Healthy human gut phageome

Manrique, Pilar

2016-09-13


http://hdl.handle.net/10138/225170
https://doi.org/10.1073/pnas.1601060113

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.
Healthy human gut phageome

Pilar Manrique*a, Benjamin Bolducb,1, Seth T. Walka, John van der Oostc, Willem M. de Vosa,*, and Mark J. Youngb,2

*Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717; bDepartment of Chemistry and Biochemistry Research, Montana State University, Bozeman, MT 59717; cLaboratory of Microbiology, Wageningen University, Stippeneng 4, 6708 WE, Wageningen, The Netherlands; dDepartment of Plant Pathology, Montana State University, Bozeman, MT 59717

The body supports a diverse community of microorganisms, the majority of which are thought to be bacteria residing in the lower gastrointestinal tract (1–4). A “healthy gut microbiome (GM)” concept, where similar, but not the same microbes provide common functions that promote human health is supported by metagenomic sequencing of microbes and comparative analyses of gene content (3). The healthy-GM structure is conserved among individuals at higher taxonomic levels (especially phylum) (2, 4), and a core GM at the species level shared by more than 90% of the individuals. Shared bacterial species have been proposed to form a “core bacterial community” that provides beneficial functions (3). This concept, however, is based almost entirely on bacterial sequence data and, with few exceptions (5–8), has not considered the role of bacteriophages.

Gut bacteria host a diverse bacteriophage community (phageome) that potentially helps determine microbial colonization and contributes to shaping GM structure and function. The active phageome of a healthy human (i.e., actively replicating as opposed to nonreplicating, integrated prophage) has been estimated to comprise 35–2,800 viruses (6, 8), the vast majority of which have no significant sequence homology to known bacteriophages. Although the phageome was found to be relatively stable within a person (6–8), comparative phageome analysis at a global scale has received little attention.

To address large-scale phageome overlap, we assembled metagenomic reads from a novel, deep bacteriophage dataset from stool samples of two healthy individuals and analyzed currently available data from 62 healthy people around the world. This analysis identified 23 bacteriophages that were present in more than one-half of all individuals. We then applied a sequence-based viral community network tool (9) to identify closely related viruses at higher taxonomic levels. This analysis identified 44 bacteriophage groups, of which 9 (20%) were shared across more than one-half of individuals around the globe. We propose shared bacteriophages as a healthy gut phageome (HGP) and that, together with their core bacterial hosts, play an essential role in maintaining GM function in the healthy human gut.

Results

Previous studies that identified a core GM in healthy people showed that increasing the depth of sequencing was critical to detect shared bacterial species (5). Thus, we conducted an ultradependent baseline study of the active human gut phageome of two healthy, unrelated adults at two sampling intervals (15 or 3 mo). We focused our efforts on DNA bacteriophages because of their relevance to host bacteria and because most (>95%) RNA viruses in the human gut are likely plant viruses (10). Total DNA was extracted from stool samples as well as purification of virus-like particles (VLPs). TEM analysis of VLPs showed a high concentration of head-tailed bacteriophages consistent with their taxonomic classification in the Caudovirales order (Podoviridae, Siphoviridae, and Myoviridae). Microviridae bacteriophages were also identified, and no eukaryotic virus sequences were found. 16S rDNA-based metagenomics (11) of total stool DNA confirmed that the bacterial community of both individuals was typical of healthy human adults, dominated by Bacteroidetes and Firmicutes. Also, as expected, the GM structure was stable between time points. Unlike previous studies, DNA extracted from VLPs was sequenced without prior amplification to avoid potential amplification bias. Sequencing yielded a total of 8.1 Gb of data, representing one of the most in-depth human gut phageome studies to date, even after deduplication of reads (Table 1).

