Diagnostics & Interactive Dynamics of the Colorectal Cancer Virome

Geoffrey D Hannigan[1], Melissa B Duhaime[2], Mack T Ruffin IV[3], Charlie C Koumpouras[1], and Patrick D Schloss[1,*]

[1]Department of Microbiology & Immunology, University of Michigan, Ann Arbor, Michigan, 48109

[2]Department of Ecology & Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, 48109

[3]Department of Family & Community Medicine, Pennsylvania State University Hershey Medical Center, Hershey, Pennsylvania, 17033

[*]To whom correspondence may be addressed.

Abstract

Viruses are assocaited with many human cancers, largely due to their mutagenic and functionally manipulative abilities. Despite this, cancer microbiome studies have almost exclusively focused on bacteria instead of viruses. We began evaluating the cancer virome by focusing on colorectal cancer, a primary cause of morbidity and mortality throughout the world, and a cancer linked to altered colonic bacterial community compositions while the virome role remains unknown. We used 16S rRNA gene, whole shotgun metagenomic, and purified virus metagenomic sequencing of stool to evaluate the differences in human colorectal cancer virus and bacterial community composition. Through random forest modeling we identified differences in the healthy and colorectal cancer virome. The cancer-associated virome consisted primarily of

temperate bacteriophages that were also bacteria-virus community network hubs. These results provide foundational evidence that bacteriophage communities are associated with colorectal cancer and likely impact cancer progression by altering the bacterial host communities.

Introduction

Due to their mutagenic abilities and propensity for functional manipulation, human viruses are strongly associated with, and in many cases cause, cancer[1–4]. Because bacteriophages (i.e. viruses that specifically infect bacteria) are crucial for bacterial community stability and composition[5–7] and have been implicated as oncogenic agents[8–11], bacteriophages have the potential to indirectly impact cancer. The gut virome (i.e. the virus community of the gut) therefore has the potential to impact health and disease. Altered human virome composition and diversity have been identified in diseases including periodontal disease[12], HIV[13], cystic fibrosis[14], antibiotic exposure[15,16], urinary tract infections[17], and inflammatory bowel disease[18]. The strong association of bacterial communities with colorectal cancer and the precedent for the virome to impact other human diseases suggest that colorectal cancer may be associated with altered virus communities.

Colorectal cancer is the second leading cause of cancer-related deaths in the United States[19]. The US National Cancer Institute estimates over 1.5 million Americans were diagnosed with colorectal cancer in 2016 and over 500,000 Americans died from the disease[19]. Growing evidence suggests that an important component of colorectal cancer etiology may be perturbations in the colonic bacterial community[8,10,11,20,21]. Work in this area has led to a proposed disease model in which bacteria colonize the colon, develop biofilms, promote inflammation, and enter an oncogenic synergy with the cancerous human cells[22]. This association also has allowed researchers to leverage bacterial community signatures as biomarkers to provide accurate, noninvasive colorectal cancer detection from stool[8,23,24]. While an understanding of colorectal cancer bacterial communities has proven fruitful both for disease classification and for identifying the underlying disease etiology, bacteria are only a subset of the colon microbiome. Viruses are another important component of the colon microbial community that have yet to be studied in the context of colorectal cancer. We evaluated disruptions in virus and bacterial community composition in a human cohort whose stool was sampled at the three relevant stages of cancer development: healthy, adenomatous, and cancerous.

Colorectal cancer progresses in a stepwise process that begins when healthy tissue develops into a precancerous polyp (i.e., adenoma) in the large intestine[25]. If not removed, the adenoma may develop into a cancerous lesion that can invade and metastasize, leading to severe illness and death. Progression to cancer can be prevented when precancerous adenomas are detected and removed during routine screening[26,27]. Survival for colorectal cancer patients may exceed 90% when the lesions are detected early and removed[26]. Thus, work that aims to facilitate early detection and prevention of progression beyond early cancer stages has great potential to inform therapeutic development.

Here we address the knowledge gap of whether virus community composition is altered in colorectal cancer and, if it is, how those differences might impact cancer progression and severity. We also aimed to evaluate the virome's potential for use as a diagnostic biomarker. The implications of this study are threefold. *First*, this work supports a biological role for the virome in colorectal cancer development and suggests that more than the bacterial members of the associated microbial communities are involved in the process. *Second*, we present a supplementary, or even alternative, virus-based approach for classification modeling of colorectal cancer using stool samples. *Third*, we provide initial support for the importance of studying the virome as a component of the microbiome ecological network, especially in cancer.

