Spring 4-26-2013

Metagenomic Analyses of the Human Gut Virome

Alejandro Reyes
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology & Biomedical Sciences

Computational and Systems Biology

Dissertation Examination Committee:
Jeffrey I. Gordon, Chair
Gautam Dantas
Justin C. Fay
Rob Mitra
Herbert W. Virgin
David Wang

Metagenomic Analyses of the Human Gut Virome

by

Alejandro Reyes

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2013

St. Louis, Missouri

Table of Contents

Chapter 3

A gnotobiotic mouse model for characterizing phage-bacterial host interactions in the human gut

Chapter 4

Preliminary analysis of the gut viromes of healthy and malnourished twins and their family members, and future directions

Appendix

List of Figures

Chapter 1

Going viral: next generation sequencing applied to human gut phage populations.

Chapter 2

Viruses in the fecal microbiota of monozygotic twins and their mothers

Chapter 3

A gnotobiotic mouse model for characterizing phage-bacterial host interactions in the human gut

Chapter 4

Preliminary analysis of the gut viromes of healthy and malnourished twins and their family members, and future directions

List of Tables

Chapter 2

Viruses in the fecal microbiota of monozygotic twins and their mothers

Chapter 3

A gnotobiotic mouse model for characterizing phage-bacterial host interactions in the human gut

Chapter 4

Preliminary analysis of the gut viromes of healthy and malnourished twins and their family members, and future directions

Acknowledgements

I will always be thankful to **Jeff** for all the help and patience he has offered during all these years. Since the start of my project we knew that I was adventuring into the unknown, we were taking a big risk that had the potential for big rewards. Jeff always had the right words to motivate and inspire me. His great vision was able to guide the project and achieve great success. Jeff teaches by example, and he is the greatest example of hard work that I have ever encountered. I will always remember his dedication, enthusiasm, patience and love for Science and as I start my career I will try to always follow his example.

I think that part of Jeff's talent as a mentor and as a scientist is his ability to recruit a very talented group of technicians and research assistants that have been, and will be, essential for the success of the lab. They assure that everything runs smoothly despite the constant attempts of trainees, like myself, to prevent that from happening. None of my research would have been possible without their help: **Jill** and **Su** were instrumental in getting all the pyrosequencing running in the lab; **Dave** and **Maria** were not only fundamental with all the gnotobiotic mouse work, but were the best company all of the many mornings I spent sampling at the mouse house. **Marty** for his help and assistance with the samples processing, I will still be extracting DNAs without him. Special acknowledgements have to be given to all the GTAC team, in particular **Jess**, who shared our excitement from the first time we reached 6 million reads out of the Illumina GAIIx, to my worried face when I wondered: 'how am I supposed to analyze the data when you are giving me 150 million reads from a single lane of Illumina HiSeq?'. The consequences of my slow learning curve programing and parallelizing in the cluster were endured by **Eric Martin** and **Brian Koebbe**, who received countless visits in their office complaining about how someone (many times me) was slowing the cluster down. From the administrative side, the patience and support of **Stephanie Amen**, as well as the dedication in graphics support of **Laura Kyro** are instrumental for the correct functioning of many aspects of the lab. Lastly, but certainly not least **Sabrina Wagoner**; it is easy to pinpoint the technical aspects in which she has contributed during my stay in the Gordon lab (the list is very long). But I cannot put into words the immense help and support she gave me

throughout these years; she became essentially my ‘lab mom’, and I have no way to thank her for all the time and patience, as well as good times (which at the end were not few) that we were able to share.

As for student and postdocs in the lab, there is a long list to be acknowledged since everyone has contributed at one time or in one way to this work. In particular I have to mention the other ‘South Mexicans’: **Vane** and **Federico**, who helped me make Spanish the second official language in the Gordon lab. The three of us arrived to the lab within the same year and as our tenure finishes, the number of shared experiences is countless, moments of joy and moments of tears, always willing to give a hug, good advice or just share a cup of coffee when needed. I really don’t think I could have made it through without their help, company and support. Along with them, the lab has a great contingent of international trainees that made every week a joyful time in the lab: the Irish guy (**Philip**), the Canadian (**Michelle**) and the Chinese (**Meng**), are not only great scientist and supportive friends but people whose company I truly enjoy, particularly our countless ‘scientific’ discussions over friday dinners. I want to also mention two great students **Mark Charbonneau** and **Nick Semenkovich**, my ‘bay mates’, who have been a great company for the last few years. I want to acknowledge current and past Gordon lab members that helped and supported me and made fundamental contributions to this work: **Pete Turnbaugh** who welcomed me to the lab and was my mentor during my rotation; **Liz Hansen**, **Nate McNulty, Brian Muegge** and **Tanya Yatsunenko** with whom I shared great moments and always were good sources of scientific advice; in particular, a large part of this work would not have been possible without Nate’s work on model communities. I couldn’t have done any of my work without the constant input from experienced postdocs; my computational mentor: **J. Faith**, and the great advice and friendship from **Andy Goodman**, **Andy Kau**, **Henning Seedorf** and **Gabe Simon**.

This work would not have been possible without the constant help, support and inspiration from **Forest Rohwer** and his entire group at SDSU. His point of view changed our understanding of phages and made them look cool. Along with the Rohwer group, I will always remember the fascinating scientific conversations with **Rob Edwards** and **Anca Segall**. Computationally, I

had the great opportunity to interact and exchange ideas with **Rob Knight**'s group at Colorado. Within Washington University, I had the privilege to collaborate with very distinguished professors including all the member of my **thesis committee**, whom I would like to thank; besides their insightful suggestions during committee meetings, I had the honor to work on different projects with every single one of them.

Finally, I have to thank all my **friends**, inside and outside of the science world that have kept me going for all these years. It is thanks to their friendship and support that I was able to wake up every morning, and even knowing that I was thousands of miles away from my home and family, I could still get motivated to do the best job possible. Finally, my **family**: my mom, my dad and my brother, who never stopped believing and supporting my career from Colombia, they have witnessed first hand many exciting, as well as frustrating moments and have never failed to give me a supporting hand.

Dedication

To my family for all their love and support during all these years.

ABSTRACT OF THE DISSERTATION

Metagenomic Analyses of the Human Gut Virome

by

Alejandro Reyes

Doctor of Philosophy in Biology and Biomedical Sciences

(Computational and Systems Biology)

Washington University in St. Louis, 2013

Professor Jeffrey I. Gordon, Chair

The human gut harbors tens of trillions of microbes belonging to all three domains of life, Bacteria, Archaea, and Eukarya; most are members of Bacteria. These organisms collaborate and compete for functional niches and physical space (habitats), together forming a continuously functioning metabolic organ that influences many aspects of host biology. The factors that drive the assembly, determine the stability, and shape the adaptive responses of the gut microbiota to a variety of perturbations are the subject of intense study as greater appreciation is gained of the importance of this microbial community for human health. My thesis focused on the viral component of the microbiota that had been less characterized than its bacterial component. I first developed and applied a series of experimental and computational tools for metagenomic analyses of viruses purified from frozen fecal samples obtained from healthy adult monozygotic twin pairs and their mothers living in the USA, over the course of a year. The virome in this population was dominated by phages and exhibited high inter-personal variation and contrasting intrapersonal stability, suggesting a prevalent temperate lifestyle rather than a predator-prey relationship that is a feature of marine microbial communities. To further characterize the role of phage in shaping gut community structure, I colonized adult germ-free mice with a defined model human gut microbiota composed of 15 sequenced human gut symbionts, seven of which harbored 10 prophages, one of which (*Bacteroides cellulosilyticus* WH2) was represented by a library of >25,000 isogenic trans-

poson mutants covering 80% of genes in its genome. Once assembled, this model microbiota was subjected to a staged phage attack with a pool of virus-like particles (VLPs) purified from the fecal microbiota of five humans from the first study. Shotgun sequencing of DNA isolated from the input human VLP preparation, cecal and fecal samples collected over time from these gnotobiotic mice, and VLPs recovered from their fecal samples, revealed a ordered and reproducible sequence of phage attack, allowing me to associate novel phages present in the input VLP preparation with bacterial hosts, and to characterize the dynamics and identify genetic determinants of prophage induction. Finally, I used the tools I developed to characterize the phages and eukaryotic viruses present in the fecal microbiota of healthy and malnourished twins living in Malawi, sampled during their first two years of life. Together, this work provided new perspectives about viral diversity and viral-bacterial host dynamics associated with the human gut microbiota.

Chapter 1

Going viral: next generation sequencing applied to human gut phage populations.

Chapter 1

Going viral: next generation sequencing applied to human gut phage populations.

Alejandro Reyes[1], Nicholas P. Semenkovich[1], Katrine Whiteson[2], Forest Rohwer[2], and Jeffrey I. Gordon[1]

[1]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis MO 63108, and [2]Department of Biology, San Diego State University, San Diego, CA 92182

This chapter corresponds to the full accepted version of a manuscript published in

Nature Reviews Microbiology 10, 607-617 (September 2012) | doi:10.1038/nrmicro2853

Abstract

Over the past decade researchers have begun to characterize viral diversity using metagenomic methods. These studies have shown that viruses, the majority of which infect bacteria (bacteriophages), are likely the most genetically diverse components of the biosphere. Here we briefly review the incipient rise of a phage biology renaissance catalyzed by recent advances in next generation sequencing. We explore how work characterizing phage diversity and their lifestyles in the gut is changing our view of ourselves as supra-organisms. Finally, we discuss how a new appreciation of phage dynamics may yield new applications for phage therapies designed to manipulate the structure and functions of our gut microbiomes.

Introduction

From Alfred Hershey and Martha Chase's studies indicating that DNA was the genetic material [1], to Francis Crick and Sydney Brenner's experiment establishing the triplet nature of the genetic code [2], bacteriophages ("phages") have helped define fundamental components of modern biology. Most of the tools for early molecular biology arose from the work of phage biologists [3]. The first genomes sequenced were from phages and other viruses. The first comparisons of multiple genomes were carried out on *Lactobacillus* and *Mycobacteria* phages. These early studies showed that there was extensive diversity in essentially every phage community. Now it is clear that viruses are the most diverse and uncharacterized components of the major ecosystems on Earth [4], and that viruses have intricate roles in ecosystem function, far beyond simple predator-prey dynamics [5] (**Box 1**).

At the same time, the clinical world has become increasingly interested in phage-based therapeutics because of the increased prevalence of antibiotic resistant bacteria [6]. The idea of using phages as therapeutic tools is not new. Félix d'Herelle, co-discoverer of phage, recognized the potential medical applications nearly a century ago [7] and his first phage therapies were tested as early as 1919 [8]. However, a rudimentary understanding of the composition and dynamic operations

of the human microbiome, a lack of knowledge of phage biology, and poor quality control during production of phages made this therapeutic approach unreliable [9].

This Review details recent advances on the rising field of viral metagenomics with an emphasis on our current views of phage communities associated with the human gut. It does not discuss many of the resistance mechanisms such as Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs), or the extremely large field of eukaryotic viral discovery that is being propelled by metagenomics, since these topics have been reviewed elsewhere [10-14]. Instead, we focus on viruses that infect bacteria. **Box 2** provides examples of some challenges facing phage metagenomics at the present time, some of which are discussed below.

Methods for viral metagenomics

The introduction of small subunit (SSU) ribosomal RNA (e.g., 16S rRNA) as a reliable prokaryotic phylogenetic marker [15-17] opened the door to remarkable insights about microbial community diversity and dynamics. Phylogenetic marker 'envy' rapidly 'infected' the psyche of phage biologists: they did not have an SSU rRNA equivalent and there was, and is, no conserved protein or gene enabling a similar characterization of all or the majority of phages present in a sample. Efforts to characterize phage diversity focused on characterizing partially conserved fragments of phage genes such as polymerases. However, this method was useful only within certain viral families [18-20]. Horizontal gene transfer further complicates the use of marker genes in phages. For example, most Caudovirales have identifiable conserved functional genes (terminase, portal proteins, capsid, etc), but horizontal transfer and recombination events generate extensive genome mosaicism that challenges phage phylogeny [21]. The arrival of next generation sequencing (NGS) together with methods for purifying virus-like particles (VLPs) set the stage for defining viral diversity based on shotgun sequencing.

Purification of VLPs - Although viral particles outnumber microbial cells 10:1 in most environments, viral DNA represents 2–5% of total community DNA [22-24]. For this reason, it is often desirable to separate viruses from microbial cells. If sample volume is large and viral density low

(such as in ocean environments), tangential flow filtration can be used to remove large particles and concentrate VLPs. For solid samples with high viral density e.g., feces, a common approach is to resuspend the material in an osmotically neutral buffer followed by one or more steps designed to remove large particles (e.g., cesium chloride density gradient ultracentrifugation [25] and subsequent filtration) (**Fig. 1**). This procedure has been successfully applied to fecal material that had been stored at -80°C for several years, without any pre-processing of the sample, indicating that VLP structures are quite stable to freezing and thawing [22].

Amplification of VLP-derived DNA - After VLPs are purified, non-encapsulated free nucleic acids are removed by treatment with DNase and RNase, and VLP nucleic acids are then isolated. The methods chosen will determine purity, influence DNA and RNA yields, and represent a selection step that can bias interpretation of virotype abundance and viral community diversity [26]. Unfortunately, the yield of DNA following extraction of nucleic acids from purified VLPs is often below the required minimum for sequencing. Therefore, a variety of amplification methods have been developed such as random amplified shotgun library RASL [27], linker-amplified shotgun library LASL [28] and others [29-32]. Caveats concerning random PCR amplification of viral DNA include inherent bias due to exponential amplification of mixed templates, uneven coverage of viral genomes, and its limitation to dsDNA templates. Another common method uses the phage-derived phi29 polymerase for multiple displacement amplification (MDA) [33]. MDA takes advantage of the high processivity of this DNA polymerase (>70,000 nucleotides) and its strong strand-displacement capability, which permits amplification of complete viral genomes. The result is a very fast method that can efficiently amplify minute amounts of both ssDNA and dsDNA. While fast, the method is not without flaws, including over-amplification of small circular ssDNA viruses [34] and potential chimera formation [35, 36]. Procedures for avoiding some of these limitations continue to be developed [37, 38], including a novel transposon-based method for rapidly generating DNA libraries from small quantities of dsDNA [39]. RNA viruses can be sequenced by reverse transcription followed by application of the protocols described above. Alternatively, Whole Transcriptome Amplification (WTA) approaches can also be used [40].

Sequencing strategies – Although sequencing costs are falling at an astonishingly rapid rate as newer technologies offer higher degrees of parallelism (greater numbers of reads per run; multiplex sequencing using sample-specific DNA barcodes), read length matters [41]. When characterizing a viral community where most of the sequences obtained are novel and enriched in regions of low complexity repeats, accurate assembly and taxonomic assignment benefit from the longest reads possible [41, 42]. The earliest NGS analyses were powered by the 454 GS20 instrument with ~100 nt reads. Advances in pyrosequencing technology, including today's FLX+ platform, have produced average read lengths that exceed 800 nt [22, 23, 43-46]. Most recently, analysis of metagenomic datasets generated by deep sequencing of total microbial community DNA with highly parallel Illumina instruments showed that the percentage of reads with similarity to known viral sequences was generally less than 0.01% [30, 47, 48]. This low value is in part due to the short read length (≤100 nt). However, the percentage increases when VLPs are purified [32]. Other studies have obtained better assignments by pre-assembling the short reads [49, 50].

In summary, viral metagenomics has opted for technologies that prioritize long read length over platforms with short read length and higher throughput. However, as the latter platforms approach 250nt read lengths and 250 million reads per lane (Illumina HiSeq 2500), and the cost per read falls, we will undoubtedly see a very rapid transition to these types of sequencers — as long as improvements in assembly algorithms keep pace (**Fig. 1**).

Computational approaches for characterizing sequenced viromes - To address the question of viral community composition, shotgun metagenomic sequences are typically compared to individual viral genomes. Although public sequence databases have expanded considerably — from 500 viral genomes as of 2007 to over 3000 full viral genomes to date — the number of deposited genomes is far less than the expected number of virotypes present in 100 liters of sea water [51]. Compounding the problem, existing databases include very few viral proteins in their training sets, meaning many taxonomic assignments are based on proteins transferred between a virus and a microbial host or that are present in prophages and described as part of a microbial genome. Databases with a particular focus on viruses are under development and include a CLAssification of

Mobile genetic Elements (ACLAME) [52] and the Phage SEED [53]. Novel data analysis pipelines are also being constructed to improve the accuracy and efficiency of homology searches [54-61].

Once taxonomic and functional assignments have been made for a given sample, a viral community profile can be created that characterizes the diversity present in that sample. Multi-dimensional reduction methods such as Principal Coordinate Analysis (PCA) and hierarchical clustering have been used to visualize similarities among viral communities, and methods such as supervised learning can help to identify discriminatory features.

Given that most of the viral metagenomic data lack similarity to entries in databases, similarity-independent methods have been developed to better understand viral community structure. One example, PHACCS (Phage Communities from Contig Spectrum), was designed to quantify virotypes [62, 63] based on the assumption that if a virotype is present in high abundance in a VLP sample, it is more likely to be assembled into a large contig. Moreover, we can posit that if assembly of a single sample dataset allows prediction of community structure and diversity, pooling two samples together and performing a cross-assembly analysis can determine inter-sample diversity (e.g., using MaxiPhi [64]). Another alternative for identifying shared viruses among different samples comes from crAss [65], an algorithm that allows for the simultaneous cross-assembly of all the samples in a dataset as opposed to the pairwise assemblies used in MaxiPhi. As more tools are developed, special attention should be given to the assembly parameters to prevent mixed assemblies and chimeras between viral genomes (**Fig. 1**).

Phages in the human gut

The gut provides an enticing place to characterize the role of phages in community assembly and dynamics. Assembly of the human gut microbiota begins at birth with evolution toward an adult-like configuration during the first three years of life [66]. The importance of early environmental exposures is emphasized by the fact that the overall phylogenetic composition of the gut microbiota of adult monozygotic twins is not significantly greater than that of dizygotic twins, and family members share a higher degree of similarity than unrelated individuals living in different

households. These patterns are robust to different cultural traditions, and the observations about mono- versus dizygotic twins apply to infants and children as well as teenagers and adults [66, 67]. Microbes in this densely populated ecosystem are engaged in a constant fight for nutrients and survival. Peristalsis moves an ephemeral menu of dietary components along the cephalocaudal axis of the intestine and microbial members face the omnipresent threat of washout from the gut 'bioreactor'. Maintaining a foothold in this ecosystem depends not only on physical interactions with the perpetually renewing mucus layer and partially digested food particles, but also on functional interactions with other community members. Preserving functional redundancy contributes to community resilience, with horizontal gene transfer providing an opportunity to constantly refashion the genomes (and by extension the pan-genomes) of species-level phylotypes. Each adult appears to harbor a persistent collection of one or at most a few hundred species in their intestines, although strain level diversity is great [24, 68, 69]. While the proportional representation of taxa changes as the community responds to various environmental perturbations, intrapersonal variation in species content is considerably less than interpersonal differences [66, 67, 70, 71].

As our knowledge of inter- and intrapersonal variations in the microbiota expands, a lagging question has been the role of phages in shaping community properties. Although a number of individual phages have been extensively characterized (providing an important genomic context against which metagenomic data can be interpreted), recently much more attention has been given to phage dynamics at a microbial community level. In 2003, the first report of a human-associated gut virome was published; it described the results of shotgun (Sanger) sequencing of VLPs isolated from a fecal sample obtained from a single healthy adult. The identifiable fraction of the virome was dominated by phages, including temperate phages ('prophages' are defined here as temperate phages in their host-incorporated state). This report estimated that there were 1200 different virotypes in the single sample analyzed, with the majority assigned to the Siphoviridae family [28]. Siphoviridae and temperate phages have subsequently been reported to be the most abundant identifiable viruses in other sampled fecal viromes, followed by members of the Podoviridae [22, 23, 72].

The prominence of Microviridae in adult human gut microbiota was initially dismissed as an artifact of the MDA method, which has a preference for ssDNA. However, a novel branch of the Microviridae has been identified recently from prophages present in the genomes from members of two genera in the Bacteroidetes: *Bacteroides* and *Prevotella* [73]. These two genera are prominently represented in the microbiota of adult human populations living in a number of diverse geographic areas [66, 74]. Another study characterizing Microviridae from healthy human donors also clustered these novel viruses with Bacteroides and Prevotella prophages [72], suggesting that Microviridae could be an important viral family in the human gut and that what was previously considered to be an exclusively lytic phage can integrate into bacterial hosts in an environment that encourages a temperate (lysogenic) viral-host lifestyle (see **Fig. 2** and below).

Marine environments can contain millions of different virotypes in a single sample [51]. None of the human fecal samples characterized thus far has had greater than 1500 virotypes. Moreover, the ratio of virotypes to species-level bacterial phylotypes in the ocean is 10:1 but closer to 1:1 in the gut [22]. Microscopy counts have further validated these estimated ratios, demonstrating an average of 10^8-10^9 VLPs per gram of feces compared to ~10^9 bacterial cells per gram of feces [72]. These findings also support the notion that phage exhibit a more temperate lifestyle in the gut, in contrast to the active kill-the-winner viral-bacterial dynamic manifest in marine environments.

The temperate lifestyle observed in the gut environment along with known bacterial mechanism of resistance such as CRISPRs allowed researchers for computational screening and discovery of novel viruses. Recently, Stern et al, [75] used datasets with deep sequencings of gut microbiomes [24] to extract CRISPR spacers present in gut-associated bacterial genomes. Those spacers were then used to query gut-associated viromes [22, 23] as well as assembled contigs. The results allowed the identification of large collections of contigs of potential viral origin, their association to known bacterial host and the wide distribution of certain phages across different datasets.

Temporal variation - To date, only three reported metagenomic studies of the gut DNA virome have characterized temporal variation [22, 23, 76]. One of these studies used VLPs isolated from

frozen fecal samples collected from four adult female monozygotic twin pairs and their mothers at three time points over a 12-month period [22]. VLP-derived virome datasets were compared to datasets of sequenced bacterial 16S rRNA genes and shotgun reads from total fecal community DNA generated from the same fecal samples used to prepare VLPs. The results disclosed that the viromes of co-twins and their mothers exhibited a significantly greater degree of interpersonal variation than did the corresponding bacterial communities. Despite the marked interpersonal variation in gut viromes and their encoded gene functions, intrapersonal diversity was extremely low, with >95% of virotypes retained over the period surveyed, and with DNA viromes dominated by a few temperate phages that exhibited remarkable genetic stability (>99% sequence conservation). These observations suggested that a temperate viral lifestyle is more prevalent in the distal intestine than a kill-the-winner dynamic (see **Fig. 2**).

Another study of temporal variation involved adults subjected to a defined diet for a period of eight days [23]. During this time, both fecal bacterial and viral communities changed in a comparable manner. Importantly, interpersonal variation at the late time points was reduced among individuals consuming the same diet, suggesting that diet has an important effect in shaping both bacterial and viral communities.

A third study examining temporal variation characterized the DNA virome of a one week old healthy infant and used DNA microarrays to compare relative viral abundances in the fecal microbiota between postnatal weeks one and two [76]. The results showed that at early stages of life, the viral population changes drastically: over half of the virotypes present at one week were undetectable at two weeks. While contrasting with the stability documented in the fecal DNA viromes of healthy adults, these results are consistent with the dynamic and rapid nature of assembly of the infant bacterial microbiota [66, 77].

Functions encoded in phage genomes – There are a number of examples of known phage-encoded host fitness factors in gut bacteria (e.g., lambda *bor* and *lom*, 933W Stx2) [78, 79], but most of these appear to be virulence determinants of one kind or another. When comparing purified VLP

viromes and fecal microbiomes in the study of monozygotic twins, phage exhibit enrichment for genes involved in anaerobic nucleotide synthesis, as well as cell wall biosynthesis and degradation [22]. Other distinctive features of phage genomes include genes that can alter bacterial receptors and prevent superinfection [80]. Interestingly, many phage receptors may be involved in carbohydrate transport and utilization. In an environment such as the gut, where carbohydrate utilization is an important fitness factor, mobilization of these genes by phages could endow their bacterial host with benefits (**Fig. 2**). There are likely a great number of bona-fide (metabolic) fitness factors encoded in phages that have yet to be characterized.

Intriguingly, new evidence suggests that carbohydrate-binding components of the human gut virome may change at an extremely high rate. A recent metagenomic study examining VLPs purified from fecal samples collected from 12 humans identified 51 hypervariable loci, areas with mutation rates that are much higher than the rest of their viral genomes [50]. Protein structural predictions revealed that some do not have homology to known folds, some have similarity to Ig-superfamily proteins, and others have similarity to C-type lectin folds, which play a key role in carbohydrate binding. Moreover, these loci appear to be specifically targeted for mutation by a reverse transcriptase-based mechanism, perhaps suggesting a critical functional advantage provided by these hypervariable loci. It is tempting to speculate that these loci confer a selective advantage to phages, enabling immune evasion through IgA binding, or improving the chances of infecting a host cell in the rapidly changing conditions of the gut through adaptable binding to relevant environmental materials or bacterial surface receptors. There is a well-documented precedent for hypervariable loci conferring a fitness advantage in Bordetella phage by allowing tropism switching in the phage receptor-binding protein [81]. Is important to emphasize that the hypothesis that these loci may allow a phage to bind IgA or environmental ligands is speculative and needs experimental validation; non-receptor phage structural proteins have been found to contain Ig-like domains, which may aid in host binding by weakly interacting with the cell surface [82].

RNA virome - The RNA virome of two healthy adults has been characterized by purifying VLPs [83]. This study showed that most RNA viruses appeared to be consumed together with food.

A pepper-associated virus (PMMV), comprised over 80% of the identifiable gut viruses. The only animal RNA virus observed was a picobirnavirus that had previously been found in the feces of healthy individuals as well as in patients with diarrhea: it has not been associated with any particular disease.

Comparative studies of gut viromes in other mammals

Comparative studies of different mammalian species represent a source of information about the effects of environmental factors, including diet [84, 85], and various host factors on phage diversity in the gastrointestinal tract. Extensive surveys have been conducted of viruses associated with different mammalian species in search of potential sources of zoonosis, the etiology of animal diseases, and to identify common mammalian viruses [13]. These studies, which includes mammals occupying very distinct habitats [29, 31, 32, 86-91], identified viruses from the same families identified in the human gut virome, further underscoring the prevalence and long-standing nature of the evolved viral-mammalian host relationship (**Box 3**). In all these studies, ssDNA viruses were ubiquitous, accompanied in some cases by positive-sense ssRNA enteric viruses [43, 90, 91].

An early survey of coliphages in cows, pigs, and humans [92] showed that they were present in titers of up to 10^7 VLPs per gram of feces and that temperate phages were the most common. Interestingly, humans and pigs (omnivores with simple guts) had higher counts of temperate coliphages than cows (herbivores with foregut fermentation chambers). In an independent study, estimates of phage diversity from bovine rumen fluid [93] suggested that up to 28,000 different virotypes could be present in titers as high as 10^9 VLPs per ml of sample, hinting at strikingly higher diversity and abundance compared to humans. In contrast to the large interpersonal variation observed in human gut viromes [22, 23], this latter study showed a high degree of similarity between the phage communities of cohabitating animals on a similar diet. Metagenomic studies in horses (herbivores containing a hindgut fermentative chamber) revealed an intermediate level of phage diversity between that documented in herbivorous foregut-fermenting ruminants and omnivorous humans with simple guts.

Together, these studies suggest that diet, gut physiology, and potentially the transit time of food, play important roles in determining the lifecycle (and diversity) of phages in the mammalian gut. Further dissection of these relationships requires manipulable, representative, and defined *in vivo* models. Moving in this direction, Maura and colleagues used mice to study the effects of a lytic enteric phage, observing stable long-term replication over three weeks [94]. As noted below, gnotobiotic mouse models may also be very informative.

Phage Therapy

Much has already been written about the history, successes, and failings of phage therapy. Most of the studies have focused on the use of lytic phages to destroy pathogenic bacteria [95] [96] (**Fig. 3**). Clinically oriented phage research began very soon after the discovery of phages, with Felix d'Herelle using phages to treat bacillary dysentery in a number of human patients [8]. This optimistic start, however, led to a number of misconceptions and missteps, both scientific and political, regarding the use of phages. d'Herelle incorrectly assumed that there was only one universally efficacious strain of lytic phage [97], though we now know phages exhibit exquisite host cell specificity. In the 1930s, pharmaceutical companies began distributing enormous amounts of lytic phages as generic antibacterial therapies, but in part because of the perceived universality of phages, they had very little knowledge of their product's components.

In retrospect, many of the commonly used phage preparations were destroyed by the organomercury preservatives added to the vials that contained them, or were contaminated with bacterial exotoxins secreted by the cultures used to generate them [98]. Inevitably these problems, along with manufacturing inconsistencies (the supposedly standardized strains of phages would change from batch-to-batch) led to distrust among the medical and scientific community.

The recent resurgence of phages as possible therapeutics has been driven by a number of factors. The alarming prevalence of antibiotic-resistant strains of pathogenic bacteria, combined with the inexorable spread of antibiotic-degrading enzymes, such as the New Delhi metallo-beta-lactamase (NDM-1), have led to calls for new therapeutic strategies [99]. From a practical standpoint,

antibiotic discovery efforts have produced few novel compounds over the past decade [100]. Phages are a promising tool as they are easy to manufacture, have good host specificity, and can be readily genetically manipulated. Moreover, resistance to phages may develop more slowly than to antibiotics, though the reasons for this are multifaceted [101]. Phage resistance can occur spontaneously in cultures (as frequently as 1 in 10^5 cells), but there can be fitness costs associated with resistance. In contrast, many forms of antibiotic resistance cannot occur spontaneously, but instead require introduction of a foreign DNA element. In many ways, addressing bacterial resistance is much easier with phages than with antibiotics because one can isolate different phages, or phages may spontaneously mutate to overcome host resistance.

Perhaps a more interesting question, in the context of community dynamics and our growing understanding of the virome and microbiome, is whether we can produce more subtle phenotypic shifts in an ecological niche. Rather than destroy a single pathogenic member of a community, lysogenic phages could be introduced to promote a community structure that is beneficial to both the human host and microbial community members (**Fig. 3**). For example, one could expand the capacity of the gut microbiome to degrade dietary components [102]. Similarly, phage could be used to introduce novel, beneficial traits to community members, such as those involving nutrient biosynthesis. In the latter circumstance, it may be difficult to introduce traits that are not purely beneficial to a lysogenic phage's bacterial host, as the energetic effects of synthesizing an unnecessary protein impose a selection pressure.

Given the potential power and replicating nature of phages, a number of questions must be addressed before they can be more widely adopted including issues related to bio-containment [101]. Although phages are frequently sold as viruses that 'can only infect bacteria', their safety has yet to be completely defined. The intravenous administration of phage (e.g., in the case of bacterial sepsis) is particularly complex given the immunogenicity of some preparations and rapid clearance of phage particles by the reticuloendothelial system of the spleen [103]. It is tempting to assume that other routes of administration, such as oral cocktails of phage to target the human gut microbiome, would not have such effects. However, phage DNA is detectable by PCR and FISH in serum

shortly after oral consumption [104]. Other studies have provided evidence of trans-placental passage of phage [105]. There is data suggesting that enzymes transcribed from phage DNA can be expressed in mammalian cells [106], this finding has even led to attempts to use phage as gene therapy vectors [107, 108].

Despite these concerns, we are exposed to millions of phages every day, including the ones from our own microbiota, without significant observable harm. In this spirit, it is interesting to consider the potential 'therapeutic' use of phages in the context of current efforts to apply microbiome-directed therapies [109]. Questions that can be asked include whether it is beneficial or detrimental for bacterial taxa being considered as candidate probiotics to possess or lack prophages or whether phages should be deliberately administered coincident with or preceding introduction of a probiotic consortium to help create niches that promote successful invasion and engraftment of the consortium.

Future directions

There has been little experimental work done on the ecology of phages *in vivo*. Germ-free mice and mice mono-colonized with different strains of *E. coli*, including strains isolated from children with diarrhea, have been used to examine replication of T4 and T7 phages [110, 111]. Gnotobiotic mouse models of the human gut microbiota may not only provide better understanding of phage-host dynamics, but may also represent a potentially valuable tool for establishing a preclinical pipeline designed to evaluate the feasibility of phage therapy. Recent work has shown that transplanting intact uncultured human gut (fecal) microbial communities into gnotobiotic mice is efficient, capturing the majority of microbial diversity and microbiome-encoded functions present in the human donor's community in the recipient animals [112, 113]. Mice with replicated human gut microbiomes can be fed diets resembling those of the human donor to explore diet-microbiome-phage interactions. An additional refinement to this approach is to transplant sequenced collections of bacteria (some containing prophage) cultured from a given donor's fecal sample into recipient mice [113]. The effects of various perturbations of gnotobiotic mice harboring these microbiota on phage-bacterial dynamics can be studied over time under highly controlled conditions.

Yet another envisioned approach is to assemble defined communities of sequenced members of the human gut microbiota in formerly germ-free mice and then to deliberately stage phage attacks. VLPs prepared from human fecal samples and introduced into mice containing these sequenced collections of cultured microbes from human donors would allow investigators to directly determine the bacterial host specificity of the VLP-associated phage, as well as the effects of (i) the presence or absence of prophage in community members, (ii) the diet administered to the animals, and (iii) the phage's contribution to mammalian host physiology, including immune function. The impact of a staged phage attack on community structure and functions can also be defined in gnotobiotic animal models over time and as a function of location along the length of the gut, where available nutrient and energy resources vary considerably.

These model systems, coupled with observational data in human fecal samples, should help us understand how phages influence metabolism in the gut and possibly how to manipulate it. Obtaining answers will almost certainly demand new and more efficient methods for deliberately curing bacterial hosts of their prophage. It may also require the application of whole genome transposon mutagenesis methods that are married to next generation sequencing platforms [114] to identify the functional contributions of genes in a given prophage. While the journey ahead will certainly be demanding, approaches are in hand or can be envisioned that will help propel the 'new age of phage' forward so that long standing questions can be addressed and new insights can be obtained.

Acknowledgements

Work from the authors' labs described in this report was supported by grants from the NIH (DK78669, DK30292, DK70977 to JIG and GM095384 to FLR), and the Crohn's and Colitis Foundation of America. A.R. is the recipient of an International *Fulbright Science and Technology Award.* N.P.S is a member of Washington University's Medical Scientist Training Program (MSTP) funded by NIH grant GM007200. Due to space limits we were not able to cite many wonderful studies relevant to the topics covered.

References

1. Hershey, A.D. & Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36**, 39-56 (1952).

2. Crick, F.H., Barnett, L., Brenner, S. & Watts-Tobin, R.J. General nature of the genetic code for proteins. *Nature* **192**, 1227-1232 (1961).

3. Carins, J., Stent, G.S. & Watson, J.D. Phage and the Origins of Molecular Biology (Cold Spring Harbor Laboratory Press, Plainview, NY, 1992).

4. Mokili, J.L., Rohwer, F. & Dutilh, B.E. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* **2**, 63-77 (2012).

5. Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13**, 278-284 (2005).

6. Fernandes, P. Antibacterial discovery and development--the failure of success? *Nat Biotechnol* **24**, 1497-1503 (2006).

7. d'Herelle, F. Sur un microbe invisible antagoniste des bacilles dysenteriques. *Comptes rendus Acad Sci Paris* **165**, 373-375 (1917).

8. Sulakvelidze, A., Alavidze, Z. & Morris, J.G., Jr. Bacteriophage therapy. *Antimicrob Agents Chemother* **45**, 649-659 (2001).

9. Levin, B.R. & Bull, J.J. Population and evolutionary dynamics of phage therapy. *Nat Rev Microbiol* **2**, 166-173 (2004).

10. Marraffini, L.A. & Sontheimer, E.J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11**, 181-190 (2010).

11. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167-170 (2010).

12. Virgin, H.W., Wherry, E.J. & Ahmed, R. Redefining chronic viral infection. *Cell* **138**, 30-50 (2009).

13. Delwart, E. Animal virus discovery: improving animal health, understanding zoonoses, and opportunities for vaccine development. *Curr Opin Virol* **2**, 1-9 (2012).

14. Haynes, M. & Rohwer, F. The Human Virome in Metagenomics of the Human Body. (ed. Nelson, K.E.) 63-77 (2011).

15. Fox, G.E. et al. The phylogeny of prokaryotes. *Science* **209**, 457-463 (1980).

16. Lane, D.J. et al. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A* **82**, 6955-6959 (1985).

17. Hugenholtz, P., Goebel, B.M. & Pace, N.R. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* **180**, 4765-4774 (1998).

18. Culley, A.I., Lang, A.S. & Suttle, C.A. High diversity of unknown picorna-like viruses in the sea. *Nature* **424**, 1054-1057 (2003).

19. Breitbart, M., Miyake, J.H. & Rohwer, F. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol Lett* **236**, 249-256 (2004).

20. Hambly, E. et al. A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage S-PM2. *Proc Natl Acad Sci U S A* **98**, 11411-11416 (2001).

21. Casjens, S.R. Comparative genomics and evolution of the tailed-bacteriophages. *Curr Opin Microbiol* **8**, 451-8 (2005).

22. Reyes, A. et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334-338 (2010).

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

24. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010).

25. Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat Protoc* **4**, 470-483 (2009).

26. Willner, D. et al. Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4547-4553 (2011).

27. Rohwer, F., Seguritan, V., Choi, D.H., Segall, A.M. & Azam, F. Production of shotgun libraries using random amplification. *BioTechniques* **31**, 108-112 (2001).

28. Breitbart, M. et al. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**, 6220-6223 (2003).

29. Shan, T. et al. The fecal virome of pigs on a high-density farm. *J Virol* **85**, 11697-11708 (2011).

30. Yozwiak, N.L. et al. Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl Trop Dis* **6**, e1485 (2012).

31. Li, L. et al. Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. *J Virol* **84**, 6955-6965 (2010).

32. Ge, X. et al. Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in china. *J Virol* **86**, 4620-4630 (2012).

33. Hutchison, C.A., Smith, H.O., Pfannkoch, C. & Venter, J.C. Cell-free cloning using φ29 DNA polymerase. *Proc Natl Acad Sci U S A* **102**, 17332 (2005).

34. Kim, K.H. et al. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl Environ Microbiol* **74**, 5975-5985 (2008).

35. Lasken, R.S. & Stockwell, T.B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol* **7** (2007).

36. Kim, K.H. & Bae, J.W. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl Environ Microbiol* **77**, 7663-7668 (2011).

37. Andrews-Pfannkoch, C., Fadrosh, D.W., Thorpe, J. & Williamson, S.J. Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. *Appl Environ Microbiol* **76**, 5039-5045 (2010).

38. Fadrosh, D.W., Andrews-Pfannkoch, C. & Williamson, S.J. Separation of single-stranded DNA, double-stranded DNA and RNA from an environmental viral community using hydroxyapatite chromatography. *J Vis Exp* (2011).

39. Marine, R. et al. Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl Environ Microbiol* **77**, 8071-8079 (2011).

40. Nakamura, S. et al. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE* **4**, e4219 (2009).

41. Wommack, K.E., Bhavsar, J. & Ravel, J. Metagenomics: read length matters. *Appl Environ Microbiol* **74**, 1453-1463 (2008).

42. Bibby, K., Viau, E. & Peccia, J. Viral metagenome analysis to guide human pathogen monitoring in environmental samples. *Lett Appl Microbiol* **52**, 386-392 (2011).

43. Ng, T.F. et al. Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS ONE* **6**, e20579 (2011).

44. Pasic, L. et al. Metagenomic islands of hyperhalophiles: the case of Salinibacter ruber. *BMC Genomics* **10**, 570 (2009).

45. Vega Thurber, R.L. et al. Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral Porites compressa. *Proc Natl Acad Sci U S A* **105**, 18413-18418 (2008).

46. Dinsdale, E.A. et al. Functional metagenomic profiling of nine biomes. *Nature* **452**, 629-632 (2008).

47. Yang, J. et al. Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J Clin Microbiol* **49**, 3463-3469 (2011).

48. Xu, B. et al. Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. *PLoS Pathog* **7**, e1002369 (2011).

49. Coetzee, B. et al. Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. *Virology* **400**, 157-163 (2010).

50. Minot, S., Grunberg, S., Wu, G.D., Lewis, J.D. & Bushman, F.D. Hypervariable loci in the human gut virome. *Proc Natl Acad Sci U S A* **109**, 3962-3966 (2012)**.**

51. Rohwer, F. & Thurber, R.V. Viruses manipulate the marine environment. *Nature* **459**, 207-212 (2009).

52. Leplae, R., Lima-Mendez, G. & Toussaint, A. ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res* **38**, D57-61 (2010).

53. Edwards, R. Phages in the SEED Viewer, Accessed: http://www.phantome.org/PhageSeed/Phage.cgi (2012).

54. Sharon, I. et al. Comparative metagenomics of microbial traits within oceanic viral communities. *ISME J* **5**, 1178-1190 (2011).

55. Ghosh, T.S., Mohammed, M.H., Komanduri, D. & Mande, S.S. ProViDE: A software tool for accurate estimation of viral diversity in metagenomic samples. *Bioinformation* **6**, 91-94 (2011).

56. Lorenzi, H.A. et al. TheViral MetaGenome Annotation Pipeline(VMGAP):an automated tool for the functional annotation of viral Metagenomic shotgun sequencing data. *Stand Genomic Sci* **4**, 418-429 (2011).

57. Meyer, F. et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).

58. Edwards, R., Breitbart, M., Elhai, J. & Sullivan, M. PHANTOME, PHage ANnotation TOols and MEthods, Accessed: http://www.phantome.org/ (2011).

59. Bhavsar, J., Polson, S.W. & Wommack, K.E. VIROME: An informatics resource for viral metagenome exploration, Accessed: http://virome.dbi.udel.edu (2009).

60. Roux, S. et al. Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27**, 3074-3075 (2011).

61. Sun, S. et al. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* **39**, D546-551 (2011).

62. Breitbart, M. et al. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**, 14250-14255 (2002).

63. Angly, F. et al. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**, 41 (2005).

64. Angly, F.E. et al. The marine viromes of four oceanic regions. *PLoS Biol* **4**, e368 (2006).

65. Dutilh, B.E. et al. Reference-independent comparative metagenomics using cross-assembly: crAss, Accessed: http://edwards.sdsu.edu/crass/ (2012).

66. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222-227 (2012).

67. Turnbaugh, P.J. et al. A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484 (2009).

68. Turnbaugh, P.J. et al. Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc Natl Acad Sci U S A* **107**, 7503-7508 (2010).

69. Hansen, E.E. et al. Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4599-4606 (2011).

70. Caporaso, J.G. et al. Moving pictures of the human microbiome. *Genome Biol* **12**, R50 (2011).

71. Costello, E.K. et al. Bacterial community variation in human body habitats across space and time. *Science* **326**, 1694-1697 (2009).

72. Kim, M.S., Park, E.J., Roh, S.W. & Bae, J.W. Diversity and abundance of single-stranded DNA viruses in human feces. *Appl Environ Microbiol* **77**, 8062-8070 (2011).

73. Krupovic, M. & Forterre, P. Microviridae goes temperate: microvirus-related proviruses reside in the genomes of Bacteroidetes. *PLoS ONE* **6**, e19893 (2011).

74. Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174-180 (2011).

75. Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res* (2012).

76. Breitbart, M. et al. Viral diversity and dynamics in an infant gut. *Res Microbiol* **159**, 367-373 (2008).

77. Palmer, C., Bik, E.M., DiGiulio, D.B., Relman, D.A. & Brown, P.O. Development of the human infant intestinal microbiota. *PLoS Biol* **5**, e177 (2007).

78. Barondess, J.J. & Beckwith, J. A bacterial virulence determinant encoded by lysogenic coliphage lambda. *Nature* **346**, 871-4 (1990).

79. Plunkett, G., 3rd, Rose, D.J., Durfee, T.J. & Blattner, F.R. Sequence of Shiga toxin 2 phage 933W from Escherichia coli O157:H7: Shiga toxin as a phage late-gene product. *J Bacteriol* **181**, 1767-78 (1999).

80. Markine-Goriaynoff, N. et al. Glycosyltransferases encoded by viruses. *J Gen Virol* **85**, 2741-2754 (2004).

81. Liu, M. et al. Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. *Science* **295**, 2091-4 (2002).

82. Fraser, J.S., Yu, Z., Maxwell, K.L. & Davidson, A.R. Ig-like domains on bacteriophages: a tale of promiscuity and deceit. *J Mol Biol* **359**, 496-507 (2006).

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

84. Ley, R.E. et al. Evolution of mammals and their gut microbes. *Science* **320**, 1647-1651 (2008).

85. Muegge, B.D. et al. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332**, 970-974 (2011).

86. Cann, A.J., Fandrich, S.E. & Heaphy, S. Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes* **30**, 151-156 (2005).

87. Donaldson, E.F. et al. Metagenomic analysis of the viromes of three North American bat species: viral diversity among different bat species that share a common habitat. *J Virol* **84**, 13004-13018 (2010).

88. Blinkova, O. et al. Novel circular DNA viruses in stool samples of wild-living chimpanzees. *J Gen Virol* **91**, 74-86 (2010).

89. Ng, T.F. et al. Metagenomic identification of a novel anellovirus in Pacific harbor seal (Phoca vitulina richardsii) lung samples and its detection in samples from multiple years. *J Gen Virol* **92**, 1318-1323 (2011).

90. Phan, T.G. et al. The fecal viral flora of wild rodents. *PLoS Pathog* **7**, e1002218 (2011).

91. van den Brand, J.M. et al. Metagenomic analysis of the viral flora of pine marten and European badger feces. *J Virol* **86**, 2360-2365 (2012).

92. Dhillon, T.S., Dhillon, E.K., Chau, H.C., Li, W.K. & Tsang, A.H. Studies on bacteriophage distribution: virulent and temperate bacteriophage content of mammalian feces. *Appl Environ Microbiol* **32**, 68-74 (1976).

93. Berg Miller, M.E. et al. Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. *Environ Microbiol* **14**, 207-227 (2012).

94. Maura, D. et al. Intestinal colonization by enteroaggregative Escherichia coli supports long-term bacteriophage replication in mice. *Environ Microbiol* (2011).

95. Fischetti, V.A., Nelson, D. & Schuch, R. Reinventing phage therapy: are the parts greater than the sum? *Nat Biotechnol* **24**, 1508-1511 (2006).

96. Lu, T.K. & Koeris, M.S. The next generation of bacteriophage therapy. *Curr Opin Microbiol* **14**, 524-531 (2011).

97. van Helvoort, T. The controversy between John H. Northrop and Max Delbruck on the formation of bacteriophage: bacterial synthesis or autonomous multiplication? *Ann Sci* **49**, 545-575 (1992).

98. Calendar, R.L. The Bacteriophages (Oxford University Press, 2005).

99. Kumarasamy, K.K. et al. Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study. *Lancet Infect Dis* **10**, 597-602 (2010).

100. Piddock, L.J. The crisis of no new antibiotics--what is the way forward? *Lancet Infect Dis* **12**, 249-253 (2012).

101. Clokie, M.R.J. & Kropinski, A.M. Bacteriophages : methods and protocols (Humana Press; Springer, London, 2009).

102. Hehemann, J.H. et al. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908-912 (2010).

103. Geier, M.R., Trigg, M.E. & Merril, C.R. Fate of bacteriophage lambda in non-immune germ-free mice. *Nature* **246**, 221-223 (1973).

104. Schubbert, R., Renz, D., Schmitz, B. & Doerfler, W. Foreign (M13) DNA ingested by mice reaches peripheral leukocytes, spleen, and liver via the intestinal wall mucosa and can be covalently linked to mouse DNA. *Proc Natl Acad Sci U S A* **94**, 961-966 (1997).

105. Schubbert, R., Hohlweg, U., Renz, D. & Doerfler, W. On the fate of orally ingested foreign DNA in mice: chromosomal association and placental transmission to the fetus. *Mol Gen Genet* **259**, 569-576 (1998).

106. Geier, M.R. & Merril, C.R. Lambda phage transcription in human fibroblasts. *Virology* **47**, 638-643 (1972).

107. Barry, M.A., Dower, W.J. & Johnston, S.A. Toward cell-targeting gene therapy vectors: selection of cell-binding peptides from random peptide-presenting phage libraries. *Nat Med* **2**, 299-305 (1996).

108. Dunn, I.S. Mammalian cell binding and transfection mediated by surface-modified bacteriophage lambda. *Biochimie* **78**, 856-861 (1996).

109. Silverman, M.S., Davis, I. & Pillai, D.R. Success of self-administered home fecal transplantation for chronic Clostridium difficile infection. *Clin Gastroenterol Hepatol* **8**, 471-473 (2010).

110. Chibani-Chennoufi, S. et al. In vitro and in vivo bacteriolytic activities of Escherichia coli phages: implications for phage therapy. *Antimicrob Agents Chemother* **48**, 2558-69 (2004).

111. Weiss, M. et al. In vivo replication of T4 and T7 bacteriophages in germ-free mice colonized with Escherichia coli. *Virology* **393**, 16-23 (2009).

112. Turnbaugh, P.J. et al. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci Transl Med* **1**, 6ra14 (2009).

113. Goodman, A.L. et al. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc Natl Acad Sci U S A* **108**, 6252-6257 (2011).

114. Goodman, A.L. et al. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* **6**, 279-289 (2009).

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

116. Lu, T.K. & Collins, J.J. Engineered bacteriophage targeting gene networks as adjuvants for antibiotic therapy. *Proc Natl Acad Sci U S A* **106**, 4629-4634 (2009).

117. Whitman, W.B., Coleman, D.C. & Wiebe, W.J. Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95**, 6578-6583 (1998).

118. Bergh, O., Borsheim, K.Y., Bratbak, G. & Heldal, M. High abundance of viruses found in aquatic environments. *Nature* **340**, 467-8 (1989).

119. Clokie, M.R., Millard, A.D., Letarov, A.V. & Heaphy, S. Phages in nature. *Bacteriophage* **1**, 31-45 (2011).

120. Fuhrman, J.A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541-548 (1999).

121. Azam, F. et al. The ecological role of water column microbes in the sea. *Mar. Ecol. Prog. Ser.* **10**, 257-263 (1983).

122. Marston, M.F. et al. Rapid diversification of coevolving marine Synechococcus and a virus. *Proc Natl Acad Sci U S A* **109**, 4544-4549 (2012).

123. Mann, N.H., Cook, A., Millard, A., Bailey, S. & Clokie, M. Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* **424**, 741 (2003).

124. Sullivan, M.B. et al. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol* **4**, e234 (2006).

125. Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M. & Chisholm, S.W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**, 86-89 (2005).

126. Anderson, R.E., Brazelton, W.J. & Baross, J.A. Is the genetic landscape of the deep subsurface biosphere affected by viruses? *Front Microbiol* **2**, 219 (2011).

127. Roossinck, M.J. The good viruses: viral mutualistic symbioses. *Nat Rev Microbiol* **9**, 99-108 (2011).

128. Roossinck, M.J. Changes in population dynamics in mutualistic versus pathogenic viruses. *Viruses* **3**, 12-19 (2011).

129. Brown, S.P., Le Chat, L., De Paepe, M. & Taddei, F. Ecology of microbial invasions: Amplification allows virus carriers to invade more rapidly when rare. *Current Biology* **16**, 2048-2052 (2006).

130. Brown, S.P., Inglis, R.F. & Tadddei, F. Evolutionary ecology of microbial wars: within-host competition and (incidental) virulence. *Evol. Appl.* **2**, 32-39 (2009).

131. Moran, N.A., Degnan, P.H., Santos, S.R., Dunbar, H.E. & Ochman, H. The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. *Proc Natl Acad Sci U S A* **102**, 16919-16926 (2005).

132. Oliver, K.M., Degnan, P.H., Hunter, M.S. & Moran, N.A. Bacteriophages encode factors required for protection in a symbiotic mutualism. *Science* **325**, 992-994 (2009).

133. Xu, P. et al. Virus infection improves drought tolerance. *New Phytol* **180**, 911-921 (2008).

134. Marquez, L.M., Redman, R.S., Rodriguez, R.J. & Roossinck, M.J. A virus in a fungus in a plant: three-way symbiosis required for thermal tolerance. *Science* **315**, 513-515 (2007).

135. Ophel, K.M., Bird, A.F. & Kerr, A. Association of Bacteriophage Particles with Toxin Production by Clavibacter toxicus, the Causal Agent of Annual Ryegrass Toxicity. *Phytopathology* **83**, 676-681 (1993).

136. Holtz, L.R., Finkbeiner, S.R., Kirkwood, C.D. & Wang, D. Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea. *Virol J* **5**, 159 (2008).

137. Finkbeiner, S.R. et al. Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog* **4**, e1000011 (2008).

138. Finkbeiner, S.R. et al. Human stool contains a previously unrecognized diversity of novel astroviruses. *Virol J* **6**, 161 (2009).

139. Phan, T.G. et al. A third gyrovirus species in human feces. *J Gen Virol* (2012).

140. Kapoor, A. et al. Multiple novel astrovirus species in human stool. *J Gen Virol* **90**, 2965-2972 (2009).

141. Kapoor, A. et al. Human bocaviruses are highly diverse, dispersed, recombination prone, and prevalent in enteric infections. *J Infect Dis* **201**, 1633-1643 (2010).

142. Victoria, J.G. et al. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J Virol* **83**, 4642-4651 (2009).

143. Rosario, K., Duffy, S. & Breitbart, M. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol* (2012).

Figure Legends

Figure 1– Experimental and computational methods for the characterization of the phage populations present in the human gut microbiota. See main text for further details. VLP, virus-like particle.

Figure 2– Potential consequences of a temperate phage lifecycle in the human gut. Viral metagenomic show that the phage population associated with the adult human gut microbiota is characterized by a relatively low number of virotypes compared to other ecosystems (e.g., soils, sediments, marine environments). These gut populations also exhibit high temporal stability of virotypes with respect to both viral community structure and nucleotide sequence conservation, and a high prevalence of temperate phages. These characteristics suggest that a temperate lifestyle is dominant in the distal human gut versus the lytic lifestyle observed in open oceans. (**a-c**) Illustration of the benefits of this temperate lifestyle on phage-host dynamics. (**a)** Integration as a prophage protects the host from superinfection, effectively 'immunizing' the bacterial host against infection from the same or closely related phages. Furthermore, the genes encoded by the viral genome may expand the niche of the bacterial host by enabling metabolism of new nutrient sources (e.g., carbohydrates), providing antibiotic resistance, conveying virulence factors, or altering host gene expression. This temperate lifecycle allows viral expansion in a 1:1 ratio with the bacterial host. If the integrated virus conveys increased fitness to its bacterial host, there will be increased prevalence of the host and phage in the microbiota. **(b)** Induction of a lytic cycle may follow a lysogenic state and can be triggered by environmental stress. As a consequence, bacterial turnover is accelerated and energy utilization optimized through 'phage shunts', where the debris remaining after lysis is used as a nutrient source by the surviving population. Furthermore, a subpopulation of bacteria that undergoes lytic induction sweeps away other sensitive species and increases the niche for survivors (i.e., bacteria that already have an integrated phage). Periodic induction of prophages can also lead to a constant diversity dynamic [115], which helps maintain community structure and functional efficiency. (**c)** Novel infections or infections of novel bacterial hosts by phages bring the

benefit of horizontally transferred genes, and create selective pressure on the hosts for diversification of their phage receptors, which are often involved in carbohydrate utilization.

Figure 3 – Potential strategies for phage therapy. (**a**) The traditional strategy has been to use a lytic phage against pathogenic bacteria. While transiently useful, this approach can lead to rapid resistance given positive selection for subpopulations ('clones') that are resistant to the lytic phage. Note that this is an antiquated approach to phage therapy that can be trivially improved by using multiple phages with non-overlapping host resistance patterns, or by selecting for phage mutants that overcome host resistance. (**b**) More recently, synergistic relationships between phage and antibiotics have been exploited, where lysogenic phages are introduced that alone do not kill the pathogen, but instead decrease its survival when used in concert with antibiotics. An example is a phage that inhibits a DNA damage repair system (SOS), which makes bacteria exquisitely sensitive to quinolone-class antibiotics [116]. (**c**) With our growing understanding of the human microbiome, it may be possible to take a more nuanced approach selectively manipulating (enhancing) microbial community functions or clearing the way for invasion by probiotic consortia. Strategies can be envisioned to benefit both microbes and their host; for example, introducing genes into phage genomes that are involved in nutrient biosynthesis (with direct benefits to the bacterial and potentially human host), or degradation of nutrients (which may stabilize the representation and niches of beneficial microbes, especially during times of acute stress).

Figures

Figure 1.

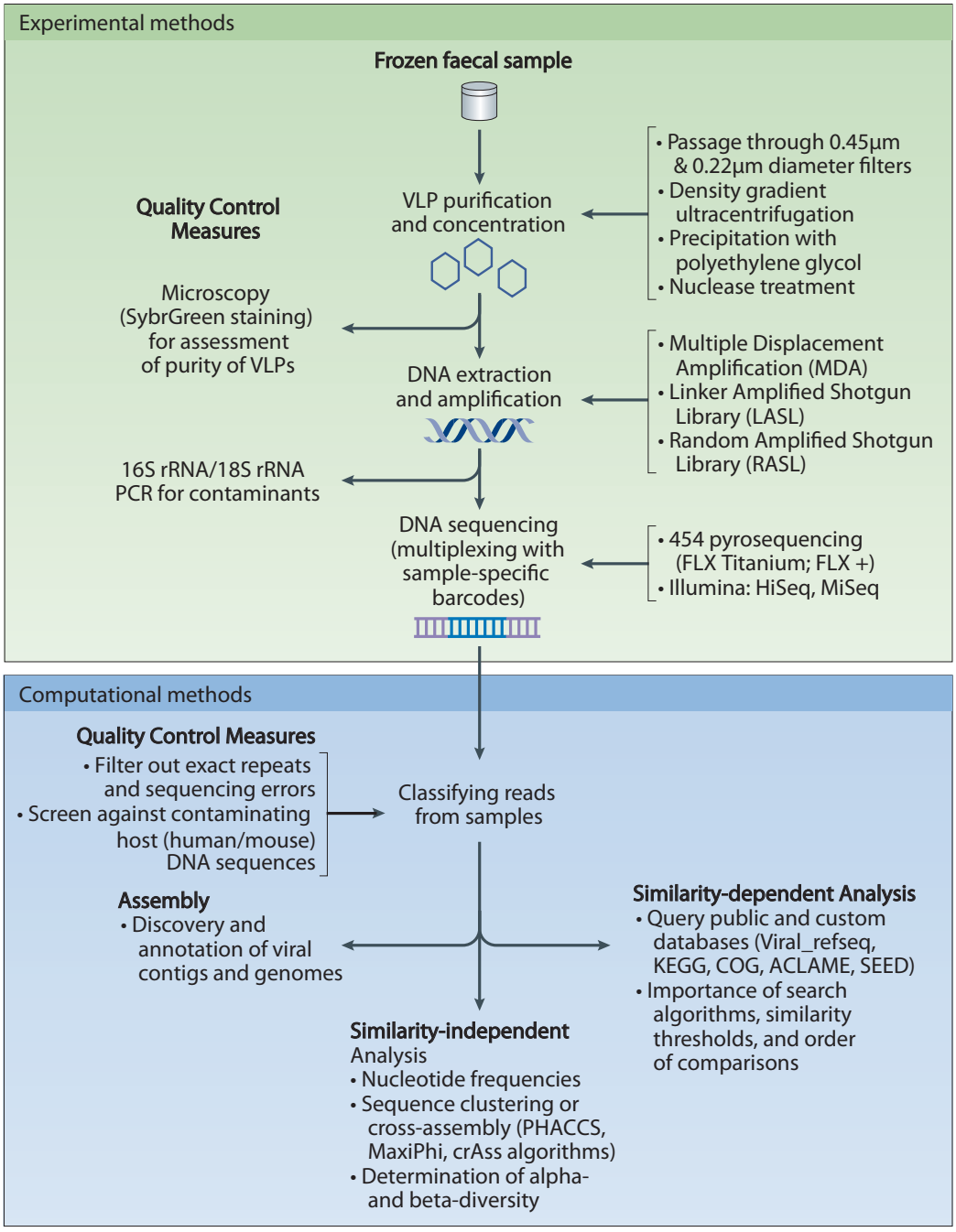


Figure 2.

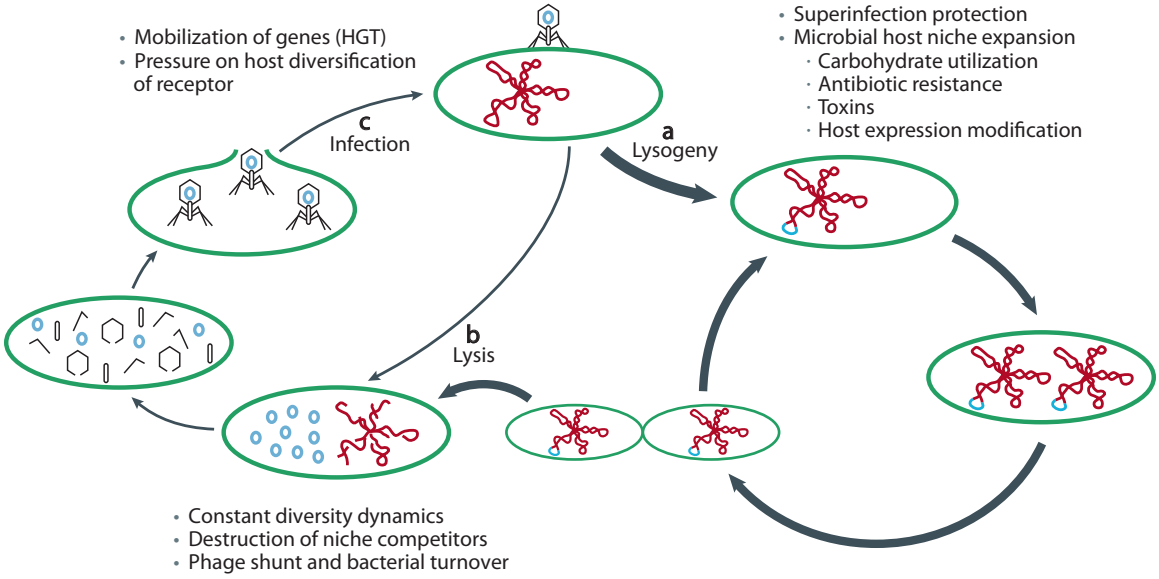
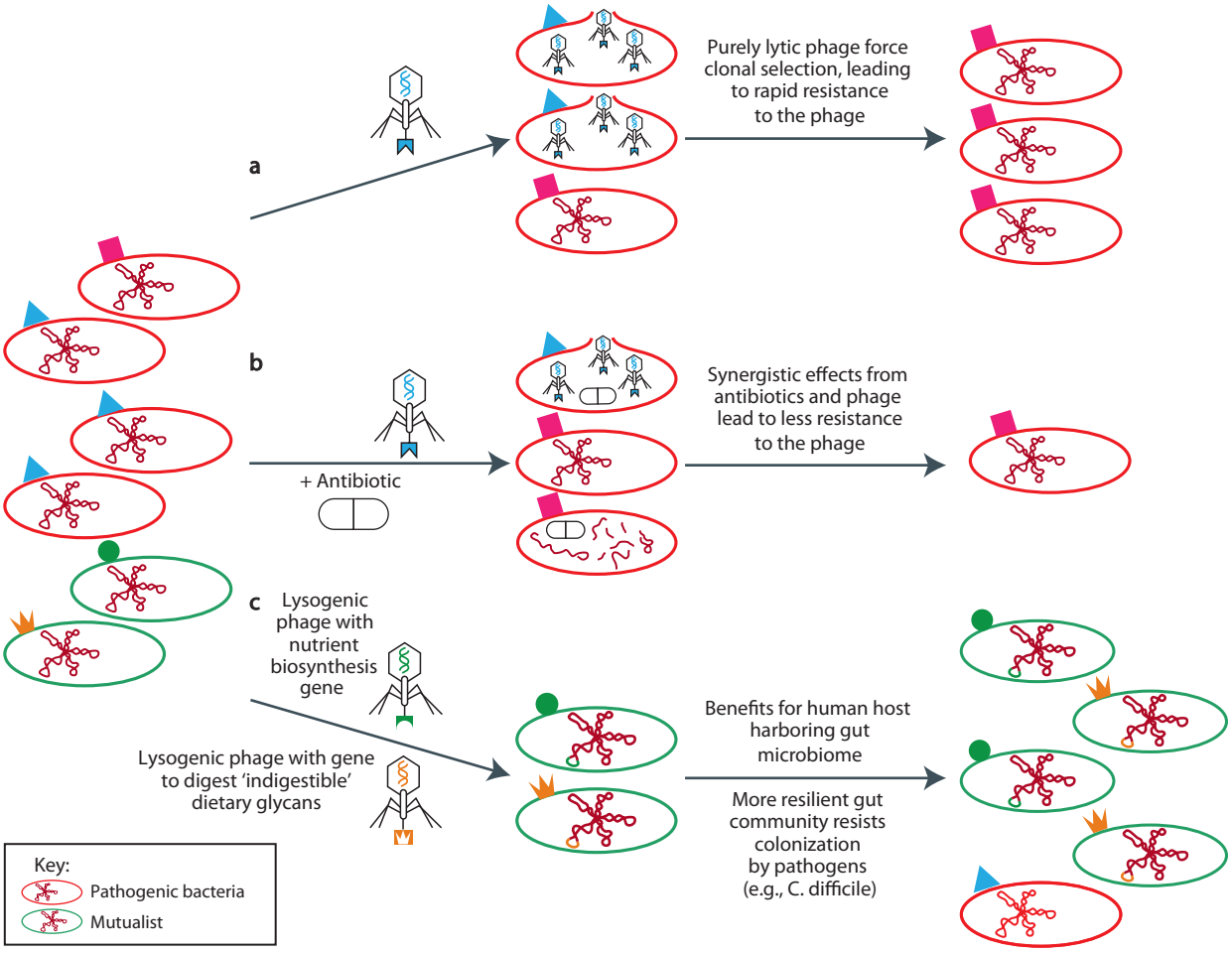

• Mobilization of genes (HGT)
• Pressure on host diversification of receptor
c
Infection
a
Lysogeny
• Superinfection protection
• Microbial host niche expansion
· Carbohydrate utilization
· Antibiotic resistance
· Toxins
· Host expression modification
b
Lysis
• Constant diversity dynamics
• Destruction of niche competitors
• Phage shunt and bacterial turnover

Figure 3.


a
Purely lytic phage force clonal selection, leading to rapid resistance to the phage
b
+ Antibiotic
Synergistic effects from antibiotics and phage lead to less resistance to the phage
c
Lysogenic phage with nutrient biosynthesis gene
Lysogenic phage with gene to digest 'indigestible' dietary glycans
Benefits for human host harboring gut microbiome
More resilient gut community resists colonization by pathogens (e.g., C. difficile)
Key:
Pathogenic bacteria
Mutualist

Box 1—Phage-bacterial host cell dynamic: lessons learned from environmental ecosystems

Most of life on Earth exists as Bacteria and Archaea in the ocean, sediments, land, and potentially the deep biosphere [117]. In the early 1980s, people became aware that there are literally millions of actively growing microbes per milliliter of seawater. Marine phages were subsequently rediscovered in 1989 and efforts to characterize the impact of phage lifecycles on planetary-scale biogeochemistry were initiated [118-121]. As an example of how dramatic phage effects can be, consider the widespread cyanobacteria clades *Prochlorococcus* and *Synechococcus*. These two unicellular algae carry out about half of the world oceans' primary production. They are infected by cyanophages: variations in cyanophage and bacterial concentrations are tied to daily and seasonal cycles [120]. In the 1990s, isolation and characterization of phages infecting *Synechococcus* and later *Prochlorococus* not only revealed that cyanophages are widespread, often infecting 40-50% of cyanobacteria, but that they kill 10-50% of their hosts daily, [119, 120] rapidly driving the diversification of their hosts as they co-evolve resistance [122] and simultaneously driving carbon into a dissolved form when bacterial cells lyse. Under lower resource conditions, cyanophages may enter a lysogenic state. A growing list of genes that are important for bacterial host metabolism and function have been found in marine phages, including photosystem genes that can increase photosynthetic output and maintain energy production during infection so that phages can lyse their host cells [51, 123-125]. Lysogeny may be an important lifestyle under a number of suboptimal conditions, including when host abundance or nutrient abundance is low [126].

Bacteria can use temperate phage to enable invasion of new habitats by sacrificing part of the population through phage lysis. The released phage will target competitors but allow bacterial kin harboring the prophage to survive since they are resistant to attack by a process called superinfection exclusion [127-130].

The concept that a virus can have a beneficial effect beyond that experienced by its host cells is not conceptually novel. Three-way symbioses have been well described in macro- and micro-ecosystems. For example, a symbiotic bacterium that inhabits pea aphids protects them

from a wasp that can otherwise lay eggs in the haemocoel of the aphids; a phage-encoded toxin expressed by the bacteria confers this protection [131, 132]. Drought, heat and cold tolerance are conferred to plants through viruses [127, 133]. In Yellowstone National Park, a fungal endophyte infecting panic grass confers thermal tolerance, allowing the grass to grow in hot geothermal soils. The fungus alone is not heat tolerant without the virus that infects it [134]. In a more phage-related example, there is a reported case of phage-associated corynetoxin synthesis in the bacterium *Rathayibacter toxicus* (formerly *Clavibacter toxicus*) that colonizes ryegrass plants; the toxin makes the grass toxic to grazing animals such as sheep [135].

Box 2 – Fundamental technical challenges in viral metagenomics

- New and better tools for recovery of VLPs from small amounts of starting microbial community biomass, and methods for less biased amplification of extracted nucleic acids before shotgun sequencing.
- Improved methods for deep-draft assemblies of full-length viral genomes. Particularly problematic are the ends of phage genomes, which can be blocked, permuted, or have hairpins.
- Automation of methods (e.g., MaxiPhi) for performing comparative metagenomics and estimating beta and gamma-diversity, in order to describe the pan-virome in a given environment.
- New and better tools for defining the host specificity of known and novel phages from either assembled genomes or VLP-derived short read sequences, and for identifying the determinants of microbial host cell range.
- Improved methods of experimentally and computationally assigning functions to 'conserved' viral genes with no known functions (illuminating the 'genetic dark matter' represented by conserved hypothetical genes).

- Better *in vitro* and *in vivo* models for determining the phage-bacterial host dynamics and its impact on energy availability and niche partitioning in a microbiota.
- Experimental and computational methods, and related visualization tools, for efficient analyses of temporal variation in model viral-bacterial communities and for measuring the effects of perturbations.
- Models for predicting the cost/benefit of having prophage present in a candidate probiotic species; e.g., weighing invasiveness and persistence in a targeted microbiota.
- Methods to test the utility of phages to directly perturb a targeted microbiota in ways that facilitate invasion by a probiotic species or species consortium.

Box 3 – Characterizing the eukaryotic gut virome in healthy individuals

The eukaryotic virome can be considered from at least three different perspectives: viruses associated with the eukaryotic component of a gut microbiota, viruses associated with various human cell populations exposed to this microbiota, and viruses associated with ingested food. Metagenomic studies of healthy individuals are dominated by bacterial viruses with eukaryotic viruses either absent [23] or present at very low abundance [22, 28, 50, 72, 76]. RNA viruses are an apparent exception: while abundant, they appear to be largely derived from ingested food [83]. Our limited view of the eukaryotic gut virome is largely derived from studies that apply viral metagenomics to identify agents associated with gastrointestinal diseases [136-142]. These studies identified known enteric viruses (Adenovirus, Rotavirus, Enterovirus and Norovirus), novel members of Bocavirus, Picobirnavirus, Cosavirus, and Anellovirus that are potential human pathogens, as well as novel viruses that may be related to diet (Gyrovirus, Nodavirus and members of the Dicistroviridae, Vigaviridae and Partitiviridae families). The high prevalence of eukaryotic ssDNA viruses in metagenomic datasets had lead to a whole new perspective of the potential importance and diversity of these small viruses [143]. Although identification of these viruses has come from symptomatic individuals, they have also been identified, at similar prevalence, in asymptomatic contacts of those with

disease. The broad representation of these eukaryotic viruses in the human gut as well as other body habitats has prompted a call to consider the functional significance of the eukaryotic human virome [12].The almost ubiquitous presence of human viruses that are not phages and that have not been associated with any disease suggests that viruses, especially those acquired in early childhood, may be essential for proper immune system development, and that particular host genotypes or immunologic constraints may cause normally benign viruses to induce disease states.

Glossary of terms

- Lytic phage cycle: State whereby fast exponential viral replication is achieved by using the bacterial host's DNA replication machinery independently of bacterial replication. This leads to the synthesis of multiple viral particles per cell and eventually lysis of the bacterium.
- Lysogenic phage: A state where linear (1:1) replication is achieved by a temperate phage through integration of its genome into the bacterial host chromosome (more rarely the phage exists as a plasmid). The integrated phage transcribes gene(s) that repress lytic action, and in some cases expresses genes that promote the fitness of the bacterial host.
- Prophage: temperate phage in a host-incorporated state.
- Transduction: transfer of DNA from one bacterial host to another by a phage; a common route for horizontal gene transfer.
- Superinfection immunity: the ability of a prophage to block superinfection of its bacterial host from another phage due to expression of genes that directly modify the phage receptor or proteins that block the receptor preventing attachment from other phage with similar specificity.
- CRISPR: A widespread genetic system in bacteria and archaea that consists of multiple copies of palindromic repeats flanking short spacers of viral or plasmid origin, which are believed to provide acquired resistance to foreign DNA.

- Kill-the-winner dynamics: A model for the population dynamics of phage–bacteria interactions that postulates that an increase in a host population (the winner) is followed by an increase in its corresponding phage predator, resulting in an increase in the rate at which the winner is killed.
- Pan-genome: The global gene repertoire of a bacterial species
- Coliphage: A bacteriophage that infects coliform bacteria, in particular *Escherichia coli*.
- Virus-like particle: Particles with physical characteristics resembling viruses but their infectivity has not been proven. Is usually used for isolation of viruses from environmental samples.
- Multiple displacement amplification: Exponential isothermal amplification of a DNA template using Phi29 DNA polymerase. The amplification is achieved by attachment of the polymerase to newly elongated fragments coupled with strong displacement activity upon extension.
- Virotype: The lowest level of taxonomical classification achievable by sequence comparison among viruses. In most cases is equivalent to species level classification but based on percent identity thresholds.
- Bacterial phylotype: The lowest level of taxonomical classification achievable by sequence comparison among bacterial species.
- Deep biosphere: The deepest oceanic regions where life is supported.

