this inhibition.

With the model system available to study CRISPR-Cas9 attack on T4, we were able to address further questions of T4 biology as described below.

3.4.5 Testing the role of T4 IP proteins

Three T4 proteins are injected into E. coli along with T4 DNA early during infection (IPI-III) and bind the T4 genome (54). IPI protects T4 from the GmrS/GmrD restriction enzyme, but the function of IPII and IPIII are unknown (55) —we thus asked whether any of the IP proteins contributed to evasion of the CRISPR-Cas9 system. This study was motivated in part by a previous report in which *Pseudomonas* phages were shown to encode protein inhibitors of a CRISPR-Cas system (56). For T4, such proteins would hypothetically be required for DNA modifications to exert their protective effect. A T4(glc-HMC) strain mutant in all three IP genes was tested by infection of strains containing the T4 targeting CRISPRs. No difference in infectivity was observed, indicating

that the IP proteins are not cofactors required to allow DNA modification to inhibit attack by CRISPR-Cas9 (Fig. S6). However, we note that there are multiple types of CRISPR-Cas systems that are quite different from each other, and T4 can infect *Escherichia*, *Shigella* and *Yersinia* (57), so it would be of interest to test possible inhibition of additional CRISPR-Cas systems from these organisms.

3.4.6 Characterization of a revertant of T4(C) with reduced sensitivity to CRISPR-Cas9

We observed a T4(C) revertant that reduced sensitivity to CRISPR-Cas9, and so characterized it further. Normally T4(C) plaques are small and turbid. During growth, we observed appearance of a new variant generating large clear plaques resembling T4(glc-HMC) plaques. Further tests showed reduced sensitivity to CRISPR-Cas9 (FigS7 C data). We sequenced the revertant phage, named $T4(C)^R$ and identified three mutations (FigS7 A and B). One point mutation eliminated the stop codon in gp42, which encodes dCMP hydroxymethylase, an enzyme necessary to synthesize HMC. A second mutation introduced a single nucleotide deletion into the deoxycytidylate deaminase gene, yielding a stop codon that truncated the encoded protein. Deoxycytidylate deaminase converts dCMP (a precursor of HMC) to dUMP. Both of these mutations favor the synthesis of HMC, which can be incorporated

in the T4(C)^R genome and then further modified to glc-HMC by alpha and beta glucosyl transferases. The third mutation was a nonsynonymous point mutation in the uncharacterized, hypothetical protein NrdC.5, and is of unknown significance.

T4(C)^R was resistant to spacers 1,2, and 3 in the CRISPR-Cas9 system, but sensitive to spacer 4 (FigS7 C), thus showing slightly greater sensitivity than wild-type T4(glc-HMC). These results and other results are consistent with the idea that T4(C)^R contains HMC or glc-HMC, though potentially not at every position in the genome due to higher cellular dCTP pools competing for incorporation. These findings again support the idea that DNA modifications can block CRISPR-Cas9 activity.

3.5 Discussion

These data show that modification of T4 DNA to HMC or glc-HMC reduces sensitivity to attack by CRISPR-Cas9. A previous study showed that adenine methylation at a 5'-GATC'3' sequence did not block CRISPR-Cas-mediated inhibition (32), and data presented here shows that low density modification with HMC also was not protective. Evidently protection against CRISPR-Cas9 attack can be achieved either by addition of bulkier glucosyl-HMC modifications or addition of a high density of less bulky HMC modifications.

While this work was in progress, and contrary to our developing data, Yaung et al. reported spacers that could in fact target glc-HMC modified T4 DNA efficiently (34). Our own tests with the reagents of Yaung et al. confirmed their conclusions. The Cas9 enzymes used were identical in both studies, specifying the RNA component as the origin of the different potency. Yaung et al. used crRNAs fused to tracrRNAs, which could have potentially improved activity by favoring RNA loading onto Cas9, or increased specific activity of the loaded sgRNA/Cas9 complex. One of the crRNAs of Yaung was notably potent even without the tracrRNA fusion, suggesting that for this spacer fusion with the crRNA did not explain potency. Another candidate explanation is that the positions of base modifications in the recognition site may be important, and that the rules for this are not fully clarified. For all spacers studied here, we have not investigated whether cleavage mediating T4 inhibition is in fact due to on target or off target cleavage, so increased off target specificity is another possible explanation for increased inhibition (58).

Classic studies on the tailed DNA phages have identified more than ten different forms of covalent DNA modification, and modification is commonly found in DNA of these viruses (7, 8). Recent metagenomic studies also emphasize the ubiquity of CRISPR systems targeting phage in natural environments such as the human microbiome (59-61). There are even examples of phage from the human

gut that themselves encoding CRISPR spacers targeting other phage from the same individual, indicating that phages may be competing with each other using the CRISPR-Cas system (60, 61). Given these observations, and data shown here that modification of T4 DNA to HMC or glc-HMC can reduce sensitivity to attack by CRISPR-Cas9, it seems probable that many of the bulkier forms of DNA modification seen in tailed DNA phage have evolved at least in part to reduce sensitivity to cleavage by CRISPR-Cas systems.

3.6 Materials and Methods

Propagation of phage strains. Manipulation of phage T4 was carried out as described in (62). Phage T4(glc-HMC), T4(HMC), and T4(C) were provided by Lindsay Black. Genotypes are listed in Table S1. T4(C) contains amber mutations in several DNA modifying genes (Table S1). The amber mutations are known to easily revert so T4(C) was propagated in the amber suppressor strain *E.coli* CR63 to prevent genotype reversion. Experiments with T4(C) were carried out in the non-suppresser *E.coli* strain DH10B ensure cytosines in T4(C) were unmodified. Experiments and propagation of T4(glc-HMC) and T4(HMC) were carried out in DH10B. Experiments with T4(IP^0) were carried out in *E.coli* B834.

CRISPR system and spacer design. Design of CRISPR spacers was carried out using custom code in R attached in Supplementary Material. CRISPR targeting plasmids were constructed using the system described by L. Marraffini and coworkers(27), which consists of two plasmids pCas9 and pCRISPR. pCas9 contains the Cas9 nuclease and tracrRNA (Addgene number 42876). Spacers in this study were cloned into the CRISPR array on pCRISPR (Addgene number 42875) using the Marraffini lab protocol available on Addgene. Comparison of work to Yaung et al. was carried out using plasmids DS-SPCas (addgene number 48645), PM-SP!TB (addgene number 48650) and plasmids provided by Yaung et al. Oligos used for cloning are listed in table S2.

Plasmid transformation assays. T4 protospacer and PAM sequences used in this study were individually cloned into pUC19 plasmids. 100ng of protospacer/PAM containing pUC19 were transformed into chemically competent *E. coli* DH10B containing a CRISPR-Cas9 system targeting the corresponding protospacer. As a transformation control, 100ng of pUC19 without a protospacer was transformed into DH10B containing a CRISPR-Cas9 expression system. Transformations were incubated at 37^{o}C for 1hr in 200uL SOC media without antibiotic selection then plated on LB 100ug/mL carbenicillin plates to select for pUC19. Efficiency of transformation was determined by dividing the number of

colony forming units observed in the protospacer containg pUC19 transformation by the number of colony forming units observed in the control pUC19 transformation.

Plaque assays. Plaque assays were used to determine the ability of phage to infect bacteria DH10B containing the CRISPR-Cas9 system. Up to 10^4 phage PFUs in a volume of 10uL were added to 200uL of log phase *E.coli* DH10B and incubated at room temperature for 10min. 3mL of 0.4% LB top agarose were added to the bacteria/phage, mixed, and poured onto LB plates containing appropriate antibiotics:100ug/mL kanamycin for pCRISPR, 50ug/mL chloramphenicol for pCas9, 100ug/mL ampicillin for DS-SPcas and 50ug/mL chloramphenicol for PM-SP!TB. Plates were incubated at 37^{o}C overnight. Three biological replicates, each with three technical replicates, were carried out per experiment. The efficiency of plaquing was determined by dividing the number of plaques on an experimental plate by the number of plaques on a control plate containing *E.coli* with no CRISPR system. A Kruskal-Wallis nonparametric comparison of means was carried out using GraphPad Prism software for each experiment.

Phage DNA isolation and sequencing. Phage lysates were grown at an MOI of 0.01 on DH10B. Phage T4 DNAs were isolated using Norgen Phage DNA isolation kit (Thorold, Canada). Chloroform-treated phage lysates were concentrated by 4% precipitation in PEG4000/500 mM NaCl, resuspended into TE buffer, and purified as recommended. The concentration of isolated T4 phage DNAs was measured using Quant-iT™ PicoGreen® dsDNA Assay Kit (Carlsbad, CA). For single molecule sequencing, purified phage DNA was fragmented to an average size of 1.5 kb via adaptive focused acoustics (Covaris, Woburn, MA). SMRTbell template sequencing libraries were prepared as previously described (63). Sequencing was carried out on an *RS II* (Pacific Biosciences, Menlo Park, CA) using P4/C2 sequencing chemistry and standard protocols for large insert libraries. Consensus sequences were generated using Quiver and kinetic data was generated with SMRT Analysis Software v2.0 (Pacific Biosciences). For further methods see SI Methods. Libraries for sequencing T4(C) and T4(C)^{R} were made using Illumina's Nextera XT DNA Sample Preparation Kit with 1 ng of input DNA, generating paired-end fragments. Metagenomic sequencing was performed on an Illumina MiSeq instrument. Paired-end reads from the MiSeq instrument were quality-trimmed. Reads were aligned using Geneious to the NCBI T4 genome sequence to form consensus sequences for T4(C) and T4(C)^{R}.

Nuclease Assays. 1ug of T4(C), T4(HMC), and T4(glc-HMC) were digested with AluI (NEB: R0137s), MspJI (NEB: R0661S), and T4 phage β-glucosyltransferase (NEB: M0357S) using NEB specified protocols.

Acknowledgments

The authors thank M. Boitano for assistance with DNA sequencing and J. Korlach for comments on the manuscript. Thanks to Alice Laughlin for assistance, to Kushol Gupta for help with Fig. 1A, and to L. Marraffini for providing pCas9 and pCRISPR plasmids, and useful discussions.

Accession Numbers

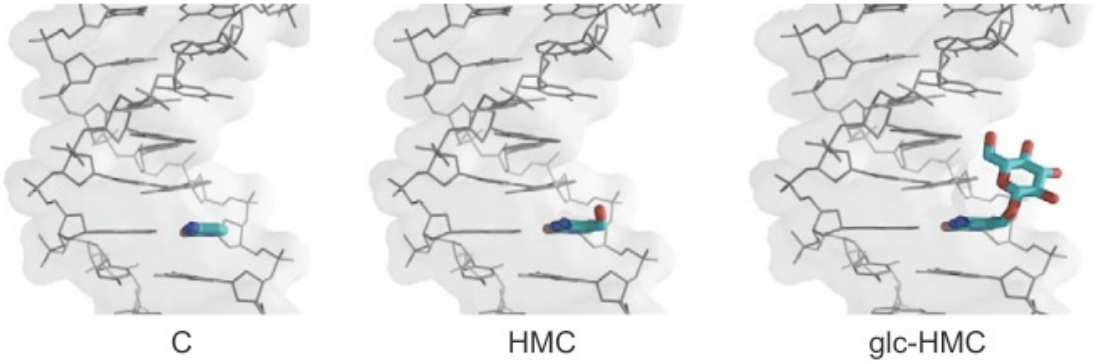
The T4 genome sequences will be deposited at NCBI upon acceptance of the paper for publications.

Funding

This work was supported by Project UH2DK083981, NIH AI39368 (GDW); Penn Digestive Disease Center (P30 DK050306); The Joint Penn-CHOP Center for Digestive, Liver, and Pancreatic Medicine; S10RR024525; UL1RR024134, K24-DK078228.

3.7 Figures

Figure 1



DNA modification in phage T4 showing C-containing DNA (left), HMC-containing DNA (middle), and glc-HMC DNA (right).

Figure 2

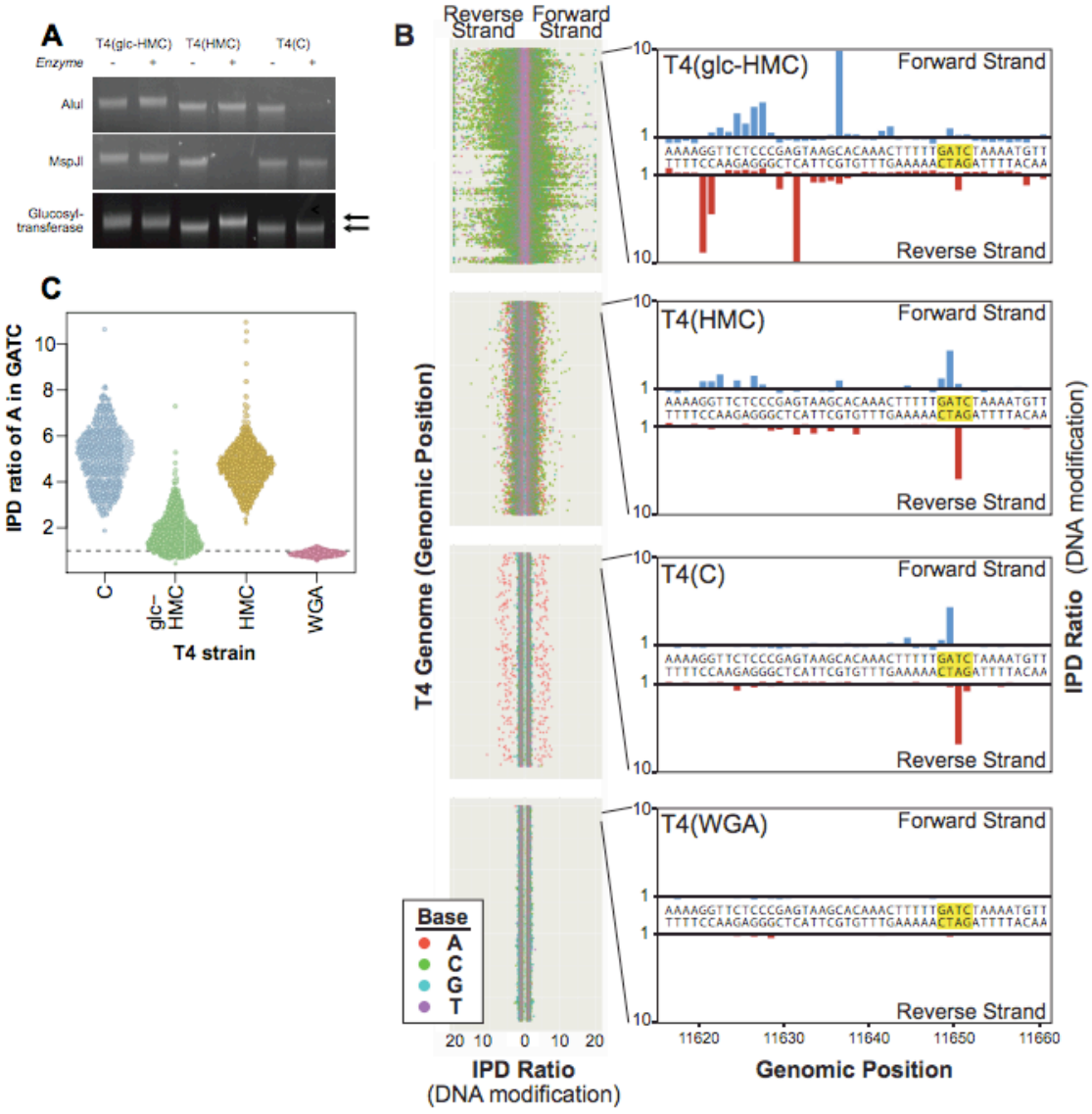


FIG 2 Characterization of phage T4 DNA modification. (A) Phage T4(glc-HMC), T4(HMC), and T4(C) DNA left untreated () or treated with () restriction enzymes AluI (top), which cleaves unmodified DNA; MspJI (middle), which cleaves HMC-containing DNA; or T4 glucosyltransferase (bottom), which increases the mobility of HMC-containing DNA by the addition of glucose groups. The arrows indicate the mobility shift due to glucose attachment. (B) Analysis of phage T4 DNA modification by single-molecule sequencing. Results are summarized for each genome by

mapping IPD ratios at each base for each of the T4 strains studied. The coloration of each base is shown by the key at the bottom left. The T4 nucleotide sequence runs from top to bottom for each of the four genomes. The distance each colored point is displaced from the center indicates the IPD ratio (scale at bottom; leftward for the reverse strand, rightward for the forward strand). Examples of interpulse distances (indicative of modification) are shown to the right for a short segment of the T4 genome. Bars indicate the magnitude of the IPD ratio (upward for the forward strand and downward for the reverse strand).A 5=GATC 3=site of DAM methylation is highlighted in yellow. (C)Violin plot showing IPD ratios of A residues at 5= GATC 3= sequences.