Sample Collection and Processing

Our study cohort consisted of 90 human subjects, 30 of whom had healthy colons, 30 of whom had adenomas, and 30 of whom had carcinomas **(Figure S1)**. Half of each stool sample was used to sequence the bacterial communities using both 16S rRNA gene and shotgun sequencing techniques. The 16S rRNA gene sequencing was performed for a previous study, and the sequences were re-analyzed using contemporary methods[8]. The other half of each stool sample was purified for virus like particles (VLPs) before genomic DNA extraction and shotgun metagenomic sequencing. In the VLP purification, cells were disrupted and extracellular DNA degraded **(Figure S1)** to allow the exclusive analysis of viral DNA within virus capsids. In this manner, the *extracellular virome* of encapsulated viruses was targeted.

Each extraction was performed with a blank buffer control to detect contaminants from reagents or other unintentional sources. Only one of the nine controls contained detectable DNA at a minimal concentration of 0.011 ng/µl, thus providing evidence of the enrichment and purification of VLP genomic DNA over potential contaminants **(Figure S2 A)**. As expected, these controls yielded few sequences and were almost entirely removed while rarefying the datasets to a common number of sequences **(Figure S2 B)**. The high quality phage and bacterial sequences were assembled into highly covered contigs longer than 1 kb **(Figure S3)**. Because contigs represent genome fragments, we further clustered related bacterial contigs into operational genomic units (OGUs) and viral contigs into operational viral units (OVUs) **(Figure S3 - S4)** to approximate organismal units.

Unaltered Diversity in Colorectal Cancer

Microbiome and disease associations are often described as being of an altered diversity (i.e., "dysbiotic"). Therefore, we first evaluated the influence of colorectal cancer on virome OVU diversity. We evaluated differences in communities between disease states using the Shannon diversity, richness, and Bray-Curtis metrics. We observed no significant alterations in either Shannon diversity or richness in the diseased states as compared to the healthy state **(Figure S5 C-D)**. There was no statistically significant clustering of the disease groups (ANOSIM p-value = 0.4, **Figure S5**). Notably, there was a significant difference between the few blank controls that remained after rarefying the data and the other study groups (ANOSIM p-value < 0.001, **Figure S6)**, further supporting the quality of the sample set. In summary, standard alpha and beta diversity metrics were insufficient for capturing virus community differences between disease states **(Figure S5)**. This is consistent with what has been observed when the same metrics were applied to 16S rRNA sequenced and metagenomic samples[8,23,24] and points to the need for alternate approaches to detect the impact of colorectal cancer disease state on these communities.

Virome Composition in Colorectal Cancer

As opposed to the diversity metrics discussed above, OTU-based relative abundance profiles generated from 16S rRNA gene sequences are effective feature sets for classifying stool samples as originating from individuals with healthy, adenomatous, or cancerous colons[8,23]. The exceptional performance of bacteria in these classification models supports a role for bacteria in colorectal cancer. We built off of these findings by evaluating the ability of virus community signatures to classify stool samples and compared their performance to models built using bacterial community signatures.

To identify the altered virus communities associated with colorectal cancer, we built and tested random forest models for classifying stool samples as belonging to individuals with either cancerous or healthy colons. We confirmed that our bacterial 16S rRNA gene model replicated the performance of the original report which used logit models instead of random forest models **(Figure 1 A)**[8]. We then compared the bacterial OTU model to a model built using OVU relative abundances. The viral model performed as well as the bacterial model (corrected p-value = 0.4), with the viral and bacterial models achieving mean area under the curve (AUC) values of 0.793 and 0.796, respectively **(Figure 1 A - B)**. To evaluate the ability of both bacterial and viral biomarkers to classify samples, we built a combined model that used both bacterial and viral community data. The combined model yielded a modest but statistically significant performance improvement beyond the viral (corrected p-value = 0.002) and bacterial (corrected p-value = 0.002) models, yielding an AUC of 0.816 **(Figure**

1 A - B). The combined features from the virus and bacterial communities improved our ability to classify stool as belonging to individuals with cancerous colons.