Chapter 2

Viruses in the fecal microbiota of monozygotic twins and their mothers

Chapter 2

Viruses in the fecal microbiota of monozygotic twins and their mothers

Alejandro Reyes[1], Matthew Haynes[2], Nicole Hanson[2], Florent E. Angly[2], Andrew C. Heath[3], Forest Rohwer[2], and Jeffrey I. Gordon[1]

[1]Center for Genome Sciences and [3]Department of Psychiatry, Washington University School of Medicine, St. Louis, MO 63108

[2]Department of Biology, San Diego State University, San Diego, CA 92182

This chapter corresponds to the full accepted version of a manuscript published in

Nature 466, 334–338 (15 July 2010) doi:10.1038/nature09199

Summary

The human gut contains viruses that infect the bacteria, archaea, and eukaryotes that comprise its microbiota. Viral diversity and lifecycles are poorly understood in this and other body habitats. Therefore, we sequenced the viromes (metagenomes) of virus-like particles isolated from fecal samples collected from adult female monozygotic twins and their mothers at three time points over a one-year period. These datasets were compared to datasets of sequenced bacterial 16S rRNA gene amplicons and total fecal community DNA. Co-twins and their mothers share a significantly greater degree of similarity in their fecal bacterial communities than do unrelated individuals. In contrast, viromes are unique to individuals regardless of their degree of genetic relatedness. Despite remarkable interpersonal variations in viromes and their encoded functions, intrapersonal diversity is very low, with >95% of virotypes retained over the period surveyed, and with viromes dominated by a few temperate phage that exhibit remarkable genetic stability. These results indicate that the fecal virome is a highly individualistic component of our gut ecosystems, even among people with identical genotypes, and that a predatory viral-microbial dynamic, manifest in a number of other characterized environmental ecosystems, is notably absent in the very distal gut.

Results

The diversity of viruses in the gut, and their role in the assembly, maintenance and adaptations of the microbiota and its pool of genes (microbiome) remain unclear. Viruses are major predators in our microbe-dominated planet. Most genetic diversity on our planet is viral. Viruses move DNA between their microbial hosts. Current estimates are that there are ≥10 VLPs per microbial cell, with the majority being dsDNA phages[1]. In many environments, the dominant ecological relationship between viruses and their microbial hosts is predatory and follows Lotka-Volterra (LV)/ Kill-the-Winner dynamics. LV is characterized by top-down control of microbial communities (i.e., microbial biomass is significantly below the carrying capacity of the habitat), rapid microbial and viral population shifts, and evidence of 'Red Queen' co-evolution (i.e., escape strategies in prey population are countered by predator adaption). One manifestation of this arms race is positive

selection on loci like bacterial O-antigens and CRISPR elements[2,3], and viral tail fibers[4]. In contrast to this predator-prey dynamic, there is another viral life cycle where temperate rather than lytic viruses are longer term contributors to microbial host phenotypes through provision of adaptive genes. This latter dynamic, which represents a more cooperative view of viral-host interactions, can change the metabolic capacities of free living bacteria, obligate intracellular mutualists[5], as well as the lifestyles of pathogens such as *Bacillus anthracis*[6]. In fact, many of the differences between closely related microbial strains arise from prophage insertions[7,8].

Recent studies of the human gut virome have focused on pathogen discovery[9], or have analyzed a few individuals without defining the microbial composition of their gastrointestinal tracts[10-12]. In this report, we characterize the fecal viromes of four pairs of adult female monozygotic (MZ) twins and their mothers. All were healthy without a history of gastrointestinal disease and none had received antibiotics in the 6-month period prior to sampling. Fecal samples were obtained at the beginning of the study, two months later (50±9 days) and 1 year later (364±10 days) (**Table S1**): all samples were frozen at -20^{o}C within 30 min after donation, placed at -80^{o}C within 36h, and subsequently maintained at -80^{o}C until use. VLPs were purified from 32 fecal specimens. Since the yield of VLP DNA from 2-5 g of fecal biomass averaged 500 ng, we performed random amplification of viral genomes contained in VLP preparations to obtain sufficient material for shotgun 454 FLX pyrosequencing. After *in silico* filtering for reads that were exact duplicates, or of low quality, or that had significant similarity to the human genome, our final data set contained 280,625,127 nt [33,043 ± 12,454 reads (mean±S.D.) per fecal VLP sample; average read length, 247±43 nt (mean±S.D.; **Table S1**)]. We verified the reproducibility of the viral DNA amplification/sequencing protocol by sequencing replicates from 5 samples (**Table S1** and *Methods***)**. One additional sample was subjected to deeper sequencing (293,654 reads; 70,157,333 nt). Bacterial taxa represented in fecal samples were characterized by pyrosequencing of amplicons generated by PCR from variable region 2 (V2) of their 16S rRNA genes (1,588–40,583 reads per sample). We had previously performed[13] shotgun sequencing of total fecal DNA isolated from the 12 frozen samples obtained at the first time point that were now used to prepare VLPs [average of 460,328±89,208 (mean±S.D.) reads per sampled microbiome; **Tables S2, S3**].

We generated a custom non-redundant viral database (NR_Viral_DB) to facilitate analysis of VLP-derived metagenomic datasets. To do so, the NCBI viral RefSeq database was downloaded and supplemented with complete viral and bacteriophage genomes from the European Bioinformatics Institute (EBI) database plus prophages identified in 396 sequenced microbial genomes (**Table S4**). To make the database non-redundant, all sequences were compared against each other and only those with <95% identity throughout their length were retained. These steps yielded 4,193 non-redundant sequences (96.2 Mb): 73.3% were eukaryotic viruses while 25.8% were phages and prophages; 76.9% of the latter were dsDNA viruses, mostly members of the order Caudovirales. The bacterial hosts of these known bacteriophages are principally members of the Proteobacteria (54%), Firmicutes (32%) and Actinobacteria (7%). The Bacteroidetes, which together with the Firmicutes constitute the dominant bacterial phyla in the adult human gut microbiota[14,15] were only represented by one viral genome and 20 prophages in this NR_Viral_DB.

A relaxed search against the NR_viral_DB (tblastx; e-value < 1e-3) showed that 81±6% (mean±S.D.) of the reads generated in this study did not match to any known viruses (**Fig. S1**). However, most of the identifiable viruses in the 32 VLP-derived viromes were prophage or phage generally classified as temperate (**Fig. 1**). The Podoviridae illustrate this point: while this family consists of both lytic and temperate members, in the fecal viral community its dominant representatives were temperate (e.g., coliphage P22-like). The predicted hosts of the identifiable phage and prophage were members of the Firmicutes and Bacteroidetes (**Fig. 1**); these phyla, and in particular the families Ruminococcaceae, Lachnospiraceae and Bacteroidaceae, comprised the most abundant bacterial taxa in the sampled fecal microbial communities, as defined by 16S rRNA gene analysis (**Fig. S2**).

Two different approaches were used to analyze 'within VLP sample' diversity (i.e., alpha diversity estimates). *First*, CD-Hit-est[16] was used to cluster reads with ≥90% sequence identity over 85% of their length. This allowed us to calculate a cluster-level Shannon index for each VLP sample (**Table S5**) and to define the expected number of virotypes per VLP preparation using procedures described in *Methods* (median of 346; range 52-2773). *Second*, PHACCS (Phage

Communities from Cross Contig Spectra) analysis (see *Methods*) indicated that each sample contained a median of 35 (range 10-984) predicted virotypes (**Table S6**; average Shannon Index of 3.32±0.71). Analysis of 16S rRNA datasets, where noise due to PCR and pyrosequencing artifacts and chimeras had been removed, indicated that there were ~800 species-level bacterial phylotypes in the first time point fecal communities[15]. Recent results obtained from deep shotgun sequencing of other fecal microbiomes suggest that the number may be <200 (ref. 14). Thus, the ratio of virotype to bacterial phylotype in the fecal microbiota of these healthy adults appears to approach 1.

Two complementary methods were employed to define the percentage of shared viral sequences between fecal samples obtained at different time points from the same individual and from different individuals (beta-diversity estimates). In the first method, we used CD-Hit-est to cluster pooled reads from all of the VLP viromes. Hellinger distances were subsequently calculated based on a matrix of the number of CD-Hit clusters vs VLP samples to determine relationships between samples. The second approach was based on contig assemblies generated from the pooled VLP viromes (Maxiphi[17]; see *Methods*). Both beta-diversity estimates showed that the same individual harbors very similar fecal viral communities over at least a one-year period: i.e., within an individual >95% of the viral genotypes were present and their relative abundances showed minimal variation (<8% permutation of the rank abundance). This is different than other comparably characterized ecosystems where almost daily changes in beta diversity occur[18].

While *intra*personal variation in viromes was minimal, *inter*personal variation was very high (**Fig. S3**). To analyze the clustering patterns generated by the distance matrices, 100 random sub-samplings of equal number of sequences per sample were used to generate the distance matrices, and a consensus UPGMA tree was subsequently produced. These trees showed that the main branching pattern clusters samples from the same individual, while there was no significant clustering of samples from the same family (**Fig. 2** and **S4**).

To further establish that temperate phages are dominant members of fecal VLP preparations, we searched for sequence identity between VLP viromes and 121 sequenced human gut microbial genomes. We identified 13 different bacterial genomes containing prophages present at

high abundance in at least one VLP sample (**Table S7**). For example, **Fig. S5** shows a region of the *Ruminococcus torques* ATCC 27756 genome that contains a predicted ~60Kb prophage. Sequence identity plots demonstrated that this prophage was prominently represented in a fecal VLP sample from the mother of family 2 [F2M.1 and F2M.1 (R)] where it comprised 18-20% of the reads at the first time point; at the two later time points (2 months and 1 year) the phage was still present in fecal VLP preparations albeit at lower abundance (0.8 and 1.4% respectively). This phage was not detectable in her twin daughters or other individuals in our study.

Given the low diversity of individual fecal viromes and the apparent representation of a few high abundance phage, we attempted to assemble partial or complete phage genomes using high stringency conditions (see *Methods*). This effort yielded 5,004 contigs ≥500 bp, 88 of which were 10Kb-71.4Kb (**Fig. S6**). All VLP-derived pyrosequencing reads from all 32 datasets were then aligned against these large contigs. The results revealed that virotypes represented by the large contigs were present mainly in only one individual where their abundance varied over time (**Table S8**). The nucleotide sequence conservation of these contigs (expressed as percent identity between reads from a given VLP sample and the contig) during the 1-year period was astonishingly high within an individual (99.74±0.26%). Only 8 of the 88 contigs were present in more than one individual from different families, with an average percent identity of 88.23±1.4% between subjects (**Table S9**).

We next defined functions encoded by the 32 sequenced fecal VLP-derived viromes by querying the KEGG and COG databases. The same procedure for functional assignments was used for reads in the 12 shotgun fecal microbiome datasets generated from the first time point sample from each individual in the 4 families. In addition, reference datasets of COG and KEGG assignments were assembled for all proteins encoded by all viral genomes present in the NR_Viral_DB, and in the 121 sequenced human gut microbial genomes (**Table S10**). Half of the bacterial genes in the sequenced genomes had assignable functions [51.6% (COG) and 54.4% (KEGG)], in contrast to 20% in the case of genes in the NR_Viral_DB [19.1% (COG) and 11.7% (KEGG)]. The percentage of fecal VLP-derived reads with significant hits to the COG and KEGG databases (BLAST cutoff, $E<10^{-5}$) was only 3.2±2.8% (mean±S.D.) and 1.7±1.9% (mean±S.D.), respectively, com-

pared to 36.0±6.9% (mean±S.D.) and 23.3±3.1% (mean±S.D.) for reads obtained from the 12 total fecal community DNA samples (**Fig. S7**). Comparison of VLP-derived viromes and the 12 microbiomes revealed significant functional differences: 'Transcription', 'Nucleotide Metabolism', and 'DNA Replication and Repair' were all over-represented in the viromes, while pathways for 'Carbohydrate Metabolism', 'Translation', 'Lipid Metabolism', 'Amino Acid Metabolism', 'Energy Metabolism', 'Membrane Transport', and 'Cellular Processes and Signaling' were significantly over-represented in the fecal microbiomes (**Fig. S8**). This result agrees with previous studies of aquatic ecosystems where viruses were significantly enriched for genes related to DNA, RNA synthesis and replication while the corresponding microbial communities were enriched for nitrogen and carbohydrate metabolism, and membrane transport[19].

Fig. 3 and **S9** provide a sample-by-sample view of the proportional representation of KEGG and COG categories in sequenced purified VLP-derived viromes and in fecal microbiomes. There is only modest *inter*personal variation in the distribution of KEGG and COG pathways in the microbiomes (R^2=0.993±0.005 for KEGG, 0.984±0.013 for COG). Moreover, the distribution of these functions is similar to their distribution in the 121 sequenced gut genomes (R^2=0.82 for KEGG, 0.95 for COG). In contrast, there is marked variation in these functional categories in the sequenced VLP-associated viromes (**Fig. 3** and **S9**), although further and deeper analysis of these differences was limited by the low percentage of viral reads that were classifiable.

The 88 VLP-derived large contigs (> 10Kb) encode 2,440 predicted proteins, 830 of which have significant similarity to viral and bacterial proteins present in the NR_Viral_DB and/or in the 121 gut genomes (blastx e-value < 1e-5). Metastats analysis identified a number of significant differences in the representation of KEGG and COG annotated functions associated with the large contigs compared to the NR_Viral_DB, including pathways related 'Glycan metabolism', 'Cell Wall Biosynthesis', and 'Transcription' (**Fig. S10**). To identify genes that may confer new and potentially advantageous functions to viruses present in the distal gut microbiota and/or to their microbial hosts, we searched this list of 2,440 proteins present in the 88 large contigs, eliminating those with homologs in the NR_Viral_DB, as well as all others whose putative functions suggested a viral origin (e.g., polymerases, capsid proteins, holins, etc). We were left with 23 proteins be-

longing to 16 protein families (**Table 1**). These proteins, seven of which use iron or sulfur in their reactions, are involved in a number of processes associated with the anaerobic gut microbiota. Homologs of these proteins present in sequenced human gut microbial genomes were subsequently retrieved, aligned and approximate maximum likelihood trees generated using FastTree[20]. The results (**Fig. 4**) indicate that some of the VLP virome-associated proteins are evolving in ways that are distinguishable from homologs present in known sequenced bacterial genomes, as is the case in the few environmental communities that have been subjected to comparable metagenomic studies[21].

Integrases are markers of temperate phage. We identified 10 ORFs with homology to integrases in the 88 contigs and 8,955 reads with significant similarity (blastp e-value <1e-4) to 785 different integrases among the 1,386,331 reads comprising the entire VLP dataset. The number of hits to different integrases per VLP sample was then used to construct a distance matrix that showed that (i) the diversity among identified integrases was significantly lower within VLP viromes purified from the same individual over time than between individuals, and (ii) there were no significant differences between individuals regardless of family relationships (**Fig. S11**).

As noted above, in most ecosystems where phage-host interactions have been studied in detail, lytic lifecycles and Red Queen dynamics appear to dominate[22]. Metagenomic studies of salterns and sludge ecosystems indicate that many of the most apparent genetic changes over time are at loci that prevent phage attachment (e.g., outer-membrane proteins and polysaccharides). Probably the clearest example of Red Queen dynamics between phage and their hosts are the CRISPR elements in sludge[2] and acid mine drainage systems[3]. Similarly, phage genomes often show evidence of changes in their tail fibers over time[4]. Given this paradigm, it is striking that we found essentially no evidence of this type of behavior in fecal phages (see *Supplementary Discussion* for analysis of CRISPRs). In contrast, we found high abundances of dominant phage, present in the same individual for extended periods of time, with no significant divergence or mutations in their genomes. The presence of integrases in the assembled VLP contigs is also consistent with the notion that they represent prominent temperate phage in the fecal microbiota.

One potential scenario is that phage production occurs via induction of prophages caused by energy limitation in the feces; at this point, fecal microbial hosts are effectively at a dead end for their associated phage, and the viruses may gain an advantage for transmission by 'going it alone'. Experimental evidence for this scenario is provided by gnotobiotic mice co-colonized for two weeks with *Marvinbryantia formatexigens*, a human gut acetogen that contains three predicted prophages in its genome, and *Bacteroides thetaiotaomicron,* a human gut-derived saccharolytic bacterium that harbors two predicted prophages. Normalized RNA-Seq counts, generated from cecal contents and fecal samples harvested from the animals at the time of their sacrifice (n=3 mice)[23], revealed that one of the three prophages in *M. formatexigens* was completely activated (all ORFs transcribed) in all fecal samples and in a subset of cecal samples (**Fig. 5)**. In the case of the remaining two *M. formatexigens* prophages, only a few genes were expressed, including a pair of adjacent ORFs encoding a HicA family toxin (BRYFOR7601) and a HicB family anti-toxin (BRYFOR7602) in one of the prophages, and a pair of genes that specify a RelE family toxin (BFYFOR9696) and a PHD family anti-toxin (BRYFOR9697) in the other prophage; these two gene pairs were constitutively expressed in all fecal samples, in all cecal samples, and during *in vitro* growth in defined medium containing a variety of carbon sources (**Fig. 5**). Co-expression of toxin-antitoxin genes is known to maintain stable integration of phage DNA in bacterial chromosomes[24,25]. Importantly, the one prophage that was fully activated (prophage 2 in **Fig. 5**) does not have a detectable toxin/anti-toxin gene pair. Only two small (2-3 gene) clusters were expressed in the two *B. thetaiotaomicron* prophages *in vivo*; all of these clusters encode predicted toxin or anti-toxin genes (**Fig. S12**). Together, these findings illustrate how a prophage may be liberated from its host cell when that cell is present in a fecal community.

Human microbiome projects have been initiated throughout the world in order to define the interrelationships between human physiologic status, and/or disease states and microbial community structure and function. Our results suggest that a potentially important dimension should be incorporated to these metagenomic studies; namely, one that targets VLPs recovered from various body habitat-associated communities (**Fig. S13** and *Supplementary Discussion*). Comparisons of the functions embedded in both dominant and sub-dominant phages present in these communities

may provide informative molecular signatures (biomarkers) of the microbiota and its human host, of microbial community responses to impending or fully manifest disease states, and of the extent to which community health or pathology endures after apparent recovery of the human host from a disease or therapeutic intervention. In addition, gnotobiotic mice harboring defined collections of human gut symbionts inoculated with VLP-derived viromes should provide informative models for further dissection of various aspects of the interactions of phage and their microbial hosts in different regions of the gut, including an assessment of whether LV dynamics operate in more proximal regions of the intestine where energy generated from dietary components may be more available.

Methods Summary

Sample collection.

The Missouri Adolescent Female Twin Study (MOAFTS[26]) is composed of female twin pairs born in the state of Missouri between 1975-1986, and their mothers. Procedures for obtaining informed consent and sample collection were approved by the Washington University Human Studies Committee.

DNA extraction and 454 Pyrosequencing.

Aliquots of frozen fecal samples (2-5 g) were processed for isolation of VLPs by serial filtration, followed by cesium chloride gradient ultracentrifugation[27]. VLPs were lysed in a solution containing Proteinase K and 10% SDS. DNA was extracted with 10% cetyltrimethylammonium bromide/0.7M NaCl and amplified using the illustra™ GenomiPhi™ V2 kit (GE Healthcare). The resulting DNA was used for multiplex shotgun 454 FLX pyrosequencing. For further details about VLP purification, extraction of VLP DNA, assembly of pyrosequencer reads, and data analysis see *Methods*.

References

1 Breitbart, M., Rohwer, F., & Abedon, S.T., Phage Ecology and bacterial pathogenesis in Phages, their role in bacterial pathogenesis and biotechnology, edited by M.K. Waldro, D.I. Friedman, & S.L Adhya (ASM Press, Washington DC, 2005), pp. 66-91.

2 Kunin, V. et al., A bacterial metapopulation adapts locally to phage predation despite global dispersal. Genome Res 18, 293-297 (2008).

3 Tyson, G.W. & Banfield, J.F., Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol 10, 200-207 (2008).

4 Angly, F. et al., Genomic analysis of multiple Roseophage SIO1 strains. Environ Microbiol 11, 2863-2873 (2009).

5 Oliver, K.M., Degnan, P.H., Hunter, M.S., & Moran, N.A., Bacteriophages encode factors required for protection in a symbiotic mutualism. Science 325, 992-994 (2009).

6 Schuch, R. & Fischetti, V.A., The secret life of the anthrax agent Bacillus anthracis: bacteriophage-mediated ecological adaptations. PLoS One 4, e6532 (2009).

7 Casjens, S., Prophages and bacterial genomics: what have we learned so far? Mol Microbiol 49, 277-300 (2003).

8 Tettelin, H. et al., Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 102, 13950-13955 (2005).

9 Finkbeiner, S.R. et al., Metagenomic analysis of human diarrhea: viral detection and discovery. PLoS Pathog 4, e1000011 (2008).

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

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

12 Breitbart, M. et al., Viral diversity and dynamics in an infant gut. Res Microbiol 159, 367-373 (2008).

13 Turnbaugh, P.J. et al., A core gut microbiome in obese and lean twins. Nature 457, 480-484 (2009).

14 Qin, J. et al., A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59-65 (2010).

15 Turnbaugh, P.J. *et al*., Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. Proc Natl Acad Sci U S A 107, 7503-7508 (2010).

16 Li, W. & Godzik, A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658-1659 (2006).

17 Angly, F.E. et al., The marine viromes of four oceanic regions. PLoS Biol 4, e368 (2006).

18 Rodriguez-Brito, B. et al., Viral and microbial community dynamics in four aquatic environments. ISME J advanced online publication, 11 Feb 2010 (DOI:10.1038/ismej.2010.1).

19 White, J.R., Nagarajan, N., & Pop, M., Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol 5, e1000352 (2009).

20 Price, M.N., Dehal, P.S., & Arkin, A.P., FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26, 1641-1650 (2009).

21 Sharon, I. et al., Photosystem I gene cassettes are present in marine virus genomes. Nature 461, 258-262 (2009).

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

23 Rey, F. E. et al. Dissecting the in vivo metabolic potential of two human gut acetogens. J Biol Chem 285, 22082-22090 (2010).

24 Magnuson, R.D., Hypothetical functions of toxin-antitoxin systems. J Bacteriol 189, 6089-6092 (2007).

25 DeShazer, D., Genomic diversity of Burkholderia pseudomallei clinical isolates: subtractive hybridization reveals a Burkholderia mallei-specific prophage in B. pseudomallei 1026b. J Bacteriol 186, 3938-3950 (2004).

26 Heath, A.C. et al., Ascertainment of a mid-western US female adolescent twin cohort for alcohol studies: assessment of sample representativeness using birth record data. Twin Res 5, 107-112 (2002).

27 Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L., & Rohwer, F., Laboratory procedures to generate viral metagenomes. Nat Protoc 4, 470-483 (2009).

28 Edgar, R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792-1797 (2004).

29 Becker, A., Kuster, H., Niehaus, K., & Puhler, A., Extension of the Rhizobium meliloti succinoglycan biosynthesis gene cluster: identification of the exsA gene encoding an ABC transporter protein, and the exsB gene which probably codes for a regulator of succinoglycan biosynthesis. Mol Gen Genet 249, 487-497 (1995).

30 Gon, S., Faulkner, M.J., & Beckwith, J., In vivo requirement for glutaredoxins and thioredoxins in the reduction of the ribonucleotide reductases of Escherichia coli. Antioxid Redox Signal 8, 735-742 (2006).

31 Padovani, D., Thomas, F., Trautwein, A.X., Mulliez, E., & Fontecave, M., Activation of class III ribonucleotide reductase from E. coli. The electron transfer from the iron-sulfur center to S-adenosylmethionine. Biochemistry 40, 6713-6719 (2001).

32 Garriga, X. et al., nrdD and nrdG genes are essential for strict anaerobic growth of Escherichia coli. Biochem Biophys Res Commun 229, 189-192 (1996).

33 Tabor, C.W. & Tabor, H., 1,4-Diaminobutane (putrescine), spermidine, and spermine. Annu Rev Biochem 45, 285-306 (1976).

Acknowledgements

We thank Sabrina Wagoner and Jill Manchester for superb technical assistance, Jeremiah Faith for help developing Phage_omics and together with Federico Rey for microbial RNA-Seq datasets, Peter Turnbaugh for assistance with fecal metagenomic studies and Beltran Rodriguez-Mueller and Dana Willner for valuable discussions. This work was supported in part by grants from the NIH (American Recovery and Reinvestment Act supplemental funding of DK78669), the Crohn's and Colitis Foundation of America, and the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation. A.R is the recipient of an International Fulbright Science and Technology Program award.

Author contributions

A.R. and J.I.G designed the experiments, A.H. recruited the patients, A.R, M.H, and N.H. generated the data, A.R., F.A., F.R, and J.I.G. interpreted the results, A.R., F.R., and J.I.G. wrote the paper.

Author information

Virome datasets are accessible in the NCBI Short Read Archive under accession number SRA012183. 16S rRNA and fecal microbiome datasets are available in GenBank under genome project ID 32089 and SRA002775. RNA-Seq data are deposited in Gene Expression Omnibus (GSE21906; see Methods for further details). The authors declare that they have no competing interests.

Figure Legends

Figure 1. **Classification of viruses present in VLP preparations generated from fecal samples collected from four families of MZ twins and their mothers.** Prophage are classified based on their bacterial host taxonomy. Prominent bacterial phyla are represented by different colors (Proteobacteria, blue; Firmicutes green; Bacteroidetes, red; Actinobacteria, black). Class-level taxa within these phyla are noted. Phage and eukaryotic viruses are sorted according to taxonomy. Nomenclature used for VLP preparations from fecal biospecimens: F, family; T1, co-twin 1; T2, co-twin 2; M, mother of co-twins. Time points (1-3), and technical replicates (R) produced from a given sample are noted. The color bar at the bottom of the figure provides a reference key for the percent coverage of a viral genome in the NR_Viral_DB by reads from given VLP virome dataset (data normalized using 14,000 randomly selected reads/dataset).

Figure 2. Beta-diversity analysis: clustering of fecal VLP-associated viromes and bacterial 16S rRNA data. Unrooted, jack-knifed (100 iterations) consensus UPGMA trees obtained from Hellinger-based distance matrices are shown for bacterial 16S rRNA data (**a**) and VLP-derived viromes (**b**). The color key provides information about the family (F), and family member. Bars represent Hellinger distances.

Figure 3. A sample-by-sample view of the proportional representation of KEGG second level pathways in sequenced VLP-associated viromes and gut microbiomes. Known or predicted proteins encoded by viruses in the NR_Viral_DB, fecal VLP-derived viromes, 121 sequenced reference human gut-associated microbial genomes, and fecal microbiomes are shown. See Fig. 1 for sample nomenclature.

Figure 4. Representative phylogenetic trees of bacterial proteins present in large contigs assembled from VLP-viromes with no homologs in the NR_Viral_DB. Multiple alignment of the indicated viral protein (highlighted in red) with all proteins from 121 human gut microbial genomes that harbored the same domain or motif was performed using Muscle[28]. Approximate

maximum likelihood trees were generated using FastTree[20]. Bars represent the number of amino acid substitutions per position.

Figure 5. Gnotobiotic mice reveal *in vivo* activation of the transcriptome of a *Marvinbryantia formatexigens* prophage. Shown are the three predicted prophages present in *M. formatexigens* and the levels of expression of their ORFs in cecal and fecal micobial communities harvested from gnotobiotic mice co-colonized with *Bacteroides thetaiotaomicron*. Expression levels for genes in each prophage genome are from normalized RNA-Seq read count data (see color key; normalization based on sequencing effort and length of each predicted ORF). Active expression is defined as a normalized read count >100. R, technical replicate shows the reproducibility of the method for performing RNA-Seq analysis. RNA-Seq data are also presented for each prophage genome in *M. formatexigens* during mid-log phase growth in defined medium containing different carbon sources (NAG, N-acetylglucosamine). Red arrows indicate the position of toxin/anti-toxin gene pairs. Green arrows denote genes with hypothetical functions that are expressed in more than 50% of the conditions tested. ORF designations for the first and last genes in the predicted genomes of each prophage are provided.

Figures

Figure 1.

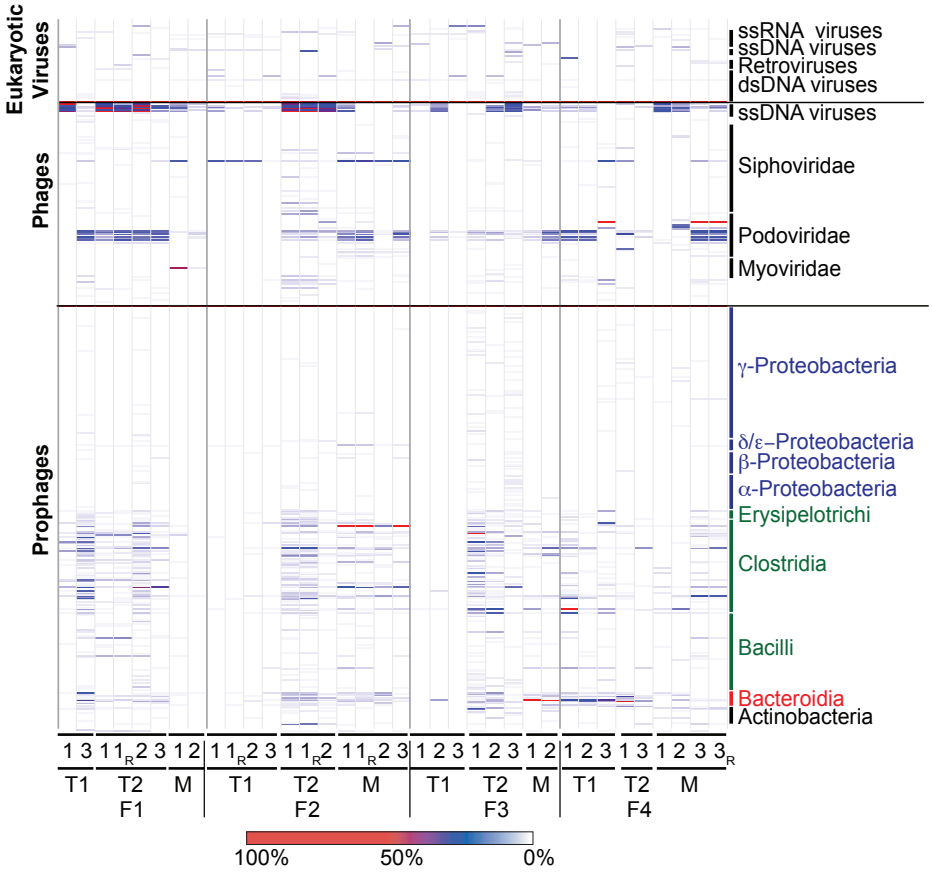


Figure 2.

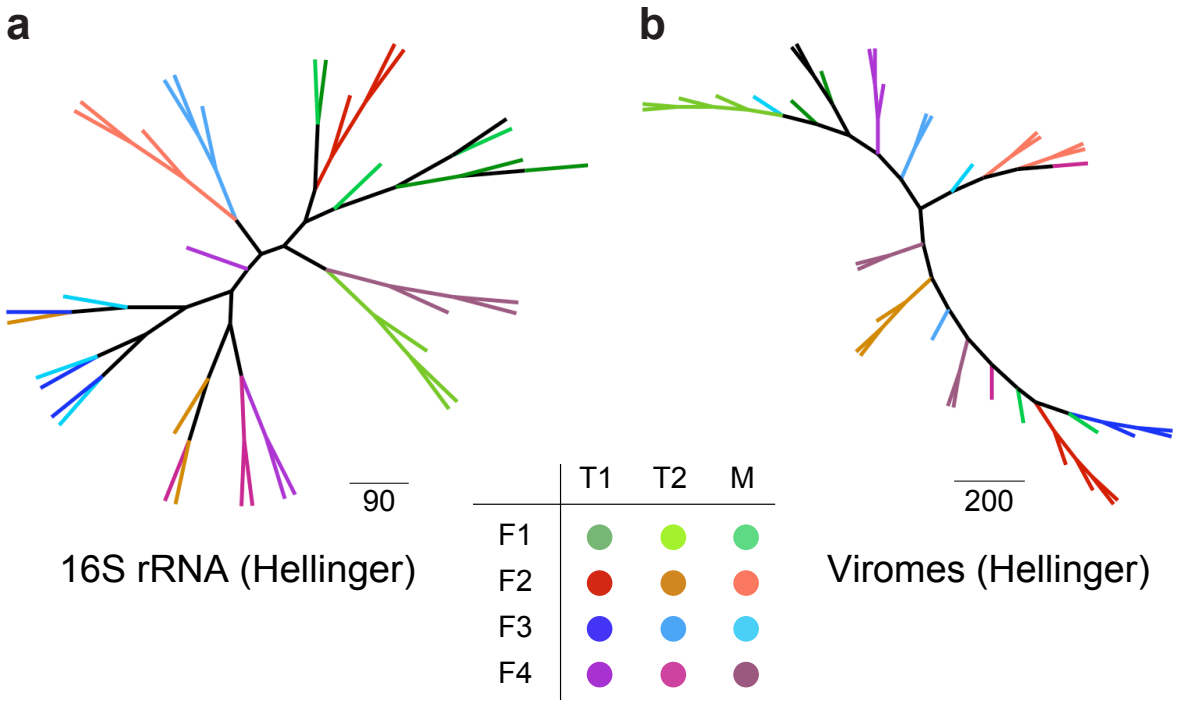

a
b
90
200
16S rRNA (Hellinger)
Viromes (Hellinger)
T1
T2
M
F1
F2
F3
F4

Figure 3.

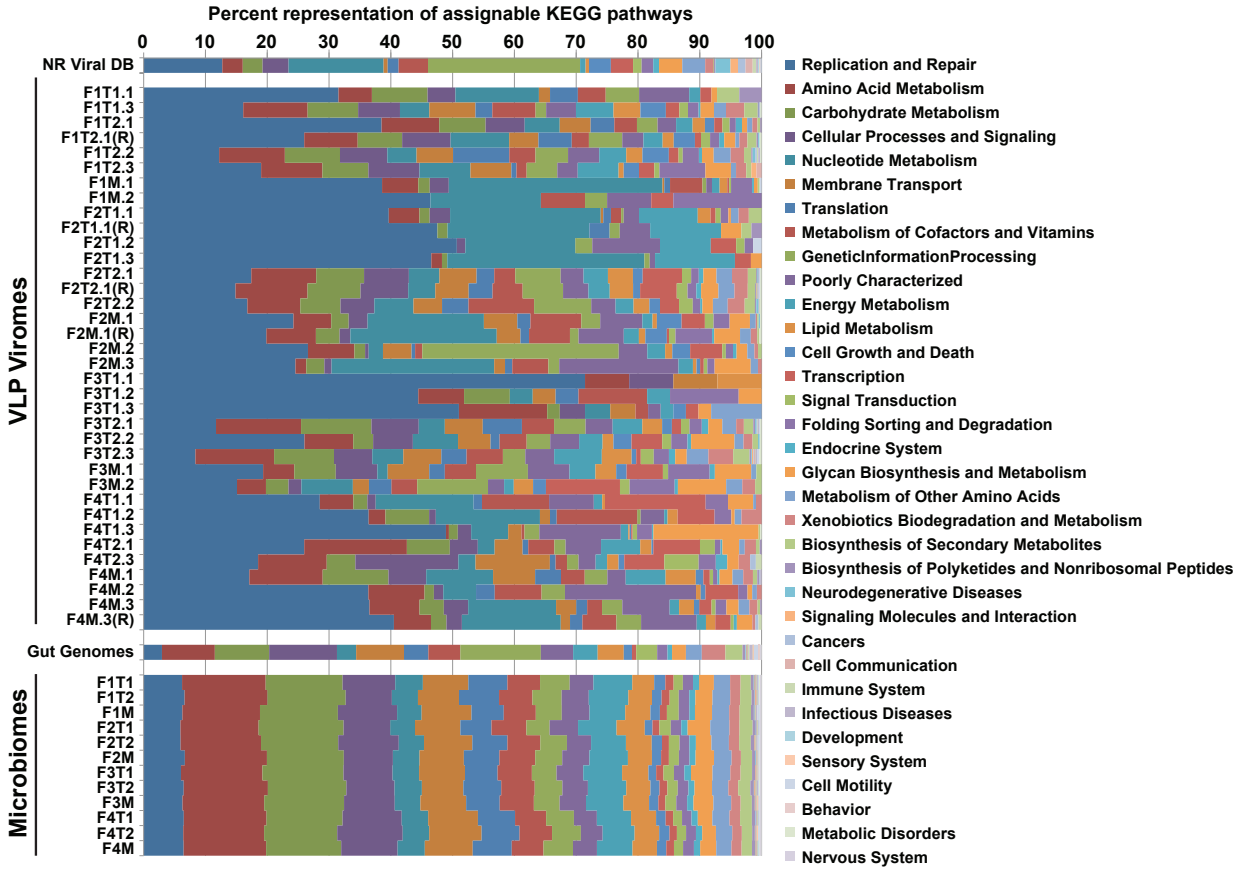

Percent representation of assignable KEGG pathways
0
10
20
30
40
50
60
70
80
90
100
NR Viral DB
VLP Viromes
F1T1.1
F1T1.3
F1T2.1
F1T2.1(R)
F1T2.2
F1T2.3
F1M.1
F1M.2
F2T1.1
F2T1.1(R)
F2T1.2
F2T1.3
F2T2.1
F2T2.1(R)
F2T2.2
F2M.1
F2M.1(R)
F2M.2
F2M.3
F3T1.1
F3T1.2
F3T1.3
F3T2.1
F3T2.2
F3T2.3
F3M.1
F3M.2
F4T1.1
F4T1.2
F4T1.3
F4T2.1
F4T2.3
F4M.1
F4M.2
F4M.3
F4M.3(R)
Gut Genomes
Microbiomes
F1T1
F1T2
F1M
F2T1
F2T2
F2M
F3T1
F3T2
F3M
F4T1
F4T2
F4M
Replication and Repair
Amino Acid Metabolism
Carbohydrate Metabolism
Cellular Processes and Signaling
Nucleotide Metabolism
Membrane Transport
Translation
Metabolism of Cofactors and Vitamins
GeneticInformationProcessing
Poorly Characterized
Energy Metabolism
Lipid Metabolism
Cell Growth and Death
Transcription
Signal Transduction
Folding Sorting and Degradation
Endocrine System
Glycan Biosynthesis and Metabolism
Metabolism of Other Amino Acids
Xenobiotics Biodegradation and Metabolism
Biosynthesis of Secondary Metabolites
Biosynthesis of Polyketides and Nonribosomal Peptides
Neurodegenerative Diseases
Signaling Molecules and Interaction
Cancers
Cell Communication
Immune System
Infectious Diseases
Development
Sensory System
Cell Motility
Behavior
Metabolic Disorders
Nervous System

Figure 4.

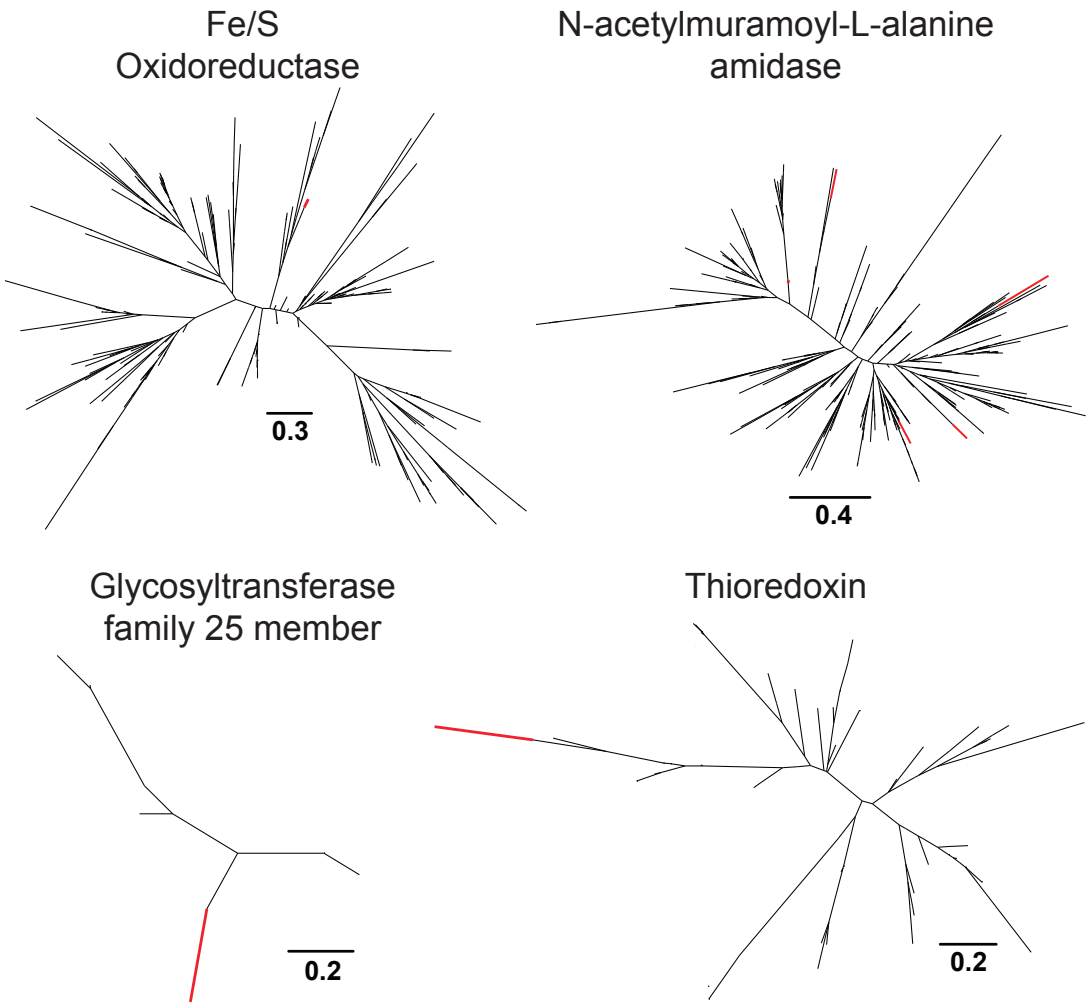
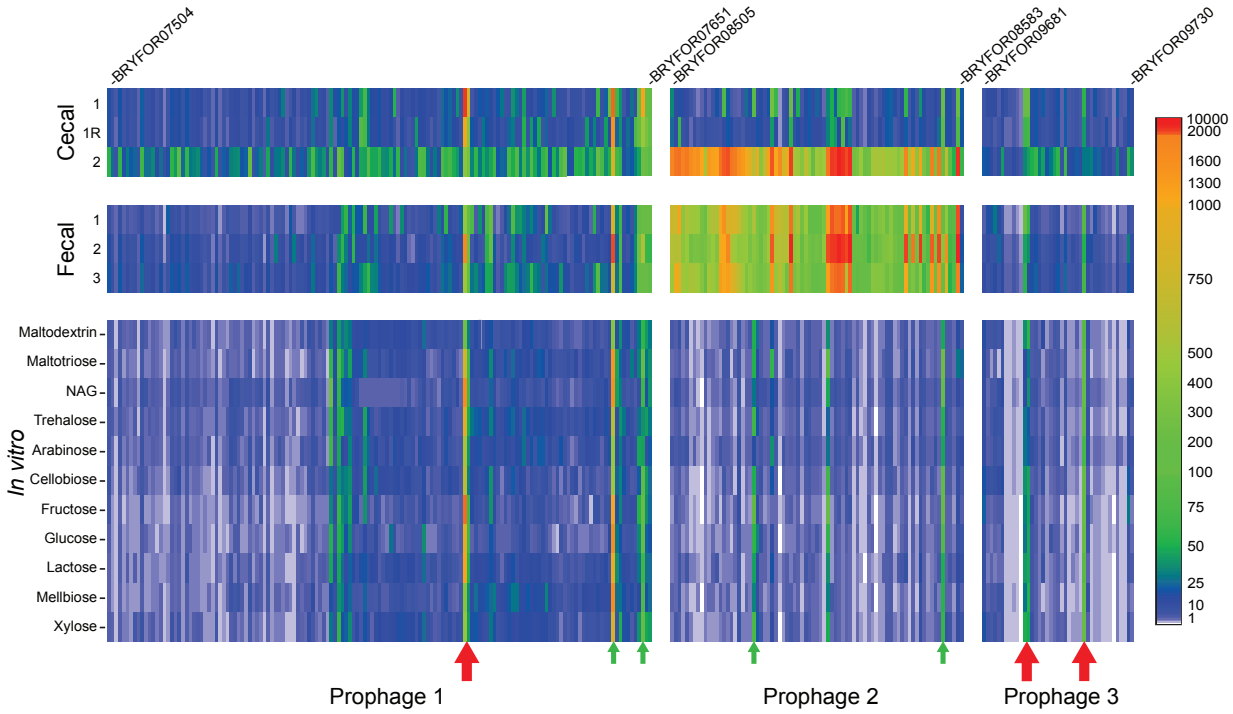

Fe/S
Oxidoreductase
0.3
N-acetylmuramoyl-L-alanine
amidase
0.4
Glycosyltransferase
family 25 member
0.2
Thioredoxin
0.2

Figure 5.


-BRYFOR07504
-BRYFOR07651
-BRYFOR08505
-BRYFOR08583
-BRYFOR09681
-BRYFOR09730
Cecal
1
1R
2
Fecal
1
2
3
In vitro
Maltodextrin
Maltotriose
NAG
Trehalose
Arabinose
Cellobiose
Fructose
Glucose
Lactose
Mellbiose
Xylose
10000
2000
1600
1300
1000
750
500
400
300
200
100
75
50
25
10
1
Prophage 1
Prophage 2
Prophage 3

Table

Table 1. Proteins encoded by 88 large viral contigs assembled from fecal VLP viromes that have no homologs in the NR_Viral_DB and whose functions are involved in processes associated with the anaerobic gut microbiota

No of ORFs	Name	Description
5	N-acetylmuramoyl-L-alanine amidase	Cleaves the amide bond between N-acetyl-muramoyl and L-amino acids in bacterial cell walls [EC 3.5.1.28]
3	Thymidylate synthase	Involved in folate biosynthesis pathway [EC 2.1.1.148]
2	6-pyruvoyl tetrahydropterin synthase	Involved in folate biosynthesis pathway [EC 4.2.3.12]
1	Anaerobic nitric oxide reductase transcription regulator	NifA
1	Fe-S Oxidoreductase	Radical SAM domain-containing protein
1	Anaerobic ribonucleoside-triphosphate reductase activating protein	Transcription factor involved in the anaerobic de-novo synthesis of nucleotides
1	ExsB	Transcription factor involved in the regulation of succinoglycan levels
1	Phosphoadenosine phosphosulfate reductase family	Involved in synthesis of cysteine [EC 1.8.4.8]
1	Ferritin Dps family protein	Participates in oxygen defense
1	Glycosyltransferase family 25	Involved in peptidoglycan biosynthesis
1	Glycosyltransferases family 2	Involved in peptidoglycan biosynthesis
1	Methylglyoxal synthase	Involved in pyruvate metabolism [EC 4.2.3.3]
1	Iron/manganese superoxide dismutases, C-terminal domain	Involved in oxygen defense [EC 1.15.1.1]
1	Thioredoxin	Involved in oxygen defense, can act as electron donor for reductases
1	S-adenosylmethionine decarboxylase	Involved in biosynthesis of polyamines
1	Cysteine desulfurase	NifS, involved in Fe-S cluster biosynthesis.

Legend – This list of proteins includes (i) two transcriptional regulators (a homolog of ExsB involved in regulation of succinoglycan levels[29]; and an anaerobic nitric oxide reductase regulator belonging to the sigma 54 family); (ii) an anaerobic ribonucleoside triphosphate reductase activat-

ing protein that uses S-adenosylmethionine (SAM), an iron-sulfur cluster, and a reductant for the *de-novo* anaerobic synthesis of nucleotides[30-32]; (iii) other SAM-related proteins (Fe-S oxidoreductase, and a SAM-decarboxylase that uses SAM for synthesis of spermidine and spermine, which in turn stimulate RNA polymerases and stabilize the DNA helix respectively[33]), (iv) three oxidative stress-related proteins (an iron/manganese superoxide dismutase, thioredoxin, and a ferritin Dps family protein); (v) a methylglyoxal synthase homolog involved in pyruvate metabolism, (vi) a thymidylate synthase and a 6-pyruvoyl tetrahydropterin synthase, both involved in folate metabolism; (vii) a member of the phosphoadenosine phosphosulfate reductase family that participates in the cysteine biosynthesis and uses thioredoxin as electron donor; (viii) cysteine desulfurase (*nifS*), which plays an important role in Fe-S cluster biosynthesis by catalyzing removal of sulfur from cysteine to produce alanine; and (ix) a group of proteins involved in peptidoglycan synthesis [a member of CAZy Glycosyltransferase family 2 (GT2), a GT25 member, and five N-acetylmuramoyl-L-alanine amidases; note that acquisition of this last group of enzymes is intriguing in light of evidence that some phages can subvert normal bacterial pathways for surface glycan biosynthesis[6]].

Methods

Purification of VLPs.

Viral purification was performed with minor modifications of the procedure described in an earlier publication[27]. In brief, a 2-5 g aliquot of each pulverized fecal sample was re-suspended in 25 mL SM buffer [100 mM NaCl, 8 mM $MgSO_4$, 50 mM Tris (pH 7.5) and 0.002% gelatin (wt/vol)]. Following centrifugation (2,500 x g for 10 min at room temperature), the resulting supernatant was removed and passed sequentially through 0.45µm and 0.22µm Whatman filters to remove residual cells. The filtrate was then adjusted with CsCl to a density of 1.12g/mL and deposited on top of a 3 mL step gradient prepared using 1mL CsCl solutions with densities of 1.7 g/mL SM buffer, 1.5 g/mL, and 1.35 g/mL. Samples were centrifuged for 2h at 60,000 x g (4ºC) in a SW41 swinging bucket rotor (Beckman). The 1.5 g/mL layer was recovered since material in this density range is known to be enriched for bacteriophages[27]. At each step of the purification procedure, an aliquot of the sample was viewed under an epifluorescence microscope after viral particles had been stained with SYBR-gold; this allowed us to document the presence of VLPs and note whether a decrease in the representation of bacterial and eukaryotic cellular elements had occurred.

Extraction of viral DNA.

After the 1.5 g/mL layer was collected from the step gradient, chloroform was added (0.2 volumes) and the solution was centrifuged for 5 min at 2,500 x g. The aqueous phase was treated with DNAse (Sigma Aldrich; final concentration 2.5U/mL) to remove residual host and bacterial DNA. To extract the virions, 0.1 volume of 2M Tris HCl/ 0.2 M EDTA, 1 volume of formamide, and 100µL of a 0.5M EDTA solution were added per 10mL of sample, and the resulting mixture was incubated at room temperature for 30 min. The sample was subsequently washed with 2 volumes of ethanol and pelleted by centrifugation for 20 min at 8,000 x g at 4ºC. The pellet was washed twice with 70% ethanol and re-suspended in 567µL of TE buffer, followed by 30µL of 10% SDS and 3µL of a 20 mg/mL solution of Proteinase K (Fisher Scientific; cat no. AC61182-0500). The mixture was

incubated for 1h at 55ºC, and 100 µL of 5M NaCl and 80µL of a solution of 10% cetyltrimethylammonium bromide/0.7M NaCl were subsequently introduced. After a 10 min incubation at 65ºC, an equal volume of chloroform was added and the mixture was centrifuged (5 min at 8,000 x g; room temperature). The resulting supernatant was transferred to a new tube and an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1) was added, followed by centrifugation (5 min at 8,000 x g; room temperature). The supernatant was recovered and an equal volume of chloroform was introduced. Following centrifugation, the supernatant was collected and 0.7 volumes of isopropanol used to precipitate the DNA. After another centrifugation step (15 min at 13,000 x g at 4ºC), the material was washed (500µL of cold 70% ethanol), air dried, and resuspended in 50µL of TE. An aliquot of the purified DNA was used as a template in polymerase chain reactions that contained universal primers directed at bacterial 16S rRNA and eukaryotic 18S rRNA genes; this assay was used to confirm the absence of detectable contaminating non-viral DNA.

Amplification of VLP-associated DNA.

Shotgun 454 pyrosequencing requires 3–5µg of DNA for library preparation. The typical yield from our fecal VLP DNA isolation procedure was 500ng/sample. Therefore, WGA (Whole Genome Amplification) was performed using reagents and protocols in the illustra™ GenomiPhi™ V2 kit (GE Healthcare) to generate sufficient material for library construction. Ten to 50ng of purified VLP DNA were mixed with 9µL of 'Sample Buffer' from the kit and heat denatured at 95ºC for 3 min. Nine microliters of 'Reaction Buffer' and 1µL of 'Enzyme Mix' were then added and the solution incubated for 90 min at 30ºC. Three separate WGA reactions were performed for each viral DNA preparation to minimize potential bias in amplification. The amplified products from each sample were subsequently pooled and purified (QIAGEN DNeasy kit).

To test for bias in the amplification and sequencing of VLP-DNA preparations that had been subjected to WGA, we analyzed the VLP sample from individual F3T1.1 where the yield of DNA was sufficient to perform shotgun pyrosequencing with un-amplified as well as with amplified subsamples. We pooled 16,567 reads derived from WGA DNA, and 18,845 reads from the

unamplified aliquot and clustered them using the procedures used for alpha diversity calculations (see *CD-Hit Clustering* below). We found that 98.4% of the sequences from the un-amplified DNA were also present in the WGA reads while 91.96% of the WGA sequences were represented in the un-amplified sample dataset. This difference could be due to sequencing of amplified low abundance viral DNAs that were not sequenced in the unamplified sample. WGA is also known to preferentially amplify small ssDNA viruses[18].

Multiplex shotgun pyrosequencing of VLP viromes.

DNAs, purified from each of 12 VLP preparations, were labeled with a different MID (Multiplex Identifiers; Roche). Equivalent amounts of the barcoded samples were then pooled prior to a run of 454 FLX pyrosequencing. Shotgun reads were filtered by removing (i) all duplicates (defined as sequences whose initial 20 nucleotides are identical and that share an overall identity of >97% throughout the length of the shortest read; duplicates are a known pyrosequencing artifact[34]), (ii) reads with degenerate bases ('N's), and (iii) sequences with significant similarity to human reference genomes (BLASTN with e-value < 1E-5) in order to ensure the de-identification of samples.

Bacterial 16S rRNA gene amplification and sequencing.

An aliquot (500mg) of each frozen pulverized fecal sample was re-suspended in a solution containing 500µL of extraction buffer [200mM Tris (pH 8.0), 200mM NaCl, 20mM EDTA], 210µL of 20% SDS, 500µL of phenol:chloroform:isoamyl alcohol (25:24:1) and 500µL of a slurry of 0.1-mm diameter zirconia/silica beads (BioSpec Products). Cells were mechanically disrupted using a bead beater (BioSpec Products) set on high for 2 min at room temperature, followed by extraction with phenol:chloroform:isoamyl alcohol and precipitation with isopropanol. DNA obtained from three separate aliquots of each fecal sample were pooled and used for amplification of bacterial 16S rRNA genes.

Approximately 330bp amplicons, spanning variable region 2 (V2) of bacterial 16 rRNA genes were generated by using PCR and (i) modified primer 8F (5´- GCCTTGCCAGCCCGCT-

CAG**TC***AGAGTTTGATCCTGGCTCAG*-3') which consisted of 454 primer B (underlined) and the universal bacterial primer 8F (italics) and (ii) modified primer 338R (5' GCCTCCCTCGC-GCCATCAGNNNNNNNNNNNN**CA***TGCTGCCTCCCGTAGGAGT* 3') which contained 454 primer A (underlined), a sample specific, error correcting 12-mer barcode[35] (N's), and the bacterial primer 338R (italics).

Four replicate polymerase chain reactions were performed for each pooled fecal DNA sample. Each 20μL reaction contained 100ng of gel purified DNA (Qiaquick, Qiagen), 8 μL 2.5X HotMaster PCR Mix (Eppendorf), and 0.3 μM of each primer. The PCR program consisted of initial denaturation at 95ºC for 2 min followed by 30 cycles of denaturation (95ºC for 20 sec), annealing (52ºC for 20 sec) and amplification (65ºC for 1 min). Replicate PCRs were subsequently pooled and purified using Ampure magnetic purification beads (Agencourt). DNA was quantified using Picogreen (Invitrogen) and an equimolar amount of each sample was used for multiplex 454 FLX amplicon pyrosequencing.

Bacterial 16S rRNA data processing and analysis.

16S rRNA reads were analyzed using QIIME[36]: fasta, quality files and a mapping file indicating the barcode sequence corresponding to each sample were used as inputs. The QIIME pipeline takes this input information and split reads by samples according to the barcode, and classifies reads into OTUs based on sequence similarity. It also performs taxonomical classification using the RDP-classifier[37], builds a *de-novo* taxonomic tree of the sequences based on sequence similarity, and creates a sample x OTUs table that can be used, together with the tree, for calculating alpha and beta diversity.

Custom Non Redundant viral database (NR_Viral_DB).

All complete viral and bacteriophage genomes available in the European Bioinformatics Institute (EBI) database were downloaded, as were all complete prophage genomes present in the SEED database and all entries for the taxid 10239 (Viruses) in RefSeq between 1Kb – 500Kb. The database

was complemented with prophage sequences identified from a survey of 396 sequenced microbial genomes (**Table S4**) using the software tool PhageFinder[38]. To make the database non-redundant, all sequences were compared against each other and only those with <95% identity throughout their length were retained.

CD-Hit Clustering.

CD-Hit-est is a software tool designed for clustering nucleotide sequences by similarity[16]. We used CD-Hit-est to cluster the pooled reads obtained from all viral samples. Hierarchical clustering was performed based on continuous reduction of the required percentage overlap between reads (from 99% to 85%) while maintaining a sequence identity of ≥90%. A sample x CD-Hit cluster table was subsequently generated, analogous to an OTU table. The table was analyzed using QIIME to generate alpha diversity estimates (as Shannon indices) as well as beta diversity matrices based on Hellinger distances.

Viral alpha and beta diversity.

These estimates were based on a pipeline composed of several software programs: GAAS[39], Circonspect[17], PHACCS[40], and MaxiPhi[17]. Circonspect (http://sourceforge.net/projects/circonspect/) was used to form cross-contig spectra of 3x coverage. To ensure stringent assembly, contigs in Circonspect were determined by Minimo (available in the AMOS package at http://sourceforge.net/projects/amos/) instead of TIGR Assembler. Contig assembly parameters were 98% similar sequences overlapping by at least 35 bp. Viral average genome length was then estimated using GAAS (http://sourceforge.net/projects/gaas/) (tblastx against complete NCBI RefSeq viral genomes, minimum 30% similarity, minimum 70% relative length).

For each pair of samples to compare, the input for MaxiPhi was their cross-contig spectrum and average genome length. Building on PHACCS (http://sourceforge.net/projects/phaccs/), MaxiPhi ran a Monte-Carlo simulation to determine how many virotypes samples had in common (percent shared), and how many of the most abundant ones changed their abundance rank (per-

cent permuted). Using VLP reads from each individual sample assembled against themselves as internal controls, the best average genome length for each beta-diversity computation was found as the length within 20% of the input value that produced the percent shared and percent permuted closest to 100 and 0% respectively for both controls.

The entire viral diversity analysis was done at several levels: between time points for each individual, between twins for each family, between twin and mother for each family, and between all families. Input viral metagenomes were pooled as necessary, e.g., for the co-twin comparison, sequences from all 3 time points from each twin were merged.

PHACCS calculates expected number of virotypes by estimating Shannon indexes from contig spectra. To compare the results with an independent method, Shannon indexes derived from CD-Hit clustering were determined: the expected number of clusters per sample was divided by the expected number of clusters per virotype, as determined by the average genome size given by GAAS and the mapping of clusters per Kb of viral contigs.

Assembly and analysis of phage genomes.

The 454 Newbler assembler Software Release 2.0.01.14 was used for assembly of viral genomes. Default parameters were employed except for minimum identity between the sequences (98%) and minimum overlap (100bp). These stringent conditions diminish the risk of false assembly between reads from different viruses[41].

We created an online tool for visualizing assembled viral contigs (http://gordonlab.wustl.edu/phage_omics/). Each contig with a length >10Kb generated from the assembly was blasted against a non-redundant set of proteins encoded by viruses present in our NR_Viral_DB (contains entries deposited in public databases as of May, 2009), as well as translated ORFs present in 121 sequenced microbial genomes representing cultured representatives of the human gut microbiota. ORF prediction was performed using glimmer3[42]. ORFs were subsequently annotated based on blastx searches (e-value < 1E-5) of the KEGG (v51), COG/String (v8[43]), PFAM (v23[44]), and TI-

GRFAM (v7[45]) databases. All features and annotations for each contig were included in a MySQL database and displayed using lightweight genome viewer[46].

All processed pyrosequencing reads from each VLP sample were blasted (blastn e-value <1e-5) against each of the contigs. Significant hits were recorded and the positions used for plotting cumulative coverage. The length of the alignment was used to calculate a normalized coverage value[47]. Percent similarity of each read to the contig was also calculated and averaged for total percent identity calculations.

Functional assignment of reads and statistical analyses.

All available pyrosequencing reads from fecal microbiomes and VLP-derived viromes were used to query (blastx e-value <1e-5) the KEGG (v51) and COG/String (v8[43]) databases. The same databases were queried (blastp, e-value < 1e-5) with known and predicted proteins encoded by the 121 reference sequenced human gut microbial genomes, and by all viral genomes (excluding prophages) in our NR_Viral_DB. After best blast hits were assigned to COG categories or KEGG second level pathways, Metastats[19] was used to identify significant functional differences ($p<0.05$) between fecal virome and microbiome datasets.

Prophage Coverage Plots.

Prophage present in the 121 gut microbial genomes were identified using PhageFinder. Each identified prophage was then extracted *in silico* together with 50Kb of flanking bacterial genomic sequences. Nucmer[48] was subsequently used to map all VLP pyrosequencer reads (defaults settings) onto this set of extracted sequences. Mummerplot, which like Nucmer is part of the Mummer package[48], was employed to generate sequence identity plots (threshold, > 80% similarity). The prophage genome coordinates of the matches and the sequence identity with VLP reads were used to generate tables of 'percent coverage' and 'fold-coverage'.

Search for integrase genes.

All integrase protein sequences were extracted from the NR_Viral_DB and from the 121 sequenced human gut-associated microbial genomes. VLP pyrosequencing reads were blasted against this database of extracted sequences (blastp, e-value <1e-4) and the best blast hit stored for every read. The number of reads from a given sample that were similar to a given integrase in the database was recorded and used to generate a Hellinger-based distance matrix between samples (QIIME).

CRISPR spacers represented in viral metagenomes.

Seventy-four available human gut microbial genomes, representing members of the most predominant bacterial families present in the human fecal microbiota, were used to search for CRISPR elements with CRISPR-Finder[49]. CRISPRs were identified in 48 of these genomes: they contained a total of 95 different repeat sequences and 2,196 spacers. The direct repeats were subsequently compiled into a database, which in turn was used to search each of our fecal microbiome datasets (Program cross_match[50]; parameters: -minmatch 7 -maxmatch 12 -gap1_only -screen -minscore 10). Spacers were then extracted from all microbiome pyrosequencer reads where at least 2 matches for the same CRISPR repeat were identified and the intervening spacer was ≥10 nucleotides. All of these spacers from the microbiomes and 121 sequenced reference genomes were pooled together, and used to screen all VLP-derived pyrosequencing reads using cross_match (parameters: -minmatch 14 -maxmatch 14 -gap1_only -screen -minscore 10). Virome reads with hits to microbiome or reference genome CRISPR spacers over >90% of the length of their spacers were recorded.

Gnotobiotic mouse experiments.

All studies with mice used protocols approved by the Washington University Animal Studies Committee. Methods for co-colonization of adult germ-free adult male C57Bl/6J mice with *Marvinbryantella formatexigens* and *Bacteroides thetaiotaomicron*, harvesting their cecal contents, preparation of rRNA-depleted RNA from cecal contents and fecal samples for subsequent cDNA synthesis, Illumina GA-IIx sequencing of cDNA, plus mapping and normalization of the result-

ing reads are described in another publication[23]. RNA-Seq datasets used for our analysis of prophage gene expression can be found under GEO accession numbers GSM544893, GSM544900, GSM544858, GSM544866, GSM544940, GSM544944 (in vivo data) and GSM544856, GSM544873, GSM544835, GSM544947, GSM544872, GSM544917, GSM544931, GSM544871, GSM544883, GSM544863, GSM544865 (in vitro data).

Statistical tests.

Statistical tests were performed and heatmaps were produced using the R package[51].

Methods References

34 Gomez-Alvarez, V., Teal, T.K., & Schmidt, T.M., Systematic artifacts in metagenomes from complex microbial communities. ISME J 3, 1314-1317 (2009).

35 Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J., & Knight, R., Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods 5, 235-237 (2008).

36 Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7, 335-336 (2010).

37 Wang, Q., Garrity, G.M., Tiedje, J.M., & Cole, J.R., Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73, 5261-5267 (2007).

38 Fouts, D.E., Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res 34, 5839-5851 (2006).

39 Angly, F.E. et al., The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput Biol 5, e1000593 (2009).

40 Angly, F. et al., PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6, 41

(2005).

41 Breitbart, M. et al., Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99, 14250-14255 (2002).

42 Delcher, A.L., Bratke, K.A., Powers, E.C., & Salzberg, S.L., Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23, 673-679 (2007).

43 Jensen, L.J. et al., STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37, D412-416 (2009).

44 Finn, R.D. et al., The Pfam protein families database. Nucleic Acids Res 36, D281-288 (2008).

45 Haft, D.H., Selengut, J.D., & White, O., The TIGRFAMs database of protein families. Nucleic Acids Res 31, 371-373 (2003).

46 Faith, J.J., Olson, A.J., Gardner, T.S., & Sachidanandam, R., Lightweight genome viewer: portable software for browsing genomics data in its chromosomal context. BMC Bioinformatics 8, 344 (2007).

47 Rohwer, F., Seguritan, V., Choi, D.H., Segall, A.M., & Azam, F., Production of shotgun libraries using random amplification. Biotechniques 31, 108-112, 114-106, 118 (2001).

48 Kurtz, S. et al., Versatile and open software for comparing large genomes. Genome Biol 5, R12 (2004).

49 Grissa, I., Vergnaud, G., & Pourcel, C., CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35, W52-57 (2007).

50 Gordon, D., Desmarais, C., & Green, P., Automated finishing with autofinish. Genome Res 11, 614-625 (2001).

51 Team, R.D.C., R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2009).

Supplementary Discussion

The value of purifying VLPs for viral metagenomic projects

When each of the 32 fecal VLP-associated viromes, sequenced to an average depth of 7.8±2.9 Mb (per sample) was used to query the 12 microbiomes sampled to an average depth of 92.2 ± 17.5 Mb, we noted that 55.8±32.4 % (mean ± SD) of viral sequences generated from the VLP preps from a given human host were detectable in that individual's sequenced fecal microbiome. When a deeply sampled VLP-virome (70.16 Mb) was used to query 0.91 Gb of pyrosequencer reads from the corresponding deeply sequenced fecal microbiome[15], the percentage of VLP-derived sequences found in the fecal community DNA sample was 76.14% (**Fig. S13).** Using the same BLAST E-value threshold cutoff, we performed a reciprocal analysis, asking what percentage of the total sequences present in each of the microbiome datasets matched to sequences present in VLP datasets generated from fecal samples collected from that human host. The results disclosed that viral reads represented 3.5±2.2% (mean ± SD) of total fecal community DNA sequences in the case of the 12 more shallowly sequenced microbiomes, and 2.5% in the case of the deeply sequenced microbiome and its corresponding deeply sequenced virome (**Fig. S13a**). These findings support a view that at the present time isolating VLPs is an efficient and direct way to characterize phage populations associated with a given (fecal) microbial community.

CRISPRs

Clusters of Interspaced Short Palindromic Repeats (CRISPR) elements are stretches of DNA composed of short palindromic repeats (23-47bp) that flank short "spacers" composed of viral DNA; their presence in a bacterial genome represents a key component of host defense against bacteriophage attack[2,3]. We used CRISPR-Finder[49] to search sequenced human gut microbial genomes; CRISPR elements were detected in 48 of 74 human gut bacterial species queried and in a prominent human gut archaeon (*Methanobrevibacter smithii*) (see **Table S10** for the list of genomes). We identified a total of 95 different direct repeats and 2,196 different spacers. These direct repeats were subsequently used to interrogate the fecal microbiome datasets to identify reads that contained at least two copies of the same direct repeat. The spacers interposed between these repeats

were subsequently extracted, and together with the spacers from the 121 sequenced human gut microbial genomes used to search for sequences with high similarity in VLP viromes (defined by Cross_match; maximum of 1 gap allowed and similarity over ≥ 90% of its length). This effort yielded 1,262 reads that were similar (≥90% identity) to a spacer sequence. Sixteen of the 38 VLP viromes (including technical replicates of samples from members of families F1-4 and the deeply sequenced virome) had hits to spacers derived from fecal microbiomes. In the 12 sequenced fecal community microbiomes for which there were corresponding VLP preparations, the only hits to the viromes were spacers represented in another individual's microbiome (**Table S11**). In this analysis of fecal microbiome datasets from a single time point, and at this depth of shotgun sequencing of the microbiome, the absence of detectable viral sequences with significant similarity to bacterial spacers in a given individual's fecal microbiome suggests that viruses to which their bacterial communities were resistant are not represented in the corresponding VLP preparation. If temperate phage dominate in the fecal microbiome, we would not expect such resistance to appear at least as judged by the representation of CRISPR spacers in viromes and microbiomes. However, additional and deeper shotgun datasets of total fecal microbial community DNA need to be generated from samples collected at all time points surveyed for each individual in order to further assess whether resistance does or does not occur.

Eukaryotic viruses represented in VLP viromes

Although 73% of the sequences in the NR_Viral_DB belong to eukaryotic viruses, none of the VLP samples yielded reads covering more than 50% of the genome of any known eukaryotic virus (tblastx, e-value < 1e-3). Eukaryotic viruses with hits throughout more than 20% of their genomes included: (i) six non-human Herpesviridae with 22–49% genome coverage; (ii) one Maculavirus (Grapevine_fleck_virus; 39% coverage of its 7,564bp genome); (iii) one Aquareovirus (Aquareovirus_A_segment_11; 26% coverage of this 783bp segment of its genome); (iv) one Parapoxivirus (Bovine_papular_stomatitis_virus; 22% coverage of its 134,431bp genome); and (v) one human Rotavirus (Human_rotavirus_G3_segment_11; 21% coverage of this 1,043bp genome segment).

Supplementary Figure Legends

Suppl. Figure 1. Percent of pyrosequencing reads generated from VLP preparations that map to the NR_Viral_DB. Sample-by-sample distribution of the percentage of VLP-derived reads with hits (tblastx, e-value < 1e-3) to the NR_Viral_DB. See the legend to **Fig. 1** in the main text for an explanation of the nomenclature used to designate samples.

Suppl. Figure 2. **Correlating family-level bacteria taxa present in fecal samples with the known bacterial hosts of bacteriophage present in the NR_Viral_DB and their identified homologs in fecal VLP metagenomic datasets**. The Universal Bacterial 16S rRNA tree in Greengenes was downloaded (http://greengenes.lbl.gov/Download/Taxonomic_Outlines/), collapsed at a family level based on NCBI taxonomy, and branches colored according to their assigned phyla. The left panel corresponds to distribution and relative abundance (1.0 = most abundant) of the different samples according to 16S rRNA data. The middle panel shows the distribution and percent coverage of bacteriophage genomes from the NR_Viral_DB by VLP-derived reads; phage genomes are classified according to their host taxonomy. The right panel shows the distribution and relative abundance of the known bacterial hosts of phage present in the NR_Viral_DB. Columns are sorted by individual and time points of fecal sampling. Green arrows point to ssDNA phage from *Chlamydia* and *Bdellovibrio* known to be preferentially amplified by WGA methods[17]. For sample abbreviations see **Fig 1** in the main text.

Suppl. Figure 3. Representative Monte Carlo simulations for cross contigs defining intrapersonal vs interpersonal variation in VLP DNA viromes. Monte Carlo simulations for the percent shared viral genotypes (virotypes) and percent permuted rank abundance of virotypes between pairs of fecal VLP samples. Colors indicate the likelihood score for a given position. Intra-personal variation is displayed in panels a-c: F4M.2 vs F4M.3 (a); F2T1.1 vs F2T1.3 (b); F1T2.2 vs F1T2.3 (c). Inter-personal variation is illustrated in panels d-f: F3M vs F3T1 (d); F2T1 vs F2T2 (e); F4T1 vs F4T2 (f).

Suppl. Figure 4. Beta-diversity analysis. Branch support for the trees displayed in **Fig. 2**. Hellinger-based UPGMA trees for bacterial 16S rRNA data and VLP-derived viromes are displayed in panels a and b, respectively. The color key provides information about family (F) and the family member.

Suppl. Figure 5. Percent similarity plots of VLP virome reads mapping to a predicted prophage in *Ruminococcus torques* ATCC 27756. The genes present within the ~60 Kbp prophage are shown in green, and those present on either strand of the flanking bacterial genome are shown in black at the bottom of the figure. Pyrosequencer reads, generated from fecal VLPs, prepared at 2 or more time points from a co-twin (T2) and her mother (M) belonging to family 2 (F2) and having ≥80% identity with prophage genes, are displayed as blue dots (each dot represents a single read with a hit to the positive strand of the prophage) or red dots (negative strand hits).

Suppl. Figure 6. Length distribution of viral contigs assembled from VLP-derived pyrosequencing reads. A frequency histogram of contig length is shown.

Suppl. Figure 7. Percentage of fecal virome and microbiome reads with significant hits to COG categories and KEGG second level pathways. Sample by sample percentage of reads with significant hits (blastx, e-value cutoff <1e-5) to (**A**) COG (STRING v7) and (**B**) KEGG (v44) databases.