Significance

Humans need a stable, balanced gut microbiome (GM) to be healthy. The GM is influenced by bacteriophages that infect bacterial hosts. In this work, bacteriophages associated with the GM of healthy individuals were analyzed, and a healthy gut phageome (HGP) was discovered. The HGP is composed of core and common bacteriophages common to healthy adult individuals and is likely globally distributed. We posit that the HGP plays a critical role in maintaining the proper function of a healthy GM. As expected, we found that the HGP is significantly decreased in individuals with gastrointestinal disease (ulcerative colitis and Crohn’s disease). Together, these results reveal a large community of human gut bacteriophages that likely contribute to maintaining human health.

Author contributions: P.M. and M.J.Y. designed research; P.M. performed research; P.M., B.B., S.T.W., J.v.d.O., W.M.d.V., and M.J.Y. analyzed data; and P.M., S.T.W., J.v.d.O., W.M.d.V., and M.J.Y. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All sequence reads reported in this paper have been deposited in the National Center for Biotechnology Information Sequence Read Archive (accession nos. SARM00415496–SARM00415499).

1Present address: Department of Microbiology, The Ohio State University, Columbus, OH 43210.

2To whom correspondence should be addressed. Email: myoung@montana.edu.

This article contains supporting information online at www.pnas.orglookup/ suppl/doi:10.1073/pnas.1601060113/DCSupplemental.

*Author contributions: P.M. and M.J.Y. designed research; P.M. performed research; P.M., B.B., S.T.W., J.v.d.O., W.M.d.V., and M.J.Y. analyzed data; and P.M., S.T.W., J.v.d.O., W.M.d.V., and M.J.Y. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All sequence reads reported in this paper have been deposited in the National Center for Biotechnology Information Sequence Read Archive (accession nos. SARM00415496–SARM00415499).

1Present address: Department of Microbiology, The Ohio State University, Columbus, OH 43210.

2To whom correspondence should be addressed. Email: myoung@montana.edu.

This article contains supporting information online at www.pnas.orglookup/ suppl/doi:10.1073/pnas.1601060113/DCSupplemental.
S1). Rank abundance analysis indicated that sufficient sequencing depth was achieved per sample to detect dominant bacteriophages (Figs. S1 and S2).

Four individual active phageome datasets were assembled together resulting in 4,301 contigs (ALL). A total of 70 complete circular and 2 nearly-full-length linear phage genomes, based on end sequence comparison were identified (Table S2; referred to as “complete” genomes). The remaining contigs were potentially complete or partial phage genomes. On average, 73% (range, 66–79%) of sequencing reads from each sample assembled into the ALL contigs, demonstrating a relatively even contribution from both individuals. Consistent with previous studies (6, 8), the diversity of the phage community was low (Shannon diversity index = 4.25; SD = 0.95), and the phage distribution was similar in each person, being dominated by a few phages, with no more than 115 genomes accounting for 75% of the normalized reads (Fig. S1). A total of 1,703 of the 4,301 contigs (40%), and 63 of the 72 (88%) complete genomes contained at least one ORF with homology to a phage protein demonstrating their viral origins. However, only 314 (7%) of ALL contigs contained a family taxon-specific phage orthologous group (mostly to Caudovirales). This indicates that the majority (60%) of active bacteriophages in the GM are novel, with only a limited subset that can be taxonomically classified. Taken together, this approach was successful in assembling complete phage genomes and showed the expected community structure.

To maximize the identification of shared phages between our two test subjects, reads from the four samples were mapped to the ALL contigs. Recruitment of a single read was used as evidence of phage presence in a particular sample, allowing the identification of both high- and low-abundance phages that were not originally assembled due to low coverage. As has been shown previously (6, 8), phageome structure was more similar within individuals over time [Bray–Curtis (BC) distance = 0.47; SD = 0.24] than between individuals (BC distance = 0.96; SD = 0.01). However, 17% of ALL contigs were present in both individuals in at least one time point, and 7% were present in both individuals at all time points. Of the 72 complete genomes, 56 (78%) were found in both individuals. These results reveal that there is a smaller but significant fraction of phage sequences shared between these two unrelated test subjects.