To determine the advantage of viral metagenomic methods over bacterial metagenomic methods, we compared the viral model to a model built using OGU relative abundance profiles from bacterial metagenomic shotgun sequencing data. This model performed worse than the other models (mean AUC = 0.505) **(Figure 1 A - B)**. Because the coverage provided by the metagenomic sequencing was not as deep as the equivalent 16S rRNA gene sequencing, we attempted to compare the approaches at a common sequencing depth. This investigation revealed that the bacterial 16S rRNA gene model was strongly driven by sparse and low abundance OTUs **(Figure S7)**. Removal of OTUs with a median abundance of zero resulted in the removal of six OTUs, and a loss of model performance down to what was observed in the metagenome-based model **(Figure S7 A)**. The majority of these OTUs had a relative abundance lower than 1% across the samples **(Figure S7 B)**. Although the features in the viral model also were of low abundance **(Figure S9 F)**, the coverage was sufficient for high model performance, likely because viral genomes are orders of magnitude smaller than bacterial genomes. Thus, the targeted 16S rRNA gene sequencing approach, which represented only a fraction of the bacterial metagenomic sequencing depth, was more effective for detecting colorectal cancer in stool samples. Despite the recent loss of enthusiasm for 16S rRNA gene sequencing in favor of shotgun metagenomic techniques, 16S rRNA gene sequencing is still a superior methodological approach for some important applications.

The association between the bacterial and viral communities and colorectal cancer was driven by a few important microbes. *Fusobacterium* was the primary driver of the bacterial association with colorectal cancer, which is consistent with its previously described oncogenic potential **(Figure 1 C)**[22]. The virome signature also was driven by a few OVUs, suggesting a role for these viruses in tumorigenesis **(Figure 1 D)**. The identified viruses were bacteriophages, belonging to *Siphoviridae*, *Myoviridae*, and “unclassified” phage taxa.

Many of the important viruses were unidentifiable (denoted "unknown"). This is common in viromes across habitats; studies have reported as much as 95% of virus sequences belonging to unknown genomic units[14,28–30]. When the bacterial and viral community signatures were combined, both bacterial and viral organisms drove the community association with cancer **(Figure 1 E)**.

Phage Influence Between CRC Stages

Because previous work has identified shifts in which bacteria were most important at different stages of colorectal cancer[8,20,22], we explored whether shifts in the relative influence of specific phages could be detected between healthy, adenomatous, and cancerous colons. We evaluated community shifts between the two disease stage transitions (healthy to adenomatous and adenomatous to cancerous) by building random forest models to compare only the diagnosis groups around the transitions. While bacterial OTU models performed equally well for all disease class comparisons, the virome model performances differed **(Figure S8 A-B)**. Like bacteria **(Figure S8 F-H)**, different virome members were important between the healthy to adenomatous and adenomatous to cancerous stages **(Figure S8 C-E)**.

After evaluating our ability to classify samples between two disease states, we performed a three-class random forest model including all disease states. The 16S rRNA gene model yielded a mean AUC of 0.771 and outperformed the viral community model, which yielded a mean AUC of 0.699 (p-value < 0.001, **Figure S9 A-C**). The microbes important for the healthy versus cancer and healthy versus adenoma models were also important for the three-class model **(Figure S9 D-E)**. The most important bacterium in the two and three class models was the same *Fusobacterium* (OTU 4) **(Figure 1 C, Figure S9 D)**. The viruses most important to the three-class model were identified as bacteriophages **(Figure 1 D, Figure S9 E)**, but not all important OVUs were of increased abundance in the diseased state **(Figure S9 F)**.

Phage Dominance in CRC Virome

Differences in the colorectal cancer virome could have been driven directly by eukaryotic viruses or indirectly by bacteriophages acting through their bacterial hosts. To better understand the types of viruses that were important for colorectal cancer, we identified the virome OVUs as being similar to either eukaryotic viruses or bacteriophages. The most important viruses to the classification model were identified as bacteriophages (**Figure S9**). Overall, we were able to identify 78.8% of the OVUs as known viruses, and 93.8% of those viral OVUs aligned to bacteriophage reference genomes. It is important to note that this could have been influenced by our methodological biases against enveloped viruses (more common of eukaryotic viruses than bacteriophage), due to chloroform and DNase treatment for purification.

We evaluated whether the phages in the community were primarily lytic (i.e. obligately lyse their hosts after replication) or temperate (i.e. able to integrate into their host's genome to form a lysogen, and subsequently transition to a lytic mode). We accomplished this by identifying three markers for temperate phages in the OVU representative sequences: 1) presence of phage integrase genes, 2) presence of known prophage genes, according the the ACLAME (A CLAssification of Mobile genetic Elements) database, and 3) nucleotide similarity to regions of bacterial genomes[29,31,32]. We found that the majority of the phages were temperate and that the overall fraction of temperate phages remained consistent throughout the healthy, adenomatous, and cancerous stages **(Figure S10 E)**. These findings were consistent with previous reports suggesting the gut virome is primarily composed of temperate phages[13,18,31,33].