Suppl. Figure 8. KEGG and COG annotations reveal significant differences in functions between fecal VLP-associated viromes and microbiomes. Only COG-categories (panel A) and KEGG second level pathways (panel B) with significant differences in their representation between fecal microbiomes and VLP-associated viromes are shown (mean ± s.e.m plotted; $p < 0.05$; two sample t-test calculated using Metastats).

Suppl. Figure 9. A sample-by-sample view of the proportional representation of COG categories in sequenced VLP-associated viromes and gut microbiomes. Blastx assignment (e-value cutoff <1e-5) of reads to functional categories. Shown from top to bottom are proteins from viruses

in the NR_Viral_DB and fecal VLP-derived viromes, plus proteins from 121 sequenced human gut-associated microbial genomes and fecal microbiomes. See **Fig. 1** for sample nomenclature.

Suppl. Figure 10. Comparison of the representation of KEGG and COG groups in proteins encoded by large VLP-derived contigs and in the NR_Viral_DB. Searches (blastp, e-value <1e-5) were performed against the STRING COG database (panel a) and KEGG (results for second level pathways are shown in panel b).

Suppl. Figure 11. Sequence diversity of integrase genes in VLP viromes. The number of pyrosequencer reads in each VLP sample with significant hits to known integrases present in the NR_Viral_DB and in prophages found in 121 human gut microbial genomes were identified and used to generate a distance matrix. Average distances among technical replicates (two shotgun datasets produced from a given VLP DNA preparation), among samples obtained from the same individual over time (intrapersonal variation), and samples obtained from co-twins, twins and their mothers or unrelated individuals, are graphed (mean ± s.e.m). The significance of differences between the groups was calculated using Student's t-test. *** $p< 0.001$; ** $p<0.01$, ns, $p>0.05$.

Suppl. Figure 12. Normalized RNA-Seq counts for predicted prophages in *Bacteroides thetaiotaomicron* VPI-5482. RNA-Seq was performed using rRNA-depleted RNA samples prepared from cecal and fecal contents harvested from gnotobiotic mice co-colonized for 2 weeks with *B. thetaiotaomicron* and *M. formatexigens* (n=3 animals). Expression levels are shown for each ORF (see color key for normalized read counts; normalization based on sequencing effort and length of each predicted ORF). Active expression is defined as a normalized read count >100. This strain of *B. thetaiotaomicron* contains two prophages. One of the prophages (labeled 1) contains a linked pair of highly expressed ORFs encoding an Xre family anti-toxin (BT4733) and a putative toxin (BT4732) while the other prophage contains a cluster of three highly expressed genes [two hypothetical proteins flanking a Xre family anti-toxin (BT4035)].

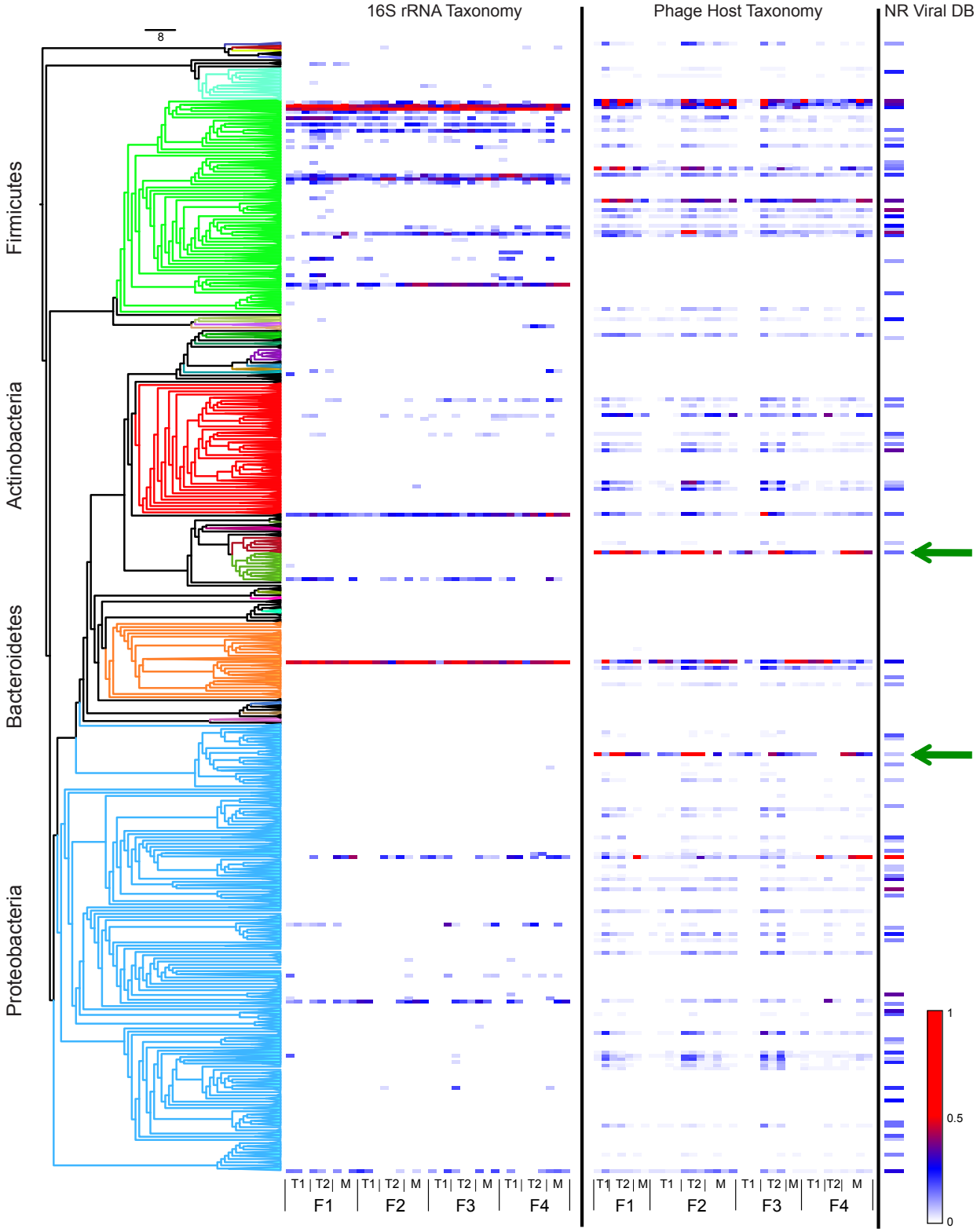
Suppl. Figure 13. Representation of VLP pyrosequencer reads in fecal microbiomes and vice versa. The percentage of reads from fecal microbiomes with significant similarity to VLP-derived

reads (blastn, e-value <1e-7) is represented as a blue wedge within the red pie charts. This wedge is expanded to the right in the form a second blue pie chart that shows the percentage of reads from each of the different time point VLP preparations that have significant similarity with reads from the fecal microbiome from time point 1. (a) The percentage of shared reads between the deeply sequenced F5T2 VLP preparation and corresponding deeply sequenced fecal microbiome (293,654 and 2,579,680 reads, respectively). (b) Data derived from shallowly sequenced fecal viromes and microbiomes.

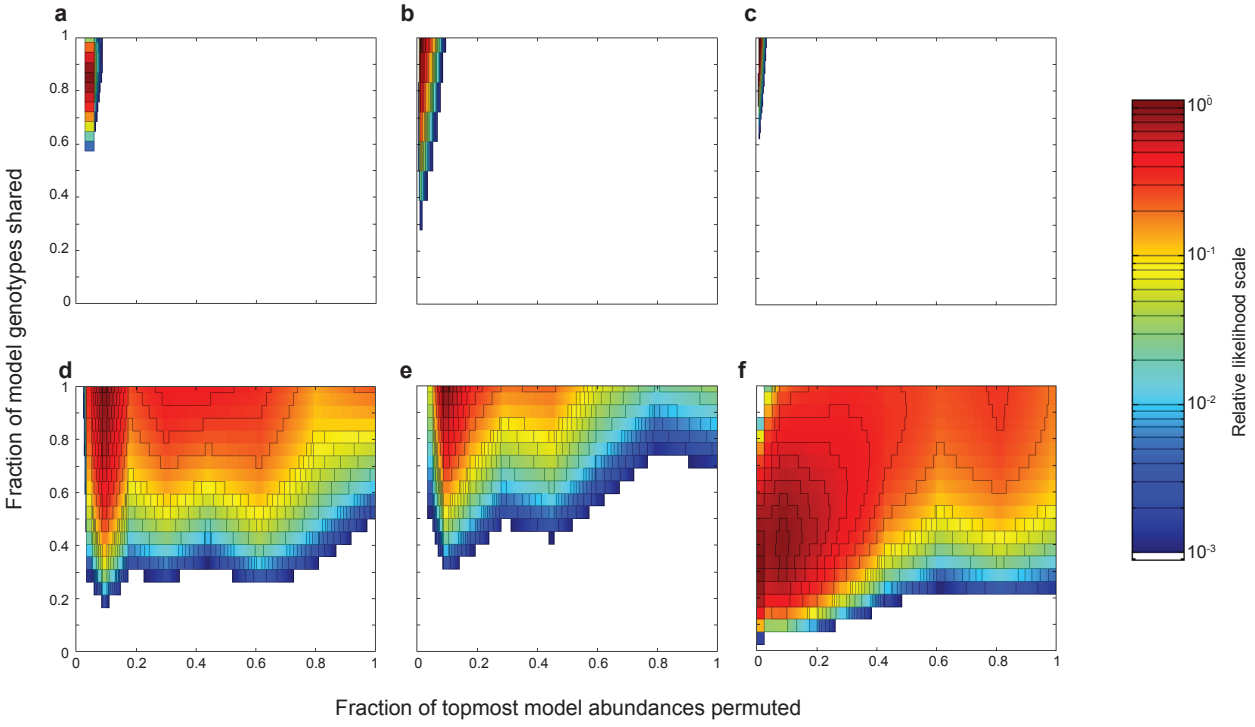
Supplementary Figures

Suppl. Figure 1.

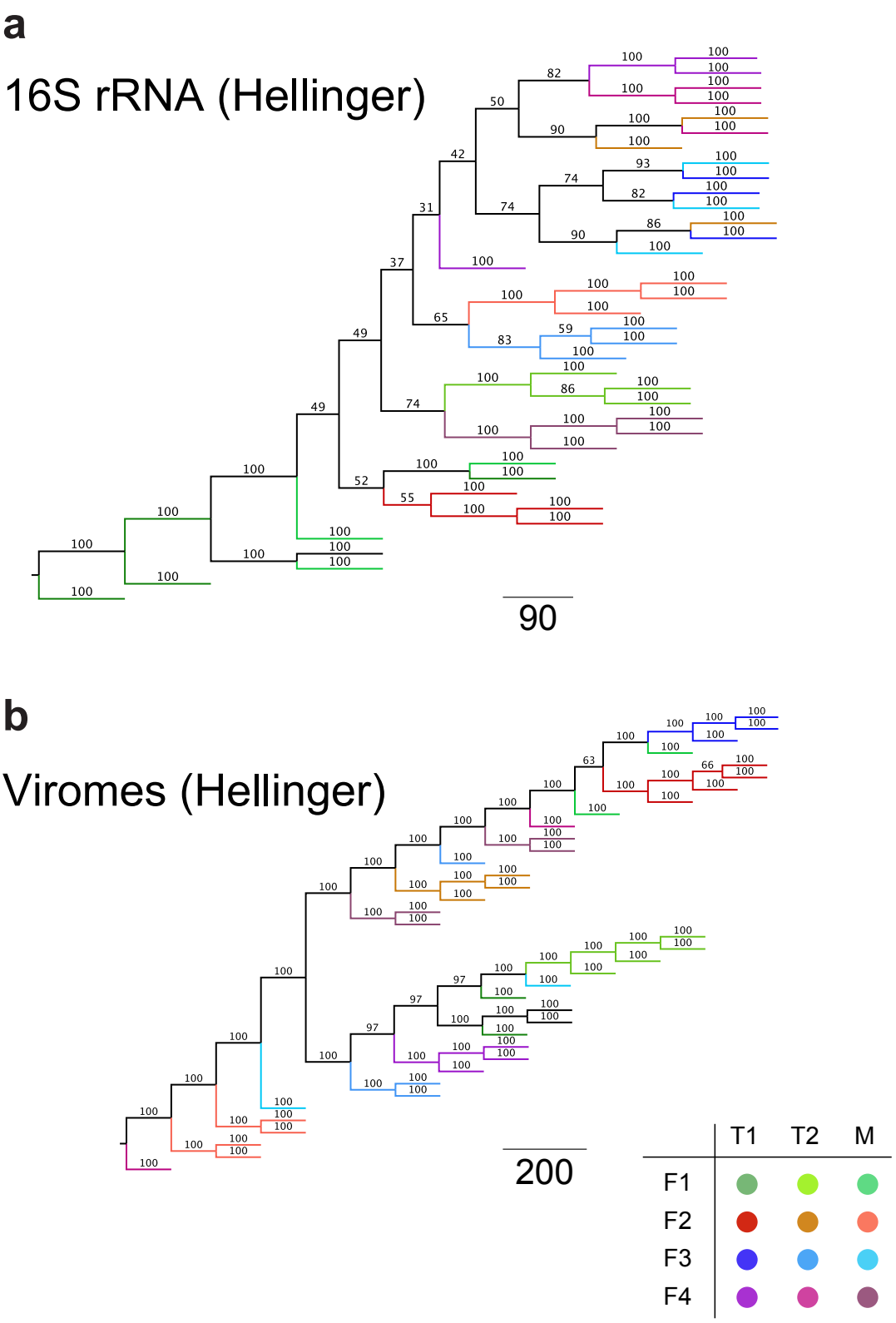
Suppl. Figure 2



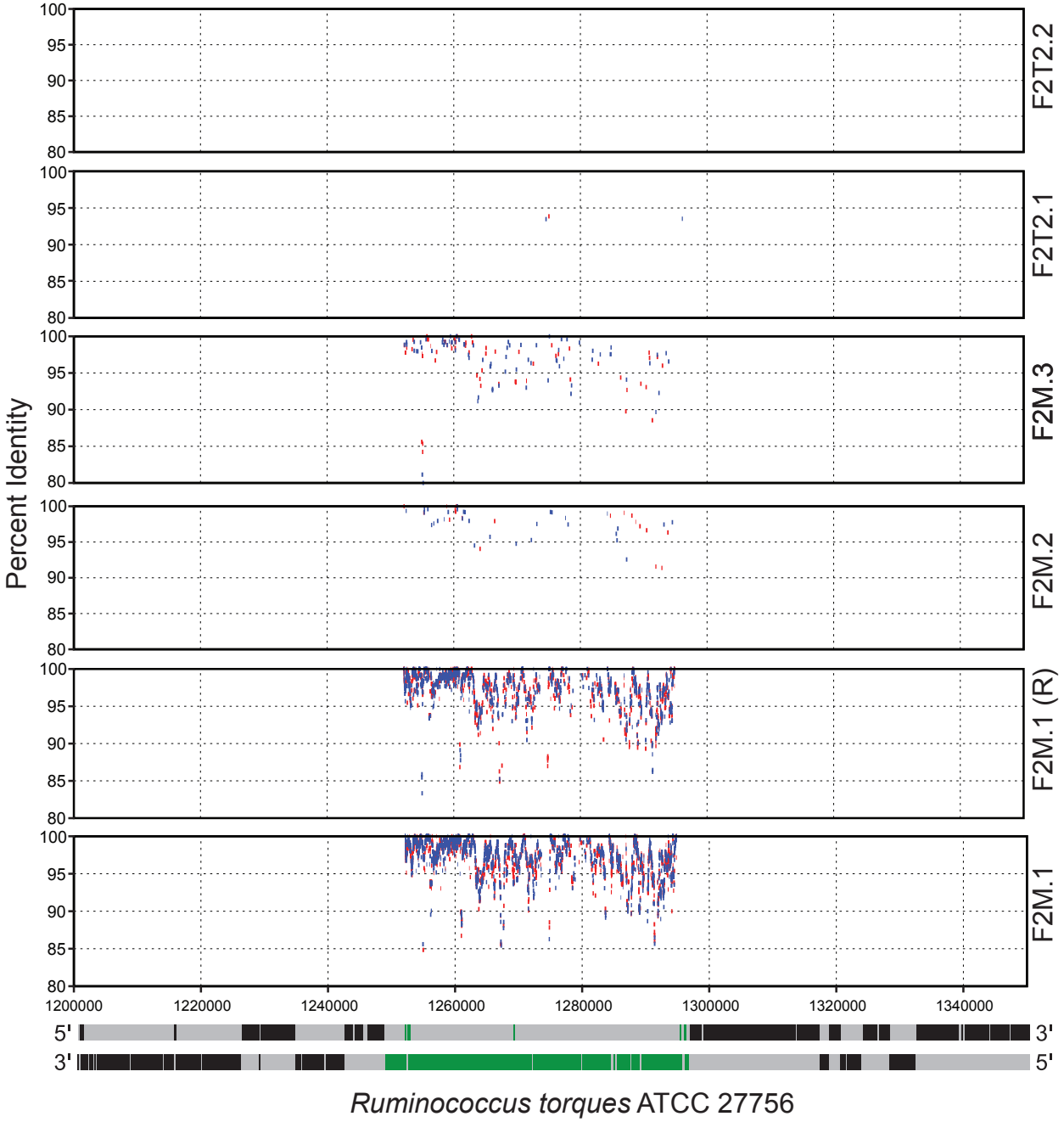
Suppl. Figure 3



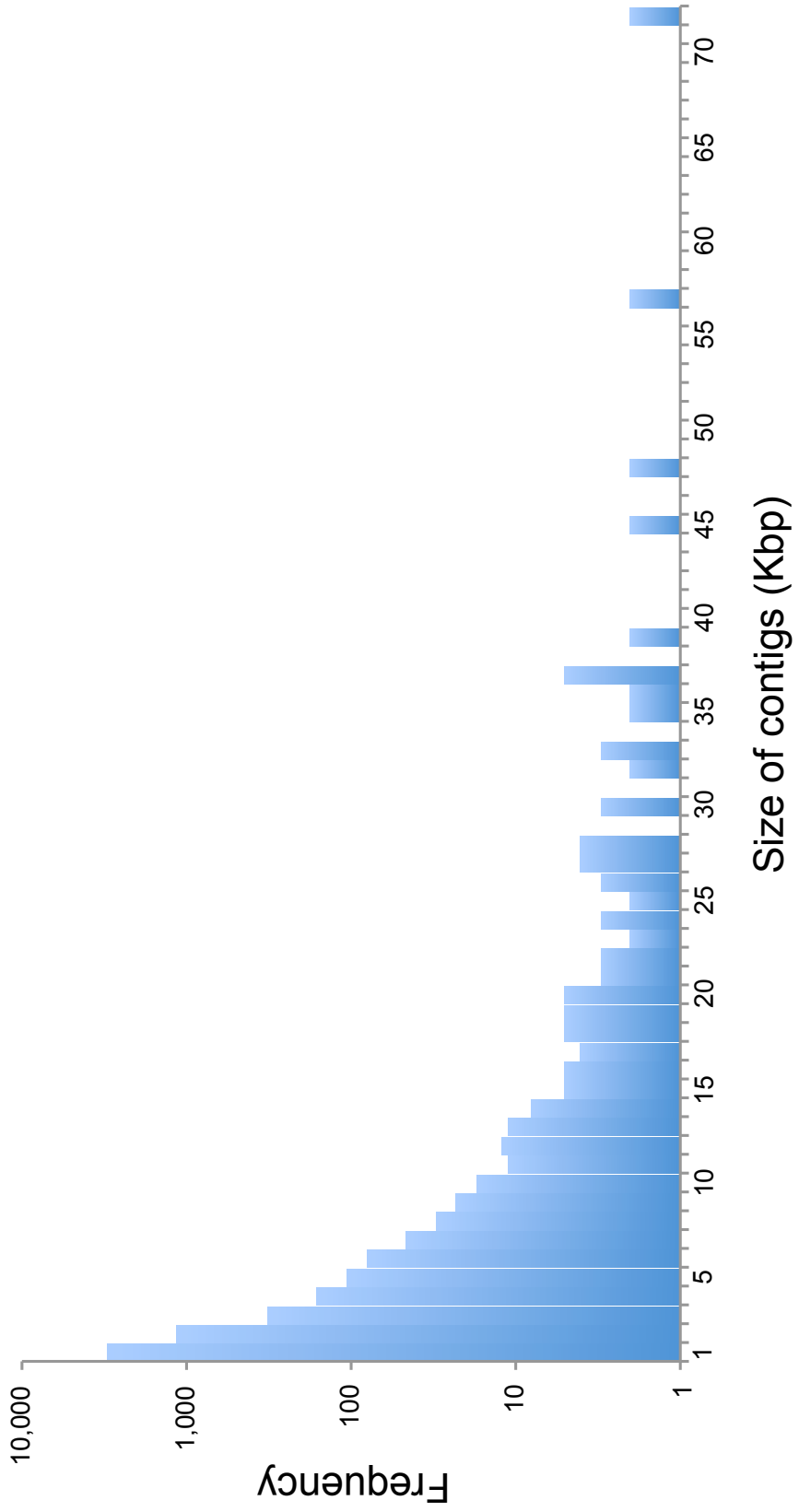
Suppl. Figure 4



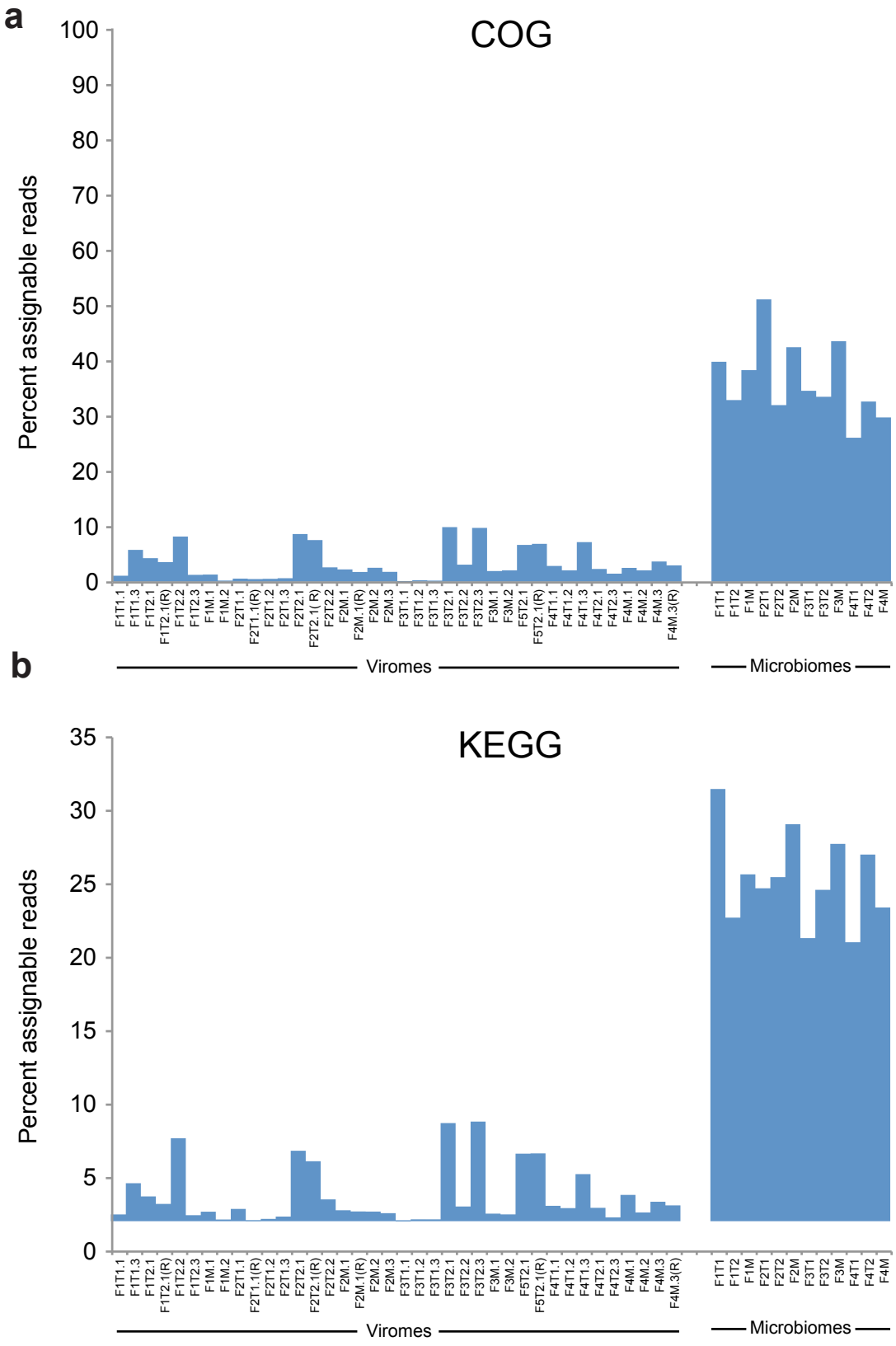
Suppl. Figure 5



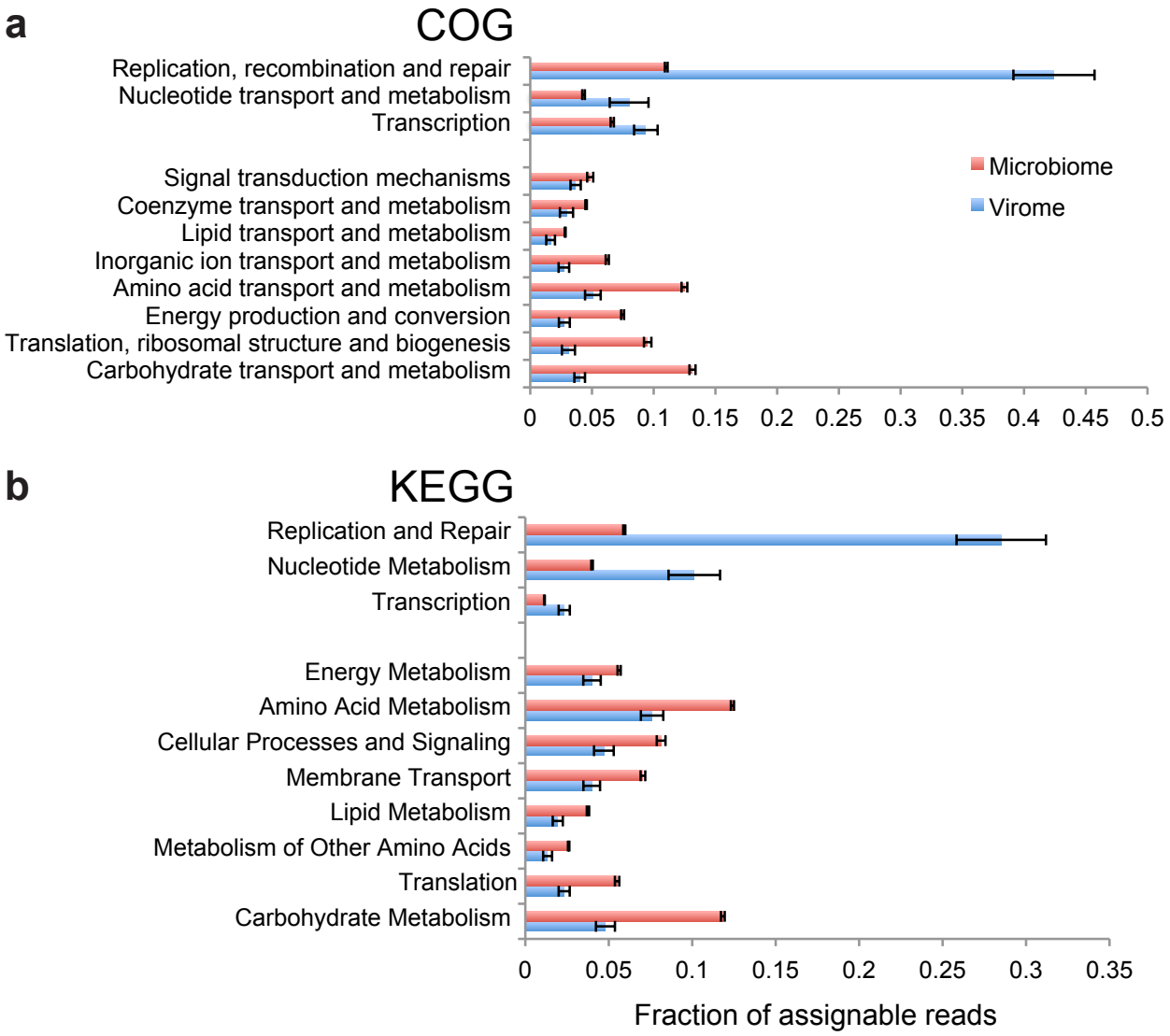
Suppl. Figure 6



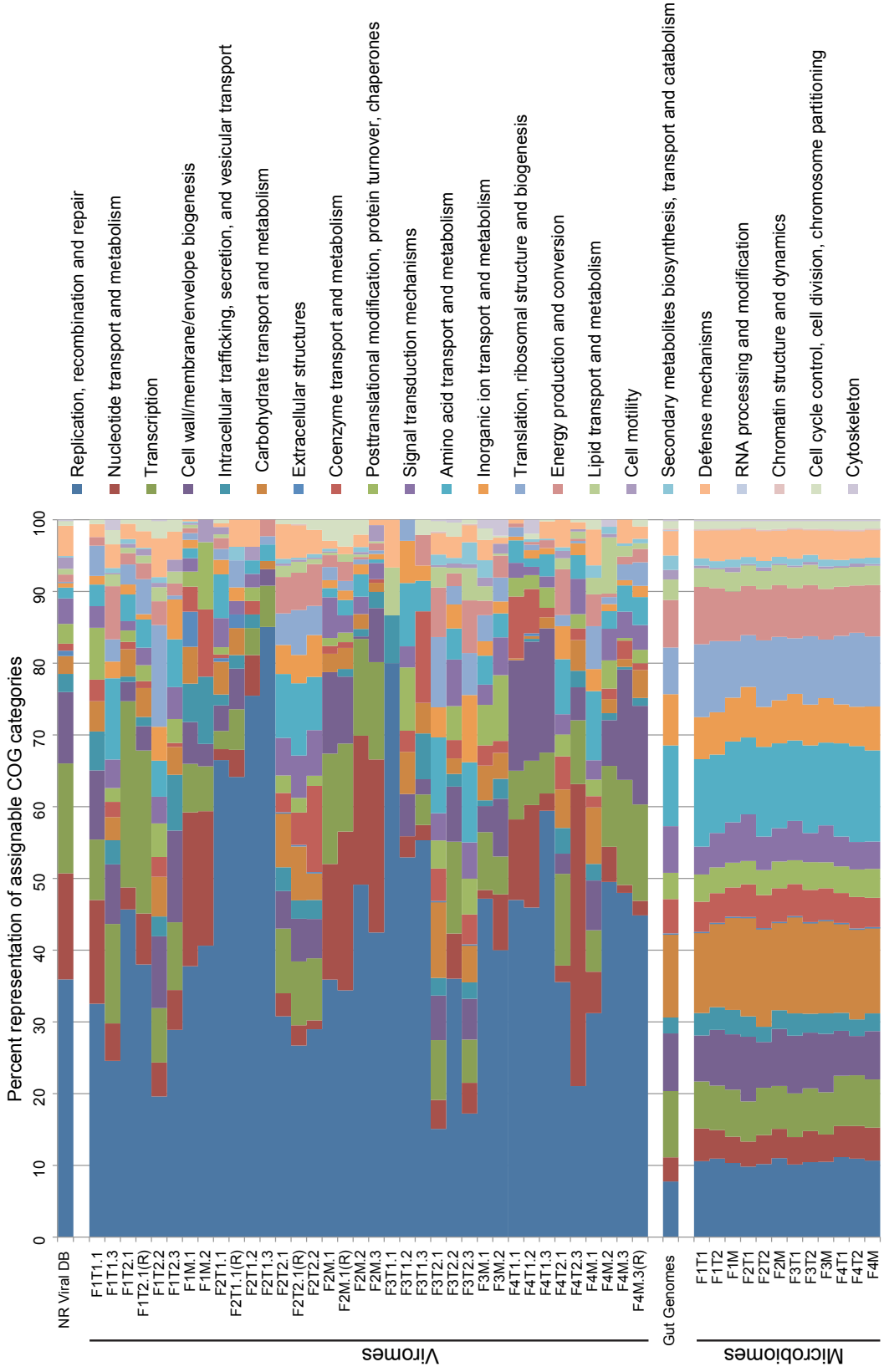
Suppl. Figure 7



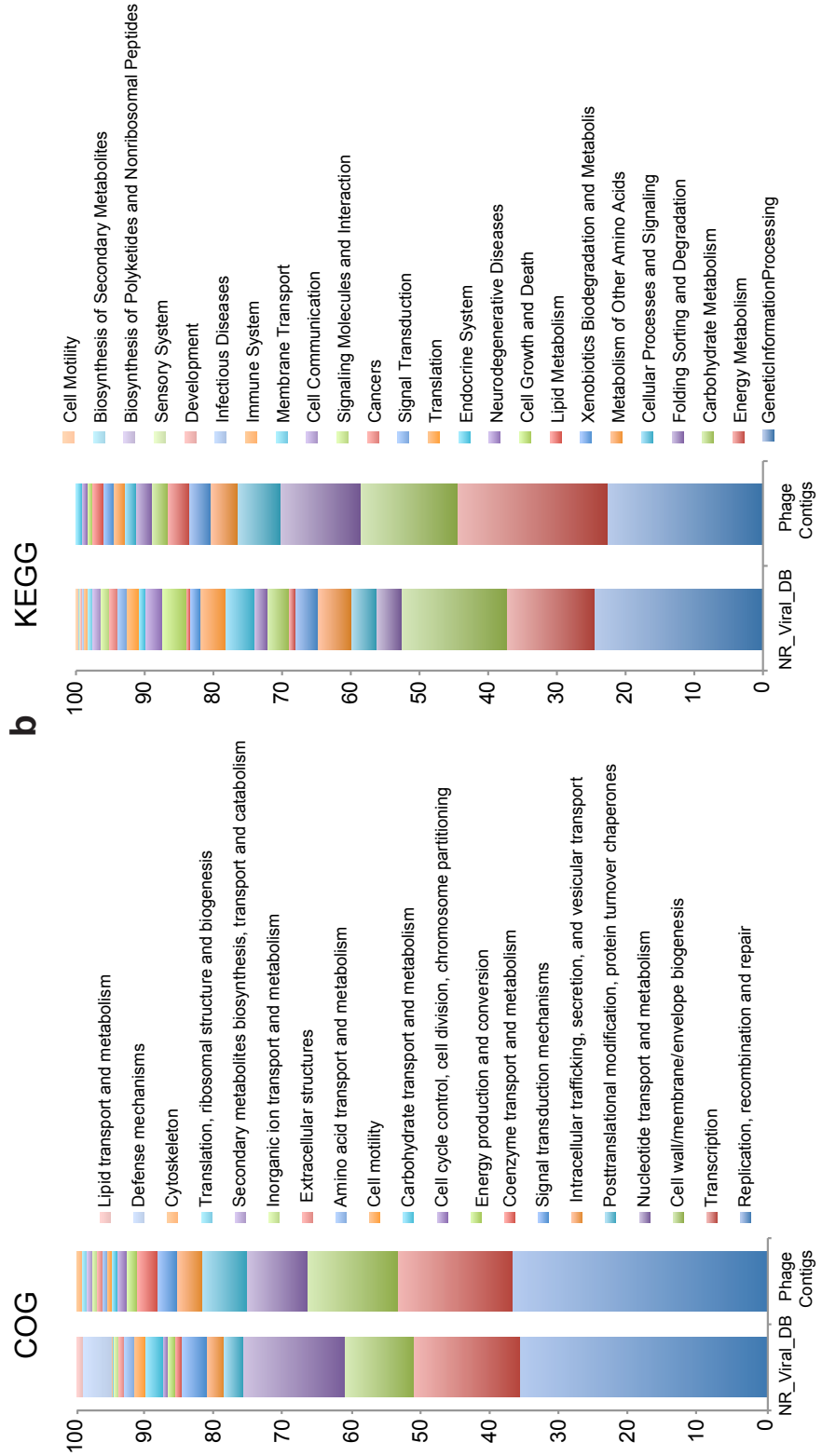
Suppl. Figure 8



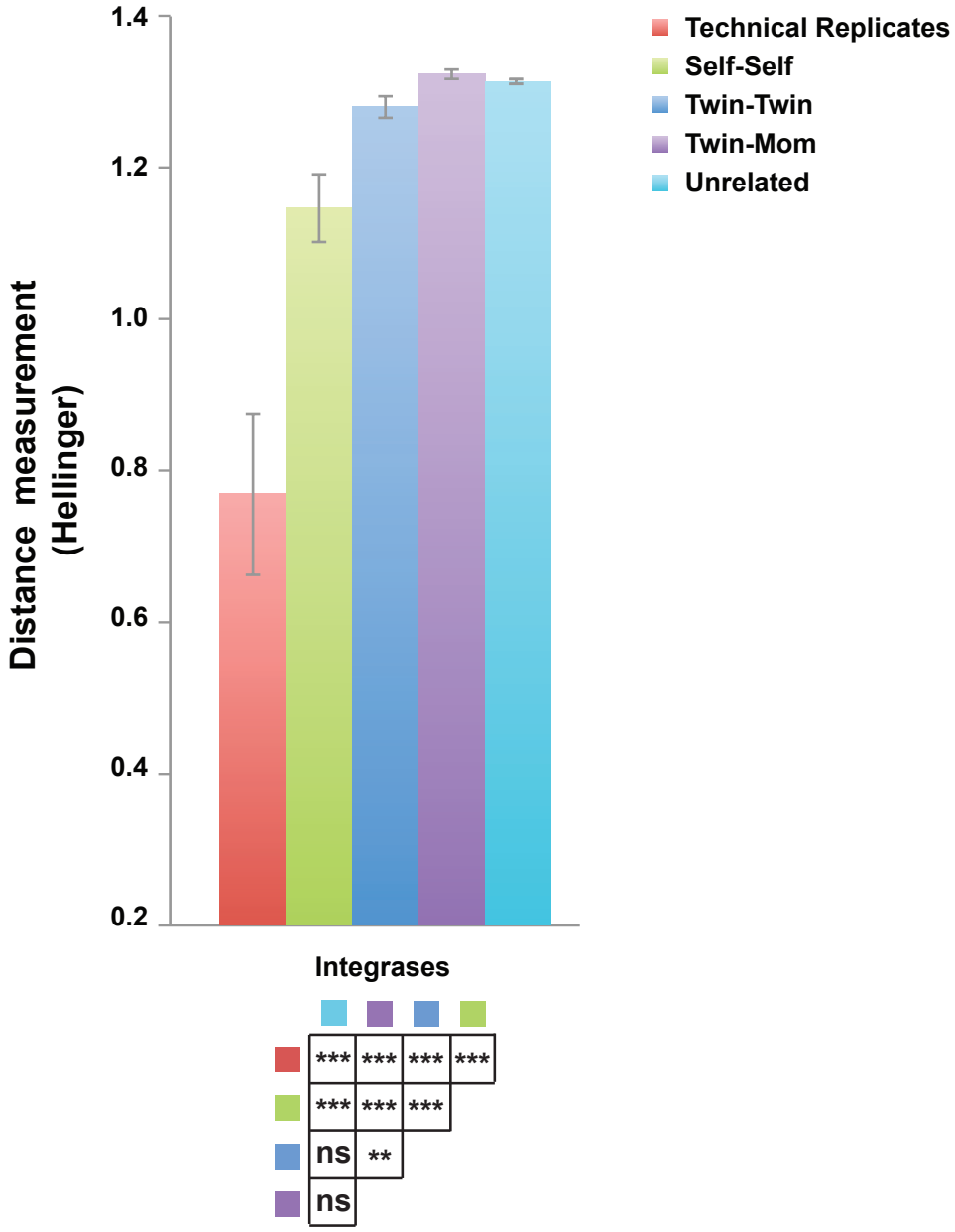
Suppl. Figure 9



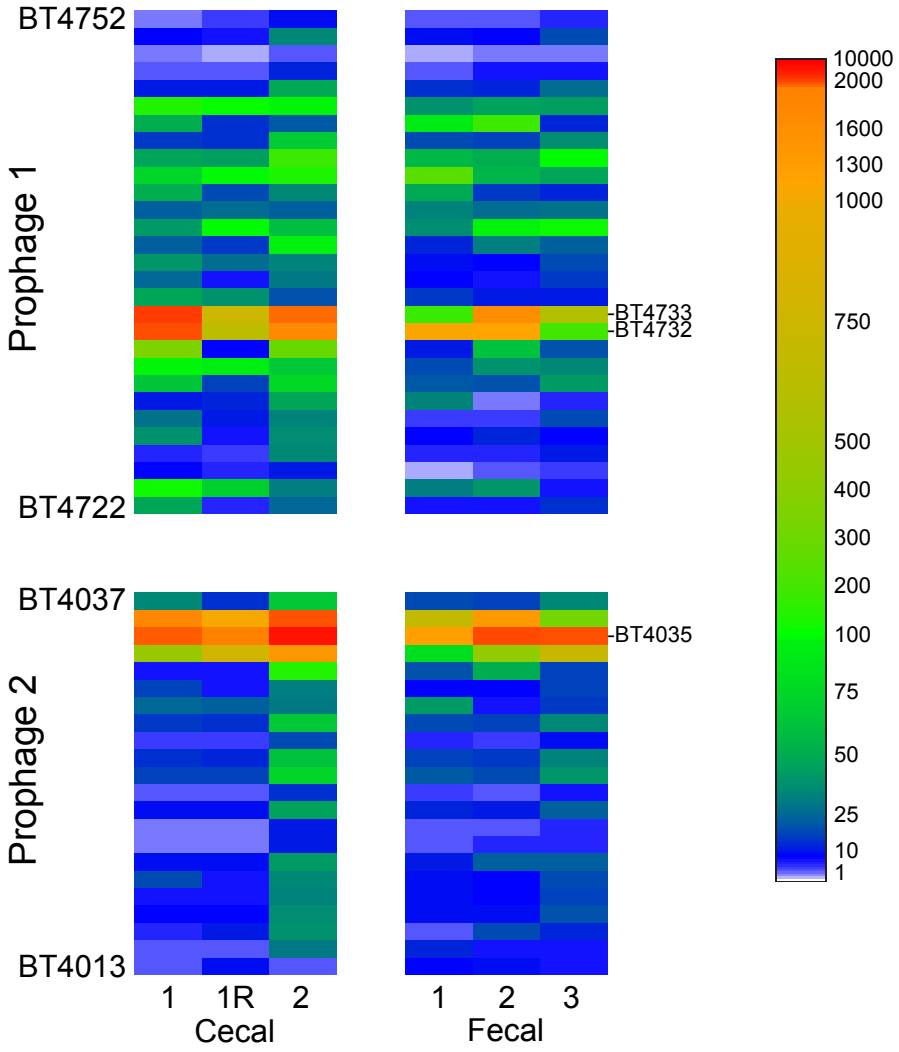
Suppl. Figure 10



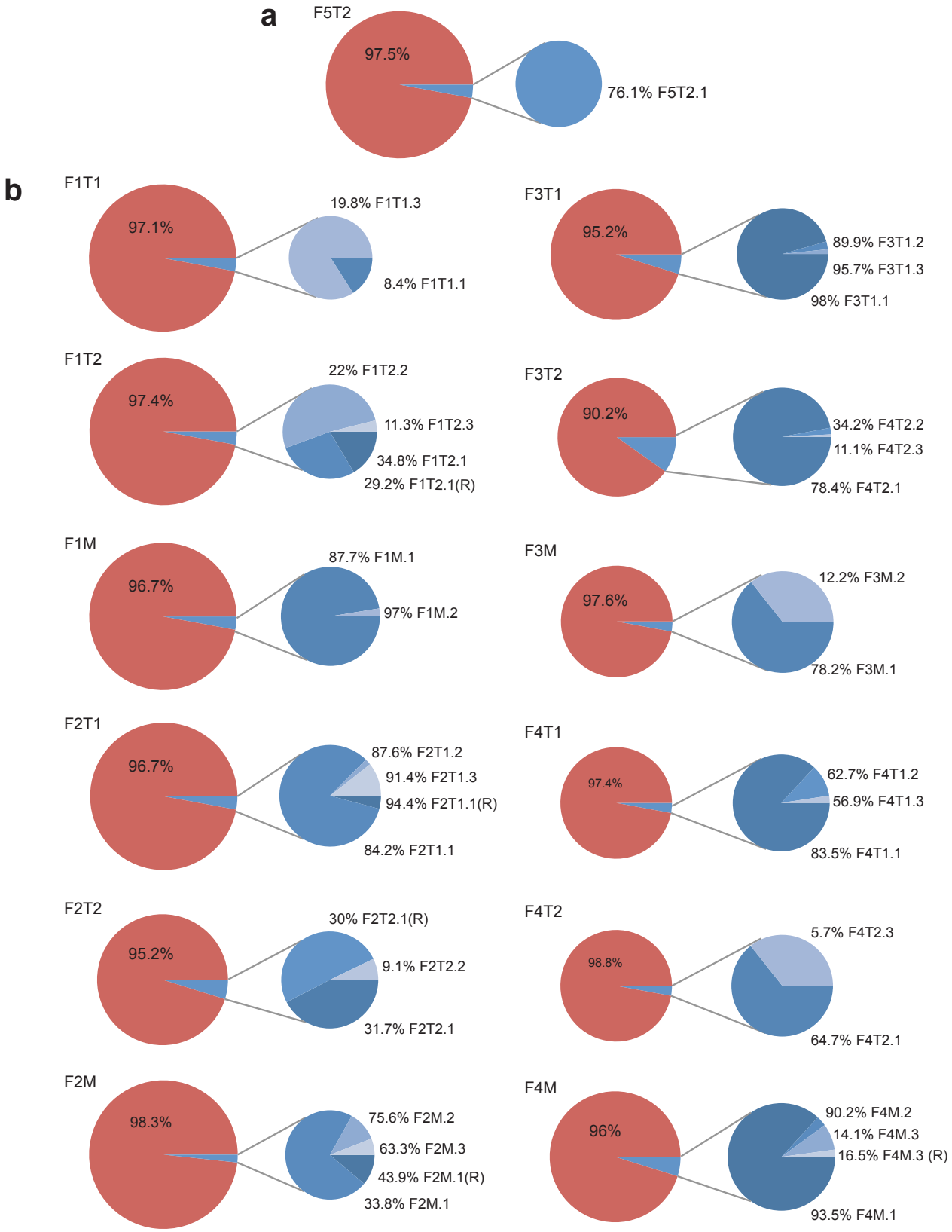
Suppl. Figure 11



Suppl. Figure 12



Suppl. Figure 13



Supplementary Tables

Suppl. Table 1. Sequencing effort for VLP preparations from 32 fecal samples obtained from 4 sets of MZ twins and their mothers. Technical replicates were performed on 6 DNA samples involving independent whole genome amplification and shotgun 454 FLX pyrosequencing. Sample F5T2.1 was subjected to deeper sequencing. NA, a fecal specimen was not available in sufficient quantity to purify VLPs at this time point.

Suppl. Table 2. Sequencing effort for bacterial 16S rRNA genes present in the fecal microbiota of study participants.

Suppl. Table 3. Shotgun sequencing effort for fecal community DNA (microbiome) samples.

Suppl. Table 4. List of 396 sequenced microbial genomes used to identify prophage sequences for the NR_Viral_DB.

Suppl. Table 5. CD-hit cluster-based alpha diversity metrics.

Suppl. Table 6. PHACCS-based alpha diversity metrics.

Suppl. Table 7. Matrix of VLP samples x reference human microbial gut genomes where significant coverage to prophages in the microbial host was identified. (a**)** Percent of the prophage genome covered by reads from a given VLP sample. (b) Fold coverage per bp of the prophage genome, normalized to 10,000 reads. Yellow highlights instances where the prophage was covered over more than 50% of its length.

Suppl. Table 8. Matrix of VLP samples x the 88 large contigs assembled from the aggregate VLP dataset showing (a) the percentage of the contig covered by reads from a given VLP sample and (b) fold-coverage per bp of each contig. Data are normalized by randomly mapping 14,000 reads per VLP sample. Yellow highlights instances where a given contig has ≥50% coverage with reads from a given VLP virome.

Suppl. Table 9. Contig x VLP sample matrix of contigs present in more than one VLP sample. The percent identity between VLP-pyrosequencer reads and the corresponding contig is shown**.** Dashes indicate that no reads corresponding to the contig were present in the VLP sample **(**threshold; minimum of 80% overlap with the contig over the length of the read).

Suppl. Table 10. List of 121 sequenced human gut-derived microbial genomes used for generating a reference list of KEGG second level pathways and COG assignments. The presence or absence of CRISPRs in the genomes of the 74 organisms used to search for these elements is noted.

Suppl. Table 11. The number of VLP reads with similarity over more than 90% of their length to CRISPR spacers present in either the fecal microbiome of the same individual or in a reference gut-associated microbial genome. Counts are normalized (10,000 reads per sample). The bottom row shows the number of spacers identified in each fecal microbiome sample (see *Supplementary Discussion* for further definition of the criteria used to select spacers from the fecal microbiome datasets to query VLP datasets).

Chapter 3

A gnotobiotic mouse model for characterizing phage-bacterial host interactions in the human gut

Chapter 3

A gnotobiotic mouse model for characterizing phage-bacterial host interactions in the human gut

Alejandro Reyes[1], Meng Wu[1], Nathan McNulty[1], Forest Rohwer[2], and Jeffrey I. Gordon[1s]

[1]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108 USA

[2]Department of Biology, San Diego State University, San Diego, CA

Summary

The microbial diversity, interpersonal variations, and dynamism of the human gut microbiota make the task of identifying the factors that define community configurations, both structural and functional, extremely challenging. Bacterial viruses (phage) are the most abundant biological group on Earth, and are more diverse than their bacterial prey/hosts. To characterize their role as agents shaping gut community structure, adult germ-free mice were colonized with a defined model community composed of 15 sequenced human gut symbionts, seven of which harbored prophages. One bacterial member (*Bacteroides cellulosilyticus* WH2) was represented by a library of >25,000 isogenic transposon mutants covering 80% of the genes in its genome. Once assembled, the community was subjected to a staged phage attack with a pool of live or heat killed virus-like particles (VLPs) purified from the fecal microbiota of five humans. Shotgun sequencing of DNA isolated from the input pooled human VLP preparation, gut microbiota samples collected over time from the gnotobiotic mice, including purified fecal VLPs, revealed an ordered, reproducible sequence of attack extending over a 25 day period involving five phage, none described previously. This system allowed us to associate phage present in the input VLP preparation with bacterial hosts, show that surviving members of the first bacterial species attacked did not contain detectable deletions, insertions or SNPs fixed in the population, determine that one of the five phages was present in four of the humans used to construct the VLP pool but that only one virotype from one donor was selected by the model human microbiota, plus characterize the prominent induction of a lambdoid prophage in *B. cellulosilyticus* that preceded introduction of the exogenous phage, and using transposon mutagenesis, establish the dramatic fitness advantage that one of its loci conferred on its bacterial host. Together, these results provide a defined community-wide view of the operations of a phage-bacterial host dynamic in the gut ecosystem.

The human gut is home to tens of trillions of microbial cells representing all three domains of life, although most are bacteria. These organisms collaborate and compete for functional niches and physical locations (habitats). Together, they form a continuously functioning microbial metabolic ‘organ’. The broad taxonomic (e.g., phylum-level) stability of the gut community observed

in individuals over time after the first 3 years of life, and the contrasting high diversity at finer taxonomical levels, fit a Constant Diversity dynamics model [1,2] where phage maintain high bacterial strain-level diversity through lysis of their host strains; the resulting emptied niche is filled with either an evolved resistant bacterial strain or a taxonomically closely related bacterial species. These dynamics have been observed in open marine environments [1]. However, recent metagenomic studies that characterized the gut virome of healthy individuals by purifying and sequencing virus-like particles (VLPs) from feces showed that the most abundant viruses are temperate phage [3-5]. In a study of adult monozygotic twins and their mothers, sampled over the course of a year, viral community structure showed high inter-personal variation with a contrasting stability within an individual over time, both at the level of sequence conservation and relative abundance [3]. These observations suggested that a temperate lifestyle rather than a predator-prey relationship characterizes the phage-host bacterial cell dynamic in a healthy distal human gut.

To improve our understanding of viral-bacterial host dynamics in the gut, we constructed a gnotobiotic mouse model containing a simplified defined microbiota composed of 15 prominent human gut-derived bacteria whose genomes have been sequenced (**Table S1**). This 15-member community was used as bait for a staged attack that involved oral gavage of VLPs purified from human fecal samples. This system allowed us to document the capture of novel viruses present in the VLP preparations by members of the defined model human gut community, while at the same time tracking induction of native prophages.

Results

Our experimental design consisted of three groups of germ-free C57BL/6J mice (n=5/group). Each group was kept in a separate gnotobiotic isolator, where each mouse was individually caged. The first group of mice was gavaged with the 15-member bacterial consortium at 8 weeks of age; 3 weeks later they were each gavaged with a pool of VLPs isolated from five fecal samples obtained from five healthy humans (‘live VLP group’). A second group of mice (‘heat-killed VLP group’) was also colonized with the 15-member bacterial community but three weeks later received a

heat-killed version of the VLPs used in the first group. The third group did not receive a gavage of bacteria at 8 weeks of age ('germ-free group'); three weeks later they were gavaged with the same live VLP pool given to group 1. Fecal samples were collected from members of each treatment group at frequent intervals (**Fig. S1**). All mice in all groups were fed the same sterilized, low fat, plant polysaccharide-rich mouse chow *ad libitum*.

Neither the bacterial gavage nor the VLP inoculum contained components that appeared to compromise gut barrier/immune function or perturb overall health status. At the time of sacrifice, none of the treatment groups exhibited any significant differences in total body weight or adiposity (as judged by epididymal fat pad weight as a percentage of total body weight) ($p=0.3957$ and $p=0.4794$, respectively; Kruskal-Wallis test; **Table S2**). FACS analysis did not reveal any significant differences between the groups in the $CD4^+$ and $CD8^+$ T cell compartments of their spleens or mesenteric lymph nodes (MLN), as judged by CD44 and CD62L T-cell activation markers, Ki-67^+ (proliferation marker), and $FoxP3^+$ ($CD4^+$ Treg cells marker) (data not shown). For the germ-free group, Illumina shotgun sequencing (50 nt reads); of ileal and colonic contents obtained at the time of sacrifice revealed that the VLP inoculum did not contain bacteria taxa (or bacterial spores) that could establish themselves in the guts of recipient animals (**Table S3**).

Fecal microbial biomass (defined as ng DNA/mg wet weight of feces) increased linearly and abruptly in the four days following introduction of the 15-member community in the 'live VLP' and 'heat-killed VLP' groups (**Fig. S2a**). Fecal DNA concentrations correlated significantly ($R^2=0.837$) with the results of fecal bacterial cell counting by flow cytometry (**Fig. S2b**). Since the genome sequences of the 15 bacterial species were known, we used COmmunity PROfiling by Sequencing (COPRO-Seq) [6], a method based on short read (50nt) shotgun sequencing of total fecal community DNA, to quantify the relative abundance of each taxon as a function of time after initial colonization and after the staged VLP attack (2,544,433 ± 96,255 (mean ± SEM) reads/sample; **Table S3**). Principal coordinates analysis (PCoA) of a Hellinger distance matrix constructed from the COPRO-Seq datasets showed that most of the variation in composition over time occurred during the period of initial community assembly (**Fig. S3a**). Changes in the relative abundance of

community members also occurred following gavage of the live but not heat-killed VLP preparation (**Figs. 1** and **S4).**

To identify which exogenously administered VLP-associated viruses might be causing the observed structural rearrangements in community configuration, we modified our previously reported method for purifying VLPs [3] so that it could be applied to individual mouse fecal samples. We then sequenced DNA isolated from the purified VLP preparations [n=27; 2 fecal pellets/VLP preparation, each amplified by MDA; multiplex 454 FLX shotgun pyrosequencing (Titanium chemistry) yielded 49,819±6,983 reads/sample (mean± SEM); **Fig. S1**, **Table S4**]. To discriminate between activation of endogenous prophages in members of the 15-member community versus novel viruses derived from the exogenously administered VLP preparation, reads generated from the input human VLPs were mapped to the sequenced genomes of community members and to the mouse genome. We used reads that did not show a significant match to either dataset, together with COPRO-Seq reads from total mouse fecal DNA that did not map to any of the 15 bacterial genomes or to mouse DNA, to characterize viral genomes that were not represented in the starting 15-member community.

In total, five viral genomes that had not been described previously were assembled and annotated from these analyses. These viruses were detected in the gut communities of mice that had received the live VLP preparation but not in the heat-killed VLP group) (**Fig. 1, Fig. S5, Table S5**). A sequential and reproducible pattern of change in the abundances of these viruses was observed among individually caged mice harboring the 15-member community.

The first virus to significantly increase in abundance (ϕHSC01, for *H*uman *S*ynthetic *C*ommunity phage 01) was a 37 kbp circular DNA virus; beginning with the time of its first detection in feces 2-3 days after animals where gavaged with the live VLP preparation, its marked increase in abundance correlated with a decrease in the abundance of a component of the model community, *Bacteroides caccae*, (R^2= -0.446; p-value $3.2x10^{-8}$ after Bonferroni correction) (**Fig. 1b, d**). Prior these changes, *B. caccae* had represented 7.29±0.5% of the microbiota on the day before that VLPs

were administered. No other community member showed a statistically significant inverse correlation, suggesting that this bacterium was a host for ϕHSC01. The drop in relative abundance in *B. caccae* abundance was abrupt, decreasing 74.5±3.7% relative to pre-treatment levels within 1 d of gavage of the live VLP preparation and returning to 75.8±5.2% of the pre-VLP gavage values within 3-6 d (**Fig. 1**). This 4-fold decrease was independently validated using qPCR (from 3.4 x 10^5 to 7.8 x10^4 genome equivalents/mg of fecal pellet; data not shown). The spike in viral abundance, just like the coincident reduction in *B. caccae* abundance, was remarkably consistent in terms of their magnitude and time course among the individually caged members of the treatment group.

The ϕHSC01 viral genome not only encodes typical phage proteins (terminase, tail protein, DNA polymerase, helicase and methyltransferase; **Table S5A**), but also (i) a protein containing a Bacteroidetes-associated BACON domain (PF13004) characterized by a carbohydrate-binding module postulated to target glycoproteins and possibly host mucin [7], and (ii) a Helix-turn-helix (HTH) transcription regulator belonging to the MerR superfamily that is sensitive to stress responses (e.g., oxygen radicals [8]).

We used deep shotgun sequencing of total fecal community DNA isolated from samples obtained 9-19 d after bacterial gavage and 9-25 d after VLP gavage to compare the *B. caccae* genome before and after attack with live versus heat killed VLPs. Pooling the sequencing reads from these four groups of samples allowed us to assemble the *B. caccae* genome at an average coverage of 30X per treatment group, giving us enough resolution to identify mutations that could be responsible for conferring viral resistance to *B. caccae*. The results did not reveal deletions, insertions or SNPs that were unique to the live VLP treatment group after the viral gavage and fixed in more than 10% of the *B. caccae* population (See *Supplementary Discussion*). One interpretation of these results is that (i) the gut environment consists of a number of microhabitats, some of which are occupied by *B. caccae* in ways that make it inaccessible to viral attack, and (ii) the rise of *B. caccae* following the staged phage attack is not due to emerging resistance to the virus but rather is a manifestation of an expansion of an uninfected population after the virus is washed out of the ecosystem. An alternative, but not mutually exclusive, possibility is that resistance is acquired in

one or more regions of the genome that are difficult to sequence and/or assemble. Incorporation of short fragments of viral DNA within a locus flanked by short spacer repeats (CRISPR elements) leads to bacterial resistance to viruses whose genomes have sequence similarity to the incorporated sequences[9]. *B. caccae* does not contain any discernable CRISPR loci or associated proteins (**Table S1**). Moreover, no other prominent community members accumulated new spacers during the course of the experiment (see *Supplementary Discussion*).

The second virus to show an increase in its abundance, ϕHSC02, had dynamics that inversely correlated with the abundance of a prominent community member *Bacteroides ovatus* (abundance prior to these changes was 15.8±0.9%). Expansion of this virus and the attendant decrease in *B. ovatus* were first detected two days after the 'crash' of *B. caccae* (i.e., within 5 d of the live VLP gavage) and coincided with the onset of recovery of the *B. caccae* population (**Fig. 1b-e**). The absence of detectable reads mapping to this virus in feces collected just 10 days after VLP gavage, and the lack of reads from this assembled viral genome mapping to any of the other bacterial genomes in the 15-member community during the course of the experiment, provided evidence for its lack of integration; i.e., ϕHSC02 was lytic. All four predicted proteins encoded by the 6.2 kb ϕHSC02 phage (**Fig. 1e**) exhibit significant similarity to the Alpavirinae. The Alpavirinae are a recently described sub-family of Bacteroides phage [10] that belong to the Microviridae, a family composed of lytic single-stranded DNA viruses previously associated with Enterobacteriaceae (e.g., *Escherichia coli* PhiX174) and obligate intracellular pathogens (e.g., *Chlamydia* and *Mycoplasma* spp. phage [11]). The genomes of two members of the Alpavirinae [11] isolated from human feces have lengths similar to that of ϕHSC02 (6,251 and 6,171 nt) and share 85% nucleotide sequence similarity over more than 90% of their genomes with ϕHSC02. VP1, the major structural protein involved in host recognition, exhibited overall amino acid similarities of 85.9 and 84.3% in pairwise comparisons to the VP1 proteins of the other two phage; in contrast, the loop in VP1 predicted to be responsible for host recognition only had 67.5% and 67.1% similarity (**Fig. S6**) suggesting that ϕHSC02 may have a host specificity distinct from the other two family members.

The 6- to 8-fold reductions in relative representation of these two members of the model human gut microbiota (*B. caccae* and *B. ovatus*) was observed during the 7 day period after ga-

vage with live VLPs (days 20-23 and 23-27 of the experiment, respectively) and was followed by a return from their nadir to a level of abundance in feces comparable to those found in the control heat-killed VLP treatment group: this rise occurred over a 7-8 day period (**Fig. 1b,d**). As these organisms increased their abundance, we documented transient decreases in *Bacteroides cellulosilyticus* plus the two *Bacteroides thetaiotaomicron* strains present in the community. At the same time, levels of *Parabacteroides distasonis, Clostridium symbiosum, Clostridium scindens,* and *Ruminococcus obeum* rose. These changes were limited to the group of mice that had received the live VLP preparation: i.e., their abundances changed significantly when compared to heat-killed VLP group (**Fig. S4).** In contrast, levels of *Eubacterium rectale, Bacteroides uniformis*, and *Dorea longicatena* remained indistinguishable between the two treatment groups (**Fig. S4**).

The rise and fall of these organisms in the live VLP group occurred during a period when three other novel viruses appeared: ϕHSC03 (153.4 kb), ϕHSC04 (104.2 kb), ϕHSC05 (95.7 kb). These viruses, first detected 7 days after VLP gavage, subsequently increased in abundance in the fecal microbiota to approximately equivalent levels and persisted during the remaining 14 d of the experiment (**Fig. S5A-F**). Unlike the distinctive negative correlation between the rise and fall of ϕHSC01 and *B. caccae* abundances, and subsequently ϕHSC02 and *B. ovatus* abundances, the simultaneous appearance and rise of ϕHSC03, ϕHSC04 and ϕHSC04, their subsequent persistence, and the coincident complex patterns of change in abundances of bacterial community members of during this latter period of experiment made it difficult for us to assign candidate bacterial hosts for these three previously undescribed phages.

During the 25 d period when this reproducible, sequential pattern of appearance of members of the exogenous human fecal VLP preparation was documented (order, ϕHSC01➔ϕHSC02➔ϕHSC03/ ϕHSC04/ϕHSC05), fecal microbial biomass, as defined by fecal DNA levels, changed no more than 2-fold in members of the live VLP treatment group. Moreover, biomass was not significantly different between recipients of live versus heat-killed VLP preparations at comparable time points in the experiment (**Fig. S2A-**Wilcoxon matched-pairs signed rank test; p-value = 0.459). These observations highlight another facet of the resiliency of the gut microbiota.

We next used our gnotobiotic mouse model to address the question of whether similar virotypes represented across different humans may have shared or distinct bacterial host specificities. To determine whether ϕHSC01, ϕHSC02, ϕHSC03, ϕHSC04, and ϕHSC05 were distributed among all of the human VLP donors or whether they were unique to particular individuals, we generated a hybrid-assembly using reads from the original VLP-derived viromes from each of the five donors as well as from the pooled VLP preparation used for gavage (see *Methods*). The hybrid assembly yielded 159 contigs greater than 2 kbp. Most contigs were derived from viruses present in single donors: four of the five viruses described above were from VLPs that originated from a fecal sample obtained from a co-twin in family 2 (F2T1.2) or family 4 (F4T1.2) (**Fig. 3**). However, one of the novel viruses, ϕHSC05, was observed in 4 of the 5 individuals used to construct the VLP pool. Mapping reads from each original human donor fecal virome to ϕHSC005 revealed that the virus recovered from the mice likely comes from individual F3T1.2 since the average percent identity (%ID) of reads from this person's virome mapping to the novel viral genome is equivalent to the percent identity obtained from VLPs isolated from mouse fecal samples [for human VLP donor F3T1.2, 87.5% of the reads with 98.45±0.02 (mean ±SEM) %ID mapped to the assembled phage; for VLPs isolated from mouse recipients of the pooled human VLP preparation at the terminal time point, 26.4% mapped reads with 98.6±0.01%ID]. In the case of the other individuals, the novel virotype detected in mice was present at significantly lower abundance (**Fig. 3**) and/or had lower percent sequence identity (for human VLP donor F2M.2, 20.2% mapped reads with 97.91±0.03%ID; F2T1.2, 72.5% mapped reads with 96.95±0.01%ID; F4T2.2, 27.2% mapped reads with 96.40±0.02%ID; F4T1.2, 0.01% mapped reads with 98.25±0.65%ID). Thus, while the nucleotide sequence similarity is very high for this virus across the five different humans used to create the pooled fecal VLP preparation, the host specificity of the different virotypes may not be the same, with resulting selection for only one in the defined model human gut microbiota. These results not only illustrate the utility of gnotobiotic mice for identifying candidate bacterial hosts for human gut-associated phage, but also for identifying differences in the properties among related virotypes derived from different human donor gut viromes.

Another question that can be addressed in a gnotobiotic animal model that is not readily determined in humans is whether the time course and magnitude of change in bacterial abundance observed in feces is representative of changes that occur along the length of the gut: i.e., do changes in the abundance of phage and their host bacterial cells occur coincidently throughout the gut, or is there a critical bio-geographical component of some of these interactions; does a phage attack limited to the proximal intestine produce a rapid and dramatic reduction in the representation of its bacterial host, followed by recovery as it is transported towards the distal colon? To address these questions, we collected intestinal contents from the proximal and distal small intestine, cecum and colon, as well as a fecal sample from all mice in the live and heat-killed treatment groups at the time of their sacrifice (25 d after the staged VLP attack). All gut samples were processed for COPRO-Seq analysis, while an aliquot of cecal contents was used to isolate VLPs for subsequent shotgun sequencing of viral DNA (**Table S4**). The results revealed no detectable phage in any of the segments in any of the mice that received the heat-killed VLP preparation (**Fig. S7a**). In the live VLP treatment group, neither ϕHSC03, ϕHSC04, nor ϕHSC05 exhibited significant differences in their relative abundances between the distal small intestine and distal colon, and between luminal contents and feces (**Fig. S7a**). Moreover, at the time of sacrifice, there were no significant biogeographic differences in the relative abundance of bacterial species within members of a treatment group or between the two treatment groups (**Fig. S7b**). Thus, at least at the time of sacrifice, feces portrayed the proportionality of these three viral-bacterial community relationships in a manner representative of more proximal regions of the gut.

Intestinal transit time in the mouse is in the order of several hours [12]. The fact that ϕHSC03, ϕHSC04 and ϕHSC05 first appeared in members of the live VLP treatment group 7 days after the single gavage of pooled human fecal VLPs suggests that an intra- and/or extracellular compartment exists that harbors components of the administered human fecal phage population. Pseudolysogeny, a state where phages exist in a host bacterial cell without multiplying or synchronizing their replication with the host [13] could represent one potential mechanism for persistence. Extracellular sequestration, for example through binding of carbohydrate binding modules, present in viral

capsular proteins [14], to mucus or epithelial cell surface glycans represents another potential mechanism for persistence. COPRO-Seq analysis of fecal samples obtained from mice in the germ-free treatment group that lacked the 15-member bacterial community and were gavaged with the live VLP preparation alone revealed no detectable VLPs at any time point, including in cecal samples harvested at the end of the experiment (**Table S4**). Therefore, we were able to conclude that persistence of ϕHSC03, ϕHSC04 and ϕHSC05 was dependent upon the presence of a bacterial cellular component.

Identification of prophage activation in the 15-member model community – The classical lysogenic state consists of a phage integrated in a host bacterial chromosome as a prophage. Prophage are thought to provide super-infection protection from other viruses that use the same or similar receptors as the integrated virus on the surface of their bacterial hosts [15,16]. Thirteen of the 15 bacterial species in the model human gut microbiota had predicted prophage in their genomes (**Table S1**). To verify these predictions and assess the capacity of these prophage to undergo induction to lytic phage, we used reads obtained from shotgun pyrosequencing of DNA isolated from two sources; (i) VLPs purified from fecal samples collected at weekly intervals from animals gavaged with the pooled live human VLP preparation, and (ii) VLPs purified from cecal samples obtained at sacrifice. Instead of mapping randomly throughout bacterial genomes (implying a background level of bacterial DNA contamination in the purified VLPs), the VLP reads mapped to one or more of the predicted prophage. In total, we identified 10 prophages derived from seven bacterial genomes that had the capacity to undergo induction *in vivo* (**Fig. S8**).

B. cellulosilyticus WH2 has two prophages, one of which (prophage 1) exhibited the greatest fold induction among these 10 prophages. Induction was observed in all mice 5-9 d after initial gavage of the 15-member consortium prior to introduction of either live or heat-killed VLPs at day 20 of the experiment. Induction occurred at the end of the period of initial bacterial community assembly, right after microbial biomass reached its peak (**Fig. S2a**), suggesting a potential role of bacterial density in the induction process. Induction correlated with a decrease in the relative abundance of its bacterial host (**Fig. 3a,b**). There was specificity: the other *B. cellulosilyticus* prophage

did not exhibit significant levels of induction at any time point surveyed during the experiment (**Fig. 3c; Fig. S8**).

The inducible 45 kbp *B. cellulosilyticus* WH2 prophage 1 has traditional a lambdoid genome architecture. Mapping VLP reads to the bacterial genome allowed us to identify the prophage insertion site at an arg-tRNA gene, with the corresponding duplicated region generating the attachment (*att*) sites (**Fig. 3e**). Lambdoid replication machinery generates concatemers of the genome that are linearized at *cos* sites before being packed inside the viral capsid. As expected from sequences obtained from VLP-derived DNA, no reads were obtained spanning the region containing the potential *cos* sites (**Fig. 3e**), highlighting the quality of the VLP purification.

Starting from the *cos* sites, the prophage 1 genome contains a group of genes that are conserved and organized in a way common to lambdoid viruses of the Siphoviridae family [17]; there was a *cos* site, followed by small and large terminases, portal protein, protease, the major head protein, several small capsid structural proteins, the major tail protein, tape measure protein, and ending with the integrase next to the *att* site. Searching the genomes of 38 human gut *Bacteroides* (**Table S6**), we identified four other species, including another strain of *B. cellulosilyticus* (DSM 14838) that contained homologs of all or most of the tail and capsid proteins in a syntenic arrangement (**Fig. 3d**). Moreover, homologs of four hypothetical genes in *B. cellulosilyticus* prophage 1 are present in 35-37 of the 38 available *Bacteroides* genomes.

The *B. cellulosilyticus* WH2 population in the 15-member community was represented by a library of 26,750 isogenic mutants with each mutant strain containing a single randomly inserted modified mariner transposon (Tn) in the bacterial genome (78.8% of predicted ORFs contain insertions covering the first 80% of each gene; average of 5.3 insertions/ORF). Because the modified Tn had engineered recognition sites for the type II restriction endonuclease MmeI at its ends, 16 nt of flanking chromosomal DNA could be excised together with the Tn after MmeI digestion of community DNA and sequenced, permitting the location and abundance of each transposon mutant in the library to be determined. Comparing the number of reads for each mutant in the

'output' population subjected to a given selection to the number of reads generated from the 'input' population provides information about the effect each transposon insertion had on the fitness of the organism under the selection condition applied [18,19].

Tn insertion sequencing (INSeq) analysis of DNA prepared from fecal samples collected before, during and after prophage 1 induction showed a dramatic enrichment for transposons located within a ~600 bp intergenic region positioned between the ORFs encoding its putative *rha* protein [20] and *cI* repressor at the time of phage induction 5 -9 days after introduction of the 15-member bacterial community) (**Fig. 4a**). This enrichment did not reflect clonal expansion of a single mutant strain within a given animal, but rather expansion of one or more of 10 independent mutants, each harboring a single transposon insertion within this intergenic region. The number and sites of these insertion mutants varied between animals (**Fig. 4a**). Moreover, no Tn insertions were observed within the ORF encoding the putative cI repressor in the input library, nor in any of the output fecal samples (**Fig. 4a**), suggesting an essential role for the repressor in bacterial host fitness.

We subsequently defined the time course of clonal expansion of bacteria containing Tn insertions in this intergenic region positioned between cI and rha to quantify the fitness effects of disrupting this part of the prophage genome. Reads mapping to this intragenic locus represented 77.9±4.6% (mean ± SEM) of all Tn reads in fecal samples collected between 5 and 9 days after bacterial gavage (range, 61.3-96.9%). This result contrasts with 0.0059% (mean value) of all Tn reads in the input library, 0.025% in samples obtained 3 days after gavage (before prophage induction had occurred), and an average of 0.0136% for any other gene in the *B. cellulosilyticus* genome throughout the experiment.