Figure 3

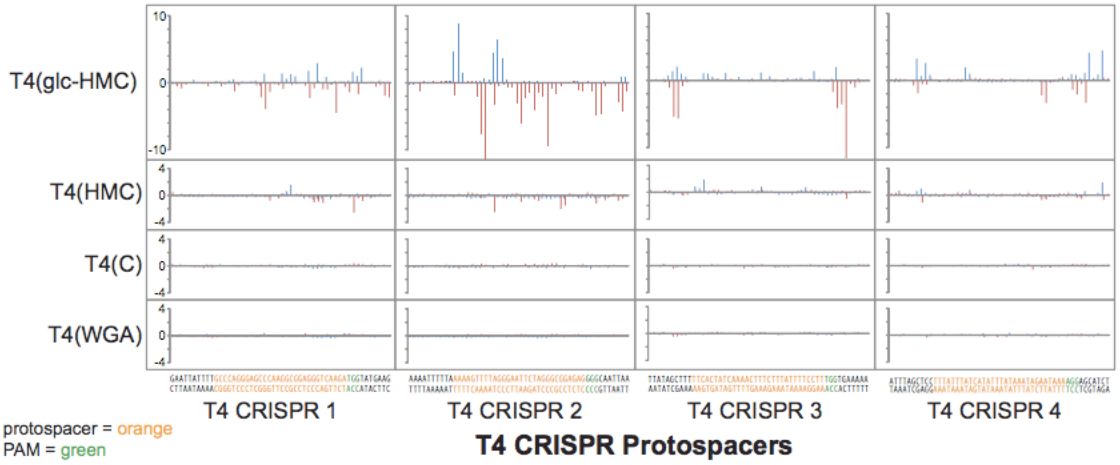


FIG 3 IPD modification profiles of T4(glc-HMC), T4(HMC), and T4(C) phage protospacers. IPD ratios for the forward strand (blue) and reverse strand (red) of T4(glc-HMC), T4(HMC), and T4(C) are depicted for the regions of the T4 genome targeted by spacers 1 to 4 along with the surrounding nucleotides. The nucleotide sequences of the phage protospacer (orange), the PAM (green), and the surrounding nucleotides (black) are along the x axis. The top strand of the protospacer is identical in sequence to the crRNA/spacer, and the bottom strand is the target strand, which is complementary to the spacer and will base pair with the crRNA.

Figure 4

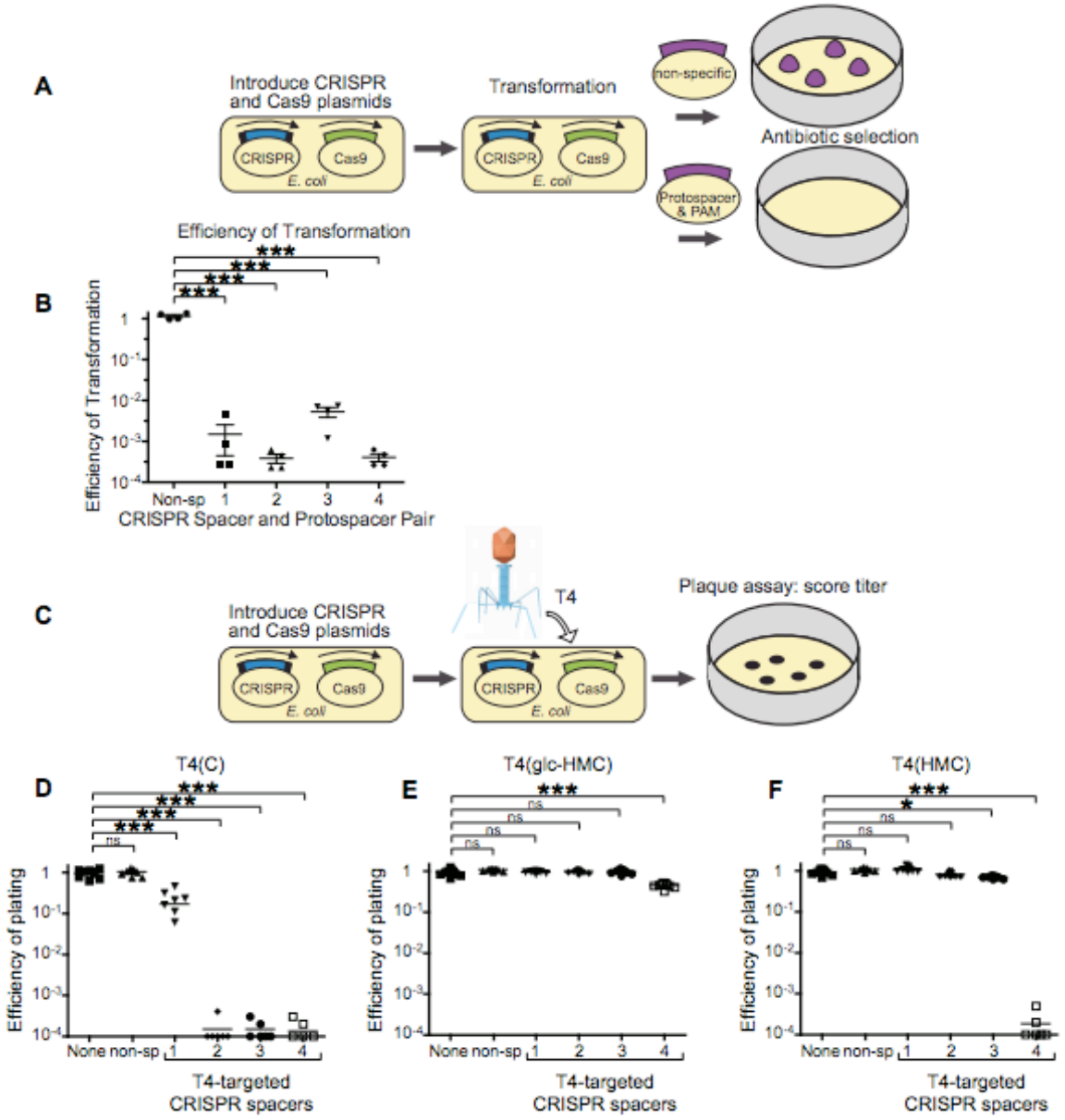


FIG 4 Glc-HMC and HMC modifications inhibit attack by the CRISPR-Cas9 system on phage T4. (A) Diagram of the strategy used to validate CRISPR spacers in a transformation assay. Bacteria containing the type II CRISPR system were transformed with a pUC19 plasmid containing either a T4 protospacer and PAM sequence or a nonspecific DNA sequence. Antibiotic selection for the

pUC19 plasmid and quantification of the efficiency of transformation reveal the efficacy of CRISPR system cleavage of unmodified DNA containing a protospacer and PAM. (B) Results of plasmid challenge tests. The efficiency of transformation is the ratio of colony counts of cells transformed with equal amounts of pUC19 that contain a protospacer targeting the plasmid (numerator) to the colony counts of cells transformed with pUC19 (denominator). (C) Diagram of plaque assays to assess inhibition of T4 infection with CRISPR-Cas9. (D to F) Results of plaque assays in which the E. coli strains indicated were infected with up to $1x10^4$ PFU of T4(C) (panel D), T4(glc-HMC) (panel E), or T4(HMC) (panel F). E. coli strains expressed Cas9 and crRNAs targeting T4 or controls. Starting from the left in each panel, None indicates no crRNA or Cas9, non-sp indicates nonspecific crRNA, 1 contained the maximum number of cytosines in the target strand and seed sequence, 2 contained the maximum number of cytosines in the target and complementary strands, 3 contained no cytosines in the target strand and seven cytosines in the complementary strand, and 4 contained the fewest cytosines in the target and complementary strands. Mean values were compared with the Kruskal-Wallis test. *, $P < 0.01$; ***, $P < 0.0001$; ns, not significant.

Figure S1

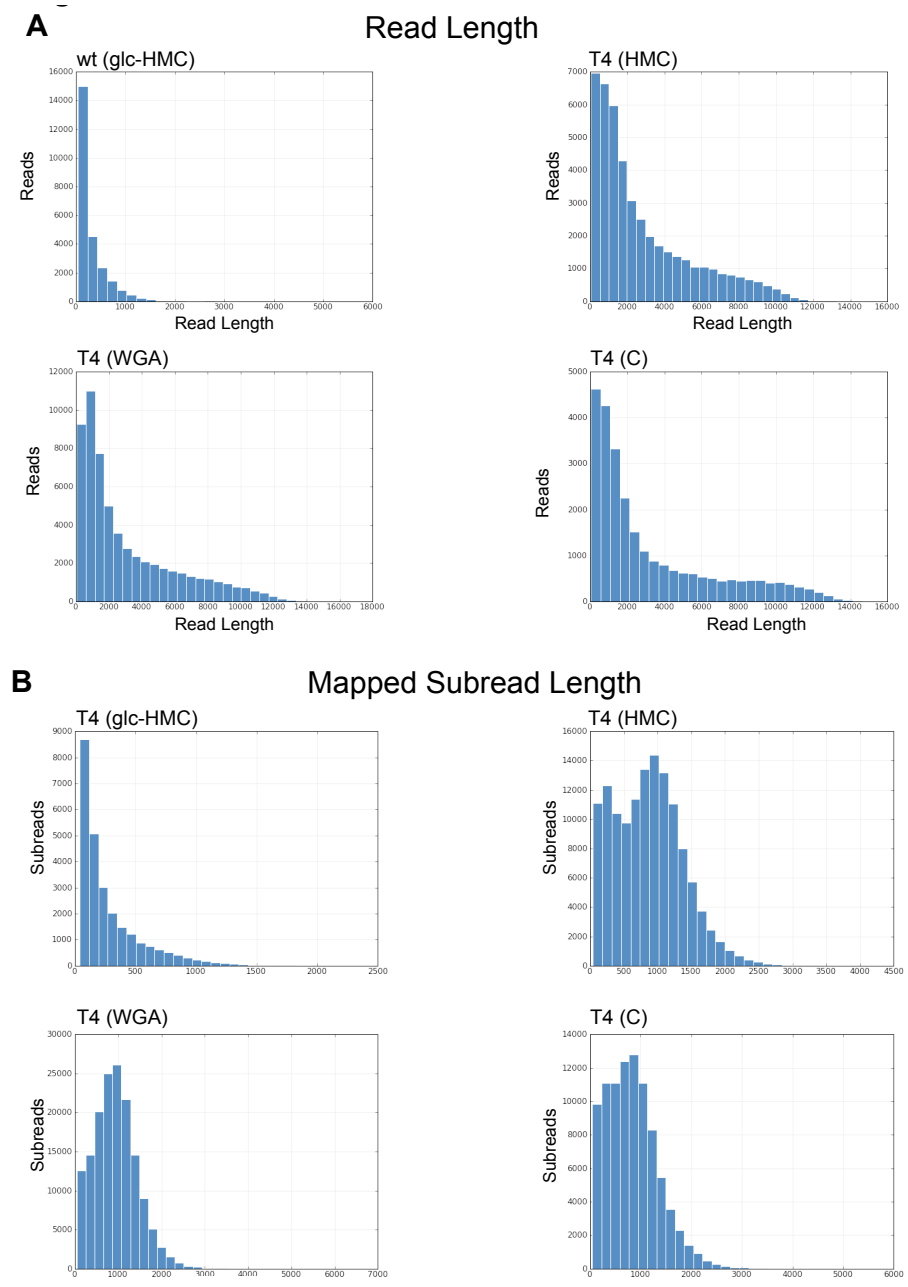


FIG S1 (A) Read lengths in the single-molecule sequence data sets. (B) Mapped subread lengths in the single-molecule sequence data sets.

Figure S2

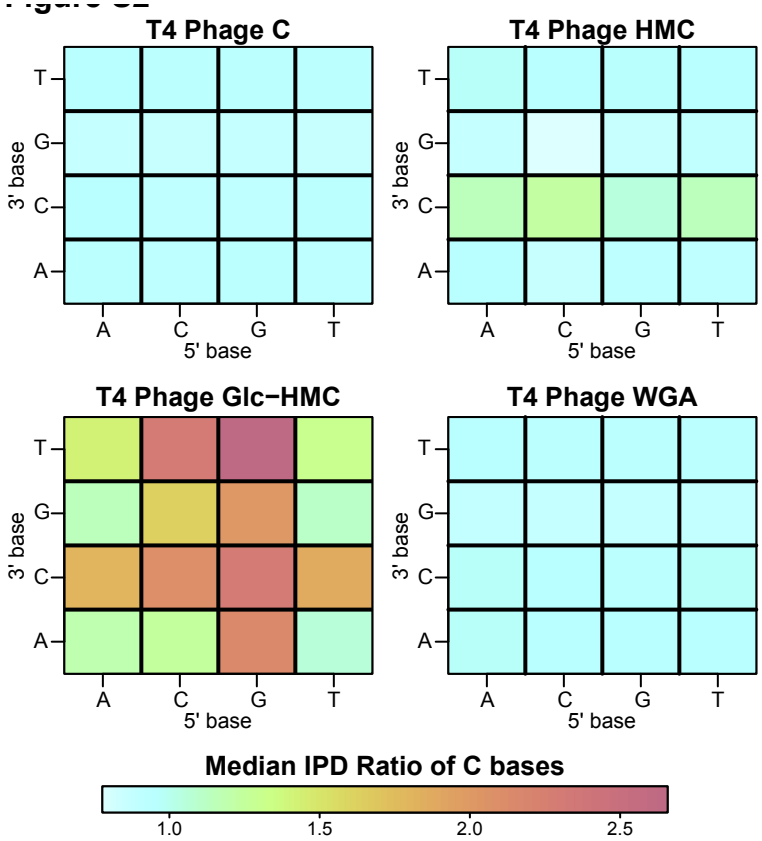


FIG S2 Heat map summarizing the effects of local sequences on IPD ratios at C residues. The base preceding the C residue in the sequence is marked 5′-base, and that following the C residue is marked 3′-base. The scale at the bottom summarizes the IPD ratios.

Figure S3

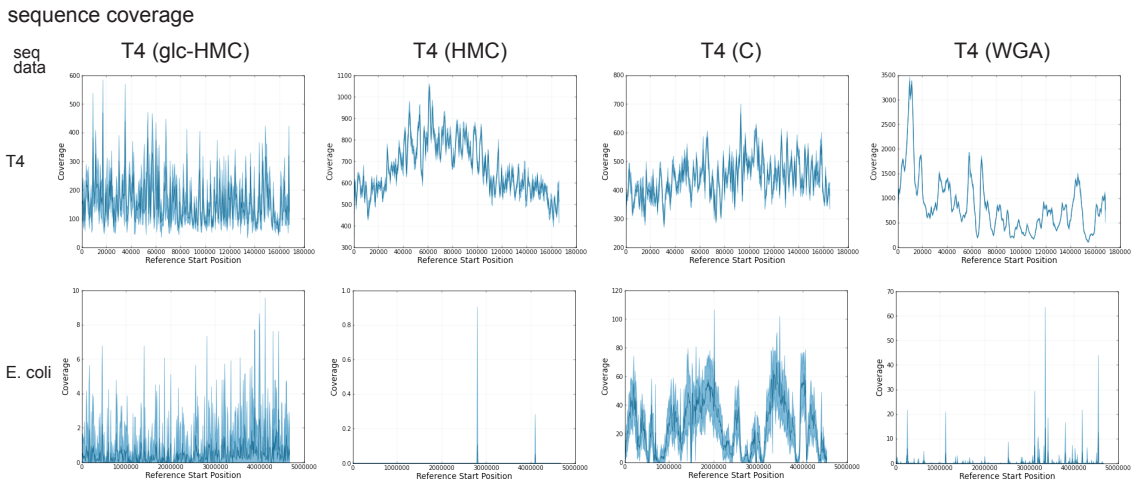


FIG S3 Sequence coverage maps for T4 strains comparing T4 (top) to *E. coli* B834 (bottom).