Given the high level of interindividual phageome variability previously reported, the identification of significant phageome overlap between our two test subjects was surprising, but consistent with the existence of a core microbial community identified by others (3, 12). This result led us to investigate the presence of ALL contigs in other human gut phageome datasets. The dataset by Norman et al. (7) was selected for comparison based on the large number of healthy (n = 62) and diseased subjects (n = 102), their geographic distribution, and the sequencing depth achieved (Table S1). The dataset was deduplicated, and 11% of the total reads were mapped to ALL contigs (Table S3). In total, 1,787 of the 4,301 ALL contigs (42%) were present in at least one person from the Norman dataset; 1,679 (39%) were present in 2–19% of people (termed “low overlap” bacteriophages), 132 (3%) were detected in 20–50% (termed “common” bacteriophages), and 23 (0.5%) were present in >50% (termed “core” bacteriophages) (Fig. S2). Of the 72 complete genomes from our two subjects, 10 (14%) were core and 23 (32%) were common bacteriophages (Fig. 1). One of the core bacteriophages is the recently identified “crAssphage” (13).

Our independent identification of crAssphage supports the robustness of our approach and highlights the discovery of a significantly expanded set of core GM bacteriophages. Given their prevalence, we refer to both core and common bacteriophages as the “healthy gut phageome,” or HGP. It is important to note that the 155 bacteriophages comprising the HGP represented only 4% of the estimated total bacteriophage community, and we presume that a larger set of HGP bacteriophages will be discovered with increasing sequencing depth of additional individuals. Also, the observation that some HGP members were detected by a single sequencing read suggests that they are either rare members of the GM or that the sequencing depth of the Norman dataset was too low to accurately estimate their true presence (Fig. S3). Regardless, these results support the presence of a broadly distributed HGP in healthy human adults.

To add confidence to these results, we reversed the process by first independently assembling the bacteriophage community from healthy individuals within the Norman dataset, followed by read recruitment from all of the individuals. We observed the same general findings. Of the 13,707 bacteriophage contigs from the 62 Norman healthy individuals, 1,408 (9%) were found in more than 20% of the individuals and 115 (0.8%) were found in more than 50% of the individuals. The average size of these contigs was 7 kb, compared with 32 kb of the HGP bacteriophages from our ALL dataset. This suggested that these contigs could be partial fragments from the core and common bacteriophages identified in the two test individuals. When we compared the HGP from the two test individuals with the Norman HGP, we identified all of the HGP bacteriophages initially identified in our ALL set. Of the 1,408 shared bacteriophages in the Norman dataset, 1,015 (71%) were highly related to the HGP. We also identified additional potential core bacteriophages (n = 10) and common bacteriophages (n = 86) (Table S4). These results show that our approach of using only two test individuals but ultradeep sequencing their bacteriophage community was sufficient to detect a HGP. We speculate that a much larger HGP will be identified by analyzing additional healthy individuals.

Erroneous mapping of sequencing reads was tested by mapping over 38 million viral metagenomic reads from other environments (marine and hot spring environments) and a synthetic dataset to ALL contigs. No reads were recruited. We used the synthetic dataset to estimate the percentage of true positives by mapping the synthetic reads back to a collection of 1,197 Caudovirales (including the 50 genomes from which the synthetic reads were derived). The error rate used to create this dataset was set at 1.5%.
(a value set by the program) compared with the estimated Illumina error rate of 0.0046% (14). We found that 93% of the contigs called were true positives (Table S5) and that increasing the number of reads required to call a contig did not significantly change this percentage (Table S5). Given the high error rate in the synthetic dataset, the frequency of miscalled contigs in our experimental datasets is likely much lower (approximately one order of magnitude), providing confidence that using even a single read to identify shared bacteriophages is a valid approach.