We subsequently performed a sliding window analysis to determine if any other 600bp region of the *B. cellulosilyticus* WH2 genome containing a Tn insertion went through a clonal expansion analogous to that documented for the cI-rha intergenic region during the first 31 days of the experiment in mice belonging to the live and heat-killed VLP treatment groups (**Fig. 4b**). The results revealed that on average any given 600 bp window with a Tn decreased its abundance over

time, usually to less than 0.001% of the *B. cellulosilyticus* WH2 population. Only 5-10% of the windows other than the cI-rha intergenic region exhibited any enrichment over time with less than 0.1% of the windows reaching levels >1% of the population. However, all of these other enrichments occurred 11 days or more following gavage when the bacterial host population was recovering from prophage induction (**Fig. 4b**). (COPROSeq analysis indicated that the relative abundance of *B. cellulosilyticus* WH2 in the community fell from 18.1±0.23% on day 3 to 12.2±1.68% on day 7, and recovered to 17.9±0.31% on day 13, **Fig. 3a,b**). Importantly, strains with the Tn-containing cI-rha intergenic region selected for during prophage 1 induction subsequently maintained high relative abundance (~4%) in the *B. cellulosilyticus* population in both the live VLP and heat-killed VLP treatment groups (**Fig. 4b**).

Together, these results indicate that prophage 1 induction is restricted in time (i.e., non-recurring over the course of the experiment), insensitive to attack of other members of the defined human gut microbiota by exogenous human fecal phage, and does not affect the long-term fitness of its *B. cellulosilyticus* WH2 host (as judged by its constant relative abundance in the community following the period of induction). One interpretation of our findings is that reads mapping to the cI-rha intergenic region originate *only* from induced viral particles. The INSeq mutant library was constructed so that there was only one Tn insertion per genome. Thus, in this first scenario, insertions in the intergenic region promote induction, with the ratio of phage genomes to prophage-containing *B. cellulosilyticus* genomes achieving values up to 95:5. However, this interpretation seems unlikely; the results shown in **Fig. 4** reveal that high numbers of Tn reads from the cI-rha intergenic region persist even after prophage induction was completed, indicating their origins from a prophage integrated into the host bacterial genome. Another interpretation that we favor is that the reads from the Tn-containing cI-rha intergenic region originated from uninduced prophage in host cells resistant to phage infection. “Suicide bomber” is a term describing a survival mechanism for bacterial cells: after a danger signal such as DNA damage is detected, native prophage are induced with resulting destruction of the host cell and production of viral particles; these phage subsequently infect other sensitive bacterial hosts except those harboring the prophage which are

protected through super-infection immunity. In this interpretation of our data, expansion of mutant strains with Tn insertions in the intergenic region between cI and rha, but not within either of these two genes, reflects their ability to sense a danger signal but their inability to induce the prophage. This renders them resistant to the phage attack that affects other members of the mutant population, allowing their rapid expansion into the emptied niche.

Prospectus

Our results illustrate how gnotobiotic mice containing defined consortia of sequenced human gut bacterial symbionts provide a tractable system for characterizing phage-bacterial host dynamics for those seeking basic principles that shape the configuration adaptations, and resiliency of the microbiota, as well as those who wish to develop new diagnostic and therapeutic approaches, including phage therapy. This model allows complex mixtures of VLPs, isolated from previously frozen fecal samples obtained from human donors representing ages, physiologic or disease states, or geographic regions/cultural traditions of interest, introduce them into mice harboring a model defined human gut microbiota community, and use the microbiota as 'filter' to identify and assemble the genomes of previously unknown (or known) phages present in the human donor viromes and link them to bacterial hosts present within the model community. Miniaturization of methods for preparing VLPs from single mouse fecal pellets collected over time provided a way for purifying these phages as they appear in an ordered sequence in the model community and at the same time verifying that they have lytic activity. The system has 'forensic' capabilities, allowing distinction of very closely related virotypes present in multiple human gut microbiota based on their differential ability to establish themselves in recipient gnotobiotic mice. These capacities not only provide a discovery pipeline that complements metagenomic surveys of the human gut virome by identifying phage 'buried' in large gut microbiome datasets, but facilitate identification of phage that can be used as experimental tools to deliberately manipulate model microbial communities, as well as new candidate therapeutic agents. These gnotobiotic models, and the associated experimental and computational approaches described in this report, provide an opportunity

to achieve deeper understanding of what a temperate viral-bacterial host dynamic means in the gut ecosystem. For example, our data reveal that lytic attacks can have a short effective time frame and suggest that determinants of the success, duration and effects of an attack go beyond the presence of a particular phage and its bacterial host and even modification of the host bacterial cells genome. Induction of prophage present in the genomes of members of the model community can be differentiated in these models from effects of introduction of and attack by (exogenous) lytic phages, while prophage-associated fitness determinants can be identified by whole genome transposon mutagenesis of their bacterial hosts.

Acknowledgements

We thank Dave O'Donnell and Maria Karlsson for their assistance with gnotobiotic mouse husbandry, Philip Ahern for help with immune characterization of tissues, Martin Meier for technical support with robotics, plus members of the Gordon label for their helpful suggestions during the course of this study. This work was supported by grants from the NIH (DK30292, DK78669) and the Crohn's and Colitis Foundation of America. A.R. is the recipient of an International Fulbright Science and Technology Program award.

Figure Legends

Fig. 1 – Sequential changes in the relative abundance of two members of the model human gut microbiota and correlation with the appearance of two novel phages. (**a**) Average relative abundance plot for each bacterial species as a function of time for either the 'Live VLP' or the heat-killed VLP treatment groups. The color key next to the plot indicates the identity of the bacterial species. **(b,c)** Plots of the relative abundance (fraction of the total community; mean ± SEM; n=5 animals/treatment group) of *B. caccae* and *B. ovatus* in the fecal microbiota of gnotobiotic mice as a function of time prior to and after gavage with live purified VLPs pooled from the fecal microbiota of five human donors, or a control heat-killed version of the same VLP preparation (time of gavage indicated by the upward pointing arrow; t=0 refers to the time of introduction of the 15-member consortium of sequenced human gut bacterial taxa into germ-free animals). The change in abundance of these *Bacteroides* occurs in a reproducible sequence among individually caged mice that received live but not heat-killed VLPs. (**d,e**) Changes in the abundance of two phages, derived from the human donor VLP sample, in the fecal microbiota of recipient gnotobiotic mice. Viral abundance negatively correlates with bacterial abundance in the group that received live but not heat-killed VLPs. Differences in the time course of change in bacterial and viral abundances are highlighted by the light green and yellow, leading us to propose that ϕHSC01 targets *B. caccae* as its host while *B. ovatus* serves as host to ϕHSC02. Insets in panels d and e are assembled genome sequences for ϕHSC01 and ϕHSC02. The location of genes in the positive strand (green) and negative strand (red) strand are shown; those that have significant sequence similarity to to known viral genes are colored blue (blast E value < 10-5; **Table S5**). The inner plot represents GC skew based on 200 bp windows (yellow, G/C ratio is greater than the average for the genome; purple, ratio is lower than the average).

Fig. 2 - Heat map of the cross-assembly between different input human VLP samples. To determine whether the viruses identified in gnotobiotic mice were distributed among all of five human VLP donors or whether they were unique to particular individuals, a cross-assembly was generated that included (i) reads from each fecal VLP-derived virome from each individual human

fecal sample used to generate the input pool of VLPs for the staged viral attack, (ii) reads from the pooled human VLPs introduced into mice by gavage, and (iii) reads generated from VLPs purified from mouse fecal samples. The cross assembly pipeline consisted of the following steps (in order): (i) normalizing coverage by clustering the reads at 95% global identity (CD-HIT) and picking up to 5 representatives per cluster, (ii) *de novo* assembly of the representative read set, (iii) mapping all raw reads to the resulting contigs, (iv) *de novo* assembly of leftover reads, (v) merging the assemblies from (ii) and (iv) and extending them by using the MIRA assembler to map back all raw reads, (vi) employing Phrap to consolidate the assembly, (vii) mapping reads back to the assembly and checking for chimeras, and (viii) checking for contig redundancy, overlapping ends and circular contigs using BLAST. Shown are the normalized abundances of the different contigs (n=159) greater than 2 kbp (rows), including those assembled from the human phage identified in gnotobiotic mice. Each column presents data for the five individual human-donor fecal VLP-derived viromes plus the pooled VLP sample introduced into gnotobiotic mice. (FX, family number; M, mother of twin pair; TX, co-twin 1 or 2 in a given monozygotic twin pair; .X, one of 3 time points where feces were collected from the human donor over the course of a year). Abundance is shown as the log(10) transformation of RPKM (reads per kb of contig per million 454 reads; see *Supplemental Methods* and main text for further details). Rows representing the five novel viruses isolated, and their corresponding human donor are enlarged for visualization purposes.

Fig. 3 - Prophage induction in *B. cellulosilyticus* WH2. (**a,b)** COPRO-Seq analysis of the relative abundance (mean ± SEM) of the bacterial host and its associated prophage 1 in the fecal microbiota of individually caged mice partitioned into two separate gnotobiotic isolators (one where animals subsequently received live human fecal derived VLPs and the other heat-killed VLPs on day 20). Relative abundance was measured independently based on reads mapping to prophages and reads mapping to all other regions of the bacterial genome. Equivalent abundances for these reads indicate that the phage is in an uninduced lysogenic state; increases in the proportional number of reads mapping to the prophage genome indicate phage induction. Prophage 1 induction and a concomitant decrease in the relative abundance of its bacterial host occurred in mice five days

after introduction of the 15-member bacterial community, after the exponential rise in community biomass had ceased (see **Fig. S2a**), but before human VLPs were administered. (**c**) Data for *B. cellulosilyticus* prophage 2 showing its lack of induction in the live VLP treatment group. (**d**) A subset of genes in prophage 1 that are conserved in the genomes of four other human gut Bacteroides. Only genes conserved in synteny with *B. cellulosilyticus* WH2 prophage 1 are shown together with the average protein similarity for their protein products. Each box represents a predicted ORF. See panel e for the color code for ORF annotations. (**e)** VLP-derived 454 pyrosequencing reads from fecal samples obtained from mice in the group that received the live VLP pool, as well as from mice that received the heat-killed control, mapped to a 150 kbp fragment of the *B. cellulosilyticus* WH2 genome containing one of its two prophage (prophage 1). The *y*-axis corresponds to the log(10) of the read coverage (blue) for a given position along the prophage genome. Note that no reads were obtained in the region containing the potential *cos* sites (downward pointing red arrow) emphasizing the quality of the VLP purification procedure (see main text). Colors represent the different products of ORFs that are characteristic of this and other lambdoid phage. The intergenic region positioned between ORFs encoding the rha protein and cI (the lambda repressor) is highlighted; strains with transposon insertions in this region dramatically increase their representation after prophage 1 induction.

Fig. 4 - INSeq reveals clonal selection for Tn mutants with insertions in the region between genes encoding the rha protein and cI homolog of prophage 1 in *B. cellulosilyticus* WH2. (a) Heatmap of log-transformed, normalized abundance of reads (reads per million, RPM) that mapped to a site of transposon insertion in the BACWH2_5233 gene (rha homolog) or the intergenic region between the rha homolog and the BACWH2_5232 gene (cI homolog). Each row represents a time point 3-9 days after bacterial gavage with the 15-member model human gut microbiota (prior to subsequent gavage with either live or heat-killed VLPs). All mice were individually caged. Fecal samples collected from five mice in each treatment group were characterized (M1-M5 in the live VLP group; M6-M10 in the heat-killed VLP control group). Each column represents the total number of normalized reads, in a window of 20 bp, obtained with a given fecal DNA sample.

Tick marks between nucleotides 7,078,320 and 7080,520 of the *B. cellulosilyticus* WH2 genome represent 100 bp increments. The bottom four rows labeled 'input' represent the read distribution for four technical replicates of the INSeq analysis for the input *B. cellulosilyticus* mutant library. There is no change on the representation of Tn insertion mutations in either of the genes in the output library. In contrast, a marked increase in the representation of Tn insertion at 10 sites in the intergenic region is seen over the time period sampled. The number and abundance of Tn mutants represented in each mouse differ. The increase in representation of Tn mutants correlates with the time of induction of prophage 1 (see Fig. 2 for induction levels. The 600bp window referred to as the cI-rha intergenic region in the main text and **panel b** is shown as a thin green line in the Figure. Note that no Tn mutants were identified in BACWH2_5232 (cI homolog). **(b)** Insertions in a 600bp intergenic region downstream of the putative cI regulator in *B. cellulosilyticus* WH2 prophage 1 provides a fitness advantage to the host that is maintained over time. The number of Tn reads per million (RPM) obtained by INSeq analysis of fecal DNA was calculated for a 600bp intergenic region between the cI regulator and the putative rha protein. The relative abundance of bacterial cells harboring this Tn insertion as a fraction of the total *B. cellulosilyticus* WH2 population is represented by the RPM value at any given time point. Mean values ± SEM for the mice in each treatment group are plotted for 13 different time points sampled during the first 31 days of the experiment. Live VLPs (panel A) or heat killed VLPs (panel B) were introduced at day 20 (downward arrow). A sliding window of 600bp was used, at intervals of 100bp, to scan the whole *B. cellulosilyticus* WH2 genome (except the 600 bp cI-rha intergenic region) in order to quantify the abundance of mutants containing Tn insertions. The resulting distribution of reads is plotted for any given time point with lines at the 25^{th}, 50^{th}, 75^{th} , 99^{th} and 100^{th} quantiles (100^{th} = maximum value of Tn reads observed within a 600 bp window). Shaded areas represent the area where 25-75% of the data (RPM/600 bp window) falls (i.e., the second and third quartiles). In general, the abundance of Tn mutations falls over time with less than 10% of the Tn-containing windows increasing in their abundance. The relative abundance of bacterial strains harboring Tn-containing cI-rha intergenic mutants was maintained at significantly higher levels than all other mutants, indicating its importance to the fitness of their bacterial hosts.

Figures

Fig 1.

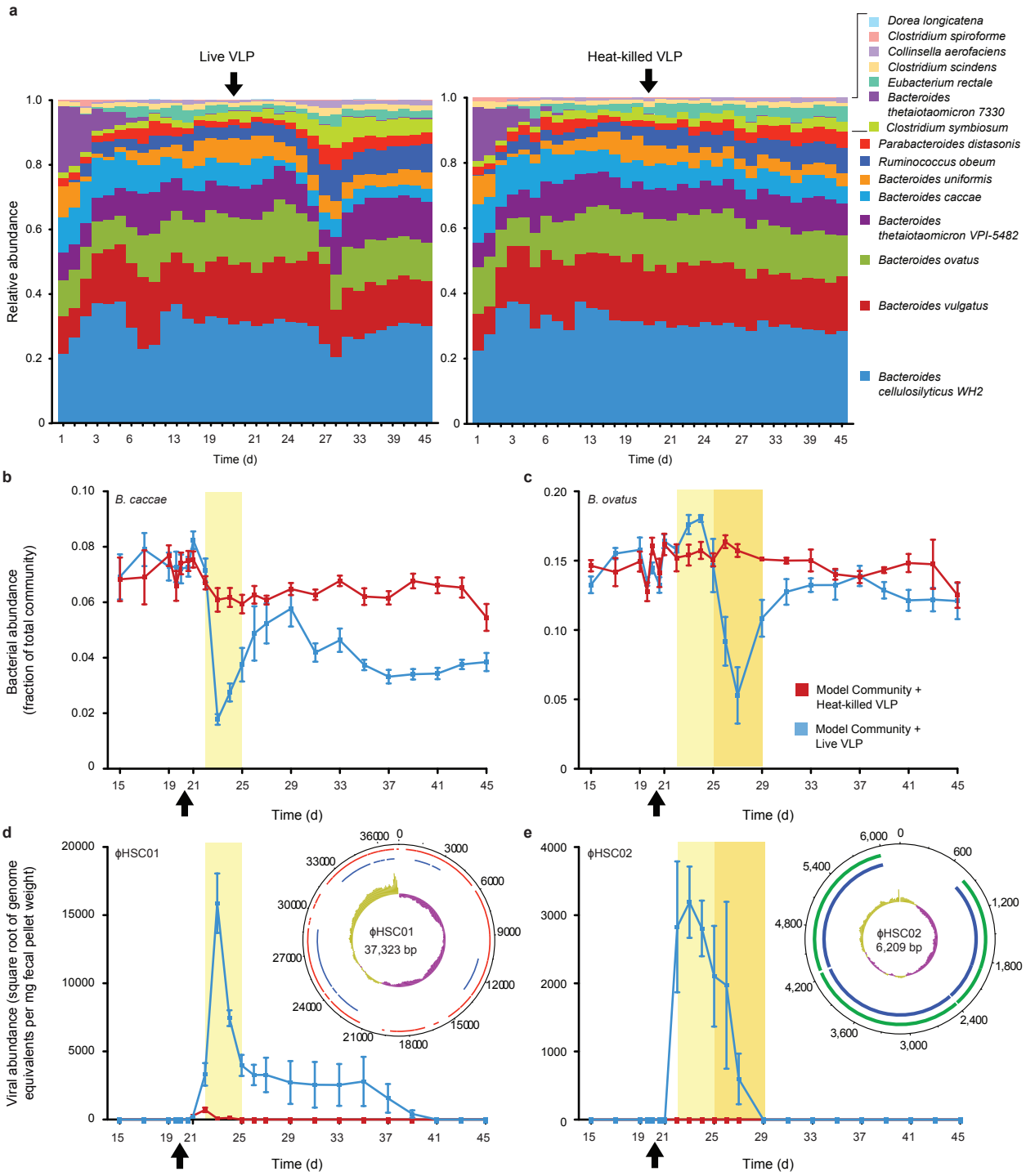


Fig. 2.

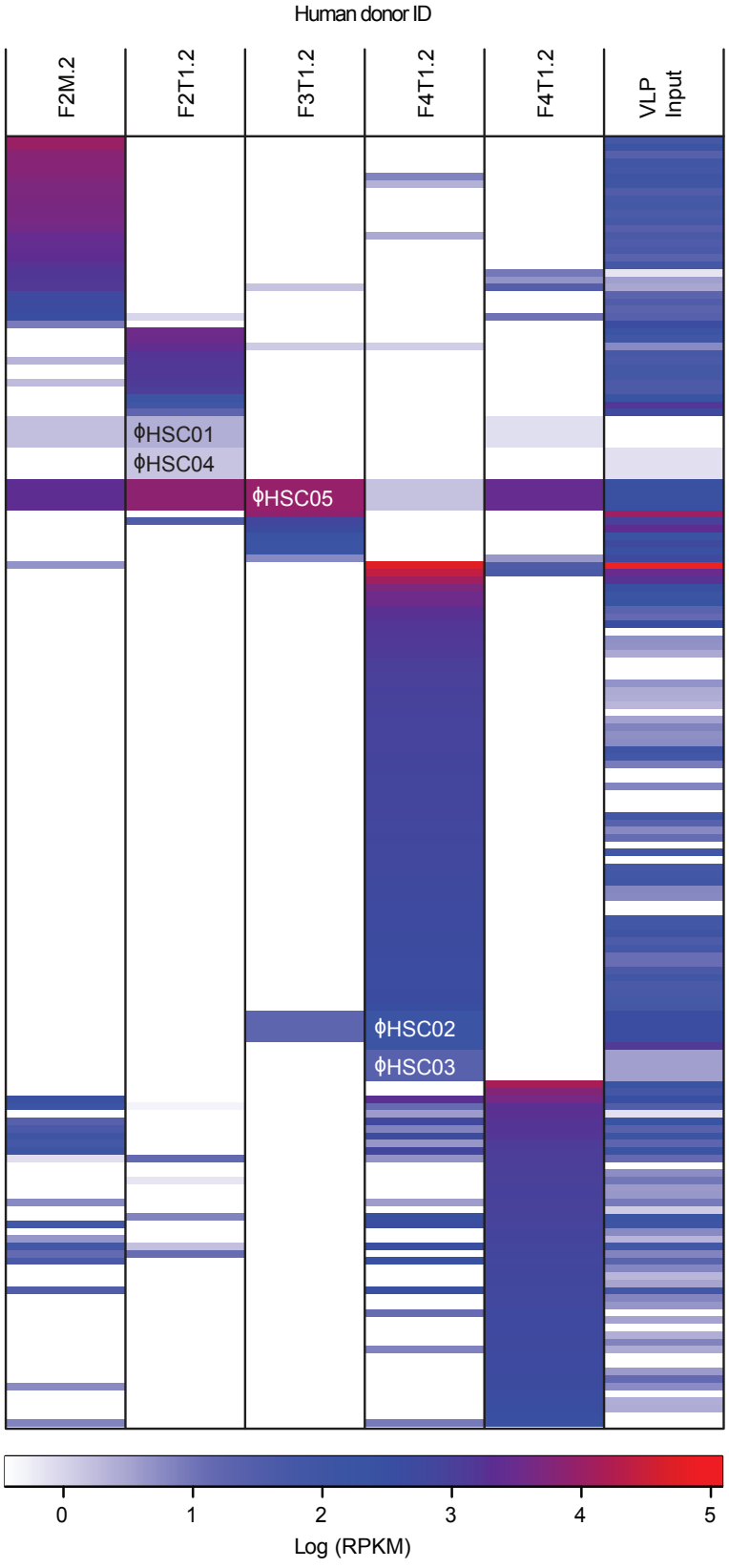

Human donor ID
F2M.2
F2T1.2
F3T1.2
F4T1.2
F4T1.2
VLP Input
ϕHSC01
ϕHSC04
ϕHSC05
ϕHSC02
ϕHSC03
0
1
2
3
4
5
Log (RPKM)

Fig. 3.

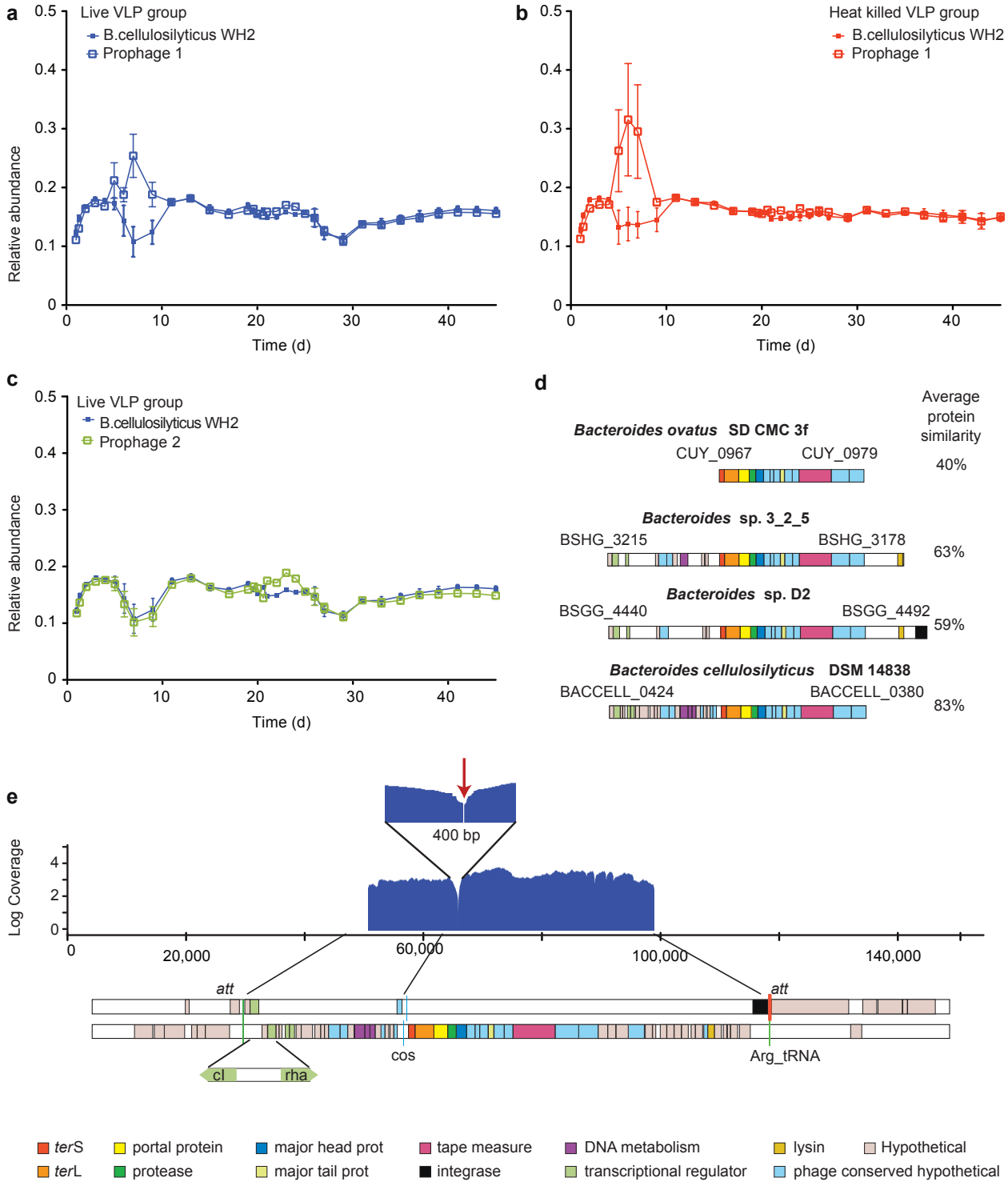
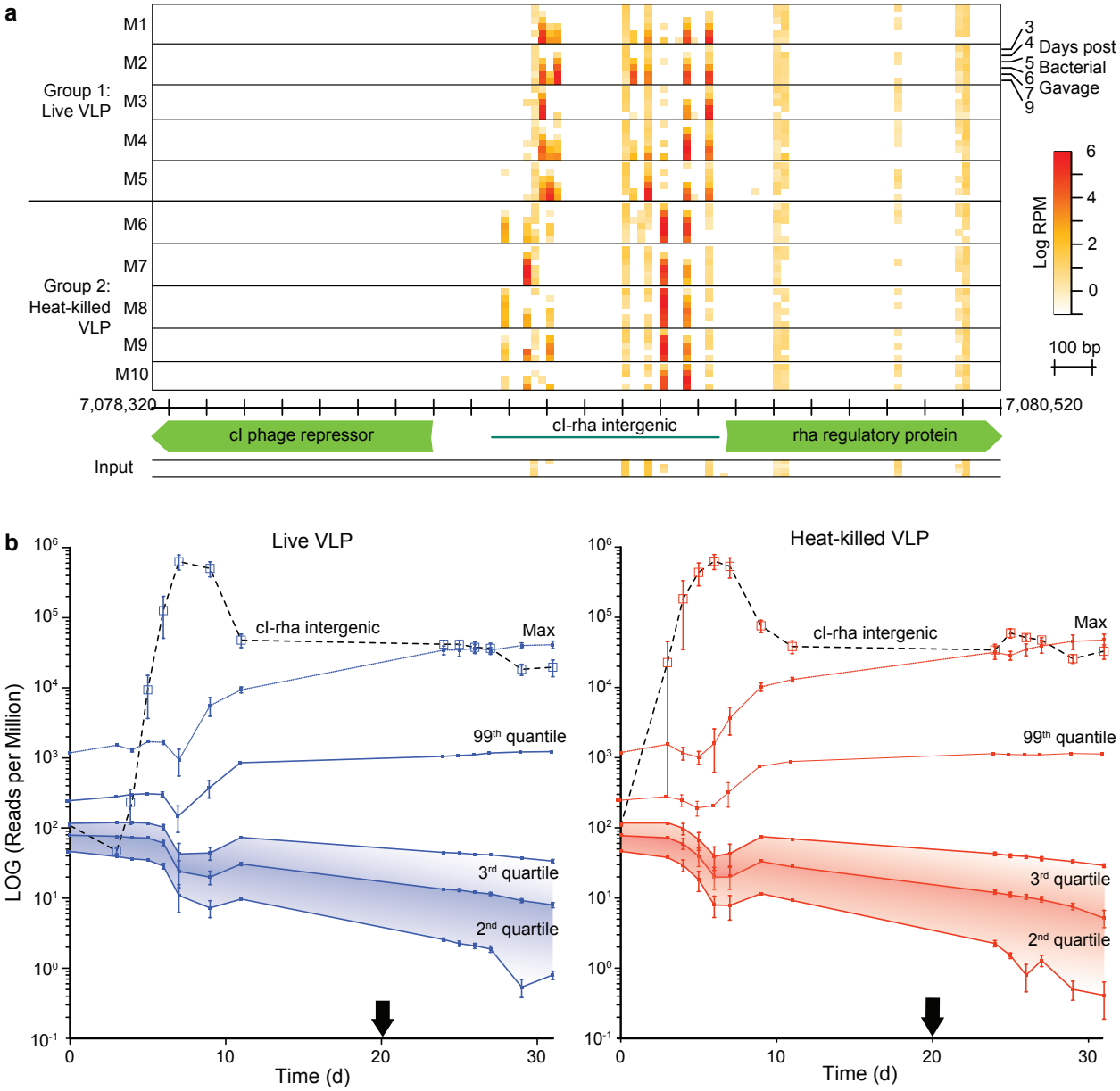

a
Live VLP group
B.cellulosilyticus WH2
Prophage 1
Relative abundance
Time (d)
b
Heat killed VLP group
B.cellulosilyticus WH2
Prophage 1
Time (d)
c
Live VLP group
B.cellulosilyticus WH2
Prophage 2
Relative abundance
Time (d)
d
Average protein similarity
Bacteroides ovatus SD CMC 3f
CUY_0967
CUY_0979
40%
Bacteroides sp. 3_2_5
BSHG_3215
BSHG_3178
63%
Bacteroides sp. D2
BSGG_4440
BSGG_4492
59%
Bacteroides cellulosilyticus DSM 14838
BACCELL_0424
BACCELL_0380
83%
e
Log Coverage
400 bp
att
att
cos
Arg_tRNA
cI
rha
terS
terL
portal protein
protease
major head prot
major tail prot
tape measure
integrase
DNA metabolism
transcriptional regulator
lysin
phage conserved hypothetical
Hypothetical

Fig. 4.


a
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
Group 1: Live VLP
Group 2: Heat-killed VLP
3
4
5
6
7
9
Days post Bacterial Gavage
Log RPM
6
4
2
0
100 bp
7,078,320
7,080,520
cI phage repressor
cI-rha intergenic
rha regulatory protein
Input
b
Live VLP
Heat-killed VLP
LOG (Reads per Million)
cI-rha intergenic
Max
99th quantile
3rd quartile
2nd quartile
Time (d)

Methods

Gnotobiotic mouse husbandry

All experiments involving mice were performed with protocols approved by the Washington University Animal Studies Committee. Germ-free mice belonging to the C57BL/6J inbred strain were maintained in plastic gnotobiotic isolators under a strict 12 h light/dark cycle and fed a standard, autoclaved low-fat/plant polysaccharide-rich chow diet (B&K Universal) *ad libitum*. Three groups of age-matched germ-free animals (n=5 per group) were kept individually caged in three separate isolators. At ~8 weeks of age, mice from two groups were colonized with a single gavage of 300 µL of supplemented TYG medium [21] containing an overnight fresh culture of 15 sequenced human gut-derived bacterial species (~$6x10^8$ CFU per strain). One of the 15 strains, *B. cellulosilyticus* WH2, was represented as a library of >25,000 isogenic transposon mutants (see INSeq analysis below).

Introduction of VLPs purified from human fecal samples into gnotobiotic mice

Five frozen de-identified fecal samples from healthy adult humans were selected for isolation of VLPs. These fecal samples had been collected previously as part of a study of the viromes of four healthy adult monozygotic twin pairs and their mothers. Recruitment and sampling of these individuals over the course of one year had been performed using a protocol approved by the Washington University HRPO [3].

We chose those fecal samples from co-twins in three families whose viromes together encompassed the broadest range of diversity present among the 32 samples in our collection [i.e., F3T1.2, F4T1.2, F2T1.2, F2M.2, F4T2.2; (F#, family number; T#, co-twin identifier; .X, time point)]. A 2-3 g aliquot of each frozen sample was re-suspended by vortexing in 10 mL of SM buffer (100mM NaCl, 8mM $MgSO_4$, 50mM Tris (pH 7.5) and 0.002% gelatin (w/v), sterilized by 0.02 µm filtration (Anotop 10, Whatman, Germany). The suspension was centrifuged at 2500 x *g* at 4°C for 10 min to pellet large particles and bacteria. The supernatant was then passed twice

through a 0.45 μm-diameter Sterivix-HV filters (Millipore, MA). A 20 μL aliquot of a 1:10 dilution of the resulting filtrate containing VLPs was used to confirm the absence of bacterial cells and for viral particle enumeration using SYBR Gold and fluorescence microscopy ($3.77 \pm 2.79 \times 10^9$ VLP/g frozen fecal sample; n= 5 samples). Aliquots (1.2 mL) from each of the 5 VLP preparations were pooled and the remaining sample stored at 4°C in 0.1 volumes of chloroform. The pooled VLPs were split into three 2 mL aliquots. One of the aliquots was further subdivided into 100 μL volumes and heat-killed by incubation at 95°C for 15 min followed by DNase treatment for 1 h [10 units of DNase Baseline Zero (Epicentre, WI)].

Three weeks after bacterial gavage, mice were gavaged with the pooled, human-derived VLPs. To ensure that the gastric pH would not affect the viability of the viral particles, mice were fasted for 12 h. Each mouse was then gavaged with 100 μL of 1M $NaHCO_3$ to keep gastric pH at 7.4 [22] followed 10 min later by 300 μL of the VLP pool (3.28×10^8 VLPs, either live or heat-killed, per animal).

Sampling the fecal microbiota of gnotobiotic mice

Fecal samples were collected at the time points shown in **Fig. S1**. Fecal samples dedicated to VLP purification were collected at least once per week. Samples were placed in 1.7 mL screw cap tubes immediately after they were produced by the animal and stored at -80°C until further processing. After mice were sacrificed, the small intestine was subdivided into two segments of equal length. Contents from the proximal and distal small intestinal segments, plus contents obtained from the cecum and colon, were snap-frozen in liquid nitrogen and stored at -80°C.

Preparation of VLP DNA from mouse fecal samples

VLP purification and DNA extraction were performed as described previously [3], with some modifications. Since a single mouse fecal pellet had too little viral mass for efficient viral purification, pairs of fecal pellets obtained from either one or two mice in a given treatment group at the same time point (30-100 mg) were resuspended in 400 μL of SM buffer [filter-sterilized, 0.02 μm pore

diameter (Whatman, Germany)]. After homogenization by vortexing for 5 min, samples were centrifuged twice at 2,500 x *g* for 10 min at 4°C to remove large particles and bacterial cells. The resulting supernatant was filtered once through a 0.45 µm Millex filter (Millipore) and twice through 0.22 µm Millex filters (Millipore). The volume of the filtrate was adjusted to 200 µL with SM buffer if needed. Each sample was treated with 20 µL of lyzosyme (100 mg/ml) for 30 min at 37°C, followed by incubation for 10 min with 0.2 volumes of chloroform. The sample was then centrifuged at 2,500 x *g* for 5 min at room temperature. The aqueous phase was collected and incubated with 3 U of DNaseI (Sigma) and 20 µL of 10X DNase buffer (50mM $MgCl_2$, 10mM $CaCl_2$*)* for 1 h at 37°C, after which enzyme activity was inactivated by incubation at 65°C for 15 min. To isolate DNA, VLPs were incubated with 10 µL of 10% SDS and 1 µL of proteinase K (Sigma, 20 mg/ml) for 20 min at 56°C, after which 35 µL of 5M NaCl and 28 µL of 10% cetyltrimethylammonium bromide (CTAB)/0.7M NaCl were added, followed by incubation at 65°C for 10 min. The sample was then mixed with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1), vortexed, and centrifuged at 8,000 x *g* for 5 min at room temperature. The resulting aqueous phase was mixed with an equal volume of chloroform and spun at 8,000 x *g* for 5 min at room temperature. The resulting aqueous phase was passed through a Qiagen MiniElute purification column (elution volume, 30 µL).

Multiple Displacement Amplification (MDA) was performed with Genomiphi v2 (GE Healthcare Life Sciences), according to the manufacturer's instructions [n=4 independent reactions/sample to prevent single amplification bias; reactions were subsequently pooled and the DNA product was purified using a Qiagen DNeasy purification kit (elution volume, 75µL)].

Other Methods

Procedures for (i) isolation of total DNA from feces and intestinal contents, (ii) quantifying microbial cell counts in fecal samples by flow cytometry, (iii) preparation of DNA libraries for Illumina HiSeq sequencing, (iv) 454 shotgun pyrosequencing of VLP-derived DNA; (v) COPRO-Seq analysis, (vi) assembly and annotation of novel viral genomes, (vii) cross-contig comparisons, (viii)

INSeq analysis of fitness determinants present in the *B. cellulosilyticus* WH2 prophage, (ix) PCR quantification of *B. caccae* abundance in the fecal microbiota of gnotobiotic mice, and (x) CRISPR analysis are described in *Supplementary Methods*.

References

1 Rodriguez-Brito, B. *et al*., Viral and microbial community dynamics in four aquatic environments. *ISME J*. 4, 739-751 (2010).

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

3 Reyes, A. *et al*., Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*. 466, 334-338 (2010).

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

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

6 McNulty, N. P. *et al*., The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. *Sci Transl Med*. 3, 106ra106 (2011).

7 Mello, L. V., Chen, X., & Rigden, D. J., Mining metagenomic data for novel domains: BACON, a new carbohydrate-binding module. *FEBS letters*. 584, 2421-2426 (2010).

8 Brown, N. L., Stoyanov, J. V., Kidd, S. P., & Hobman, J. L., The MerR family of transcriptional regulators. *FEMS Microbiol Rev*. 27, 145-163 (2003).

9 Barrangou, R. *et al*., CRISPR provides acquired resistance against viruses in prokaryotes. *Science*. 315, 1709-1712 (2007).

10 Krupovic, M. & Forterre, P., Microviridae goes temperate: microvirus-related proviruses reside in the genomes of Bacteroidetes. *PLoS ONE*. 6, e19893 (2011).

11 Roux, S., Krupovic, M., Poulet, A., Debroas, D., & Enault, F., Evolution and Diversity of the Microviridae Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads. *PLoS ONE*. 7, e40418 (2012).

12 Kashyap, P. C. *et al*., Complex Interactions Among Diet, Gastrointestinal Transit, and Gut Microbiota in Humanized Mice. *Gastroenterology*. 2013).

13 Los, M. & Wegrzyn, G., Pseudolysogeny. *Advances in virus research*. 82, 339-349 (2012).

14 Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D., & Bushman, F. D., Hypervariable loci in the human gut virome. *Proc Natl Acad Sci U S A*. 109, 3962-3966 (2012).

15 Villafane, R., Zayas, M., Gilcrease, E. B., Kropinski, A. M., & Casjens, S. R., Genomic analysis of bacteriophage epsilon 34 of Salmonella enterica serovar Anatum (15+). *BMC microbiology*. 8, 227 (2008).

16 Hyman, P. & Abedon, S. T., Bacteriophage host range and bacterial resistance. *Advances in applied microbiology*. 70, 217-248 (2010).

17 Brussow, H. & Desiere, F., Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. *Mol Microbiol*. 39, 213-222 (2001).

18 Goodman, A. L. *et al*., Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe*. 6, 279-289 (2009).

19 Goodman, A. L. *et al*., Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc Natl Acad Sci U S A*. 108, 6252-6257 (2011).

20 Henthorn, K. S. & Friedman, D. I., Identification of related genes in phages phi 80 and P22 whose products are inhibitory for phage growth in Escherichia coli IHF mutants. *J Bacteriol*. 177, 3185-3190 (1995).

Supplementary Methods

Isolation of total DNA from feces and intestinal contents

Microbial community DNA was extracted in a semi-automated 96-sample format. Starting with a single fecal pellet (30-70 mg) or 70-400 mg of frozen intestinal contents (e.g., from the cecum), material was transferred to a 1.7 mL screw cap tube containing 250 µL of sterile 0.1 mm zirconia beads (BioSpec Products), and a steel ball (3.97 mm diameter). Using a Tecan Genesis Series Robot (Tecan), 800 µL of a 500:210 mixture of 2X buffer A (200mM NaCl, 200mM Tris, 20mM EDTA) and 20% SDS was added, followed by 563 µL of phenol:chloroform:isoamyl alcohol (25:24:1; Ambion). The tubes were capped and mixed using a BioSpec Minibeadbeater-96 (4 min). The tubes were then centrifuged at 3,200 x *g* for 4 min. A total of 480 µL of the aqueous phase from each sample was subsequently transferred to an Axygen P-DW-11-C 96-well plate with a Genesis series robot. Using the Biomek FX laboratory automation workstation (Beckman Coulter), 180 µL of the stored aqueous phase was transferred along with 720 µL of a 675:45 mixture of Qiagen buffer PM: 3M sodium acetate (pH 5) to a QiaQuick 96 plate stacked on a Nunc 260251 plate. After mixing 10 times by pipetting, the plate was centrifuged at 3,200 x *g* for 4 min. The plate was washed twice with Qiagen PE buffer (900 µL per well) and centrifuged at 3,200 x *g* for 2 min. Following another centrifugation step to remove leftover ethanol, the multi-well plate was placed on a vacuum manifold to remove residual liquid from the membranes. To elute the DNA, 100 µL of Buffer EB (Qiagen) was added to each well, and the plate was left standing for 2 min before centrifugation at 3,200 x *g* for 2 min. Purified DNA and leftover aqueous phase were stored at -20^{o}C. DNA concentration was measured using the Qubit Quant-IT dsDNA BR protocol as recommended by the manufacturer (Invitrogen, Carlsbad, CA).

Quantifying microbial cell counts in fecal samples by flow cytometry

To quantify the number of microbial cells per fecal pellet and to compare the results with fecal DNA yields, we used the Bacteria Counting Kit (Cat No B7277, Life Technologies) and the manu-

facturer's protocol, with some modifications. In brief, each frozen fecal pellet was weighed and transferred to a sterile 2 mL screw cap tube. 500 µL of sample buffer (1X TE, 0.9% NaCl) was added, and the pellet was resuspended by vortexing for 5 min. The slurry was allowed to settle for 5 min at room temperature and a 10 µL aliquot of the clarified supernatant was transferred to a new tube. The remaining suspension was stored at -20°C for extraction of total community DNA. Serial dilutions of the clarified sample were performed to a final dilution of 1:10,000 in 200 µL of sample buffer containing 1X SYTO BC dye and the kit's standard counting control beads at a concentration of 250 beads/µL. The sample was then passed through a nylon mesh (60 µm pore diameter) to remove large particles that could potentially clog the cytometer (note that the control beads [6µm diameter] are larger than the bacterial cells and the number of beads present was not affected by passage through the nylon mesh). The flow thru was then split into three equal parts and each replicate was counted for 1 min in the MXP flow cytometer. Forward and side scatter data were collected along with fluorescence data in the 525 nm channel. Gates were drawn on the total bacterial cell population, the SYTO BC positive population., and the control beads. Quantifying the control beads allowed us to calculate the number of bacterial cells (positive fluorescent cells) per mg (wet weight) of fecal material.

To extract DNA from these samples, we added zirconia and steel beads to the resuspended pellet described above, along with 300 µL of extraction buffer (118 mM NaCl, 435 mM Tris, 44mM EDTA, 547mM SDS), so that the final concentration of components was equivalent to 2X buffer A (see above) and 20% SDS. At this point, the protocol for extracting total community DNA was performed exactly as described above.

Preparation of DNA libraries for Illumina HiSeq sequencing

100 µL of total DNA in TE pH 7.0 (5 ng/µL), was fragmented by sonication in thin-walled 0.2 mL 8-strip PCR tubes using a BioruptorXL multi-sample sonicator (Diagenode) set on 'high'; sonication occurred over the course of 20 min using successive cycles of 30 sec 'on' followed by 30 sec 'off '. Sonicated samples were subsequently cleaned up using the MinElute 96 UF PCR Purifica-

tion Kit (Qiagen) per the manufacturer's instructions. Each sonicated DNA sample in each well of the 96-well plate was eluted with 22 µL nuclease-free sterile water. For end repair and A-tailing, 20 µL of sonicated DNA was mixed with 5 µL of a mixture containing 2.5 µL of 10X T4 DNA ligase buffer (NEB), 1 µL of 1mM dNTPs (NEB), and 0.5 µL of each of the following enzymes: T4 polymerase (NEB, 3 U/µL), T4 polynucleotide kinase (NEB, 10 U/µl), and Taq polymerase (Invitrogen, 5 U/µL). The solution was mixed by vortexing and then incubated for 30 min at 25°C followed by 20 min at 75°C. Customized Illumina adapters containing maximally distant 4-6 bp barcodes (**Table S7**) were ligated to the A-tailed DNA in a 27 µL reaction by adding 1 µL of 25µM adapter mix plus 1 µL of T4 DNA ligase (2,000,000 U/mL; NEB). Adapter mix was prepared by mixing 12.5 µL of a 100µM stock of each adapter oligo (constituting the forward and reverse strands) and 25 µL of oligo buffer (1X TE, 0.1M NaCl), incubating the mixture at 95°C for 1 min, then slowly decreasing the temperature (0.1°C/second) until reaching 4°C. After a 30 min incubation at 16°C, 2.5 µL of 50mM EDTA was added to stop the ligation reaction. Sets of 24 samples, all harboring different adapter sequences, were pooled, and the pool was purified using MinElute PCR purification columns according to the manufacturer's instructions (15 µL final eluate volume). A 10 µL aliquot of the purified pool was subjected to 2% agarose gel electrophoresis. DNA of approximately 200 bp was excised and purified using QIAquick Gel Extraction Kit and MinElute purification columns (Qiagen; 12 µL final elution volume). Finally, 1 µL of the size-selected library was used as the template in an enrichment PCR (19 cycles of 98°C for 10 s, 67°C for 30 s and 72°C for 30 s, using Phusion HF Master mix (NEB) and Illumina's standard amplification primers (**Table S7**) in a final volume of 25 µL). The PCR product was purified using MinElute PCR purification columns and DNA was quantified using the Quant-iT dsDNA High-Sensitivity (HS) Assay Kit (Invitrogen). Libraries concentrations were then normalized and an equimolar pool was subjected to multiplex sequencing with an Illumina HiSeq 2000 instrument (2.5pM/lane; read length, 50 nt).

454 shotgun pyrosequencing of VLP-derived DNA

Libraries for 454 shotgun pyrosequencing were generated using a protocol similar to that used for the Illumina libraries. The protocol differed slightly in that sonication was performed for 8 min instead of 20 min. Each library was purified using Agencourt AMPure XP beads (Beckman Coulter) and quantified using the recommended 454 rapid library barcoded fluorescent-labeled adapters and a plate reader (Synergy2, Biotek, Winooski, VT). A total of 24 independent adapters were synthesized harboring 24 different multiplex identifier (MID) barcodes. After quantification, normalized pools of 24 samples were sequenced using 454 FLX Titanium chemistry. Initial quality filtering of the raw data consisted of parsing the sequenced reads by their MID followed by removal of short reads (less than 60 nt), reads with three or more ambiguous ('N') bases anywhere in the sequence, reads with two continuous 'N' bases, and replicate reads (reads where the first 20 nt were >97% identical).

COPRO-Seq analysis pipeline

A combination of pre-existing standalone software and custom perl scripts were used for performing COPRO-Seq analysis on our data in a computer cluster environment running Sun Grid Engine. For raw Illumina reads, sequences were demultiplexed using the 4 bp barcode at the beginning of each read. The short barcode lacks the complexity required for error correction; however, with low error rates at the beginning of Illumina reads and a Hamming distance between any two barcodes of at least 2, requiring a perfect match to a barcode sequence allowed us to efficiently and accurately assign the reads to specific samples. After dividing the sequences by sample, we were able to run the downstream analysis in parallel. This downstream processing consisted of first screening for adapter sequences that indicated the presence of either short inserts or adapter dimer [screening was performed using cross_match (version 1.090518; Green P, 2009, http://www.phrap.org) with slight modifications to the default parameters (-gap1_only -minmatch 6 -minscore 10 -gap_init -3 –screen)]. Next, reads were either trimmed back when an adapter sequence was seen or removed entirely in the event that: (i) the sequence was shorter than 35 nt after removing

the adapter sequence, or (ii) more than 3 N's were found anywhere in the read. The 'clean' reads were used to query the 15 sequenced bacterial genomes in the model human microbiota, as well as their predicted prophage, using FR-HIT v.0.7 [23] with one modification to the default parameters (-c 90). The mapping file was parsed as described previously [24]; in particular, reads mapping uniquely to a single genome were used to determine relative abundances, subsequently, exact ties or reads mapping equally to more than one genome were weighted based on the relative abundance of the genomes involved. Raw counts were then normalized to reads/kb/million mapped reads. Reads that did not map to any bacterial genome were then mapped against other known plasmids from these bacterial species, PhiX174 (used as the internal control on the Illumina sequencer), and to the reference mouse genome. All remaining reads were used for *de novo* hybrid assembly for identification of novel viruses (see below).

Normalized counts for the different bacterial species and virotypes present in each sample were used to calculate relative abundances or Genome Equivalents (GE; the latter by normalizing the percentage of mapped reads to the length of the genome, and the DNA yield/sample). Matrices of either relative or absolute abundances were generated for principal coordinates analysis (PCoA) and generation of 3D plots using scripts in QIIME [25] (version 1.3.0-dev). Statistical analysis and figure generation were performed in Prism (v6).

Assembly and annotation of novel viral genomes

Reads generated from either VLP-derived DNA (454 FLX pyrosequencer using Titanium chemistry) or from total community DNA (Illumina HiSeq2000 instrument) that did not map to any of the reference genomes were pooled together and submitted for a hybrid assembly using MIRA V3.4.0 [26] with minimum overlap of 70nt (454) or 20nt (Illumina) and a minimum relative score of 90%. Note that raw data from 454 pyrosequencing of VLPs was parsed with essentially the same pipeline used for Illumina reads except for the adapter-mapping step that is incorporated by default in 454 sfftools. Given that 454 FLX reads are longer than Illumina reads, FR-HIT parameters used were –c 90 for percent similarity and –m 40 for minimal length of the alignment. A *de novo* assem-

bly was performed for the first round. After assembly, contigs were individually analyzed using Tablet v.1.12.09.03 [27] to identify potential chimeras in the assembly, and checked using blast for overlapping ends either within contigs (indicating a complete, circular phage genome) or between contigs (indicating a potential link between two contigs). Raw reads were then mapped to the contigs using mapping-based assembly in MIRA, which allows extension of previously assembled contigs via the incorporation of new reads at the edges of these contigs. The process was repeated five times yielding, in the final iteration, a total of five large circular contigs covering most of the non-mapped reads from both the 454 and Illumina datasets. These novel viral genomes were used in conjunction with the reference bacterial genomes as a new reference dataset for analysis of the relative abundances of the viruses and host bacteria (from the VLP and total fecal community DNA shotgun sequencing datasets).

The five viral genomes were annotated first with Glimmer v.3.02 [28] trained on all open reading frames (ORFs) predicted from viral ref_seq (NCBI). Predicted ORFs were then blasted against COG (STRING v.9.0), KEGG (v58), ACLAME (v0.4), CDD (online version), NCBI nr (retrieved 13/08/2012), and Phantom (retrieved 01/09/2012) (blastp; threshold < 1e-5, no low redundancy filter).

Cross-contig comparison

To determine which human donors were the source of the viral genomes assembled from the mouse fecal VLP and COPRO-Seq datasets, we first retrieved FASTA files of pyrosequencing datasets that we had previously generated from each of the five purified human donor VLP DNA preparations used to create the input pool for our mouse experiments (GenBank ID SRX028824; [3]). The sequences were pooled together with 454 FLX Titanium sequences obtained from the pooled VLP input (71,546 reads). The assembly pipeline consisted of the following steps: (i) CD-HIT v.4.6 [29] was used to cluster reads at 90% global identity from each cluster and the top 5 sequences were taken as representatives; (ii) these reads were used for *de novo* assembly using Newbler v.2.6 (454 Life Sciences) with default parameters; (iii) FR-HIT was used to map all raw reads to the

assembled contigs at 90% identity; (iv) reads that did not map were pooled and re-assembled using Newbler; (v) contigs >500 bp after both rounds of assembly were pooled together along with sequences from the five novel viral genomes and all raw reads were mapped using MIRA in mapping assembly mode, which allows for extension at the edges of contigs; (vi) the extended contigs were assembled using Phrap; (vii) the Phrap output files: 'contigs', 'singlets' and 'problems' were concatenated, re-named and used to map the reads using FR-HIT; (viii) chimeras were identified as sudden drops in coverage given by the FR-HIT mapping; contigs were split on chimeric junctions; (ix) contigs were sorted by size and then the program Megablast was used to compare 'all-against-all' and to identify contigs that were fully contained within another contig and contigs with overlapping ends; and (x) the final set of contigs over 2 kbp was used as a reference set for mapping (FR-HIT) all raw reads from each of the fecal VLP-derived viromes from each of the individual human fecal samples utilized to generate the input pool of VLPs for the staged viral attack, the pooled VLPs used to gavage the mice, and mouse fecal VLP DNA. A matrix of reads per kbp of contig sequence per million reads of sample was generated. The matrix was then log transformed, and a heat map was built using R [30].

INSeq analysis of fitness determinants present in the *B. cellulosilyticus* WH2 prophage

Whole genome transposon mutagenesis of *B. cellulosilyticus* WH2 was performed using protocols described in earlier publications [18,21]. Total fecal community DNA was isolated from fecal pellets obtained from mice belonging to treatment groups 1 and 2 at time points between 3 and 13 d after they had received a single gavage of this library of 25,000 transposon mutants together with the other bacterial members of the defined community. Each DNA sample was adjusted to 500 ng in 15 µL of TE buffer and then digested with MmeI (4 U, NEB) in a 20 µL reaction supplemented with 10 pmoles of a 12 bp DNA with an MmeI restriction site (to improve the efficiency of restriction enzyme digestion) [18]. The reaction was incubated for 1 h at 37°C and then terminated (80°C for 20 min). MmeI-digested DNA was subsequently purified using 125 µL of AMPure beads (after washing the beads once with 100 µL of sizing solution (1.2 M NaCl and 8.4% PEG 8000)). The digested

DNA was added to the beads and the solution incubated at room temperature for 5 min. Beads were pelleted with a Magnetic Particle Collector (MPC), washed twice (each time using a mixture composed of 20 µL TE (pH 7.0) and 100 µL sizing solution, with bead recovery via MPC after each wash), followed by two ethanol washes (180µL 70% ethanol/wash) and air-drying for 10 min. Samples were resuspended in 18 µL TE (pH 7.0) and the DNA was removed after pelleting beads with the MPC. Ligation of adapters was performed in a 20 µL reaction that contained 16 µL of purified DNA, 1 µL of T4 Ligase (2000 U/µL; NEB), 2 µL 10X ligase buffer and10 pmoles of barcoded adapter (incubation for 1 h at 16^{o}C). Ligations were subsequently diluted with TE (pH 7.0) buffer to a final volume of 50 µL, mixed with 60 µL of AMPure beads, and incubated at room temperature for 5 min. Beads with bound DNA were pelleted using the MPC and washed twice with 70% ethanol as above. After allowing the ethanol to evaporate for 10 min, 35 µL of nuclease-free water was added and the mixture was incubated at room temperature for 2 min before collecting the beads with the MPC. Enrichment PCR was performed in a final volume of 50 µL using 32µL of the cleaned up sample DNA, 10 µL 10X Pfx buffer, 2 µL 10mM dNTPs, 0.5 µL 50mM $MgSO_4$, 2 µL of 5µM amplification primers (Forward primer: 5'CAAGCAGAAGACGGCATACG3', Reverse primer: 5'AATGATACGGCGACCACCGAACACTCTTTCCCTACACGA3'), and 1.5 µL Pfx polymerase (2.5 U/µL; Invitrogen) 22 cycles of denaturation at 94^{o}C for 15 s, annealing at 65^{o}C for 1 min and extension at 68^{o}C for 30 s. The 134 bp PCR product from each reaction was purified [4% metaphore gel; MiniElute Gel extraction kit (Qiagen)] in a final volume of 20 µL, and was quantified (Qubit, dsDNA HS Assay Kit; Invitrogen). Reaction products were then combined in equimolar amounts into a pool that was subsequently adjusted to 10nM and sequenced (Illumina Hi-Seq2000 instrument).

Illumina 50 nt short reads were separated by sample using 4 or 7 nt sample-specific barcodes. The remaining read contains either the 5' or 3' end of the Tn sequence along with 16-17nt of flanking genomic DNA. After trimming of the Tn sequences, reads were mapped to the *B. cellulosilyticus* WH2 genome, allowing up to 1 nucleotide mismatch. The read counts derived from mapping the 5' or 3' termini of the transposon were added and a normalized count of reads per mil-

lion for each insertion position was generated. For the sliding window analysis, normalized read counts were added up for every 600bp window throughout the reference genome, while sliding the window in increments of 100bp. The final distribution of counts per window for any given sample obtained at any given time point were analyzed using the quantile function on R [30] to determine the different abundance levels. Final plots of the quantiles over time were generated using Prism v6.0a.

PCR quantification of *B. caccae* abundance in the fecal microbiota of gnotobiotic mice

B. caccae abundance was evaluated using total fecal community DNA isolated from samples obtained 22 or 23 days after gavage of the 15-member community (i.e., 2 and 3 days post viral gavage). Purified *B. caccae* genomic DNA was used to generate a standard curve. The standard curve and amplification plots for the samples [generated using the PCR primer set for region dpg_11, see **Table S7**) were used to calculate the number of genome copies per mg of fecal pellet obtained at each selected time point from members of both the live and heat killed VLP treatment groups].

CRISPR analysis

The genome sequences of the bacterial species introduced into gnotobiotic mice were searched using CRISPR-Finder [31]. All Illumina HiSeq reads generated for COPRO-Seq analyses of fecal DNA samples were screened for CRISPR repeats using tre-agrep (v0.8) for fuzzy string matching. Any read covering the repeat with at most 3 mismatches was retrieved and subsequently screened for the presence of known spacers using cross-match. Once spacers that were present in the sequenced genome were filtered out, any remaining spacer should correspond to new accumulated spacers; as noted in *Supplementary Discussion*, none were observed.

Supplementary Discussion

A search for fixed mutations in the *B. caccae* genome that could potentially confer viral resistance after the staged attack with pooled human VLPs

Reads derived from total community shotgun sequencing that mapped to the *B. caccae* genome were selected from fecal samples obtained 11-21 d after introduction of the 15-member model human microbiota and 11-25 d following VLP gavage. Reads were initially grouped into four datasets: (i) live VLP, pre-VLP gavage, (ii) live VLP, post-VLP gavage, (iii) heat-killed VLP, pre-VLP gavage, and (iv) heat killed VLP, post-VLP gavage. A mutation conferring resistance should be fixed in the population represented in dataset (ii).

All the reads from (i)-(iv) were pooled and used for mapping assembly (MIRA) while keeping track of the dataset from which the reads originated. Thirty-fold coverage of the genome was obtained on average per treatment. We identified no single nucleotide substitution specific to any given treatment group and fixed in over 10% of the *B. caccae* population (as defined by the sequence coverage at a given nucleotide position). A total of nine loci were identified where there was low or no coverage in one compared to the other treatment group, suggesting possible deletions. These loci were selected after filtering for: (i) repetitive regions that can mask true SNPs due to variation between the repeats, and (ii) regions with an overall coverage (in all four treatments) of less than 25-fold. We designed a pair of PCR primers flanking each of the nine loci in order to amplify a product between 270-370 bp (**Table S7**). DNA from each of the five mice per treatment group were selected as templates for the PCR: i.e., (i) fecal DNAs from all mice in treatment group 1 that had been sampled 17 d following gavage with the 15-member community (live VLP, pre-VLP gavage group), (ii) fecal DNAs from all mice in treatment group 1 prepared from samples obtained 21 d after viral gavage (live VLP, post-VLP gavage group), (iii) fecal DNAs from all mice from treatment group 2 sampled 17 days post bacterial gavage (heat-killed VLP, pre-VLP gavage group), and (iv) fecal DNAs from all mice in treatment group 2 sampled 21 d after viral gavage (heat-killed VLP, post-VLP gavage group).

Amplifications were performed using a MX3000P qPCR instrument (Stratagene) in 25 µL reaction mixtures that contained 1X Platinum Taq HiFi PCR buffer (Invitrogen), 0.3µM PCR primers (**Table S7**), 2mM $MgSO_4$, 0.2mM dNTPs, 0.1 mg/ml BSA, 0.01% Tween 20, 0.08X SYBR Green I (Invitrogen), 0.02 U/µL Platinum Taq (Invitrogen) and total fecal community DNA

(10ng). Amplification was performed using 35 cycles of: denaturation at 94°C for 30 s, annealing at 50°C for 30 s and elongation at 68°C for 1 min, with fluorescence recorded at the end of elongation. After amplification, a standard dissociation curve was generated to assess the specificity of the amplification, followed by a10 min incubation at 68°C to ensure that the amplified products had an A-overhang. PCR products were subsequently purified using the MinElute 96 UF Purification Kit (Qiagen; 20µL final elution volume). DNA was quantified (Qubit dsDNA HS Assay Kit; Invitrogen) and pooled by treatment group and primer set (100 ng/sample). A total of 500 ng of amplified DNA for each treatment group and primer set was added to a 50 µL reaction mixture that contained 5 µL of 1µM Illumina adapters (**Table S7**), 5 µL of 10X T4 ligase buffer, and 1 µL T4 Ligase. Ligation was performed for 30 min at 16°C followed by heat inactivation at 65°C for 10 min. The adapter ligated DNAs were purified using AMPure beads as described above for the INSeq protocol (and eluted in 30 µL), quantified (Qubit dsDNA BR Assay Kit; Invitrogen) and a 4 ng aliquot was used per treatment/primer set for enrichment PCR, as described above for preparation of Illumina libraries. Barcoded amplicons were subjected to 2% agarose gel electrophoresis to confirm their molecular weight and purified using MinElute PCR purification columns (Qiagen). After quantification, the different products were pooled in equimolar amounts and sequenced using an Illumina MiSeq instrument (paired-end, 250 nt reads).

Sequencing reads were split by barcode; for each sample, reads were trimmed by quality and the paired-end products were assembled using FLASH [32] (note that assembly of the paired end reads is expected under these conditions since the largest PCR product is 370 bp and reads are 242 bp after removing the 8 bp barcode on each end). Even after quality trimming of the reads, overlaps >30 bp were observed, allowing assembly of >80% of the data. Once the amplicons were assembled, they were pooled with the reference sequence and cd-hit-est [29] was used to generate 100% identical clusters of sequences. A table of read counts per treatment per cluster was generated; the percentage of reads in any cluster that were different than the reference cluster never exceeded 5% of the total reads, indicating that no single mutation was fixed in *B. caccae* at the time points and conditions surveyed.

CRISPR analysis

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) play a role in conveying resistance to foreign DNA through a mechanism that involves accumulation of spacers in the bacterial genome that map to short segments in the invading DNA [9]. Only 6 of the 15 members of the model microbiota contained CRISPR regions; five were members of the Firmicutes; the other, *Parabacteroides distasonis*, is a member of the Bacteroidetes (**Table S1**). With the sequence coverage observed over time, none of the CRISPR regions appeared to have accumulated new spacer sequences. However, this finding has to be interpreted cautiously since none of these bacterial species reached a relative abundance greater than 10% at any time point after VLP gavage, thereby limiting our coverage.

Supplementary Figures

Fig. S1 - Experimental design. Three groups of 8-week old male C57BL/6J germ-free mice were gavaged with a 15-member community of sequenced human gut bacterial symbionts and/or a pool of virus-like particles (VLPs) that had been isolated from frozen fecal samples obtained from five healthy human adults. Each group of mice was kept in a separate gnotobiotic isolator and each mouse in each isolator was individually caged. Each vertical line represents a day. Arrows denote the time of gavage with the 15-member bacterial community (black) and with live or heat-killed VLPs. Time points selected for sampling the fecal microbiota of each mouse in each treatment group are represented as circles. Samples were subjected to shotgun sequencing of total community DNA (COPRO-Seq, yellow circles) and/or shotgun sequencing of purified VLP DNA (blue circles).

Fig. S2 - Measurements of fecal microbial biomass in mice gavaged with live or heat-killed human VLPs. (**A)** Plot of fecal DNA yields (mean ± SEM) over time. Microbial biomass achieves a maximum value four days after introduction of the 15-member consortium of human gut bacterial symbionts. (**B)** Correlation between fecal DNA concentration/mass ratios and bacterial cell concentration as measured by flow cytometry (n=94 fecal samples). A Pearson correlation of 0.8371 and an associated p-value of 7×10^{-26} indicated a significant association.

Fig. S3 - Community dynamics of members of the 15-member community as a function of time after colonization with the species consortium and type of VLP treatment. PCoA of a matrix of Hellinger distances measured using shotgun COPRO-Seq data. Samples are plotted in a two-dimensional representation of the first vector (49%) of variance against time. Most of the variation is explained by community assembly during the first week of the experiment. The arrow corresponds to the time of the VLP gavage. Red and blue dots represent individual samples from the two different treatment groups (see **Fig. 1 and S4** for additional analysis of changes in the relative abundance of community members as a function of time and the different treatments).

Fig. S4 - Changes in relative abundance of bacterial taxa in gnotobiotic mice containing the 15-member model human gut microbiota prior to and after attacked with a pool of purified live- or heat-killed human fecal VLPs. Relative abundance (mean ± SEM) is plotted as a function of time for the 13 bacterial species not shown in **Fig. 1**. Data are based on COPRO-Seq analysis of fecal samples collected from gnotobiotic animals beginning 15 days after gavage with the 15-member model community. The upper pointing arrow at day 20 indicates the time when the live or heat-killed human fecal-derived VLP preparation was administered to animals with a single gavage. Some but not all of these community members change their representation (either increase or decrease) in after introduction of live as opposed to heat-killed VLP preparations.

Fig. S5 - Abundance, genome annotation, and associated bacterial host dynamics of three human phage identified in the fecal microbial communities of gnotobiotic mice. (**a-f**) The assembled annotated phage genomes, their changes in abundance in the fecal microbiota of mice belonging to the live and heat-killed VLP treatment groups. The circular representation of each phage genome illustrates the location of genes in positive (green) and negative (red) strand and which genes were found to be significantly similar to known viral genes (blue; blast E value < 10-5; **Table S5**). The inner plot represents GC skew based on 200 bp windows (yellow, G/C ratio is greater than the average for the genome; purple, ratio is lower than the average). Symbols and line patterns differentiate mice within a given treatment group.

Fig. S6 - Sequence variability within the host recognition region of the ϕHSC02 VP1 protein. (**a)** Three dimensional space filling model of the viral capsid of a member of Microviridae (PDB identifier SpV4 PDB:1KVP) composed of polymers of VP1 (monomer highlighted in red). (**b)** Individual monomer of the VP1 protein. The fold highlighted in green is the region predicted to be involved in host recognition [11]. (**c)** Segment (450aa) of a multiple protein alignment of the VP1 proteins from the ϕHSC02 virus and two related members of Alphavirinae: human_gut_33_017 and human_gut_22_017 [11] starting from residue 150. Residues lying in the green background box correspond to the region highlighted in green in panel B. Alignment conservation is shown by font

colors which differentiate identical amino acids (blue) from similar (red) or non-related residues (black).

Fig. S7 – Relationship between the abundance of bacterial species and three human-derived phage (ϕHSC03, ϕHSC04, ϕHSC05) along the length of the gut of gnotobiotic recipients of the pooled live human VLP preparation at the time of sacrifice. (**a)** Viral abundance for each mouse in the indicated treatment group as defined by shotgun sequencing of DNA isolated at the time of sacrifice from different portions along the length of the gut. The five columns for each mouse (M1-M10) are organized from left to right as follows: proximal small intestine (PSI); distal small intestine (DSI), cecum (Ce), colon (Co), fecal sample (Fe). Note that no virus was detected in mice that had received the heat killed human fecal VLP preparation. ND: Not detected (Due to the lower density of bacteria in the proximal gut, samples from the proximal half of the small intestine may harbor phages at levels below what could be detected by COPRO-Seq at the depth of sequencing employed). (**b)** Relative abundance for each of the 15 bacterial members of the model human gut microbiota along the length of the gut of mice belonging to the two VLP treatment groups. The order of the columns for each mouse is the same as in panel a.

Fig. S8 - Shotgun 454 pyrosequencing data generated from VLPs isolated from mouse fecal samples validates predicted prophage regions in bacterial genomes. Mapping of VLP-derived shotgun sequencing reads to the genomes of the 15-member model human bacterial community assembled in gnotobiotic mice allows identification of prophage that are induced. The number of reads mapping to bacterial genomes outside the prophage region was low in all mice. Therefore, all hits to any bacterial genome (outside of a prophage region) are depicted in black. Each bar represents data generated from two fecal pellets obtained from one or two mice within a given treatment group at a given time point. Cecal samples were collected at time of sacrifice and are numbered according to the mouse ID in each treatment group.

Supplementary Figures

Fig. S1.

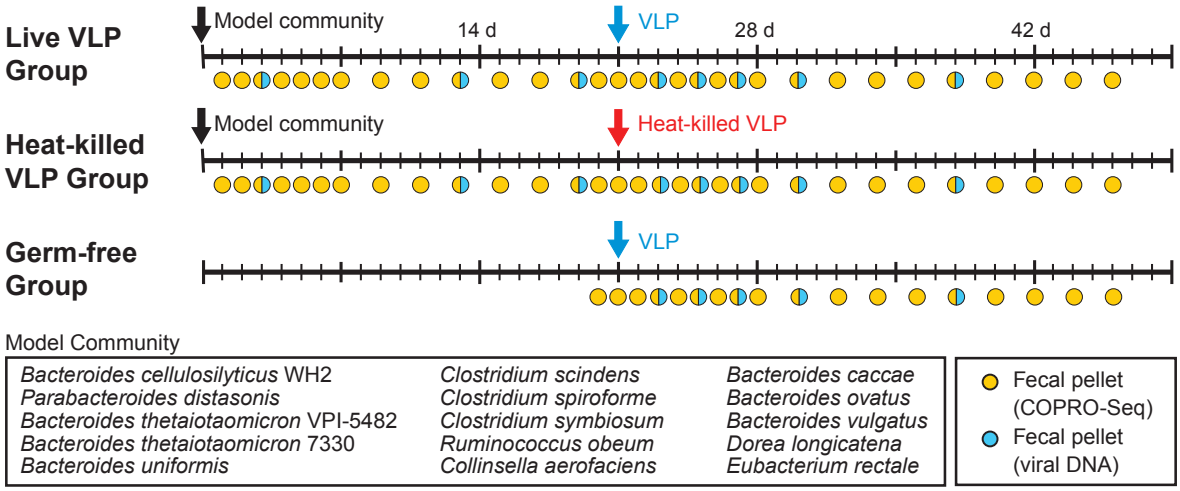


Fig. S2.

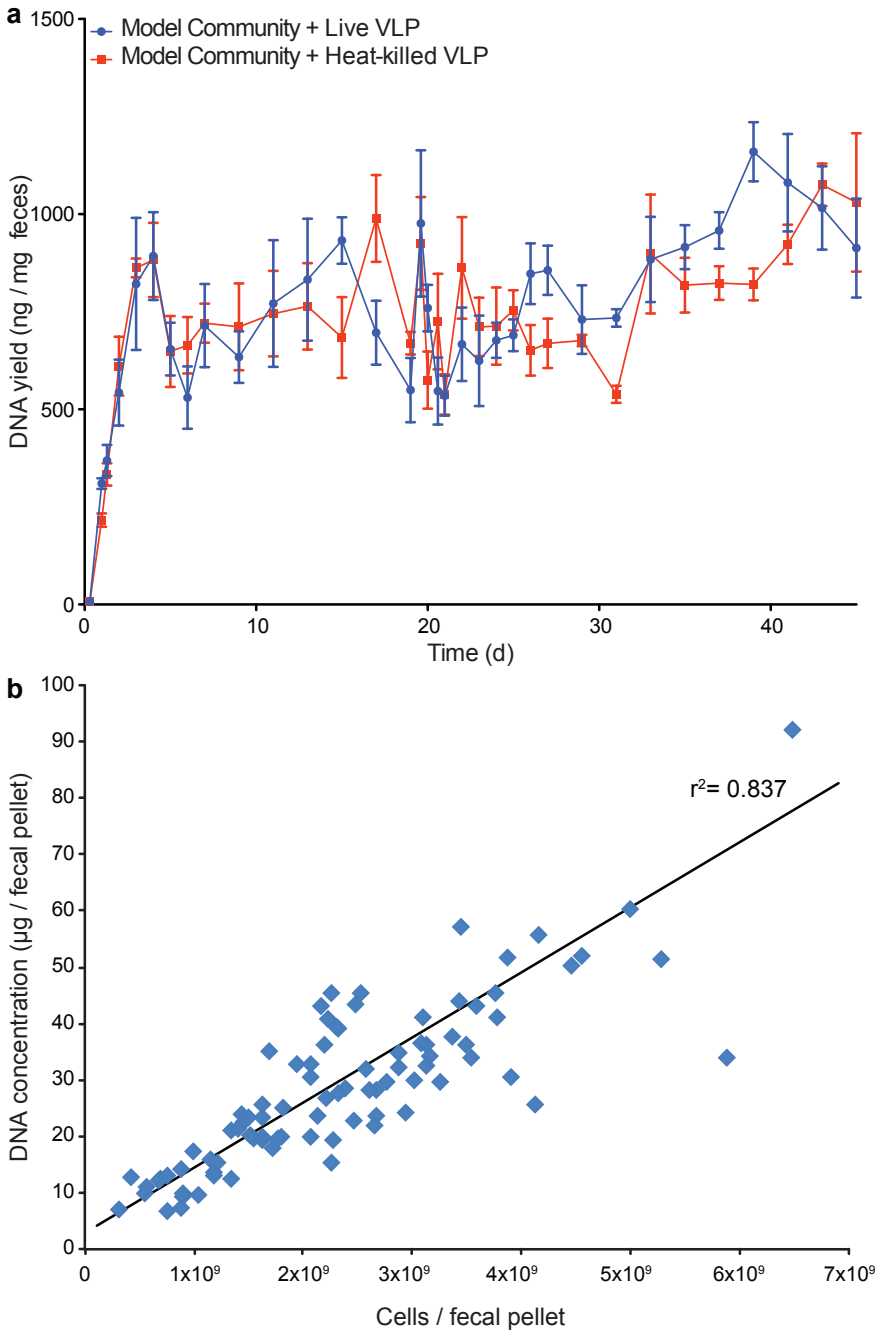

a
Model Community + Live VLP
Model Community + Heat-killed VLP
DNA yield (ng / mg feces)
Time (d)
b
r²= 0.837
DNA concentration (µg / fecal pellet)
Cells / fecal pellet

Fig. S3.