Figure S4

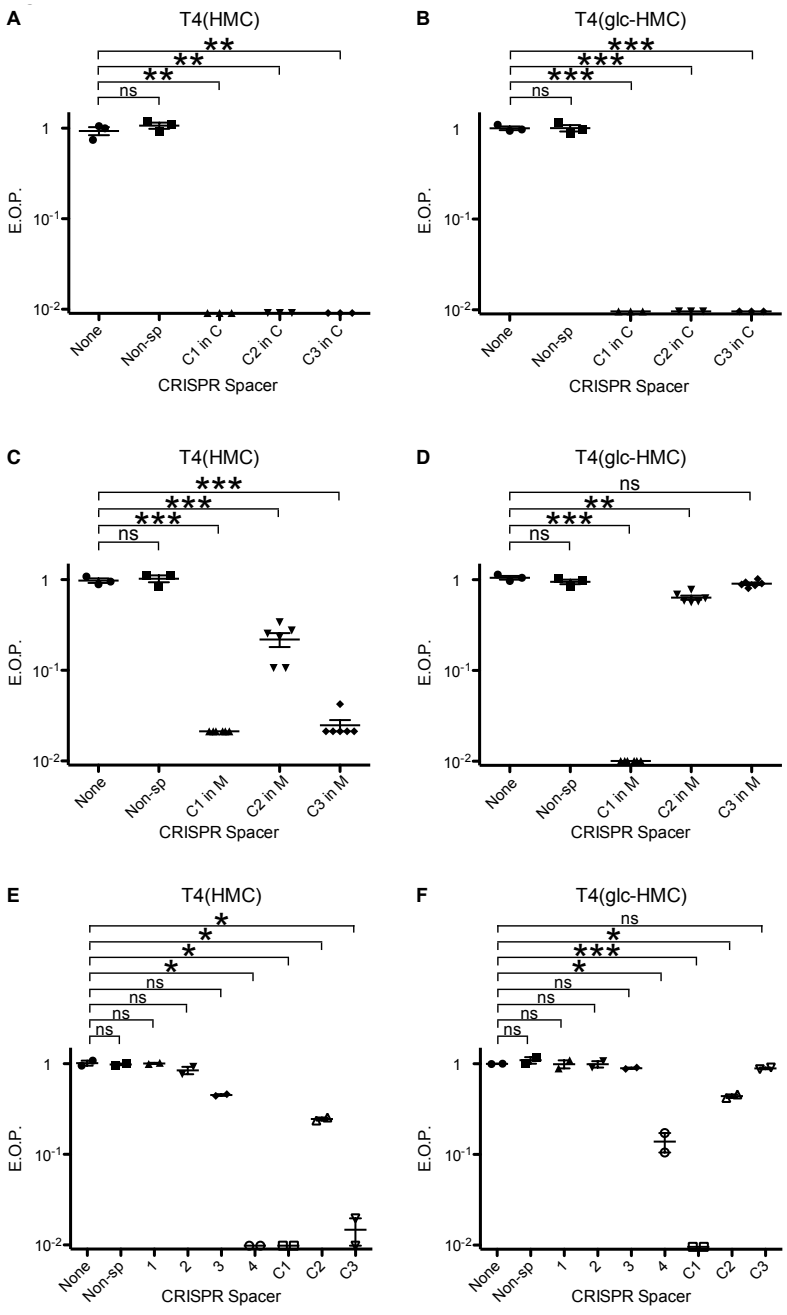


FIG S4 Comparison of our work with that of Yaung et al. (A and B) Results of plaque assays in which the*E. coli* strains indicated containing previously studied CRISPR spacers (C1 to C3) in the

Yaung-Church CRISPR system (in C) were infected with up to 100 PFU of T4(HMC) (panel A) or T4(glc-HMC) (panel B). CRISPR spacer labeling: None, no crRNA or Cas9; non-sp, nonspecific crRNA. E.O.P., efficiency of plating. (C and D) Results of plaque assays in which the *E. coli* strains indicated containing previously studied CRISPR spacers (C1 to C3) in the Marraffini CRISPR system (in M) were infected with up to 100 PFU of T4(HMC) (panel C) or T4(glc-HMC) (panel D). CRISPR spacer labeling: None, no crRNA or Cas9; non-sp, nonspecific crRNA. (E and F) Plaque assay results of the Church laboratory Cas9 expression vector with the Marraffiini CRISPR array on the T4 CRISPR spacers studied here. The Church Cas9 expression vector and the Marraffini CRISPR array containing the spacers studied in this investigation (spacers 1 to 4) and the previously studied spacers (C1 to C3) were tested for efficacy against T4(HMC) (panel E) and T4(glc-HMC) (panel F). Mean efficiency of transformation was compared to that of a nonspecific control with a *t* test. **, $P < 0.001$; ***, $P < 0.0001$; ns, not significant.

Figure S5

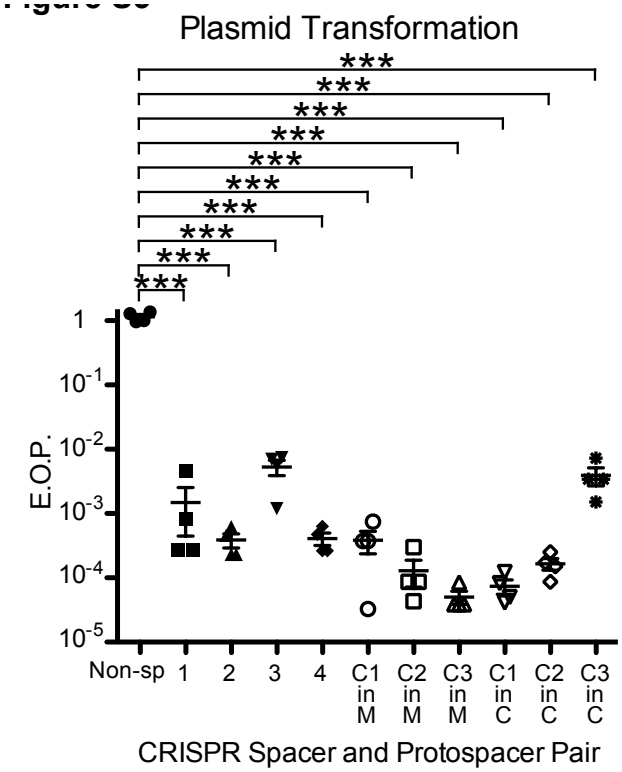


FIG S5 Results of plasmid transformation assay comparing the efficacies of all of the CRISPR spacers studied in this investigation against unmodified DNA. Efficiency of transformation was normalized to 1 by the transformation of a plasmid not targeted by the CRISPR system (control). Spacers 1 to 4 are from this study. C1 to C3 are the spacers from Yaung et al. cloned into the CRISPR-Cas9 system developed by the Church lab (in C) or the Marraffini lab (in M). Mean

efficiency of transformation was compared to the nonspecific control with a *t* test. ***, $P < 0.0001$. E.O.P., efficiency of plating.

Figure S6

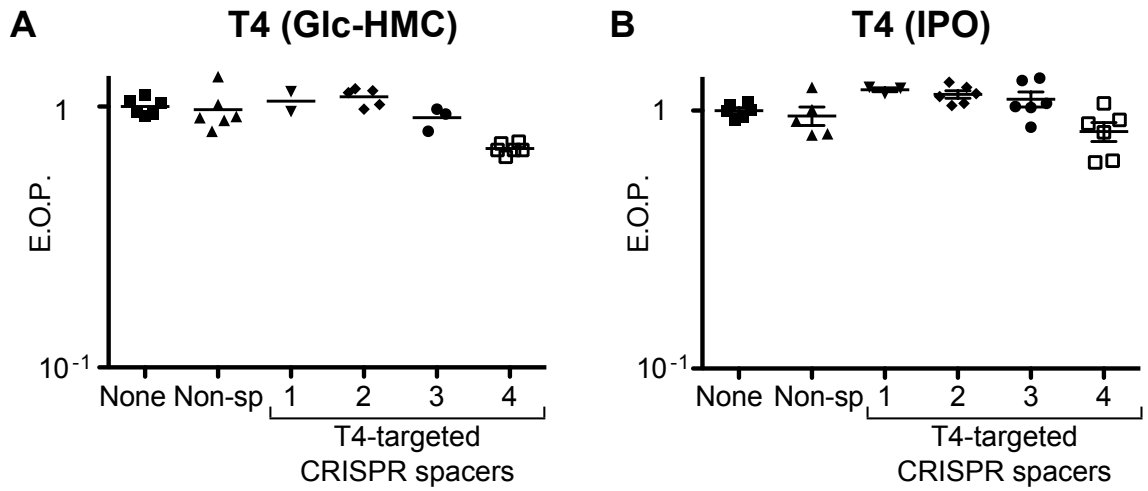


FIG S6 Mutation of IP1-3 genes encoding the three T4 proteins that are injected along with the phage DNA does not reduce resistance to attack by CRISPR-Cas9. Shown are replicate infections with wild-type T4(glc-HMC) (A) and the triple mutant T4(IP0) (B), which show no differences in infectivity for the strains tested. Plaque assay with approximately 100 PFU of T4(glc-HMC) or T4(IP0) infecting *E. coli*not expressing CRISPR-Cas9 (None) or *E. coli* expressing CRISPR-Cas9 with spacers targeting a protospacer in the T4 genome with the maximum number of cytosines in the target sequence and seed sequence (1), the maximum number of cytosines in the target and complementary strands (2), the fewest cytosines in the target strand (3), the fewest cytosines in the target and complementary strands (4), or a nonspecific spacer that does not target the T4 genome (non-sp). The mean efficiency of plating (E.O.P.) for infection of cells with each spacer was compared to that of no-CRISPR-Cas9 control with a *t* test. No statistically significant differences were found.

Figure S7

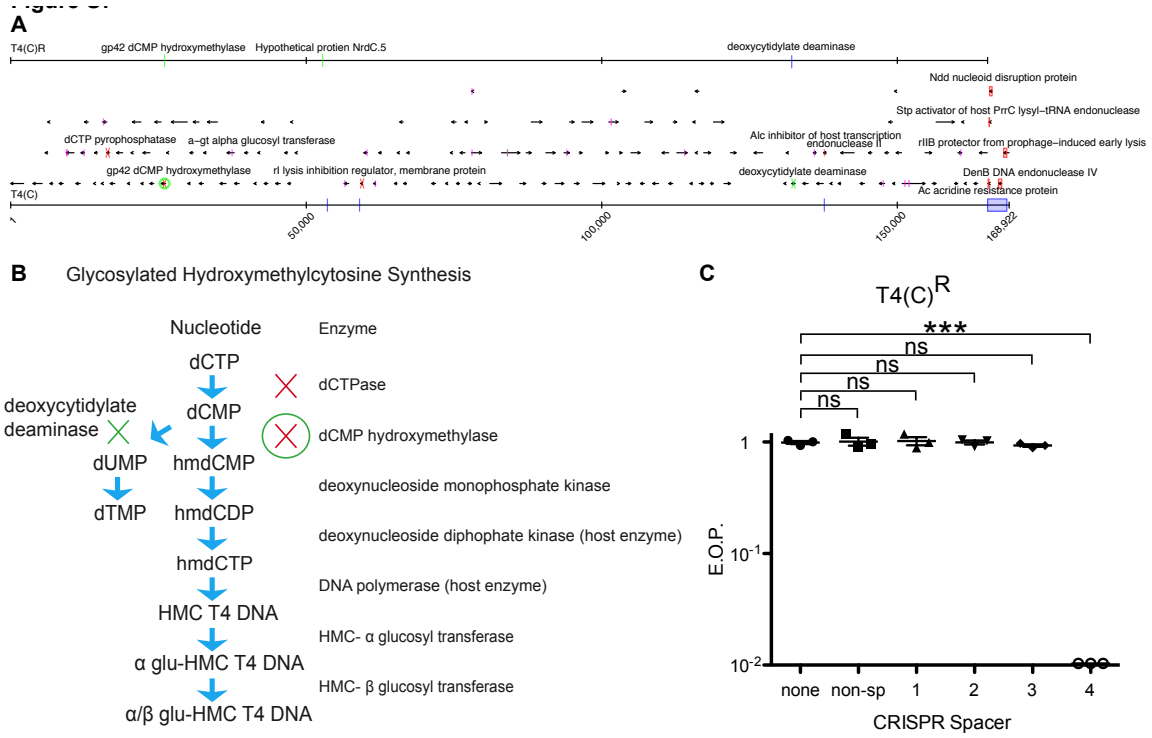


FIG S6 Phenotype of a revertant of T4(C) named T4(C)^{R} with reduced CRISPR sensitivity. (A) Genetic map of T4(C) and the revertant obtained by Illumina deep sequencing. The bottom black line represents the T4(glc-HMC) genome length, and black arrows indicate genes and the direction of transcription. Variations in the T4(C) genome compared to T4(glc-HMC) are blue for single nucleotide deletions, red for large deletions, and pink for nonsynonymous substitutions; red ×'s represent early stop codons. The top black line represents the T4(C)^{R} genome. Variations in T4(C)^{R} compared to T4(C) are green for nonsynonymous mutations and blue for deletions. The green circle indicates reversion of a stop codon, and the green × indicates a stop codon. (B) Glycosylated hydroxymethylcytosine synthesis pathway in T4(glc-HMC) phage and mutations in T4(C) and T4(C)^{R}. Proteins mutated in T4(C) are shown by red ×'s. T4(C)^{R} acquired a mutation designated by a green × and reverted a previous amber mutation denoted by a green circle. (C) Reduced sensitivity to CRISPR attack in the revertant. The mean efficiency of plating (E.O.P.) for infection of cells with each spacer was compared with that of the no-CRISPR-Cas9 control with a *t* test. ***, $P < 0.0001$; ns, not significant.

3.8 Tables

Table S1

Table S1 Genotypes of bacteria and bacteriophages used in this study

Phage (Nomenclature used in this paper)	Alternative/ historical name	Genotype
T4(Glc-HMC)	T4	wild type
T4(HMC)	T4147	αgt1, βgt7
T4(C)	T4GT7	amC87g42, amE51g56, rNB5060, alc
IP^0	IP^0	IPI (HA35) IP2 (amber HA100) IP3 (amber HA9)

Bacteria		Genotype
DH10B		F– mcrA Δ(mrr-hsdRMS-mcrBC)
CR63		λ-, serU60(AS), lamB63
B834		hsdRB, hsdMb, Sup0, rgl+
K803		rk-, mk-, rgl-, supE44

Table S2

Table S2. Oligonucleotides used in this study

Oligo Name	Oligo DNA sequence 5'-3'	Function	Source
T4_Top_70718_70747	AAACAAAAGTTTTAGGGAATTCTAGGGCGGAGAGG	top oligo for cloning T4 spacer 1 into pCRISPR	this study
T4_Bottom_70718_70747	AAAACCTCTCCGCCCTAGAATTCCCTAAAACTTTT	bottom oligo for cloning T4 spacer 1 into pCRISPR	this study
T4_ModA_Top_13205_13234	AAACGCCCAGGGAGCCCAAGGCGGAGGGTCAAGAG	top oligo for cloning T4 spacer 2 into pCRISPR	this study
T4_ModA_Bottom_13205_13234	AAAACTCTTGACCCTCCGCCTTGGGCTCCCTGGGC	bottom oligo for cloning T4 spacer 2 into pCRISPR	this study
T4_30.2_top_127687_127716	AAACTTCACTATCAAAACTTTCTTTATTTTCCTTG	top oligo for cloning T4 spacer 3 into pCRISPR	this study
T4_30.2_bottom_127687_127716	AAAACAAGGAAAATAAAGAAAGTTTTGATAGTGAA	bottom oligo for cloning T4 spacer 3 into pCRISPR	this study
T4_Top_116439_116468	AAACTTTATTTATCATATTTATAAATAGAATAAAG	top oligo for cloning T4 spacer 4 into pCRISPR	this study
T4_Bottom_116439_116468	AAAACTTTATTCTATTTATAAATATGATAAATAAA	bottom oligo for cloning T4 spacer 4 into pCRISPR	this study
Church1_top_pPCRISPR	AAACATATCGAAAGCAATCAGGTTG	top oligo for cloning T4 spacer C1 into pCRISPR	adapted from Yaung et al. 2014
Church1_bottom_pCRISPR	AAAACAACCTGATTGCTTTCGATAT	bottom oligo for cloning T4 spacer C1 into pCRISPR	adapted from Yaung et al. 2014
Church2_top_pPCRISPR	AAACAAGAACTTCCAACCGGTAATG	top oligo for cloning T4 spacer C2 into pCRISPR	adapted from Yaung et al. 2014
Church2_bottom_pCRISPR	AAAACATTACCGGTTGGAAGTTCTT	bottom oligo for cloning T4 spacer C2 into pCRISPR	adapted from Yaung et al. 2014
Church3_top_pPCRISPR	AAACGATGCTGATGCTGAACTGTCG	top oligo for cloning T4 spacer C3 into pCRISPR	adapted from Yaung et al. 2014
Church3_bottom_pCRISPR	AAAACGACAGTTCAGCATCAGCATC	bottom oligo for cloning T4 spacer C3 into pCRISPR	adapted from Yaung et al. 2014
70_puc19_top	AATTCAAAAGTTTTAGGGAATTCTAGGGCGGAGAGGGGA	top oligo for cloning T4 protospacer 1 into puc19	this study
70_puc19_bottom	AGCTTCCCCTCTCCGCCCTAGAATTCCCTAAAACTTTTG	bottom oligo for cloning T4 protospacer 1 into puc19	this study
ModA_puc19_top	AATTCGCCCAGGGAGCCCAAGGCGGAGGGTCAAGATGGA	top oligo for cloning T4 protospacer 2 into puc19	this study
ModA_puc19_bottom	AGCTTCCATCTTGACCCTCCGCCTTGGGCTCCCTGGGCG	bottom oligo for cloning T4 protospacer 2 into puc19	this study
30.2_puc19_top	AATTCTTCACTATCAAAACTTTCTTTATTTTCCTTTGGA	top oligo for cloning T4 protospacer 3 into puc19	this study
30.2_puc19_bottom	AGCTTCCAAAGGAAAATAAAGAAAGTTTTGATAGTGAAG	bottom oligo for cloning T4 protospacer 3 into puc19	this study
11_puc19_top	AATTCTTTATTTATCATATTTATAAATAGAATAAAAGGA	top oligo for cloning T4 protospacer 4 into puc19	this study
11_puc19_bottom	AGCTTCCTTTTATTCTATTTATAAATATGATAAATAAAG	bottom oligo for cloning T4 protospacer 4 into puc19	this study
30_Chruch1_top_puc19	AATTCAACACCACAAATATCGAAAGCAATCAGGTTAGGA	top oligo for cloning T4 protospacer C1 into puc19	adapted from Yaung et al. 2014
30_Church1_bottom_puc19	AGCTTCCTAACCTGATTGCTTTCGATATTTGTGGTGTTG	bottom oligo for cloning T4 protospacer C1 into puc19	adapted from Yaung et al. 2014
30_Church2_top_puc19	AATTCTTCCGATCCGAAGAACTTCCAACCGGTAATGGGA	top oligo for cloning T4 protospacer C2 into puc19	adapted from Yaung et al. 2014
30_Church2_bottom_puc19	AGCTTCCCATTACCGGTTGGAAGTTCTTCGGATCGGAAG	bottom oligo for cloning T4 protospacer C2 into puc19	adapted from Yaung et al. 2014
30_Church3_top_Puc19	AATTCTCACGGTATGGATGCTGATGCTGAACTGTCTGGA	top oligo for cloning T4 protospacer C3 into puc19	adapted from Yaung et al. 2014
30_Church3_bottom_puc19	AGCTTCCAGACAGTTCAGCATCAGCATCCATACCGTGAG	bottom oligo for cloning T4 protospacer C3 into puc19	adapted from Yaung et al. 2014

3.9 References

1. **Luria SE, Human ML.** 1952. A nonhereditary, host-induced variation of bacterial viruses. J Bacteriol **64:**557-569.