To investigate whether HGP bacteriophages play a role in maintaining human health, we compared the prevalence of core bacteriophages in healthy individuals to those with irritable bowel disease (IBD) [n = 66 with ulcerative colitis (UC); n = 36 with Crohn’s disease (CD)] (7) (Fig. 2). Although core bacteriophages are found on average in 62% of healthy individuals, their prevalence was reduced by 42% and 54% on average in UC and CD patients, respectively (value of $P < 0.0001$) (Fig. 2A). Moreover, healthy subjects harbored on average 62% of the 23 core bacteriophages; however, in UC and CD patients, the average is significantly reduced to 37% and 30%, respectively (value of $P < 0.0001$) (Fig. 2B). These results suggest that the HGP is significantly perturbed in diseased patients.

To analyze the HGP structure, we conducted a network analysis using all contigs. The network groups viruses based on sequence homology to identify related viruses and to organize short contigs from partial genomes that dominate metagenomic datasets (9). A total of 240 statistically supported groups were identified (clustering coefficient value of 0.667 and modularity value of 0.675), with 44 containing five or more contigs (Fig. 3A and Table S6). Groups with less than five contigs represented only 11% of the total reads and were excluded to focus on the largest groups of related sequences. Of the 72 complete genomes, 45 (63%) were present and belonged to 17 different groups. One of these groups contained crAssphage and three additional complete crAssphage-related genomes. Previously classified phage genomes were mapped onto the network to test the accuracy of the grouping, and in nearly all cases (25 of 28), phages belonging to the same family grouped together. We also evaluated 172 artificially fragmented Caulovirales genomes. The fragmented genomes clustered as expected, with 20 of 26 being family-specific groups. Only six groups had fragments from more than one family, which is consistent with the high level of horizontal gene transfer shown between bacteriophages (15). These results suggest that the 44 network-defined groups can be considered roughly as family-level groups. Only 14 of the 44 network-defined groups contained classified bacteriophages, supporting that novel bacteriophages are abundant in the human gut. The network analysis proved useful in reducing the complexity of metagenomic datasets by organizing contigs into biologically relevant groups. Whether the members of these groups carry out similar functions in the gut is a subject for future study.

The network analysis supports the global distribution of the HGP. The majority (41 of 44) of network-defined groups were present in at least one of our test subjects (Table S7): in the Norman study (Fig. 3C), 12 groups were present in 2–20% of individuals (low overlap group); 13 were present in 20–50% (common groups); and 9, including the crAssphage and 2 other groups with complete genome representatives, were present in >50% of healthy individuals (core groups, Fig. 3B). Overall, the viral network represents a more encompassing view of the viral community than can be seen by only looking at the limited number of specific taxa classified by fragmented sequence information and supports the existence of a definable HGP.

**Discussion**

The overall objective of this study was to investigate the presence of a HGP. We propose that the phageome of healthy people can be divided into three classes (Fig. 3D): (i) the core, found in more than one-half of all individuals; (ii) the common, shared among many individuals; and (iii) the low overlap/unique, found in a limited number of individuals. The core is composed of at least 23 bacteriophages in this study and is significantly reduced in UC and CD patients, suggesting that these bacteriophages play a role in maintaining human health. Our results do not account for temporal dynamics in either the GM or phageome. Therefore, our estimates of sharing (>50%, 20–50%, 2–20%) for these bacteriophage categories are conservative and may need to be refined with temporal data from more subjects.