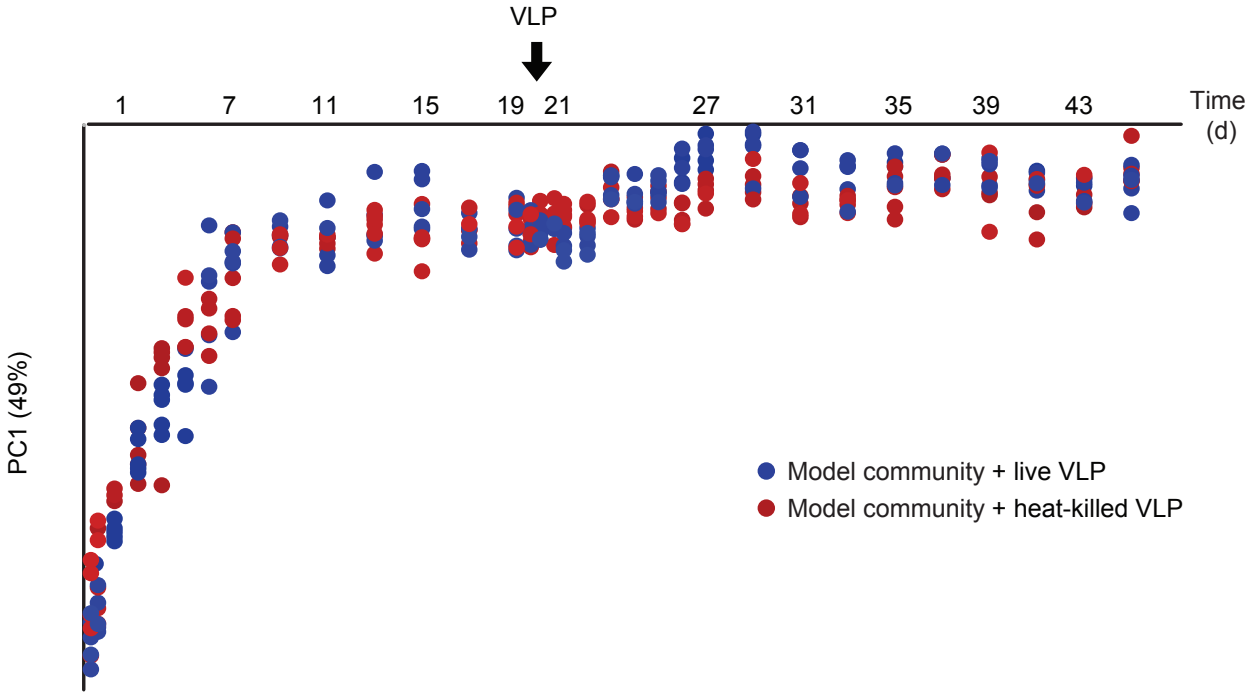

VLP
1
7
11
15
19
21
27
31
35
39
43
Time (d)
PC1 (49%)
Model community + live VLP
Model community + heat-killed VLP

Fig. S4.

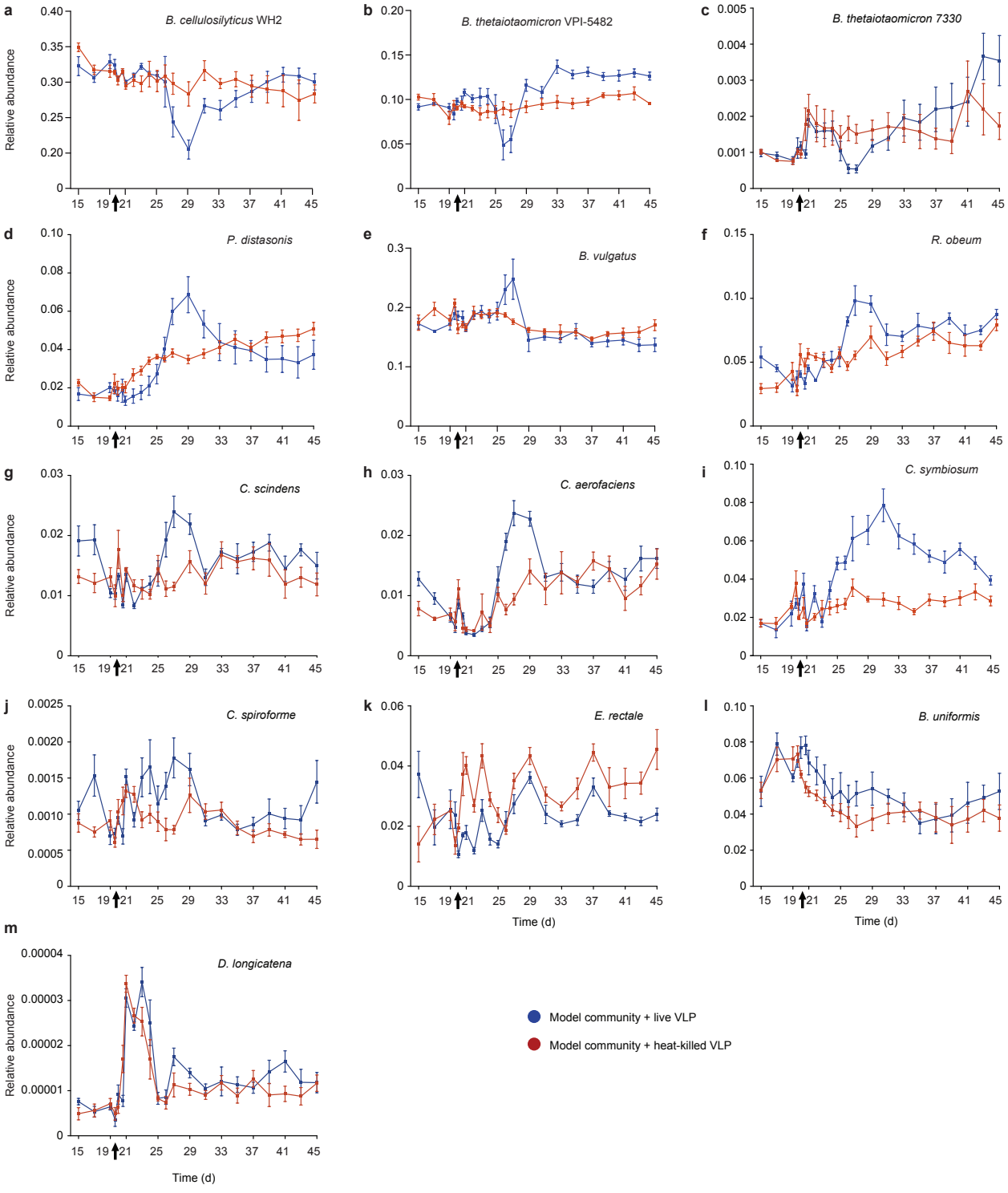

a
B. cellulosilyticus WH2
b
B. thetaiotaomicron VPI-5482
c
B. thetaiotaomicron 7330
d
P. distasonis
e
B. vulgatus
f
R. obeum
g
C. scindens
h
C. aerofaciens
i
C. symbiosum
j
C. spiroforme
k
E. rectale
l
B. uniformis
m
D. longicatena
Relative abundance
Time (d)
Model community + live VLP
Model community + heat-killed VLP

Fig. S5.

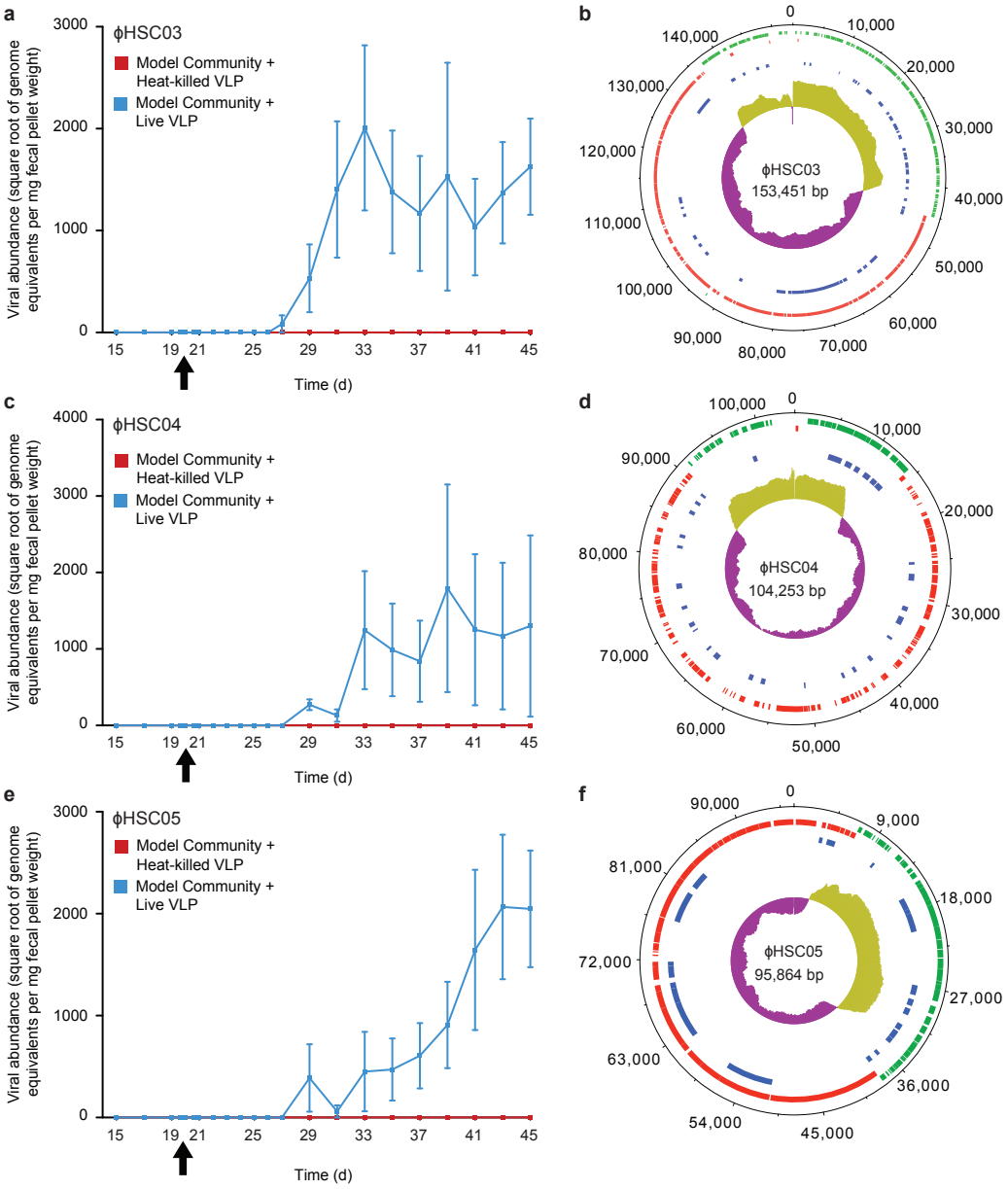

a
ϕHSC03
Model Community + Heat-killed VLP
Model Community + Live VLP
Viral abundance (square root of genome equivalents per mg fecal pellet weight)
Time (d)
b
ϕHSC03
153,451 bp
c
ϕHSC04
Model Community + Heat-killed VLP
Model Community + Live VLP
Viral abundance (square root of genome equivalents per mg fecal pellet weight)
Time (d)
d
ϕHSC04
104,253 bp
e
ϕHSC05
Model Community + Heat-killed VLP
Model Community + Live VLP
Viral abundance (square root of genome equivalents per mg fecal pellet weight)
Time (d)
f
ϕHSC05
95,864 bp

Fig. S6.

A
B
C
ϕHSC02
Human_gut_33
Human_gut_22
RADLAYKLLSYLGYGNFI PSPPSNSTRWWSTSLKKEASPTGYTQQYIQNNYVNLFPLLAYQKIYQDFFRWSQWER 225
RADLSFKLLSYLGYGNLI ETPPSLGNRWWSTSLKNTDDGANYTQQYIQNTIVN FPLLTYQKIYQDFFRWPQWEK 225
RADLAFKLFSYLGYGNVWSEFSSSNRWWSTSLKGGGS- - - YTQQYVQDSYVNLFPLLAYQKIYQDFFRWSQWEN 216
ANPSSYNVDYYSGVSSSLVTVLPDYTSDYWKSDTMFDLKYCNWNKDMLMGILPDSQFGDVAII DFETPGG- DLKA 299
SNPSSYNVDYFSGSSPSIVSDLPAASSAYWKSDTMFDLKYCNWNKDMLMGVLPNSQFGDVAVLDI SSSGDSDVVL 300
SNPSSYNVDYFTGVSPLLVSSLPEASNAYWKSPTMFDLKYCNWNKDMLMGVLPNSQFGDVAVLDI DNSGKPDVVL 291
GFRTTDGKFI SAVTNAPLTTENSSG- LSTPGVTSGSTVALKSPLISDLSALQSQFSVLALRQAEALQRWKEISQ 373
GVDPHK- - - - STLGISAITSKSAVVPLFALDASTSNPVSVGSKLHLDLSSIKSQFNVLALRQAEALQRWKEISQ 371
GLGNAN- - - - STVGVASYVSSNTASIPFFALKASSANTLPVGSTLRVDLASLKSQFTVLALRQAEALQRWKEISQ 362
SGDSDYREQIRKHFGVNLPQSLSNLCTYIGGISRNLDISEVVNNNLAAEGDTAVIAGKGVGAGNGSFTYTTDEHC 448
SGDSDYREQILKHFGVKLPQALSNLCTYIGGISRNLDISEVVNNNLAAEEDTAVIAGKGVGTGNGSFTYTTNEHC 446
SGDSDYREQIRKHFGVNLPQALSNMCTYIGGISRNLDISEVVNNNLAAEGDTAVIAGKGVGAGNGSFTYTTNEHC 437
VVMCIYHAVPLLDYTITGQDGQLLVTDAESLPIPEFDNIGMEVLPMTQIFNSPKASIVNLFNAGYNPRYFNWKTK 523
VIMGIYHAVPLLDYTLTGQDGQLLVTDAESLPIPEFDNIGLEVLPMAQIFNSSLATAFNLFNAGYNPRYFNWKTK 521
VVMCIYHAVPLLDYTITGQDGQLLVTDAESLPIPEFDNIGMETLPMTQIFNSPKASIVNLFNAGYNPRYFNWKTK 512
LDVVNGAFTTTLKSWSPVTESLLSGWFGFGYSQDDVNK TKVVLNYKFFKVNPSVLDPIFGVNADSTWDTDQLL 598
LDVINGAFTTTLKSWSPVSESLLSGWARFGAS- -DSKTGTK VLNYKFFKVNPSVLDPIFGVKADSTWDTDQLL 594
LDVINGAFTTTLKSWSPVTESLLSGWFGFGYSEGDVNSQNKVVLNYKFFKVNPSVLDPIFGVAADSTWDSDQLL 587

Fig. S7.

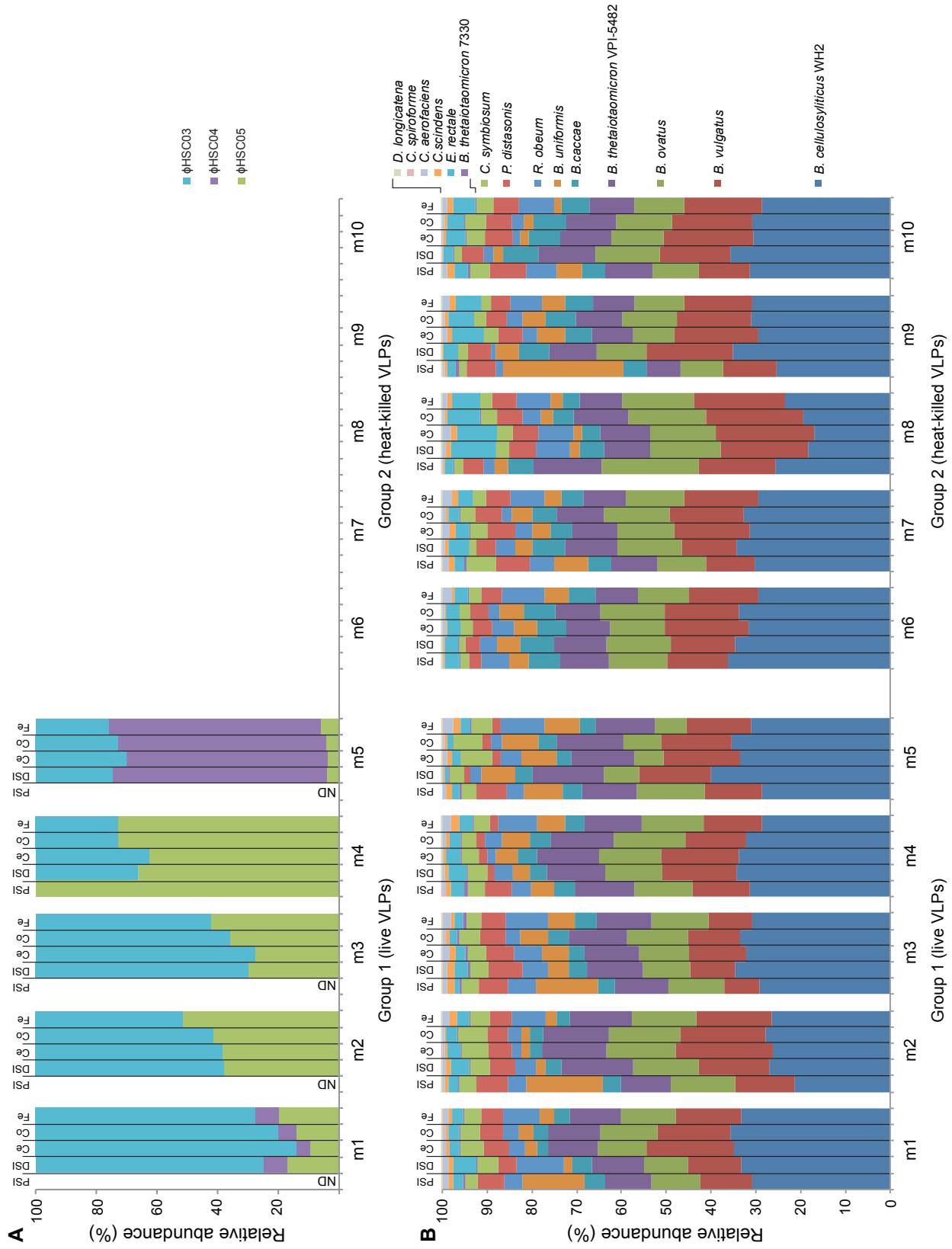
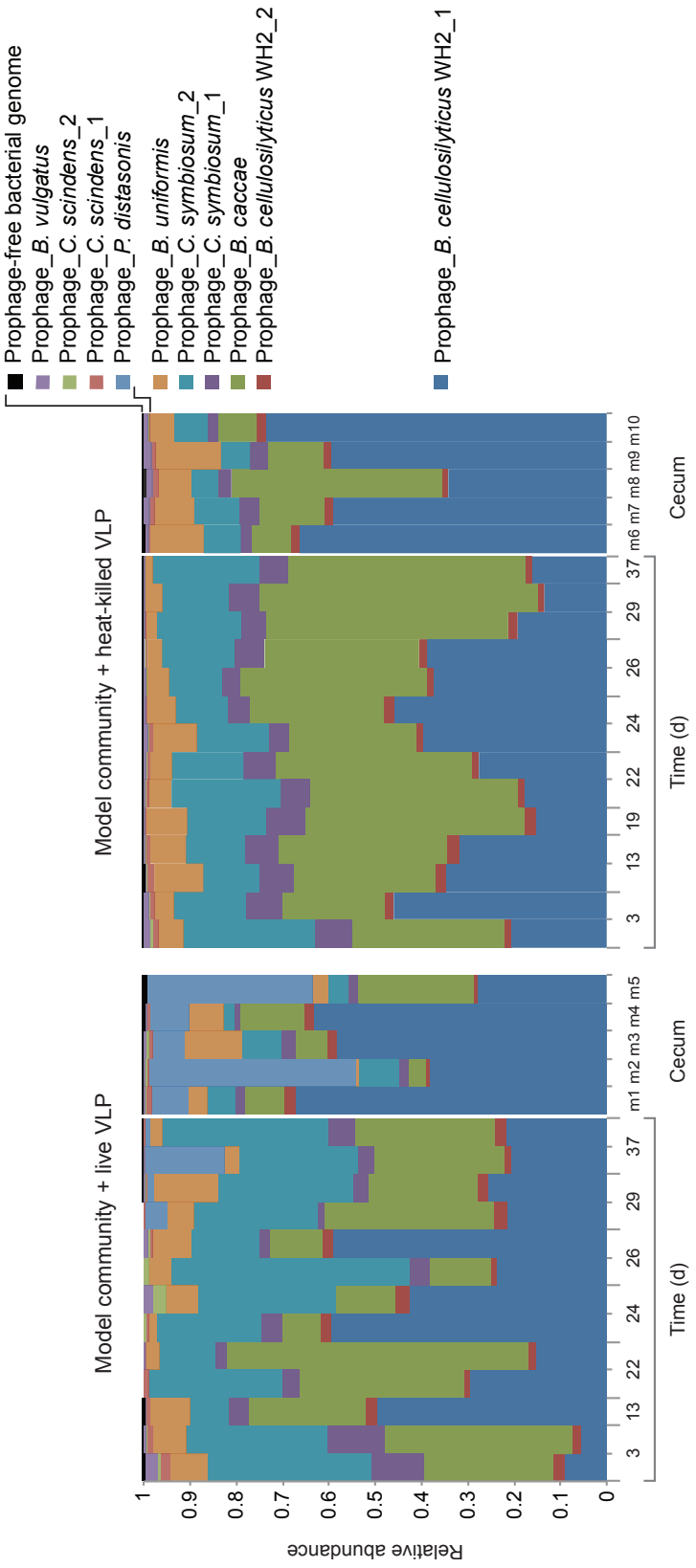

A
B
Relative abundance (%)
ϕHSC03
ϕHSC04
ϕHSC05
D. longicatena
C. spiroforme
C. aerofaciens
C. scindens
E. rectale
B. thetaiotaomicron 7330
C. symbiosum
P. distasonis
R. obeum
B. uniformis
B. caccae
B. thetaiotaomicron VPI-5482
B. ovatus
B. vulgatus
B. cellulosylticus WH2
Group 1 (live VLPs)
Group 2 (heat-killed VLPs)
m1
m2
m3
m4
m5
m6
m7
m8
m9
m10
PSI
DSI
Ce
Co
Fe
ND

Fig. S8


Prophage-free bacterial genome
Prophage_B. vulgatus
Prophage_C. scindens_2
Prophage_C. scindens_1
Prophage_P. distasonis
Prophage_B. uniformis
Prophage_C. symbiosum_2
Prophage_C. symbiosum_1
Prophage_B. caccae
Prophage_B. cellulosilyticus WH2_2
Prophage_B. cellulosilyticus WH2_1
Model community + live VLP
Model community + heat-killed VLP
Relative abundance
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
3
13
22
24
26
29
37
Time (d)
m1 m2 m3 m4 m5
m6 m7 m8 m9 m10
Cecum

Supplementary Tables

Table S1 - Genomic features of the 15-member community of human gut bacterial symbionts, including their prophage and CRISPR elements.

Table S2 – Body and epididymal fat pad weights at the time of sacrifice.

Table S3 - Samples used for Community Profiling Sequencing (COPRO-Seq).

Table S4 - Description of samples used for VLP-purification and 454 shotgun pyrosequencing.

Table S5 – Annotation of assembled human gut-derived phage genomes identified in the fecal microbiota of gnotobiotic mice. Annotations are based on the best blast hit (blastp, e-value cut-off 10^{-5}) of each predicted protein against six independent protein databases (COG (STRING v.9.0), KEGG (v58), ACLAME (v0.4), CDD (online version; cut-off 10^{-2}), NCBI nr (retrieved 13/08/2012), and Phantom (retrieved 01/09/2012)). (**a**) phage ϕHSC01. (**b**) phage ϕHSC02. (**c**) phage ϕHSC03. (**d**) phage ϕHSC04. (**e**) phage ϕHSC05.

Table S6 – List of *Bacteroides spp* genomes retrieved from HMP webserver for homology search.

Table S7 – List of primers, barcodes and adapters used.

Supplementary References

21 Goodman, A. L., Wu, M., & Gordon, J. I., Identifying microbial fitness determinants by insertion sequencing using genome-wide transposon mutant libraries. *Nat Protoc*. 6, 1969-1980 (2011).

22 Singh, A. & McFeters, G. A., Survival and virulence of copper- and chlorine-stressed Yersinia enterocolitica in experimentally infected mice. *Appl Environ Microbiol*. 53, 1768-1774 (1987).

23 Niu, B., Zhu, Z., Fu, L., Wu, S., & Li, W., FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics*. 27, 1704-1705 (2011).

24 Rey, F. E. *et al*., Dissecting the in vivo metabolic potential of two human gut acetogens. *J Biol Chem*. 285, 22082-22090 (2010).

25 Caporaso, J. G. *et al*., QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 7, 335-336 (2010).

26 Chevreux, B., Wetter, T., & Suhai, S., Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)*. 99, 45-56 (1999).

27 Milne, I. *et al*., Using Tablet for visual exploration of second-generation sequencing data. *Briefings in bioinformatics*. doi 10.1093/bib/bbs012 (2012).

28 Delcher, A. L., Harmon, D., Kasif, S., White, O., & Salzberg, S. L., Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*. 27, 4636-4641 (1999).

29 Li, W. & Godzik, A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 22, 1658-1659 (2006).

30 Team, R. C., R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2012).

31 Grissa, I., Vergnaud, G., & Pourcel, C., CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res*. 35, W52-57 (2007).

32 Magoc, T. & Salzberg, S. L., FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 27, 2957-2963 (2011).

Chapter 4

Preliminary analysis of the gut viromes of healthy and malnourished twins and their family members, and future directions

Introduction

As noted in Chapters 2 and 3, surveys of DNA viruses associated with the gut microbiota of healthy adults have revealed a high prevalence of bacteriophages, in particular lysogenic phages [1-3]. There are a number of reasons why a temperate lifestyle can be beneficial. For instance, prophages provide the bacterial host with superinfection immunity, preventing other similar viruses to infect the host; phages are capable of transferring bacterial genes and functions with potential fitness advantages to new hosts. Additionally, 'suicide bomber' refers to a strategy whereby a bacterial host under stress (e.g. from signals that lead to DNA damage) induces its prophage leading to lysis of other susceptible bacterial hosts except those that harbor the same virus and hence are resistant to the infection; this opens niche for the surviving resistant bacteria, and provides nutrients that are the products of bacterial cell lysis [4,5].

One consequence of a lysogenic lifestyle is that alpha-diversity (the number of different virotypes) in a given sample is low. Coupled with its high stability over time in healthy individuals, the gut virome is a potential biomarker of alterations in the health status of the human host. The configuration of the bacterial component of the human gut microbiota is influenced by a number factors, including host diet [6-8], cultural traditions, and age [9,10]. The microbiota of healthy individuals evolves to an adult like configuration during the first three years of life [9].

Under-nutrition is the leading cause of child mortality worldwide [11], Moderate Acute Malnutrition (MAM) is defined by a weight-for-height z score (WHZ) between two and three standard deviations below the median established by the World Health Organization childhood growth standards [12]. Severe acute malnutrition (SAM) refers to either marasmus, characterized by extreme wasting with WHZ scores below 3 standard deviations, or kwashiorkor, which is associated with generalized edema, hepatic steatosis, skin rashes and ulcerations, and anorexia [13,14]. Studies of twins discordant for SAM revealed that (i) undernutrition is associated with a more immature microbiota, (ii) transplantation of fecal microbiota obtained from co-twins with kwashiorkor and their healthy co-twins into separate groups of germ-free mouse recipients transmits discordant weight

loss and microbiota-associated metabolic dysfunction to the recipient animals; and (iii) treatment with a therapeutic food produces transient rescue of these metabolic abnormalities but the system regresses after treatment is withdrawn. Together, these findings suggest that the microbiome/microbiota is causally related to undernutrition [10]. Based on these findings, I hypothesized that the gut virome may be a useful biomarker for identifying features of the microbiota associated with SAM and its degree of repair by various therapeutic interventions including food-based therapies.

Results

In this Chapter, I describe preliminary findings from a metagenomic study of the fecal viromes of eight monozygotic and 12 dizygotic Malawian twin pairs between 0-30 months of age (**Supp Table 1, Fig. 1**). Twelve of these twin pairs were discordant SAM: in six cases, a co-twin in the discordant pair developed kwashiorkor while the other co-twin remained healthy; in the other six cases, a co-twin in the discordant pair developed marasmus and the other co-twin remained healthy. Based on current clinical practices in Malawi, when one co-twin in a discordant pair is diagnosed with SAM both co-twins are treated with a peanut-based Ready-to-Use Therapeutic Food (RUTF). In this study, following a period of treatment for 4-8 weeks, children were returned to their traditional diets and followed until they reached 36 months of age. Fecal samples, collected at time points encompassing the period just before, during and after the nutritional intervention, were selected for the current metagenomic study of the gut virome (**Fig. 1**).

Extracting VLPs from small amounts of fecal microbiota required modifications to the original protocol described in Chapter 2. The new protocol, developed for extracting VLPs from mouse fecal samples (see Chapter 3), is suitable for purifying VLP-derived DNA from as little as 50mg of frozen feces (see Chapter 3).

The current study sequenced 231 VLP DNA samples (**Table S1**; 56,178±2629 (average ± SEM) shotgun 454 pyrosequencer reads per sample; Titanium chemistry). Previously, using raw pyrosequencing reads from samples only allowed analysis and interpretation of approximately 20% of the data (Chapter 2). However, I was able to assemble large contigs, in part due to the

extreme diversity between virotypes and the low number of virotypes per sample. Applying improved assembly strategies revealed that a large percentage of the shotgun reads could be assembled yielding complete or almost complete viral genomes. I obtained a total 17,676 contigs ≥500nt after collapsing contigs that were more than 90% similar over 90% of the contig length. As described below, these contigs allowed me to map 85±9% of the reads (**Fig. 2A**).

Analyzing the size distribution of the contigs as a function of coverage (**Fig. 2B**), I identified contigs up to 200,000bp with a large number of contigs between 50-100kb (i.e., the expected length of common bacteriophages). There was an over-representation of circular 3Kb and 6Kb contigs assigned to the Anelloviridae or Microviridae viral families (both ssDNA circular viruses). As noted in Chapter 2, whole genome amplification (MDA) of purified VLP DNA has inherent biases for single stranded circular DNAs [15].

De-replication and clustering of contigs assembled from all individual samples produced an 'overall' VLP-derived contig set. Mapping pyrosequencing reads to this full set of contigs allowed me to generate a 'sample-by- VLP-derived contig' abundance matrix equivalent to an OTU table. This matrix, which was normalized to contig length and the depth of sequencing per sample, was used for intra- and inter-sample diversity analysis.

An initial analysis of the inter-sample variation was performed using hierarchical clustering from a Hellinger distance matrix (**Fig. 3**) calculated from the contig abundance table. The results showed clear clustering by twin pairs, indicating a shared virome composition. There was no statistically significant clustering by health status or by zygosity. Only two branches were composed of samples from different families; the first branch was composed of mothers and older siblings and was independent of family of origin, suggesting that the mature/adult virome has a significantly different composition than the fecal viromes of infants between 0 and 3 years of age. The other branch containing samples from different families was composed of six of the 11 youngest (<5 months old) donors, suggesting that very early environmental exposures leads to a virome profile which is significantly different from that observed later during microbiota assembly. A few

other individual samples did not cluster and appeared as isolated branches; these samples corresponded in most cases to the youngest or oldest samples from a given twin-pair, re-enforcing the important role that age plays in determining an individual's early gut virome profile.

I compared the average Hellinger distances measured within an individual over time as well as between co-twins, and between twins and their mothers or older siblings (**Fig. 4A**). Highest similarity was observed within an individual over time, followed by between co-twins independent of their zygosity. Significantly greater differences were observed between a co-twin and their mother or older sibling, or between any two unrelated individuals. These results agree with what I had previously observed in monozygotic healthy adult twin pairs living in the USA (Chapter 2), except that for healthy adult twin pairs, the co-twin-co-twin distance were almost as different as the co-twin-mother distance, while in infants co-twin-co-twin distance although significantly different from self-self are still more similar than co-twin-mom or co-twin-sibling distances. A potential explanation is that individual adults have distinctive viromes that are very stable over time and two non-cohabitating adults lack a history shared environmental exposures while the infants in this study have a deeply shared history of environmental exposures.

Self-self and co-twin-co-twin comparisons revealed greater similarities in the case of individuals with SAM (kwashiorkor or marasmus) compared to healthy individuals (**Fig. 4B**). This observation is in agreement with the finding that there is delayed maturation of the microbiota of individuals with kwashiorkor [10] and hence less change as a function of time. This observation could also be accentuated by an increase in the number of virotypes present in a given individual, since having more viruses at any time point will increase the probability of sharing virotypes over time and with related individuals (see below for the results of alpha-diversity analysis).

Principal coordinate analysis is a widely used statistical method that aims to summarize the principal drivers of variation among samples in orthogonal vectors which are then sorted by their capacity to explain variation in the dataset. The first principal coordinate (PC1) explains the highest percentage of variation. In the Malawian infant/child virome, PC1 and PC2 accounted for

8.4% of the total variation (**Fig. 5**). When each axis was plotted as a function of age, it becomes obvious that PC1 separates late time points (> 18 months old) while PC2 separates the early time points (< 18 months old) and that together they define an age gradient for the fecal virome similar to what is observed for the bacterial component of the fecal microbiota [9].

The correlation between virome and age led me to search for viral contigs or virotypes that are biomarkers of a given age group. Applying mutual information analysis, I identified a set of 453 contigs whose presence or absence were significant discriminators of age independent of their sample of origin. **Fig. 6** shows six different clusters of age-responsive clusters; half decrease their representation over time while the other half increases their abundance, including the 'latest' cluster that is only correlated with VLP samples obtained from mothers or older siblings and a cluster composed only of contigs associated with the Anelloviridae family (see below).

Mutual information analysis was then used to identify VLP-derived contigs significantly associated with a given family, individual, or the 'mothers/siblings' cluster described in **Fig. 3**. The result was 1245 contigs that were significant discriminators of families or individuals (**Fig. 7**); within this set of contigs, 20 were significant discriminators of mothers and older siblings, and were essentially absent from any of the infant samples. Furthermore, there was a difference in the number of discriminatory contigs for kwashiorkor or marasmus families compared to healthy families. I postulated that this could reflect larger viral alpha-diversity in the former families or that there were virotypes significantly associated with SAM. Therefore, I estimated alpha-diversity based on the number of observed VLP-derived contigs in an abundance matrix normalized to the length of the contig as well as rarefied to 10,000 pyrosequencing reads per sample. In agreement with the discriminatory contig heatmap, diversity within samples was lower for twin pairs that remained healthy through the study than for twin pairs where one of the co-twins developed kwashiorkor or marasmus (**Fig. 8A,** Kruskal-Wallis non-parametric multiple test). This observation was also confirmed when quantifying the number of discriminatory contigs for healthy versus SAM (**Fig. 8B**), although the differences between healthy and marasmus were not statistically different, probably due to the limited number of observations (i.e. the number of discriminatory contigs

per twin-pair as opposed to the observed number of VLP contigs per sample). Combining these observations with the fact that the discriminatory contigs (i) are not conserved for all twins pairs containing a co-twin with SAM (marasmus or kwashiorkor) but rather are specific to individuals or twin-pairs, and (ii) are present at most time points for a given twin-pair and not just when malnutrition is manifest in a co-twin, suggests that the higher viral alpha-diversity observed at early ages in these families can be used as a biomarker of significant risk for disease.

The question that still remains is if there are any significant changes or effects on the virome due to severe malnutrition episodes or in response to the RUTF treatment Given that most of the variation on the dataset is explained by family relationships and age, any further analysis needs to control or take into account these two factors. The disadvantage of such analysis is that it reduces the number of samples per treatment, and hence the statistical power to identify significant correlations. Nonetheless, a preliminary principal coordinate analysis was performed on each family aimed at identifying changes as a function of time and health status. In this case, I analyzed changes in PC1 (**Fig. 9**). The position for healthy co-twins along PC1 correlates with age, while in twin pairs discordant for SAM this behavior is not always observed; certain twin-pairs exhibit variation along the axis even in opposite directions as a function of time. Together, these observations are consistent with the notion that healthy twin-pairs undergo a more stable maturation process of their microbiota, including their virome.

Annotation of the infant Malawian fecal virome

Having demonstrated that contig occurrence can be significantly associated with individual families, that their presence/abundance can varies as a function of age, that differences in health status correlate with differences in alpha-diversity among fecal virome samples, and that beta-diversity measurements are capable of tightly clustering samples derived from the same individual over time, I next sought to describe the types of viruses represented by the assembled contigs.

Previous attempts to annotate viral sequences and viral reads derived from human gut associated samples have generally classified less than 20% of the data, even when using non-redundant

viral databases [2]. However, the availability of larger contigs in this study increased my chances of achieving accurate annotation.

As noted above, my dataset was enriched for circular contigs (potential indicators of complete viral genomes) at ~6Kb and ~3Kb (**Fig. 2B**). Recent reports of human gut viromes have shown the presence of ~6Kb circular contigs [3,16] associated with a recently described subfamily of the Microviridae known as Alpavirinae [17]. The Microviridae are ssDNA circular bacteriophages with a previously characterized lytic life cycle; they were first associated with intracellular bacterial pathogens and gamma-Proteobacteria such as *Escherichia coli* (i.e. coliphage PhiX 174). The Alpavirinae were initially described as a new subfamily based on characterization of prophages present in the genomes of Bacteroides [17], a commonly represented genus in the human gut microbiota. Together, these observations led me to attempt to identify contigs that could correspond to full or partial genomes associated with the Microviridae and its subfamily Alpavirinae. This effort yielded 395 contigs with significant similarity between their predicted proteins and proteins encoded by known members of the Microviridae (**Supp. Table 2**, and **Fig. 10A**): these 395 contigs including most of the circular contigs observed in the ~6Kb range, confirming their high abundance in the human gut. When analyzing their distribution among the different human families and individuals (**Fig. 10B**), I observed that a few contigs were highly prevalent among different families while most others were markers of individuals or even specific time points for a given person. Furthermore, a large number (17%) of these contigs were only present in mothers and older siblings, indicating that Microviridae and in particular Alpavirinae viruses are associated with a mature gut microbiota.

The other set of circular contigs highly represented in the current dataset were ~3Kb. Retrieval of Anelloviridae specific proteins from public databases followed by blasting against the full set of assembled contigs produced 2414 contigs with significant similarity to members of the Anelloviridae family (e.g. Torque Teno Viruses TTV), including most of the ~3kb circular contigs (**Fig. 10A**). The Anelloviridae family is composed of non-enveloped, ssDNA circular viruses that were first described in 1997 [18]. This family is dominated by TTV (Torque teno virus), TTMV

(Torque Teno Mini Virus) and TTMDV (Torque Teno Midi virus) and is currently represented by 39 distinct genotypes, characterized by their high genetic variability. Since their initial description, anelloviruses have been identified as chronic infecting viruses with no proven causal role in disease [14,18]. Initially isolated from serum, they have subsequently been found in many tissues and body sites, including fecal samples. PCR based studies have identified anelloviruses with prevalence of up to 70% in Africa and Asia and 40% in the USA and Europe [19]. Anelloviruses have been found in young infants [18], including documentation of co-infection with different anelloviruses in feces [20]. The great number of different Anellovirus genomes identified in the present study and the degree of co-infection and persistence observed are new findings (**Fig. 10C**). Interestingly, the abundance of Anellovirus in the current dataset decreases as a function of age, with almost complete absence from fecal samples obtained from mothers and older siblings (**Fig. 6**). Moreover, the low abundance of these viruses in family F57, the oldest twin pairs in the study, recruited at ~20 months of age and sampled until they were ~30 months old (**Fig. 1**), supports the conclusion that the prevalence of Anellovirus in fecal samples obtained from this cohort decreases rapidly around the second year of age.

My preliminary annotation of viral contigs focused mainly on those contigs that showed significant discriminatory power among different families. Blastx has been performed between contig sequences and a de-replicated protein database composed of viral proteins from the NCBI viral refseq database and prophage annotated proteins deposited on Phantome (http://www.phantome. org/). Although comparisons were made against all known viral proteins, the only significant hits were to ssDNA viruses (Microviridae, Anelloviridae, Circoviridae, Geminiviridae) or dsDNA bacteriophages (Caudovirales) (**Fig. 11**). The presence of other eukaryotic viruses in these fecal samples at lower abundances has been confirmed by amplification and sequencing of two novel polyoma viruses recovered from two of the samples by members of David Wang's laboratory [21,22]. In total, 168 contigs in this subset of 1245 discriminatory contigs did not show any significant similarity to any currently known viral protein. The distribution of the contigs among the different families and categories revealed that for mothers and siblings, all the discriminatory contigs were

either Microviridae bacteriophages or caudovirales, indicating that adult microbiota is dominated by bacteriophages, as was previously reported. There was no significant association of any other viral family to either a given human family or to health status, suggesting that the increased diversity observed in marasmus or kwashiorkor twin-pairs cannot be attributed to a specific viral type.

Global conservation of the human gut virome

As noted above, my previous analyses comparing the viromes of healthy USA monozygotic twin pairs and their mothers showed that virotypes were highly conserved within an individual over time but were highly variable between individuals, suggesting that the viromes constituted personal fingerprints [2] (Chapter 2). This observation is a consequence of the low alpha-diversity (low predicted number of virotypes per individual) and high beta-diversity among individuals (very low overlap of the virotypes in terms of presence and abundance between individuals). However, in the staged phage attack of a model defined human gut microbiota described in Chapter 3, one of the captured phages was present at high abundance in 4 of the 5 human donors used to generate the VLP pool used for the attack, indicating that there are shared viruses or virotypes among different individuals. To examine this notion further, I identified a set of 348 contigs that were present in at least 20% of the Malawian fecal samples (**Fig. 12**), indicating that their corresponding viruses can be shared among different individuals representing different families. Although I have not corroborated that these families share the exact viral strain with the same host specificity, the abundance matrix was built based on reads with sequence identity to a given contig of ≥95% over at least 50% of the length of the read. Furthermore, some of these VLP-derived contigs are also conserved among mothers and older siblings indicating their presence in infant as well as mature gut communities.

A further step in assessing the global diversity of these viral contigs was to compare those assembled from the Malawi dataset to the 88 large contigs (>10kb) I assembled during my analysis of the fecal viromes of adult USA monozygotic twin pairs [23]. A total of 11 of these 88 contigs were also present in the Malawian dataset (**Supp. Table 2**). As an example, alignment of two of

the contigs from the USA twin virome study with contigs from the Malawian samples shows the high degree of sequence conservation and synteny among the respective contigs (**Fig. 13**). In addition to the 11 contigs identified from the previous study of adult healthy USA twins, two of the novel viral genomes captured as part of the staged gnotobiotic mouse phage attack (Chapter 3) were also identified among the Malawi contigs (**Supp. Table 2**). Phage ϕHSC02, belonging to the Microviridae family, and phage ϕHSC05, which was present in 4 out of the 5 adult VLP donors. This latter phage was identified in 22 samples from 13 individuals (including 7 samples from the twin pair in Family 47).

As reported on our lab's analysis of the fecal microbiomes of Malawian twins discordant for kwashiorkor [10], transplantation of their intact uncultured gut communities to separate groups of germ-free mice transmitted discordant weight loss and metabolic phenotypes: these discordant phenotypes were manifest when recipient mice received a macro- and micronutrient deficient Malawian diet, where partially rescued by RUTF and then re-emerged when animals were returned to a representative Malawian diet. This led me to characterize the degree to which the viromes associated with these human donor samples could be transmitted to recipient mice. Therefore, I analyzed VLPs obtained from mice colonized with the microbiota from human donor F56T1 (healthy co-twin) or F56T2 (co-twin with kwashiorkor), or F57T1 (kwashiorkor co-twin) or F57T2 (healthy co-twin). Mapping the VLP-derived pyrosequencing reads revealed reads that mapped to contigs present in the human donor's fecal virome (**Fig. 14**). Interestingly, in some cases, viral contigs identified in mice sampled at varying time points after gavage were not found in the original donor microbiota sample, suggesting that the corresponding viruses were either present at levels below the limits of detection in the human sample, or were present as prophages in the bacteria at the moment of gavage and were later induced.

Conclusions and future directions

The combined development of tools for isolating and characterizing viral particles from small quantities of previously frozen fecal samples, together with improvement and/or development of

novel computational methods for analyzing viral metagenomic data has allowed me to characterize an important component of the human gut microbiota. I have found that phage dominate the virome present in the guts of healthy adults living in the USA, that these phages manifest a temperate rather than the lytic lifestyle observed in other environments such as the open ocean. This temperate lifestyle is associated with low alpha-diversity: i.e. the virome is composed of a few abundant virotypes whose nucleotide sequences are highly conserved within an individual over time. In contrast, Malawi twins under two years of age possess a fecal virome that is more variable over time and shares a high degree of similarity among co-twins independent of their zygosity, highlighting the importance of early environmental exposures. A remarkable finding was the great variety of eukaryotic viruses present at high abundance in the fecal microbiota of young Malawian infants that rapidly decreased after 24 months of age.

Despite the lack of viral contigs or virotypes that were significantly associated with the onset or resolution of episodes of SAM, I observed an overall expansion of the viromes in twin-pairs discordant for severe acute malnutrition (manifested by the larger number of contigs present). Although I am not aware of any metagenomic studies of humans describing a comparable expansion of the virome, expansion of the virome but not the bacterial component of the gut microbiota has been documented in non-human primates during infection with Simian Immunodeficiency Virus (SIV) [24].

Advances in DNA sequencing and analysis, combined with gnotobiotic mouse models of the human gut microbiota, provide a powerful catalyst for advancing research related to the gut virome. The result should be new and important insights about the mechanisms by which phage and other viruses shape the structure and expressed functions of the microbiota. The phage diversity present the planet is immense and largely unknown, but as we continue sampling and analyzing diverse environments, conserved proteins and viral modules will be identified that should enable a better viral taxonomy. The use of gnotobiotic mouse models and the type of staged phage attacks described in Chapter 3 of this thesis should provide new understanding of the factors that define the tropism of viruses, and how they persist within the dynamic gut ecosystem. These animal mod-

els coupled with tools such as whole genome transposon mutagenesis of bacterial hosts and high throughput Tn insertion site mapping promise to shed light on the *in vivo* conditions that modulate/ mediate prophage induction and the role of prophage in maintaining community structure and function. Finally, as noted in the Introductory chapter of this thesis, these approaches have value to both fundamental as well as applied science efforts and may yield new approaches to diagnosis as well as new strategies for microbiota-directed therapeutics.

Figure legends

Figure 1. Experimental design. Twenty Malawi families were selected to study the gut viromes of mono- and dizygotic twin pairs during the first 30 months after birth. Samples used for VLP-isolation and sequencing are depicted as black squares as a function of time. Color boxes are used to differentiate mother and older sibling samples and to denote time points where individuals developed marasmus, kwashiorkor, moderate malnutrition or were subjected to RUTF treatment.

Figure 2. Assembly contigs from shotgun datasets of Malawian fecal viromes. **(A)** Histogram of the percent of shotgun pyrosequencing reads per VLP sample that was incorporated into the assembly of ~17,000 contigs (> 500nt long). (**B)** Improvement in assembly strategy and the depth of sequencing allowed assembly of contigs up to 200,000 nt and a median coverage of up to 10,000X. Individual contigs of more than 500nt are classified as linear contigs (blue dots) or as circular contigs when there was evidence of overlapping ends (red dots). Note the distribution of circular contigs centers at 3Kb, 6Kb, and between 40-100Kb. Contig length and median coverage are shown as log 10 transformed values for visualization purposes. Median coverage is expressed in integers, hence the step increase of coverage under 10X. Median coverage corresponds to the median number of pyrosequencing reads per nucleotide position along the length of a given contig.

Figure 3. Hierarchical clustering of samples indicates a strong familial clustering. UPGMA tree generated from Hellinger distances between samples. Colors and cartoon edges represents branches from twin pairs in a single family, the only exception was F121 and F57 where a few of the later time points for both twins clustered separately. The first branch from top to bottom is not collapsed and represents samples from mothers and siblings. Another branch that includes samples from different families is composed of fecal VLPs from individuals less than 5 months old. The few other samples that did not cluster with their families are shown in black; they usually correspond to the first or last time point for a given twin pair. Initials in parenthesis next to the family ID indicate whether the twin-pair remained healthy (H) throughout the study or whether one of the co-twins in a discordant pair developed kwashiorkor (K) or marasmus (M) .

Figure 4. Inter and Intra-personal dissimilarities. **(A)** Comparison of Hellinger distances within the same individual over time (self-self comparison) or between individuals: co-twin- co-twin (in mono- or dizygotic pairs), co-twin-mother or co-twin-sibling, or unrelated individuals. A one-way ANOVA was performed and a Tukey multiple-comparison test was used to determine the significance in the observed difference between the means of the datasets. Different letters indicate groups with significant different means ($p < 0.05$). **(B)** Comparisons of intra- and interpersonal variation: samples are grouped according to the health status of their donors. A lower case (a) indicates the datasets with no significant difference in their means when compared to the self-self group from concordant healthy co-twins ($p > 0.05$), indicating that self-self and co-twin - co-twin distances in SAM groups were significantly lower than on healthy co-twins.

Figure 5. Principal coordinate analysis based on the Hellinger distances between Malawian fecal virome samples. Sample coordinates on the first two principal coordinates (PC1 and PC2) are represented as a function of time. PC1 separates viromes from donors older than 18 months or age, while PC2 shows a gradient for younger donor samples (< 18 months). Colors represent different 6-month age bins. For visualization purposes, older siblings and mothers were arbitrarily collapsed into a single age bin.

Figure 6. Subset of VLP-derived contigs having significant association with donor age. Heatmap of 453 contigs identified by mutual information analysis to be significantly associated with age. Samples are sorted by age and VLP derived contigs were clustered based on their appearance patterns. Six clusters were identified where the presence of the VLP-derived contigs decreases or increases as a function of age. The top cluster was only composed of VLP-contigs with significant similarity to the Anelloviridae family. Color intensity represents abundance of a given contig in a given sample, expressed as the log of the normalized RPMM [Reads per Million base pair (of the contig) per million reads (of the sample)] .

Figure 7. VLP-derived contigs constitute significant discriminators of twin-pairs. A heatmap of contigs that are significantly (estimated pvalue < 1e-5) associated with either a given individual

or twin-pair are shown. Columns represent individual samples and each row corresponds to a single VLP-derived contig. The leftmost set of samples correspond to the samples on **Fig. 3** that were clustered together without a familial relationship, corresponding to samples under 5 months of age. The rightmost set of samples corresponds to all the mothers and siblings present on the dataset. All other samples are organized by family; within each family they are organized from left to right by co-twin (T1, T2; separated by tick marks) and by ascending age. Intensity of the heatmap represents abundance as described in **Fig. 6**.

Figure 8. Twin-pairs discordant for kwashiorkor or marasmus exhibit an increase in the number of VLP-derived contigs. **(A)** Alpha-diversity analysis of a matrix of reads normalized to contig length and rarefied to 10,000 reads per sample. Average ± SEM of the number of observed VLP-derived contigs are plotted for each category. A Dunn's non-parametric multiple comparison rank test was performed; * corresponds to a significant difference (p-value < 0.05) between the means. **(B)** The average ± SEM for discriminatory contigs per twin-pair are plotted for each category. The same test of significance was performed. Although the difference in mean appears larger than in panel A, the limited number of observations (n=6 per category) and the variation likely contribute to the lack of statistical significance to the difference between concordant healthy co-twins and co-twins discordant for marasmus .

Figure 9. Principal coordinate analysis for Malawi twin-pairs as a function of age. The coordinates on the principal axis of variation (PC1) are plotted as a function of time for each individual twin pair. Twin-pairs are separated based on whether they remain healthy throughout the study or became discordant for kwashiorkor or marasmus. Time points where a given individual developed marasmus, kwashiorkor, moderate malnutrition or were subjected to RUTF treatment are shown in different colors.

Figure 10. Abundance of Microviridae and Anelloviriridae families in the assembled dataset. **(A)** VLP-derived contigs, represented as dots, are plotted as a function of contig length and median coverage. The figure is built as in **Fig. 2B**. In this case, contigs with significant similarity to Micro-

viridae (green) or Anelloviridae (orange) are highlighted. Note the correlation between the circular contigs at 6Kb with Microviridae and 3Kb with Anelloviridae. **(B)** Heatmap of contigs annotated as belonging to the Microviridae family. The prevalence of these contigs is higher in mothers and older siblings**. (C)** Heatmap of the abundances of Anelloviridae in the different samples and families. Anelloviruses are absent in mothers and older siblings. Sample order and color intensities for Heatmaps in panels B and C are as described in **Fig. 7**.

Figure 11. Taxonomic classification of VLP-derived contigs associated with particular families or individuals. The set of 1245 VLP-derived contigs that show significant association with a particular individual or twin-pair, was annotated by blasting against a viral non-redundant protein database (threshold blastx e-value < 1e-5). Annotations were binned at the family level for ssDNA viruses and the order level for dsDNA viruses.

Figure 12. Heatmap of VLP-derived contigs present in at least 20% of samples. A total of 348 contigs that were present in at least 46 (20%) of the analyzed samples were selected and clustered. Many of the contigs are present in multiple samples irrespective of family, age or health status. Sample order and color intensities are the same as in **Fig. 7**.

Figure 13. Multiple alignment between Malawi VLP-derived contigs and healthy adult USA twin (MOAFTS) derived VLP contigs. (A) Alignment between MOAFT contig 2803 (~12,000bp) and five other Malawian contigs, all originated from different families. Colored blocks in the alignment represent significant regions of similarity. Histograms within each block represent the percent identity of the particular sequence. Note that the complete 13 kb region is present in same orientation and synteny in all five Malawi derived contigs. **(B)** Alignment between MOAFTS-derived contig 1331 (~13 kb) and eight different Malawian-derived contigs. In this case, the contig is divided into three different colored blocks since in the case of some Malawian contigs the fragment spans the start and end of the assembled contig, suggesting a complete circular viral genome. In addition, six of the Malawian contigs are ~58Kb (in the range of the average size of common bacteriophage genomes).

Figure 14. VLP-contigs transferred to germ-free mice. Heatmap of the relative abundance of contigs in VLPs prepared from fecal microbiota of co-twins in families 56 and 57 and in VLPs isolated from the fecal microbiota of gnotobiotic mouse recipients of these human donor samples. For each human donor, the first set of columns (surrounded by a black box) represent the abundances of contigs present at each of the time points (sorted by age, left to right). The black arrow indicates the specific time point sample used to gavage germ-free mice. Following the human samples, one or more columns represent the abundances of VLP-derived contigs identified in the fecal microbiota of gnotobiotic mice sampled at different time points. Color intensities are as described in **Fig. 6**.

Figures

Figure 1.

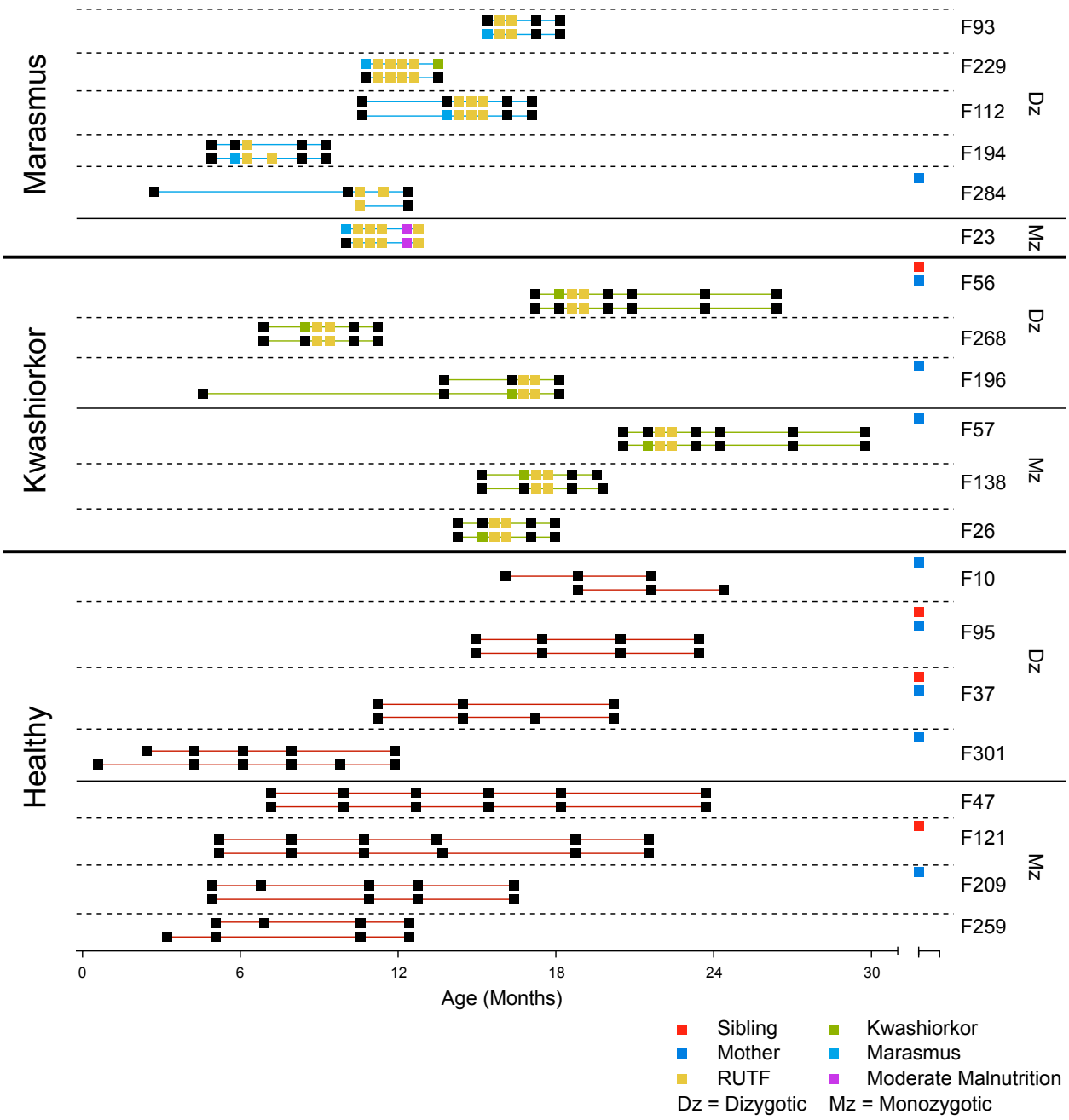


Figure 2.

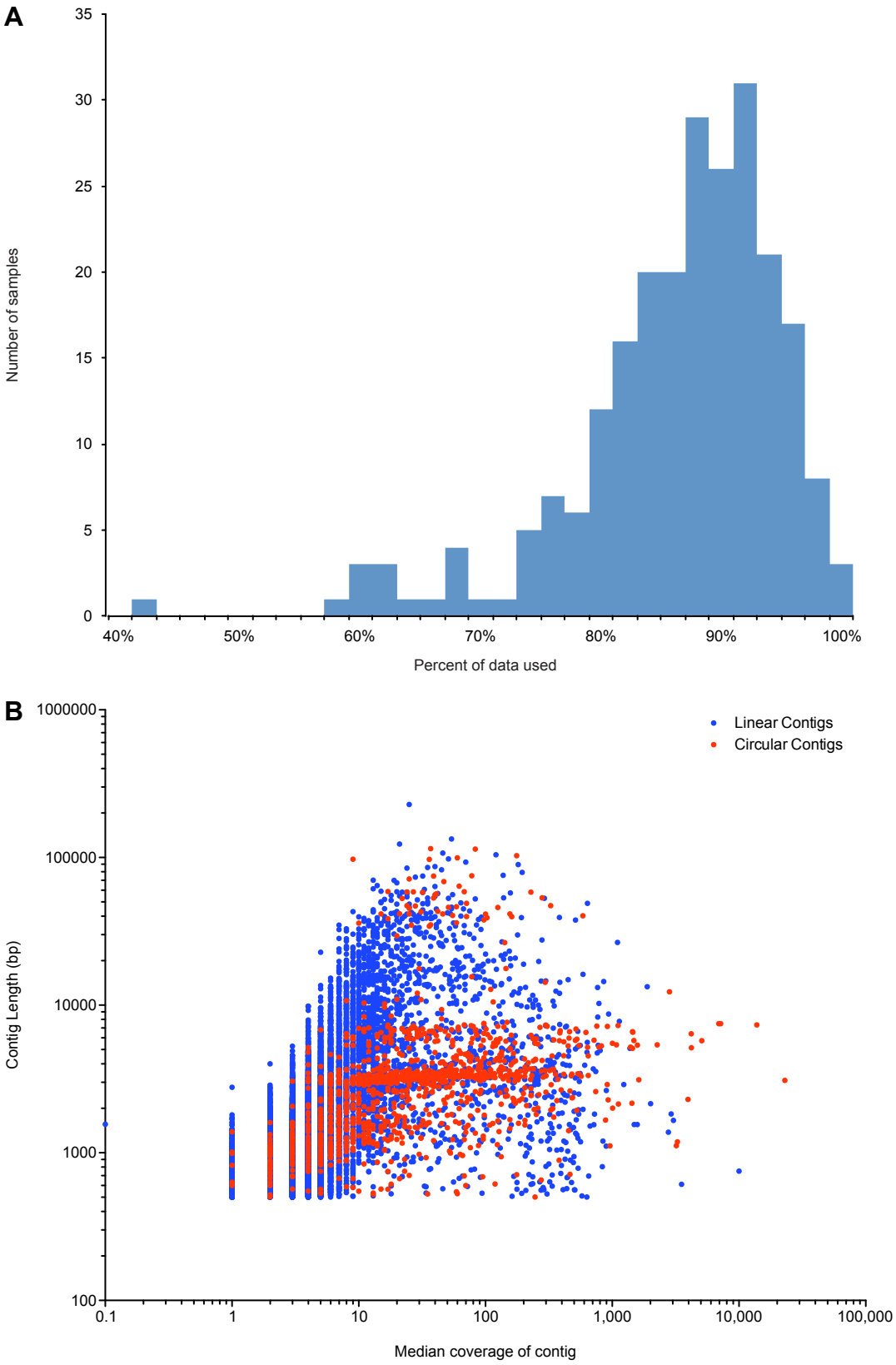

A
Number of samples
Percent of data used
40%
50%
60%
70%
80%
90%
100%
0
5
10
15
20
25
30
35
B
Contig Length (bp)
Median coverage of contig
Linear Contigs
Circular Contigs
100
1000
10000
100000
1000000
0.1
1
10
100
1,000
10,000
100,000

Figure 3.

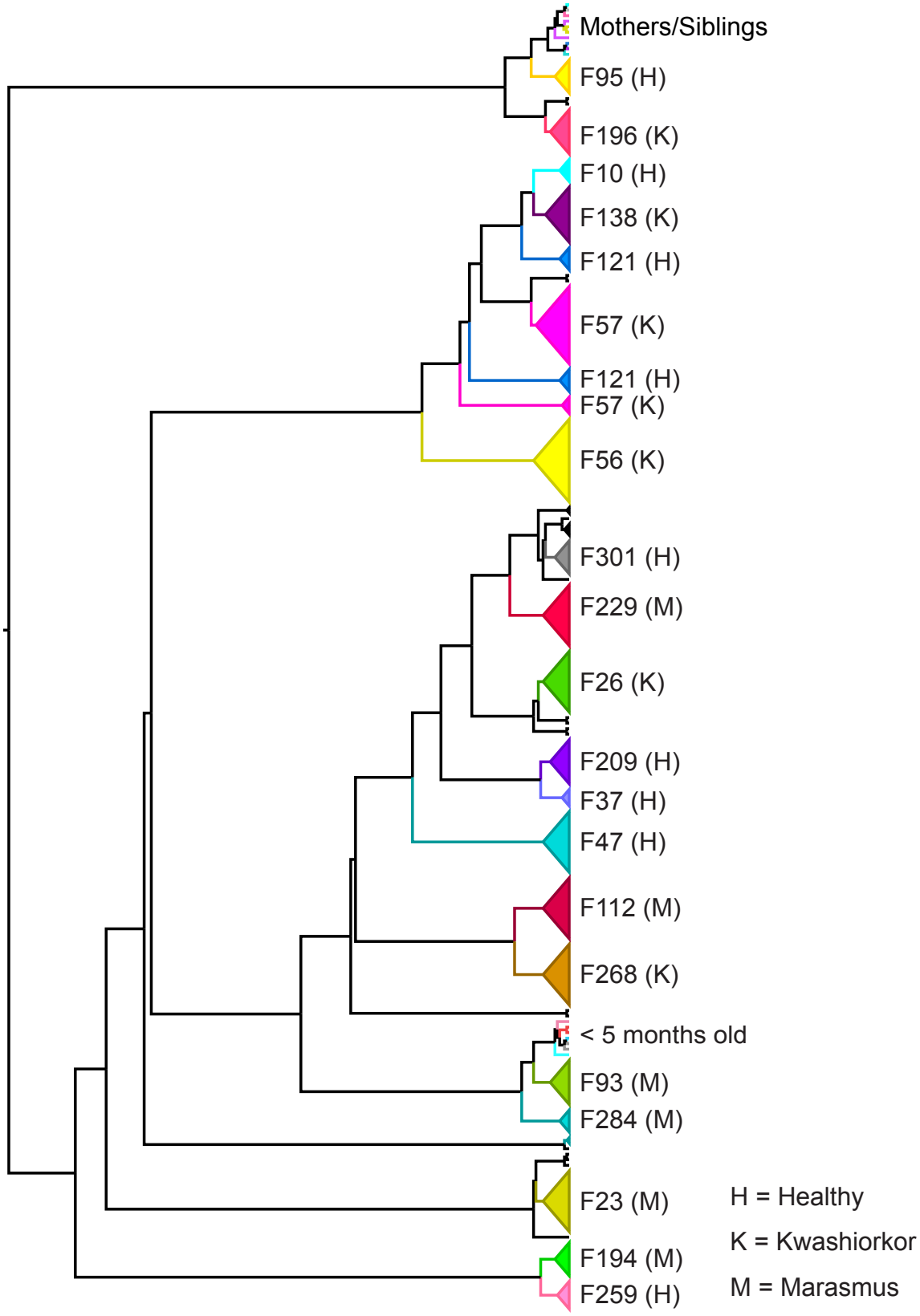

Mothers/Siblings
F95 (H)
F196 (K)
F10 (H)
F138 (K)
F121 (H)
F57 (K)
F121 (H)
F57 (K)
F56 (K)
F301 (H)
F229 (M)
F26 (K)
F209 (H)
F37 (H)
F47 (H)
F112 (M)
F268 (K)
< 5 months old
F93 (M)
F284 (M)
F23 (M)
F194 (M)
F259 (H)
H = Healthy
K = Kwashiorkor
M = Marasmus

Figure 4.

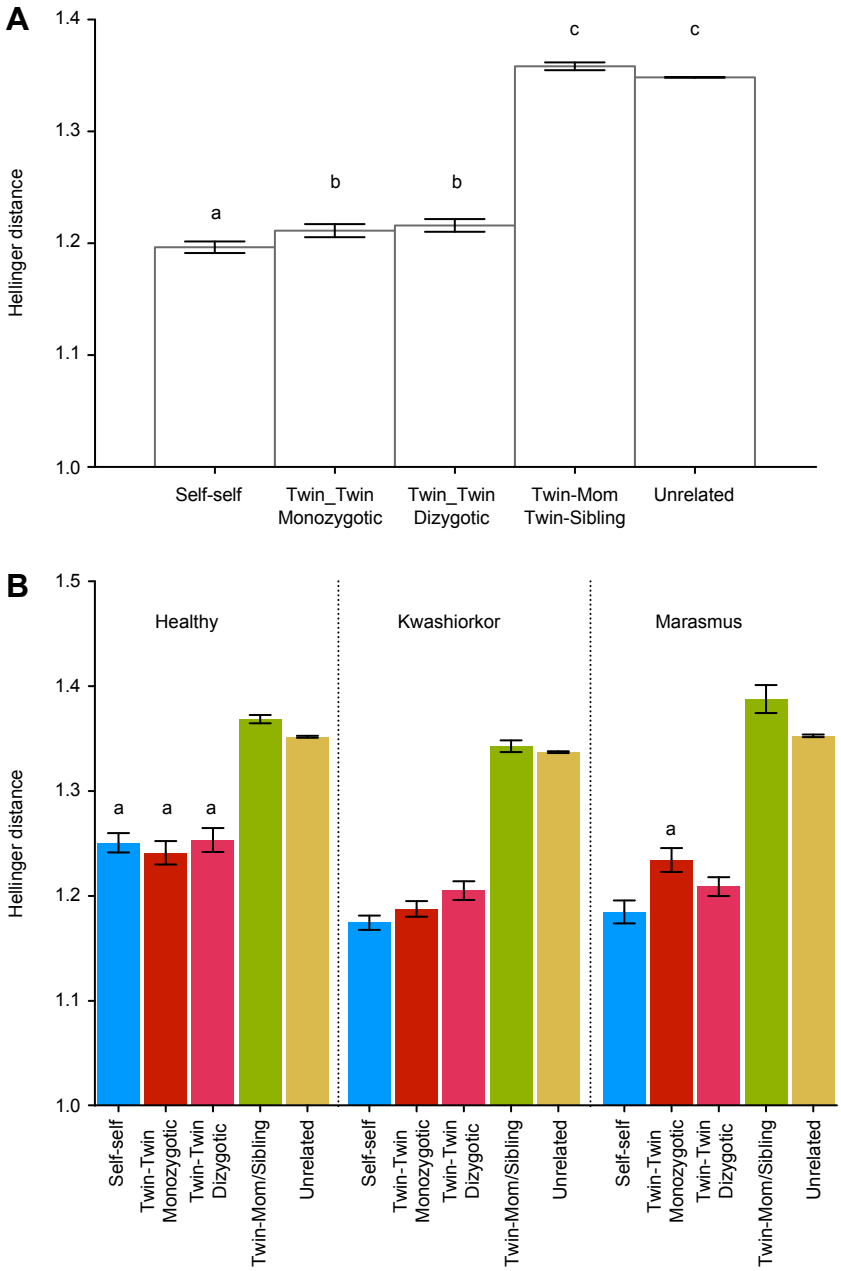

A
Hellinger distance
1.4
1.3
1.2
1.1
1.0
a
b
b
c
c
Self-self
Twin_Twin Monozygotic
Twin_Twin Dizygotic
Twin-Mom Twin-Sibling
Unrelated
B
Hellinger distance
1.5
1.4
1.3
1.2
1.1
1.0
Healthy
Kwashiorkor
Marasmus
a
a
a
a
Self-self
Twin-Twin Monozygotic
Twin-Twin Dizygotic
Twin-Mom/Sibling
Unrelated

Figure 5.

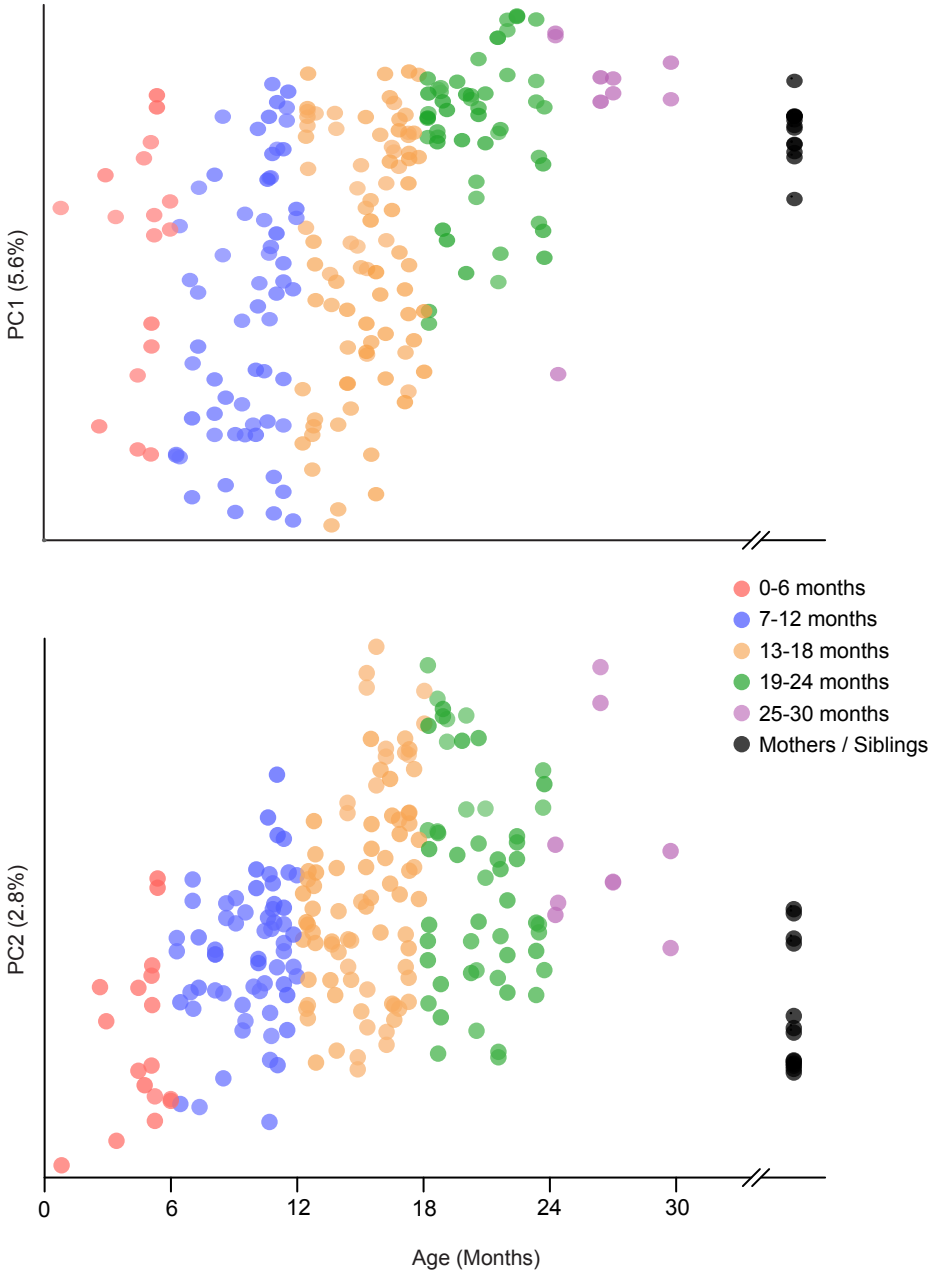

PC1 (5.6%)
PC2 (2.8%)
0-6 months
7-12 months
13-18 months
19-24 months
25-30 months
Mothers / Siblings
0
6
12
18
24
30
Age (Months)

Figure 6.