2. **Hattman S.** 2009. The first recognized epigenetic signal: DNA glucosylation of T-even bacteriopages. Epigenetics **4:**150-151.

3. **Karam JD, Drake JW.** 1994. Molecular biology of bacteriophage T4. American Society for Microbiology, Washington, DC.

4. **Josse J, Kornberg A.** 1962. Glucosylation of deoxyribonucleic acid. III. alpha- and beta-Glucosyl transferases from T4-infected Escherichia coli. J Biol Chem **237:**1968-1976.

5. **Pratt EA, Kuo S, Lehman IR.** 1963. Glucosylation of the deoxyribonucleic acid in hybrids of coliphages T2 and T4. Biochim Biophys Acta **68:**108-111.

6. **Kornberg A, Baker T.** 1991. DNA Replication. W. H. Freeman and Company, New York.

7. **Warren RA.** 1980. Modified bases in bacteriophage DNAs. Annu Rev Microbiol **34:**137-158.

8. **Loenen WA, Raleigh EA.** 2013. The other face of restriction: modification-dependent enzymes. Nucleic Acids Res doi:10.1093/nar/gkt747.

9. **Gommers-Ampt JH, Van Leeuwen F, de Beer AL, Vliegenthart JF, Dizdaroglu M, Kowalak JA, Crain PF, Borst P.** 1993. beta-D-glucosyl-hydroxymethyluracil: a novel modified base present in the DNA of the parasitic protozoan T. brucei. Cell **75:**1129-1136.

10. **Kriaucionis S, Heintz N.** 2009. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science **324:**929-930.

11. **Barrangou R, Coute-Monvoisin AC, Stahl B, Chavichvily I, Damange F, Romero DA, Boyaval P, Fremaux C, Horvath P.** 2013. Genomic

impact of CRISPR immunization against bacteriophages. Biochem Soc Trans **41:**1383-1391.

12. **Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J.** 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science **321:**960-964.

13. **Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P.** 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science **315:**1709-1712.

14. **Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A.** 1987. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. J Bacteriol **169:**5429-5433.

15. **Tyson GW, Banfield JF.** 2008. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol **10:**200-207.

16. **Jansen R, Embden JD, Gaastra W, Schouls LM.** 2002. Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol **43:**1565-1575.

17. **Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV.** 2006. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct **1:**7.

18. **Pourcel C, Salvignol G, Vergnaud G.** 2005. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology **151:**653-663.

19. **Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E.** 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol **60:**174-182.

20. **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.

21. **Sapranauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V.** 2011. The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res **39:**9275-9282.

22. **Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, Ma E, Anders C, Hauer M, Zhou K, Lin S, Kaplan M, Iavarone AT, Charpentier E, Nogales E, Doudna JA.** 2014. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science **343:**1247997.

23. **Sorek R, Lawrence CM, Wiedenheft B.** 2013. CRISPR-mediated adaptive immune systems in bacteria and archaea. Annu Rev Biochem **82:**237-266.

24. **Westra ER, Swarts DC, Staals RH, Jore MM, Brouns SJ, van der Oost J.** 2012. The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. Annu Rev Genet **46:**311-339.

25. **Wiedenheft B, Sternberg SH, Doudna JA.** 2012. RNA-guided genetic silencing systems in bacteria and archaea. Nature **482:**331-338.

26. **Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E.** 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science **337:**816-821.

27. **Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA.** 2013. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol **31:**233-239.

28. **Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F.** 2013. Multiplex genome engineering using CRISPR/Cas systems. Science **339:**819-823.

29. **Barrangou R, May AP.** 2015. Unraveling the potential of CRISPR-Cas9 for gene therapy. Expert Opin Biol Ther **15:**311-314.

30. **Yin H, Xue W, Chen S, Bogorad RL, Benedetti E, Grompe M, Koteliansky V, Sharp PA, Jacks T, Anderson DG.** 2014. Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype. Nat Biotechnol **32:**551-553.

31. **Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM.** 2013. RNA-guided human genome engineering via Cas9. Science **339:**823-826.

32. **Dupuis ME, Villion M, Magadan AH, Moineau S.** 2013. CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. Nat Commun **4:**2087.

33. **Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, Cradick TJ, Marraffini LA, Bao G, Zhang F.** 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol **31:**827-832.

34. **Yaung SJ, Esvelt KM, Church GM.** 2014. CRISPR/Cas9-mediated phage resistance is not impeded by the DNA modifications of phage T4. PLoS One **9:**e98811.

35. **Lehman IR, Pratt EA.** 1960. On the structure of the glucosylated hydroxymethylcytosine nucleotides of coliphages T2, T4, and T6. J Biol Chem **235:**3254-3259.

36. **Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, et al.** 2009. Real-time DNA sequencing from single polymerase molecules. Science **323:**133-138.

37. **Song CX, Clark TA, Lu XY, Kislyuk A, Dai Q, Turner SW, He C, Korlach J.** 2012. Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nat Methods **9:**75-77.

38. **Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW.** 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods **7:**461-465.

39. **Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC, Jabado OJ, Deikus G, Clark TA, Luong K, Murray IA, Davis BM, Keren-Paz A, Chess A, Roberts RJ, Korlach J, Turner SW, Kumar V, Waldor MK, Schadt EE.** 2012. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nat Biotechnol **30:**1232-1239.

40. **Schadt EE, Banerjee O, Fang G, Feng Z, Wong WH, Zhang X, Kislyuk A, Clark TA, Luong K, Keren-Paz A, Chess A, Kumar V, Chen-Plotkin A, Sondheimer N, Korlach J, Kasarskis A.** 2013. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. Genome Res **23:**129-141.

41. **Clark TA, Spittle KE, Turner SW, Korlach J.** 2011. Direct detection and sequencing of damaged DNA bases. Genome Integr **2:**10.

42. **Doutriaux MP, Wagner R, Radman M.** 1986. Mismatch-stimulated killing. Proc Natl Acad Sci U S A **83:**2576-2578.

43. **Deschavanne P, Radman M.** 1991. Counterselection of GATC sequences in enterobacteriophages by the components of the methyl-directed mismatch repair system. J Mol Evol **33:**125-132.

44. **Au KG, Welsh K, Modrich P.** 1992. Initiation of methyl-directed mismatch repair. J Biol Chem **267:**12142-12148.

45. **Hattman S.** 1970. DNA methylation of T-even bacteriophages and of their nonglucosylated mutants: its role in P1-directed restriction. Virology **42:**359-367.

46. **Wilson GG, Young KY, Edlin GJ, Konigsberg W.** 1979. High-frequency generalised transduction by bacteriophage T4. Nature **280:**80-82.

47. **Young KK, Edlin GJ, Wilson GG.** 1982. Genetic analysis of bacteriophage T4 transducing bacteriophages. J Virol **41:**345-347.

48. **Young KK, Edlin G.** 1983. Physical and genetical analysis of bacteriophage T4 generalized transduction. Mol Gen Genet **192:**241-246.

49. **Lin H, Black LW.** 1998. DNA requirements in vivo for phage T4 packaging. Virology **242:**118-127.

50. **Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, van der Oost J, Brouns SJ, Severinov K.** 2011. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U S A **108:**10098-10103.

51. **Wang T, Wei JJ, Sabatini DM, Lander ES.** 2014. Genetic screens in human cells using the CRISPR-Cas9 system. Science **343:**80-84.

52. **Rollins MF, Schuman JT, Paulus K, Bukhari HS, Wiedenheft B.** 2015. Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa. Nucleic Acids Res doi:10.1093/nar/gkv094.

53. **Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E.** 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature **471:**602-607.

54. **Benchetrit LC, Bachrach U.** 1980. Studies on phage internal proteins. VI. Interaction of bacteriophage T4 internal proteins with T4 DNA in vivo and in vitro. Rev Bras Pesqui Med Biol **13:**41-45.

55. **Rifat D, Wright NT, Varney KM, Weber DJ, Black LW.** 2008. Restriction endonuclease inhibitor IPI* of bacteriophage T4: a novel structure for a dedicated target. J Mol Biol **375:**720-734.

56. **Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR.** 2013. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. Nature **493:**429-432.

57. **Tetart F, Repoila F, Monod C, Krisch HM.** 1996. Bacteriophage T4 host range is expanded by duplications of a small domain of the tail fiber adhesin. J Mol Biol **258:**726-731.

58. **Pattanayak V, Lin S, Guilinger JP, Ma E, Doudna JA, Liu DR.** 2013. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat Biotechnol **31:**839-843.

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

60. **Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD.** 2011. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res **21:**1616-1625.

61. **Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD.** 2013. Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A **110:**12450-12455.

62. **Karam JD, Drake JW, Kreuzer KN.** 1994. Molecular Biology of Bacteriophage T4. ASM, Washington, D. C.

63. **Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, Fomenkov A, Turner SW, Korlach J, Roberts RJ.** 2012. The methylomes of six bacteria. Nucleic Acids Res **40:**11450-11462.

CHAPTER 4: Phage predation in the human gut microbiome

Alexandra Bryson*, Christel Chehoud*, Anatoly Dryga, Abigail Lauder, Jacque Young, Seth Zost, Elizabeth Loy, Eric Chen, Hongzhe Li, Richard Roberts, Samuel Minot, Tyson Clark, Jonas Korlach, Scott Sherrill-Mix, Frederic D. Bushman

*these authors contributed equally

4.1 Contributions

My contributions to this paper include the DNA modification analysis, 16S bacteria analysis, Jaccard analysis, gold particle staining of phage, writing the paper, and working with Christel Chehoud to organize qPCR, staining, and sequencing data collected by other coauthors.

4.2 Abstract

Studies of the human microbiome have specified the types of micro-organisms present but interactions between organisms are less well studied. Here we report analysis of the dynamics of gut bacteriophage and their hosts in one healthy human male over four years of sampling. We used multiple sequencing methods, including long-read single-molecule sequencing, to specify many of the major phage and bacterial lineages present, and their changes in abundance over time. RNA viruses were less common than DNA viruses and corresponded to probable transients in food. Analysis of single-molecule

sequencing data identified motif-specific covalent DNA modifications in 73% of phage contigs and 56% of bacterial contigs. The patterns of shared base modifications between phage and bacteria could be used to propose hosts for some of the phage studied, which allowed querying of the pairs for evidence of predator-prey cycles. Assessment of viral numbers using particle counts or qPCR for phage genomes standardized with metagenomic data suggested population sizes ranging from 2.8×10^{10} to 3.7×10^{12} viral particles per gram of stool. These estimates allowed investigation of predation rates--phage must multiply at rates that maintain the steady state density in gut despite the continuous outward flow of luminal contents. Using this assumption, we were able to estimate predation rates ranging from 0.2% to 14% of bacteria killed by phage predation per day in the human gut. Although these data emphasize that different approaches to estimation yield a wide range of values, we can use these approaches to begin to investigate the dynamics of predation in the human gut—for example to analyze variation among human populations, disease states, and responses to therapies such as antibiotics.

4.3 Introduction

The composition of the human gut microbiota has now been studied extensively, but dynamics associated with growth and predation are much less well understood. Studies of bacteria in marine and freshwater environments suggest that phage often outnumber their hosts, and that substantial fractions of the bacterial population are killed per day by phage predation(1-3). These high turnover rates must have a strong effect on the composition and dynamics of bacterial communities, but little data is available for human gut.

Studies of bacteriophage populations are complicated by the extremely high numbers and diversity of global phage populations. Earth is believed to host 10^{31} viral particles, and this is paralleled by very large numbers of viral types. In contrast, the NCBI viral database only contains 6,693 reference viral genomes (NCBI accessed 3/10/16). Consequently, sequence samples of environmental viral populations typically show only modest sporadic matches to database viral genomes (though for viruses that are human pathogens the coverage is much greater). Another challenge is that bacteriophage from vertebrate guts have typically proven difficult to culture outside the gut environment (4), further limiting experimental characterization.

However, it is possible to investigate bacteriophage populations in the vertebrate gut by purifying viral particles from stool, sequencing the encapsulated genomes, assembling reads to generate contigs corresponding to complete or partial genomes, then interrogating the contigs in longitudinal analysis. Several studies have used this approach to characterize phage communities in humans (5-8) or reconstructed human-derived communities in gnotobiotic mice (4).

Here we take such an approach to investigate longitudinal dynamics over four years in a closely-studied gut phage community from a healthy human male (subject 1014;(5)). Previously viral communities were studied using Illumina short read sequencing over 2.5 years (5). Here we acquired single-molecule sequencing data on the phage and bacterial communities, which we used to improve contig assembly. The single-molecule data also allowed quantification of covalent DNA modification, which provide a novel means of associating potential phage-host pairs. We used metagenomic sequence data, qPCR, and fluorescence microscopy to estimate the sizes of the phage and bacterial populations, and found that estimates ranged widely depending on the quantification methods used. Modeling the gut as a steady state with our inferred phage bacterial dynamics predicts that between 0.2% and 14% of gut bacteria are killed per day by phage predation.

4.4 Results and Discussion

4.4.1 Sequence data acquisition

DNA sequence data to characterize phage and bacterial communities in subject 1014 were derived from several sources. We previously reported phage contigs from 2.5 years of sampling from this subject derived from Illumina HiSeq short-read sequencing. We also reported deep shotgun sequence analysis for whole stool, allowing comparison to the full gut community. Here we add single-molecule sequencing data using the Pacific Biosciences technology, and further Illumina MiSeq sequencing of viral fractions of later time points out to 4 years. To characterize bacterial prey species further, whole stool DNA was also analyzed by PacBio sequencing, and 16S rRNA gene tag sequencing was used to characterize bacterial communities over the four year time span plus an additional year. A complete list of samples studied is in Table S1.

For all virome samples, viral particles were purified by filtration, and preparations treated with nucleases to remove free nucleic acids. Samples were treated with chloroform during purification, which disrupts membranes and is important for achieving high purity, so only non-enveloped viruses were recovered. We and others have observed that enveloped viruses are relatively uncommon in stool (5, 9, 10), so we expect only modest losses due to this step. Virome samples were verified to be depleted in 16S rRNA gene copies, which

indicates minimal bacterial DNA contamination. To characterize the RNA virome, RNA was purified from particles from two time points, reverse transcribed and analyzed by Illumina DNA sequencing.

4.4.2 DNA virome contigs

A total of 3358 contigs corresponding to DNA viruses were generated from the merged Illumina and PacBio sequence data (information on contigs is summarized in Table S2). For the merged contigs, the N50 was 25,633. The maximum length was 217,304. Contigs were aligned to viral databases, and 16% found annotations based on matching ORFs (open reading frames) to ORFs of reference viral genomes. Of these, the major groups were Siphoviridae (57%), Myoviridae (22%), and Podoviridae (7%). Contigs annotated as Microviridae yielded large numbers of reads after GenomiPhi amplification of DNA samples, as expected because these small single stranded circular DNA viruses are preferentially amplified by this method.

No convincing matches were detected to viruses infecting animal cells, though this pipeline has yielded well-known animal-cell viruses in studies of other sample types (11, 12). We thus infer that most or all of the contigs detected here likely derive from bacteriophage.

Many viral genomes are either circular or terminally redundant—thus an indication of completion of a viral contig sequence is closure of the sequence as a circle. Of the 3358 contigs, 51 closed as circles. Thus we infer that most of these are complete sequences, though apparent circularity could also be obtained if genomes contained internal direct repeats with high sequence identity. The linear contigs represented either linear viral genomes or incomplete genomes.

Viral ORFs were identified and the encoded proteins assessed for similarity to proteins of known viruses (Figure S1). Genes commonly annotated as phage structural proteins (capsid, baseplate, tail, portal and others), functions important in nucleic acid manipulation (recombinase, resolvase, terminase, repressor and others), and proteins important in host cell manipulation (lysozyme, beta lactamases, and restriction-modification). Only 22% (2407 / 10899) of viral ORFs found any taxonomic annotation by comparison using a BLAST e-value threshold of 10^{-5}.