The role of bacteriophages in the human GM is poorly understood. In most bacteria-dominated ecosystems, bacteriophages play a key role in shaping community structure and function (16,
Thus, it is expected that the human gut ecosystem is strongly affected by the activity of bacteriophages. Microbe–virus interactions are often dominated by lytic interactions in a prey–predator dynamic (18). However, community sequence analysis of the human gut indicates that lysogenic bacteriophages (prophages) dominate over lytic bacteriophages (6, 8, 19). It is thought that the majority of bacterial cells contain at least one prophage, and that selective activation of a subset of these prophages is what can be detected as the active phageome in stool samples. The dynamics that arise from a lytic–lysogenic balance in the gut are still unknown. However, the active gut bacteriophages have been shown to influence bacterial population dynamics in a murine gut microbial ecosystem model (20), and to confer advantages to certain microbiome members during stress, such as antibiotic exposure (21). We presume that the active phageome of healthy people plays a critical role in maintaining the structure and function of a healthy gut ecosystem. Although the HGP represents a small proportion of the overall potential gut bacteriophage community, elucidating the role of the active phageome in a healthy gut will be essential to fully understanding its potential roles in disease or rehabilitation following episodes of microbiome imbalance.

Identification of a human HGP was somewhat surprising. Currently, two somewhat contrasting views of human gut phageome exist. On one hand, there is evidence of more phageome similarity among relatives and household members compared with unrelated individuals (7, 8). On the other hand, there is evidence that some
phages are common to unrelated people (13, 19, 20, 22). Stern et al. (19) identified 991 bacteriophage sequences from cellular metagenomes that were shared across 124 individuals. The majority of these bacteriophages appeared to be integrated prophages, and only a small fraction was present as VLPs. We found an expanded set of shared phages as well as phages that were unique to individuals. It is likely that the core HGP is hosted by core gut bacteria detected by others (3, 12). Our work provides a more comprehensive view and accurate estimate of phageome overlap. We presume even more overlap will be discovered with greater sequencing depth, the application of similar network analytical approaches, and consideration of more individuals.

In humans, viruses are often viewed as deleterious disease-causing agents, but the existence of an HGP suggests that the net influence of bacteriophages in the human gut is not deleterious, but rather beneficial. Despite large differences in the species composition of gut bacteria between people, there is a reproducible, higher-level taxonomic signature (4). This higher-level structure is thought to be a marker of a healthy gut ecosystem. We posit that bacteriophages play an essential role in this ecosystem and that the human HGP, arising from the large pool of cellular prophages, likely contributes to the proper function of the healthy GM by influencing the stability and maintenance of a healthy ecosystem through active lysis and predator-prey dynamics. Our results increase the known repertoire of shared bacteriophages in healthy humans and expand the number of complete core HGP genomes from one (crAssphage) to nine representatives. Network analysis identified groups of related viruses, and provided a powerful tool to organize and reduce the complexity of fragmented metagenomic data. We suspect that our efforts (sequencing and computation) underestimated the full extent of the human HGP, and that larger studies will generate even more accurate estimates of HGP prevalence, which will allow for a better understanding of the cosmopolitan human phageome. Future studies are needed to address the role of the HGP in health and disease, and to assess the potential use of core bacteriophages for clinical therapeutics and controlled manipulation of the human gut ecosystem.

Materials and Methods

Sample Collection. Samples were collected from two individuals of 26 and 55 y at two time points over a period from time 0–3 or 15 mo, respectively, under a protocol approved by the Internal Review Board of Montana State University. Both individuals were healthy, with normal bowel movement (at least once every 2 d and no more than three times a day), and had not received antibiotics for >2 mo before the collection of the samples. Collections were carried out with subject’s informed consent under an approved institutional review board protocol.

VLP Isolation, DNA Extraction, and Sequencing. VLPs were purified based on Minot et al. (6) (modification: 0.45-μm filtration). The 1.5 gcm⁻³ fractions were pooled, dialyzed, and DNase treated, and DNA was extracted using PureLink Viral DNA/RNA extraction kit (Invitrogen) and sequenced using MiSeq Paired End Illumina Technology.