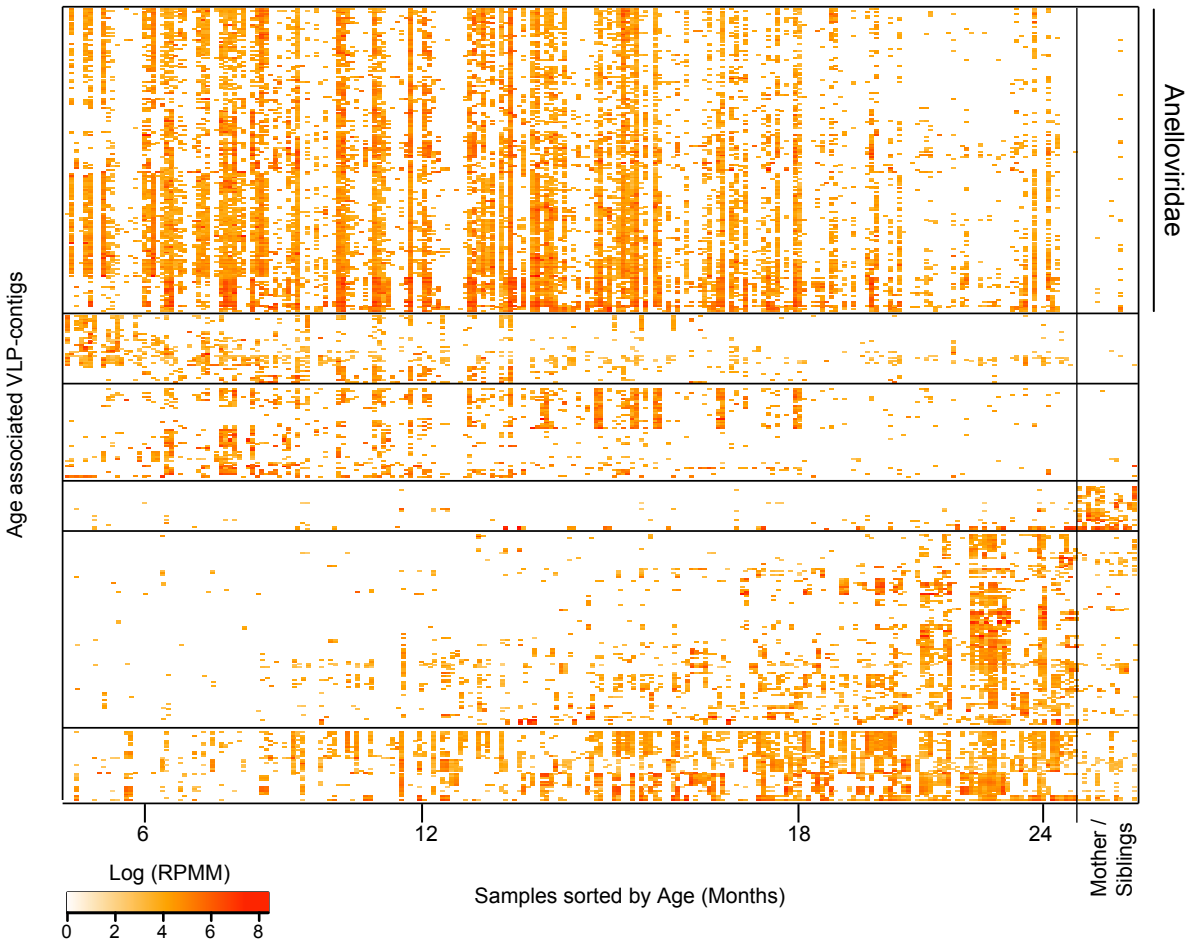

Anelloviridae
Age associated VLP-contigs
6
12
18
24
Mother / Siblings
Samples sorted by Age (Months)
Log (RPMM)
0
2
4
6
8

Figure 7.

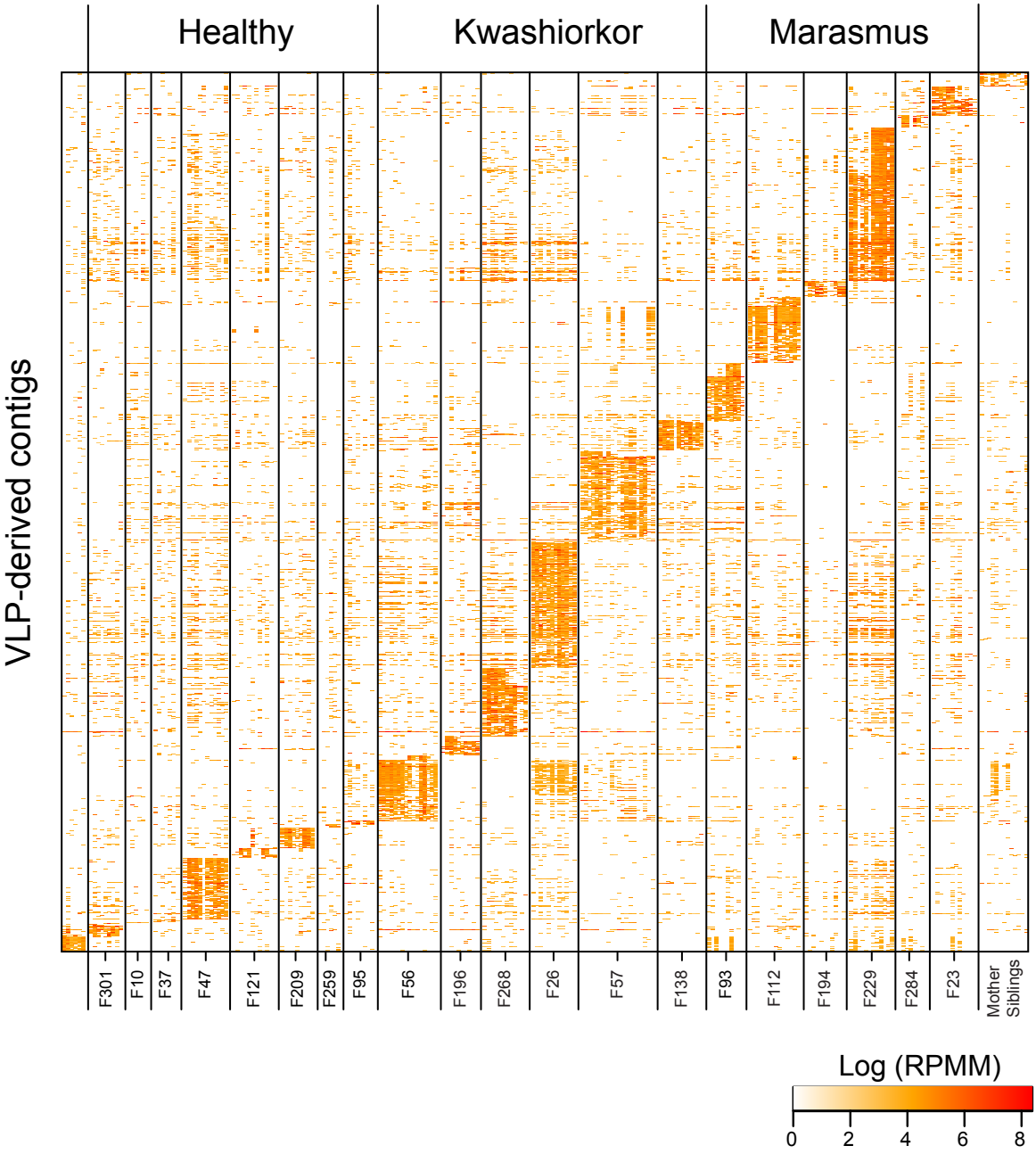

Healthy
Kwashiorkor
Marasmus
VLP-derived contigs
F301
F10
F37
F47
F121
F209
F259
F95
F56
F196
F268
F26
F57
F138
F93
F112
F194
F229
F284
F23
Mother
Siblings
Log (RPMM)
0
2
4
6
8

Figure 8.

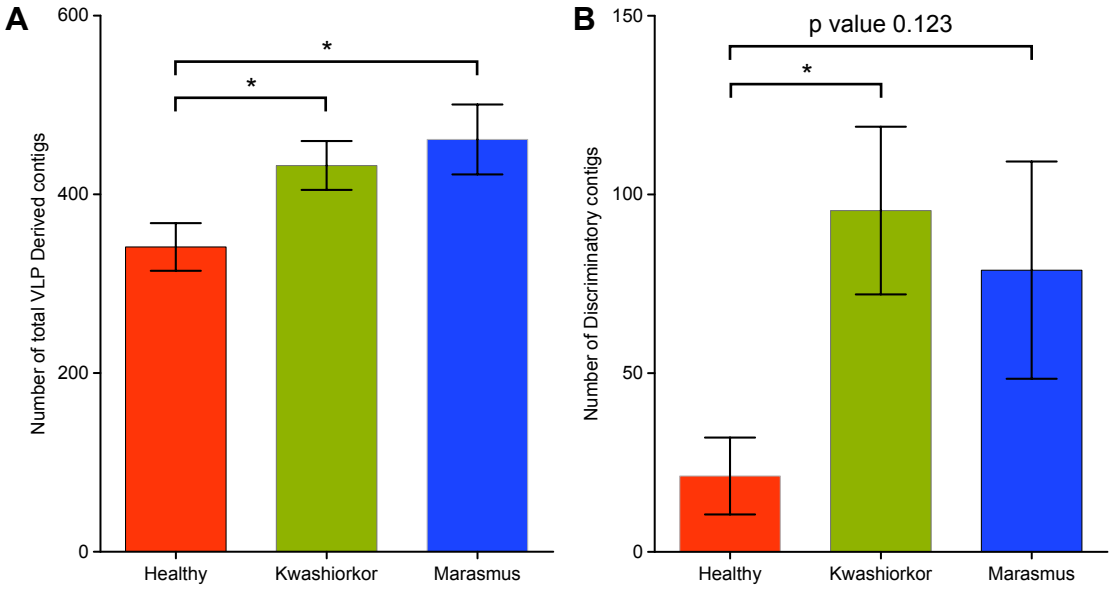

A
600
400
200
0
Number of total VLP Derived contigs
*
*
Healthy
Kwashiorkor
Marasmus
B
150
100
50
0
Number of Discriminatory contigs
p value 0.123
*
Healthy
Kwashiorkor
Marasmus

Figure 9.

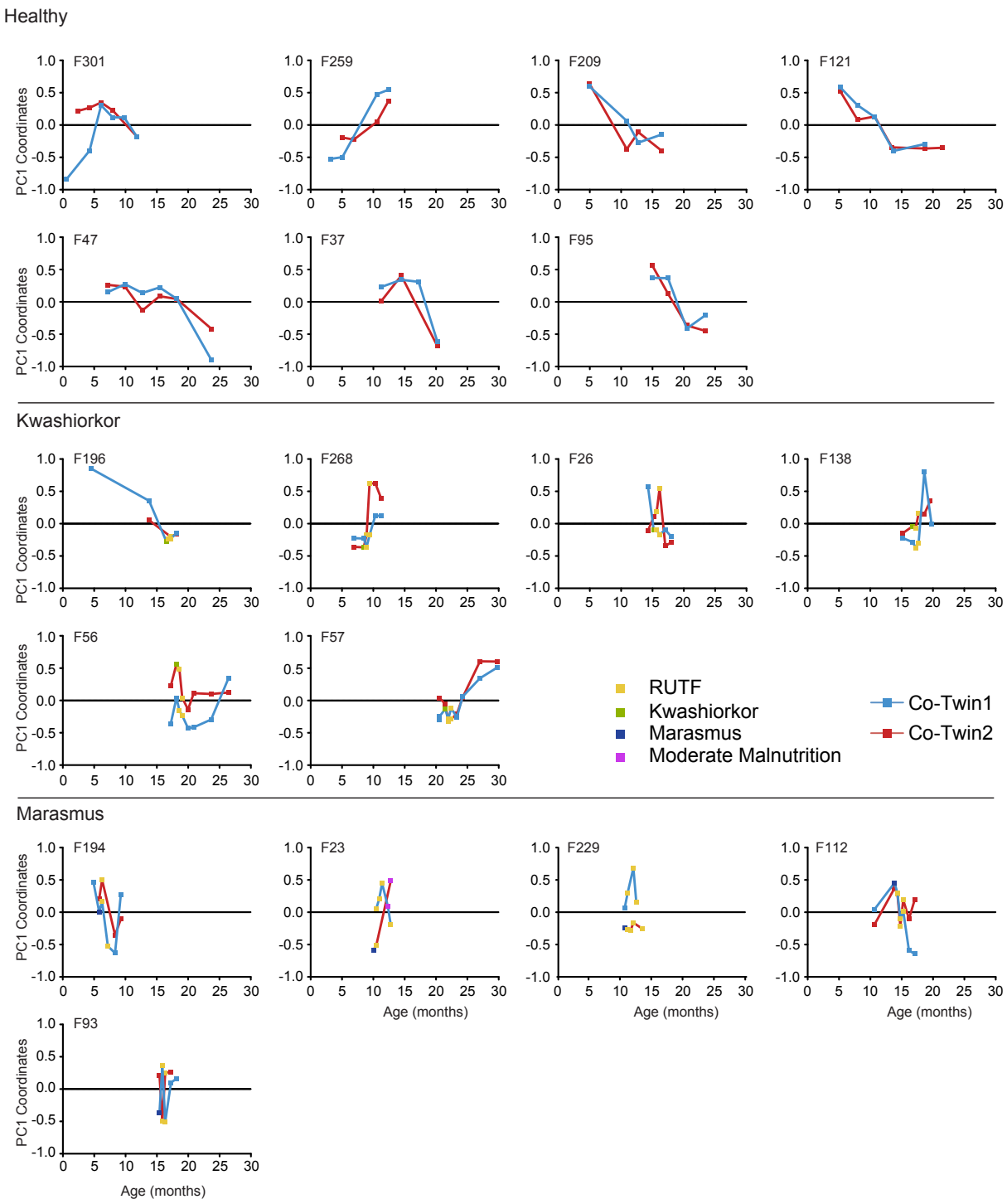

Healthy
F301
F259
F209
F121
F47
F37
F95
Kwashiorkor
F196
F268
F26
F138
F56
F57
Marasmus
F194
F23
F229
F112
F93
PC1 Coordinates
Age (months)
RUTF
Kwashiorkor
Marasmus
Moderate Malnutrition
Co-Twin1
Co-Twin2

Figure 10.

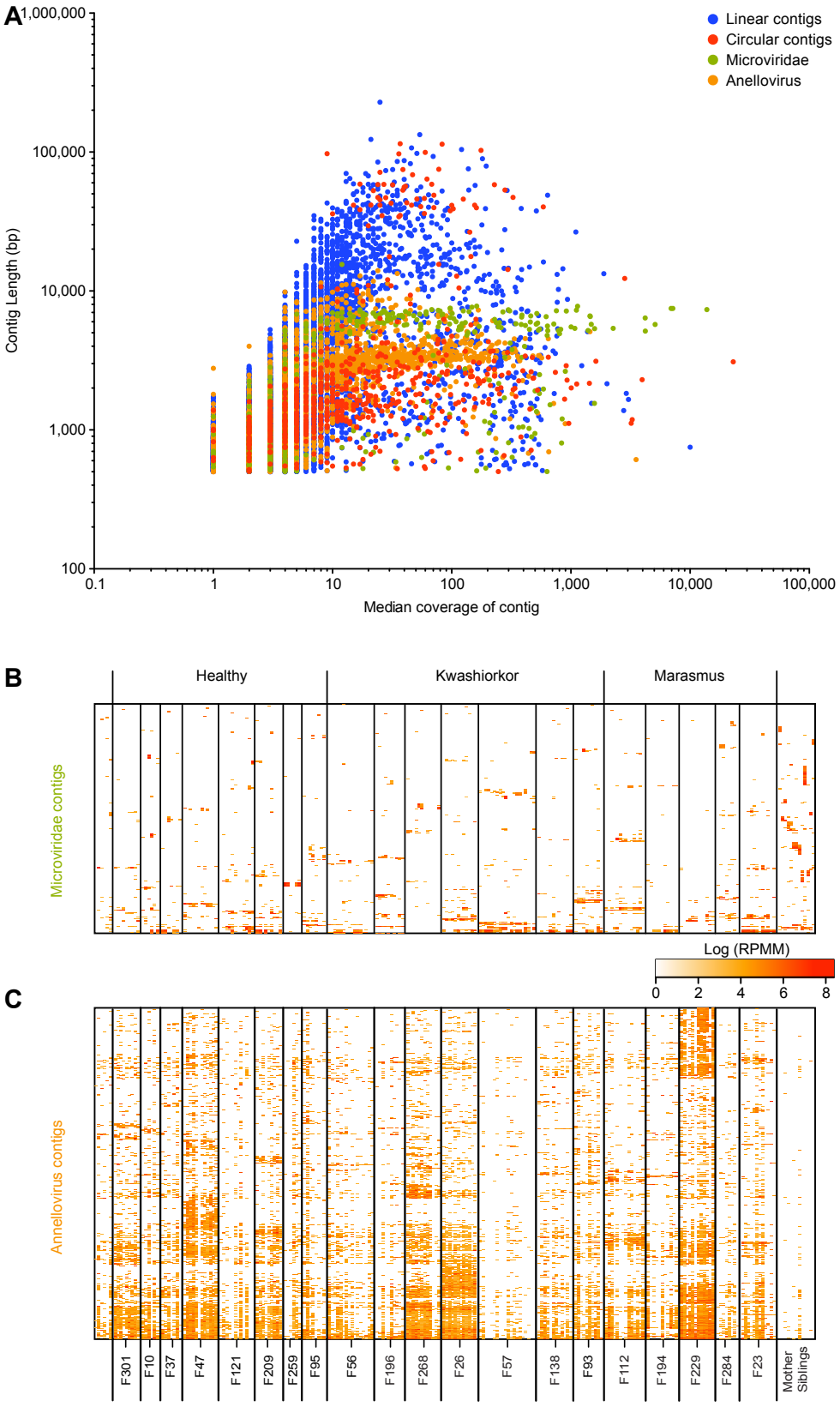

A
Linear contigs
Circular contigs
Microviridae
Anellovirus
Contig Length (bp)
1,000,000
100,000
10,000
1,000
100
0.1
1
10
100
1,000
10,000
100,000
Median coverage of contig
B
Healthy
Kwashiorkor
Marasmus
Microviridae contigs
Log (RPMM)
0
2
4
6
8
C
Annellovirus contigs
F301
F10
F37
F47
F121
F209
F259
F95
F56
F196
F268
F26
F57
F138
F93
F112
F194
F229
F284
F23
Mother
Siblings

Figure 11.

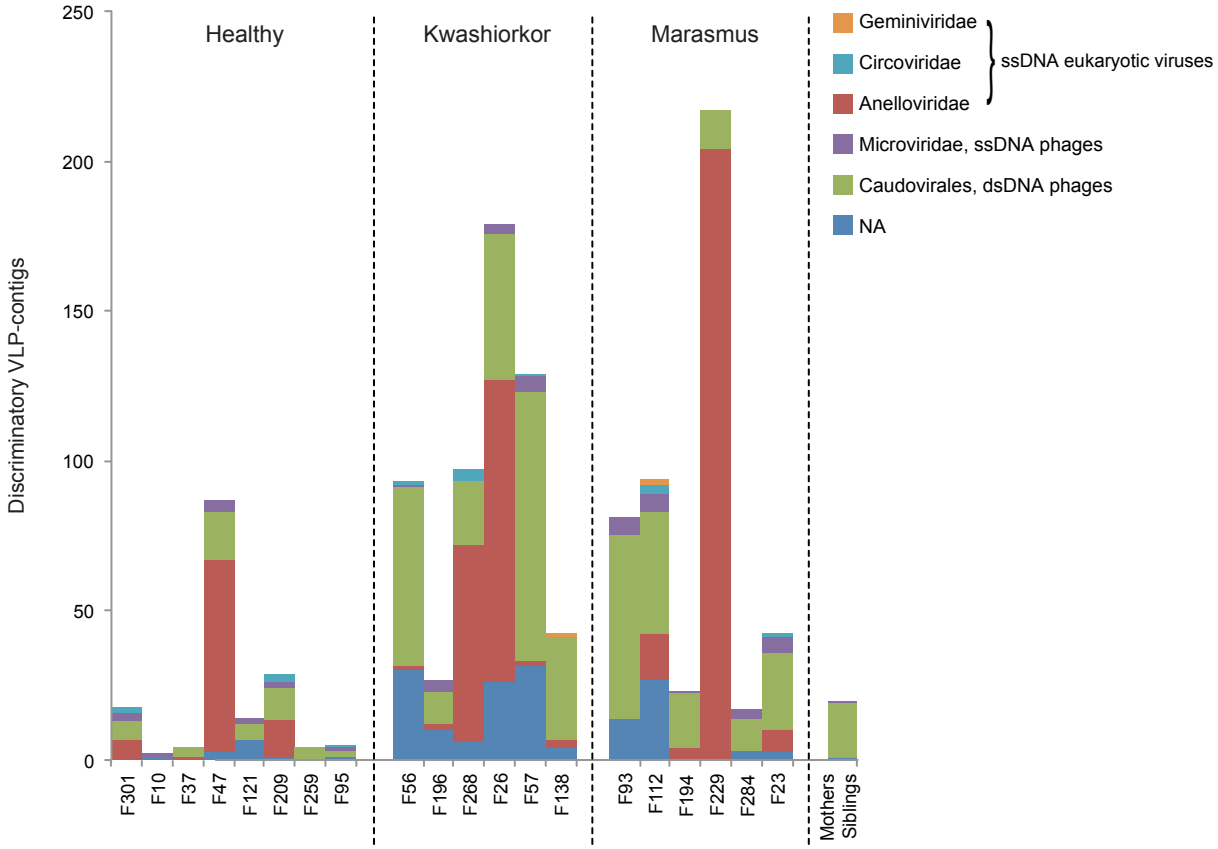

Healthy
Kwashiorkor
Marasmus
Discriminatory VLP-contigs
250
200
150
100
50
0
F301
F10
F37
F47
F121
F209
F259
F95
F56
F196
F268
F26
F57
F138
F93
F112
F194
F229
F284
F23
Mothers
Siblings
Geminiviridae
Circoviridae
Anelloviridae
ssDNA eukaryotic viruses
Microviridae, ssDNA phages
Caudovirales, dsDNA phages
NA

Figure 12.

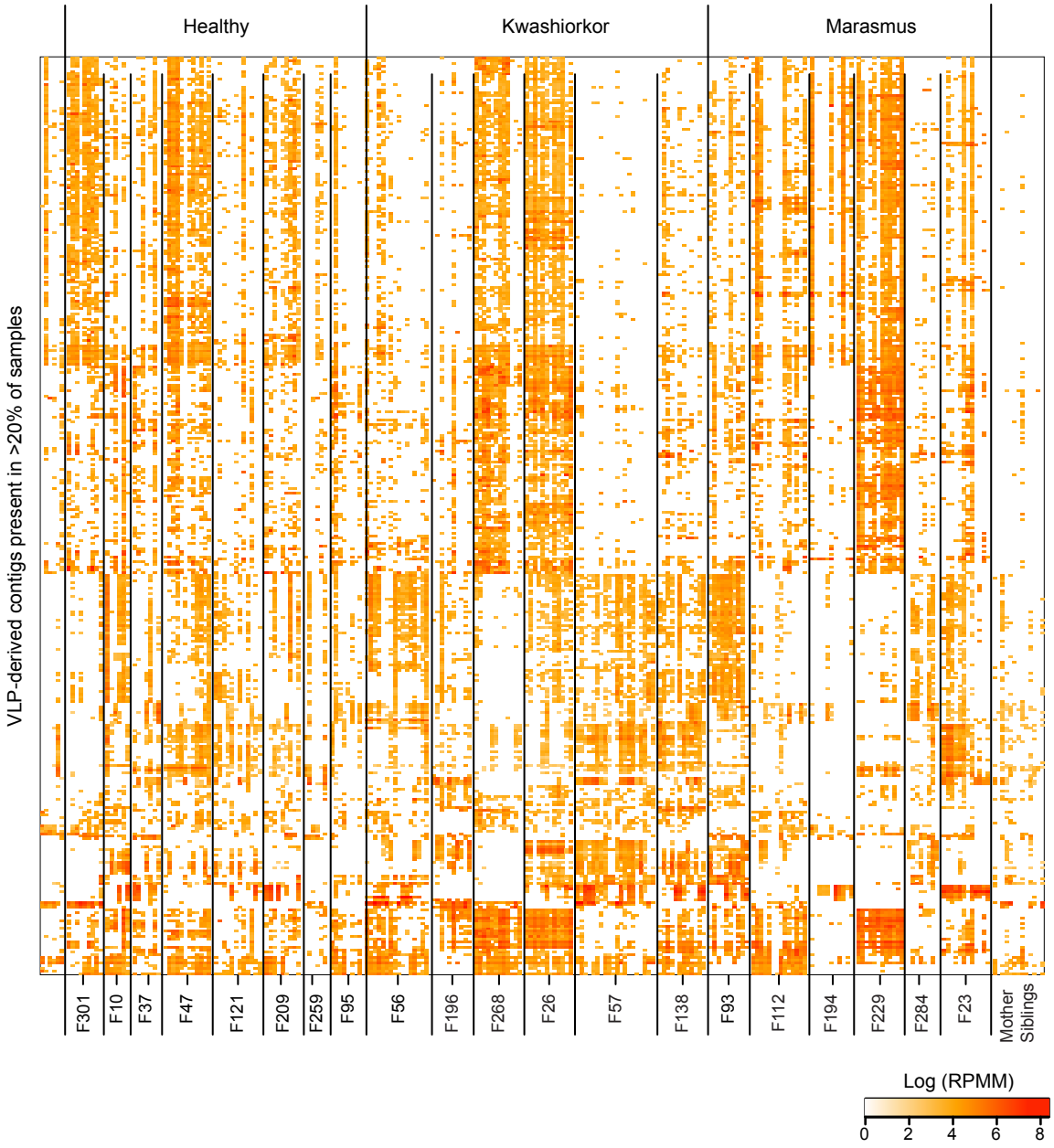

Healthy
Kwashiorkor
Marasmus
VLP-derived contigs present in >20% of samples
F301
F10
F37
F47
F121
F209
F259
F95
F56
F196
F268
F26
F57
F138
F93
F112
F194
F229
F284
F23
Mother
Siblings
Log (RPMM)
0
2
4
6
8

Figure 13.

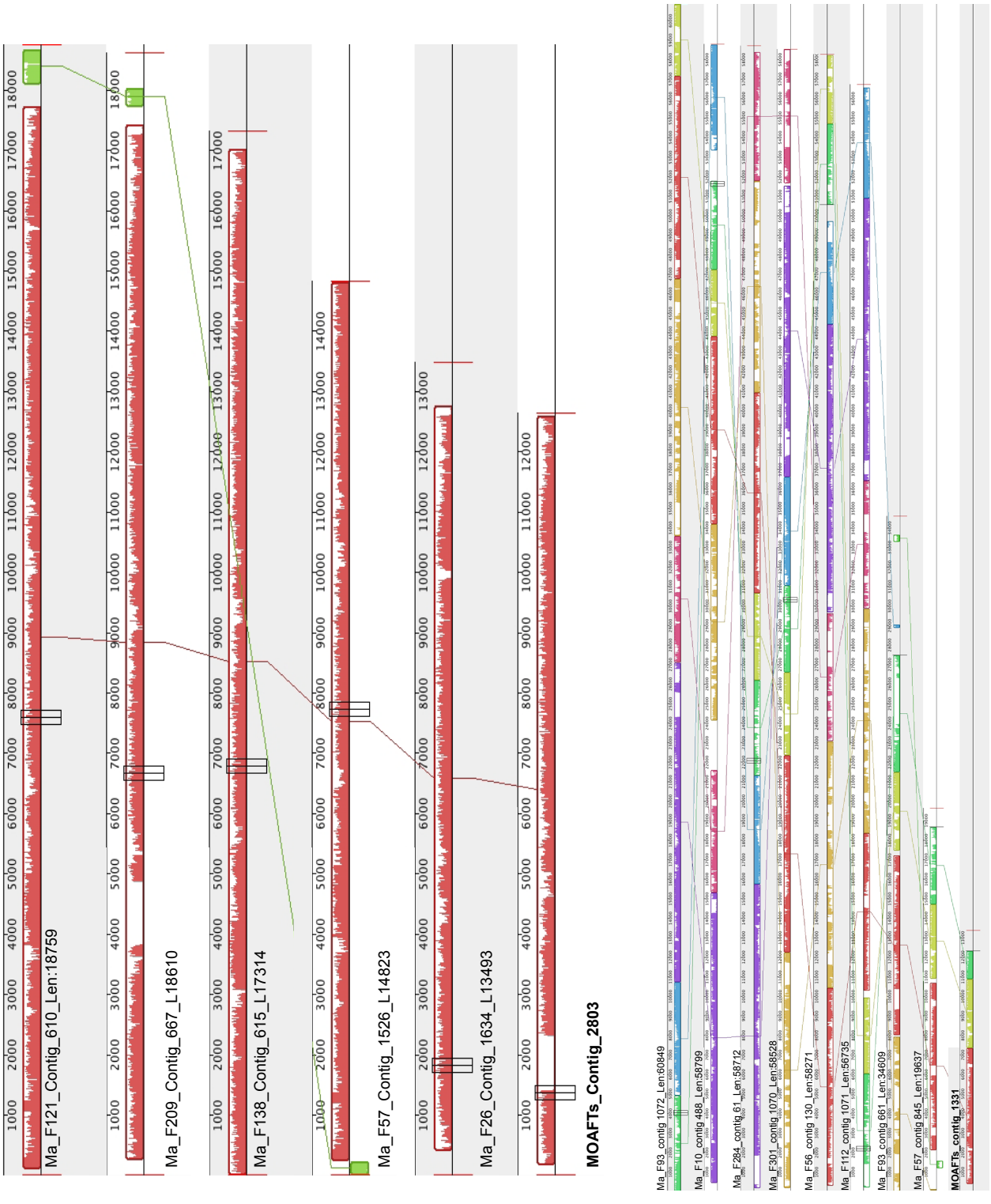
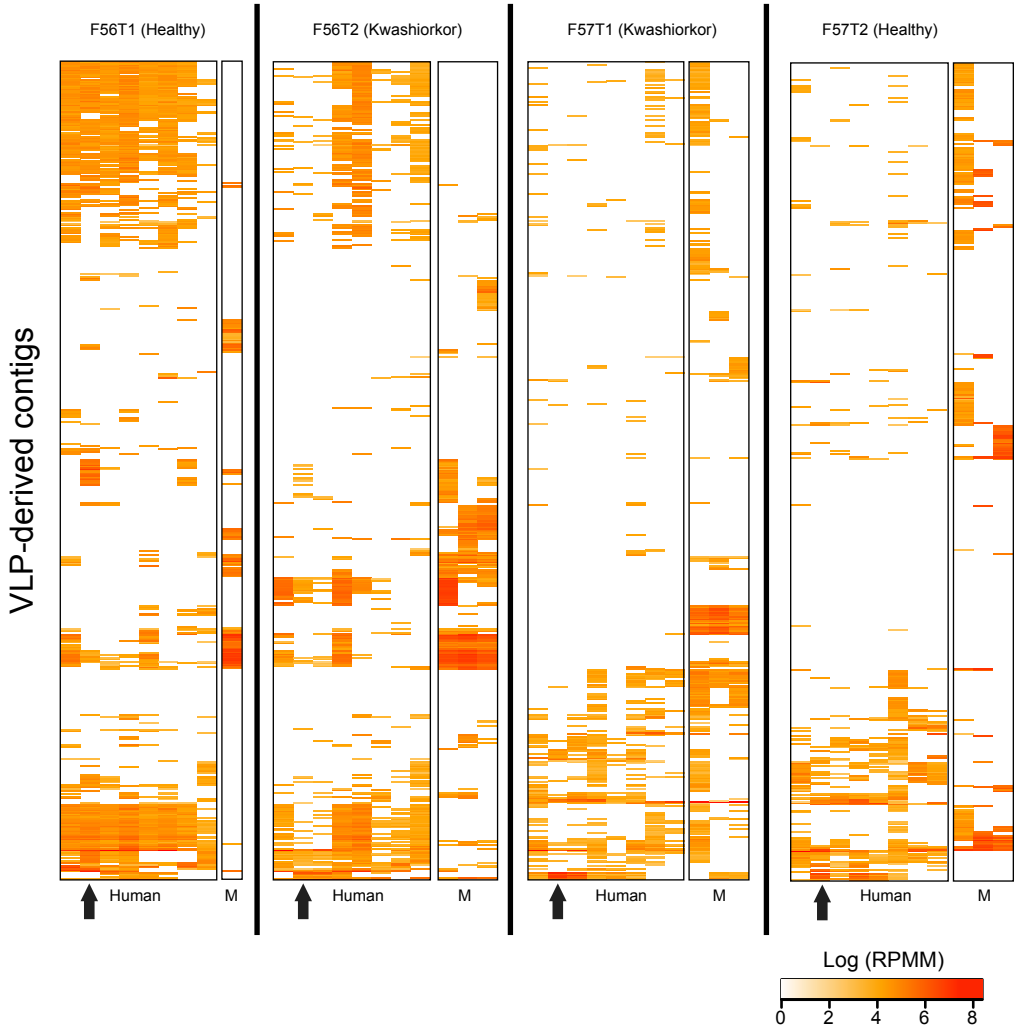

A
Ma_F121_Contig_610_Len:18759
Ma_F209_Contig_667_L18610
Ma_F138_Contig_615_L17314
Ma_F57_Contig_1526_L14823
Ma_F26_Contig_1634_L13493
MOAFTs_Contig_2803
B
Ma_F93_contig 1072_Len:60849
Ma_F10_contig 488_Len:58799
Ma_F284_contig 61_Len:58712
Ma_F301_contig 1070_Len:58528
Ma_F56_contig 130_Len:58271
Ma_F112_contig 1071_Len:56735
Ma_F93_contig 661_Len:34609
Ma_F57_contig 845_Len:19637
MOAFTs_contig_1331

Figure 14.


F56T1 (Healthy)
F56T2 (Kwashiorkor)
F57T1 (Kwashiorkor)
F57T2 (Healthy)
VLP-derived contigs
Human
M
Log (RPMM)
0
2
4
6
8

Supplementary Tables

Table S1. Sample information and pyrosequencing effort of generated viromes.

Table S2. Assembled contigs statistics and preliminary annotation.

References

1 Breitbart, M. *et al*., Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol*. 185, 6220-6223 (2003).

2 Reyes, A. *et al*., Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*. 466, 334-338 (2010).

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

4 Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F., & Gordon, J. I., Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat Rev Microbiol*. 1-11 (2012).

5 Breitbart, M., Marine viruses: truth or dare. *Ann Rev Mar Sci*. 4, 425-448 (2012).

6 Turnbaugh, P. J., Backhed, F., Fulton, L., & Gordon, J. I., Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe*. 3, 213-223 (2008).

7 Wu, G. D. *et al*., Linking long-term dietary patterns with gut microbial enterotypes. *Science*. 334, 105-108 (2011).

8 Faith, J. J., McNulty, N. P., Rey, F. E., & Gordon, J. I., Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science*. 333, 101-104 (2011).

9 Yatsunenko, T. *et al*., Human gut microbiome viewed across age and geography. *Nature*. 486, 222-227 (2012).

10 Smith, M. I. *et al*., Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science*. 339, 548-554 (2013).

11 United Nations (U.N.) Inter-Agency Group for Child Mortality Estimation, Levels & Trends in Child Mortality Report, Available at http://www.childinfo.org/files/Child_Mortality_Report_2011.pdf, (2011).

12 WHO, UN Children's Fund, WHO Child Growth Standards and the Identification of Severe Acute Malnutrition in Infants and Children, Available at http://www.who.int/nutrition/publications/severemalnutrition/9789241598163/en/index.html, (2009).

13 Williams, C. D., Oxon, B. M., & Lond, H., Kwashiorkor: a nutritional disease of children associated with a maize diet. *Lancet*. 226, 1151 (1935).

14 Virgin, H. W., Wherry, E. J., & Ahmed, R., Redefining chronic viral infection. *Cell*. 138, 30-50 (2009).

15 Kim, K. H. & Bae, J. W., Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl Environ Microbiol*. 77, 7663-7668 (2011).

16 Roux, S., Krupovic, M., Poulet, A., Debroas, D., & Enault, F., Evolution and Diversity of the Microviridae Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads. *PLoS ONE*. 7, e40418 (2012).

17 Krupovic, M. & Forterre, P., Microviridae goes temperate: microvirus-related proviruses reside in the genomes of Bacteroidetes. *PLoS ONE*. 6, e19893 (2011).

18 Okamoto, H., History of discoveries and pathogenicity of TT viruses. *Curr Top Microbiol Immunol*. 331, 1-20 (2009).

19 Bernardin, F., Operskalski, E., Busch, M., & Delwart, E., Transfusion transmission of highly prevalent commensal human viruses. *Transfusion*. 50, 2474-2483 (2010).

20 Kapusinszky, B., Minor, P., & Delwart, E., Nearly constant shedding of diverse enteric viruses by two healthy infants. *J Clin Microbiol*. 50, 3427-3434 (2012).

21 Siebrasse, E. A. *et al*., Identification of MW polyomavirus, a novel polyomavirus in human stool. *J Virol*. 86, 10321-10326 (2012).

22 Lim, E. S. *et al*., Discovery of STL polyomavirus, a polyomavirus of ancestral recombinant origin that encodes a unique T antigen by alternative splicing. *Virology*. 436, 295-303 (2013).

23 Goodman, A. L. *et al*., Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc Natl Acad Sci U S A*. 108, 6252-6257 (2011).

24 Handley, S. A. *et al*., Pathogenic simian immunodeficiency virus infection is associated with expansion of the enteric virome. *Cell*. 151, 253-266 (2012).

Appendix

List of Appendices

Appendix 1

Efrem S. Lim, Alejandro Reyes; Martin Antonio, Debasish Saha, Usman N. Ikumapayi, Mitchell Adeyemi, O Colin Stine, Rebecca Skelton, Daniel C. Brennan, Rajhab S. Mkakosya, Mark J. Manary, Jeffrey I. Gordon, David Wang

Discovery of STL polyomavirus, a polyomavirus of ancestral recombinant origin that encodes a unique middle T antigen by alternative splicing.

Virology, 2013 Feb 20; 436 (2):295-303.

http://dx.doi.org/10.1016/j.virol.2012.12.005

Appendix 2

Erica A. Siebrasse, Alejandro Reyes, Efrem S. Lim, Guoyan Zhao, Rajhab S. Mkakosya, Mark J. Manary, Jeffrey I. Gordon, David Wang

Identification of MW Polyomavirus, a Novel Polyomavirus in Human Stool.

J Virol. 2012 Oct; 86(19):10321-6.

http://jvi.asm.org/content/86/19/10321

Appendix 3

Andrew L. Goodman, George Kallstrom, Jeremiah J. Faith, Alejandro Reyes, Aimee Moore, Gautam Dantas, Jeffrey I. Gordon

Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice.

PNAS 2011, Apr 12;108(15):6252-7.

http://www.pnas.org/content/108/15/6252

Appendix 4

Kevin J. Forsberg[†], Alejandro Reyes [†], Bin Wang, Elizabeth M. Selleck, Morten O.A. Sommer, Gautam Dantas.

The Shared Antibiotic Resistome of Soil Bacteria and Human Pathogens.

Science, 2012; 337 (6098): 1107-1111.

http://www.sciencemag.org/content/337/6098/1107

Virology 436 (2013) 295–303

Contents lists available at SciVerse ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Discovery of STL polyomavirus, a polyomavirus of ancestral recombinant origin that encodes a unique T antigen by alternative splicing

Efrem S. Lim [a], Alejandro Reyes [b], Martin Antonio [e], Debasish Saha [e], Usman N. Ikumapayi [e], Mitchell Adeyemi [e], O. Colin Stine [f], Rebecca Skelton [c], Daniel C. Brennan [c], Rajhab S. Mkakosya [g], Mark J. Manary [d,h], Jeffrey I. Gordon [b], David Wang [a,*]

[a] Departments of Molecular Microbiology and Pathology & Immunology, Washington University School of Medicine, 660 S. Euclid Avenue, St. Louis, MO, USA
[b] Center for Genome Sciences and Systems Biology, Washington University School of Medicine, 660 S. Euclid Avenue, St. Louis, MO, USA
[c] Department of Internal Medicine, Washington University School of Medicine, 660 S. Euclid Avenue, St. Louis, MO, USA
[d] Department of Pediatrics, Washington University School of Medicine, 660 S. Euclid Avenue, St. Louis, MO, USA
[e] Medical Research Council Unit, PO Box 273, Banjul, The Gambia
[f] Department of Epidemiology and Public Health, University of Maryland School of Medicine, 660 W Redwood St., Baltimore, MD, USA
[g] Department of Pathology, University of Malawi College of Medicine, Private Bag 360, Chichiri, Blantyre 3, Malawi
[h] Department of Community Health, University of Malawi College of Medicine, Private Bag 360, Chichiri, Blantyre 3, Malawi

ARTICLE INFO

Article history:
Received 15 August 2012
Returned to author for revisions
8 October 2012
Accepted 5 December 2012
Available online 29 December 2012

Keywords:
Polyomavirus
Virus discovery
Tumor antigen
Alternative splicing
Recombination

ABSTRACT

The family *Polyomaviridae* is comprised of circular double-stranded DNA viruses, several of which are associated with diseases, including cancer, in immunocompromised patients. Here we describe a novel polyomavirus recovered from the fecal microbiota of a child in Malawi, provisionally named STL polyomavirus (STLPyV). We detected STLPyV in clinical stool specimens from USA and The Gambia at up to 1% frequency. Complete genome comparisons of two STLPyV strains demonstrated 5.2% nucleotide divergence. Alternative splicing of the STLPyV early region yielded a unique form of T antigen, which we named 229T, in addition to the expected large and small T antigens. STLPyV has a mosaic genome and shares an ancestral recombinant origin with MWPyV. The discovery of STLPyV highlights a novel alternative splicing strategy and advances our understanding of the complex evolutionary history of polyomaviruses.

Introduction

Polyomaviruses are small circular double stranded DNA viruses. Members of the family *Polyomaviridae* have been isolated from a variety of human specimen types, as well as from other hosts, including primates, rodents and birds. The genomes of polyomaviruses range in size from 4754 to 5387 bp and can be divided by transcriptional criteria into an early region, a late region and a non-coding control region (Van Ghelue et al., 2012). The early region specifies the large tumor antigen (LTAg) and small tumor antigen (STAg). Rodent polyomaviruses encode an additional middle tumor antigen (MTAg) (Gottlieb and Villarreal, 2001). The late region encodes the structural proteins VP1, VP2 and VP3. Additionally, avian polyomaviruses harbor a unique VP4 upstream of VP1, VP2 and VP3, that is absent from mammalian polyomaviruses (Johne and Müller, 2007). The VP4 of avian polyomaviruses differs from the similarly named SV40 VP4, which is an opening reading frame within the VP2 transcript (Daniels et al., 2007). Finally, a limited subset of polyomaviruses encode an agnoprotein 5′ of the late region (Van Ghelue et al., 2012).

Alternative splicing is a critical mechanism for regulating expression of different gene products from the early region of polyomaviruses (Huang and Carmichael, 2009). Three major forms of the T antigen (LTAg, MTAg and STAg) have been described, as have multiple alternative forms of the LTAg. While most polyomaviruses express STAg from an unspliced mRNA transcript, STAg from rodent polyomaviruses is encoded from an alternatively spliced transcript. Alternative splicing also results in LTAg and STAg sharing approximately 80 amino acid (aa) residues at their N-terminus. Additionally, SV40 encodes a 17 kT protein that shares the first 131 aa residues with LTAg, followed by 4 aa residues due to differential splicing (Zerrahn et al., 1993). JCPyV encodes 3 additional proteins (T′135, T′136 and T′165), which also similarly share 132 N-terminal aa residues with LTAg, but differ at the C-terminus due to alternative splicing patterns (Trowbridge and Frisque, 1995). Similarly, MCPyV encodes a 57 kT antigen by

* Correspondence to: Departments of Molecular Microbiology and Pathology & Immunology, Washington University School of Medicine, Campus Box 8230, 660S. Euclid Ave., St. Louis, MO 63110, USA. Fax: +1 314 362 1232.
E-mail address: davewang@borcim.wustl.edu (D. Wang).

0042-6822/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.virol.2012.12.005

alternative splicing of the LTAg transcript (Shuda et al., 2008). Moreover, a truncated form of LTAg expressed by BKPyV results from alternative splicing of its LTAg transcript (Abend et al., 2009). Furthermore, alternative splicing can also result in the translation of a middle tumor antigen (MTAg), although this is thought to be a feature unique to rodent polyomaviruses (Huang and Carmichael, 2009). While the 17 kT, 57kT, T′- and truncated T antigens share their splice donor site with their cognate LTAg transcripts, the splice donor site of MTAg is shared by STAg instead. Importantly, many of these proteins expressed from alternatively spliced early products have been shown to have transforming potential (Bollag et al., 2000; Boyapati et al., 2003).

Eight new polyomaviruses have been discovered in human clinical specimens within the last five years (Allander et al., 2007; Feng et al., 2008; Gaynor et al., 2007; Schowalter et al., 2010; Scuda et al., 2011; Siebrasse et al., 2012; van der Meijden et al., 2010), leading to new insights in the fundamental biology and pathogenesis of polyomavirus. These discoveries, along with the discoveries of additional animal polyomaviruses have led to a recent taxonomic proposal for classification of polyomaviruses (Johne et al., 2011). According to this proposal, polyomaviruses are broadly classified into three genera, *Avipolyomavirus*, *Wukipolyomavirus*, and *Orthpolyomavirus*, primarily based on phylogenetic analyses on full-length polyomavirus genomes (Johne et al., 2011). Polyomaviruses have co-evolved with their hosts over long evolutionary timescale (Pérez-Losada et al., 2006; Sharp and Simmonds, 2011). While several intra-strain recombinations have been observed (Chen et al., 2004; Hatwell and Sharp, 2000), large scale recombination of polyomaviruses is generally thought to be rare (Bhattacharjee, 2010; Crandall et al., 2006). However, emerging studies highlight several lineages with inconsistency amongst topologies constructed from different regions of the genome, indicative of recombination events (Sauvage et al., 2011; Schowalter et al., 2010; Siebrasse et al., 2012).

Here, we describe a novel polyomavirus provisionally named STL polyomavirus (STLPyV). We sequenced the complete genomes of STLPyV strains from Malawi and Saint Louis and found that they differed by 5.2% at nucleotide level. We show that the early region of STLPyV, in addition to the unspliced STAg, undergoes alternative splicing that would encode for LTAg and a unique 229 amino acid T antigen (229T) unlike that of previously characterized rodent MTAg or various modified forms of LTAg, such as 17 kT, T′ or truncated T antigens. STLPyV is most closely related to MW polyomavirus (MWPyV) and we establish that their common ancestor had a recombinant origin. We also developed a sensitive degenerate PCR assay that detects both STLPyV and MWPyV and used this assay to screen for STLPyV and MWPyV in clinical specimens from Saint Louis, USA and Gambia. Both viruses were detected in pediatric stool specimens, but there was no statistically significant evidence of association between either virus with diarrheal cases. These results contribute to a deeper understanding of polyomavirus diversity, evolution and gene content.

Results

Discovery of the novel STL polyomavirus

As part of a broader effort to characterize the human gut virome in health and disease, we performed shotgun 454 pyrosequencing of DNA amplified by multiple displacement amplification (MDA) from the stool of a healthy 15 month child in Malawi. Prior to DNA extraction, the stool was processed by CsCl ultracentrifugation to generically enrich for viral particles as described previously (Reyes et al., 2010). Two reads from this sample shared limited sequence identity to known polyomaviruses (Fig. 1A). At the time the sequences were generated, the first read (308 nt) shared only 46% amino acid identity to the LTAg of Squirrel monkey polyomavirus, its top scoring hit after tBlastx search to the Genbank nt database. The second read (87 nt) mapped to a separate region of the LTAg in California sea lion polyomavirus based (81% amino acid identity). We recently reported the discovery of another novel polyomavirus, MWPyV (Siebrasse et al., 2012); sequence analyses indicated that these two reads were most closely related to, but clearly distinct from, MWPyV. From these two initial reads, we designed primers to PCR amplify products that span the two initial pyrosequencer reads in either direction. The resulting complete circular genome of 4776 bp was sequenced to more than 3X coverage (Fig. 1A). Whole genome comparison to all known polyomaviruses indicated that this virus shared the highest nucleotide sequence similarity (64.2%) to MWPyV, which is less than the 81% criterion outlined in species classification guidelines from the International Committee on Taxonomy of Viruses (ICTV) (Johne et al., 2011). We gave this new species of polyomavirus the provision name of STLPyV.

We also amplified, cloned, and sequenced to greater than 3X coverage a second full-length STLPyV strain (WD972). This strain was amplified, using a similar PCR strategy as the index strain, from a fecal specimen obtained from a child who had been enrolled a study of diarrheal diseases conducted at Saint Louis Children's hospital (described in detail below). The complete genome of STLPyV strain WD972 was 4775 bp, which included a one nucleotide deletion in the non-coding control region compared to the index Malawi strain. Whole genome sequence analyses indicated that the overall nucleotide sequence identity between the two STLPyV strains was 94.8% with divergence of up to 13% within regions of VP1 and LTAg (Fig. 1B).

Genome annotation and alternative splicing of a unique T antigen mRNA transcript

The STLPyV genome organization and sizes of its predicted open reading frames were characteristic of known polyomaviruses (Fig. 1A and Table 1). Further, the non-coding control region of STLPyV (nucleotide positions 1–352) contained features typical of polyomaviruses (Van Ghelue et al., 2012). The *ori* region encoded five consensus T antigen binding pentanucleotide sequence G(A/G)GGC (nucleotide positions 29, 35, 44, 51 and 249) and one non-canonical GTGGC pentanucleotide sequence (nucleotide position 147) (Cantalupo et al., 2005). In addition, there was a 17 nucleotide stretch of AT-rich region including the putative TATA box (nucleotide positions 59–75).

The early proteins encoded by polyomaviruses are translated from alternatively spliced transcripts. Typically, gene predictions for LTAg are made based upon the presence of conserved splice donor and splice acceptor sites defined by alignment of other polyomaviruses. However, STLPyV lacked the consensus splice donor sites commonly utilized by most polyomaviruses. Moreover, the rarer splice donor sites identified were incongruent with the predicted LTAg open reading frame due to an out-of-frame arrangement when paired with the corresponding splice acceptor sites. Therefore, we experimentally determined the splice sites used by STLPyV early mRNA transcripts. To do this, we cloned the genomic region encoding the putative LTAg of STLPyV (nucleotide positions 4776–2452) into an expression vector, transiently transfected the plasmids into 293T cells, and performed RT-PCR on total cell RNA extracts. Primers were designed to span a 666 bp fragment of the early region judged most likely to harbor the splice donor and acceptor sites. As expected, PCR using a plasmid template containing STLPyV genomic DNA yielded the predicted band, while RNA extracted from mock-transfected cells did not amplify any products (Fig. 2A). Surprisingly, we detected three bands following RT-PCR of STLPyV transfected extracts (Fig. 2A,

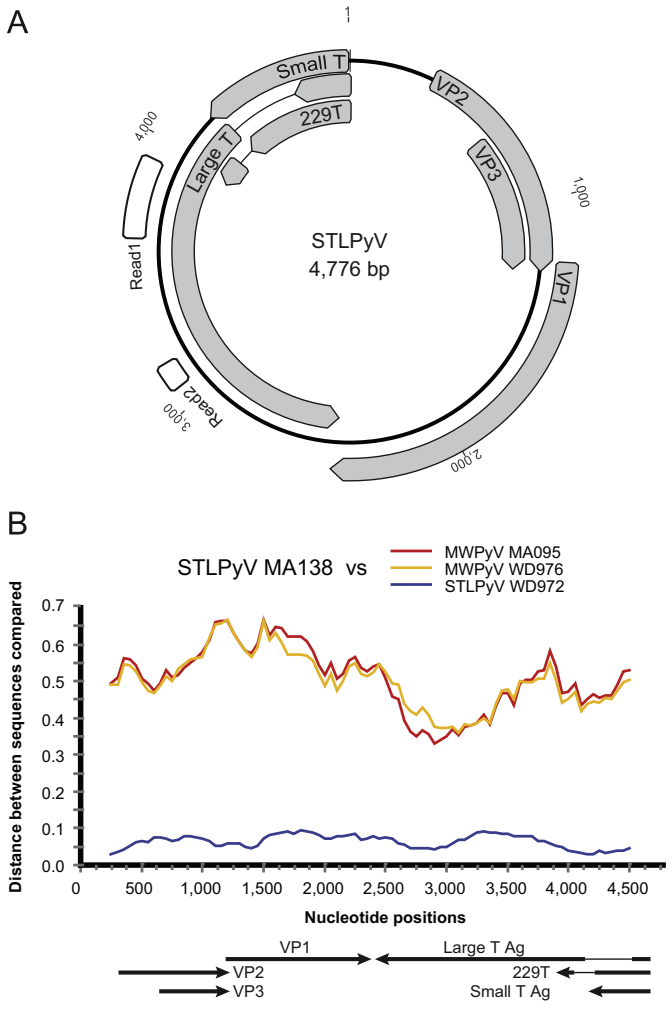


Fig. 1. *STLPyV is a novel polyomavirus.* (A) Diagram of STLPyV genome based on strain MA138. Positions of the two reads obtained from the initial 454 pyrosequencing of DNA generated by MDA of purified virus particles from the index Malawian case are indicated in white. Predicted open reading frames of STLPyV are shown in gray. (B) Diversity plots of nucleotide sequences are shown between STLPyV MA138 and STLPyV WD972 (blue). Their closest relatives MWPyV MA095 (red) and MWPyV WD976 (orange) are included as reference. The proportion of DNA sequence difference is indicated (0.1 = 10%).

Table 1
Putative proteins encoded by STLPyV (strain MA138).

Protein	Putative coding region(s)	Predicted size (aa)	Calculated mass (kDa)	Range (aa) in other polyomaviruses
STAg	4776–4189	195	23.0	124–199
229T	4776–4206, 4090–3972	229	27.2	N/A
LTAg	4776–4534, 4188–2452	660	75.9	599–817
VP1	1242–2447	401	43.4	343–497
VP2	353–1264	303	33.2	241–415
VP3	677–1264	195	21.9	190–272

STLPyV). Each of the three bands was cloned and sequenced and subsequently determined to be a unique product. The largest band represented the unspliced mRNA transcript that yielded a 195 aa open reading frame expected of STAg. The smallest band represented an mRNA transcript generated by excision of a 345 bp intron, yielding an open reading frame with features consistent with those expected for a LTAg (660 aa). The LTAg of STLPyV contained a putative pRb-binding motif (LSCNE beginning at aa residues 105), an N-terminal DnaJ domain (HPDKGG commencing at aa residue 42), and a predicted Bub1 binding site (WDQWW beginning at aa residue 90)—features that are highly conserved in polyomaviruses (Van Ghelue et al., 2012).

The intermediate band corresponded to an mRNA transcript derived by splicing of a 115 bp intron that yields a unique open reading frame of 229 aa; the first 190 aa were shared with the STAg, the 191st aa encompassed the splice junction and the remaining 38 aa were derived from the second exon. Based on its predicted amino acid length, we named this putative protein 229T. The splice sites that generated 229T were unique from the splice sites utilized by the STLPyV LTAg. Although there is precedence in rodent polyomaviruses for splicing of STAg, in those instances the splice acceptor site is identical to the LTAg splice acceptor, which was not the case for STLPyV. 229T also differed from the 17 kT and T′ antigens of SV40 or BKPyV which share common splice sites with their LTAg

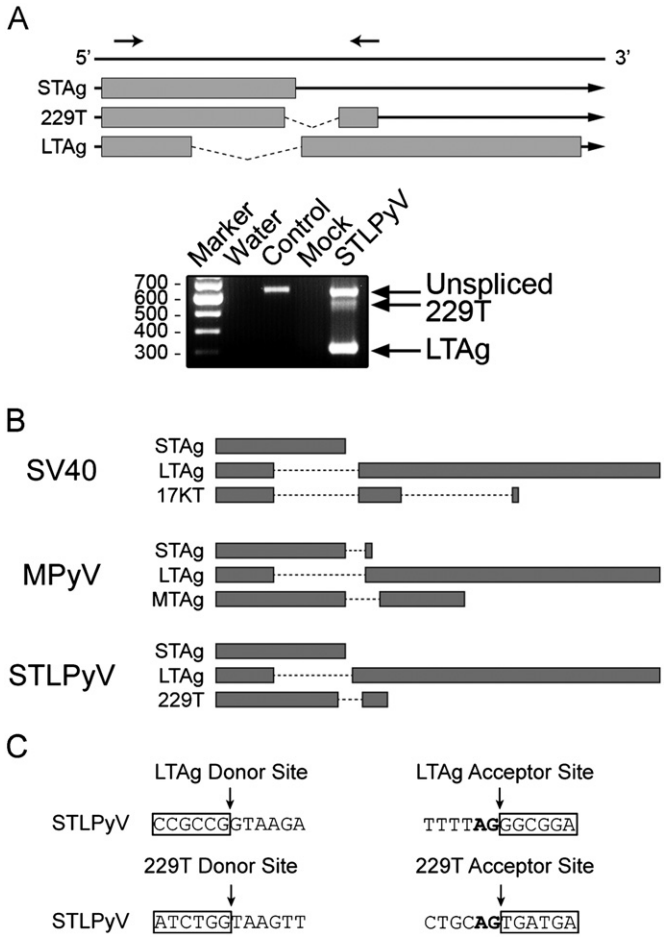


Fig. 2. *Alternative splicing of STLPyV early region.* (A) Schematic shows mRNA transcripts expressed from the early region of STLPyV. Exonic regions are indicated by gray boxes, intronic regions by lines. Arrows above the diagram indicate the position of primers. 293T cells were transfected with a plasmid expressing STLPyV LTAg for 48 h. Cells were harvested and lysed for RT-PCR analysis. RT-PCR reactions were performed with water (Water), STLPyV LTAg plasmid (Control), RNA harvested from mock-transfected (Mock), STLPyV LTAg transfected 293T cells (STLPyV). Data are representative of at least three independent experiments. The three alternatively splice mRNA transcripts were confirmed with a different set of primers (data not shown) and verified in HeLa cells (data not shown). (B) Diagram showing the coding region of T antigens, separated by intron regions in dashed lines. The T′- and truncated *T* antigens of JCPyV and BKPyV are spliced from LTAg transcripts in a similar manner as SV40 17KT. (C) Sequences of the splice donor and acceptor sites for STLPyV LTAg and 229T transcripts (strain MA138). Exon sequences are highlighted in boxes. The nucleotide sequences of the splice donor and acceptor sites of STLPyV shown are conserved in the three STLPyV strains characterized.

transcripts (Fig. 2B). To verify the alternatively spliced transcripts, we designed a second set of primers for RT-PCR. Cloning and sequencing independently confirmed the splice junctions, demonstrating that the results were not due to an artifact of the initial primers (data not shown). The splicing patterns were also confirmed with both primer pairs in HeLa cells indicating that the splice variants were not cell line specific (data not shown).

We next sequenced a fragment of the STLPyV genome in the region spanning the splice sites (nt 4731–4066) from an additional sample from another child in the Saint Louis study found to be positive for the presence of STLPyV (using a PCR assay described below) so that we could perform a sequence comparison with the two STLPyV strain. While we observed up to a 3.2% difference in the nucleotide sequence across the 666 bp region between the three strains, both 5′ splice and 3′ sites for LTAg and 229 T were entirely conserved. Thus, we conclude that the early region of STLPyV has unique alternative splicing features.

Phylogenetic analysis indicates ancient recombinant origin of STL polyomavirus

To estimate the phylogenetic relationship of the new STLPyV to other polyomaviruses, we used the LTAg, VP1 and VP2 sequences to construct phylogenies using both Bayesian (BI) and maximum likelihood (ML) methods. Both methods yielded trees with similar topologies. Midpoint rooting of the large T antigen phylogeny, which was the most alignable region across all polyomaviruses, positioned the avian polyomaviruses (*Avipolyomavirus* genera) basal to mammalian polyomaviruses (Fig. 3A). This is consistent with other studies (Pérez-Losada et al., 2006; Sauvage et al., 2011). Furthermore, the outgroup positioning of the *Avipolyomavirus* is consistent with marked differences in the genomic organization, pathogenesis and biology of the avian polyomaviruses (Johne et al., 2011). Therefore, the avian polyomaviruses were used as an outgroup to root the phylogenies.

The phylogenies of LTAg, VP1 and VP2 showed that both STLPyV strains are closely related to MWPyV. STLPyV and MWPyV strains formed a monophyletic clade with high confidence (Fig. 3). However, the topologies of the LTAg and VP1 derived trees were different. Within the LTAg region, STLPyV and MWPyV were closely related to HPyV6 and HPy7 and formed a clade with several *Orthopolyomavirus* species including TSPyV, MCPyV, MPyV and HPyV9 (Fig. 3A). Strikingly, in the VP1 region, STLPyV and MWPyV were most closely related to HPyV6, HPyV7, WUPyV and KIPyV, but clustered with different *Orthopolyomavirus* species (SV40, BKPyV, JCPyV and SA12) (Fig. 3B). Finally, the VP2-derived tree indicated that STLPyV and MWPyV cluster with the similar *Orthopolyomavirus* species as in the LTAg derived phylogeny, whereas HPyV6 and HPyV7 are closely related to WUPyV and KIPyV. The VP2 region of polyomaviruses has many insertions and deletions; therefore, it is important to be cautious about over-interpreting the different branching orders. Nonetheless, the discordant phylogenetic relationship of the STLPyV and MWPyV clade in LTAg and VP1 derived trees are indicative of recombinant events in their common ancestor. HPyV6 and HPyV7 are also likely recombinant lineages, although the phylogenies suggest that the evolutionary histories of HPyV6 and HPyV7 as compared to that of STLPyV and MWPyV are likely distinct because the relationships of all four viruses cannot be adequately explained by the same recombination events. Thus, our results indicate that STLPyV and MWPyV share an ancestral recombinant origin.

Molecular characterization demonstrates STLPyV presence in Saint Louis and Gambia

In order to determine the prevalence of STLPyV, we developed a PCR assay targeted against the large T antigen region. Since STLPyV is closely-related to the newly-identified MWPyV (Fig. 3A), we designed an assay capable of detecting both STLPyV and MWPyV, thus allowing us to characterize the prevalence of both novel polyomaviruses as well as to identify additional variants of these two viruses. We confirmed that the assay amplified a specific 481 bp (STLPyV) or 484 bp (MWPyV) PCR product using plasmids encoding the large *T* antigen of STLPyV or MWPyV (Fig. 4A). To validate our PCR assay, we examined pediatric fecal specimens sent for bacterial culture to the clinical microbiology laboratory at the Saint Louis Children's Hospital. These samples were collected from patients, primarily with diarrhea, who were examined between July 2009 and June 2010. This cohort was previously screened for the presence of MWPyV by a real-time PCR assay and 12 specimens were found to be positive (Siebrasse et al., 2012). In addition to the 12 samples, the new PCR assay described in this report identified two additional samples that were positive for MWPyV, for a total of 14 samples (Fig. 4B, right column). In the previous study, these two samples yielded detectable Ct values in the real time assay, but were below the cutoff value used to define positive samples (data not shown).

Seven samples from the Saint Louis children's study were found to be positive for STLPyV (Fig. 4B, left column) demonstrating that STLPyV can also be found in a geographically separate region from the initial strain identified in Malawi. Interestingly, three samples contained both STLPyV and MWPyV (samples WD972, WD976 and WD1226). These three samples were collected from the same individual, a five year-old lung transplant recipient who presented with persistent, recurring diarrhea. The first two samples were taken on consecutive days while the third sample was collected four months later. The other four STLPyV samples came from four different individuals.

Previous reports indicate that the prevalence of polyomaviruses is elevated in immunocompromised patients. Hence, we sought to understand the prevalence of STLPyV and MWPyV in adult transplant recipients by screening specimens collected from kidney transplant patients at Washington University in Saint Louis. Different types of biospecimens (fecal, urine, nasopharyngeal swab and sera) from individuals were examined with our PCR assay. We did not detect STLPyV in fecal samples ($n=237$), plasma ($n=261$) or nasopharyngeal swabs ($n=261$) (Fig. 4C). We found that one urine sample was positive for STLPyV. Interestingly, we did not detect MWPyV in any of the specimens. The absence of STLPyV and MWPyV in fecal samples from the adult study was unexpected since the prevalence of STLPyV and MWPyV in the children we surveyed from St. Louis was about 1% and 2.2% respectively (Fig. 4B). To verify the integrity of the specimens, we screened urine samples for the presence of JCPyV using a published real time PCR assay (Siebrasse et al., 2012). Forty-five samples were positive for JCPyV (data not shown); the 12% prevalence of JCPyV in this study is consistent with estimates from other reports suggesting that the low/lack of STLPyV and MWPyV was not due to compromised specimen conditions. Thus, the prevalence of STLPyV and MWPyV in stool samples from our study of adult kidney recipients is lower than that observed in the children's study.

Since STLPyV and MWPyV were readily detected in the stool samples from children, we examined whether STLPyV and MWPyV are associated with childhood diarrhea by analyzing fecal samples collected from Gambia as part of the ongoing Global Enteric Multicenter Study (GEMS) (Kotloff et al., 2012). For each enrolled case with diarrhea, one healthy control without diarrhea (matched for age, gender and time of presentation) was randomly selected from the community. We screened 332 cases and 389 controls for the presence of STLPyV and MWPy, finding one sample from the controls and none from the cases to be positive for STLPyV, and five samples from cases and five from controls to be positive for MWPyV. Based on these results, we concluded that neither STLPyV nor MWPyV had a statistically significant association with diarrheal cases (Fig. 4D, right column).

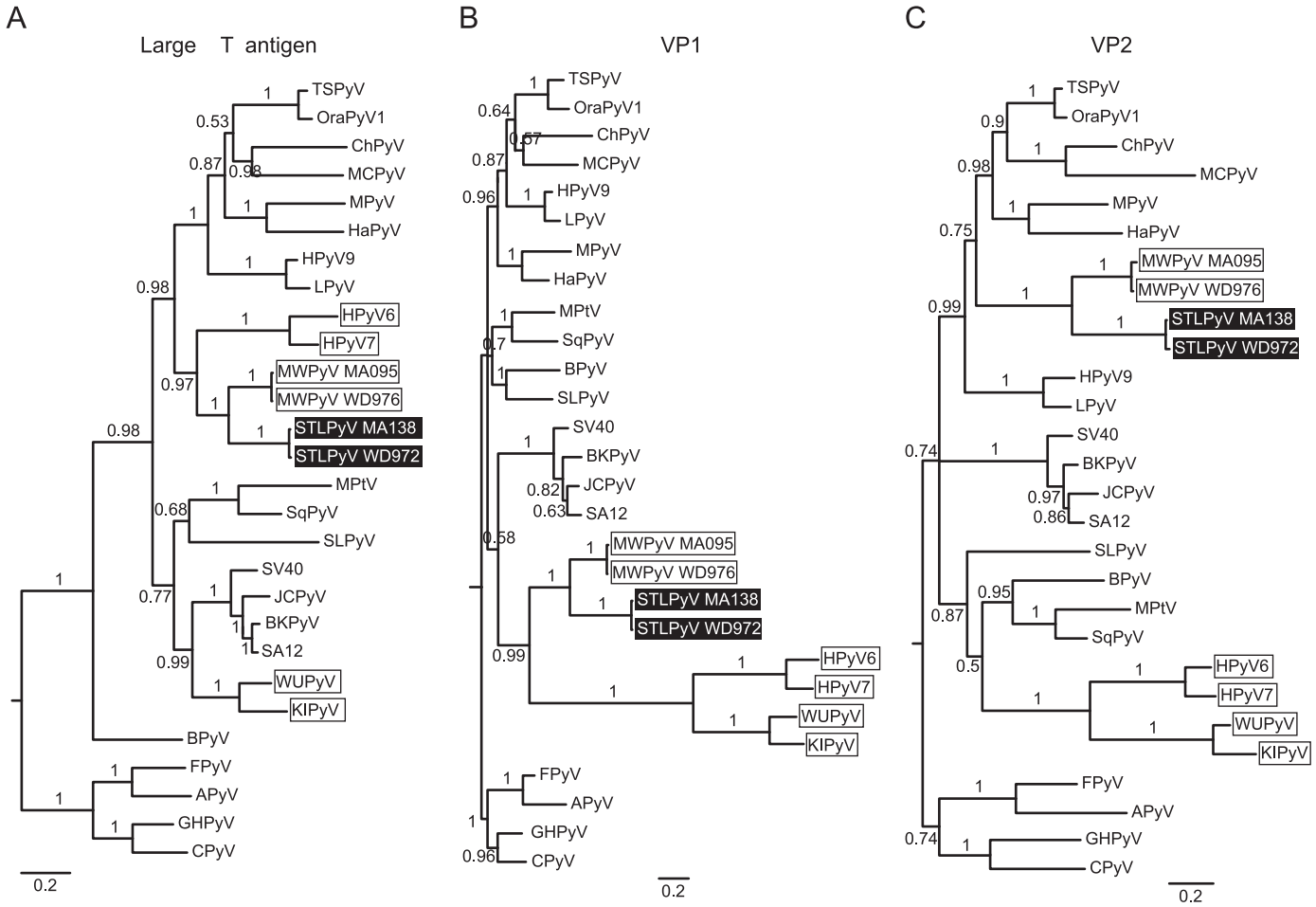


Fig. 3. *Phylogenetic analysis of STLPyV*. Phylogenetic relationships of 28 diverse polyomavirus sequences were inferred from alignment of protein sequences from LTAg (A), VP1 (B) and VP2 (C). Avian polyomaviruses were used as an outgroup to root the phylogenies. STLPyV strains are highlighted in black; MWPyV strains and *Wukipolyomavirus* whose members show discordant phylogenetic relationships are indicated by boxes. Internal branch labels indicate Bayesian posterior probabilities. The ML method yielded trees with similar topologies.

We next examined the diversity of STLPyV and MWPyV strains based on the amplicon sequences generated from the positive samples from the GEMS and St. Louis studies described above. Phylogenetic analyses indicated that STLPyV and MWPyV sequences formed distinct clades with high confidence (Fig. 5). The STLPyV strains formed two monophyletic clades that differed by at least 5% nucleotide sequence identity between clades, hence we tentatively designated them as two genotypes. MWPyV strains formed three monophyletic clades with high confidence, with at least 5% difference in inter-clade nucleotide sequence identity, which we designated as three genotypes. A previous study of MWPyV identified members of two of these clades (Siebrasse et al., 2012); in the present study, additional members of those two clades were identified as well as the existence of a third clade. By comparison, the global variation of WUPyV strains is approximately 1.2%, while the BKPyV strains vary by up to 5.3% at the nucleotide sequence (Bialasiewicz et al., 2010; Krumbholz et al., 2009). Thus, this indicates that there is diverse strain variation in STLPyV and MWPyV.

Discussion

Novel T antigen alternative splicing

Alternative splicing of the early mRNA transcripts of a few canonical polyomaviruses is well documented. In recent years, as application of molecular methods dramatically increased the rate of discovery of new polyomaviruses, many of the genome annotations have been based on strictly computational analyses and predictions; there have been few studies that have experimentally defined early region splicing patterns in these newly identified viruses. Because of a lack of consensus splice donor and acceptor sites, we experimentally determined the splice junctions in the early transcript of STLPyV, leading us to identify a unique splicing pattern. As the predicted protein generated by this splicing event consists of 229 aa, we have termed this protein 229T. The splicing pattern that yields STLPyV 229T is distinct from the one that yields rodent MTAgs and modified versions of LTAg described in other polyomavirus species (Fig. 2B). STLPyV 229T does not arise from secondary splicing of the LTAg transcript, as seen in 17kT/T′ antigens. Furthermore, STLPyV 229T splicing is different from the splicing pattern of rodent polyomaviruses, which rely on the same splice donor site for STAg and MTAg; instead, STLPyV 229T uses a distinct downstream splice donor and acceptor site from its LTAg alternative splicing. The putative open reading frame of STLPyV 229T included the N-terminal DnaJ domain (HPDKGG) motif, which recruits and activates ATPase activity of DnaKs in studies of other polyomaviruses (Campbell et al., 1997; Srinivasan et al., 1997; Sullivan et al., 2000). No additional conserved protein motifs were detected using Motif scan and Prosite. Interestingly, MTAg of MPyV has been shown to be the major transforming protein (Rassoulzadegan et al., 1982; Treisman et al., 1981). While we do not have functional evidence for

A

B

Saint Louis Children's Study

	STLPyV	MWPyV
Number of positive samples N = 634	7 (1.1%)	14 (2.2%)

C

Saint Louis Adult Transplant Study

Specimen type	STLPyV	MWPyV
Stool (N = 237)	0	0
Plasma (N = 261)	0	0
NP (N = 261)	0	0
Urine (N = 373)	1 (0.3%)	0

D

Gambia Diarrheal Case-Control Study

	Case	Control	
STLPyV present	0	1	OR = 0 (NS)
STLPyV absent	332	389	Fisher exact test p = 1.0 (NS)

	Case	Control	
MWPyV present	5	5	OR = 1.177, CI = 0.2-4.7 (NS)
MWPyV absent	327	385	Fisher exact test p = 1.0 (NS)

Fig. 4. *Prevalance of STLPyV and MWPyV.* (A) PCR analysis of STLPyV and MWPyV is shown for control plasmids of the respective LTAg, or representative samples found to be negative, and positive for STLPyV or MWPyV. Bands corresponds to a STLPyV (481 bp) or MWPyV (484 bp) PCR product. PCR products for all positive samples were cloned and sequenced verified. (B) Prevalence of STLPyV and MWPyV in 634 stool specimens collected from a study of children from Saint Louis, as defined by the direct PCR assay described in panel A. Numbers in parenthesis indicate frequency. (C) Results of PCR screening for STLPyV and MWPyV in feces, plasma, nasopharyngeal swabs (NP), and urine specimens collected from a cohort of adult USA kidney transplant recipients. Numbers in parenthesis beside each specimen type indicate the number of specimens screened. (D) Prevalence of STLPyV and MWPyV in a Gambian diarrheal case control study of children is shown, and is based on the PCR screening assay. Odds ratio (OR), 95% confidence intervals (CI) and Fisher exact test indicate that there is no statistically evidence that STLPyV or MWPyV are significantly associated with diarrheal cases (NS=not significant).