4.4.3 RNA virome contigs

Viruses containing RNA genomes were interrogated by purifying RNA from particle preparations from two time points, reverse transcription, and sequence analysis. Reads were quality filtered and aligned to a viral database. The major lineages detected based on extent of coverage of the target genome

all annotated as *Tobamoviruses*, which are non-enveloped helical plant viruses. The *Tobamovirus* genus contains species such as the Tobacco mosaic virus, Pepper mild mottle virus, Tomato mosaic virus, and Rehmannia mosiac virus. At the first time point queried (day 181), 2,768 reads aligned to Pepper mild mottle virus, and 545 aligned to Tomato mosaic virus. At the second time point (day 852), 548 reads aligned to Rehmannia mosiac virus and 862,802 aligned to Tomato mosaic virus. No convincing alignments were detected to RNA viruses infecting human cells or RNA phage, though human RNA viruses were readily detected in spiked-in positive controls (data not shown). We thus infer that the major RNA viruses in our fecal specimens are non-enveloped plant viruses ingested with food and not long term residents of the gut virome, consistent with published studies of other human subjects (13).

4.4.4 DNA Phage populations analyzed longitudinally

DNA Phage and bacterial contigs detected over the 4-year period of sampling are shown in Figure 1. The phage populations were analyzed by Illumina shotgun sequencing of GenomiPhi-treated samples, so viruses with small, circular genomes such as Microviridae are enriched. Many of the viral types detected were seen at multiple time points, suggesting stability in the major types present. During the first year of sampling, one ~6 Kb contig annotating as

Microviridae predominated, but by day 851 this variant was replaced as the major form by another ~6Kb variant that could not be classified. To evaluate stability in the phage population, we scored shared community membership between all pairwise time point comparisons using Jaccard index values. Figure S2A shows the Jaccard index for each pairwise comparison and indicates that community composition changed only slowly over the four years studied.

Bacteria were characterized by sequencing 16S rRNA gene tags (V1V2 region). Overall, the predominant families were present from the start to the end of the study. Prominent lineages included the *Firmicutes* families *Ruminococcaceae*, *Lachnospiraceae*, and *Clostridiaceae*. Lesser amounts were seen for *Bacteroidetes* families *Bacteroidaceae*, *Poryphoromonadaceae*, and *Prevotellaceae*. Only trace amounts of *Proteobacteria* were detected. Thus the bacterial community is rich in anaerobes, as expected for a healthy adult (14). We used Jaccard index values to score shared bacterial community membership between all pairwise time points (Figure S2B) which suggested that shared community membership was generally stable but changed slowly over time.

4.4.5 DNA modification analyzed in phage and bacterial metagenomic samples

Bacteria commonly encode nucleases targeting DNA of phage, plasmids and other invaders. These include restriction enzymes (15-28), CRISPR/Cas systems (29-35), and other nucleases (36, 37). DNA phages recruit DNA modifying enzymes from their hosts to protect their DNA, so that they are insensitive to host nucleases, and in addition encode further modifying enzymes to protect their DNA. Previous studies have specified more than ten types of chemical modification in phage DNA (38).

Here we take advantage of modification patterns to propose associations between phage and their bacterial hosts. In metagenomic samples, it is usually not possible to determine which phage infect which bacteria in the population. We thus sought to associate phage-host pairs in the human gut microbiome though matching modified motifs in our single-molecule Pacific Biosciences sequencing. In this technology, single polymerase molecules traverse a single DNA molecule, but the presence of many forms of covalent DNA modification slows the kinetics, allowing modified base detection. We thus applied single-molecule sequencing to metagenomic phage and whole stool fractions, and scored covalent DNA modifications common to the two, reasoning that sharing of patterns between phage and hosts would indicate potential phage-host pairs.

Single-molecule real-time sequencing has previously been used to identify modifications of more than 230 individual prokaryotic genomes(39) (40)--this is the first study to analyze paired modification profiles in bacteria and phage in human metagenomic samples.

We developed a pipeline for statistical analysis of kinetic sequencing data, in which modified bases were scored in either motifs known to be modified from previous studies (41-44) or all k-mers of lengths 2-5 nucleotides. To identify modified bases, we fitted a Gaussian mixture model on IPD ratio, quality score, and sequencing coverage. We assumed that the IPD ratios from the modified bases and unmodified bases follow two separate, normal distributions. By fitting the Gaussian mixture model, we obtained a posterior probability indicating how likely the base is to be modified. We fitted the Gaussian mixture model for G, A, T, and C separately. We compared these to a list of 467 DNA motifs previously known to be modified (41-44). Each nucleotide within each k-mer was analyzed for modification, yielding 6366 k-mer-position combinations. The occurrences of each known motif and k-mer were identified in every bacterial and phage contig, and we recorded the mean IPD and mean posterior probability of modification for each motif of interest. We used an FDR correction on the posterior probability to establish a list of modified motifs. Because our method evaluates each motif independently on each contig, some larger k-mer motifs may show up as

modified because they contain an embedded smaller motif. We thus parsed larger modified motifs for embedded smaller motifs within the same contig, and recorded the smaller motifs.

As a control to validate our motif finding algorithm, we used a set of T4 phage variants studied previously (45). T4 DNA was compared containing glycosylated hydroxymethylcytosine T4(glc-HMC), hydroxymethylcytosine T4(HMC) or unmodified cytosines T4(C) (Figure S3). In all three strains, adenines are methylated within G**A**TC (A bolded to emphasize modification) motifs. The G**A**TCs in T4(glc-HMC) are know to be modified less frequently than T4(HMC) and T4(C)(45, 46), probably due to crowding on the DNA, and this was seen in results from our pipeline. Modifications at G**A**TC were detected in T4(C) and T4(HMC), but not T4(glc-HMC) because T4(glc-HMC) only modifies some G**A**TC motifs and our pipeline reports motifs with complete modification. Not every C in the T4(glc-HMC) and T4(HMC) appear modified by SMRT sequencing, so modified Cs also are not identified here. We tested a whole genome amplified (WGA) version of T4(glc-HMC), which contains no DNA modifications and this was confirmed in our data.

Contigs where a motif occurred at least three times were considered for analysis. Of the 467 known motifs we queried, 56 were identified among 78 phage genomes (Figure 2). Among the bacterial contigs, 1,868 had at least one

of 138 known motifs (Figure 3). The bacteria and phage shared 31 known modified motifs. In the k-mer analysis, 1,004 modified k-mer motifs were discovered in 288 phage contigs, and we observed 2,285 modified k-mer motifs in 2,526 bacterial contigs (Table S3). Merging the known motif and k-mer data revealed that 73% of phage contigs and 56% of bacterial contigs contain fully modified motifs. Table S3 contains the complete list of contigs and their associated modified motifs. This dataset identifies 2,804 candidate novel modified motifs.

4.4.6 Phage-host pairs linked via DNA modifications

We next used covalent DNA modification data to link bacteria and phage pairs, based on the idea that phage and bacterial DNAs in the same cell will have been exposed to the same DNA modifying enzymes and thus share modification patterns. Overall, 443 modified motifs were shared between phage and bacteria (found in Table S3) suggesting many possible phage-host pairs. Figure 4 depicts a subset of these possible phage-host pairs, where phage contigs were filtered by requiring matching to one ORF annotated as a phage protein or showing the top NCBI blast hit to be a phage genome. Motifs with G**A**TCs were removed from this analysis because they are widely distributed biologically and common

among bacteria and phage, making it difficult to determine specific phage-host pairs using this motif.

Five phage contigs could be linked to bacterial contigs though five motifs (Figure 4). Bacteria of the Bacteriodes, Anaerostipes, and Enterococcus genera could be linked to specific phage contigs, as well as four unattributed bacteria. In one case (CAAAA**A** motif) the phage motif was found in bacteria of multiple phyla, indicating either interphyletic mobility of restriction/modification operons or convergent evolution of enzyme specificities. The CAAAA**A** motif has previously been reported modified in *Clostridium difficile* (47). One of the bacterial contigs that contained the CAAAA**A** motif annotated as *Lachnospiraceae*, which belongs to the same order *Clostridialese*.

For each of these pairs, we assessed possible Lotka-Volterra predator-prey cycles in longitudinal abundance (48, 49). In this analysis, we assumed that increases in predator would precede decreases on prey species, and vice versa. Such cycles can be observed in a plot comparing predator and prey abundances as a counter-clockwise progression of longitudinal samples (50). Thus we searched for an enrichment of left turns in the prey-predator plot. In contrast, reverse (clockwise) cycles have recently been proposed to occur in natural systems where there is co-evolutionary tradeoff of costly offense and defense between predator and prey (50).

Samples were compared for the periods of sampling where three or more contiguous days were acquired, days 180-183, 851-853 and 879-883. As three points are required to determine turn angle, we had six total turns to compare to the null hypothesis of a 50% chance of a left or right turn. None of the pairs in Figure 4 showed such cycles. In the larger set of possible phage/host pairs, aggregating the bacterial data at the Family level (n=21), two examples of clockwise cycling were detected, and one example of counterclockwise cycling. However, these results are not significant after an FDR correction for multiple comparisons, emphasizing that future experimental designs should incorporate more time points and time scales of analysis.

4.4.7 Dynamics of phage and bacteria in the human gut

Given counts of phage, their hosts, and the ratio between them, then we can make initial estimates of the predation rate. We thus devised two ways of counting phage and bacteria, and a third way of estimating the phage:host ratio from metagenomic sequence data.

In the first approach, we used quantitative PCR (qPCR) to estimate numbers. DNA was purified from weighed stool samples, then 16S qPCR was used to measure the total numbers of bacterial 16S rRNA gene copies per gram

of stool. The number of cells was then inferred assuming four 16S rRNA gene copies per cell (51), yielding a median of 1.9×10^{13} (Table 1).

For phage, we took advantage of qPCR analysis of phage genomes combined with information on the proportions of each phage in metagenomic data. We devised qPCR assays for 8 phage contigs and measured the copy numbers in DNA from weighed stool samples (Table S4). The total phage population size was then estimated by dividing the qPCR estimate by the proportion of the quantified phage in the metagenomic analysis of unfractionated stool DNA at each of three time points. All estimates were averaged to yield a global median of 3.7×10^{12} for the phage population size (Table 1).

The second method relied on classical staining of aliquots of weighed samples with fluorescent dyes, DAPI for bacteria and cyber gold for phage. This yielded 2.8×10^{11} cells for bacteria and 2.8×10^{10} particles for phage. The numbers for phage reported here are ~10-fold higher than those reported previously by Kim et al (52). It is unknown whether this reflects a true difference between subjects or methodological differences. We suspect that the cyber gold staining for phage undercounts particles due to inefficient staining of phage with smaller genomes and single stranded DNA phage.

We calculated the geometric mean values for phage and bacteria using each of the two methods, yielding phage:host ratios of 1:10 to 1.3:1 (Table 2).

Another estimate of the phage/host ratio could be made based solely on metagenomic sequence data. As described above, we have a well-studied contig set describing phage in subject 1014 and shotgun metagenomic sequence data from unfractionated stool at three time points. We could thus ask what fraction of the total metagenomic reads are contributed by sequences aligning to phage contigs. We found that an average of 4% of the total reads matched phage contigs (range 3.5 to 4.4%). The bacteria are estimated, on average, to have genome sizes 176 times longer than phage (53). Scaling the percentage of reads by the ratio of genome sizes gives (0.04 phage reads)(176)=7.0 for the phage/host ratio (Table 2). The proportion of phage called in the metagenomic ratio estimate may be high, because some of the phage sequences will be present as prophages in bacterial genomes. The metagenomic ratio could be low due to possible contamination of phage sequences with bacterial sequences, and also due to addition of phage not yet represented as contigs. In summary, our estimated phage:host ratios ranged from 1:10 to 7:1.

4.4.8 Estimating the phage predation rate

The data in Table 1 allow estimation of the rate of predation by phage on gut bacteria. Replication rates of phage and bacteria in gut must be matched, so that the population is replaced continuously despite material flowing out of the

gut. To estimate predation rates, we need to know burst sizes of gut phage, and the rate of flow of material through the gut. Using these data, we can estimate the proportion of bacteria killed per day by phage to meet the required replacement rate to maintain a steady state.

From environmental data on phage replication in marine and lake ecosystems, the average size of a phage burst is 50 particles/burst (range 25^{th} percentile 27, 75^{th} percentile 75; Table S5). Values for burst sizes are higher for laboratory measurements using bacteria grown in rich medium (170 phage per burst range 25^{th} percentile 115, 75^{th} percentile 260; Table S5), but we favor use of the environmental estimates for gut to reflect phage replication under non-optimal growth conditions. The transit time in the individual studied was estimated at 24 hours.

To estimate predation rates, we used three different values for the phage host ratio (Table 2). Using targeted qPCR, we estimate that there are 8.3×10^{12} phage and 1.6×10^{12} bacteria, though with wide confidence intervals. To produce this number of phage, given the average burst size of 50, 1.7×10^{11} bacteria must die each day. This corresponds to 10% of all gut bacteria killed per day by phage predation. Using staining data, we estimate there are 2.8×10^{10} phage and 3.0×10^{11} bacteria (Table 2) and conclude that 0.2% of all bacteria are killed due to phage predation each day. Using the metagenomic ratio of phage to

bacteria of 7:1, we obtain a predation rate of 14% of bacteria killed per day. Thus our values range from 0.2% to 14% of bacteria killed per day depending on the estimation method used.

4.4.9 Summary and prospectus

Here we present a four-year study of the dynamics of phage and bacterial populations in the gut of one adult human male. We carried out single-molecule sequencing of phage and bacterial populations, augmenting a large short-read data set acquired previously (5), which allowed us to characterize new aspects of the dynamics of phage replication. Analysis of DNA modification using the single-molecule sequencing data indicated that 73% of phage genomes and 56% of bacterial genomes were fully modified at a motif. We expect the percent of modified genomes would increase if we considered motifs that are only partially modified. A total of 443 modified motifs were shared between phage and bacteria (Table S3), suggesting potential phage:host paires, and five phage-host pairs could be called using conservative criteria. Predation rates could be assessed using data on phage and bacterial population sizes, burst sizes of phage in the environment, and transit time of material through the gut, leading to the suggestion that between 0.2 and 14% of bacteria in gut are killed by phage predation per day. Several factors could have affected our estimates. For

example, purification of phage particles from stool samples is unlikely to be 100% efficient. It has been suggested that phage decay rates can be substantial in natural environments (54) —if this is true in gut, our proposed production rates are minimal estimates. Conversely, it is possible that defecation and exposure of anaerobic bacteria to oxygen may result in phage induction and particle production. If so, our measurements are overestimates. Our estimates of predation in gut varied associated with the method used, but all estimates suggested substantial predation rates. Although methods of quantification need further refinement, these data pave the way for analysis of phage:host dynamics in medically important settings such as bacterial infections, inflammatory autoimmune diseases, and antibiotic use.

4.5 Methods

Research subject. This work was carried out under an IRB approved protocol (5). Transit time was self-reported by the research subject and parallels measurements with Sitz markers on healthy adults eating a high fiber diet (55).

Single-molecule sequencing. For single-molecule sequencing, purified DNA was fragmented to an average size of 1.5 kb via adaptive focused acoustics

(Covaris, Woburn, MA). SMRTbell template sequencing libraries were prepared as previously described (5). Sequencing was carried out on an *RS II* (Pacific Biosciences, Menlo Park, CA) using P4/C2 sequencing chemistry and standard protocols for large insert libraries. Consensus sequences were generated using Quiver.

Sequencing of bacterial 16S rRNA gene tags. DNA was extracted from fecal samples in triplicate using the MBIO powersoil kit. The V1V2 region of the 16S rRNA gene fragment was amplified using Golay-barcoded universal primes BSF8(27F) and BSR357(228R), listed in Table S4. The adaptors added to the 16S specific primers allows the amplicons to be sequenced using the Illumina MiSeq and HiSeq platforms. Each PCR reactions included 0.19uL of AccuPrime Taq DNA Polymerase High Fidelity (Thermo Fisher Scientific Inc, Waltham, MA), 7.21 µL PCR-grade water, 2.5 µL 10X buffer II, 5 µL of each forward and reverse primer (2 µM), and 5 µL template DNA. PCRreactions were prepared in a PCR clean room. Reactions were run on an Applied Biosystems GeneAmp PCR System 9700 (Thermo Fisher Scientific Inc, Waltham, MA) with the following cycling conditions: initial denaturation at 95°C for 5 min followed by 30 cycles of denaturation at 95°C for 30 seconds, annealing at 56°C for 30 seconds, and extension at 72°C for 90 seconds, with a final extension of 8 min at 72°C.

Amplicions were pooled and bead purified using Agencourt AMPure XP (Beckman Coulter, Indianapolis, IN) with the manufacturer's protocol. Reaction products were sequenced using the Illumina MiSeq technology. The 16S rRNA gene reads were annotated using QIIME.

