nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-018-0210-y

Murine colitis reveals a disease-associated bacteriophage community

Breck A. Duerkop[1,7*], Manuel Kleiner[2,7*], David Paez-Espino[3], Wenhan Zhu[4], Brian Bushnell[3], Brian Hassell[5], Sebastian E. Winter[4], Nikos C. Kyrpides[3] and Lora V. Hooper[5,6*]

The dysregulation of intestinal microbial communities is associated with inflammatory bowel diseases (IBD). Studies aimed at understanding the contribution of the microbiota to inflammatory diseases have primarily focused on bacteria, yet the intestine harbours a viral component dominated by prokaryotic viruses known as bacteriophages (phages). Phage numbers are elevated at the intestinal mucosal surface and phages increase in abundance during IBD, suggesting that phages play an unidentified role in IBD. We used a sequence-independent approach for the selection of viral contigs and then applied quantitative metagenomics to study intestinal phages in a mouse model of colitis. We discovered that during colitis the intestinal phage population is altered and transitions from an ordered state to a stochastic dysbiosis. We identified phages specific to pathobiotic hosts associated with intestinal disease, whose abundances are altered during colitis. Additionally, phage populations in healthy and diseased mice overlapped with phages from healthy humans and humans with IBD. Our findings indicate that intestinal phage communities are altered during inflammatory disease, establishing a platform for investigating phage involvement in IBD.

IBDs, including Crohn's disease and ulcerative colitis, are chronic disorders characterized by persistent inflammation of the intestine[1]. IBD imposes health and economic burdens worldwide[2], inspiring intense investigation into the host and environmental factors involved in IBD. The intestinal microbiota is integral to intestinal health[3–7] and has been implicated in IBD[8]. Efforts to define the composition and function of intestinal microbial communities during health and disease have yielded key insights into how the immune system maintains a beneficial relationship with the microbiota[9].

Studies aimed at understanding the contribution of the microbiota to inflammatory diseases have focused primarily on the bacterial component. However, metagenomic studies have revealed an abundant viral component of the microbiota[10–13]. While both eukaryotic and prokaryotic viruses are part of the microbiota, prokaryotic viruses, known as phages, dominate. Intestinal phages exist in two states: as infectious lytic particles or as lysogenic prophage elements integrated within bacterial chromosomes[14]. Stimuli such as nutrient cues and reactive oxygen species, common signals associated with intestinal inflammation, can induce prophage excision and replication from bacterial chromosomes[15–18]. Increased phage infectivity during inflammation points to the immune system and inflammation as drivers of prophage excision and phage particle biosynthesis in the intestine[19]. However, the exact signals and mechanisms that promote prophage excision and lytic viral particle production in the intestine remain to be determined and the role of phages in intestinal inflammation is unclear.

Phages associate with intestinal mucosal surfaces and elevated levels of certain double-stranded DNA phages have been implicated in IBD[20–23]. However, there is a limited understanding of the effects of inflammation on phage communities and whether phages play a role in the development of IBD. To begin to address these questions we used a quantitative metagenomics approach to define the abundances of host-associated phage populations in an experimental mouse model of T-cell-mediated colitis. We observed that the abundance of intestinal phages is altered during colitis, and that phages that infect members of the microbiota with pathogenic potential (pathobionts) are elevated during colitis. Our data show that the host inflammatory response correlates with compositional changes in intestinal phage communities, and highlight the need for further investigation of phage biology during intestinal disease.

Results

A mouse model of intestinal colitis for the isolation of virus-like particles (VLPs). To determine whether colitis alters the intestinal phage community we studied a mouse model of T-cell adoptive transfer that exhibits chronic inflammation of the colon (Supplementary Fig. 1a)[24,25]. *Rag1*$^{-/-}$ mice were injected with CD4^{+} CD45RB^{High} T cells or saline, housed in individual cages and monitored for 6 weeks (42 days) for weight loss and inflammatory markers (Supplementary Fig. 1b,c). Faeces were collected from each mouse at days 0, 14 and 42 (Supplementary Fig. 1a), and genomic DNA from the whole metagenome and VLPs was isolated. Whole metagenome and VLP DNAs were sequenced without amplification. Similar to a previous study, the resulting viromes are among the first quantitative viromes available from the microbiota[26]. A similar number of VLPs were recovered from faeces when counted by epifluorescence microscopy (Supplementary Fig. 1d). Bacterial composition analysis of the faecal microbiota showed an increased abundance of Proteobacteria at 42 days (Supplementary Fig. 1e). Similar results were obtained in mice cohoused with their treatment groups for 30 days after T-cell transfer (Supplementary Fig. 2a). These data support prior reports showing that Proteobacterial communities expand during intestinal inflammation[27–29].

[1]Department of Immunology and Microbiology, University of Colorado School of Medicine, Aurora, CO, USA. [2]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC, USA. [3]Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA. [4]Department of Microbiology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [5]Department of Immunology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [6]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA. [7]These authors contributed equally to this work: Breck A. Duerkop, Manuel Kleiner. *e-mail: breck.duerkop@ucdenver.edu; manuel_kleiner@ncsu.edu; lora.hooper@utsouthwestern.edu

Generation and verification of de novo VLP contigs. To alleviate the limitation of a priori knowledge of virus gene sequence features during virus contig identification we developed a sequence-independent approach. Our method employs the quantitative comparison of contigs between whole metagenomes and VLP metagenomes, making use of information obtained from physical VLP enrichments. We used sequence-based approaches to validate these contigs as viral (see Methods). We used a quantitative metagenomic approach to detect viruses whose abundances differ between healthy and inflamed mice using an established method of read recruitment to the de novo assembled contigs (Supplementary Fig. 3)[26,30]. Although we used a read-mapping ratio approach to limit contigs of potential bacterial origin within our VLP contig data set, we cannot rule out the presence of bacterial genomic DNA sequences in viral contigs originating from induced prophages. Our custom VLP database consisted of 1,104 contigs ranging in size from 2 kb to 486 kb. Approximately 80% of contigs were 2–25 kb in length, 16% were 25–100 kb and the remaining 4% were >100 kb.

To validate the viral origin of the contigs we employed two separate approaches. First, we compared their open reading frames to a set of 25,281 viral protein families (VPFs) from known viruses described by Paez-Espino and colleagues[31]. We identified 278 viral contigs with >70% of genes belonging to VPFs (Supplementary Table 1). We found 600 and 118 viral contigs with 35-70% and less than 35% of their genes belonging to VPFs, respectively (Supplementary Table 1). The remaining 108 contigs contained genes that did not match any VPF. Contrary to the standard application of VPFs for the extraction of viral sequences from metagenomes, which requires meeting stringent criteria, we considered our viral contigs validated if they contained at least one VPF. We further validated our VLP contigs using VirSorter, a tool that identifies viral sequences from complex microbial DNA samples[32]. VirSorter identified 156 of the VLP contigs as phages and, of these, 152 were also identified using VPFs (Supplementary Table 2).

We next compared our contigs to the IMG/VR viral database, which contains over 264,000 metagenomic viral contigs and 8,000 viral isolate genomes[33]. We used a sequence-based classification framework to group closely related viral sequences[31]. Of the 1,104 contigs in our database, 382 viral contigs (representing 34% of the total contigs) grouped within 142 viral clusters (Supplementary Table 3). All viral contigs from IMG/VR that clustered with the contigs from this study originated from intestinal viruses (Supplementary Table 3). The remaining 722 viral contigs did not group with any viral contigs in IMG/VR or our curated database and were considered singletons.

Quantitative metagenomics reveal alterations to phage communities during colitis. To determine whether compositional changes occur within the phage community during colitis, we mapped our VLP reads to all publicly available Caudovirales genomes downloaded from the National Center for Biotechnology Information (NCBI), all viral isolate genomes and metagenomes in IMG/VR and our curated VLP contig database.