Community Context of Influential Phages

Because the link between colorectal cancer and the virome was driven by bacteriophages, we hypothesized that the influential phages were primarily predators of the influential bacteria, and thus influenced their relative abundance through predation. If this hypothesis were true, we would expect a correlation between the relative abundances of influential bacteria and phages. Instead, we observed a strikingly low correlation between bacterial and phage relative abundances **(Figure 2 A,C)**. Overall, there was an absence of correlation between the most influential OVUs and bacterial OTUs **(Figure 2 B)**. This evidence supported our null hypothesis that the influential phages were not primarily predators of influential bacteria.

Given these findings, we hypothesized that the most influential phages were acting by infecting a wide range of bacteria in the overall community, instead of just the influential bacteria. In other words, we hypothesized that the influential bacteriophages were community hubs (i.e. central members) within the bacteria and phage interactive network. We investigated the potential host ranges of all phage OVUs using a previously developed random forest model that relies on sequence features to predict which phages infected which bacteria in the community **(Figure 3 A)**[34]. The predicted interactions were then used to identify phage community hubs. We calculated the alpha centrality (i.e. measure of importance in the ecological network) of each phage OVU's connection to the rest of the network. The phages with high centrality values were defined as community hubs. Next, the centrality of each OVU was compared to its importance in the colorectal cancer classification model. Phage OVU centrality was significantly and positively correlated with importance to the disease model (p-value = 0.02, R = 0.14), suggesting that phages important in driving colorectal cancer also were more likely to be community hubs **(Figure 3 B)**. Together these findings supported our hypothesis that influential phages were hubs within their microbial communities and had broad host ranges.

Model for Virome & Cancer Progression

Because of their propensity for mutagenesis and capacity for modulating their host functionality, many viruses are oncogenic[1–4]. Some bacteria also have oncogenic properties, suggesting that bacteriophages may play an indirect role in promoting carcinogenesis by influencing bacterial community composition and dynamics[8–10]. Despite their carcinogenic potential and the strong association between bacteria and colorectal cancer, a mechanistic link between virus colorectal communities and colorectal cancer has yet to be evaluated. Here we show that, like colonic bacterial communities, the colon virome was altered in patients with colorectal cancer relative to those with healthy colons. Our findings support a working hypothesis for oncogenesis by phage-modulated bacterial community composition.

We have begun to delineate the role the colonic virome plays in colorectal cancer **(Figure 4 A)**. We found that basic diversity metrics of alpha diversity (richness and Shannon diversity) and beta diversity (Bray-Curtis dissimilarity) were insufficient for identifying virome community differences between healthy and cancerous states. By implementing a more sophisticated machine learning approach (random forest classification), we detected strong associations between the colon virus community composition and colorectal cancer. The colorectal cancer virome was composed primarily of bacteriophages. These phage communities were not exclusively predators of the most influential bacteria, as demonstrated by the lack of correlation between the abundances of the bacterial and phage populations. Instead, we identified influential phages as being community hubs, suggesting phages influence cancer by altering the greater bacterial community instead of directly modulating the influential bacteria. Our previous work has shown that modifying colon bacterial communities alters colorectal cancer progression and tumor burden in mice[10,20]. This provides a precedent for phage indirectly influencing colorectal cancer progression by altering the bacterial community composition. Overall, our data support a model in which the bacteriophage community modulates the bacterial community, and through those interactions indirectly influences the bacteria driving colorectal cancer progression **(Figure 4 A)**. Although our evidence suggested phages indirectly influenced colorectal cancer development, we were not able to rule out the role of phages directly interacting with the human host[35,36].

In addition to modeling the potential connections between virus communities, bacterial communities, and colorectal cancer, we also used our data and existing knowledge of phage biology to develop a working hypothesis for the mechanisms by which this may occur. This was done by incorporating our findings into the current model for colorectal cancer development **(Figure 4 B)**[22]. We hypothesize that the process begins with broadly infectious phages in the colon lysing and thereby disrupting the existing bacterial communities. This shift opens novel niche space that enabled opportunistic bacteria (such as *Fusobacterium nucleatum*) to colonize. Once the initial influential founder bacteria establish themselves in the epithelium, secondary opportunistic bacteria are able to adhere to the founders, colonize, and establish a biofilm. Phages may play a role in biofilm dispersal and growth by lysing bacteria within the biofilm, a process important for effective biofilm growth[37]. The oncogenic bacteria may then be able to transform the epithelial cells and disrupt tight junctions to infiltrate the epithelium, thereby initiating an inflammatory immune response. As the adenomatous polyps developed and progressed towards carcinogenesis, we observed a shift in the phages and bacteria whose relative abundances were most influential. As the bacteria enter their oncogenic synergy with the epithelium, we conjecture that the phages continue mediating biofilm dispersal. This process would thereby support the colonized oncogenic bacteria by lysing competing cells and releasing nutrients to other bacteria in the form of cellular lysates. In addition to highlighting the likely mechanisms by which the colorectal cancer virome is interacting with the bacterial communities this model will guide future research investigations of the role the virome plays colorectal cancer.