Annotation of Pac Bio Sequenced Bacterial contigs. Taxonomic classification of bacterial contigs was performed with BROCC pipeline (56). Contigs were aligned against NCBI nt database with blastn program(e value = 1e-5, max_target_seqs = 100 and outfmt=7). BROCC then uses BLAST sequence alignment results for taxonomic annotation by first filtering BLAST hits for sufficient coverage and identity and then uses voting to classify contigs for required taxonomical level. Minimum coverage for the hit (min_cover) was 20% and minimum identity of the hit was 20% to pass quality filtering for the alignment. Minimum identity for classification for species and genus levels was set to 60 and 40% respectively. Standard taxonomic ranks has been assigned with NCBI taxonomy database and are used in the further analysis. Details of the BROCC filtering, voting and parameters description can be found in (56) and source code for the BROCC pipeline is available at [https://github.com/kylebittinger/brocc].

Matching Phage-Host Pairs by Modified Motifs. 433 motifs were shared between bacteria and phage contigs. The subset shown in figure 4 was stringently filtered to find the best matches. Phage contigs were required to host one ORF annotated as a phage protein or showing the top NCBI blast hit to be a phage genome. Motifs with G**A**TCs were removed from this analysis because they are widely distributed biologically and common among bacteria and phage, making it difficult to determine specific phage-host pairs for this motif. Motifs from our list of previous known motifs that occurred at least three times within a contig were used in this analysis. A slightly higher cutoff of five motif occurrences per contig was used for k-mer motifs. This was done to be more conservative in determining novel modified motifs. Bacterial contigs were annotated with BROCC.

Quantitative PCR. Quantitative PCR was carried out as previously described using the Syber green method for bacteriophage and Taqman for 16S rRNA genes using the primers described in Table S4. For each amplicon, amplification products were cloned into bacterial plasmids and quantified for use as standards in the qPCR reactions.

Counting phage and bacteria using fluorescence staining. Phage DNA was isolated from viral particles as previously described by Minot et al.(5). Fluorescent staining of phage particles was carried out as described in (57). Bacteria from stool was quantified using DAPI-staining and visualized using flourescent microscopy (58).

Known motifs and k-mers. Lists of modified motifs from NEB REBASE (http://rebase.neb.com/rebase/rebase_methylase_recseqs.txt) and R. Roberts (unpublished) were combined to form the known motif database. Motifs listed as <genuine>y indicates modified motifs that have been experimentally validated or identified with high confidence through Pac Bio Sequencing. Only motifs marked <genuine>y were used from this list. Another list of known motifs came from the NEB REBASE website in the list of all enzymes, sorted by organism (http://rebase.neb.com/rebase/rebase.files.html). The motifs are listed by restriction recognition sites where the sequence is targeted by a nuclease/methylase pair. These lists were merged and redundant motifs were removed. Only motifs with one modification site per motif were analyzed.

Statistical analysis to identify sites of modification in metagenomic data. We first filtered out low quality bases and contigs by 1) removing the 100 bases at the beginning and the end of each contig, 2) filtering out short contigs with length less than 1000 bp, 3) removing contigs with extreme GC content (GC%<15% or GC%>85%) and removing contigs with more than 50% low quality bases. Then we fit a Gaussian mixture model on the IPD ratio. For each base, there are two possible modification states, i.e. the base is either modified or unmodified. We assume that the IPD ratio from the modified bases and unmodified bases follow two separate normal distributions. Each normal distribution has its own mean and standard deviation. By fitting the Gaussian mixture model, we obtained a posterior probability indicating how likely the base is to be modified. We fitted the Gaussian mixture model for A, T, C, G separately and the posterior probability was adjusted by FDR control. To identify modified motifs, the occurrences of known motifs and k-mers were identified and recorded for each contig along with the median IPD and median probability of being modified for the base of interest within the motif. An FDR Benjamini hochberg correction was applied to determine the list of modified motifs in each data set (phage, bacteria, or T4 strains).

Contig assembly and annotation. Three sets of contigs were combined: 1)

Pacific Bioscience unitigs, 2) Illumina HiSeq contigs previously published(5) and 3) Illumina MiSeq contigs built from a single time point (10). These contigs were merged using Mimimo, an overlap consensus assembler, with default parameters. The finalized merged contigs were annotated by length, circularity, open reading frames, putative viral family classification, and presence of integrase genes, as previously described in Chehoud et al., 2015 (10). The viral contig summary and annotation is in Table S6. Bacterial contigs were annotated using BROCC (56). All of the bacterial genera found in the bacterial contigs by BROCC annotation were also identified in the 16S rRNA gene tag analysis in Figure 1.

4.6 Figures

Figure 1

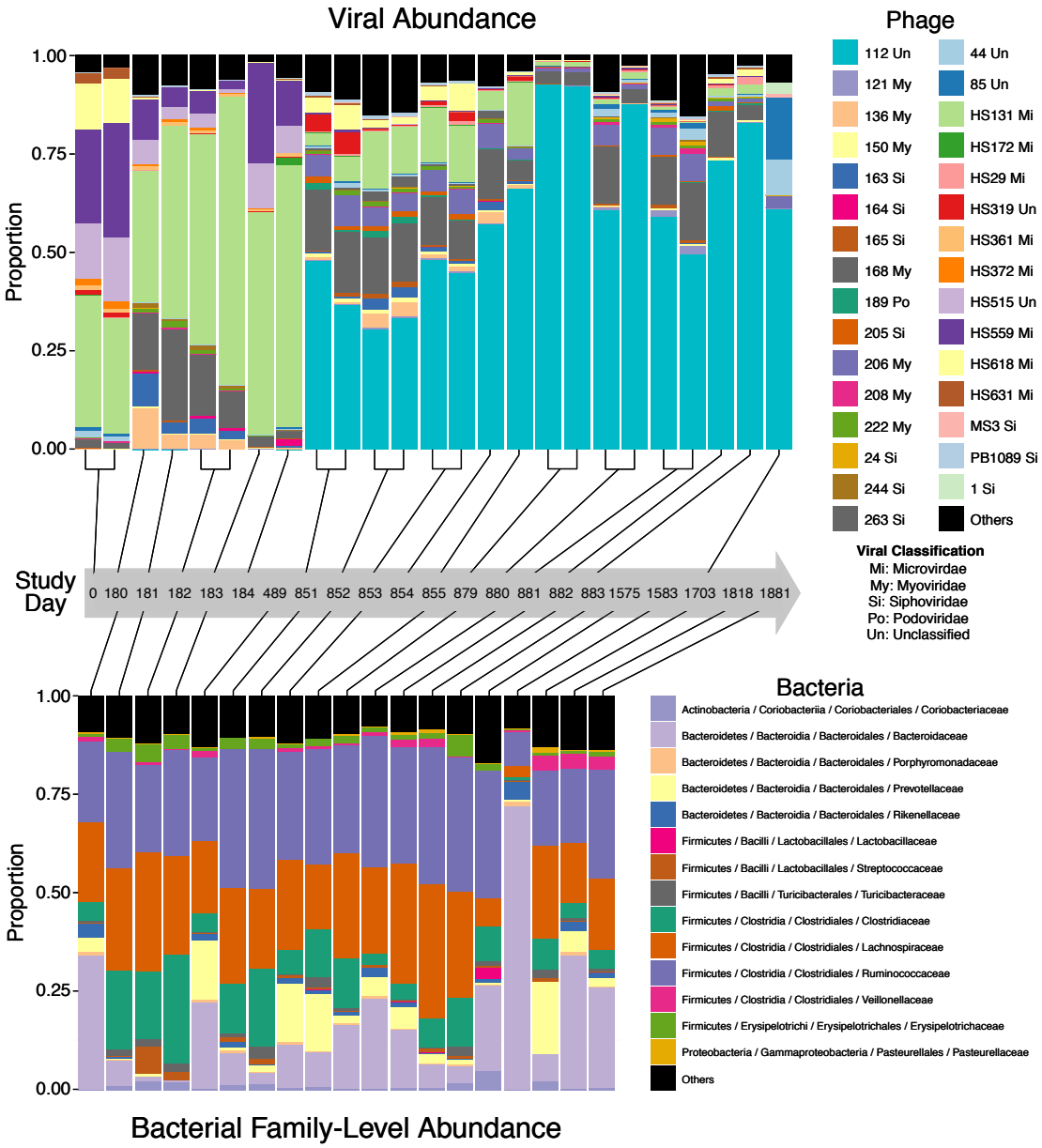


Figure 1. Longitudinal variation in phage and bacterial communities over four years of sampling. A) Longitudinal abundance of phage contigs. The abundance of phage contigs is shown in

relative proportions over time. The gray arrow indicates sequential days that fecal samples were collected. Sample collection began on day zero and ended on day 1881. Solid lines connecting the plot to the sample timeline indicate time points sequenced for viral particles using shotgun metagenomics. Family level classification was assigned when possible. B) Longitudinal abundance of bacterial contigs. The abundance of bacterial families over time is shown in relative proportions. Solid lines connecting the plot to the sample timeline indicate time points where 16s sequencing was done. Family level classification was assigned using QIIME.

Figure 2

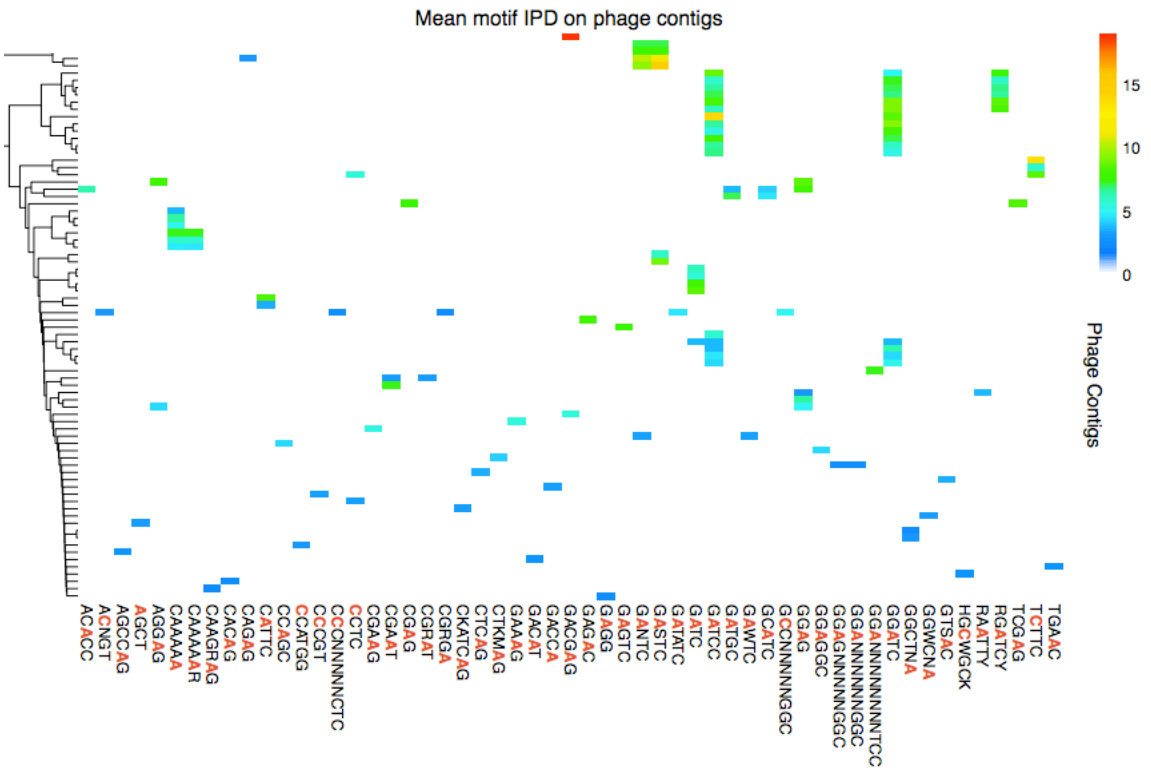


Figure 2. Sites of DNA modification in phage DNA inferred from single-molecule sequencing data. The heat map shows the mean IPD value for individual phage contigs (rows) at the modified base (denoted in red) within the motifs listed in the columns. The motifs depicted here come from the list of previously studied motifs.

Figure 3

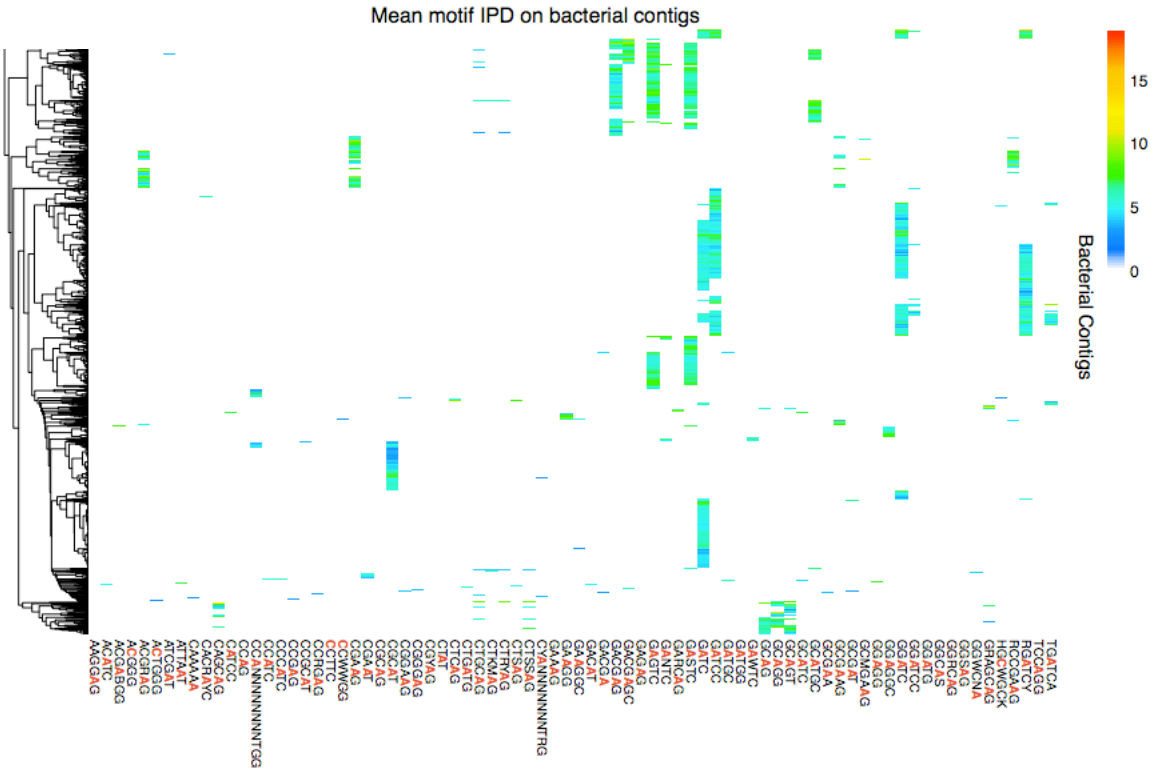


Figure 3. Sites of DNA modification in bacterial DNA inferred from single-molecule sequencing data. The heat map shows the mean IPD value for individual bacterial contigs (rows) at the modified base (denoted in red) within the motifs listed in the columns. Only modified motifs that occurred in more than two bacterial contigs are show here. The motifs depicted here come from the list of previously studied motifs.