Mapping to phage reference genomes from NCBI showed divergence of the phage community in T-cell-treated animals at 42 days, suggesting that disease status correlates with an altered faecal phage community (Fig. 1a). Enterobacteriaceae-specific phages were enriched in T-cell-treated mice compared to healthy controls at day 42 (Fig. 1b and Supplementary Table 4), including phages resembling those that infect Proteobacteria that expand during inflammation (Supplementary Figs. 1f and 2b). Reads that mapped to phages predicted to infect Enterococci were also more abundant in T-cell-treated animals on day 42 (Fig. 1b and Supplementary Table 4). These data correlated with an increased proportion of phage reads that mapped to known viruses in colitic mice (Supplementary Fig. 4). The presence of some phages correlated with an increased abundance of prospective host bacteria, suggesting that phages interact with known intestinal pathobionts during colitis (Supplementary Figs. 1f and 2b). In addition, VLP reads that mapped to Caudovirales phage families, including the Siphoviridae, Myoviridae and Podoviridae, were elevated in T-cell-treated animals (Fig. 1c). These data accord with the altered phage abundances in human IBD patients[21]. The VLP reads that mapped to the Caudovirales genome data set accounted for less than 0.01% of the total VLP reads per group of mice, indicating that most of the phage-specific VLP reads are unique to our data set or fall below our identity cut off.

To improve VLP read mapping for our samples we mapped the VLP reads to the IMG/VR database, as well as the 1,104 viral contigs extracted from assemblies of our data set. IMG/VR contains viral isolate genomes and metagenomes from ocean, soil, sewage and animal hosts[33]. A total of 1,068 viral genomes from IMG/VR recruited reads from our mouse VLP samples (Supplementary Table 5). Of these viral genomes, 958 originate from intestinal samples including 406 from humans, 417 from mouse, 102 from rat and 33 from other mammals (Supplementary Fig. 5 and Supplementary Table 5). This increased our read-mapping percentage by two orders of magnitude (1.2%).

We next mapped each sample's VLP reads to our curated database of 1,104 viral contigs (Fig. 2a). Using the curated contig data set allowed us increase our read-mapping average to 22% (Supplementary Table 6), yet 78% of the reads were not mapped to any database. Comparative read-mapping analysis of day 0 samples showed that both T-cell-treated and control animals had a similar viral community. However, as colitis progressed the viral community shifted, with the most marked shift occurring 42 days after T-cell treatment (Fig. 2a–c). This shift in viral community composition was also observed in animals that remained cohoused with their treatment group for the duration of the experiment (Supplementary Fig. 6). Thus, T-cell-mediated colitis was associated with a divergence of the intestinal viral community.

The shift in phage community composition at day 42 reflected a decrease in viral contig abundances (Supplementary Table 7) and a significant loss of diversity, leading to an increased dissimilarity between healthy and diseased mice (Fig. 2c). Rarefaction analysis revealed sufficient read-mapping coverage for all samples, ruling out low sampling as a cause of the divergence (Supplementary Fig. 7). In addition to altered phage abundances there was a marked divergence in the bacterial community during colitis (Supplementary Fig. 8). We found that some viral contigs differed between control and diseased mice at day 42 (Fig. 2d). Of the contigs that were statistically significant, 35 contigs were at least 10-fold more abundant and 41 contigs were at least 8-fold less abundant during colitis.

When comparing the presence of various contigs between healthy and colitic animals, we observed a stochastic pattern consistent with a random distribution of contigs within animals at various time points (Fig. 3a). Four percent of the VLP contigs were shared between healthy animals as compared to only 1% in animals with colitis (Fig. 3a). Furthermore, the colitic animals contained a greater number of singleton viral contigs (Fig. 3b). Although contigs shared between healthy animals accounted for a small proportion of the total VLPs, these data suggest that the viral community diverges as the intestine transitions from a healthy to dysbiotic state.

At day 42 there was a dramatic reduction in the number of reads recruited to VLP contigs in colitic mice. This was most pronounced for mouse number two (Fig. 2a). Twenty percent of the mapped reads from this animal aligned with four highly abundant contigs. Three of these contigs were among the most differentially abundant contigs at the height of disease (Fig. 2d). An 18.4-kb contig recruited 8.3% of the reads and the other three contigs of 136.6 kb, 145.5 kb and 176.1 kb recruited the remaining 11.7% of the reads (Supplementary Fig. 9 and Supplementary Table 8). Although these

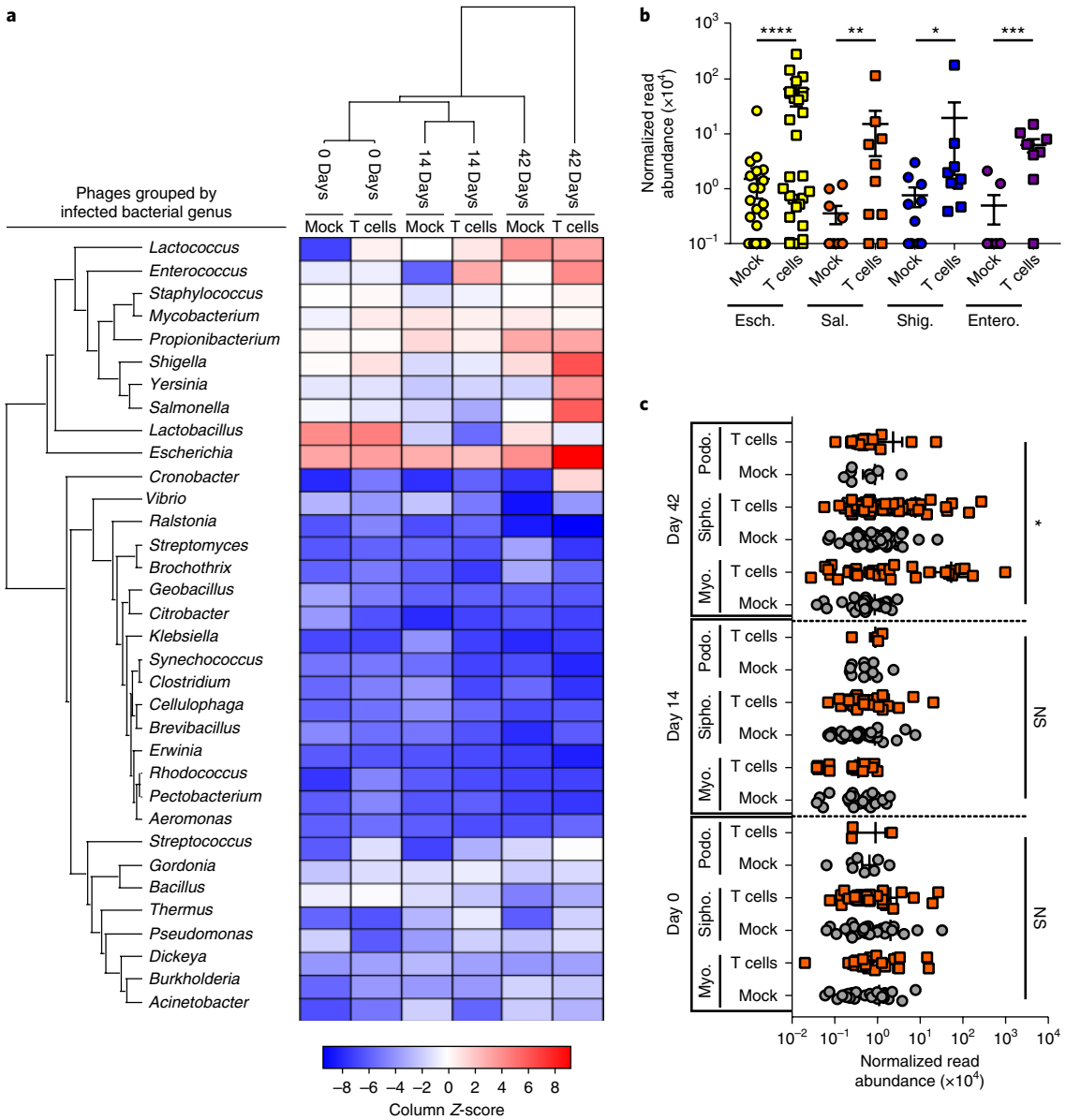

Fig. 1 | Caudovirales phage abundances are altered during T-cell-mediated colitis. a, Heat map indicating the abundance of VLP sequencing reads with at least 97% identity to select sequenced Caudovirales phage genomes deposited in NCBI. Relative read abundances were summed by phage host genus and *Z*-scores were normalized by column. **b**, Quantitative read-mapping abundances for select groups of phages shown in panel **a**. Esch., *Escherichia* phages; Sal., *Salmonella* phages; Shig., *Shigella* phages; Entero., *Enterococcus* phages. ****$P = 0.0003$, ***$P = 0.003$, **$P = 0.02$, *$P = 0.05$ by two-tailed nonparametric Mann–Whitney test. **c**, Mapped read abundances for the Caudovirales phage families Siphoviridae (Sipho.), Myoviridae (Myo.) and Podoviridae (Podo.) at days 0, 14 and 42 post T-cell transfer. Mock, saline-injected control mice; T cells, mice receiving $CD4^+$ $CD45RB^{Hi}$ T cells. *Z*-scores show how many s.d.s a value in a column is above and below the population mean. The data represent three independent mice per group. Error bars represent ± s.e.m. *$P = 0.007$ by one-way analysis of variance. NS, not significant.

contigs recruited a large proportion of the VLP reads from T-cell-treated animals, read recruitment to these contigs varied between animals (Supplementary Fig. 9).