Conclusions

In addition to the diagnostic ramifications for understanding the colorectal cancer microbiome, our findings suggest that viruses, while understudied and currently under-appreciated in the human microbiome, are likely to be an important contributor to human disease. Viral community dynamics have the potential to provide an abundance of information to supplement those of bacterial communities. Evidence has suggested that the virome is a crucial component to the microbiome and that bacteriophages are important players. Bacteriophage and bacterial communities cannot maintain stability and co-evolution without one another[6,38]. Not only is the human virome an important element to consider in human health and disease[12–18], but our findings support that it is likely to have a significant impact on cancer etiology and progression.

Materials and Methods

This study was approved by the University of Michigan Institutional Review Board and all subjects provided informed consent. All study sequences are available on the NCBI Sequence Read Archive under the BioProject ID `PRJNA389927`. All associated source code is available at the following GitHub repository. Further detailed methods can be found in the supplemental methods.

https://github.com/SchlossLab/Hannigan_CRCVirome_Nature_2017

Acknowledgments

The authors thank the Schloss lab members for their underlying contributions, and the Great Lakes-New England Early Detection Research Network for providing the fecal samples that were used in this study. GD Hannigan was supported in part by the Molecular Mechanisms in Microbial Pathogenesis Training Program (T32 AI007528). PD Schloss was supported by funding from the National Institutes of Health (P30DK034933). MT Ruffin was supported by funding from the National Institutes of Health (5U01CA86400). The authors declare no competing interests.