Figure 4

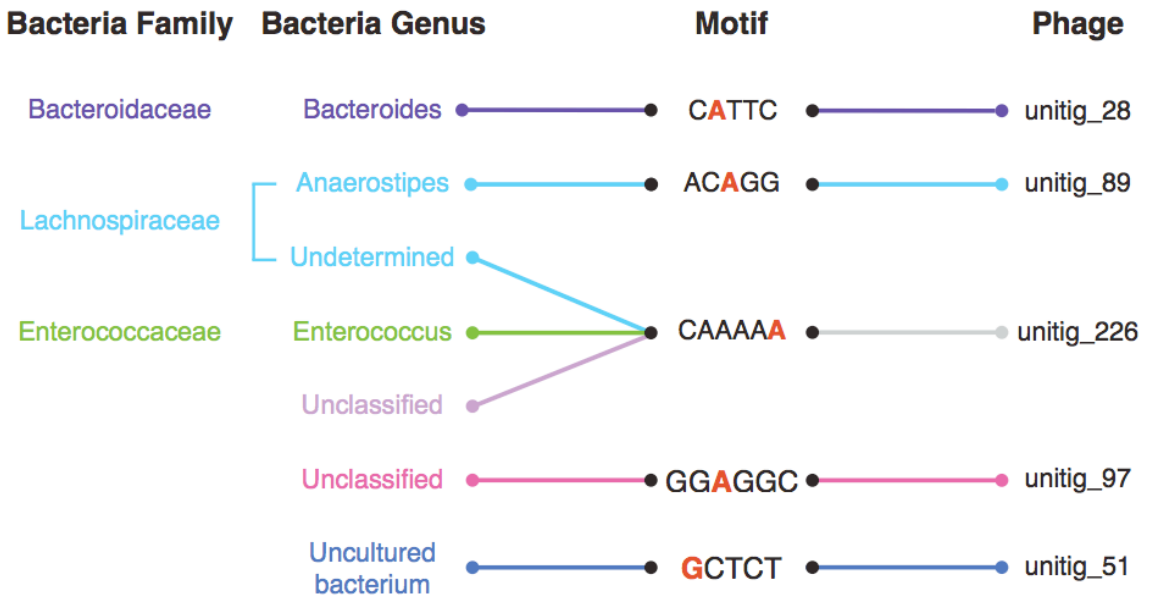
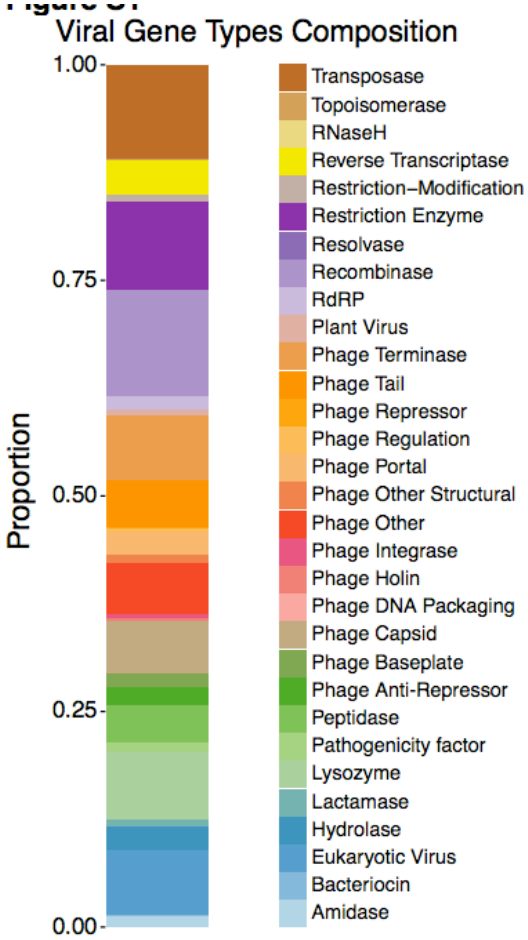


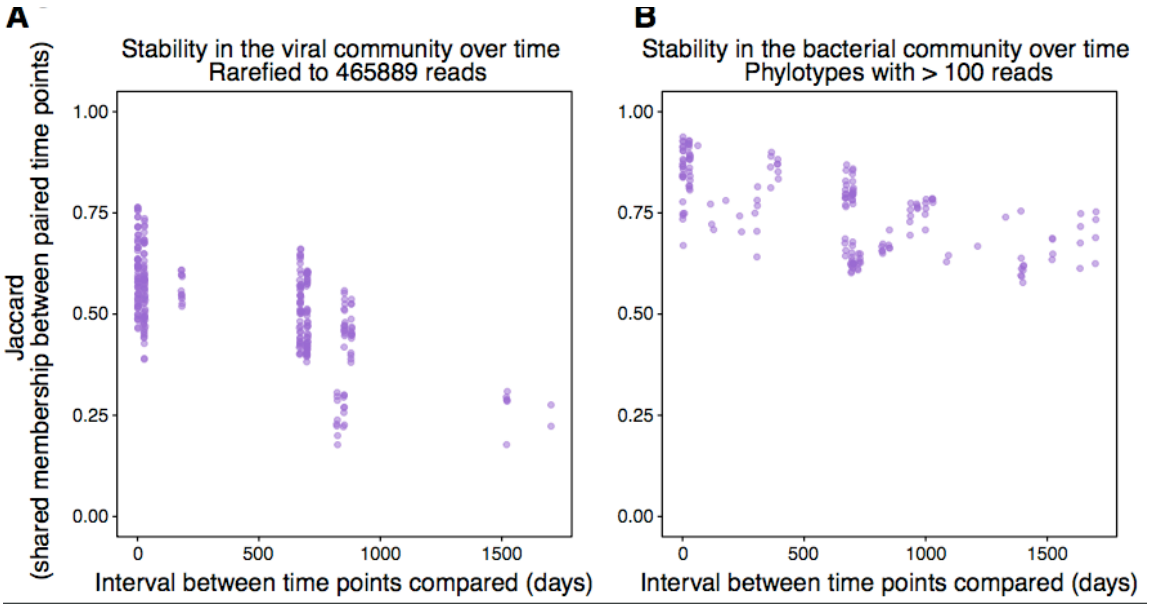
Figure 4. Associating phage host pairs via modification patterns. Multiple modified motifs showed up in both the phage and bacteria data sets, which may be a way to link phage to their bacterial host. Depicted here are phage contigs (containing at least one phage gene marker) that shared a modified motif pattern with a bacterial contig. The lines indicated which phage contigs and bacteria contigs share a modified motif. Red text indicates the predicted modified base within each motif. The motifs shown here have modified patterns that match between the bacteria and phage, and the nucleotide in red indicates where the strong IPD signal occurs and is the predicted modified base. Note that strong IPD signals can occur nucleotides neighboring a modified bases, thus the actual modified nucleotide may occur on an adjacent base.

Figure S1



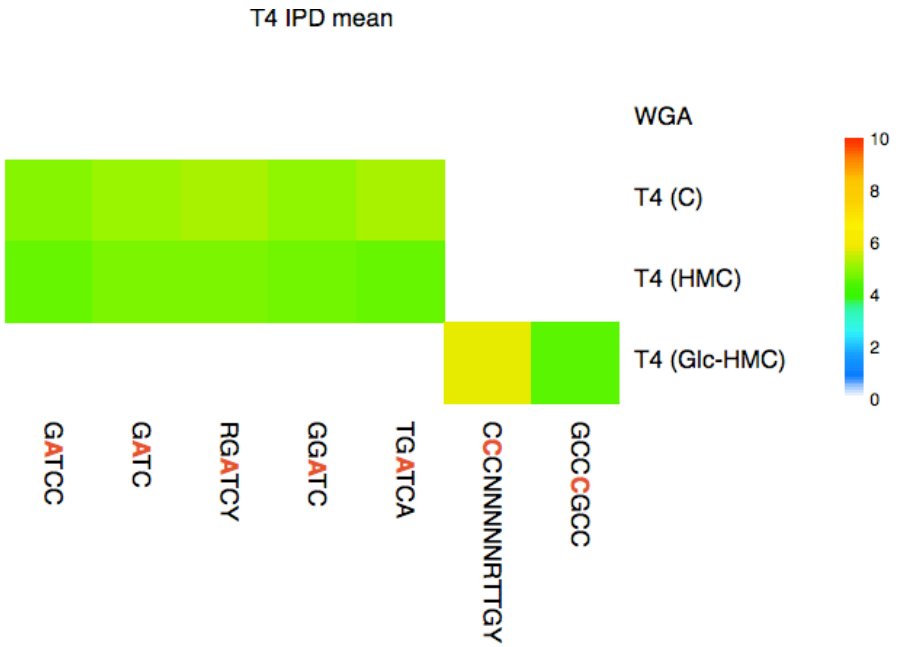
Supplemental Figure 1: Viral gene type composition. The predicted ORFs in the viral contigs that were found any taxonomic annotation by comparison using BLAST e-value threshold of 10^{-5} are shown here grouped by gene type.

Figure S2



Supplemental Figure 2: Shared community membership over time. We scored shared community membership between all pairwise time point comparisons using Jaccard index values. The X-axis indicates how much time (in days) occurred between two sample time points. At the far left of the graph are samples that were collected only one day apart, where as the comparisons on far right depict samples taken 4years apart. A) Viral Jaccard analysis. B) Bacterial Jaccard analysis.

Figure S3



Supplemental Figure 3: Analysis control of custom modified motif finder. T4 Phage T4(glc-HMC), T4(HMC), T4(C), and T4(WGA) have known modifications and serve as a control for our custom modified motif finder. T4(glc-HMC) is methylated at G**A**TC and **C**'s are modified to glucosyl hydroxymethylcytosine. T4(HMC) is methylated at G**A**TC and **C**'s are modified to hydroxymethylcytosine. T4(C) is methylated at G**A**TCs only. T4(WGA) has been whole genome amplified from T4(glc-HMC) DNA to make an unmodified genome. The heatmap depicts the mean IPD value for the modified base in the motif context denoted at the bottom of the figure. The rows indicate individual contigs, and the columns contain motifs that have previously been know to be modified in a living organism. Our motif finding pipeline only finds motifs that are always modified, the G**A**TC does not appear in the T4(glc-HMC) because it is only modified a fraction of the time. The motifs depicted here come from the list of previously studied motifs.

4.7 Tables

Table 1

Table 1. Estimating bacterial and viral counts in the human gut

Method	Median	Geometric Mean	25th Percentile	75th Percentile
Quantitative PCR				
Bacteria 16S*	1.86×10^{13}	1.59×10^{12}	7.03×10^{11}	3.05×10^{13}
Virus	3.74×10^{12}	8.28×10^{12}	9.82×10^{11}	4.55×10^{13}
Staining				
DAPI	2.81×10^{11}	3.01×10^{11}	2.16×10^{11}	3.21×10^{11}
CyberGold dsDNA Virus	2.82×10^{10}	2.77×10^{10}	2.18×10^{10}	4.16×10^{10}

*Assumed 4 copies of 16S per bacterial genome

Table 2

Table 2. Estimating Phage to Host Ratios and Predation Rate

Method	Phage : Host	Predation Rate
Staining	1 : 10	0.2 %
QPCR	1.3 : 1	10 %
Metagenomics	7 : 1	14 %

Table S1

Table S1. Summary of Samples Used in this Study

	0	180	181	182	183	184	489	851	852	853	854	855	879	880	881	882	883	1471	1545	1575	1583	1703	1818	1881
Illumina HiSeq																								
Viruses	✔	✔	✔	✔	✔	✔		✔	✔	✔	✔	✔	✔	✔	✔	✔	✔							
All DNA				✔					✔								✔							
Illumina MiSeq																								
16S Bacteria		✔	✔	✔	✔		✔	✔	✔	✔		✔	✔	✔	✔	✔	✔			✔	✔	✔	✔	✔
Viruses (VLPs)																						✔		
Pac Bio SMRT																								
Viruses		✔	✔	✔	✔	✔	✔	✔	✔	✔		✔	✔	✔	✔	✔	✔							
All DNA				✔					✔								✔			✔				
Staining																								
DAPI																		✔						
Gold Staining																			✔					
Quantitative PCR																								
16S Bacteria		✔	✔	✔	✔		✔	✔	✔	✔		✔	✔	✔	✔	✔	✔			✔	✔	✔	✔	✔
Viral				✔					✔								✔							

Table S2

Table S2. Merged Contigs Summary Statistics

	Number of Contigs	N50 Contigs	Max Contig Length	Avg. Contig Length	Sum of Contig Lengths
Illumina HiSeq (2)	959	29693	167677	11613	11136652
PacBio	1640	1068	81592	1132	1858087
Illumina MiSeq (22)	2095	1478	39567	1019	2135037
Merged	**3358**	**25633**	**217304**	**4094**	**13748809**

Table S3

See supplemental attachments. Modified phage and bacteria contigs. All phage and bacteria contigs with modified motifs (including known motifs and kmers) are listed with the mean and median IPD value for each modified motif.

Table S4

Supplemental Table 4. Oligonucleotides used in this study.

	Forward Primer	Reverse Primer
16S Amplicon		
27F	AGAGTTTGATCCTGGCTCAG	-
R338	-	TGCTGCCTCCCGTAGGAGT
BSR357	-	CTGCTGCCTYCCGTA
Viruses		
HS559	GGGCAAATCGTAACTTTTCCTG	TGCTGGTGCTACTACTTCC
239	GTAGAACAGGTGGTTATGCG	CGGATTACTGAAGACAAAACGG
243	ATAATCCCCTGTAGCGGAG	TGCCTGTGAGTATCCTCTGG
HS0	GGCTGAACCGCATACATAG	AAGATTGGGCTAAGTCACAC
300	CTTGCGTTTTTGGCTGTC	CCCAGAAGATGAGAAGTGC
HS377	ATGCTTTCTGTGGCGTGTTCCG	AGGTTGACAAGGTGTTTGAG
HS587	AAAGGATTGCTCCAGAAGTG	GCGTAAAAACCGATAAGGCTGATG
24	GGTCGTATGTTTCGGTCTG	CGTTTGGTTTTCTATCCTTGCC

Table S5

Supplemental Table 5. (A) Phage burst size measurements from laboratory experiments.

Phage	Genome	Host	Burst Size Average	Burst Size Range	Reference (PMID)
Inoviridae					
M13	DNA	*E. coli*	NA (buds)		16756387
Leviviridae					
MS2	DNA	*E. coli*	400		16756387
R17	DNA	*E. coli*	3570		16756387
Microviridaea					
phiX174	DNA	*E. coli*	180		16756387
Myoviridae					
T4	DNA	*E. coli*	125	100-150	16756387
T2	DNA	*E. coli*	135	wide	10835374
Mu	DNA	*E. coli*	200		16756387
P1	DNA	*E. coli*	160		16756387
P4	DNA	*E. coli*	300		16756387
S3	DNA	*Stenotrophomonas sp.*	100		18952876
phiEF24C	DNA	*Enterococcus faecalis*	115	110-120	18096017
ZZ1	DNA	*Acinetobacter baumanni*	200		22838726
Podoviridae					
T3	DNA	*E. coli*	200		16756387
T7	DNA	*E. coli*	260		16756387
Siphoviridae					
T1	DNA	*E. coli*	180	wide	10835374
Lambda	DNA	*E. coli*	135	115-154	19171945
phi80	DNA	*E. coli*	600		16756387
T5	DNA	*E. coli*	290		16756387
S1	DNA	*Stenotrophomonas sp.*	75		18952876
S4	DNA	*Stenotrophomonas sp.*	75		18952876
Spherical					
KHP30	DNA	*Helicobacter pylori*	13		23475617
Tectiviridae					
PRD1	DNA	*E. coli*	50		16756387

Supplemental Table 5. (B) Phage burst size measurements in the environment.

Site	Environment	Burst Size	Reference (PMID)
Lake Cretail	freshwater lake	44	19878265
Rimov Reseroir	freshwater lake	27	12024261
Eutrophic Lakes	freshwater lake	46	9758799, 11931167
Gulf of Mexico	costal seawater	11	16535459
Gulf of Mexico	costal seawater	54	16535459
Mediterranean Sea	surface seawater	109	16269692
Mediterranean Sea	surface seawater	105	16269692
Mediterranean Sea	surface seawater	41	16269692
Mediterranean Sea	sediment water interface	15	16269692

Table S6

See supplemental attachments. Merged contig summary.

4.8 References

1. **Proctor LM, Okubo A, Fuhrman JA.** 1993. Calibrating estimates of phage-induced mortality in marine bacteria: Ultrastructural studies of marine bacteriophage development from one-step growth experiments. Microb Ecol **25:**161-182.

2. **Bettarel Y, Sime-Ngando T, Amblard C, Dolan J.** 2004. Viral activity in two contrasting lake ecosystems. Appl Environ Microbiol **70:**2941-2951.

3. **Pradeep Ram AS, Sime-Ngando T.** 2010. Resources drive trade-off between viral lifestyles in the plankton: evidence from freshwater microbial microcosms. Environ Microbiol **12:**467-479.

4. **Reyes A, Wu M, McNulty NP, Rohwer FL, Gordon JI.** 2013. Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. Proc Natl Acad Sci U S A **110:**20236-20241.

5. **Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD.** 2013. Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A **110:**12450-12455.

6. **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.

7. **Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD.** 2011. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res **21:**1616-1625.

8. **Minot S, Wu GD, Lewis JD, Bushman FD.** 2012. Conservation of gene cassettes among diverse viruses of the human gut. PLoS One **7:**e42342.

9. **Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G, Fleshner P, Stappenbeck TS, McGovern DP, Keshavarzian A, Mutlu EA, Sauk J, Gevers D, Xavier RJ, Wang D, Parkes M, Virgin HW.** 2015. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell **160:**447-460.

10. **Chehoud C, Dryga A, Hwang Y, Nagy-Szakal D, Hollister EB, Luna RA, Versalovic J, Kellermayer R, Bushman FD.** 2016. Transfer of Viral Communities between Human Individuals during Fecal Microbiota Transplantation. MBio **7**.

11. **Young JC, Chehoud C, Bittinger K, Bailey A, Diamond JM, Cantu E, Haas AR, Abbas A, Frye L, Christie JD, Bushman FD, Collman RG.** 2015. Viral metagenomics reveal blooms of anelloviruses in the respiratory tract of lung transplant recipients. Am J Transplant **15:**200-209.

12. **Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, Bittinger K, Bailey A, Friedman ES, Hoffmann C, Albenberg L, Sinha R, Compher C, Gilroy E, Nessel L, Grant A, Chehoud C, Li H, Wu GD, Bushman FD.** 2015. Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. Cell Host Microbe **18:**489-500.

13. **Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW, Hibberd ML, Liu ET, Rohwer F, Ruan Y.** 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol **4:**e3.

14. **Flores GE, Caporaso JG, Henley JB, Rideout JR, Domogala D, Chase J, Leff JW, Vazquez-Baeza Y, Gonzalez A, Knight R, Dunn RR, Fierer N.** 2014. Temporal variability is a personalized feature of the human microbiome. Genome Biol **15:**531.