We were able to assign phage taxonomy or bacterial host information to two of these four contigs. We assigned phage taxonomy or host information using a list of signature phage orthologous groups (POGs), phage-encoded transfer RNAs and clustered regularly interspaced short palindromic repeat (CRISPR) spacers that serve as a historical record of phage–bacteria interactions[34–36]. The 145.5-kb contig contained a tRNA 100% identical to the genus *Acholeplasma*, a member of the *Mollicutes*. This contig had a region of ~8 kb that resembled known phage genes (Supplementary Fig. 9 and Supplementary Table 8). The 136.7-kb contig matched a POG belonging to Spounaviridae, a subfamily of the *Myoviridae* composed of phages known to infect pathogens such as *Staphylococcus aureus*[37].

Of the remaining 35 statistically significant contigs of high abundance, only 3 other contigs could be assigned taxonomy or host information (Supplementary Table 9). A tRNA broadly connected one contig to either the Firmicutes or Tenericutes, one contig was assigned to a POG representing the genus *Betalipothrixviruses* that infect Archaea[38] and the third contig was linked to the Clostridiales through a CRISPR spacer. Although we assigned taxonomy or host information to only a few significantly different contigs, this was expected based on previous studies[21,39]. This highlights the power of our approach for identifying VLPs independent of existing database information. Of the 41 contigs with significantly reduced

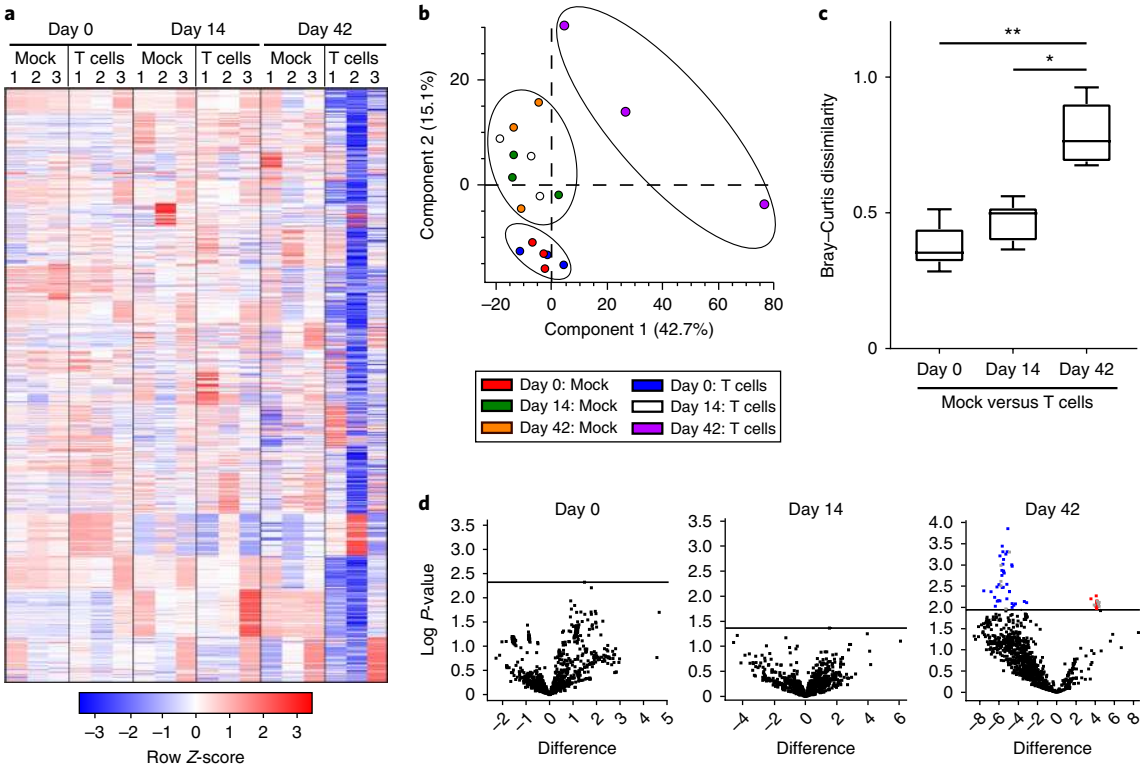

Fig. 2 | Phage community alterations during colitis were identified using a curated VLP contig database. a, Heat map showing the differential abundance analysis of VLP sequencing reads mapped to a curated database consisting of 1,104 VLP contigs during colitis. Numbers 1–3 represent individual mice within each group. Each row representing a unique phage contig was *Z*-score normalized. **b**, Principal component analysis showing that phage community abundances are different between healthy and T-cell-treated animals. Ellipses were manually added to guide the reader and do not indicate confidence intervals. **c**, Bray–Curtis dissimilarity β diversity metric quantifying the dissimilarity of the phage communities between healthy and T-cell-treated animals across the time series. The closer the value is to 1, the more dissimilar the communities are in terms of species composition. The whiskers represent the 5th to 95th percentile of an all against all comparison with the central line in each box representing the median value of the data set. **d**, Volcano plots of two-tailed *t*-tests testing for changes to viral contig abundances during colitis. The multiple-testing problem was corrected by calculating a permutation-based false discovery rate (FDR). An FDR cut off of 0.05 was used. Horizontal lines indicate the FDR-based *P* value cut off for statistical significance for each time point when comparing the contig abundances between healthy and diseased animals. Data points highlighted in blue and red indicate contigs that are significantly reduced or enriched in colitic animals, respectively, and grey data points indicate contigs for which phage host taxonomic assignments could be made. The data represent three independent mice per group. $^{**}P = 0.0001$, $^{*}P = 0.01$ by Kruskal–Wallis test and a post hoc Dunn's test for multiple comparisons.

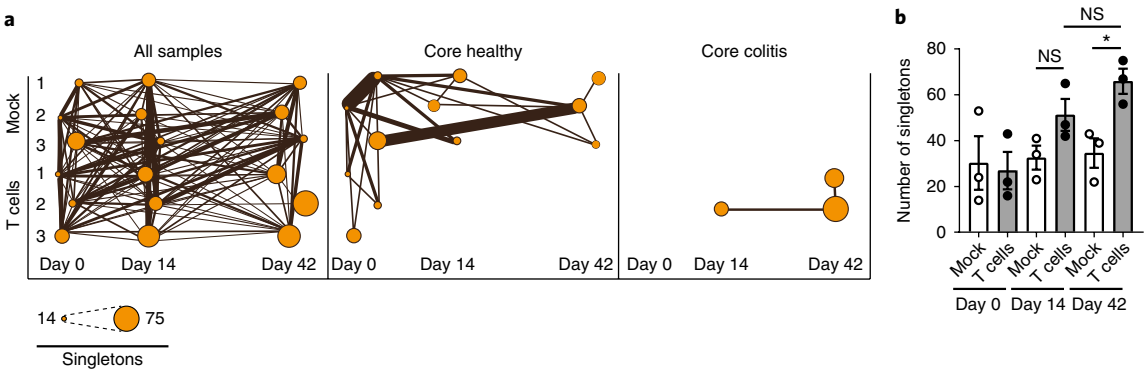

Fig. 3 | Colitic animals share fewer VLP connections relative to healthy animals. a, Connected VLP contigs (both shared and singleton contigs) across all time points and animals, from only healthy animals (Core healthy), which includes animals before T-cell treatment and T-cell-treated animals 14 and 42 days post T-cell administration (Core colitis). Circle diameter indicates the number of singleton VLP contigs that do not cluster with contigs in IMG/VR or to contigs within the curated database. Line thickness represents the number of shared contigs between animals and time points. **b**, Abundance of singleton VLP contigs between mock and T-cell-treated animals across the time series. The data represent three independent mice per group. Error bars represent ± s.e.m. $^{*}P = 0.02$ by two-tailed unpaired Student's *t*-test.