Figures

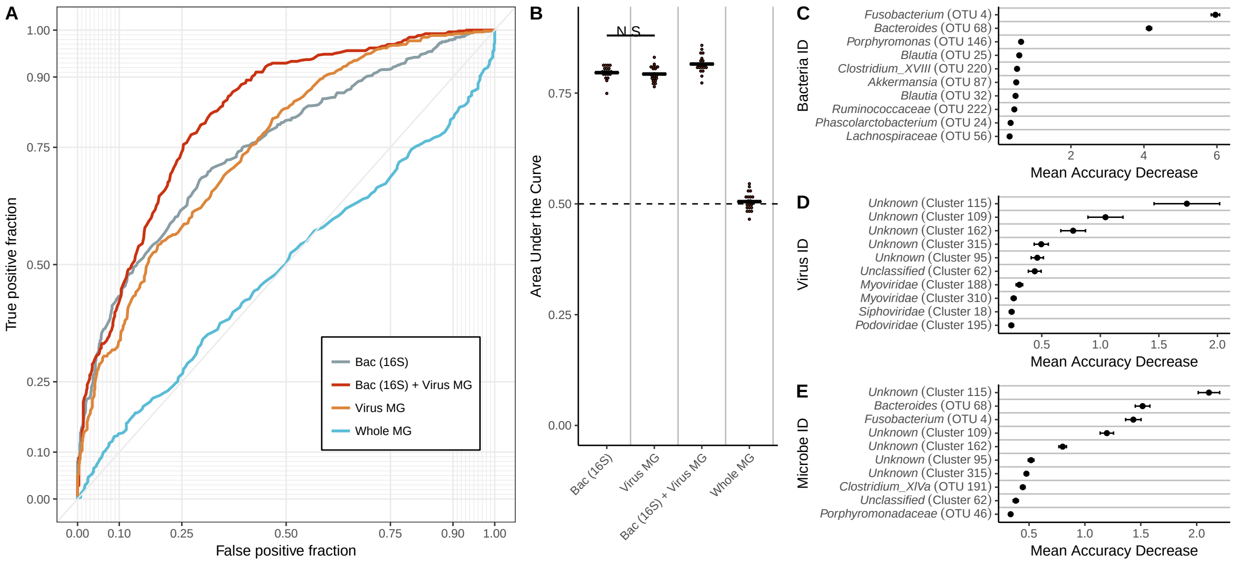


Figure 1: *Results from healthy vs cancer classification models built using virome signatures, bacterial 16S rRNA gene sequence signatures, whole metagenomic signatures, and a combination of virome and 16S rRNA gene sequence signatures. A) An example ROC curve for visualizing the performance of each of the models for classifying stool as coming from either an individual with a cancerous or healthy colon. B) Quantification of the AUC variation for each model, and how it compared to each of the other models based on 15 iterations. A pairwise Wilcoxon test with a false discovery rate multiple hypothesis correction demonstrated that all models are significantly different from each other (p-value $<$ 0.01). C) Mean decrease in accuracy (measurement of importance) of each operational taxonomic unit within the 16S rRNA gene classification model when removed from the classification model. Mean is represented by a point, and bars represent standard error. D) Mean decrease in accuracy of each operational virus unit in the virome classification model. E) Mean decrease in accuracy of each operational genomic unit and operational taxonomic unit in the model using both 16S rRNA gene and virome features.*

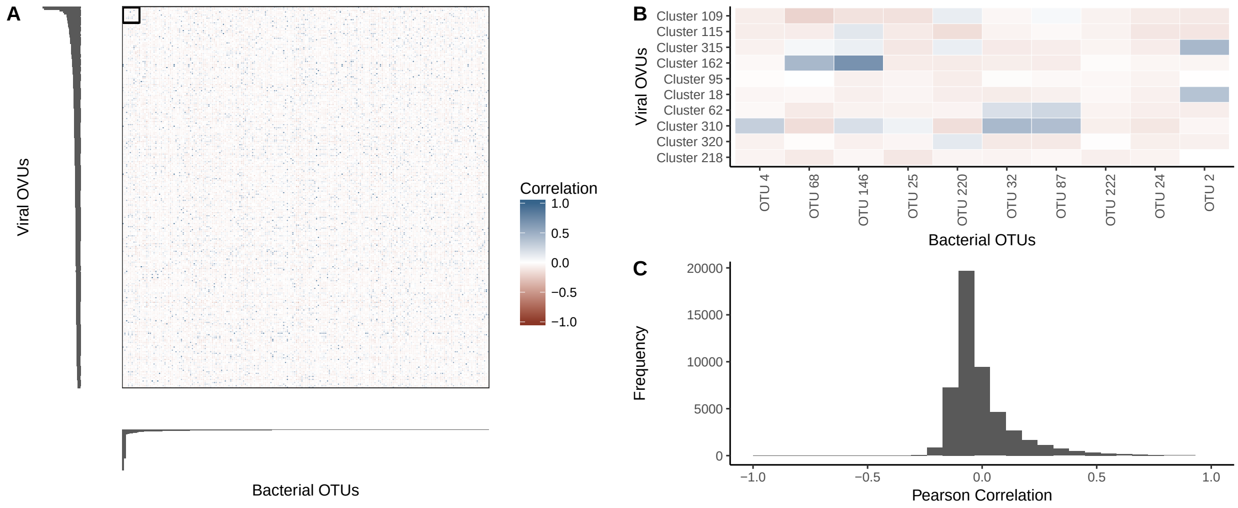


Figure 2: *Relative abundance correlations between bacterial OTUs and virome OVUs. A) Pearson correlation coefficient values between all bacterial OTUs (x-axis) and viral OVUs (y-axis) with blue being positively correlated and red being negatively correlated. Bar plots indicate the viral (left) and bacterial (bottom) operational unit importance in their colorectal cancer classification models, such that the most important units are in the top left corner. B) Magnification of the boxed region in pannel (A), highlighting the correlation between the most important bacterial OTUs and virome OVUs. The most important operational units are in the top left corner of the heatmap, and the correlation scale is the same as pannel (A). C) Histogram quantifying the frequencies of Pearson correlation coefficients between all bacterial OTUs and virome OVUs.*