15. **Dussoix D, Arber W.** 1962. Host specificity of DNA produced by Escherichia coli. II. Control over acceptance of DNA from infecting phage lambda. J Mol Biol **5:**37-49.

16. **Lederberg S, Meselson M.** 1964. Degradation of Non-Replicating Bacteriophage DNA in Non-Accepting Cells. J Mol Biol **8:**623-628.

17. **Meselson M, Yuan R.** 1968. DNA restriction enzyme from E. coli. Nature **217:**1110-1114.

18. **Kelly TJ, Jr., Smith HO.** 1970. A restriction enzyme from Hemophilus influenzae. II. J Mol Biol **51:**393-409.

19. **Smith HO, Wilcox KW.** 1970. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. J Mol Biol **51:**379-391.

20. **Dong A, Zhou L, Zhang X, Stickel S, Roberts RJ, Cheng X.** 2004. Structure of the Q237W mutant of HhaI DNA methyltransferase: an insight into protein-protein interactions. Biol Chem **385:**373-379.

21. **Xu QS, Kucera RB, Roberts RJ, Guo HC.** 2004. An asymmetric complex of restriction endonuclease MspI on its palindromic DNA recognition site. Structure **12:**1741-1747.

22. **Zheng Y, Roberts RJ, Kasif S.** 2004. Identification of genes with fast-evolving regions in microbial genomes. Nucleic Acids Res **32:**6347-6357.

23. **Yang Z, Horton JR, Maunus R, Wilson GG, Roberts RJ, Cheng X.** 2005. Structure of HinP1I endonuclease reveals a striking similarity to the monomeric restriction enzyme MspI. Nucleic Acids Res **33:**1892-1901.

24. **Roberts RJ.** 2005. How restriction enzymes became the workhorses of molecular biology. Proc Natl Acad Sci U S A **102:**5905-5908.

25. **O'Driscoll J, Heiter DF, Wilson GG, Fitzgerald GF, Roberts R, van Sinderen D.** 2006. A genetic dissection of the LlaJI restriction cassette reveals insights on a novel bacteriophage resistance system. BMC Microbiol **6:**40.

26. **Horton JR, Zhang X, Maunus R, Yang Z, Wilson GG, Roberts RJ, Cheng X.** 2006. DNA nicking by HinP1I endonuclease: bending, base flipping and minor groove expansion. Nucleic Acids Res **34:**939-948.

27. **Rasko T, Der A, Klement E, Slaska-Kiss K, Posfai E, Medzihradszky KF, Marshak DR, Roberts RJ, Kiss A.** 2010. BspRI restriction endonuclease: cloning, expression in Escherichia coli and sequential cleavage mechanism. Nucleic Acids Res **38:**7155-7166.

28. **Zheng Y, Cohen-Karni D, Xu D, Chin HG, Wilson G, Pradhan S, Roberts RJ.** 2010. A unique family of Mrr-like modification-dependent restriction endonucleases. Nucleic Acids Res **38:**5527-5534.

29. **Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P.** 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science **315:**1709-1712.

30. **Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV.** 2015. An updated

evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol **13:**722-736.

31. **Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C.** 2009. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology **155:**733-740.

32. **Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, Marraffini LA.** 2015. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature **519:**199-202.

33. **Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadan AH, Moineau S.** 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature **468:**67-71.

34. **Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, van der Oost J, Brouns SJ, Severinov K.** 2011. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U S A **108:**10098-10103.

35. **Yosef I, Goren MG, Qimron U.** 2012. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res **40:**5569-5576.

36. **Horton JR, Wang H, Mabuchi MY, Zhang X, Roberts RJ, Zheng Y, Wilson GG, Cheng X.** 2014. Modification-dependent restriction endonuclease, MspJI, flips 5-methylcytosine out of the DNA helix. Nucleic Acids Res **42:**12092-12101.

37. **Bair CL, Black LW.** 2007. A type IV modification dependent restriction nuclease that targets glucosylated hydroxymethyl cytosine modified DNAs. J Mol Biol **366:**768-778.

38. **Warren RA.** 1980. Modified bases in bacteriophage DNAs. Annu Rev Microbiol **34:**137-158.

39. **Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A, Fries R, Froula J, Kang DD, Malmstrom RR, Morgan RD, Posfai J, Singh K, Visel A, Wetmore K, Zhao Z, Rubin EM, Korlach J, Pennacchio LA, Roberts RJ.** 2016. The Epigenomic Landscape of Prokaryotes. PLoS Genet **12:**e1005854.

40. **Leonard MT, Davis-Richardson AG, Ardissone AN, Kemppainen KM, Drew JC, Ilonen J, Knip M, Simell O, Toppari J, Veijola R, Hyoty H, Triplett EW.** 2014. The methylome of the gut microbiome: disparate Dam methylation patterns in intestinal Bacteroides dorei. Front Microbiol **5:**361.

41. **Roberts RJ, Macelis D.** 1996. REBASE--restriction enzymes and methylases. Nucleic Acids Res **24:**223-235.

42. **Roberts RJ, Vincze T, Posfai J, Macelis D.** 2003. REBASE: restriction enzymes and methyltransferases. Nucleic Acids Res **31:**418-420.

43. **Roberts RJ, Vincze T, Posfai J, Macelis D.** 2007. REBASE--enzymes and genes for DNA restriction and modification. Nucleic Acids Res **35:**D269-270.

44. **Roberts RJ, Vincze T, Posfai J, Macelis D.** 2015. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res **43:**D298-299.

45. **Bryson AL, Hwang Y, Sherrill-Mix S, Wu GD, Lewis JD, Black L, Clark TA, Bushman FD.** 2015. Covalent Modification of Bacteriophage T4 DNA Inhibits CRISPR-Cas9. MBio **6:**e00648.

46. **Hattman S.** 1970. DNA methylation of T-even bacteriophages and of their nonglucosylated mutants: its role in P1-directed restriction. Virology **42:**359-367.

47. **van Eijk E, Anvar SY, Browne HP, Leung WY, Frank J, Schmitz AM, Roberts AP, Smits WK.** 2015. Complete genome sequence of the Clostridium difficile laboratory strain 630Deltaerm reveals differences from strain 630, including translocation of the mobile element CTn5. BMC Genomics **16:**31.

48. **Lotka AJ.** 1920. Analytical Note on Certain Rhythmic Relations in Organic Systems. Proc Natl Acad Sci U S A **6:**410-415.

49. **Volterra V.** 1928. Variations and Fluctuations of the Number of Individuals in Animal Species living together. Journal du Conseil **3:**3-51.

50. **Cortez MH, Weitz JS.** 2014. Coevolution can reverse predator-prey cycles. Proc Natl Acad Sci U S A **111:**7486-7491.

51. **Stoddard SF, Smith BJ, Hein R, Roller BR, Schmidt TM.** 2015. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and

archaea and a new foundation for future development. Nucleic Acids Res **43:**D593-598.

52. **Kim MS, Park EJ, Roh SW, Bae JW.** 2011. Diversity and abundance of single-stranded DNA viruses in human feces. Appl Environ Microbiol **77:**8062-8070.

53. **Hatfull GF.** 2008. Bacteriophage genomics. Curr Opin Microbiol **11:**447-453.

54. **Bongiorni L, Magagnini M, Armeni M, Noble R, Danovaro R.** 2005. Viral production, decay rates, and life strategies along a trophic gradient in the North Adriatic Sea. Appl Environ Microbiol **71:**6644-6650.

55. **Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD.** 2011. Linking long-term dietary patterns with gut microbial enterotypes. Science **334:**105-108.

56. **Dollive S, Peterfreund GL, Sherrill-Mix S, Bittinger K, Sinha R, Hoffmann C, Nabel CS, Hill DA, Artis D, Bachman MA, Custers-Allen R, Grunberg S, Wu GD, Lewis JD, Bushman FD.** 2012. A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. Genome Biol **13:**R60.

57. **Williamson KE, Radosevich M, Wommack KE.** 2005. Abundance and diversity of viruses in six Delaware soils. Appl Environ Microbiol **71:**3119-3125.

58. **Tarnowski BI, Spinale FG, Nicholson JH.** 1991. DAPI as a useful stain for nuclear quantitation. Biotech Histochem **66:**297-302.

Chapter 5: Conclusion and Future Directions

5.1 Conclusion and Future Directions

This thesis sheds light on how communities of phage and bacteria interact through the CRISPR system and covalent DNA modifications. In chapter 2 we find high nucleotide substitution rates in human gut phage over time, which may be in part due to CRISPR pressure(1). In chapter 3 we explore the role of covalent DNA modifications in protecting phage from the CRISPR-Cas9 system. Specifically, we find that glc-HMC and HMC allow T4 phage to escape the CRISPR-Cas9 system. In Chapter 4 we estimate phage predation rates on bacteria of the human gut and link potential phage-host pairs though DNA modification patterns. This is the first study to look at global modification patters of phage and bacteria in the human gut microbiome.

In a 2.5 year longitudinal study of the human gut virome, we observed nucleotide substitution rates up to 4% in the strictly lytic ssDNA Microviridae phage. Phage with dsDNA genomes showed modest rates of nucleotide variation over time, which is consistent with temperate phage whose genomes are replicated by an accurate bacterial polymerases. One possible driver for nucleotide changes in phage is to escape pressure from CRISPR systems(2). Seven bacterial contigs with CRISPR arrays targeting phage contigs were identified. All but one targeted phage contained between one and four

protospacers, and the remaining phage contained 27 protospacers. In one of the targeted phage, a point mutation in the protospacer arose and went to fixation indicating a fitness advantage for this genotype. We also found a phage contig harboring a CRISPR array that targeted another phage. There has been one previous report of a phage encoded CRISPR(3). Our results suggest that both phage and bacteria use CRISPR systems to compete with phage.

Phage have been shown to use covalent DNA modifications to protect themselves from restriction endonucleases(4-6), however the ability of DNA modifications to block CRISPR nuclease activity was previously unknown. One study had shown that a single adenine-N6-methyl group was not sufficient to block CRISPR activity(7), however phage contain multiple unique and unusual DNA modifications (many of which are significantly larger than methyl groups)(8). We used T4 phage, which replaces all of its cytosine bases with glc-HMC and two mutant T4 phage, which contain either HMC or unmodified C's to further address this question. Our results show that glc-HMC and HMC in high concentrations are sufficient to block CRISPR-Cas9 activity(9).

In future studies I would like to test additional combinations of CRISPR systems and DNA modifications. At least 16 CRISPR system subtypes and 10 unique phage DNA modifications have been discovered(8, 10). It is likely that the diverse CRISPR systems will behave differently when confronted with various

DNA modifications. It is also important to establish a mechanism for how DNA modifications block CRISPR activity. The *in vivo* assays we used could not distinguish if CRISPR activity is blocked at the DNA binding step or the DNA cleavage step. In vitro assays with type I and type II CRISPR systems have been recently established and can be used to address this question. I would also like to systematically test which nucleotide positions are most important in blocking CRISPR activity. Crystal structures and nuclease protection assays for types I -III CRISPR systems indicate that the Cas proteins do not come in direct contact with every nucleotide in the protospacer(11-16). Even when modified, I expect some nucleotides within the protospacer will not contact the Cas proteins and will not provide protection. Our study uses the naturally occurring T4 protospacers, so it was not possible to test all permutations of modified nucleotide positions. However, we can use synthetic oligos and *in vitro* enzymatic assays to address this hypothesis.

The acquisition of spacers from modified DNA is another exciting area of research that has yet to be addressed. I hypothesize that some DNA modifications can block spacer acquisition which can be tested using T4 phage and a genetically modified *E. coli* strain. *E. coli* K12 contains a type I CRISPR system that is repressed by H-NS (heat-stable nucleoid-structure: a global transcriptional repressor) and LeuO(a transcription factor)(17). Knocking out H-

NS and LeuO result in a fully functional CRISPR system. I would culture the H-NS and LeuO *E. coli* knockout with T4(glc-HMC), T4(HMC), and T4(C) separately and sequence the newly acquired spacers. *E. coli* K12 has two CRISPR arrays, so I would PCR amplify both arrays and deep sequence the amplicons. I anticipate spacers to be acquired without bias from the T4(C) infection, and the T4(HMC) and T4(gls-HMC) infections would prevent, reduce or bias spacer acquisition to regions of the T4 genome most devoid of modified bases.

We used staining, qPCR, an metagenomics to study phage predation dynamics in the human gut microbiome and estimate that between 0.2 and 14% of the gut bacteria community are killed by phage per day. In this study it became clear that each assay has advantages and disadvantages leading to an overall estimate with wide confidence intervals. This suggests that current methods to evaluate phage and bacteria populations have significant limitations. Using multiple techniques to query microbial communities should be used to help reduce bias.

Analysis of the gut microbiome DNA modifications indicated that 73% of phage genomes and 56% of bacterial genomes contain motifs that are completely modified. A total of 443 modified motifs were shared between phage and bacteria, suggesting potential phage-host pairs, and five phage-host pairs

could be called using conservative criteria. The computer pipeline we developed is the first to query modification patters among metagenomic phage and bacterial communities. The pipeline currently only identifies motifs that are modified at each motif occurrence. For future directions of this project, I would like to expand the pipeline to identify motifs that are only modified a fraction of the time. It is known that some phage modify motifs only 20-50% of the time(8). Our pipeline identified modifications on all four bases (G, A, T, and C); so, I would also like to use high-performance liquid chromatography coupled with mass spectrometry to confirm the presence of modifications on each of the bases.

The work in this thesis indicates that both bacteria and phage within the human gut microbiome use CRISPR systems to target phage. One way phage may protect themselves from the CRISPR system is through covalent DNA modifications such as HMC and glc-HMC, so we conducted the first global covalent modifications analysis of bacteria and phage within the human gut microbiome. We found that the majority of bacteria and phage contain modified motifs. The fact that many modified motifs were shared between bacteria and phage highlights how important DNA modifications are to the survival of these intertwined communities. This work provides insight into how phage and bacteria interact with each other within the human gut microbiome, a crucial step in the development of therapies for dysbiotic microbiomes.

5.2 References

1. **Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD.** 2013. Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A **110:**12450-12455.

2. **Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, van der Oost J, Brouns SJ, Severinov K.** 2011. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U S A **108:**10098-10103.

3. **Seed KD, Lazinski DW, Calderwood SB, Camilli A.** 2013. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature **494:**489-491.

4. **Kelly TJ, Jr., Smith HO.** 1970. A restriction enzyme from Hemophilus influenzae. II. J Mol Biol **51:**393-409.

5. **Meselson M, Yuan R.** 1968. DNA restriction enzyme from E. coli. Nature **217:**1110-1114.

6. **Smith HO, Wilcox KW.** 1970. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. J Mol Biol **51:**379-391.

7. **Dupuis ME, Villion M, Magadan AH, Moineau S.** 2013. CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. Nat Commun **4:**2087.

8. **Warren RA.** 1980. Modified bases in bacteriophage DNAs. Annu Rev Microbiol **34:**137-158.

9. **Bryson AL, Hwang Y, Sherrill-Mix S, Wu GD, Lewis JD, Black L, Clark TA, Bushman FD.** 2015. Covalent Modification of Bacteriophage T4 DNA Inhibits CRISPR-Cas9. MBio **6:**e00648.

10. **Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV.** 2015. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol **13:**722-736.

11. **Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, Wiedenheft B, Pul U, Wurm R, Wagner R, Beijer MR, Barendregt A, Zhou K, Snijders AP, Dickman MJ, Doudna JA, Boekema EJ, Heck AJ, van der Oost J, Brouns SJ.** 2011. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat Struct Mol Biol **18:**529-536.

12. **Wiedenheft B, van Duijn E, Bultema JB, Waghmare SP, Zhou K, Barendregt A, Westphal W, Heck AJ, Boekema EJ, Dickman MJ, Doudna JA.** 2011. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc Natl Acad Sci U S A **108:**10092-10097.

13. **Wiedenheft B, Lander GC, Zhou K, Jore MM, Brouns SJ, van der Oost J, Doudna JA, Nogales E.** 2011. Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature **477:**486-489.

14. **Jiang F, Doudna JA.** 2015. The structural biology of CRISPR-Cas systems. Curr Opin Struct Biol **30:**100-111.

15. **Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, Ma E, Anders C, Hauer M, Zhou K, Lin S, Kaplan M, Iavarone AT, Charpentier E, Nogales E, Doudna JA.** 2014. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science **343:**1247997.

16. **Staals RH, Agari Y, Maki-Yonekura S, Zhu Y, Taylor DW, van Duijn E, Barendregt A, Vlot M, Koehorst JJ, Sakamoto K, Masuda A, Dohmae N, Schaap PJ, Doudna JA, Heck AJ, Yonekura K, van der Oost J, Shinkai A.** 2013. Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus. Mol Cell **52:**135-145.

17. **Westra ER, Pul U, Heidrich N, Jore MM, Lundgren M, Stratmann T, Wurm R, Raine A, Mescher M, Van Heereveld L, Mastop M, Wagner EG, Schnetz K, Van Der Oost J, Wagner R, Brouns SJ.** 2010. H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. Mol Microbiol **77:**1380-1393.