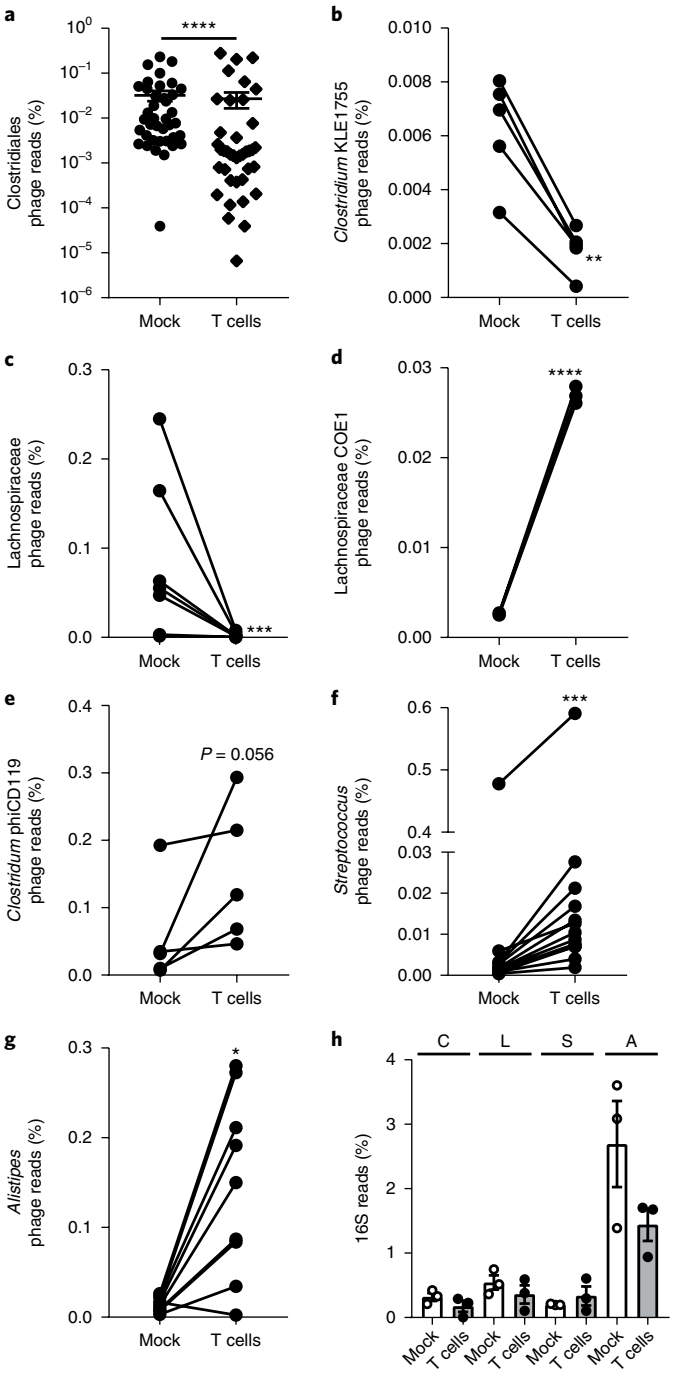

Fig. 4 | Phage taxonomy and host bacterial assignments for curated VLP contigs reveal differential abundances of phages that infect both commensals and pathobionts during colitis. a, All phage contigs linked to the order Clostridiales. Each data point represents a unique Clostridiales phage contig found in both healthy and diseased animals. **b**–**e**, Family- and genus-specific phages constituting lower taxonomic ranks of the Clostridiales. **f**,**g**, Abundances of Streptococcal phages and *Alistipes* phages. **h**, Bacterial 16S rDNA sequencing read abundances. C, Clostridiales; L, Lachnospiraceae; S, Streptococci; A, *Alistipes*. In panels **b**–**g**, lines connect shared contigs between healthy and diseased animals. The data represent three independent mice per group. Error bars represent ± s.e.m. $^{****}P < 0.0001$, $^{***}P = 0.0002$, $^{**}P = 0.008$, $^{*}P = 0.004$ by two-tailed nonparametric Mann–Whitney test.

abundance, CRISPR spacer and tRNA matching revealed 6 to be potential Clostridiales phages and 2 were identified as phages that infect the Lachnospiraceae family within the Clostridiales. These data indicate that colitis is associated with a reduction in Clostridiales-like phages.

VLP identification reveals changes in the phage community during colitis. We next determined taxonomic or host assignments for all 1,104 contigs in the curated VLP database. We were able to assign 116 contigs (11%) to a specific phage taxonomic group or host designation using one or more of the 3 identification methods described above. The number of total assigned contigs is smaller than the number of contigs predicted to be of viral origin using VPFs because numerous VPFs only contain viral hallmark genes and lack taxonomic or host information. CRISPR spacer matching resulted in the greatest number of assigned contigs (50 contigs), followed by tRNAs (41 contigs) and POGs (36 contigs) (Supplementary Fig. 10). Both tRNA and CRISPR spacer matching identified 11 shared contigs, accounting for 27% and 22% of the total contigs identified using those methods, respectively. Of the 36 contigs identified using POGs, 26% belonged to the Myoviridae, 18% to the Siphoviridae and 37% to the Podoviridae. The remaining 19% of contigs matched several archaeal viruses belonging to the Fuselloviridae and Lipothrixviridae (Supplementary Fig. 10a).

POG matching identified numerous phages with taxonomic signatures of known Enterobacteriaceae-like phages. This is consistent with the increased abundance of Proteobacteria in the metagenomes of T-cell-treated mice and the mouse VLP reads mapping with higher abundance to Proteobacterial phage genomes (Supplementary Fig. 1e and Fig. 1a). Since viruses within a given POG can infect very different bacterial hosts, we cannot completely exclude that the contigs classified as Enterobacteriaceae-like phages are from viruses that infect bacteria outside the Proteobacteria. Additionally, POGs identified phages taxonomically related to phiCD119, phiKZ, phiHp1, AHJD and Spounavirinae phages that infect pathobionts including *Clostridium difficile*, *Pseudomonas aeruginosa, Haemophilus influenza* and *S. aureus*[37,40–42] (Supplementary Fig. 10a and Supplementary Table 9). We were able to obtain order, family and in some cases genus-specific identification of multiple phage contigs using tRNA and CRIPSR spacer matching. As expected, the most highly represented phages in our database matched hosts in the phyla Bacteroidetes and Firmicutes. Firmicute-specific phages were dominated by those predicted to infect Clostridiales bacteria including phages specific for Lachnospiraceae, Ruminococcaceae and Clostridia. The next most prevalent were phages specific for the Streptococci. Bacteroidetes phages were almost exclusively represented by *Alistipes* phages (Supplementary Fig. 10b,c and Supplementary Table 9).

To identify differences in the abundance of host-assigned phage contigs we grouped the contigs based on host identity, summed the read-mapping-derived abundances and averaged these values for healthy and colitic animals. At the order level there was a significant decrease in the abundance of Clostridiales phages in T-cell-treated animals (Fig. 4a). This supports our finding, from the analysis of the differentially abundant contigs, that a reduction in Clostridiales phages occurs in colitis. When these data were stratified into family- and genus-specific contigs we discovered that phages predicted to infect *Clostridium sp*. KLE 1755 and some Lachnospiraceae bacteria decreased, whereas Lachnospiraceae phages that target strain COE1 were more abundant during colitis (Fig. 4b–d). Additionally, Clostridiales phages predicted to infect *C. difficile* were more abundant in the T-cell-treated animals (Fig. 4e).