Bacteriophage Sequences Assemblies. Bacteriophage sequences from two test individuals were initially assembled using MIRA 4.0 (23) and VICUNA 24 (quality trimmed reads using Trimmomatic v0.32 (25), 25-bp overlap at a minimum of 90% identity, and a minimum contig size of 1,000 bp). Contigs generated by MIRA and VICUNA were further assembled to each other using Geneious with high-stringency parameters [a read needed to be aligned to contig/base pair of contig]. The number of reads per contig was normalized to the length of the contig (number of total base pairs aligned to contig/base pair of contig). The percentage of total normalized reads to a contig was considered the percentage of the bacteriophage community accounted for by a particular contig. BC distances and between communities and Shannon index were calculated using vegan package in R.

Bacteriophage Community Analysis. Reads were recruited to contigs using Geneious with high-stringency parameters [a read needed to be >98% identical across its entire length (300 bp) to be recruited]. The number of reads per contig was normalized to the length of the contig (number of total base pairs aligned to contig/base pair of contig). The percentage of total normalized reads to a contig was considered the percentage of the bacteriophage community accounted for by a particular contig. BC distances and between communities and Shannon index were calculated using vegan package in R.

Taxonomic Classification and Detection of Bacteriophage Genes. ORFs were predicted using the WebMGA server. Proteins longer than 100 aa with less than 1% of the length of ambiguities were queried against ACLAME database (e value < 1e-10) and contigs were classified based on Waller (28).

Detection of Contigs in Publicly Available Metagenomes. Reads from the Norman dataset (7) were quality trimmed using Trimmomatic v0.32 (25) and deduplicated using cd-hit-dup (29). Trimmed unique reads were mapped to ALL contigs using high-stringency parameters [90% identity across the full length of the maximum read length (250 bp) using Geneious]. Reads shorter than this length will not be aligned; thus, these parameters only allow for high-quality deduplicated reads to be recruited to contigs.

Read Recruitment Controls. Because read length from the hot springs datasets (9) was very heterogeneous (340–855 bp), reads were aligned at 90% over the minimum length of these datasets. Reads from marine metagenomes [deep ocean hydrothermal vent (SRR3133481)] were mapped with the same aligning parameters as the reads from the Norman dataset, adjusting the read length to the size of the reads from this dataset (202 bp). A synthetic dataset was generated from randomly selected 50 genomes from a list of 1,197 complete Caudovirales genomes downloaded from National Center for Biotechnology Information with Grinder (30). The generated 2,000,000 Illumina reads (250 bp) were mapped to ALL contigs and to the set of 1,197 genomes using the same parameters as for Norman reads (see above).

Network Analysis. The network analysis was carried out based on Bolduc et al. (9). Briefly, network analysis is based on an all-versus-all BLASTN/neuviz search that incorporates both the number and strength of matches between contigs to group-related sequences into discrete groups. BLASTN analysis was performed with default parameters, except e-value cutoff was set to 1e-10 and max target sequences to 10,000. Matches were filtered based on the parameters of high-scoring segment pair (HSP) length > 200 (minimum alignment length) and 75% nucleotide identity. The network algorithm measures the number and weight (HSP and e value) of these connections and clusters sequences into groups of highly related sequences. The network was manually curated and visualized using Gephi with ARF view. A network control was performed on 172 complete Caudovirales genomes associated with bacterial host from the human gut (31). Genomes were fragmented into 9,000 with 50-bp overlap using an in-house python script (9), and network analysis was performed using the same parameters used in the ALL dataset network.

Identification of HGP Bacteriophages in Norman Assemblies. HGP bacteriophages were identified via network analysis. Contigs from our ALL and Norman dataset were pooled together. An all-versus-all BLASTN analysis was performed and network analysis was applied (see above). Contigs that were in the same group as HGP bacteriophages identified in our analysis were considered closely related HGP bacteriophages.

ACKNOWLEDGMENTS. We thank Dr. Herbert W. Virgin and Dr. Scott A. Handley of Washington University School of Medicine for sharing their data with us. This work was supported by National Institutes of Health Grant R01–1R01GM117361.