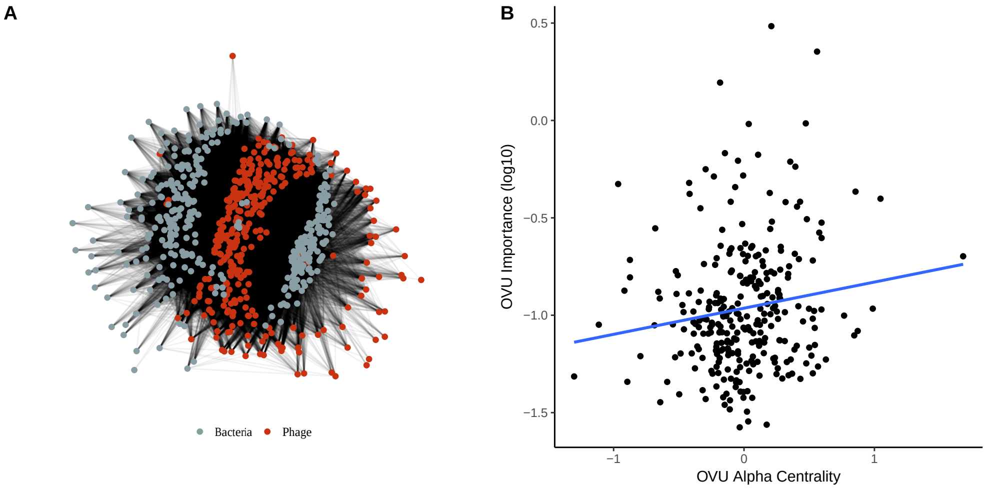
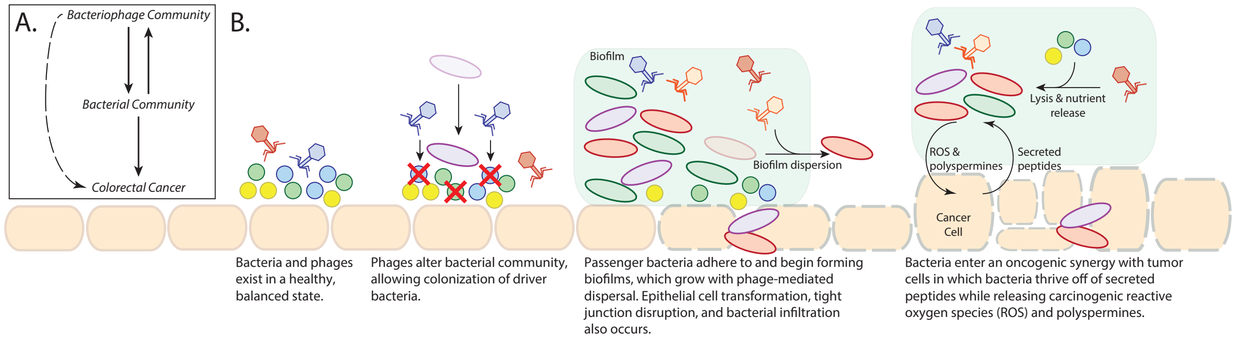


Figure 3: *Community network analysis utilizing predicted interactions between bacteria and phage operational genomic units. A) Visualization of the community network for our colorectal cancer cohort. B) Scatter plot illustrating the correlation between importance (mean decrease in accuracy) and the degree of centrality for each OVU. A linear regression line was fit to illustrate the correlation (blue) which was found to be statistically significantly and weakly correlated (p-value = 0.0173, R = 0.14).*



Bacteria and phages exist in a healthy, balanced state.

Phages alter bacterial community, allowing colonization of driver bacteria.

Passenger bacteria adhere to and begin forming biofilms, which grow with phage-mediated dispersal. Epithelial cell transformation, tight junction disruption, and bacterial infiltration also occurs.

Bacteria enter an oncogenic synergy with tumor cells in which bacteria thrive off of secreted peptides while releasing carcinogenic reactive oxygen species (ROS) and polyspermines.

Figure 4: *Final model and working hypothesis from this study. A) Basic model illustrating the connections between the virome, bacterial communities, and colorectal cancer. B) Working hypothesis of how the bacteriophage community is associated with colorectal cancer and the associated bacterial community.*

Supplemental Materials

Materials & Methods

Analysis Source Code & Data Availability

All study sequences are available on the NCBI Sequence Read Archive under the BioProject ID `PRJNA389927`.

All associated source code is available at the following GitHub repository:

https://github.com/SchlossLab/Hannigan_CRCVirome_Nature_2017

Study Design and Patient Sampling

This study was approved by the University of Michigan Institutional Review Board and all subjects provided informed consent. Design and sampling of this sample set have been reported previously[8]. Briefly, whole evacuated stool was collected from patients who were 18 years of age or older, able to provide informed consent, have had colonoscopy and histologically confirmed colonic disease status, had not had surgery, had not had chemotherapy or radiation, and were free of known co-morbidities including HIV, chronic viral hepatitis, HNPCC, FAP, and inflammatory bowel disease. Samples were collected from four geographic locations: Toronto (Ontario, Canada), Boston (Massachusetts, USA), Houston (Texas, USA), and Ann Arbor (Michigan, USA). Ninety patients were recruited to the study, thirty of which were designated healthy, thirty with detected adenomas, and thirty with detected carcinomas.