Streptococcus sp. phages were more abundant in T-cell-treated animals, supporting a prior report on intestinal phages from IBD patients[21], as well as phages predicted to infect *Alistipes* (Fig. 4f,g). The 16S ribosomal RNA gene read abundances for the host bacteria

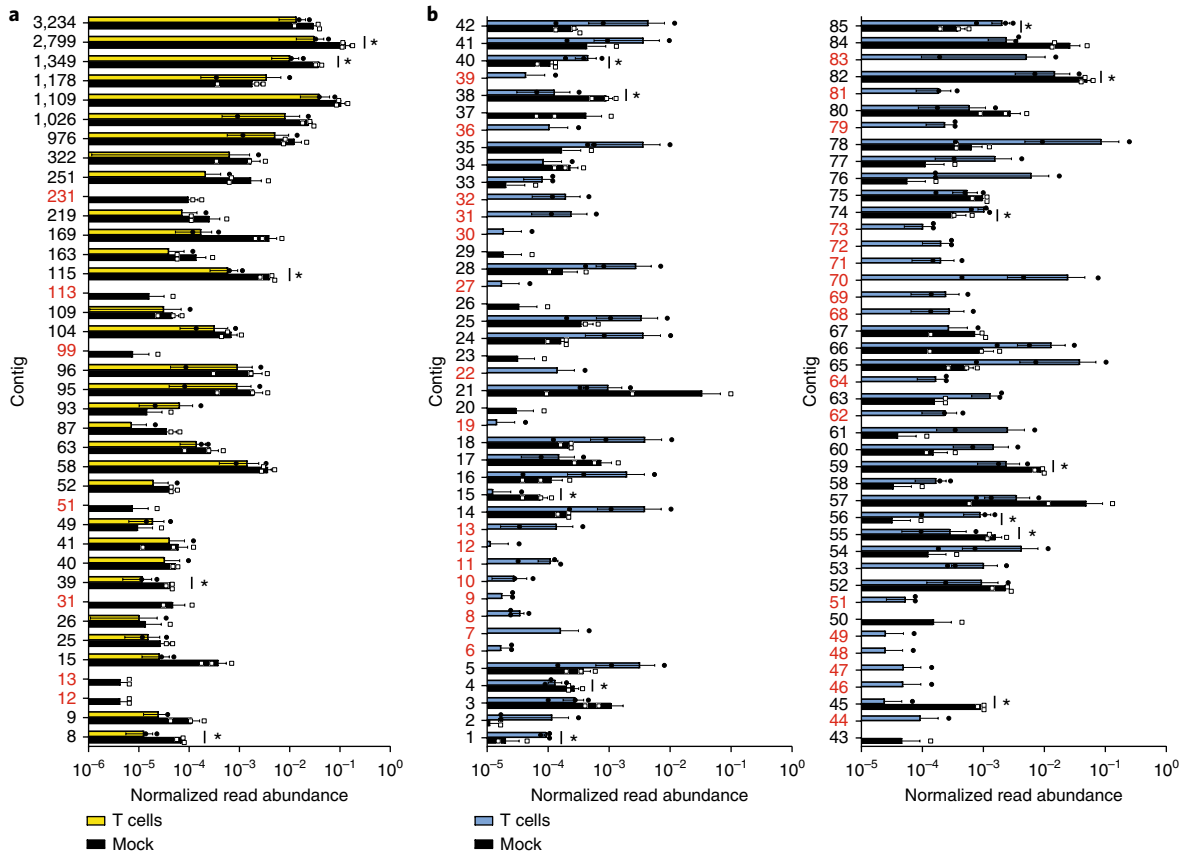

Fig. 5 | VLP reads from healthy and colitic animals share identity to human-associated intestinal phages. a, VLP reads from healthy (Mock) and colitic (T cells) mice 42 days post T-cell treatment were mapped against a cohort of conserved phage contigs found in healthy humans. **b**, VLP reads from healthy (Mock) and colitic (T cells) mice 42 days post T-cell treatment were mapped against phage contigs assembled from IBD patients. Panel **b** shows the top 85 IBD phage contigs with read recruitment. Normalized read abundance is the total number of mapped reads divided by the length of the contig in base pairs. Contigs highlighted in red indicate contigs that are only found in healthy or colitic mice. The data represent three independent mice per group. Error bars represent ±s.e.m. *P <0.1 by a two-sample unpaired t-test with a permutation-based FDR cut off of 0.1 (Supplementary Table 10). Scatter plot data for the T cell (black circles) and Mock (open squares) treated animals are overlaid with their representative bar graphs.

infected by these phages were not statistically different between healthy and colitic animals, although *Alistipes* showed a trend toward lower abundance during colitis (Fig. 4h). This suggests that shifts in phage abundance are not solely dependent on bacterial host density. Together, these data support the idea of a defined phage community during colitis.

Mouse VLP reads map to phages from human phageomes. To determine whether mouse VLP sequencing reads matched intestinal phages from humans, we mapped our VLP reads from healthy and colitic animals to groups of phage contigs derived from healthy humans and humans with IBD[21,39]. A total of 38 contigs, accounting for 25% of the healthy human phageome, recruited reads from both healthy and colitic animals (Fig. 5a). Comparative analysis of the read-mapping frequency revealed multiple healthy human phage contigs that recruited fewer reads from colitic animals and seven phages that were exclusively in healthy animals (Fig. 5a). Consistent with the observation that more contig connections were maintained under healthy conditions across time points (Fig. 3a), these data suggest that the healthy component of the murine phageome overlaps with that of humans and that the abundances of these phages decrease during colitis.

We next assessed the abundance of VLP reads that mapped to a set of 519 VLP contigs generated from the Chicago cohort of Crohn's and ulcerative colitis patients described by Norman and colleagues[21]. A total of 169 contigs, accounting for 33% of VLP contigs generated from IBD patients, recruited reads from either healthy or colitic animals (Fig. 5b). Phages from diseased humans recruited more reads from colitic animals and in some cases only human IBD phage contig-specific reads were represented in colitic mice (Fig. 5b). Thus, the murine phageome overlaps with that of humans during both intestinal health and disease.

Discussion

Using a metagenomics approach, we discovered that intestinal phage communities in colitic animals undergo compositional shifts that are similar to those observed in human IBD patients[21]. Accordingly, we observed a decrease in phage community diversity and an expansion of subsets of phages in animals with colitis. We observed decreased abundance of certain phages, such as Clostridiales phages, during colitis. Phage populations that expanded during colitis were frequently connected to pathobiont hosts that benefit from or are linked to intestinal inflammation. Not all colitis-specific phage expansion reflected increased abundances of assigned bacterial hosts, suggesting that alternative mechanisms drive altered phage abundances during colitis.

We hypothesize that inflammation or other host defences alter phage abundances during colitis. Such stresses could produce ecological disturbances in the intestinal environment, driving alterations within the viral community. Disturbances leading to the stochastic assembly of microbial communities have been described in other ecological networks[43,44], suggesting a plausible explanation

for the high level of variation and divergence observed in the viral communities of colitic animals.

Changes in viral community structure during colitis could have several underlying causes. For instance, most intestinal phages are integrated into bacterial chromosomes as lysogenic prophages[10,11]. Host inflammatory products could enhance prophage excision leading to increased phage abundances. Nutrient availability may also impact prophage excision. During inflammation, the metabolic landscape of the intestine dramatically changes and bacteria adapt to this environment[45]. Furthermore, prophage excision from bacterial chromosomes and progression to the lytic cycle are linked to nutrient availability[14,18]. As dysbiotic bacterial communities establish during colitis and competition for resources ensues, increased prophage induction could account for the emergence of specific phages.

Our findings suggest that members of the Spounaviridae subfamily of phages could serve as informative markers for colitis. Additionally, during colitis there was an overall heightened abundance of phages predicted to infect Streptococci and phages that target *Alistipes* bacteria. Streptococcal phages are abundant in human IBD patients[21] thus highlighting the possibility of using these phages as disease markers. Elevated abundance of *Alistipes* phages was unexpected as these phages associate with members of the Bacteroidetes whose abundances decline during dysbiosis[46,47]. Thus, prophage excision in the Bacteroidetes might be driven by colitis and could be an indicator of disease.

Overall, we observed the most dramatic changes in phages predicted to infect pathobionts. Since phages infect bacteria with strain-level specificity[48–50], elevated abundances of specific phages during IBD could serve as a proxy for strain-level resolution of disease-causing bacteria. This correlated with both reduced and increased abundances of phage DNAs in colitic animals that had similarity to phage genomes from healthy humans and humans with IBD, respectively. Previous studies have shown that there is a high degree of interindividual variation among human intestinal phage communites[10,11]. We were surprised to discover that there is overlap of murine intestinal phage reads with phages from both healthy humans and humans with IBD. These data suggest that experimental colitis in mice is a suitable model for the study of phage–bacterial interactions in the intestine, and that the findings from this model could be relevant to human disease.