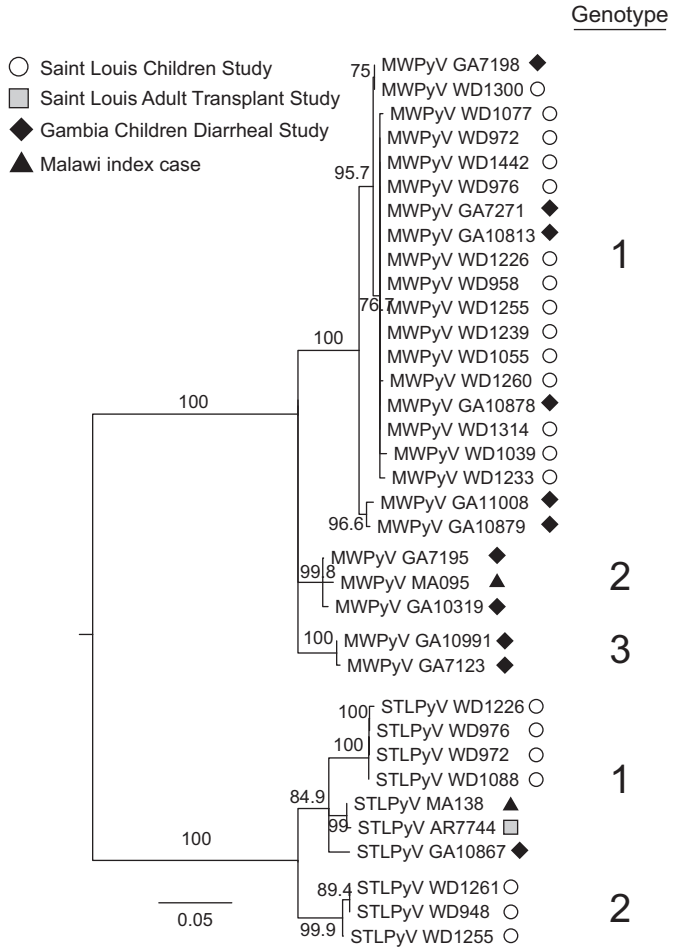


Fig. 5. *STLPyV and MWPyV strains are diverse.* Midpoint-rooted neighbor-joining phylogeny inferred from the nucleotide sequences of 9 STLPyV strains and 25 MWPyV strains identified in the three cohorts (Fig. 4) and the Malawi index case. Primer sequences were removed to yield a 437 nt alignment. The two genotypes of STLPyV and three genotypes of MWPyV, with strong bootstrap support and $>5\%$ inter-clade sequence difference, are indicated on the right. The ML method yielded a tree with similar topologies.

the role of STLPyV 229T, this raises provocative questions about the role of STLPyV 229T and whether an analogous T antigen might be present but unrecognized in other polyomaviruses.

Complex evolutionary history of polyomaviruses

The evolutionary history of polyomaviruses is complex. Phylogenetic analysis of individual genes yielded distinct topologies depending on the locus we analyzed. One constant was the consistent placement of STLPyV with MWPyV at all loci, suggesting that STLPyV shares an ancestral recombinant origin with MWPyV. The recombinant nature of these two viruses poses a conundrum for the current taxonomic scheme for polyomaviruses, which cannot accommodate the complex phylogenetic relationships of polyomaviruses that are becoming apparent. For example, in addition to the clear recombinant nature of STLPyV and MWPyV, HPyV6 and HPyV7 (currently members of *Wukipolyomavirus* genera) are evidently recombinant species as well, as shown by their discordant phylogenies; they do not form a monophyletic clade with WUPyV and KIPyV when the LTAg is analyzed (Fig. 3A and B, compare HPyV6, HPyV7 to WUPyV, KIPyV). Second, species of the *Orthpolyomavirus* taxa are paraphyletic regardless of designating either *Avipolyomavirus* or *Wukipolyomavirus* as the outgroup to root the phylogenies.

Given these issues, the current taxonomic system may need further refinement with the broad goal of making genus level taxonomic assignments concordant with phylogenetic structure. More broadly, these results underscore the importance of assessing phylogeny using multiple loci. Finally, unlike RNA viruses and ssDNA viruses, dsDNA viruses like polyomaviruses have a low mutation rate as they rely on the host polymerase for replication (Duffy et al., 2008). Hence, the clear evidence for recombination within the family *Polyomaviridae* is significant because it suggests

that polyomaviruses can compensate for their limited mutational changes by recombination to gain new, and possibly more advantageous, genetic information. Indeed, from the multiple sequence alignments, we observed many clade-specific insertions and deletions in the late region proteins (data not shown). Moreover, a more extreme recombination between two viruses of *Polyomaviridae* and *Papillomaviridae* has been previously described (Woolford et al., 2007), corroborating that recombination in *Polyomaviridae* can occur. Thus, recombination events are important milestones in the evolutionary history of polyomaviruses.

Insight into STLPyV epidemiology

In addition to the index case from Malawi, we detected STLPyV in fecal specimens collected from Saint Louis and Gambia, demonstrating that it is geographically widespread. Through the use of a degenerate PCR assay, we were able to define simultaneously the prevalence of both STLPyV and MWPyV in multiple patient cohorts. In each of the cohorts, STLPyV was found in fewer fecal samples compared to MWPyV (Fig. 4). Strikingly, both STLPyV and MWPyV were detected in three serial fecal samples collected from the same patient. This patient had received a lung transplant three years previously; however it is unclear whether immunosuppression contributed to this observation. Further screening to determine whether this observation is a statistical anomaly or whether it reflects some underlying biological linkage between the two viruses is warranted.

Although both viruses were detected in fecal specimens collected from pediatric patients, we did not detect either virus in 237 fecal specimens collected from adult renal transplant patients. Screening of nasopharygeal specimens, serum and urine from these patients yielded a single urine sample positive for STLPyV. By contrast we detected JCPyV in 45 (12%) of the urine samples tested. At this point in time, it is still not known whether either STLPyV or MWPyV causes *bona fide* infection in humans. If we assume that STLPyV does in fact infect humans, its lower prevalence than JCPyV in adult renal transplant patients may be the result of one or more of the following factors: (1) STLPyV infection may be relatively rare; (2) acute STLPyV infection may occur primarily in children; (3) STLPyV may have a distinct tropism that is not reflected by the samples tested; (4) the life cycle of STLPyV may not involve persistence and then reactivation in the context of immunosuppression.

While there are many outstanding questions regarding the potential role of STLPyV in human infection, STLPyV and the recently described MWPyV are the first two polyomaviruses to be discovered in human stool samples, raising the question as to whether they may have a primary gastrointestinal tropism. Together with the frequent detection of WUPyV and KIPyV in respiratory specimens (Babakir-Mina et al., 2011), and the apparent primary tropism of HPyV6 and HpyV7 for the skin (Schowalter et al., 2010), these findings demonstrate the ubiquity of polyomaviruses in different human specimens. As disease associations have been established for neurotropic polyomaviruses (JCPyV), renal-tropic polyomaviruses (BKPyV) and skin-tropic polyomaviruses (Merkel, TSPyV), it remains to be seen whether disease associations for the other novel polyomaviruses will emerge.

Materials and methods

Clinical specimens

The index stool specimen was obtained in January 2009 from a healthy, 15 month-old male living in Malawi as part of a human gut microbiome survey of healthy and malnourished children (Yatsunenko et al., 2012). The sample was processed by CsCl ultracentrifugation as described previously (Reyes et al., 2010).

This study was approved by the College of Medicine Research Ethics Committee of the University of Malawi; the Human Research Protection Office (HRPO) of Washington University in St. Louis, Missouri, USA; the Institutional Review Board of the University of Maryland Baltimore, Baltimore, Maryland, USA; the Joint MRC/Gambia Government Ethics Committee and by the Ethics committee of the London School of Hygiene and Tropical Medicine, UK.

A panel of 514 stool samples (Saint Louis Children's cohort) were collected previously from children age 0–18 years old, primarily with diarrheal diseases from July, 2009 through June, 2010 (Siebrasse et al., 2012). A total of 237 fecal samples, 261 plasma samples, 261 nasopharyngeal swabs and 373 urine specimens were obtained from adult kidney transplant recipients at Washington University's O'Brien Center Kidney Translational Research Core using a protocol approved by the University's HRPO. Patients enrolled in this study were defined as 'new' transplant recipients if the samples were obtained within 12 months after they received a kidney or as 'pre-existing' transplant recipients if sampling occurred more than 12 months after the transplant. In addition, we surveyed 722 fecal samples (332 cases and 390 controls) that had been collected, using a protocol approved by the Ethics committee, from children aged 0–5 years old who living in The Gambia as part of the Global Enterics Multi-Center Study (GEMS), a study of diarrhea etiologies (Kotloff et al., submitted for publication). For each enrolled child with diarrhea, one healthy control child without diarrhea matched to the case by age, gender, and time of presentation is randomly selected from the censused community in which the case resides, matched to the case by age, gender, and time of presentation.

Sample preparation for high throughput sequencing

From the index Malawi case, DNA was purified from the fecal sample, amplified by rolling circle amplification and subjected to FLX Titanium pyrosequencing as previously described (Reyes et al., 2010). Unique high quality reads with no detectable similarity to the reference human genome or NCBI nt database by BLASTn were analyzed by BLASTx alignment against the NCBI non-redundant (nr) protein database.

Amplification of complete genomes and early region splice sites

The complete genomes of the STLPyV MA138 and WD972 strains were PCR amplified in three overlapping fragments, cloned and sequenced. DNA from the samples was initially subjected to rolling circle amplification using the Illustra GenomiPhi V2 kit (GE Healthcare) prior to PCR amplification. The following primers were used: (i) STLPyVMA138-P1F (5′–GCATTCATAGGGTTTCAGAC–3′) with STLPyVMA138-P5r (5′–GTAGCAGCTGCTATATTAG–3′); (ii) STLPyVMA138-Q2F (5′–CCTTCAGGCCTGGATTGTTTTGTTACAGC–3′) with STLPyVMA138-P5r; and (iii) STLPyVMA138-P2r (5′–GGCTGGAAAACATAGAGTG–3′) with STLPyVMA138-P3F (5′–CTAATATAGCAGCTGCTAC–3′). WD972 strain was amplified using the following PCR primers: (i) STLPyVWD972-P3F (5′–CTAATATA GCAGCTGCTACAGTTG–3′) with STLPyVWD972-P5r (5′–CAACTGT AGCAGCTGCTATATTAG–3′), (ii) and STLPyVWD972-P3F with STLPyVMA138-P2r. Amplicons were cloned into pCR4 using TOPO TA cloning kit (Invitrogen) and sequenced; for the STLPyV MA138 strain primers included, STLPyVMA138-Seq1F (5′–CTGATGTTGTTGATGTGGTGGCAACCTG–3′), STLPyVMA138-Seq1r (5′–GTACATC ACATGTTTCCAAGCATGAAGGC–3′), STLPyVMA138-Seq2F (5′–CTC CAGGTAAGGGTTCTGAGCCATCTG–3′), STLPyVMA138-Seq2r (5′–GCTTTCAAATTGTGGACCAAGCAATTCACC–3′), STLPyVMA138-Seq3F

(5′–GGTAAAGTTGGAGCAGCAGGAGTTGC–3′), STLPyVMA138-Seq3r (5′–TGAGAGTAGTAACTTCGGCAGCGAGC–3′), STLPyVMA138-Seq4F (5′–ACTGCATCAGGGCCTACTTGAATGTC–3′), and STLPyVMA138-Seq4r (5′–AGTTTCAGGTGATGCTGCTGCCATTG–3′), while primers for the STLPyV WD972 strain consisted of STLPyVWD972-Seq1F (5′–GTACATCACATGTTTCCAAGCATGAAG–3′), STLPyVWD972-Seq1r (5′–CTTCATGCTTGGAAACATGTGATGTAC–3′), STLPyVWD972-Seq2F (5′–GGTTGTCCTGATACTCTGGGCATTTGG–3′), STLPyVWD972-Seq2r (5′–GGAGGACAGTTAATATTTAAGGTTCTGCC–3′), STLPyVWD972-Seq3F (5′–GTAGTCCTGCATTTTCAGCTGCCTGTAG–3′), and STLPyVWD972-Seq3r (5′–CAGTGACAGCAACTCCTGCTGCTCCAAC–3′).

A segment of STLPyV viral genome that encompasses the early region splice sites was verified from an additional sample (WD1226) found to be positive by PCR screening in the Saint Louis Children's study. PCR amplification was performed on the rolling circle amplified sample, using primers STLPyVConSpliceF (5′–CTTCTRGGGCTTCCAGAAGAYTCCTG–3′) in combination with STLPyVConSpliceR (5′–TGGGATGCAGAGGTCCCTTCATCATC–3′).

Cells and plasmids

293T cells were maintained at 37 °C in Dulbecco's modified Eagle's medium, supplemented with 10% bovine growth serum, and 1% penicillin-streptomycin, under an atmosphere of 5% CO_2/95% air. The large *T* antigen region of STLPyV MA138 from nucleotide positions 4776–2452 was cloned and ligated into a pCDNA3.1 expression vector.

RT-PCR

293T cells were seeded at 1.7×10^5 cells/ml. Plasmid DNA (500 ng) was transfected with TransIT-LT1 reagent (Mirius) according to the manufacturer's recommendations. Forty-eight hours post-transfection, cells were washed with Dulbecco's phosphate-buffered saline and removed with 0.05% trypsin. RNA was purified from the cells using a RNeasy Mini Kit (Qiagen). Reverse transcription was performed with the OneStep RT-PCR kit (Qiagen) according to the manufacturer's instructions. STLPyV mRNA was amplified with "forward primer" STLSplice1F (5′–CTTCTGGGGCTTCCAGAAGATTCCTG–3′) in combination with "reverse primer" STLSplice1r (5′–GGATGCAGAGGTCCCTTCATCATCAC–3′). A second set of primers was used to independently verify the splice junction [STLSplice2F (5′–GATCAAGCTCTCTCTAGGCAAGAAGC–3′) in combination with STLSplice2r (5′–GGTGTTGAATTCTGAGTAGAAGAATCAGG–3′)]. Bands corresponding to the unspliced, middle *T* antigen spliced, and large *T* antigen spliced transcripts were purified by QIAquick gel extraction kit (Qiagen), cloned into pCR4 vector using TOPO TA cloning kit (Invitrogen) and clones from multiple transformed bacterial colonies were sequenced.

Diversity plots and phylogenetic analyses

Nucleotide sequences of the full-length genome from the STLPyV MA138, STLPyV WD972, MWPyV MA095 and MWPyV WD976 strains were aligned by MUSCLE (Edgar, 2004), and minor editing was done manually. Diversity plots were generated with Simplot (Lole et al., 1999), employing sliding windows of 300 nt in 50 nt steps, with Kimura (2-parameter) correction.

Phylogenetic trees were constructed from alignments of the LTAg, VP1 and VP2 protein sequences from 28 polyomaviruses: avian polyomavirus (NC_004764, APyV), crow polyomavirus (NC_007922, CPyV), finch polyomavirus (NC_007923, FPyV), goose hemorrhagic polyomavirus (NC_004800, GHPyV), trichodysplasia spinulosa-associated polyomavirus (NC_014361, TSPyV), bornean orangutan polyomavirus (NC_013439, OraPyV1), chimpanzee polyomavirus (NC_014743, ChPyV), Merkel cell polyomavirus (HM0-11557, MCPyV), murine polyomavirus (NC_001515, MPyV), hamster polyomavirus (NC_001663, HaPyV), human polyomavirus 9 (NC_015150, HPyV9), B-lymphotropic polyomavirus (NC_004763, LPyV), simian virus 40 (NC_001669, SV40), BK polyomavirus (NC_001538, BKPyV), JC polyomavirus (NC_001699, JCPyV), baboon polyomavirus (NC_007611, SA12), California sea lion polyomavirus (NC_013796, SLPyV), bovine polyomavirus (NC_001442, BPyV), murine pneumotropic virus (NC_001505, MPtV), squirrel monkey polyomavirus (NC_009951, SqPyV), human polyomavirus 6 (NC_014406, HPyV6), human polyomavirus 7 (NC_014407, HPyV7), KI polyomavirus (NC_009238, KIPyV), WU polyomavirus (NC_009539, WUPyV), MW polyomavirus MA095 strain, (JQ898291, MWPyV MA095), MW polyomavirus WD976 strain (JQ898292, MWPyV WD976), and the two strains of STL polyomaviruses (STLPyV MA138 and STLPyV WD972). Alignments were performed with a probabilistic, multiple sequence alignment algorithm fast statistical alignment (FSA) (Bradley et al., 2009). Phylogenies were constructed with MrBayes v3.2.1 (Huelsenbeck and Ronquist, 2001) using a Bayesian MCMC inference (BI), and PhyML v3.0 (Guindon and Gascuel, 2003) by the maximum likelihood (ML) method. MrBayes analyses (RtREV+I+G+F) were run for 4,000,000–6,000,000 steps with a sample frequency set to 500 and a 25% burn-in period. Convergence and mixing were assessed with Tracer v1.5 (Drummond and Andrew, 2009). Analyses were performed at least twice. Support for ML trees (LG+I+G+F) was assessed by 1000 nonparametric bootstraps. The two methods yielded trees with similar topologies.

For the phylogenetic analysis of STLPyV and MWPyV sequences obtained from screening, nucleotide sequences were aligned by Muscle (Edgar, 2004) and primer sequences were trimmed from the alignment. A phylogeny was constructed by the neighbor-joining method using the Jukes Cantor method of correction (Drummond et al., 2011). Consistent results were obtained with the ML method.

Diagnostic PCR amplification

Standard precautions to avoid end product contamination were taken for all PCR assays, including the use of PCR hoods and maintaining separate areas for PCR set up and analysis. For every 88 samples tested, seven no-template negative controls were interspersed between the actual samples. Accuprime hot start Taq (Invitrogen) was used to amplify 5 μl of extracted samples using the following PCR program: 95 °C for 5 min, 40 cycles of 95 °C for 30 sec, 55 °C for 30 sec, 72 °C for 29 sec, followed by 72 °C for 10 min. STLPyV and MWPyV were detected with the "forward primer" STLMWScreenF (5′–GRATGAAAYRCWWTTACAGGTTGCCACC–3′) in combination with the "reverse primer" STLMWScreenr (5′– GTGGWAAAACAACTGTAGCWGCTGC–3′) that together generate a 481 bp (STLPyV) or 484 bp (MWPyV) amplicon from the 3′-end of the large T antigen coding region. Products were visualized following electrophoresis on 1.25% agarose gels. Amplicons were purified by QIAquick gel extraction kit (Qiagen), cloned into pCR4 using TOPO TA cloning kit (Invitrogen) and clones from multiple transformed bacterial colonies were sequenced to verify their identity.

Accession numbers

The sequences of the two complete genomes of STLPyV have been entered into the GenBank database under accession numbers JX463183 (strain MA138) and JX463184 (strain WD972). Amplicon sequences from the STLPyV and MWPyV strains have been deposited under accession numbers JX463185 – JX463215.

Acknowledgments

This work was supported in part by the NIH Grant U54 AI057160 to the Midwest Regional Center of Excellence for Biodefense and Emerging Infectious Disease Research, NIH Grant K24 DK002886 and NIH DK079333 (DCB), a Grant from the Roche Organ Transplantation Foundation, The Medical Research Council (UK) and the Bill & Melinda Gates Foundation (OPP1016839). DW holds an investigator in the pathogenesis of infectious disease award from the Burroughs Wellcome Fund. We thank Irma Bauer for performing the JCPyV screening.

References

Abend, J.R., Joseph, A.E., Das, D., Campbell-Cecen, D.B., Imperiale, M.J., 2009. A truncated T antigen expressed from an alternatively spliced BK virus early mRNA. J. Gen. Virol. 90, 1238–1245.

Allander, T., Andreasson, K., Gupta, S., Bjerkner, A., Bogdanovic, G., Persson, M.A., Dalianis, T., Ramqvist, T., Andersson, B., 2007. Identification of a third human polyomavirus. J. Virol. 81, 4130–4136.

Babakir-Mina, M., Ciccozzi, M., Perno, C.F., Ciotti, M., 2011. The novel KI, WU, MC polyomaviruses: possible human pathogens? New Microbiol. 34, 1–8.

Bhattacharjee, S., 2010. Evolutionary interrelationships among polyomaviruses based on nucleotide and amino acid variations. Indian J. Biotechnol. 9, 252–264.

Bialasiewicz, S., Rockett, R., Whiley, D.W., Abed, Y., Allander, T., Binks, M., Boivin, G., Cheng, A.C., Chung, J.Y., Ferguson, P.E., Gilroy, N.M., Leach, A.J., Lindau, C., Rossen, J.W., Sorrell, T.C., Nissen, M.D., Sloots, T.P., 2010. Whole-genome characterization and genotyping of global WU polyomavirus strains. J. Virol. 84, 6229–6234.

Bollag, B., Prins, C., Snyder, E.L., Frisque, R.J., 2000. Purified JC virus T and T′ proteins differentially interact with the retinoblastoma family of tumor suppressor proteins. Virology 274, 165–178.

Boyapati, A., Wilson, M., Yu, J., Rundell, K., 2003. SV40 17KT antigen complements dnaj mutations in large *T* antigen to restore transformation of primary human fibroblasts. Virology 315, 148–158.

Bradley, R.K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I., Pachter, L., 2009. Fast statistical alignment. PLoS Comput. Biol. 5, e1000392.

Campbell, K.S., Mullane, K.P., Aksoy, I.A., Stubdal, H., Zalvide, J., Pipas, J.M., Silver, P.A., Roberts, T.M., Schaffhausen, B.S., DeCaprio, J.A., 1997. DnaJ/hsp40 chaperone domain of SV40 large T antigen promotes efficient viral DNA replication. Genes Dev. 11, 1098–1110.

Cantalupo, P., Doering, A., Sullivan, C.S., Pal, A., Peden, K.W., Lewis, A.M., Pipas, J.M., 2005. Complete nucleotide sequence of polyomavirus SA12. J. Virol. 79, 13094–13104.

Chen, Y., Sharp, P.M., Fowkes, M., Kocher, O., Joseph, J.T., Koralnik, I.J., 2004. Analysis of 15 novel full-length BK virus sequences from three individuals: evidence of a high intra-strain genetic diversity. J. Gen. Virol. 85, 2651–2663.

Crandall, K.A., Pérez-Losada, M., Christensen, R.G., McClellan, D.A., Viscidi, R.P., 2006. Phylogenomics and molecular evolution of polyomaviruses. Adv. Exp. Med. Biol. 577, 46–59.

Daniels, R., Sadowicz, D., Hebert, D.N., 2007. A very late viral protein triggers the lytic release of SV40. PLoS Pathog. 3, e98.

Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T., Wilson, A., 2011. Geneious v5.4. Available from: 〈http://www.geneious.com/〉.

Drummond, A.J., Andrew, R., 2009. Tracer v1.5.

Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9, 267–276.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Feng, H., Shuda, M., Chang, Y., Moore, P.S., 2008. Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science 319, 1096–1100.

Gaynor, A.M., Nissen, M.D., Whiley, D.M., Mackay, I.M., Lambert, S.B., Wu, G., Brennan, D.C., Storch, G.A., Sloots, T.P., Wang, D., 2007. Identification of a novel polyomavirus from patients with acute respiratory tract infections. PLoS Pathog. 3, e64.

Gottlieb, K.A., Villarreal, L.P., 2001. Natural biology of polyomavirus middle T antigen. Microbiol Mol Biol. Rev. 65, 288–318, second and third pages, table of contents.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.

Hatwell, J.N., Sharp, P.M., 2000. Evolution of human polyomavirus JC. J. Gen. Virol 81, 1191–1200.

Huang, Y., Carmichael, G.G., 2009. RNA processing in the polyoma virus life cycle. Front Biosci. 14, 4968–4977.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Johne, R., Buck, C.B., Allander, T., Atwood, W.J., Garcea, R.L., Imperiale, M.J., Major, E.O., Ramqvist, T., Norkin, L.C., 2011. Taxonomical developments in the family Polyomaviridae. Arch. Virol 156, 1627–1634.

Johne, R., Müller, H., 2007. Polyomaviruses of birds: etiologic agents of inflammatory diseases in a tumor virus family. J. Virol. 81, 11554–11559.

Kotloff, K.L., Blackwelder, W.C., Nasrin, D., Nataro, J.P., Farag, T.H., van Eijk, A., Adegbola, R.A., Alonso, P.L., Breiman, R.F., Golam Faruque, A.S., Saha, D., Sow, S.O., Sur, D., Zaidi, A.K., Biswas, K., Panchalingam, S., Clemens, J.D., Cohen, D., Glass, R.I., Mintz, E.D., Sommerfelt, H., Levine, M.M., 2012. The Global Enteric Multicenter Study (GEMS) of diarrheal disease in infants and young children in developing countries: epidemiologic and clinical methods of the case/control study. Clin. Infect. Dis. 55 (Suppl. 4), S232–S245.

Krumbholz, A., Bininda-Emonds, O.R., Wutzler, P., Zell, R., 2009. Phylogenetics, evolution, and medical importance of polyomaviruses. Infect. Genet. Evol. 9, 784–799.

Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H.W., Ray, S.C., 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. 73, 152–160.

Pérez-Losada, M., Christensen, R.G., McClellan, D.A., Adams, B.J., Viscidi, R.P., Demma, J.C., Crandall, K.A., 2006. Comparing phylogenetic codivergence between polyomaviruses and their hosts. J. Virol. 80, 5663–5669.

Rassoulzadegan, M., Cowie, A., Carr, A., Glaichenhaus, N., Kamen, R., Cuzin, F., 1982. The roles of individual polyoma virus early proteins in oncogenic transformation. Nature 300, 713–718.

Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., Gordon, J.I., 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466, 334–338.

Sauvage, V., Foulongne, V., Cheval, J., Ar Gouilh, M., Pariente, K., Dereure, O., Manuguerra, J.C., Richardson, J., Lecuit, M., Burguière, A., Caro, V., Eloit, M., 2011. Human polyomavirus related to African green monkey lymphotropic polyomavirus. Emerg. Infect Dis. 17, 1364–1370.

Schowalter, R.M., Pastrana, D.V., Pumphrey, K.A., Moyer, A.L., Buck, C.B., 2010. Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. Cell Host Microbe 7, 509–515.

Scuda, N., Hofmann, J., Calvignac-Spencer, S., Ruprecht, K., Liman, P., Kühn, J., Hengel, H., Ehlers, B., 2011. A novel human polyomavirus closely related to the african green monkey-derived lymphotropic polyomavirus. J. Virol. 85, 4586–4590.

Sharp, P.M., Simmonds, P., 2011. Evaluating the evidence for virus/host coevolution. Curr. Opin. Virol. 1, 436–441.

Shuda, M., Feng, H., Kwun, H.J., Rosen, S.T., Gjoerup, O., Moore, P.S., Chang, Y., 2008. T antigen mutations are a human tumor-specific signature for Merkel cell polyomavirus. Proc. Nat. Acad. Sci. USA 105, 16272–16277.

Siebrasse, E.A., Bauer, I., Holtz, L.R., Le, B.M., Lassa-Claxton, S., Canter, C., Hmiel, P., Shenoy, S., Sweet, S., Turmelle, Y., Shepherd, R., Wang, D., 2012. Human polyomaviruses in children undergoing transplantation, United States, 2008-2010. Emerg. Infect. Dis. 10, 1676–1679.

Siebrasse, E.A., Reyes, A., Lim, E.S., Zhao, G., Mkakosya, R.S., Manary, M.J., Gordon, J.I., Wang, D., 2012. Identification of MW polyomavirus, a novel polyomavirus in human stool. J. Virol. 19, 10321–10326.

Srinivasan, A., McClellan, A.J., Vartikar, J., Marks, I., Cantalupo, P., Li, Y., Whyte, P., Rundell, K., Brodsky, J.L., Pipas, J.M., 1997. The amino-terminal transforming region of simian virus 40 large T and small t antigens functions as a J domain. Mol. Cell Biol. 17, 4761–4773.

Sullivan, C.S., Cantalupo, P., Pipas, J.M., 2000. The molecular chaperone activity of simian virus 40 large T antigen is required to disrupt Rb-E2F family complexes by an ATP-dependent mechanism. Mol. Cell Biol. 20, 6233–6243.

Treisman, R., Cowie, A., Favaloro, J., Jat, P., Kamen, R., 1981. The structures of the spliced mRNAs encoding polyoma virus early region proteins. J. Mol. Appl. Genet. 1, 83–92.

Trowbridge, P.W., Frisque, R.J., 1995. Identification of three new JC virus proteins generated by alternative splicing of the early viral mRNA. J. Neurovirol. 1, 195–206.

van der Meijden, E., Janssens, R.W., Lauber, C., Bouwes Bavinck, J.N., Gorbalenya, A.E., Feltkamp, M.C., 2010. Discovery of a new human polyomavirus associated with trichodysplasia spinulosa in an immunocompromized patient. PLoS Pathog. 6, e1001024.

Van Ghelue, M., Khan, M.T., Ehlers, B., Moens, U., 2012. Genome analysis of the new human polyomaviruses. Rev. Med. Virol. 6, 354–377.

Woolford, L., Rector, A., Van Ranst, M., Ducki, A., Bennett, M.D., Nicholls, P.K., Warren, K.S., Swan, R.A., Wilcox, G.E., O'Hara, A.J., 2007. A novel virus detected in papillomas and carcinomas of the endangered western barred bandicoot (*Perameles bougainville*) exhibits genomic features of both the Papillomaviridae and Polyomaviridae. J. Virol. 81, 13280–13290.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., Heath, A.C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J.G., Lozupone, C.A., Lauber, C., Clemente, J.C., Knights, D., Knight, R., Gordon, J.I., 2012. Human gut microbiome viewed across age and geography. Nature 486, 222–227.

Zerrahn, J., Knippschild, U., Winkler, T., Deppert, W., 1993. Independent expression of the transforming amino-terminal domain of SV40 large I antigen from an alternatively spliced third SV40 early mRNA. EMBO J. 12, 4739–4746.

JVI Journals.ASM.org

Identification of MW Polyomavirus, a Novel Polyomavirus in Human Stool

Erica A. Siebrasse,[a] Alejandro Reyes,[b] Efrem S. Lim,[a] Guoyan Zhao,[a] Rajhab S. Mkakosya,[d] Mark J. Manary,[c] Jeffrey I. Gordon,[b] and David Wang[a]

Departments of Molecular Microbiology and Pathology & Immunology,[a] Center for Genome Sciences and Systems Biology,[b] and Department of Pediatrics,[c] Washington University School of Medicine, St. Louis, Missouri, USA, and Department of Pathology, College of Medicine, University of Malawi, Chichiri, Blantyre, Malawi[d]

We have discovered a novel polyomavirus present in multiple human stool samples. The virus was initially identified by shotgun pyrosequencing of DNA purified from virus-like particles isolated from a stool sample collected from a healthy child from Malawi. We subsequently sequenced the virus' 4,927-bp genome, which has been provisionally named MW polyomavirus (MWPyV). The virus has genomic features characteristic of the family *Polyomaviridae* but is highly divergent from other members of this family. It is predicted to encode the large T antigen and small T antigen early proteins and the VP1, VP2, and VP3 structural proteins. A real-time PCR assay was designed and used to screen 514 stool samples from children with diarrhea in St. Louis, MO; 12 specimens were positive for MWPyV. Comparison of the whole-genome sequences of the index Malawi case and one St. Louis case demonstrated that the two strains of MWPyV varied by 5.3% at the nucleotide level. The number of polyomaviruses found in the human body continues to grow, raising the question of how many more species have yet to be identified and what roles they play in humans with and without manifest disease.

Over the past 5 years, seven novel polyomaviruses have been discovered in humans, including KI polyomavirus (KIPyV) (2), WU polyomavirus (WUPyV) (13), Merkel cell polyomavirus (MCPyV) (11), human polyomavirus 6 (HPyV6) (40), human polyomavirus 7 (HPyV7) (40), trichodysplasia spinulosa-associated polyomavirus (TSPyV) (45), and human polyomavirus 9 (HPyV9) (42). Polyomaviruses also infect a wide variety of mammalian and avian hosts, including the recently described novel polyomaviruses of bats (*Myotis* species) (32), sea lions (*Zalophus californianus*) (8, 47), multimammate mice (*Mastomys* species) (33), canaries (*Serinus canaria*) (19), orangutans (*Pongo* species) (17), squirrel monkeys (*Saimiri* species) (46), chimpanzees (*Pan troglodytes* subsp. *verus*) (28), and gorillas (*Gorilla gorilla*) (28).

Viruses in the *Polyomaviridae* family typically possess ~5,000-bp circular, double-stranded DNA genomes. The genome can be divided into three parts—the regulatory region, the early region, and the late region. The regulatory region, also called the noncoding control region (NCCR), contains the origin of replication and promoters for the early and late regions. Transcription occurs bidirectionally from the regulatory region. The early region is expressed from a common primary transcript and is alternatively spliced to produce the large T antigen (LTAg) and small T antigen (STAg) prior to viral replication. LTAg and STAg typically share the first ~80 amino acids. The late region is expressed after viral replication has begun and encodes the structural proteins VP1, VP2, and VP3. VP1, the major structural protein, typically comprises over 70% of the viral particle and is the antigenic portion of the virus to which most natural antibodies are made (20).

Disease associations have been established for some of the human polyomaviruses. The two well-studied human polyomaviruses BK polyomavirus (BKPyV) and JC polyomavirus (JCPyV) are important human pathogens. BKPyV is known to cause BK nephropathy, which can lead to renal allograft failure, and hemorrhagic cystitis, while JCPyV is the etiological agent of progressive multifocal leukoencephalopathy (PML). Both viruses are ubiquitous worldwide, with seroprevalence rates of 55 to 85% for BKPyV and 44 to 77% for JCPyV (25). Following primary infection in childhood, BKPyV and JCPyV establish persistent latent infections that can periodically reactivate, leading to shedding of infectious virus in the urine (36). Primary infection and periodic reactivation are typically asymptomatic unless the host is immunocompromised, in which case life-threatening illness can occur (36). MCPyV is associated with Merkel cell carcinoma (MCC), a rare but aggressive skin cancer. MCPyV DNA is found in ~80% of MCC tumors and is clonally integrated into a subset of these (14). TSPyV has been linked to trichodysplasia spinulosa, a very rare skin condition associated with immunosuppression following organ transplantation (24). It is unclear if the other human polyomaviruses play a role in disease.

The recently discovered human polyomaviruses have all been identified through the use of molecular methods for detection of viral nucleic acids. WUPyV and KIPyV were discovered using high-throughput Sanger sequencing (2, 13). MCPyV was identified using digital transcriptome subtraction, which entails pyrosequencing of a cDNA library followed by subtraction of human reads to identify novel viral sequences (11). HPyV6 and TSPyV were discovered using rolling circle amplification (RCA) (40, 45), and consensus PCR primers were utilized to find HPyV7 and HPyV9 (40, 42).

We used shotgun pyrosequencing of purified virus-like particles (VLPs) recovered from a fecal sample to discover a novel polyomavirus in the stool of a healthy child from Malawi. The virus was also detected in 12 additional stool samples from the United States, indicating it has a wide geographic distribution. As

Received 18 May 2012 **Accepted** 19 June 2012
Published ahead of print 27 June 2012
Address correspondence to David Wang, davewang@borcim.wustl.edu.

doi:10.1128/JVI.01210-12

stool is not a sterile site, it is currently unknown whether this polyomavirus actively infects humans. Finally, we compared the whole genome nucleotide sequences of the index Malawi case and a case from St. Louis and found these two strains to have 5.3% nucleotide variation.

MATERIALS AND METHODS

Human studies. This study was approved by the College of Medicine Research and Ethics Committee of the University of Malawi and the Human Research Protection Office of Washington University in St. Louis. The index stool specimen was obtained from a healthy, breast-fed, 15-month-old female living in Mayaka, Malawi, in September 2008 as part of a global gut microbiome survey (48).

A total of 514 stool specimens from St. Louis were tested for MWPyV. Stool samples were from children, age 0 to 18 years, with diarrhea and were submitted to the St. Louis Children's Hospital, St. Louis, MO, microbiology laboratory for bacterial culture from July 2009 to June 2010.

Sample preparation and 454 pyrosequencing. VLPs were purified as described earlier (37) with minor modifications. In brief, 50 mg of a frozen fecal sample was resuspended in 400 μl of SM buffer (100 mM NaCl, 8 mM $MgSO_4$, 50 mM Tris [pH 7.5], and 0.002% gelatin [wt/vol]). Following centrifugation (2,500 × *g* for 10 min at 4°C) and filtration through 0.45- and 0.22-μm-pore-size Millex filters (Millipore) to remove bacterial cells and large particles, the sample was treated with chloroform (0.2 volumes) for 10 min and centrifuged for 5 min at 2,500 × *g*. The aqueous phase was treated with Baseline-Zero DNase (2.5 U/ml) (Epicentre) for 1 h at 37°C to remove free DNA, followed by an incubation at 65°C for 15 min to inactivate the enzyme. To extract VLP-associated DNA, the solution was treated with 10 μl 10% SDS and 3 μl proteinase K (20 mg/ml) for 20 min at 56°C. Subsequently, 35 μl of 5 M NaCl and 28 μl of a solution of 10% cetyltrimethylammonium bromide-0.7 M NaCl were introduced. After a 10-min incubation at 65°C, an equal volume of chloroform was added, and the mixture was centrifuged for 5 min at 8,000 × *g* at room temperature. The supernatant was transferred to a new tube, and an equal volume of phenol-chloroform-isoamyl alcohol (25:24:1) was added, followed by centrifugation for 5 min at 8,000 × *g* at room temperature. The supernatant was collected, and the DNA was purified using Qiagen MiniElute columns by following the manufacturer's instructions, with a final elution volume of 35 μl.

Purified VLP-derived DNA (1 μl) was used as an input in a 20-μl RCA reaction mixture using the Illustra GenomiPhi V2 kit (GE Healthcare) as recommended by the manufacturer (n = four independent reactions). After 90 min of amplification, the four reactions were pooled and purified using the Qiagen DNeasy kit. DNA (500 ng) was subjected to 454 FLX Titanium pyrosequencing.

Analysis of pyrosequencing reads. The individual 454 reads were analyzed using a custom bioinformatics pipeline as previously described (7). In brief, unique, high-quality reads were aligned against the reference human genome and the GenBank nucleotide database using BLASTn. Reads with no hits or hits with an E-value greater than e^{-5} were then aligned using BLASTx to the GenBank nr (nonredundant) database, and reads aligning to viral sequences with the lowest E value were identified.

Complete genome sequencing. PCR primers were designed to span the gaps between the six reads showing significant similarity to polyomaviruses generated by pyrosequencing to obtain an initial whole-genome sequence. The sequences for these primers are available upon request. The complete MWPyV genome derived from the index Malawi case (designated strain MA095, GenBank JQ898291) was sequenced to greater than 3× coverage using four sets of overlapping PCR primers. They were (listed 5′ to 3′) ACTTAAACCATGTTCCTGACTCTGT (ES087) and AC AGAGATTACAGCACCCATATACT (ES091), GCATCTGCCCTGGTA CAAACA (ES088) and CAGACAACTCAGAAGTTTCCACCTC (ES092), GAAGTAGAAGGAGAGGAAAATGCCG (ES089) and TGCTGTTGAG GATACACAACAAGAC (ES093), and AGGCTGCTTAAAGGCCTATG AATG (ES090) and CTGAAACACCAGTTGCTCCAGC (ES094). Amplicons from independent PCRs were cloned into pCR4 (Invitrogen) and bidirectionally sequenced. The complete genome from St. Louis sample WD976 (strain WD976; GenBank accession no. JQ898292) was amplified and sequenced to greater than 3× coverage in the same manner, using the same primer pairs.

Genome annotation. Open reading frames (ORFs) were predicted using NCBI ORF Finder. The LTAg and STAg ORFs were manually scanned for conserved splice donor and acceptor sites. Conserved motifs in the TAgs and in the NCCR were identified using NCBI CD-Search software (30) and by manual identification. Prediction of putative binding sites for transcription factors was performed using AliBaba software, version 2.1 (15). The NCCR region was scanned for palindrome patterns using the EMBOSS palindrome software (38).

Phylogenetic analysis. Protein sequences associated with the reference genomes for 27 polyomaviruses were obtained from GenBank; these included baboon polyomavirus (NC_007611; SA12) (6), bat polyomavirus (NC_011310; BatPyV) (32), B-lymphotropic polyomavirus (NC_004763; LPyV) (34), BKPyV (NC_001538) (43), Bornean orangutan polyomavirus (NC_013439; OraV1) (17), bovine polyomavirus (NC_001442; BPyV) (41), California sea lion polyomavirus (NC_013796; SLPyV) (8), hamster polyomavirus (NC_001663; HaPyV) (9), JCPyV (NC_001699) (12), MCPyV (HM011557) (40), murine pneumotropic virus (NC_001505; MPtV) (31), murine polyomavirus (NC_001515; MPyV) (16), simian virus 40 (NC_001669; SV40) (27), squirrel monkey polyomavirus (NC_009951; SqPyV) (46), Sumatran orangutan polyomavirus (FN356901; OraV2) (17), TSPyV (NC_014361) (45), HPyV6 (NC_014406) (40), HPyV7 (NC_014407) (40), KIPyV (NC_009238) (2), WUPyV (NC_009539) (13), avian polyomavirus (NC_004764; APyV) (39), canary polyomavirus (GU345044; CaPyV) (19), crow polyomavirus (NC_007922; CPyV) (23), finch polyomavirus (NC_007923; FPyV) (23), goose hemorrhagic polyomavirus (NC_004800; GHV) (22), chimpanzee polyomavirus (NC_014743; ChPyV) (10), and HPyV9 (NC_015150) (42). The predicted open reading frames for MWPyV LTAg, VP1, and VP2 were aligned with the corresponding proteins from the 27 known polyomaviruses using Fast Statistical Alignment (FSA) software, version 1.15.2 (5). For the LTAg analysis, unalignable regions were removed, and the remainder of the alignment was concatenated. Maximum likelihood trees were generated using PhyML, version 3.0 (18), with 1,000 bootstrap replicates and the best model as determined by Prot Test software, version 2.4 (1); these were RtRev for VP1 and LG for VP2 and LTAg.

Nucleic acid extraction. Stools which had been frozen at −80°C were diluted approximately 1:6 in phosphate-buffered saline (PBS) and filtered through 0.45-μm-pore-size membranes prior to extraction. Total nucleic acids were extracted using an Ampliprep Cobas automated extractor (Roche) and eluted in a volume of 75 μl. The samples were arrayed in a 96-well plate for storage at −80°C.

Real-time PCR screening of the St. Louis cohort. A TaqMan real-time PCR assay was designed to target the MWPyV LTAg using Primer Express software (Applied Biosystems). Primers and probe used for this assay were 5′-TGAGAAGGCCCCGGTTCT-3′ (ES105), 5′-GAGGATGG GATGAAGATTTAAGTTG-3′ (ES106), and 5′-FAM-CCTCAT-CACTGGGAGC-MGBNFQ-3′ (ES107) (where FAM is 6-carboxyfluorescein). The resulting amplicon was 73 bp. Standard curves were generated using serial 10-fold dilutions ranging from 5×10^6 to 5 copies of a positive-control plasmid (plasmid K-p31) per reaction. The 25-μl PCR mixtures consisted of 5 μl of extracted sample, 1× universal TaqMan real-time PCR master mix (Applied Biosystems), 12.5 pmol of each primer, and 4 pmol of the probe. Samples were tested in a 96-well-plate format, with eight water negative controls (one per row) and one positive control containing 50 copies of plasmid per plate. The cycling conditions were 50°C for 2 min, 95°C for 10 min, and 45 cycles of 95°C for 15 s followed by 60°C for 1 min. Reactions were run on an ABI 7500 real-time thermocycler (Applied Biosystems). The threshold of all plates was set at a

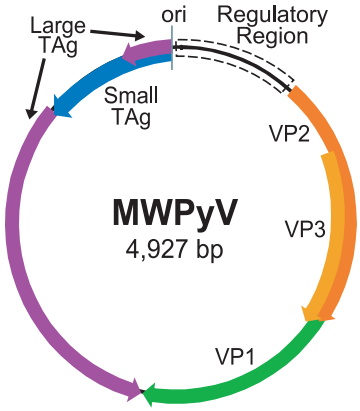


FIG 1 Genome organization of MWPyV. ori, origin of replication.

standard value, and the data were analyzed using the ABI software. Samples were counted as positive if their threshold cycle (C_T) value was <35.

Nucleotide sequence accession numbers. The sequences reported here were deposited in GenBank under the accession numbers JQ898291 (index case, strain MA095) and JQ898292 (St. Louis case, strain WD976).

RESULTS

Discovery of a novel polyomavirus by pyrosequencing. MW polyomavirus was discovered in a stool sample from a child from Malawi; the sample was collected in September 2008 as part of a global gut microbiome survey project (48). Following purification of VLPs by passage through 0.45- and 0.22-μm-pore-sized filters and subsequent DNase treatment, DNA was extracted from the VLPs and amplified using the highly processive phi29 polymerase. The resulting material was subjected to 454 pyrosequencing. Six reads were identified with limited similarity to known polyomaviruses. Three of the initial six reads could be assembled into one 959-bp contig, with the highest scoring BLASTx hit possessing 36% amino acid identity to LPyV STAg. The other three reads all aligned to the VP1 protein of known polyomaviruses by BLASTx and shared 64%, 48%, and 59% amino acid identity to JCPyV VP1, TSPyV VP1, and JCPyV VP1, respectively.

Complete genome sequencing and genome analysis. A series of PCR primers was designed based on the initial six reads. Sequencing of the resulting amplicons yielded a complete genome of 4,927 bp (Fig. 1). ICTV has set the demarcation criteria for proposed new polyomaviruses at 81% nucleotide identity over the whole genome (21). Based on the limited sequence similarity to any known polyomaviruses, we named the novel virus MW polyomavirus (MWPyV) after its discovery in Malawi. The overall GC content of MWPyV was 37%, which is very similar to those of WUPyV (39%), BKPyV (39%), and JCPyV (40%). The MWPyV genome organization was characteristic of the known polyomaviruses and included an early region coding on one strand for LTAg and STAg and a late region coding on the opposite strand for the structural proteins VP1, VP2, and VP3. The sizes of the predicted ORFs were comparable to those of known polyomaviruses (Table 1).

The TAgs and VP2 were separated by a regulatory region, which had an A/T-rich tract on the late side of the putative replication origin. The core origin of replication contained three repeats of the consensus pentanucleotide LTAg binding site, G(A/G)GGC (35) (two GAGGC and one GGGGC), and one nonconsensus binding site, TAGGC. Several polyomaviruses (BKPyV, JCPyV, WUPyV, KIPyV, and SV40) contain an imperfect palindrome sequence followed by additional LTAg binding sites to the early side of the four binding sites. Palindrome patterns were identified in MWPyV, but no additional LTAg binding sites were detected in this area. The regulatory region contained several predicted transcription factor binding sites, including multiple binding sites for four factors known to play a role in BKPyV viral transcription and regulation: Sp1, nuclear factor I (NFI), AP1, and C/EBP (29). Multiple binding sites were also identified for HNF-3, USF-2, and Oct-1. Many other transcription factors were predicted to bind to only one site.

Analysis of the MWPyV LTAg ORF revealed a conserved splice donor site immediately after amino acid 80; the position of this site was similar to that found in WUPyV, BKPyV, and JCPyV, which occur after amino acids 84, 81, and 81, respectively. Two consensus splice acceptor sites were identified, which would yield introns of 355 or 463 bp and proteins of 668 or 632 amino acids, respectively. Examination of the protein sequence of the 632-amino-acid form showed that it lacked the Rb-binding motif, which was contained in the excised intron. In contrast, the predicted 668-amino-acid protein included the conserved Rb-binding motif. Based on this analysis, we predicted the LTAg to be 668 amino acids.

MWPyV LTAg possessed conserved features common to other polyomavirus LTAgs, including a DNaJ domain containing the conserved region 1 (cr1) sequence and the highly conserved hexapeptide motif HPDKGG. These domains were followed by conserved region 2 (cr2), which contained the Rb-binding motif LxCxE (LSCNE in MWPyV), a putative nuclear localization signal (NLS), a canonical DNA binding domain, and a zinc finger region. Closer inspection of the zinc finger region revealed a conserved C_2H_2 zinc finger motif with the sequence C324, C327, H334, H339. There are typically three highly conserved amino acids N terminal to the first cysteine (C324) in this motif, including a tyrosine 10 amino acids away, an aspartic acid located 18 amino acids away, and an alanine present 25 amino acids away (35). In MWPyV, the aspartic acid and alanine residues were conserved, while the tyrosine was not and was replaced by a leucine. A conserved leucine-rich hydrophobic region C terminal to the aspartic acid was also present. Following the zinc finger region, the MWPyV LTAg contained the highly conserved ATPase-p53 binding domain, including the two conserved motifs GPXXXGKT and GXXXVNLE. There was no sequence corresponding to the host range domain present in SV40, BKPyV, SA12, and JCPyV (35).

In most polyomaviruses, STAg is encoded by a single unspliced ORF. In HaPyV and MPyV, the STAg transcript is spliced. Analysis of the MWPyV early region did not reveal an obvious splice donor site, so the STAg was predicted to be 199 amino acids. As

TABLE 1 Putative proteins encoded by MWPyV (strain MA095)

Protein	Putative coding region(s)	Predicted size (aa)	Calculated mass (kDa)	Range (aa) in other polyomaviruses
STAg	4927–4328	199	23.4	124–198
LTAg	4927–4688, 4332–2566	668	77.0	599–817
VP1	1353–2564	403	43.6	343–497
VP2	431–1363	310	34.2	241–415
VP3	761–1363	200	22.8	190–272

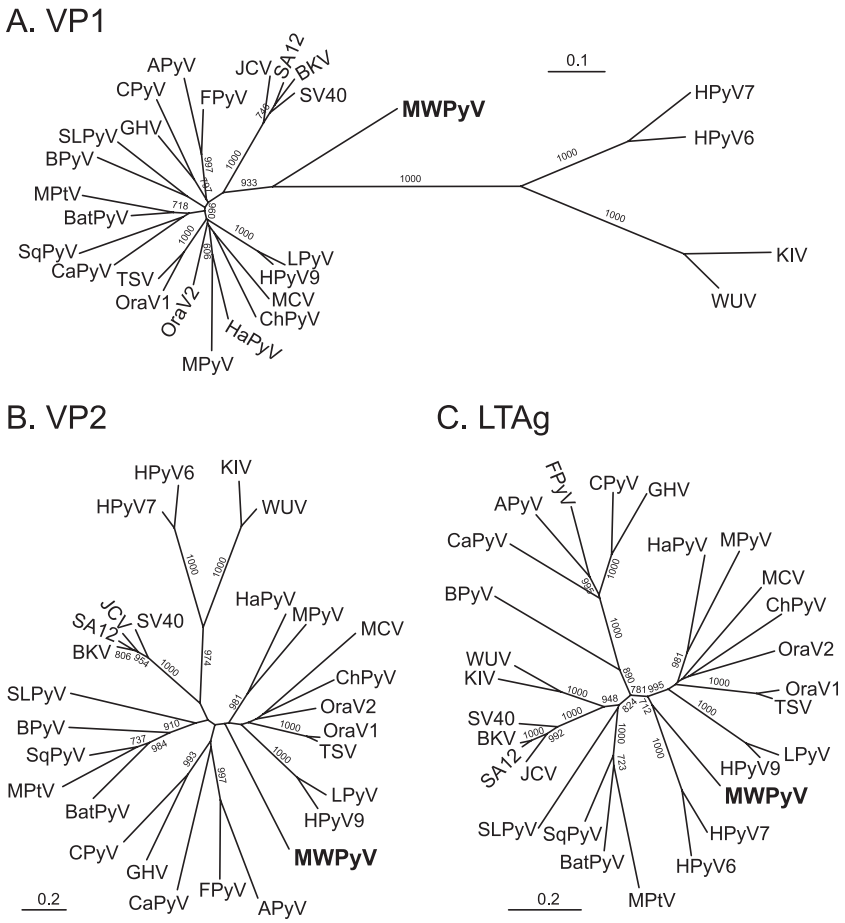


FIG 2 Phylogenetic analysis of MW polyomavirus. Amino-acid-based trees were generated using the maximum likelihood method with 1,000 bootstrap replicates. Bootstrap values less than 700 are not shown. (A) VP1; (B) VP2; (C) LTAg.

LTAg and STAg share the first 80 amino acids, the STAg also contained the DNaJ domain. In the unique C-terminal part of STAg, there was a conserved cysteine-rich motif, $CX_5CX_{7-8}CXCX_2CX_{21-22}CSCX_2CX_3WFG$. This motif was conserved in MWPyV with the exception of the initial cysteine residue and the serine residue, which were an isoleucine and a phenylalanine, respectively.

MPyV and HaPyV encode a middle T antigen (MTAg) generated by alternative splicing; the MWPyV genome was scanned for splicing motifs similar to those used by MPyV and HaPyV. No obvious splice sites that would generate an appropriately sized third T antigen protein were identified, suggesting that MWPyV likely does not encode an MTAg.

Some polyomaviruses, including JCPyV and BKPyV, also encode an agnoprotein in the late region between the NCCR and the VP2 start codon. Analysis of the MWPyV sequence in this region yielded one 45-amino-acid ORF on the same strand as the structural proteins. However, because this ORF was not conserved in the other completely sequenced MWPyV strain, strain WD976 (described later in this report), we do not believe that MWPyV encodes an agnoprotein.

Phylogenetic analysis. Maximum likelihood analysis of the VP1, VP2, and LTAg proteins demonstrated that MWPyV was highly divergent from all known polyomaviruses (Fig. 2). Analysis of VP1 sequences showed that MWPyV is midway between the *Wukipolyomavirus* and *Orthopolyomavirus* genera (Fig. 2A). In contrast, based on VP2 and LTAg sequences, MWPyV clustered with the clade containing HPyV9, LPyV, HaPyV, MPyV, TSPyV, MCV, ChPyV, and the orangutan polyomaviruses (Fig. 2B and C). The discordant phylogenetic relationships suggest that MWPyV might have been derived from an ancestral recombination event.

Prevalence of MWPyV. A TaqMan real-time PCR assay targeting the MWPyV LTAg was designed and validated using a positive-control plasmid; based on the standard curve, the MWPyV assay demonstrated a reliable detection limit of approximately five copies per reaction, yielded a linear regression R^2 value of 0.99, and was 93% efficient. This real-time PCR assay was used to screen a cohort consisting of 514 stool samples from children at St. Louis Children's Hospital presenting with diarrhea. Twelve samples (2.3%) from the St. Louis cohort tested positive for MWPyV (Table 2).

Three of the MWPyV-positive samples were obtained from a 5-year-old lung transplant recipient over a period of 4 months from August to December 2009 (patient 1; Table 2). This patient had received a lung transplant 3 years earlier and at the time of sampling presented with persistent, recurrent diarrhea. Two of the samples, WD972 and WD976, were obtained on consecutive days in August 2009, and both samples were positive for *Escherichia coli*

TABLE 2 Specimens and patients testing positive for MWPyV

Sample	Patient	Age	Sex	C_T	Date (mo/day/yr)	Tested positive	Tested negative
WD972	1	5 yr 0 mo	M	21.68	8/10/09	*E. coli* serotype O Rough	Enteric pathogen culture[a] (except *E. coli*), *Giardia*, *Cryptosporidium*, ova & parasite screen (O&P)
WD976	1			21.85	8/11/09	*E. coli* serotype O Rough	Enteric pathogen culture (except *E. coli*), *Giardia*, *Cryptosporidium*, *C. difficile*, O&P
WD1226	1			28.21	12/22/09		Enteric pathogen culture, O&P
WD1239	2	1 yr 0 mo	M	31.76	12/29/09		Enteric pathogen culture, *C. difficile*
WD1314	3	1 yr 5 mo	F	30.37	2/11/10		Enteric pathogen culture, rotavirus
WD1300	4	1 yr 8 mo	F	34.99	2/4/10		Enteric pathogen culture, rotavirus, viral culture
WD1260	5	1 yr 4 mo	F	29.91	1/12/10		Enteric pathogen culture, O&P
WD958	6	1 yr 8 mo	M	31.89	8/3/09		Enteric pathogen culture, *C. difficile*, O&P
WD1039	7	4 yr 3 mo	F	31.42	9/14/09		Enteric pathogen culture, *C. difficile*
WD1233	8	4 yr 9 mo	M	32.33	12/25/09		Enteric pathogen culture, rotavirus
WD1055	9	5 yr 5 mo	M	30.90	9/22/09		Enteric pathogen culture
WD1442	10	3 yr 0 mo	M	32.41	5/24/10	*C. jejuni*	Enteric pathogen culture (except *C. jejuni*)

[a] Enteric pathogen culture includes *Salmonella*, *Shigella*, *E. coli* O157, *E. coli* Shiga toxins not 0157, *Yersinia*, *Aeromonas*, *Plesiomonas*, and *Campylobacter*.

serotype O Rough. The patient again presented with diarrhea in December 2009 (sample WD1226), but this sample had no growth in the enteric pathogen culture (including *E. coli*) and was negative for ova and parasites (Table 2). The other nine samples came from nine individual patients, ranging in age from 1 to 5 years (Table 2). Eight of the nine patients were negative for all organisms tested except MWPyV. Only patient 10 (sample WD1442) was positive for *Campylobacter jejuni*.

Strain variation. To assess the extent of sequence variation between the St. Louis and Malawi isolates, we sequenced the complete genome of MWPyV from St. Louis sample WD976 to greater than 3× coverage. The two whole-genome sequences diverged by 5.3% at the nucleotide level. Strain WD976 had two insertions (11 bp and 1 bp) in the NCCR, which resulted in a genome size of 4,939 bp. The vast majority of the polymorphisms in the coding regions resulted in synonymous mutations. One notable mutation changed the size of the STAg ORF. The predicted TAA stop codon identified in the MA095 strain was mutated to AAA in WD976, resulting in a protein prediction of 206 amino acids, seven amino acids longer than the index genome's STAg.

DISCUSSION

We used a pyrosequencing strategy to identify a novel polyomavirus present in human stool. The initial discovery was in a stool specimen collected from a healthy child in Malawi. Further screening by real-time PCR demonstrated the presence of the virus in 12 stool samples collected from a cohort of patients in St. Louis, MO. These data demonstrated that MWPyV is geographically widespread in human populations and can be found on two continents. As the ICTV polyomavirus subgroup currently has no systematic naming convention for novel polyomaviruses, we chose to name this new virus using a two-letter convention following the model of BKPyV, JCPyV, KIPyV, and WUPyV; we made this decision for two reasons. First, we did not employ the numerical system used in the naming of HPyV6, HPyV7, and HPyV9 because we have not yet formally demonstrated that this virus infects humans and to avoid potential conflicts in temporal priority in describing novel polyomaviruses. Second, both MCPyV and TSPyV are named based on putative disease associations, but no disease association currently exists for our new virus. Therefore, we chose a two-letter abbreviation reflecting the geographic location of the index case.

The ICTV polyomavirus subgroup recently defined two mammalian genera, *Orthopolyomavirus* and *Wukipolyomavirus*, within the family *Polyomaviridae* based primarily on phylogenetic analysis of the late genes (combined VP1 and VP2) (21). Classification of MWPyV into one of these two genera is confounded by the distinct phylogenetic tree topologies that were generated for the VP1 and VP2 proteins (Fig. 2). The different topologies suggest that MWPyV is derived from an ancestral recombination event. Such recombination among polyomaviruses has been previously suggested (40).

We sequenced two complete genomes of MWPyV, one from the index child in Malawi and one from a child in St. Louis. A high degree of strain variation (5.3%) was observed between these two MWPyV strains, which is comparable to the ~5% sequence divergence present in strains of BKPyV (26). It contrasts sharply with the very limited variation (<1.2%) seen with WUPyV worldwide (3). The primers and probe used in the aforementioned MWPyV real-time PCR assay were perfectly conserved in both strains and thus detect both stains with equal efficiency. However, it remains to be determined whether even greater variation in MWPyV can be discovered when broader consensus sequence-based assays are used. Others have speculated that sequence variation in BKPyV and JCPyV plays a role in viral pathogenesis and disease severity (4, 44). If MWPyV is ultimately found to be a pathogen, it will be interesting to determine whether there are strain-dependent pathogenic phenotypes. Among the differences we observed were a 7-amino-acid extension of the STAg and an 11-bp insertion in the NCCR in the WD976 strain versus the index Malawi strain. The functional consequences of these alterations remain to be defined.

One critical question is whether MWPyV is a bona fide infectious agent of humans and if so, what disease(s), if any, might be associated with MWPyV infection. The detection of MWPyV in stools of children with diarrhea, many of which have no known infection etiology, raises the possibility that MWPyV might play a role in human diarrhea. Alternatively, it is possible that MWPyV does not cause infection in the gastrointestinal tract but has a

tropism for other human organ systems and is shed in stool as a mode of transmission or simply as a by-product. It is also possible that MWPyV is a dietary contaminant and does not actively infect humans. Approaches to answer whether MWPyV is an infectious agent include serological studies to determine whether the host mounts an antibody-based immune response to MWPyV and additional screening of specimens collected from sterile sites, such as serum or cerebrospinal fluid. Further studies will be needed to define whether MWPyV has additional tropisms in the human body and to assess potential associations with human disease.

ACKNOWLEDGMENTS

This work was supported in part by NIH grant U54 AI057160 to the Midwest Regional Center of Excellence for Biodefense and Emerging Infectious Disease Research and the Bill and Melinda Gates Foundation. E.A.S. is supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

REFERENCES

1. **Abascal F, Zardoya R, Posada D.** 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics **21**:2104–2105.
2. **Allander T, et al.** 2007. Identification of a third human polyomavirus. J. Virol. **81**:4130–4136.
3. **Bialasiewicz S, et al.** 2010. Whole-genome characterization and genotyping of global WU polyomavirus strains. J. Virol. **84**:6229–6234.
4. **Boldorini R, et al.** 2009. Genomic mutations of viral protein 1 and BK virus nephropathy in kidney transplant recipients. J. Med. Virol. **81**:1385–1393.
5. **Bradley RK, et al.** 2009. Fast statistical alignment. PLoS Comput. Biol. **5**:e1000392. doi:10.1371/journal.pcbi.1000392.
6. **Cantalupo P, et al.** 2005. Complete nucleotide sequence of polyomavirus SA12. J. Virol. **79**:13094–13104.
7. **Cantalupo PG, et al.** 2011. Raw sewage harbors diverse viral populations. mBio **2**(5):e00180–12. doi:10.1128/mBio.00180–11.
8. **Colegrove KM, et al.** 2010. Polyomavirus infection in a free-ranging California sea lion (*Zalophus californianus*) with intestinal T-cell lymphoma. J. Vet. Diagn. Invest. **22**:628–632.
9. **Delmas V, Bastien C, Scherneck S, Feunteun J.** 1985. A new member of the polyomavirus family: the hamster papovavirus. Complete nucleotide sequence and transformation properties. EMBO J. **4**:1279–1286.
10. **Deuzing I, et al.** 2010. Detection and characterization of two chimpanzee polyomavirus genotypes from different subspecies. Virol. J. **7**:347.
11. **Feng H, Shuda M, Chang Y, Moore PS.** 2008. Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science **319**:1096–1100.
12. **Frisque RJ, Bream GL, Cannella MT.** 1984. Human polyomavirus JC virus genome. J. Virol. **51**:458–469.
13. **Gaynor AM, et al.** 2007. Identification of a novel polyomavirus from patients with acute respiratory tract infections. PLoS Pathog. **3**:e64. doi:10.1371/journal.ppat.0030064.
14. **Gjoerup O, Chang Y.** 2010. Update on human polyomaviruses and cancer. Adv. Cancer Res. **106**:1–51.
15. **Grabe N.** 2002. AliBaba2: context specific identification of transcription factor binding sites. In Silico Biol. **2**:S1–S15.
16. **Griffin BE, Fried M, Cowie A.** 1974. Polyoma DNA: a physical map. Proc. Natl. Acad. Sci. U. S. A. **71**:2077–2081.
17. **Groenewoud MJ, et al.** 2010. Characterization of novel polyomaviruses from Bornean and Sumatran orang-utans. J. Gen. Virol. **91**:653–658.
18. **Guindon S, et al.** 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59**:307–321.
19. **Halami MY, et al.** 2010. Whole-genome characterization of a novel polyomavirus detected in fatally diseased canary birds. J. Gen. Virol. **91**:3016–3022.
20. **Imperiale MJ, Major EO.** 2007. Polyomaviruses. *In* Knipe DM, Howley PM (ed), Fields virology, 5th ed, vol 2. Lippincott Williams & Wilkins, Philadelphia, PA.
21. **Johne R, et al.** 2011. Taxonomical developments in the family Polyomaviridae. Arch. Virol. **156**:1627–1634.
22. **Johne R, Muller H.** 2003. The genome of goose hemorrhagic polyomavirus, a new member of the proposed subgenus Avipolyomavirus. Virology **308**:291–302.
23. **Johne R, Wittig W, Fernandez-de-Luco D, Hofle U, Muller H.** 2006. Characterization of two novel polyomaviruses of birds by using multiply primed rolling-circle amplification of their genomes. J. Virol. **80**:3523–3531.
24. **Kazem S, et al.** 2012. Trichodysplasia spinulosa is characterized by active polyomavirus infection. J. Clin. Virol. **53**:225–230.
25. **Knowles WA.** 2006. Discovery and epidemiology of the human polyomaviruses BK virus (BKV) and JC virus (JCV). Adv. Exp. Med. Biol. **577**:19–45.
26. **Krumbholz A, Bininda-Emonds OR, Wutzler P, Zell R.** 2009. Phylogenetics, evolution, and medical importance of polyomaviruses. Infect. Genet. Evol. **9**:784–799.
27. **Lebowitz P, Weissman SM.** 1979. Organization and transcription of the simian virus 40 genome. Curr. Top. Microbiol. Immunol. **87**:43–172.
28. **Leendertz FH, et al.** 2011. African great apes are naturally infected with polyomaviruses closely related to Merkel cell polyomavirus. J. Virol. **85**:916–924.
29. **Liang B, Tikhanovich I, Nasheuer HP, Folk WR.** 2012. Stimulation of BK virus DNA replication by NFI family transcription factors. J. Virol. **86**:3264–3275.
30. **Marchler-Bauer A, et al.** 2011. CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res. **39**:D225–D229.
31. **Mayer M, Dorries K.** 1991. Nucleotide sequence and genome organization of the murine polyomavirus, Kilham strain. Virology **181**:469–480.
32. **Misra V, et al.** 2009. Detection of polyoma and corona viruses in bats of Canada. J. Gen. Virol. **90**:2015–2022.
33. **Orba Y, et al.** 2011. Detection and characterization of a novel polyomavirus in wild rodents. J. Gen. Virol. **92**:789–795.
34. **Pawlita M, Clad A, zur Hausen H.** 1985. Complete DNA sequence of lymphotropic papovavirus: prototype of a new species of the polyomavirus genus. Virology **143**:196–211.
35. **Pipas JM.** 1992. Common and unique features of T antigens encoded by the polyomavirus group. J. Virol. **66**:3979–3985.
36. **Randhawa P, Vats A, Shapiro R.** 2006. The pathobiology of polyomavirus infection in man. Adv. Exp. Med. Biol. **577**:148–159.
37. **Reyes A, et al.** 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature **466**:334–338.
38. **Rice P, Longden I, Bleasby A.** 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. **16**:276–277.
39. **Rott O, Kroger M, Muller H, Hobom G.** 1988. The genome of budgerigar fledgling disease virus, an avian polyomavirus. Virology **165**:74–86.
40. **Schowalter RM, Pastrana DV, Pumphrey KA, Moyer AL, Buck CB.** 2010. Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. Cell Host Microbe **7**:509–515.
41. **Schuurman R, Sol C, van der Noordaa J.** 1990. The complete nucleotide sequence of bovine polyomavirus. J. Gen. Virol. **71**(Pt 8):1723–1735.
42. **Scuda N, et al.** 2011. A novel human polyomavirus closely related to the African green monkey-derived lymphotropic polyomavirus. J. Virol. **85**:4586–4590.
43. **Seif I, Khoury G, Dhar R.** 1979. The genome of human papovavirus BKV. Cell **18**:963–977.
44. **Sunyaev SR, Lugovskoy A, Simon K, Gorelik L.** 2009. Adaptive mutations in the JC virus protein capsid are associated with progressive multifocal leukoencephalopathy (PML). PLoS Genet. **5**:e1000368. doi:10.1371/journal.pgen.1000368.
45. **van der Meijden E, et al.** 2010. Discovery of a new human polyomavirus associated with trichodysplasia spinulosa in an immunocompromized patient. PLoS Pathog. **6**:e1001024. doi:10.1371/journal.ppat.1001024.
46. **Verschoor EJ, et al.** 2008. Molecular characterization of the first polyomavirus from a New World primate: squirrel monkey polyomavirus. J. Gen. Virol. **89**:130–137.
47. **Wellehan JF, Jr, et al.** 2011. Characterization of California sea lion polyomavirus 1: expansion of the known host range of the Polyomaviridae to Carnivora. Infect. Genet. Evol. **11**:987–996.
48. **Yatsunenko T, et al.** 9 May 2012. Human gut microbiome viewed across age and geography. Nature **486**:222–227.

Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice

Andrew L. Goodman[1], George Kallstrom, Jeremiah J. Faith, Alejandro Reyes, Aimee Moore, Gautam Dantas, and Jeffrey I. Gordon[2]

Center for Genome Science and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108

Contributed by Jeffrey I. Gordon, February 24, 2011 (sent for review January 21, 2011)

The proportion of the human gut bacterial community that is recalcitrant to culture remains poorly defined. In this report, we combine high-throughput anaerobic culturing techniques with gnotobiotic animal husbandry and metagenomics to show that the human fecal microbiota consists largely of taxa and predicted functions that are represented in its readily cultured members. When transplanted into gnotobiotic mice, complete and cultured communities exhibit similar colonization dynamics, biogeographical distribution, and responses to dietary perturbations. Moreover, gnotobiotic mice can be used to shape these personalized culture collections to enrich for taxa suited to specific diets. We also demonstrate that thousands of isolates from a single donor can be clonally archived and taxonomically mapped in multiwell format to create personalized microbiota collections. Retrieving components of a microbiota that have coexisted in single donors who have physiologic or disease phenotypes of interest and reuniting them in various combinations in gnotobiotic mice should facilitate preclinical studies designed to determine the degree to which tractable bacterial taxa are able to transmit donor traits or influence host biology.

gut bacterial diversity | nutrient–microbe interactions | translational medicine pipeline for human microbiome

Efforts to dissect the functional interactions between microbial communities and their habitats are complicated by the long-standing observation that, for many of these communities, the great majority of organisms have not been cultured in the laboratory (1). Methodological differences between culture-independent and culture-based approaches have contributed to the challenge of deriving a realistic appreciation of exactly how much discrepancy exists between the culturable components of a microbial ecosystem and total community diversity. Table S1 gives examples of these methodological differences.

The largest microbial community in the human body resides in the gut: Its microbiome contains at least two orders of magnitude more genes than are found in our *Homo sapiens* genome (2). Culture-independent metagenomic studies of the human gut microbiota are identifying microbial taxa and genes correlated with host phenotypes, but mechanistic and experimentally demonstrated links between key community members and specific aspects of host biology are difficult to establish with these methods alone. The goals of the present study were (*i*) to evaluate the representation of readily cultured phylotypes in the human gut microbiota; (*ii*) to profile the dynamics of these cultured communities in a mammalian gut ecosystem; and (*iii*) to determine whether a clonally arrayed, personalized strain collection could be constructed to serve as a foundation for reassembling varying elements of a human's gut microbiota in vitro or in vivo.

Results

To estimate the abundance of readily cultured bacterial phylotypes in the distal human gut, primers were used to amplify variable region 2 (V2) of bacterial 16S ribosomal RNA (rRNA) genes present in eight freshly discarded fecal samples obtained from two healthy, unrelated anonymous donors living in the United States (n = 1 complete sample per donor at t = 1, 2, 3, and 148 d). Amplicons were subjected to multiplex pyrosequencing, and the results were compared with those generated from DNA prepared from ~30,000 colonies cultured from each sample under strict anaerobic conditions for 7 d at 37 °C on a rich gut microbiota medium (GMM) composed of commercially available ingredients ("cultured" samples; details of the culturing technique are given in *SI Materials and Methods*, and a description of GMM is given in Table S2). The resulting 16S rRNA datasets were de-noised to minimize sequencing errors (3, 4), reads were grouped into operational taxonomic units (OTUs) of ≥97% nucleotide sequence identity (ID), and chimeric sequences were removed (*SI Materials and Methods*).

In total, 632 distinct 97%ID OTUs were observed in the complete samples, and 316 were identified in the cultured samples. The average abundance of cultured OTUs in the complete samples was 0.4%, but the average abundance of uncultured OTUs (i.e., those observed in the complete but not the cultured samples) was significantly lower (0.06%; $P < 10^{-6}$ by an unpaired, two-tailed Student's t test, not assuming equal variances) (5).