Methods

Animals. Two-week-old male C57BL6/J *Rag1*$^{-/-}$ mice[51] were weaned and cohoused together in the same cage for four weeks. At 6 weeks of age 3 C57BL6/J *Rag1*$^{-/-}$ mice were administered 5×10^5 naïve CD4$^+$ CD45RB^{High} T cells via retro-orbital injection. A second group of three C57BL6/J *Rag1*$^{-/-}$ mice were given a saline injection as a control. All mice were housed in either separate cages or by treatment group for the duration of the experiment. The sample size of three mice per group was used because this was the minimum number of mice needed for reliable statistical analyses. Four male wild-type C57BL6/J mice aged six to eight weeks were used as spleen donors for T-cell isolation. Male germ-free C57BL6/J *Rag1*$^{-/-}$ mice were reared in sterile isolators as described previously[52]. Faecal samples from three eight-week-old germ-free *Rag1*$^{-/-}$ were pooled in duplicate. After cohousing for four weeks experimental mice were randomly assigned to treatment groups. It was not possible to blind researchers to the treatment conditions of the animals before experimentation. The Institutional Animal Care and Use Committee of University of Texas Southwestern Medical Center approved all animal protocols.

T-cell isolation. The spleens from four male wild-type C57BL6/J mice were dissected and added to RPMI (BD Biosciences) containing 10% foetal bovine serum (Gemini Bio-Products). The spleens were homogenized using the frosted ends of sterile microscope slides and passed through a 70-µm cell strainer. CD4 cells were negatively selected using a Dynabeads Untouched Mouse CD4 Cells Kit (Invitrogen). Purified mouse CD4 cells were Fc-blocked using purified rat anti-mouse CD16/CD32, clone 2.4G2, (BD Biosciences) followed by two-colour staining using anti-mouse CD4-FITC, clone RM4-5 and anti-mouse CD45 RB-APC, clone C363-16A (Thermo Fisher). CD4$^+$ CD45RB^{High} cells were collected on a MoFlo XDP cell-sorting machine at the University of Texas Southwestern Flow Cytometry Core Facility.

Metagenomic DNA isolation. VLPs and their DNA were isolated as described previously[53]. Briefly, faecal pellets were homogenized in 2 ml salt magnesium plus buffer[50] and 1.3 ml sample volume was clarified by centrifugation (2,500 *g* for 5 min followed by 5,000 *g* for 15 min at 4 °C). The supernatant was filtered through a 0.45-µm Millex-HV syringe filter (EMD Millipore) and treated with 100 units of DNase I (Roche) and 15 units of RNase A (Sigma) for 1 hr at 37 °C. VLPs were not further purified by caesium chloride gradient centrifugation. Although caesium chloride purification is efficient at removing potential contaminating nucleic acids, this method is not suited for quantitative metagenomic comparison of a large number of samples due to its limited throughput[53]. Additionally, this method discriminates against specific phages and can vary in reproducibility due to the necessity of removing the VLPs from precise locations within the gradient that are difficult to normalize between samples[53]. Therefore, we proceeded to isolate VLP DNA from filtered and DNase/RNase-treated samples by adding 100 mM EDTA, 50 µg ml^{-1} Proteinase K (New England Biolabs) and 0.5% SDS to each sample and to incubate them at 56 °C for 30 min. The samples were extracted with an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1) (Sigma) followed by an additional extraction using just chloroform. The DNA was precipitated with isopropanol and cleaned using a MinElute Reaction Cleanup Kit (Qiagen). Following a rigorous VLP purification strategy can greatly reduce the possibility of contaminating prokaryotic host genomic DNA that can lead to false positive viral gene identification from the VLP fraction[54,55]. For the isolation of total metagenomic DNA, the remaining 0.7 ml faecal material was treated with 5 mg ml^{-1} lysozyme (Amersco) at 37 °C for 30 min. The samples were transferred to 2 ml Lysing Matrix B tubes (MP Biomedicals) and were disrupted by bead-beating in a Bullet Blender 5 (Next Advance) at a speed setting of 9 for 1 min followed by a 2-min incubation on ice. This was repeated five times. Samples were centrifuged at 16,300 *g* for 1 min and DNA from the supernatant was extracted and purified as described above.

Metagenomic DNA sequencing. DNA sequencing was performed using Illumina HiSeq 2500 Rapid Run at 150 cycles in paired-end mode. DNA concentration and quality were determined using the Qubit dsDNA HS kit (Invitrogen) and a Bioanalyzer 2100 (Agilent). DNA libraries were prepared using a KAPA High Throughput Library Preparation Kit with standard PCR library amplification (KAPA Biosystems). TruSeq adapters (Illumina) with 6-bp indices were used to enable multiplexed sequencing. Illumina libraries were size selected using AMPure XP beads (Beckman Coulter). Library quality was measured on a Bioanalyzer 2100 by product size and concentration. On average, 43 million paired-end whole metagenome reads and 30 million paired-end VLP reads were generated for each sample. Sequencing reads were demultiplexed using the 6-bp TruSeq adapter sequences allowing for one mismatch using the Illumina CASAVA software. DNA quality control, HiSeq 2500 sequencing and library preparation were performed by the University of Texas Southwestern Eugene McDermott Center for Human Growth and Development Next Generation Sequencing Core.

Decontamination and processing of reads. Read decontamination and trimming were performed using the BBMap short read aligner v36.99[56]. Raw reads were mapped to the internal Illumina control phiX174 (J02482.1) and the mouse (mm10) and human (hg38) reference genomes using the bbsplit algorithm with default settings. The resulting unmapped reads were adapter trimmed and low-quality reads and reads of insufficient length were removed using the bbduk algorithm with the following parameters: ktrim = lr, k = 20, mink = 4, minlength = 20, qtrim = f. These reads went through a second round of decontamination by mapping to a set of reference contigs assembled from samples obtained from the faeces of germ-free C57BL6/J *Rag1*$^{-/-}$ mice using bbsplit. These final unmapped reads constituted the cleaned M and P read sets.

Read-based quantification of taxa. To assess microbial community composition, we used phyloFlash Version 2.0[57] to reconstruct 16S and 18S rRNA genes from the Illumina read data sets and quantified taxa by read mapping. 16S and 18S rRNA genes were classified with phyloFlash against the SILVA SSU Ref NR 99 database 123[58]. Taxa abundances were summed at the phylum level and the genus level for specific genera.

Metagenomic assembly. Whole metagenome, VLP and germ-free mouse decontamination reference assemblies were performed using the IDBA-UD assembler v1.1.1[59]. Before assembly, paired-end reads were interleaved using the fq2fa --merge --filter command within IDBA-UD. Assemblies were generated in the absence of error correction using the default parameters.

Construction of the curated VLP contig database. VLP reads were assembled into 18 sets of contigs, with 1 set per animal per time point. VLP contigs smaller than 2 kb were removed from each sample's contig set. Using these pruned contig sets, each sample's VLP and whole metagenome reads were mapped to their corresponding sample's contigs using the bbmap algorithm within BBMap with the following settings: ambiguous = random, qtrim = lr, minid = 0.97. Read-mapping coverage for each contig was generated by setting the 'covstats' parameter. The percentage of reads that mapped to each contig was used to calculate a read-

mapping ratio per contig (VLP/whole metagenome). We assumed that if this ratio was greater than 1 (that is, more VLP reads than whole metagenome reads mapped to a contig), then these contigs were more highly represented in the viral fraction and probably represented genuine viral contigs. From each of the 18 read-mapping sets we selected and pooled the top 100 contigs based on the largest ratios. We removed redundancy by sequence clustering using cd-hit-est[60] at an identity threshold of 99%, resulting in a collection of 1,104 unique contigs that constituted our curated viral contig database.

Quantitative read mapping. The complete genomes of 1,993 Caudovirales genomes downloaded from NCBI, a group of 154 phage genomes found in healthy humans from the Manrique study[39], contigs obtained from a Chicago cohort of IBD patients from the Norman study[21] and our curated viral contig database generated from healthy and colitic mice were used as mapping references using the bbmap algorithm. For the Caudovirales genomes and the curated viral contig database we used the following bbmap parameters: ambiguous=random, qtrim=lr, minid=0.97. For the Manrique and Norman data sets we decreased the identity threshold to minid=0.90, to account for increased heterogeneity between mouse and human samples[39]. Read-mapping statistics for each contig were generated by setting the 'scafstats' parameter. For the Caudovirales genomes, the Manrique data set and the Norman data set the total numbers of unambiguous reads per contig were tallied and averaged across animal treatments and time points. For the curated viral contig database, the percentages of ambiguous and unambiguous mapped reads were combined for each contig and compared between treatments and time points.

Confirming viral contigs, establishing level of similarity with existing viruses and defining host associations. We used a curated set of 25,281 VPFs[31] and the VirSorter database[32] to compare open reading frames within the 1,104 viral contigs from this study for viral protein similarity. VPFs are a recently developed tool yet to be widely used for virus discovery. VPFs have been benchmarked for high specificity for viruses and only in rare cases a VPF may include a prokaryotic host protein. Additionally, we obtained taxonomic information for viral contigs by using phage orthologous groups (POGs) that are recognized as taxon-specific signatures for virus classification[35]. For all Hidden Markov Models searches we used the algorithm hmmsearch with trusted cut offs from the hmmer web server[61] and an e-value below $1e^{-05}$. To cluster viral sequences, we used the public viral database IMG/VR composed of over 264 thousand metagenomic viral contigs and almost 8 thousand virus isolates[33]. We used the sequence-based classification framework described by Paez-Espino and colleagues[31] for grouping closely related viral sequences. The method relies on both nucleotide identity and the fraction of total alignment for pairwise comparisons of viral sequences. We used BLASTn in the Blast+ package[62] to find hits with an e-value cut off of $1e^{-10}$ and at least 50% identity across ≥50% of the shortest sequence length. We used three different computational methods to predict host–virus connections: (1) Grouping with isolated viruses, metagenomic viral contigs and host or taxonomic association using POGs. (2) Prokaryotic CRISPR spacer matching to viral contigs requiring at least 95% identity over the whole spacer length. To acquire CRISPR spacers from our experimental data set each animal's whole metagenome reads from each time point were combined and contigs were assembled using IDBA-UD. Binning of metagenome assemblies to retrieve genome bins was performed using MetaWatt[63] with default settings except avg read length=75, contig length threshold=500 and minimum bin seed size=5,000. Taxonomy was assigned to each bin based on MetaWatt taxonomy binning. CRISPR repeat patterns and spacers were extracted from bins using metaCRT[64] with the following settings minNR=4, minRL=19, maxRL=48. Spacers were extracted from CRISPRs and aligned to the curated viral contig database using the seal algorithm in the BBMap software suite with the k parameter set to 19. (3) Matching viral tRNA sequences to bacterial tRNAs. We identified tRNAs from our viral contig set using ARAGORN v1.2[65] and compared these against isolate microbial genomes in IMG/M[66] using BLASTn. Only hits with a 100% identity over 100% of the length were considered. There are multiple methods available for phage host identification from VLP metagenomes[67]. We chose to use CRISPRs and tRNAs to make host connections because CRISPR matching does not depend on pre-existing databases for sequence comparison and tRNAs are common features of phage genomes that provide host-specific information without the need for a reference phage genome.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. The DNA sequencing reads and the curated VLP contigs database associated with this study are deposited at the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession number PRJEB22710.

Received: 13 December 2017; Accepted: 27 June 2018;
Published online: 23 July 2018

References

1. Abraham, C. & Cho, J. H. Inflammatory bowel disease. *N. Engl. J. Med.* **361**, 2066–2078 (2009).
2. Molodecky, N. A. et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology* **142**, 46–54 (2012).
3. Hooper, L. V., Midtvedt, T. & Gordon, J. I. How host–microbial interactions shape the nutrient environment of the mammalian intestine. *Annu. Rev. Nutr.* **22**, 283–307 (2002).
4. Barthel, M. et al. Pretreatment of mice with streptomycin provides a *Salmonella enterica* serovar Typhimurium colitis model that allows analysis of both pathogen and host. *Infect. Immun.* **71**, 2839–2858 (2003).
5. Ivanov, I. I. et al. Specific microbiota direct the differentiation of IL-17-producing T-helper cells in the mucosa of the small intestine. *Cell Host Microbe* **4**, 337–349 (2008).
6. Mazmanian, S. K., Liu, C. H., Tzianabos, A. O. & Kasper, D. L. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* **122**, 107–118 (2005).
7. Reeves, A. E., Koenigsknecht, M. J., Bergin, I. L. & Young, V. B. Suppression of *Clostridium difficile* in the gastrointestinal tracts of germfree mice inoculated with a murine isolate from the family Lachnospiraceae. *Infect. Immun.* **80**, 3786–3794 (2012).
8. Sheehan, D., Moran, C. & Shanahan, F. The microbiota in inflammatory bowel disease. *J. Gastroenterol.* **50**, 495–507 (2015).
9. Belkaid, Y. & Hand, T. W. Role of the microbiota in immunity and inflammation. *Cell* **157**, 121–141 (2014).
10. Minot, S. et al. The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Res.* **21**, 1616–1625 (2011).
11. Reyes, A. et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).
12. Dinsdale, E. A. et al. Functional metagenomic profiling of nine biomes. *Nature* **452**, 629–632 (2008).
13. Oh, J. et al. Temporal stability of the human skin microbiome. *Cell* **165**, 854–866 (2016).
14. Feiner, R. et al. A new perspective on lysogeny: Prophages as active regulatory switches of bacteria. *Nat. Rev. Microbiol.* **13**, 641–650 (2015).
15. Winter, S. E., Lopez, C. A. & Baumler, A. J. The dynamics of gut-associated microbial communities during inflammation. *EMBO Rep.* **14**, 319–327 (2013).
16. Stecher, B. The roles of inflammation, nutrient availability and the commensal microbiota in enteric pathogen infection. *Microbiol. Spectr.* 3, MBP-0008-2014 (2015).
17. Banks, D. J., Lei, B. & Musser, J. M. Prophage induction and expression of prophage-encoded virulence factors in group A *Streptococcus* serotype M3 strain MGAS315. *Infect. Immun.* **71**, 7079–7086 (2003).
18. Duerkop, B. A., Clements, C. V., Rollins, D., Rodrigues, J. L. & Hooper, L. V. A composite bacteriophage alters colonization by an intestinal commensal bacterium. *Proc. Natl Acad. Sci. USA* **109**, 17621–17626 (2012).
19. Diard, M. et al. Inflammation boosts bacteriophage transfer between *Salmonella* spp. *Science* **355**, 1211–1215 (2017).
20. Wang, W. et al. Metagenomic analysis of microbiome in colon tissue from subjects with inflammatory bowel diseases reveals interplay of viruses and bacteria. *Inflamm. Bowel Dis.* **21**, 1419–1427 (2015).
21. Norman, J. M. et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447–460 (2015).
22. Lepage, P. et al. Dysbiosis in inflammatory bowel disease: A role for bacteriophages? *Gut* **57**, 424–425 (2008).
23. Barr, J. J. et al. Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc. Natl Acad. Sci. USA* **110**, 10771–10776 (2013).
24. Ostanin, D. V. et al. T cell transfer model of chronic colitis: Concepts, considerations, and tricks of the trade. *Am. J. Physiol. Gastrointest. Liver Physiol.* **296**, G135–146 (2009).
25. Powrie, F., Correa-Oliveira, R., Mauze, S. & Coffman, R. L. Regulatory interactions between $CD45RB^{high}$ and $CD45RB^{low}$ $CD4^{+}$ T cells are important for the balance between protective and pathogenic cell-mediated immunity. *J. Exp. Med.* **179**, 589–600 (1994).
26. Kang, D. W. et al. Microbiota Transfer Therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: An open-label study. *Microbiome* **5**, 10 (2017).
27. Berry, D. et al. Phylotype-level 16S rRNA analysis reveals new bacterial indicators of health state in acute murine colitis. *ISME J.* **6**, 2091–2106 (2012).
28. Hughes, E. R. et al. Microbial respiration and formate oxidation as metabolic signatures of inflammation-associated dysbiosis. *Cell Host Microbe* **21**, 208–219 (2017).
29. Lupp, C. et al. Host-mediated inflammation disrupts the intestinal microbiota and promotes the overgrowth of Enterobacteriaceae. *Cell Host Microbe* **2**, 204 (2007).

30. Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**, e3817 (2017).
31. Paez-Espino, D., Pavlopoulos, G. A., Ivanova, N. N. & Kyrpides, N. C. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nat. Protoc.* **12**, 1673–1682 (2017).
32. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: Mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
33. Paez-Espino, D. et al. IMG/VR: A database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res.* **45**, D457–D465 (2017).
34. Bailly-Bechet, M., Vergassola, M. & Rocha, E. Causes for the intriguing presence of tRNAs in phages. *Genome Res.* **17**, 1486–1495 (2007).
35. Kristensen, D. M. et al. Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J. Bacteriol.* **195**, 941–950 (2013).
36. Paez-Espino, D. et al. Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
37. Uchiyama, J. et al. In silico analysis of AHJD-like viruses, *Staphylococcus aureus* phages S24-1 and S13', and study of phage S24-1 adsorption. *Microbiologyopen* **3**, 257–270 (2014).
38. Prangishvili, D., Forterre, P. & Garrett, R. A. Viruses of the Archaea: A unifying view. *Nat. Rev. Microbiol.* **4**, 837–848 (2006).
39. Manrique, P. et al. Healthy human gut phageome. *Proc. Natl Acad. Sci. USA* **113**, 10400–10405 (2016).
40. Govind, R., Fralick, J. A. & Rolfe, R. D. Genomic organization and molecular characterization of *Clostridium difficile* bacteriophage PhiCD119. *J. Bacteriol.* **188**, 2568–2577 (2006).
41. Mesyanzhinov, V. V. et al. The genome of bacteriophage phiKZ of *Pseudomonas aeruginosa*. *J. Mol. Biol.* **317**, 1–19 (2002).
42. Esposito, D. et al. The complete nucleotide sequence of bacteriophage HP1 DNA. *Nucleic Acids Res.* **24**, 2360–2368 (1996).
43. Ferrenberg, S. et al. Changes in assembly processes in soil bacterial communities following a wildfire disturbance. *ISME J.* **7**, 1102–1111 (2013).
44. Lee, S. H., Sorensen, J. W., Grady, K. L., Tobin, T. C. & Shade, A. Divergent extremes but convergent recovery of bacterial and archaeal soil communities to an ongoing subterranean coal mine fire. *ISME J.* **11**, 1447–1459 (2017).
45. Winter, S. E. et al. Host-derived nitrate boosts growth of *E. coli* in the inflamed gut. *Science* **339**, 708–711 (2013).
46. Turnbaugh, P. J. et al. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027–1031 (2006).
47. Jiang, W. et al. Dysbiosis gut microbiota associated with inflammation and impaired mucosal immune function in intestine of humans with non-alcoholic fatty liver disease. *Sci. Rep.* **5**, 8096 (2015).
48. Duplessis, M. & Moineau, S. Identification of a genetic determinant responsible for host specificity in *Streptococcus thermophilus* bacteriophages. *Mol. Microbiol.* **41**, 325–336 (2001).
49. Holmfeldt, K., Middelboe, M., Nybroe, O. & Riemann, L. Large variabilities in host strain susceptibility and phage host range govern interactions between lytic marine phages and their *Flavobacterium* hosts. *Appl. Environ. Microbiol.* **73**, 6730–6739 (2007).
50. Duerkop, B. A., Huo, W., Bhardwaj, P., Palmer, K. L. & Hooper, L. V. Molecular basis for lytic bacteriophage resistance in Enterococci. *mBio* **7**, e01304–01316 (2016).
51. Mombaerts, P. et al. RAG-1-deficient mice have no mature B and T lymphocytes. *Cell* **68**, 869–877 (1992).
52. Cash, H. L., Whitham, C. V., Behrendt, C. L. & Hooper, L. V. Symbiotic bacteria direct expression of an intestinal bactericidal lectin. *Science* **313**, 1126–1130 (2006).
53. Kleiner, M., Hooper, L. V. & Duerkop, B. A. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genom.* **16**, 7 (2015).
54. Roux, S., Krupovic, M., Debroas, D., Forterre, P. & Enault, F. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* **3**, 130160 (2013).
55. Enault, F. et al. Phages rarely encode antibiotic resistance genes: A cautionary tale for virome analyses. *ISME J.* **11**, 237–247 (2017).
56. BBMap: short read aligner and other bioinformatic tools v36.99 (Bushnell, B., 2018); https://sourceforge.net/projects/bbmap/
57. phyloFlash v.2.0 (Gruber-Vodicka, H., Pruesse, E.A. & Seah, B., 2018); https://github.com/HRGV/phyloFlash
58. Quast, C. et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–596 (2013).
59. Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
60. 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).
61. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–37 (2011).
62. Camacho, C. et al. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
63. Strous, M., Kraft, B., Bisdorf, R. & Tegetmeyer, H. E. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front. Microbiol.* **3**, 410 (2012).
64. Rho, M., Wu, Y. W., Tang, H., Doak, T. G. & Ye, Y. Diverse CRISPRs evolving in human microbiomes. *PLoS Genet.* **8**, e1002441 (2012).
65. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).
66. Chen, I. A. et al. IMG/M: Integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2017).
67. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.* **40**, 258–272 (2016).

Acknowledgements

We would like to thank C. Boyd, T. Leal and K. Ruhn for assistance with animals and X. Dong and F. Santoriello for bioinformatics assistance. This work was supported by NIH R01DK070855 (L.V.H.), the Howard Hughes Medical Institute (L.V.H.), NIH K01DK102436 (B.A.D.), start-up funds from the University of Colorado School of Medicine (B.A.D.), the Government of Canada's Banting Postdoctoral Fellowship (M.K.) and the NC State Chancellor's Faculty Excellence Program Cluster on Microbiomes and Complex Microbial Communities (M.K.). This work was partly conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under contract number DE-AC02-05CH11231.

Author Contributions

B.A.D., M.K. and L.V.H. designed the study. B.A.D., M.K., D.P.E. and B.H. performed experiments. B.A.D., M.K., D.P.E. and B.B. performed bioinformatic analyses. B.A.D., M.K., D.P.E., W.Z., S.E.W., N.C.K. and L.V.H. analysed data. B.A.D., M.K. and L.V.H. wrote the paper with input from all of the authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-018-0210-y.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to B.A.D. or M.K. or L.V.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Corresponding author(s): Breck A. Duerkop, Manuel Kleiner & Lora V. Hooper

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

   Describe how sample size was determined.

   Sample sizes were chosen based on within experiment replicates of at least three mice per group in order to perform sample to sample statistical comparisons including student's t-tests, ANONVA analysis, and Kruskal-Wallis test follwed by a post hoc Dunn's test.

2. Data exclusions

   Describe any data exclusions.

   No data were excluded from the analysis

3. Replication

   Describe the measures taken to verify the reproducibility of the experimental findings.

   Reproducibility was performed on a per animal basis as determined by whole metagenome sequencing and microbiome analysis.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   When possible animals were either litter mates or were co-housed together following weaning. Mice were then randomly assigned to treatment or control groups.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Blinding was not used. Absolute metrics based on microbial quantity were used to determine differences between groups and thus could not be inherently biased by knowledge of treatment conditions.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	BBMap v36.99, phyloFlash v2.0, IDBA-UD v1.1.1, HmmerWeb version 2.19.0, MetaWatt v3.5.3, metaCRT v1.0, ARAGORN v1.2.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	There are no restrictions for the availability of unique materials from this study.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	Antibodies were used for cell sorting. Rat anti-mouse CD16/CD32, clone 2.4G2, (BD Biosciences) CD16/CD32 was detected on mouse spleen cells and demonstration of FCγR-mediated non-specific staining according to the manufacturer. Anti-mouse CD4-FITC, clone RM4-5 (Thermo Scientific). Antibody target was verified by Relative expression to ensure the antibody binds to the antigen stated. Anti-mouse CD45 RB-APC, clone C363-16A (Thermo Scientific). Antibody has been tested by flow cytometric analysis of mouse splenocytes according to the manufacturer's website.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	Mouse, C57BL6/J Rag1 -/-, 8 weeks of age, males Mouse, C57BL6/J, 6-8 weeks of age, males

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	This study did does not include human subjects