To evaluate the representation of readily cultured taxa in the human gut microbiota at varying phylogenetic levels, we assigned taxonomic designations to each 97%ID OTU (*SI Materials and Methods*). Each 16S rRNA read from the complete fecal sample was scored as "cultured" if it had a taxonomic assignment that also was identified in the corresponding cultured population. If a 97%ID OTU in the complete sample could not be placed in any known taxonomic group, it was scored as "cultured" only if the same 97%ID OTU was observed in the cultured sample. This analysis indicated that 99% of the 16S rRNA reads derived from the complete fecal samples from either donor belong to phylum-, class- and order-level taxa that are also present in the corresponding cultured sample; 89 ± 4% of the reads are derived from readily cultured family-level taxa, and 70 ± 5% and 56 ± 4% belong to readily cultured genus- and species-level taxa, respectively (Fig. 1*A Upper*). Two alternate taxonomic binning

Author contributions: A.L.G., G.K., J.J.F., G.D., and J.I.G. designed research; A.L.G., G.K., J.J.F., and A.M. performed research; A.R. contributed new reagents/analytic tools; A.L.G., J.J.F., A.R., and J.I.G. analyzed data; and A.L.G. and J.I.G. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The sequences reported in this paper have been deposited in the National Center for Biotechnology Information Sequence Read Archive (accession nos. SRA026269, 026270, and 026271).

Freely available online through the PNAS open access option.

[1]Present address: Section of Microbial Pathogenesis and Microbial Diversity Institute, Yale University, New Haven, CT 06536.

[2]To whom correspondence should be addressed. E-mail: jgordon@wustl.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1102938108/-/DCSupplemental.

methods, the Ribosomal Database Project (RDP) Bayesian classifier v2.0 and an arbitrary %ID cutoff, produced similar results (Fig. S1 *A–F*). Control experiments described in *SI Materials and Methods* indicate that at least 98% of the reads generated from 30,000 pooled colonies are not derived from nongrowing or lysed bacteria (the percentage of reads from the original fecal samples that are derived from dead cells is unknown).

Unsupervised hierarchical clustering of the complete and cultured microbial communities, across the two donors and four time points, revealed that cultured samples cluster separately from those that had not been cultured. Both phylogenetic and nonphylogenetic metrics segregate cultured samples by donor, suggesting that the distinctiveness of each donor's microbiota is preserved in their collections of readily cultured representatives (Fig. S1 *G* and *H*).

We performed shotgun DNA pyrosequencing to determine the degree to which predicted functions contained in the composite genomes of the complete human fecal microbial communities were represented in the corresponding collection of cultured microbes [$n = 4$ samples (one complete and one cultured from each of two donors); 119,842 ± 43,086 high-quality shotgun reads per microbiome; average read length, 366 nt]. On average, 90% of the 2,302 distinct KEGG Orthology (KO) annotations identified in the two uncultured samples also were observed in the cultured communities (Fig. 1*B*, Fig. S2 *A* and *B*, and Table S3). This high percentage of functional representation also was observed when the microbiomes were subjected to alternate annotation schemes. On average, 94% of 929 enzyme commission (EC) assignments and 95% of 216 level 2 KEGG pathways associated with the complete fecal samples also were detected in the cultured communities (Fig. S2 *C–F* and Tables S4 and S5).

To compare the functions represented in the complete and cultured microbiota independent of annotation, we captured antibiotic-resistance genes from their microbiomes in *Escherichia coli* expression vectors. Each *E. coli* library contained ~1 GB of 1.5- to 4-kB fragments of microbiome DNA and was screened against a panel of 15 antibiotics and clinically relevant antibiotic combinations (Table S6). Genes encoding resistance to the same 14 antibiotics were captured in libraries prepared from complete and cultured fecal samples (Fig. S2*G* and Table S7). In one example, a screen for DNA fragments that confer resistance to the aminoglycoside amikacin produced candidate genes from the microbiomes of both complete and cultured microbial communities from Donor 1 but not from Donor 2. Two genes conferring amikacin resistance (either the 16S rRNA methylase *rmtD* or the aminoglycoside phosphotransferase *aphA-3*) were identified in 70% of the DNA fragments captured in selections for this phenotype. Direct culturing of the original fecal communities in the presence of amikacin confirmed that this resistance function is significantly enriched in the readily cultured microbiota of Donor 1 compared with Donor 2 ($P < 0.005$ based on triplicate samples; unpaired, two-tailed Student's t test assuming equal variances) (Fig. S2*H*). PCR analysis showed that many of the amikacin-resistant fecal strains harbor *rmtD* or *aphA-3*. Sequencing the 16S rRNA genes of a subset of these isolates indicated that *rmtD* is present in strains of *Bacteroides uniformis, B. caccae,* and *B. thetaiotaomicron* in this donor (although, notably, not in the sequenced type strains of these species) and that *aphA-3* is contained in the genome of a member of the genus *Desulfotomaculum* (order Clostridiales).

To determine whether a community composed of an individual's readily cultured bacteria exhibits behavior in vivo mirroring that of the individual's complete microbial community, 9-wk-old C57Bl6/J germfree mice were colonized with a complete or cultured microbiota from each of the two human donors ($n = 5$ recipient mice per sample type). A fecal sample from each donor was divided after collection, and one aliquot was gavaged directly into one group of recipient mice; the other aliquot was cultured on GMM plates for 7 d, as above, harvested, and introduced into a second group of recipient animals. Mice were maintained on a standard autoclaved low-fat, plant polysaccharide-rich (LF/PP) chow diet before and 4 wk after gavage. 16S rRNA analysis of fecal samples collected from these mice at the end of the 4-wk period indicated that the complete and the cultured communities were influenced similarly by host selection: 91 ± 3% of the 16S rRNA reads identified from mice colonized with a human donor's complete fecal microbiota were derived from genus-level taxa that also were identified in the mice colonized with the cultured microbial community from the same donor (Fig. 1*A Lower*). Importantly, control experiments demonstrated that the harvested, actively growing colonies gavaged into each germfree mouse are able to prevent nongrowing species that might be present on GMM plates from establishing themselves in recipient animals (details are given in *SI Materials and Methods*).

Luminal material was collected from the proximal, central, and distal portions of the small intestine, cecum, and colon of mice colonized with either the complete or cultured communities from each of the two human donors. V2-directed bacterial 16S rRNA sequencing revealed similar geographic variations in community structures (Fig. 1*C* and Fig. S3 *A–C*).

To determine whether the similarities in community composition in vivo extend to similarities in community gene content, the same fecal DNA samples that had been prepared from these mice after 4 wk on the LF/PP diet for 16S rRNA analyses were subjected to shotgun pyrosequencing ($n = 4$ samples; 87,357 ± 30,710 reads per sample). As with the 16S rRNA analysis, comparisons of the representation of KOs in the various microbiome samples revealed an even greater correlation between complete and cultured communities after they had been subjected to in vivo selection than before their introduction into mice (Fig. 1*B*, Fig. S2 *A–F*, and Tables S3–S5).

Previous comparisons of adult germfree mice with those that harbor gut microbial communities (either conventionally raised animals or formerly germfree animals colonized from mouse or human donors) have shown that the presence of a complete gut microbiota is associated with increased adiposity (6, 7). In comparison, colonization of germfree mice with a single, readily cultured, prominent human gut symbiont (*Bacteroides thetaiotaomicron*) or with a defined community of 12 bacterial species prominently represented in the distal human gut (8) is insufficient to restore epididymal fat pad stores to levels observed in conventionally raised animals (data not shown). To assess whether a complex community of cultured microbes could restore epididymal fat pad weights to the levels associated with complete microbial communities, we evaluated mice colonized with the complete or the cultured fecal communities from the two human donors. All animals displayed significantly greater fat pad to body weight ratios than germfree controls, and no significant difference was observed in adiposity between mice colonized with the donors' complete or cultured microbiota (Fig. S3*D*).

We have reported that mice colonized with a complete human microbiota undergo marked changes in microbial community structure (even after a single day) when shifted from LF/PP chow to a high-fat, high-sugar Western diet (7). To test whether a microbial community consisting only of cultured members recapitulates this behavior in vivo, the four groups of gnotobiotic mice colonized with the complete or cultured microbes from two unrelated human donors were monitored by fecal sampling before, during, and after a 2-wk period when they were placed on the Western diet (samples were collected at days 4, 7, and 14 of the first LF/PP phase, then 1 d before and 3, 7, and 14 d after initiation of the Western diet phase, and finally 1, 3, 8, and 15 d after the return to the LF/PP diet). 16S rRNA-based comparisons of fecal communities were performed using both phylogenetic and nonphylogenetic distance metrics. With either

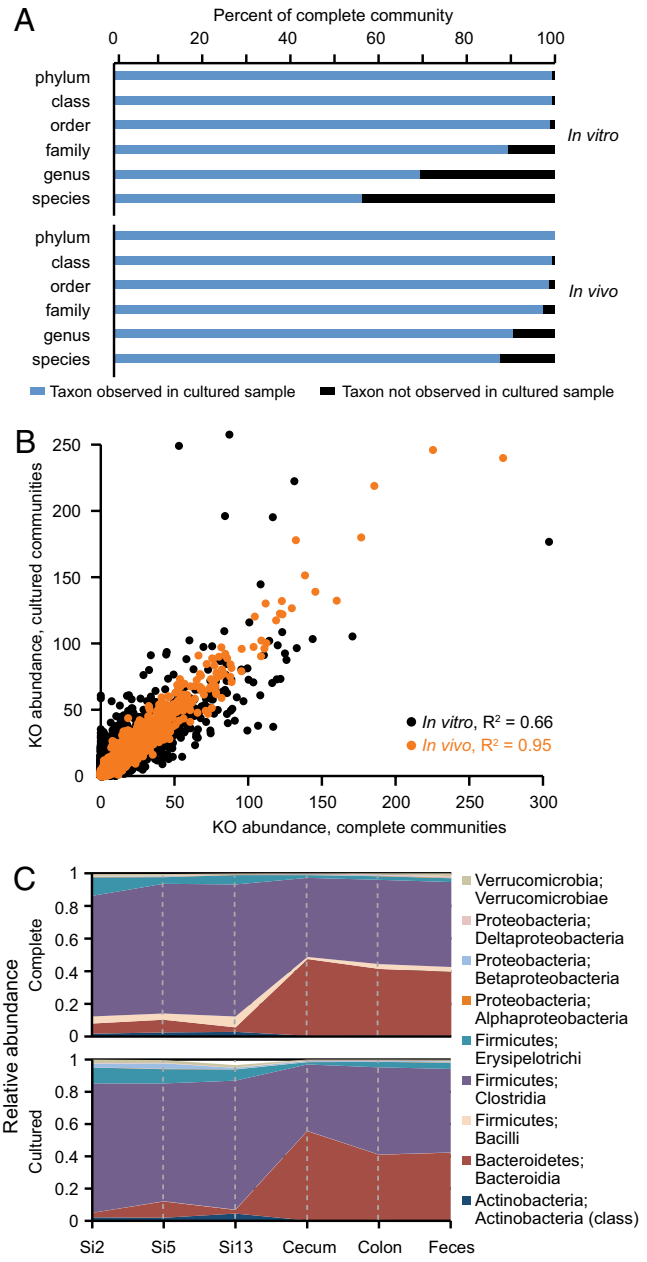


Fig. 1. Comparison of the taxonomic representation of bacterial species and gene content in complete versus cultured human fecal microbial communities before and after their introduction into gnotobiotic mice. (*A*) 16S rRNA sequences from complete microbiota were compared with those identified from microbial communities cultured from the same human donors. At each taxonomic level, the proportion of reads in the complete community belonging to a taxonomic group observed in the cultured sample is shown in blue; the proportion of reads belonging to a taxonomic group not observed in the cultured sample (or lacking taxonomic assignment) is shown in black. Data shown are the average of two unrelated human donors. In vitro samples refer to comparisons between human fecal samples and plated material. In vivo samples refer to comparisons between gnotobiotic mice colonized with a complete human fecal microbiota and mice colonized with the readily cultured microbes from the same human fecal sample. (*B*) Annotated functions identified in the microbiomes of complete and cultured human gut communities. Each point represents a KO designation plotted by relative abundance (average across two donors, per 100,000 sequencing reads). Black points represent KO comparisons between the in vitro samples; orange points represent comparisons between in vivo samples. (*C*) The distribution of taxa and their relative abundance along the length of the intestine are similar in gnotobiotic mice colonized with complete or cultured human gut communities. Relative abundances of class-level taxa at six locations are shown; data represent the average of mice colonized from two unrelated donors. Si, small intestine divided into 16 equal-size segments and sampled at Si-2 (proximal), Si-5 (middle), and Si-13 (distal). PCoA suggests that gut biogeography, rather than donor or culturing, explains the majority (58%) of variance between samples (Fig. S3 *A–C*).

metric, principal coordinates analysis (PCoA) revealed that mice colonized with the complete or cultured samples maintain communities that cluster first by donor (principal coordinate 1; PC1) and that the complete and cultured communities from both donors respond to the diet shift in a similar manner [principal coordinate 2 (PC2); Fig. 2 *A–C* and Fig. S4 *A* and *B*]. Like the transplanted complete microbiota examined here and in previous reports (7), the cultured microbiota responded to this Western diet by increasing the relative proportion of representatives of one class of Firmicutes (the Erysipilotrichi) and decreasing the relative proportion of the Bacteroidia class (Fig. S4*C*). Notably, of the 18 species-level phylotypes significantly affected by diet shift in the mice containing the complete microbiota of both human donors, 14 were detected and demonstrated the same statistically significant response in mice colonized with readily cultured taxa (Fig. 2*D* and Fig. S5).

The fecal microbiomes of LF/PP-fed mice harboring complete or cultured communities from each of the two unrelated donors were compared with microbiomes sampled after these mice consumed the Western diet for 14 d (101,222 ± 24,271 reads per sample). The representation of level 2 KEGG pathway functions was highly concordant on both diets, with one exception: Genes encoding phosphotransferase system (PTS) pathways for carbohydrate transport were significantly overrepresented in Western diet-fed mice harboring complete or cultured communities from either donor (Fig. 2*E* and Fig. S6 *A–D*). Higher-resolution KO-level annotations confirmed that the diet-based PTS pathway enrichment reflected increased representation of multiple carbohydrate transporters (Fig. S6 *E* and *F*). These results emphasize that the similar taxonomic restructuring of complete and cultured communities in response to diet is accompanied by similar changes in community gene content.

Because human gut communities composed of readily cultured members exhibit responses to host diet that mirror those characteristic of a complete microbiota, we explored the possibility that gnotobiotic mice can be used as biological filters to recover collections of readily cultured microbes, obtained from selected human hosts, that are enriched for certain properties [e.g., the ability to prosper (bloom) when exposed to specific foods or food ingredients]. To this end, fecal samples from mice colonized with the complete or corresponding cultured human gut microbial communities from the two unrelated donors and fed the LF/PP or Western diets were collected directly into anaerobic medium and then plated on prereduced GMM plates (Fig. 3*A*). After 7-d incubation, V2-directed 16S rRNA profiling of these plated microbial collections confirmed that these populations of cultured microbes can be reshaped deliberately in vivo and then recovered in vitro (Fig. 3*B* and Fig. S7). On either diet, cultured populations showed significantly greater resemblance to the in vivo communities from mice consuming the same diet than to the in vivo communities from the same mice consuming the alternate diet ($P < 10^{-11}$, unpaired two-tailed Student's *t* test of within-donor distances shown in Fig. 3*B*, assuming equal variances).

We next turned our attention to developing ways to dissect these recovered populations that display properties of interest in vivo. The strict anaerobic techniques used here, compounded with highly diverse colony morphologies across taxa, complicate efforts to pick and isolate individual colonies at a scale sufficient to capture the bacterial diversity represented on the culture plates. Therefore, we used a most probable number (MPN)

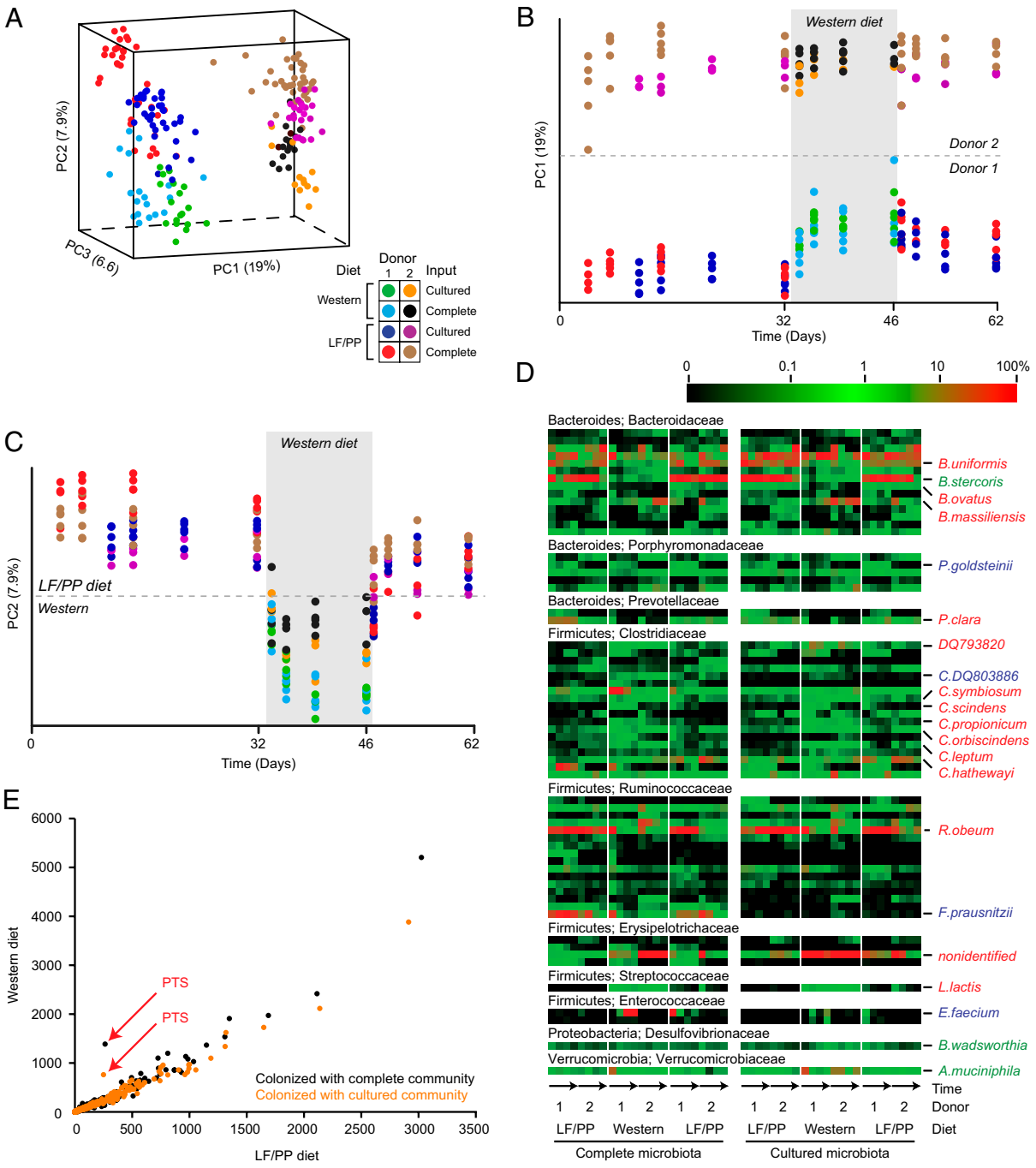


Fig. 2. Human gut microbial communities composed only of cultured members exhibit in vivo dynamics similar to those observed in their complete counterparts. (*A*) PCoA of UniFrac distances between 16S rRNA datasets generated from fecal samples from gnotobiotic mice, colonized with complete or cultured human fecal microbial communities from two unrelated donors and sampled over time. From day 33–46, mice were switched from their standard LF/PP chow to a high-fat, high-sugar Western diet. Time series analysis of community structure as viewed along the first two principal coordinates from *A* shows that interpersonal (donor) differences separated communities on PC1 (*B*), and host diet separated communities on PC2 (*C*). Principal coordinate 3 (PC3) separated samples from mice colonized with complete communities from those colonized with cultured populations (Fig. S4*A*). Nonphylogenetic distance metrics produced similar results (Fig. S4 *D*–*H*). (*D*) Evidence that the community response to diet is driven by readily cultured bacteria and that members of the same taxonomic group manifest distinct responses to diet perturbations. Species-level taxa significantly influenced by diet (Student's *t* test $P \leq 0.01$ after Bonferroni correction; $n = 97$ taxa tested) in either the complete communities (blue names), the cultured communities (green names), or both (red names) are plotted over time (arrows). Each column represents the average relative abundance in fecal samples harvested from three to five individually caged mice that were sampled at various times: (*i*) during the initial LF/PP diet phase; (*ii*) during the subsequent shift to the Western diet; and (*iii*) upon return to LF/PP chow. Members of family-level groups with at least one diet-responsive species are shown (excluding rare species with average abundance <0.1% across each time point). The names of all taxa are shown in Fig. S5. (*E*) The functional gene repertoire in the fecal microbiomes of humanized gnotobiotic mice. Each point represents a KEGG level 2 pathway; the number of hits to each pathway per 100,000 shotgun pyrosequencing reads is plotted for mice consuming LF/PP (*x* axis) or Western (*y* axis) diets. Data represent the averages of mice colonized with microbial communities from two unrelated donors. The results show that the fecal microbiome associated with the Western diet is enriched for genes in pathways associated with carbohydrate phosphotransferase systems (PTS; red arrows) both in mice colonized with complete (uncultured) human gut communities (black points) and mice colonized with communities of readily cultured members (orange points). Donor-specific data and results from alternate annotation schemes are shown in Fig. S6 and Tables S3–S5.

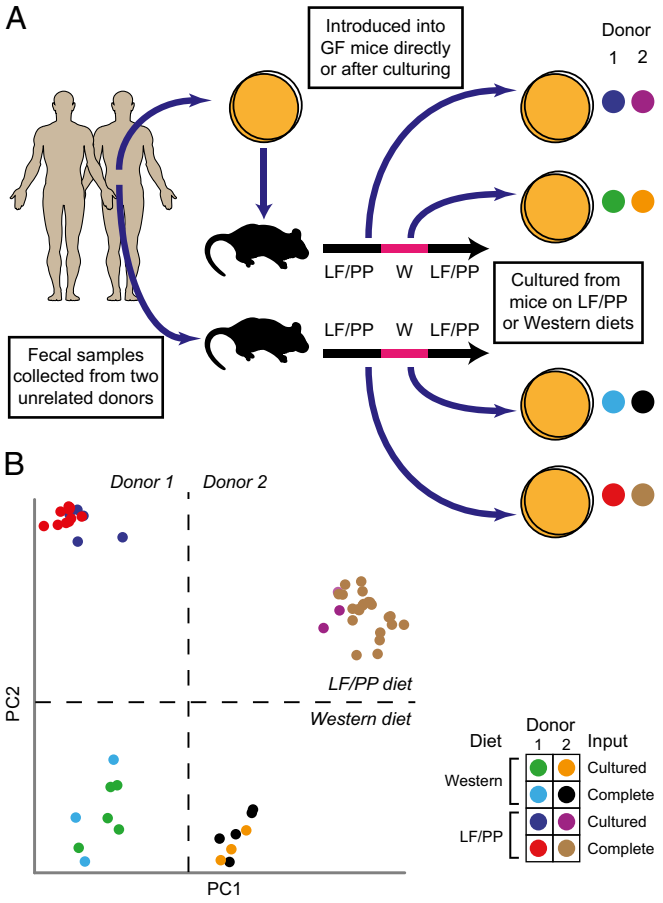


Fig. 3. The community composition of microbes cultured from humanized gnotobiotic mice can be reshaped by altering host diet. (*A*) Culture collections were generated from fecal samples obtained from gnotobiotic mice colonized with complete or cultured gut microbial communities from either of two unrelated human donors and maintained on LF/PP or Western diets. (*B*) PCoA of nonphylogenetic (binary Jaccard) distances between cultured samples indicates that manipulation of host diet can be used to shape the composition of communities recovered in culture from these animals. Analysis of phylogenetic (UniFrac) distances between samples produced similar clustering by donor and host diet (Fig. S7).

technique for creating arrayed species collections in a multiwell format without colony picking. We first empirically determined the dilution point for a fecal sample that yields 70% empty wells (no detectable growth) after inoculation into 384-well trays and 7-d anaerobic incubation. Assuming that the distribution of cells into the wells follows a Poisson distribution, a dilution that leaves 70% of wells empty should yield nonclonal wells (that is, wells that received more than one cell in the inoculum) only 5% of the time; the remainder should be clonal (Fig. S8 *F* and *G*). At this dilution, ten 384-well trays should yield ~1,000 clonal wells. We developed the two-step barcoded pyrosequencing scheme outlined in Fig. 4*A* to assign a 16S rRNA sequence to the isolate(s) present in each turbid well.

We used this approach to create an archived, personalized culture collection of ten 384-well trays from one of the human donors. 16S rRNA sequences could be assigned to more than 99% of growth-positive wells (Table S8). One advantage of clonally arrayed collections is that the effects of 16S rRNA primer bias encountered using DNA templates prepared from complex microbial populations are minimized when wells contain a single taxon. This point is illustrated by the known bias of most commonly used primers against Bifidobacteria spp. (9). Members of this genus were better represented among the set of 16S rRNA genes produced from individual wells than among those observed in complex communities harvested from GMM plates.

After the archived trays had been frozen under anaerobic conditions and stored at −80 °C for 7 mo, recovery of organisms from wells exceeded 60%. Full-length 16S rRNA sequences generated from these recovered strains matched the assignments from the barcoded pyrosequencing data in every case, suggesting that the dilutions did follow a Poisson distribution as predicted. Like 16S rRNA-based community profiling, such collections may miss rare, but important, members of the microbiota; seeding additional 384-well trays with the diluted sample will capture additional phylotypes (Fig. S8*H*). In total, this individual's culture collection contained 1,172 taxonomically defined isolates from four different phyla, seven classes, eight orders, 15 families, 23 genera, and 48 named bacterial species. Novel isolates were encountered at the family-, genus-, and species- levels (Table S8), and 69% of the complete community had a genus-level representative in the arrayed collection (Fig. 4*B*). As a frame of reference, we identified a total of 159 human fecal or gut bacterial species from humans worldwide (including pathogens) in the German Resource Centre for Biological Material (DSMZ) culture collection (*SI Materials and Methods*). As such, personalized microbiota collections can complement those of interna-

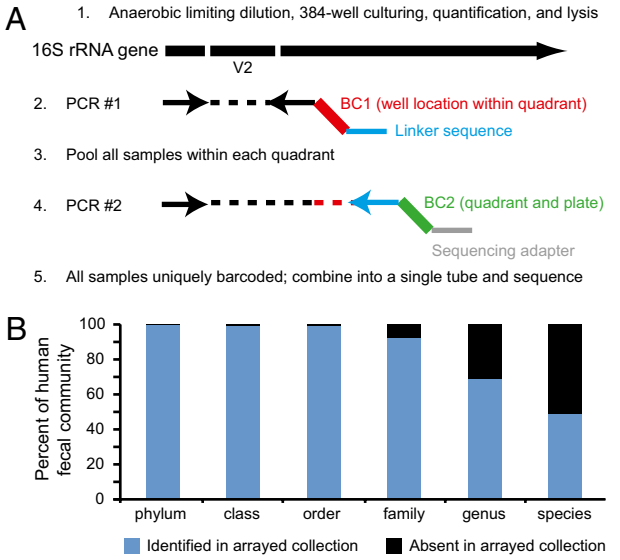


Fig. 4. Personal culture collections archived in a clonally arrayed, taxonomically defined format. (*A*) After limiting dilution of the sample into 384-well trays to the point at which most turbid wells are clonal, a two-step, barcoded pyrosequencing scheme allows each culture well to be associated with its corresponding bacterial 16S rRNA sequence. In the first round of PCR, one of the V2-directed 16S rRNA primers incorporates 1 of 96 error-correcting barcodes (BC1, highlighted in red) that designates the location (row and column) within a quadrant of the 384-well tray where the sample resides. The primer also contains a 12-bp linker (blue). All amplicons generated from all wells in a given quadrant from a single plate then are pooled and subjected to a second round of PCR in which one of primers, which targets the linker sequence, incorporates another error-correcting barcode (BC2; green) that designates the quadrant and plate from which the samples were derived, plus an oligonucleotide (gray) used for 454 pyrosequencing. Amplicons generated from the second round of PCR then are pooled from multiple trays and subjected to multiplex pyrosequencing. This approach allows unambiguous assignment of 16S rRNA reads to well and plate locations using a minimum number of barcodes and primers; e.g., 96 BC1 primers and 96 BC2 primers allow 96^2 (9,216) wells to be analyzed. (*B*) Representation of the original (complete) microbial community in the arrayed strain collection.

tional repositories by capturing strains that coexist in a shared habitat where community structure and host parameters can be measured.

Our ability to capture this level of diversity after MPN dilution in these arrayed collections indicates that it is unlikely that interspecies syntrophic relationships by themselves are sufficient to explain the diversity observed on the GMM agar plates. On the other hand, these personalized arrayed culture collections should help identify obligate syntrophic relationships (e.g., by analyzing the patterns of co-occurrence of taxa in wells harboring more than one phylotype or by comparing arrayed collections in which one set of trays contains a candidate syntroph deliberately added to all wells).

Discussion

We find that it is possible to capture a remarkable proportion of a person's fecal microbiota using straightforward anaerobic culturing conditions and easily obtained reagents. Variations in culturing conditions, including components that are not commercially available (e.g., sterile rumen or human fecal extracts) and other approaches for more closely approximating a native gut habitat, undoubtedly will allow additional members of the human gut microbiota to be cultured in vitro (10, 11). These personal culture collections can be generated from humans representing diverse cultural traditions and various physiologic or pathophysiologic states. A key opportunity is provided when anaerobic culture initiatives are combined with gnotobiotic mouse models, thereby allowing culture collections to be characterized and manipulated in mice with defined (including engineered) genotypes who are fed diets comparable to those of the human donor, or diets with systematically manipulated ingredients. Temporal and spatial studies of these communities can be used to identify readily cultured microbes that thrive in certain physiological and nutritional contexts, creating a discovery pipeline for new probiotics and for preclinical evaluation of the nutritional value of food ingredients. Based on their in vivo responses, clonally archived cultured representatives of a person's microbiota can be selected for complete genome sequencing (including multiple strains of a given species-level phylotype) to identify potential functional variations that exist or evolve within a species occupying a given host's body habitat. Coinciding with the introduction of yet another generation of massively parallel DNA sequencers, this approach should also allow vast scaling of current sequencing efforts directed at characterizing human (gut) microbial genome diversity, evolution, and function. In addition, recovered organisms could also be used as source material for functional metagenomic screens (bio-prospecting). Guided by the results of metagenomic studies of human microbiota donors, components of a personalized collection that have coevolved in a single host can be reunited in varying combinations in gnotobiotic mice, potentially after genome-wide transposon mutagenesis of selected taxa of interest (8), for further mechanistic studies of their interactions and impact on host phenotypes.

Materials and Methods

Culturing of Fecal Microbiota. The Washington University Human Studies Committee reviewed the study design. Freshly discarded fecal samples from two anonymous unrelated human donors were transferred into an anaerobic chamber (Coy Laboratory Products) within 5 min of their collection as described in *SI Materials and Methods* and Tables S2 and S9.

Gnotobiotic Mouse Husbandry. All experiments involving mice were performed with protocols approved by the Washington University Animal Studies Committee. Germfree adult male C57BL/6J mice were maintained in plastic gnotobiotic isolators. Colonization, housing, diet manipulations, and control experiments to evaluate the contribution of uncultured cells to microbial communities from gnotobiotic mice are described in *SI Materials and Methods*.

16S rRNA Sequencing and Analysis. The V2 region of bacterial 16S rRNA genes was subjected to PCR amplification (DNA extraction and PCR protocols are described in *SI Materials and Methods*). Metadata for all 500 samples, including barcodes, are provided in Table S10. All 16S rRNA pyrosequencing datasets have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (accession no. SRA026269).

16S rRNA sequences were filtered to remove low-quality or chimeric sequences, de-noised, and analyzed using QIIME v1.1 (3) with parameters described in *SI Materials and Methods*. To quantify the representation of cultured and uncultured lineages in microbial communities, the presence or absence of each phylum-, class-, order-, family-, genus-, or species-level phylotype assigned to sequences in the complete sample(s) was determined in the cultured sample(s). Taxonomy was assigned with both SILVA-VOTE (*SI Materials and Methods* and Table S11) and RDP Bayesian classifiers. Data were normalized by the abundance of each taxonomic group in the original (uncultured) sample. For analysis of microbial communities from mice, taxonomic groups observed in fewer than two replicate animals were omitted. Protocols for creation and 16S rRNA sequencing of arrayed culture collections are described in *SI Materials and Methods* and Tables S12 and S13.

Shotgun Pyrosequencing. Aliquots (500 ng) of DNA prepared from selected complete and cultured microbiota were sheared and ligated to the default 454 Titanium multiplex identifiers (MIDs; Roche Rapid Library Preparation Method Manual, GS FLX Titanium Series, October 2009) and sequenced using 454 Titanium pyrosequencing chemistry. After filtering of low-quality or host DNA sequences, reads were queried against the KEGG KO database (v52) using parameters described in *SI Materials and Methods*; metadata for each sample are provided in Table S14, and all shotgun pyrosequencing datasets are available in the NCBI SRA under accession no. SRA026270. Procedures for bio-prospecting for antibiotic resistance genes in complete and cultured microbial communities are described in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank B. Muegge, S. Wagoner, D. O'Donnell, M. Karlsson, M. Gonzalez, A. Keel, T. Ellison, B. Wang, J. Symington, V. Wagner, M. Dunne, and R. Knight for assistance and comments. This work was supported by National Institutes of Health Grants DK30292, DK70977, and DK78669 (to J.I.G), F32AI078628 and K01DK089121 (to A.L.G.), and T32-HD043010 (to A.M.) and by the Crohn's and Colitis Foundation of America.

1. Razumov AS (1932) *Mikrobiologija* 1:131–146.
2. Qin J, et al.; MetaHIT Consortium (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59–65.
3. Caporaso JG, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335–336.
4. Reeder J, Knight R (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat Methods* 7:668–669.
5. Walker AW, et al. (2010) Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J* 5:220–230.
6. Bäckhed F, et al. (2004) The gut microbiota as an environmental factor that regulates fat storage. *Proc Natl Acad Sci USA* 101:15718–15723.
7. Turnbaugh PJ, et al. (2009) The effect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice. *Sci Transl Med* 1:6–14.
8. Goodman AL, et al. (2009) Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* 6:279–289.
9. Hill JE, et al. (2010) Improvement of the representation of bifidobacteria in fecal microbiota metagenomic libraries by application of the cpn60 universal primer cocktail. *Appl Environ Microbiol* 76:4550–4552.
10. Nichols D, et al. (2010) Use of ichip for high-throughput in situ cultivation of "uncultivable" microbial species. *Appl Environ Microbiol* 76:2445–2450.
11. Bollmann A, Lewis K, Epstein SS (2007) Incubation of environmental samples in a diffusion chamber increases the diversity of recovered isolates. *Appl Environ Microbiol* 73:6386–6390.

GOS scaffolds, finding 46 MpnS and 20 HepD homologs, using a protein basic local alignment search tool (BLASTP) cutoff value of 10^{-10} (table S2). No HppE homologs were observed. None of the HepD homologs were identified when *N. maritimus* MpnS was used as the query sequence; likewise, none of the MpnS homologs were identified when HepD was used as a query. Thus, BLASTP clearly differentiates between the two homologous groups, supporting the assignment of the recovered sequences as MpnS and HepD proteins, respectively. To independently support these functional assignments, we constructed maximum-likelihood phylogenetic trees including biochemically validated MpnS, HepD, and HppE proteins (Fig. 3A and fig. S9). We also used a hierarchical clustering method to examine all putative and validated MpnS, HepD, and HppE proteins (fig. S10). In both cases, robust support for the functional assignments was obtained. Thus, we conclude that the recovered GOS MpnS homologs are likely to be methylphosphonate synthases.

Additional support for the function of the MpnS homologs was revealed by analysis of neighboring genes found in GOS DNA scaffolds (Fig. 3B and table S3). Many of the nearby open reading frames are homologous to those found in the *N. maritimus* gene cluster, including the phosphonate biosynthetic genes *ppm*, *ppd*, and *pdh*, as well as homologs of the sulfatases and nucleotidyl transferase genes, suggesting that the GOS scaffolds encode genes for the synthesis of similar MPn esters. Several other genes found on the scaffolds provide evidence for the identity of the organisms in which they are found. One of the scaffolds includes a 23*S* ribosomal RNA gene that can be confidently placed within the SAR11 clade between *Pelagibacter* species (fig. S11), whereas two of the *manC* genes are nearly identical to ones found in *Pelagibacter sp.* HTCC7211. Although the *mpnS* gene is absent in sequenced *Pelagibacter* genomes, these data strongly support the conclusion that some members of this genus have the capacity to synthesize MPn.

Relatives of *Nitrosopumilus* and *Pelagibacter* are among the most abundant organisms in the sea, with global populations estimated at 10^{28} for both ammonia-oxidizing Thaumarchaeota (*14*) and members of the SAR11 clade (*22*). Thus, the observation of *mpnS* in some members of these genera is consistent with the idea that MPn synthesis is prevalent in marine systems. To provide direct support for this notion, we measured the abundance of the *mpnS* gene relative to the abundance of typical single-copy genes as previously described (*23*). We also quantified the occurrence of the *ppm* gene to provide an estimate of the relative occurrence of phosphonate synthesis in general (table S4). Based on these data, we estimate that ~16% of marine microbes are capable of phosphonate biosynthesis, whereas 0.6% have the capacity to synthesize MPn. Because the GOS samples are confined to the upper few meters of the ocean, extrapolation of this analysis to the deeper ocean should be viewed with some skepticism. Nevertheless, the upper 200 m of the world's oceans are thought to contain ~3.6×10^{28} microbial cells, with an average generation time of ~2 weeks (*24*). Thus, even with the relatively modest abundance of MPn biosynthesis suggested by our data, it seems quite possible that these cells could provide sufficient amounts of MPn precursor to account for the observed methane production in the aerobic ocean via the C-P lyase–dependent scenario suggested by Karl *et al*. (*2*).

References and Notes

1. J. E. Rogers, W. B. Whitman, Eds., *Microbial Production and Consumption of Greenhouse Gases: Methane, Nitrogen Oxides, and Halomethanes* (American Society for Microbiology, Washington, DC, 1991).
2. D. M. Karl *et al*., *Nat. Geosci.* **1**, 473 (2008).
3. W. S. Reeburgh, *Chem. Rev.* **107**, 486 (2007).
4. C. G. Daughton, A. M. Cook, M. Alexander, *FEMS Microbiol. Lett.* **5**, 91 (1979).
5. A. Martinez, G. W. Tyson, E. F. Delong, *Environ. Microbiol.* **12**, 222 (2010).
6. I. N. Ilikchyan, R. M. L. McKay, J. P. Zehr, S. T. Dyhrman, G. S. Bullerjahn, *Environ. Microbiol.* **11**, 1314 (2009).
7. L. L. Clark, E. D. Ingall, R. Benner, *Am. J. Sci.* **299**, 724 (1999).
8. L. L. Clark, E. D. Ingall, R. Benner, *Nature* **393**, 426 (1998).
9. W. W. Metcalf, W. A. van der Donk, *Annu. Rev. Biochem.* **78**, 65 (2009).
10. S. A. Borisova, B. T. Circello, J. K. Zhang, W. A. van der Donk, W. W. Metcalf, *Chem. Biol.* **17**, 28 (2010).
11. J. A. Blodgett, J. K. Zhang, W. W. Metcalf, *Antimicrob. Agents and Ch.* **49**, 230 (2005).
12. A. C. Eliot *et al*., *Chem. Biol.* **15**, 765 (2008).
13. B. T. Circello, A. C. Eliot, J. H. Lee, W. A. van der Donk, W. W. Metcalf, *Chem. Biol.* **17**, 402 (2010).
14. M. B. Karner, E. F. DeLong, D. M. Karl, *Nature* **409**, 507 (2001).
15. M. Könneke *et al*., *Nature* **437**, 543 (2005).
16. Z. Shao *et al*., *J. Biol. Chem.* **283**, 23161 (2008).
17. H. M. Seidel, S. Freeman, H. Seto, J. R. Knowles, *Nature* **335**, 457 (1988).
18. Materials and methods are available as supplementary materials on *Science* Online.
19. J. C. Tebby, Ed., *CRC Handbook of Phosphorus-31 Nuclear Magnetic Resonance Data* (CRC Press, Boca Raton, FL, 1991).
20. R. L. Hilderbrand, Ed., *The Role of Phosphonates in Living Systems* (CRC Press, Boca Raton, FL, 1983).
21. S. Yooseph *et al*., *PLoS Biol.* **5**, e16 (2007).
22. R. M. Morris *et al*., *Nature* **420**, 806 (2002).
23. E. C. Howard, S. Sun, E. J. Biers, M. A. Moran, *Environ. Microbiol.* **10**, 2397 (2008).
24. W. B. Whitman, D. C. Coleman, W. J. Wiebe, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6578 (1998).

Acknowledgments: This work was supported by the NIH (grants GM PO1 GM077596 and F32 GM095024), the Howard Hughes Medical Institute (HHMI), and NSF (grants MCB-0604448, OCE-1046017, and MCB-0920741). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute of General Medical Sciences, NIH, NSF, or HHMI. The authors thank L. Zhu (University of Illinois at Urbana–Champaign) for valuable help with NMR experiments.

Supplementary Materials

www.sciencemag.org/cgi/content/full/337/6098/1104/DC1
Materials and Methods
Figs. S1 to S11
Tables S1 to S4
References (*25*–*39*)

31 January 2012; accepted 10 July 2012
10.1126/science.1219875

The Shared Antibiotic Resistome of Soil Bacteria and Human Pathogens

Kevin J. Forsberg,[1]* Alejandro Reyes,[1]* Bin Wang,[1,2] Elizabeth M. Selleck,[3] Morten O. A. Sommer,[4,5]† Gautam Dantas[1,2]†

Soil microbiota represent one of the ancient evolutionary origins of antibiotic resistance and have been proposed as a reservoir of resistance genes available for exchange with clinical pathogens. Using a high-throughput functional metagenomic approach in conjunction with a pipeline for the de novo assembly of short-read sequence data from functional selections (termed PARFuMS), we provide evidence for recent exchange of antibiotic resistance genes between environmental bacteria and clinical pathogens. We describe multidrug-resistant soil bacteria containing resistance cassettes against five classes of antibiotics (β-lactams, aminoglycosides, amphenicols, sulfonamides, and tetracyclines) that have perfect nucleotide identity to genes from diverse human pathogens. This identity encompasses noncoding regions as well as multiple mobilization sequences, offering not only evidence of lateral exchange but also a mechanism by which antibiotic resistance disseminates.

The continued evolution and widespread dissemination of antibiotic resistance genes in human pathogens is a preeminent clinical challenge (*1*). Environmental reservoirs have long been implicated as a source of resistance found in human pathogens (*2*). However, apart from certain opportunistic bacterial pathogens, among which the same species can be found in the environment or infecting humans (*3*), examples of resistance genes from environmental bacteria with high identity to those of pathogens are rare (*4*, *5*). The two documented examples are of *Kluyvera* and *Shewanella* isolates, which are found free-living in environmental settings (*5*, *6*) yet have resistance genes (CTX-M β-lactamase and qnrA genes, respectively) with high identity (100% identity in clinical *Kluyvera* isolates) to those of pathogens (*4*, *5*). The limited examples of resistance genes shared between environmental microbes and human pathogens raise questions regarding the clinical impact of environmental resistance. For instance, whether shared resistance

is confined to genes of particular mechanisms (such as enzymatic β-lactam cleavage) or applies to many genes with diverse mechanisms of resistance is unknown. Additionally, whether a single horizontal gene transfer (HGT) event between environment and clinic can result in the de novo acquisition of a multidrug-resistant phenotype is unclear. The two previous reports of high-identity resistance genes shared between environmental and pathogenic bacteria did not find evidence of colocalized resistance genes or of syntenic mobilization elements (*4*, *5*), hallmarks of transferable multidrug resistance (*7*, *8*). Determining the clinical impact of environmental resistance requires a deeper profiling of environmental reservoirs for the organisms and genotypes most likely to exchange resistance with human pathogens.

Soil, one of the largest and most diverse microbial habitats on earth, is increasingly recognized as a vast repository of antibiotic resistance genes (*9*–*13*). Not only does soil come into direct contact with antibiotics used extensively in rearing livestock (*14*) and plant agriculture (*15*), but it is also a natural habitat for the Actinomycete genus *Streptomyces*, whose species account for the majority of all naturally produced antibiotics (*16*). Despite numerous studies demonstrating that soil contains resistance genes with biochemical mechanisms similar to those in common pathogens (*3*, *11*–*13*), the sequence identities of these genes diverge from those of pathogens (*17*), providing little evidence that these resistomes have more than an evolutionary relationship. Therefore, whether soil has recently contributed to or acquired resistance genes from the pathogenic resistome remains an open question, and accordingly, the role of soil in the current global exchange of antibiotic resistance remains poorly defined.

To examine the capacity of nonpathogenic, soil-dwelling organisms to exchange antibiotic resistance with human pathogens, we sought to select for organisms prone to this exchange. Because many major clinical pathogens are Proteobacterial (*18*), we cultured multidrug-resistant Proteobacteria from the soil (*19*), with the aim of enriching for resistance genes shared between soil and human pathogens. We interrogated the resistome of the resulting culture collection using functional metagenomic selections, which are ideally suited to characterize acquirable resistance because they identify any gene sufficient to confer resistance to a new host (such as a pathogen) (*20*). To facilitate the rapid and efficient functional characterization of metagenomic libraries, we developed a massively parallel, multiplexed functional selection platform that enables simultaneous sequencing, de novo assembly, and functional annotation of hundreds of resistance fragments from many independent selections (termed PARFuMS: Parallel Annotation and Re-assembly of Functional Metagenomic Selections) (fig. S1) (*19*).

We applied PARFuMS to a collection of 95 soil-derived cultures ("AB95"), representing bacteria with high-level resistance to various antibiotics. Cultures were obtained from 11 U.S. soils (table S1), passaged serially through minimal and rich media containing one of 18 antibiotics at 1000 mg/L (tables S2 and S3) (*21*), and subjected to 16*S* ribosomal DNA (rDNA) profiling (*19*). We confirmed that the culture collection was enriched for Proteobacteria and dominated by traditional soil-dwelling organisms (such as *Pseudomonas* and *Pandoraea*) (fig. S2). Equal proportions of the 95 cultures were pooled, and bulk genomic DNA was extracted. One- to 3-kb fragments of this metagenomic DNA were cloned into an expression vector and transformed into *Escherichia coli*. The resulting 2.57-Gb metagenomic library was selected on solid culture medium containing 1 of 12 antibiotics representing amino acid derivatives, aminoglycosides, amphenicols, β-lactams, and tetracyclines, at concentrations to which the host-strain was susceptible (table S4). Resistance was detected against all 12 antibiotics, and resistance-conferring fragments were sequenced, assembled, and annotated by using PARFuMS, yielding 161 contigs (N50 > 1.7 kb). Of the 252 open reading frames (ORFs) identified, 110 (44%) could confidently be annotated as antibiotic resistance genes (by similarity to a known resistance gene, which was consistent with functional selection), whereas another 62 (25%) were categorized as resistance-related (Fig. 1, A to C, and table S5).

Of the 110 resistance genes, 18 had 100% amino acid identity to entries in GenBank, and another 32 were highly similar (≥90% identity). Thus, although we recovered several genes previously identified, most of the resistance genes discovered (54%) were formerly unknown (Fig. 1D). For instance, we identified a gene conferring D-cycloserine resistance from an AB95 isolate (most closely related to *Serratia ficara*) for which sequence alone could not predict resistance function (*19*). The ORF was 92% identical to a protein of unknown function from *Serratia proteamaculans* 568 (CP000826) (Fig. 2A) and enabled *E. coli* to tolerate high concentrations of D-cycloserine (128 μg/mL) (Fig. 2B). The D-cycloserine resistance protein had low-level identity to a drug/metabolite transporter (46% identity over 91% of the sequence; YP_001583420), indicating that the gene may have efflux-related function, which is consistent with known mechanisms of D-cycloserine tolerance (*22*).

Of the 110 AB95 resistance genes, 55 were β-lactamases. The majority of these sequences clustered with class C β-lactamases and were dissimilar to entries currently in GenBank (fig. S3), which is a common result from metagenomic experiments (*11*, *20*, *23*). AB95 β-lactamases were highly divergent from those of the antibiotic-

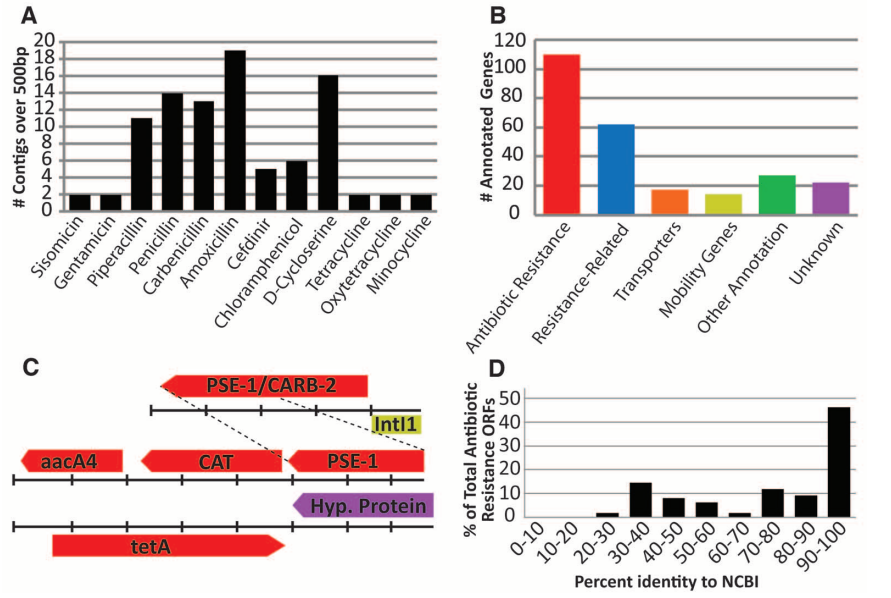


Fig. 1. Functional selection of the AB95 soil metagenomic library with 12 antibiotics (*19*). (**A**) Bar chart depicting the number of distinct contigs over 500 base pairs (bp) recovered from selection with each of the 12 antibiotics. (**B**) Functional classification of ORFs predicted by PARFuMS, across all selections. (**C**) Three representative metagenomic fragments; colors match catergorizations depicted in (B). The distance between tick marks is 300 bp, and dashed lines indicate common sequence on two distinct fragments. (**D**) Amino acid identity between antibiotic-resistance ORFs and the closest hit from GenBank, across all selections.

[1]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108, USA. [2]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63108, USA. [3]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110, USA. [4]Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark. [5]Novo Nordisk Foundation Center for Biosustainability, DK-2900 Hørsholm, Denmark.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: dantas@wustl.edu (G.D.); msom@bio.dtu.dk (M.O.A.S.)

producing *Streptomyces*, indicating ancient evolutionary relationships (fig. S3 and table S6). However, several β-lactamases with >99% identity to sequences from both soil and enteric organisms were recovered (fig. S3).

We identified 16 sequences, from 10 selections, with 100% nucleotide identity to antibiotic resistance genes previously sequenced from clinical isolates of many common human pathogens (Table 1). A bacterium was considered pathogenic only if it was isolated from an infection in a diseased human host. The 16 sequences represent seven different genes, conferring resistance to five classes of antibiotics (β-lactams, aminoglycosides, amphenicols, sulfonamides, and tetracyclines) (Table 1). We discovered multiple examples of syntenic, soil-derived resistance genes shared with many common pathogens. For example, a chloramphenicol-acetyltransferase with 99.7% identity to *K. pneumoniae* clinical isolates was adjacent to both an aminoglycoside-acetyltransferase and a β-lactamase identical to genes found in many pathogens (JX009248). Additionally, an insert from two selections contained *aadB* (an aminoglycoside-adenyltransferase) adjacent to *qacEΔ1* (an efflux pump conferring antiseptic resistance) and *sul1* (a dihydropteroate synthase conferring sulfonamide resistance) in a class 1 integron-like structure (JX009286). All three genes and much of the surrounding integron (>2 kb) are 100% identical to numerous clinical pathogens. The seven soil-derived resistance genes (Table 1) are globally distributed amongst human pathogens: Clinical isolates from many countries and all major continents contain genes with perfect nucleotide identity to genes from this set (fig. S4).

To identify soil isolates from the AB95 culture collection harboring the aforementioned resistance genes, we performed polymerase chain reactions using primers specific to the boundaries of the predicted ORFs (*19*). We identified two organisms isolated from farmland soil containing six of the resistance genes identical to pathogens, as well as two additional genes with over 99% identity to those in pathogens (tables S7 and S8) (*19*). We confirmed that seven genes were present in an organism most closely related to *Pseudomonas sp.* K94.23 [a member of the *P. fluorescens* complex (*24*)], three originated from a strain most similar to *Ochrobactrum anthropi*, and two were in both genomes (*19*). *P. fluorescens* is not believed to cause human infection (*25*), and there are only limited examples of *O. anthropi* subgroups known to infect humans (*26*). Rather, these two organisms are predominantly found in environmental settings (*25*, *27*). The substantial phylogenetic divergence between these traditionally nonpathogenic soil isolates and numerous human pathogens (table S9) contrasts with the 100% identity of numerous resistance genes found in both groups, confirming that these genes moved between species via HGT.

Three ORFs from *O. anthropi* and *P. fluorescens*, conferring β-lactam, aminoglycoside, and amphenicol resistance and representing one gene shared by both organisms and one specific to each, were cloned from their genomic DNA, expressed in *E. coli*, and verified for resistance to seven antibiotics (*19*). In all cases, the ORFs conferred resistance at concentrations 16-fold greater than that of an empty-vector control and enabled growth in a minimum of 128 μg/mL (and up to 2048 μg/mL) of antibiotic (Table 2). These results mirror the minimum inhibitory concentrations of the source soil strains (Table 2), demonstrating that the resistance genes retain functionality even when removed from all native genomic context, emphasizing their broad host-range compatibility.

Perfect nucleotide identity between full-length resistance genes from distinct species implies that recent HGT has occurred between these organisms (*28*)—evidence that has not been previously reported between a nonpathogenic soil-dwelling organism and human pathogens. The seven resistance genes we discovered encompass all major mechanistic classes of antibiotic resistance (*29*) and are identical to genes found in diverse human pathogens, representing both Gram-negative and -positive bacteria. Moreover, for five of the soil-derived contigs that share resistance genes with pathogens, at least 80% of the contig is identical to sequence from a clinical isolate, encompassing coding and noncoding regions alike (the maximum span of identity is 2.28 kb) (table S10). In support of recent mobilization, we found 11 distinct sequences annotated as either an integrase or transposase from six antibiotic selections. Two *intI1* integrases were adjacent to resistance genes from both our organisms and pathogens, indicating a shared mechanism of HGT between soil and pathogenic bacteria. Four of the contigs assembled from our set are over 99% identical to a large span of sequence, found in numerous pathogens, that contains a high density of resistance genes and is flanked by multiple mobility elements (Fig. 3). This cluster of resistance genes exhibits extensive modularity; many combinations of the individual resistance elements are present in a multitude of clinical pathogens.

The closest homologs to each AB95 resistance gene include pathogenic resistance genes that are chromosomal as well as plasmid-borne, implying a diverse genetic organization of these

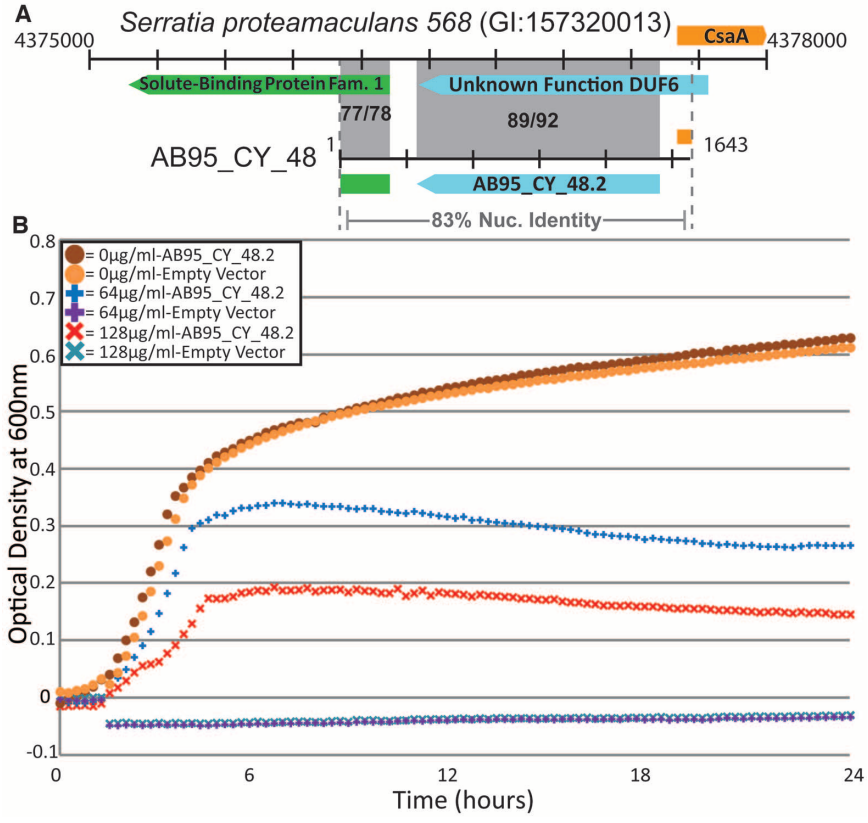


Fig. 2. A gene conferring resistance to D-cycloserine was captured for which sequence was unable to predict resistance function. (**A**) Resistance-conferring fragment AB95_CY_48 compared with its closest hit from the National Center for Biotechnology Information (NCBI) nucleotide collection. ORFs of the same color indicate homologous sequence; both nucleotide and amino acid percent identities are given in shaded regions (nucleotide/amino acid). Base-pair coordinates flank sequences, and the distance between each tick mark is 300 bp. (**B**) Measurements of absorbance at 600 nm, taken every 15 min, depict growth of *E. coli*, containing either AB95_CY_48.2 or an empty vector at clinically relevant concentrations of D-cycloserine. Measurements are corrected for background absorbance from media-only controls and are averages of three trials (*19*).

genes. Four of the pathogen-identical genes from *P. fluorescens*, conferring resistance to the aminoglycosides, tetracyclines, amphenicols, and sulfonamides, were identified in a plasmid preparation, implicating conjugation or transformation as potential mechanisms of HGT (table S11). Additionally, we discovered nine integrases/transposases proximal to resistance genes not yet identified in pathogens, indicating that additional resistance genes from these soil bacteria may be available for HGT with pathogens.

Table 1. Nonredundant antibiotic resistance genes with 100% identity to known human pathogens.

Gene name	GenBank ID	Number of selections*	Antibiotic class	Annotation [mechanism]	Pathogens hit (GI number)
AB95_PI_68.1	JX009363	4	β-lactam	blaP1 [enzymatic degradation]	*A. baumannii* (94960156), *K. pneumoniae* (114147191), *P. aeruginosa* (117321883), *S. typhimurium* (12719011), *P. mirabilis* (157674381)†
AB95_CH_13.1	JX009364	1	Amphenicol	Chloramphenicol efflux [efflux]	*A. baumannii* (169147133), *P. aeruginosa* (260677483)
AB95_TE_2.2	JX009366	3	Tetracycline	tetA(G) [efflux]	*A. baumannii* (169147133), *S. typhimurium* (12719011)
AB95_TE_1.1	JX009365	3	Tetracycline	tetA [efflux]	A. baumannii (169147133), *E. coli* (312949035), *K. pneumoniae* (290792160), *S. typhimurium* (37962716)†
AB95_GE_3.3	JX009367 JX009373	2	Aminoglycoside	aadB [covalent modification]	*E. cloacae* (71361871), *K. pneumoniae* (206731403), *P. aeruginosa* (37955767), *S. typhimurium* (17383994)†
AB95_GE_3.1	JX009368 JX009374	2	Sulfonamide	sul1 [target modification]	*C. diptheriae* (323714042) *E. cloacae* (71361871), *K. pneumoniae* (206731403), *P. aeruginosa* (37955767), *S. typhimurium* (17383994), *Yersinia pestis* (165913934)†
AB95_CH_21.1	JX009369	1	Aminoglycoside	aacA4 [covalent modification]	*A. baumannii* (164449567), *K. pneumoniae* (238865601), *P. aeruginosa* (219872982), *S. typhi* (34014739)†

*Number of selections in which the entirety of a given gene was captured. †More pathogens exist for which 100% nucleotide identity was observed than listed

Given the extensive interspecific transfer of antibiotic resistance, and our data suggesting recent exchange between soil bacteria and clinical pathogens, we sought to identify routes of dissemination between these reservoirs. Possibilities include direct exchange between soil microbes and human pathogens or indirect transfer via reservoirs such as the human intestinal microbiota. Many resistance genes from the intestinal microbiota are identical to those found in diverse human pathogens (*20*), and accordingly, we compared the AB95 resistance genes with a set of resistance genes from cultured intestinal isolates (*20*), a collection of 128 representative gut organisms (table S12), and resistance genes from fecal metagenomes (*19*, *20*). Most AB95 resistance genes were dissimilar to sequences from any intestinal data set, with the average amino acid identity ranging from 30.2 to 45.5% (fig. S5). However, the two cultured data sets contained perfect matches to distinct AB95 resistance genes (table S13). One such AB95 gene (JX009365) was not only identical to *tetA* from an intestinal isolate, but also to numerous pathogens, including *A. baumannii*, *E. coli*, *K. pneumoniae*, and *S. typhimurium*, indicating potential interconnections between the resistomes of the human gastrointestinal tract, soil, and clinical pathogens.

The exchange of resistance between soil and pathogens emphasizes the clinical importance of the soil resistome, regardless of whether resistance genes are moving from soil to the clinic, or vice versa. Transmission from soil to clinic establishes soil as a direct source of pathogenic resistance genes. Movement of resistance from pathogens into soil means pathogens can transfer resistance

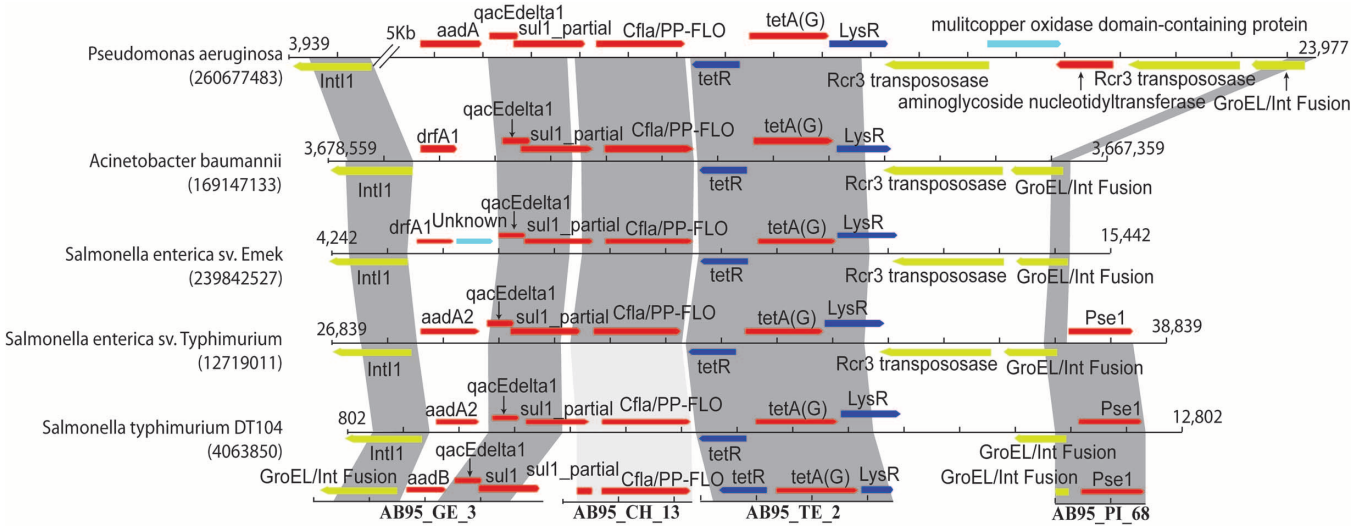


Fig. 3. Comparison of four AB95-derived resistance fragments to five human pathogenic isolates. The four fragments are depicted along the bottom, and shading indicates high nucleotide identity between the fragments and pathogens (NCBI GenInfo numbers identify each pathogenic isolate). Dark gray shading indicates >99% identity; light gray shading indicates ~88% identity. Base-pair coordinates flank pathogenic sequences, and the distance between each tick mark is 800 bp. Red ORFs represent resistance genes, yellow represents mobility elements, dark blue represents resistance-associated regulatory elements, and light blue represents other functions.

Table 2. Minimum inhibitory concentrations of various antibiotics toward both multidrug resistant soil isolates and *E. coli* clones expressing selected resistance genes (all concentrations are μg/mL). AX, amoxicillin; CA, carbenicillin; PE, penicillin; PI, piperacillin; CF, cefdinir; CH, chloramphenicol; SI, sisomicin; GE, gentamicin; MN, minocycline; OX, oxytetracycline; TE, tetracycline; and blank cells indicate inhibitory concentrations were not determined.

	AX	CA	PE	PI	CF	CH	SI	GE	MN	OX	TE
Ochrobactrum soil isolate	>2048	>2048	>2048	>2048	<16	512	512	512	<4	256	64
Pseudomonas soil isolate	>2048	>2048	>2048	>2048	>1024	1024	>1024	>1024	8	128	32
AB95_PI_68.1	>2048	>2048	2048	2048							
AB95_CH_33.1						256					
AB95_GE_3.3							>1024	>1024			
E. coli + empty vector control	<16	<32	64	16	<8	<8	<8	<8	<8	8	4

to soil organisms, of which many can cause nosocomial infection and may emerge as pathogens, akin to the rise of *A. baumannii*.

Powered by PARFuMS, a method for characterizing functional selections at <1% of the cost of traditional approaches (*19*), we describe antibiotic resistance genes found in nonpathogenic soil-dwelling bacteria and of all major mechanistic classes (*29*) with perfect nucleotide identity to many diverse human pathogens. We also show that multiple resistance genes are colocalized within long stretches of perfect nucleotide identity and are flanked by mobile DNA elements. These findings not only provide evidence for recent HGT of multidrug resistance cassettes between soil and clinic, but also a mechanism through which this exchange may have occurred.

The *Ochrobactrum* and *Pseudomonas* isolates originated from farmland soils fertilized with manure from antibiotic-treated livestock. However, our current study design did not enable a statistically significant association of pathogen-identical resistance genes to specific soils. Rather, our results highlight the fact that soil and pathogenic resistomes are not distinct, emphasizing the clinical importance of environmental resistance. Our new method provides the increased throughput required to power future studies to identify soil (*11*), aquatic (*5*), and other (*20*) environments prone to resistance exchange with human pathogens and to understand how specific anthropogenic practices influence the likelihood of this dissemination (*3*, *23*).

References and Notes

1. C. A. Arias, B. E. Murray, *N. Engl. J. Med.* **360**, 439 (2009).
2. R. Benveniste, J. Davies, *Proc. Natl. Acad. Sci. U.S.A.* **70**, 2276 (1973).
3. J. L. Martínez, *Science* **321**, 365 (2008).
4. L. Poirel, P. Kämpfer, P. Nordmann, *Antimicrob. Agents Chemother.* **46**, 4038 (2002).
5. L. Poirel, J. M. Rodriguez-Martinez, H. Mammeri, A. Liard, P. Nordmann, *Antimicrob. Agents Chemother.* **49**, 3523 (2005).
6. J. J. Farmer 3rd *et al.*, *J. Clin. Microbiol.* **13**, 919 (1981).
7. T. Stalder, O. Barraud, M. Casellas, C. Dagot, M. C. Ploy, *Front. Microbiol.* **3**, 119 (2012).
8. B. M. Marshall, S. B. Levy, *Clin. Microbiol. Rev.* **24**, 718 (2011).
9. V. M. D'Costa *et al.*, *Nature* **477**, 457 (2011).
10. V. M. D'Costa, K. M. McGrann, D. W. Hughes, G. D. Wright, *Science* **311**, 374 (2006).
11. H. K. Allen, L. A. Moe, J. Rodbumrer, A. Gaarder, J. Handelsman, *ISME J.* **3**, 243 (2009).
12. J. J. Donato *et al.*, *Appl. Environ. Microbiol.* **76**, 4396 (2010).
13. R. I. Aminov, R. I. Mackie, *FEMS Microbiol. Lett.* **271**, 147 (2007).
14. H. Heuer, H. Schmitt, K. Smalla, *Curr. Opin. Microbiol.* **14**, 236 (2011).
15. P. S. McManus, V. O. Stockwell, G. W. Sundin, A. L. Jones, *Annu. Rev. Phytopathol.* **40**, 443 (2002).
16. T. Kieser, M. J. Bibb, M. J. Buttner, K. F. Chater, D. A. Hopwood, *Practical Streptomyces Genetics* (John Innes Foundation, Norwich, UK, ed. 1, 2000).
17. J. Davies, D. Davies, *Microbiol. Mol. Biol. Rev.* **74**, 417 (2010).
18. H. W. Boucher *et al.*, *Clin. Infect. Dis.* **48**, 1 (2009).
19. Materials and methods are available as supplementary materials on *Science* online
20. M. O. Sommer, G. Dantas, G. M. Church, *Science* **325**, 1128 (2009).
21. G. Dantas, M. O. Sommer, R. D. Oluwasegun, G. M. Church, *Science* **320**, 100 (2008).
22. V. L. Clark, F. E. Young, *Antimicrob. Agents Chemother.* **11**, 871 (1977).
23. H. K. Allen *et al.*, *Nat. Rev. Microbiol.* **8**, 251 (2010).
24. F. Rezzonico, G. Défago, Y. Moënne-Loccoz, *Appl. Environ. Microbiol.* **70**, 5119 (2004).
25. M. W. Silby, C. Winstanley, S. A. Godfrey, S. B. Levy, R. W. Jackson, *FEMS Microbiol. Rev.* **35**, 652 (2011).
26. S. Romano *et al.*, *BMC Microbiol.* **9**, 267 (2009).
27. P. S. G. Chain *et al.*, *J. Bacteriol.* **193**, 4274 (2011).
28. C. S. Smillie *et al.*, *Nature* **480**, 241 (2011).
29. C. Walsh, *Nature* **406**, 775 (2000).

Acknowledgments: We thank R. Mitra for initial discussions regarding simulations of metagenomic assembly; T. Druley for discussions surrounding the use of the six β-lactamase control fragments; J. Fay for discussions on dating horizontal gene transfer; J. Gordon for support, thoughtful discussion, and as the advisor to A.R.; A. Moore for naming PARFuMS; and the Genome Technology Access Center at Washington University in St. Louis for generating Illumina sequence data. This work was supported by awards to G.D. through the Children's Discovery Institute (award MD-II-2011-117), the International Center for Advanced Renewable Energy and Sustainability at Washington University, and the National Academies Keck Futures Initiatives, Synthetic Biology–SB2. M.O.A.S. received funding from the Lundbeck foundation and the European Union FP7-HEALTH-2011-single-stage grant agreement 282004, EvoTAR. K.J.F. is a NSF graduate research fellow (award DGE-1143954). A.R. is the recipient of an International Fulbright Science and Technology Award. The data reported in this paper are described in the supplementary materials. Raw sequencing reads have been deposited to MG-RAST with accession nos. 4489630-39, 4489641-43, 4489645-46, 4489648-49, 4489650-51, 4489653-57, 4489659, 4489661-63, 4489665, and 4489667-68. Assembled sequences have been deposited to GenBank with accession nos. JX009202 to JX009380. The authors declare no competing financial interests.

Supplementary Materials

www.sciencemag.org/cgi/content/full/337/6098/1107/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S19
References (*30*–*39*)

20 February 2012; accepted 22 June 2012
10.1126/science.1220761

TLR13 Recognizes Bacterial 23*S* rRNA Devoid of Erythromycin Resistance–Forming Modification

Marina Oldenburg,[1]* Anne Krüger,[1]* Ruth Ferstl,[2]*† Andreas Kaufmann,[3] Gernot Nees,[3] Anna Sigmund,[1] Barbara Bathke,[4] Henning Lauterbach,[4] Mark Suter,[4,5] Stefan Dreher,[2] Uwe Koedel,[6] Shizuo Akira,[7] Taro Kawai,[7] Jan Buer,[1] Hermann Wagner,[2] Stefan Bauer,[3] Hubertus Hochrein,[4]* Carsten J. Kirschning[1]*‡

Host protection from infection relies on the recognition of pathogens by innate pattern-recognition receptors such as Toll-like receptors (TLRs). Here, we show that the orphan receptor TLR13 in mice recognizes a conserved 23*S* ribosomal RNA (rRNA) sequence that is the binding site of macrolide, lincosamide, and streptogramin group (MLS) antibiotics (including erythromycin) in bacteria. Notably, 23*S* rRNA from clinical isolates of erythromycin-resistant *Staphylococcus aureus* and synthetic oligoribonucleotides carrying methylated adenosine or a guanosine mimicking a MLS resistance–causing modification failed to stimulate TLR13. Thus, our results reveal both a natural TLR13 ligand and specific mechanisms of antibiotic resistance as potent bacterial immune evasion strategy, avoiding recognition via TLR13.

Toll-like receptor 2 (TLR2), TLR4, and TLR9 are major host sensors of Gram-negative bacteria, and TLR2 is thought to be the central detector of Gram-positive bacteria, whereas other pattern-recognition receptors (PRRs) such as TLR7 contribute to bacteria sensing as well (*1*–*7*). However, the high sensitivity of mice lacking expression of these TLRs to Gram-positive