16S rRNA Gene Sequence Data Acquisition & Processing

The 16S rRNA gene sequences associated with this study were previously reported[8]. Sequence (fastq) and metadata files were downloaded from:

http://www.mothur.org/MicrobiomeBiomarkerCRC

The 16S rRNA gene sequences were analyzed as described previously, relying on the mothur software package (v1.37.0)[39,40]. Briefly, the sequences were de-replicated, aligned to the SILVA database[41], screened for chimeras using UCHIME[42], and binned into operational taxonomic units (OTUs) using a 97% similarity threshold. Abundances were normalized for uneven sequencing depth by randomly sub-sampling to 10,000 sequences, as previously reported[23].

Whole Metagenomic Library Preparation & Sequencing

DNA was extracted from stool samples using the PowerSoil-htp 96 Well Soil DNA Isolation Kit (Mo Bio Laboratories) using an EPMotion 5075 pipetting system. Purified DNA was used to prepare a shotgun sequencing library using the Illumina Nextera XT library preparation kit according to the standard kit protocol, including 12 cycles of limited cycle PCR. The tagmentation time was increased from five minutes to ten minutes to improve DNA fragment length distribution. The library was sequenced using one lane of the Illumina HiSeq4000 platform and yielded 125 bp paired end reads.

Virus Metagenomic Library Preparation & Sequencing

Genomic DNA was extracted from purified virus-like particles (VLPs) from stool samples, using a modified version of a previously published protocol[29,31,43,44]. Briefly, an aliquot of stool (~0.1 g) was resuspended in SM buffer (Crystalgen; Catalog #: 221-179) and vortexed to facilitate resuspension. The resuspended stool was centrifuged to remove major particulate debris then filtered through a 0.22-µm filter to remove smaller contaminants. The filtered supernatant was treated with chloroform for ten minutes with gentle shaking, so as to lyse contaminating cells including bacteria, human, fungi, etc. The exposed genomic DNA from the lysed cells was degraded by treating the samples with 5U of DNase for one hour at 37C. DNase was deactivated by incubating the sample at 75C for ten minutes. The DNA was extracted from the purified virus-like particles (VLPs) using the Wizard PCR Purification Preparation Kit (Promega). Disease classes were staggered across purification runs to prevent run variation as a confounding factor. As for whole community metagenomes, purified DNA was used to prepare a shotgun sequencing library using the Illumina Nextera XT preparation kit according to the standard kit protocol. The tagmentation time was increased from five minutes to ten minutes to improve DNA fragment length distribution. The PCR cycle number was increased from twelve to eighteen cycles to address the low biomass of the samples, as has been described previously[29]. The library was sequenced using one lane of the Illumina HiSeq4000 platform and yielded 125 bp paired end reads.

Metagenome Quality Control

Both the viral and whole community metagenomic sample sets were subjected to the same quality control procedures. The sequences were obtained as de-multiplexed fastq files and subjected to 5' and 3' adapter trimming using the CutAdapt program (v1.9.1) with an error rate of 0.1 and an overlap of 10[45]. The FastX toolkit (v0.0.14) was used to quality trim the reads to a minimum length of 75 bp and a minimum quality score of 30[46]. Reads mapping to the human genome were removed using the DeconSeq algorithm (v0.4.3) and default parameters[47].

Contig Assembly & Abundance

Contigs were assembled using paired end read files that were purged of sequences without a corresponding pair (e.g. one read removed due to low quality). The Megahit program (v1.0.6) was used to assemble contigs for each sample using a minimum contig length of 1000 bp and iterating assemblies from 21-mers to 101-mers by 20[48]. Contigs from the virus and whole metagenomic sample sets were concatenated within their respective groups. Abundance of the contigs within each sample was calculated by aligning sequences back to the concatenated contig files using the bowtie2 global aligner (v2.2.1), with a 25 bp seed length and an allowance of one mismatch[49]. Abundance was corrected for contig reference length and the number of contigs included in each operational genomic unit. Abundance was also corrected for uneven sampling depth by randomly sub-sampling virome and whole metagenomes to 1,000,000 and 500,000 reads, respectively, and by removing samples with fewer total reads than the threshold. Thresholds were set for maximizing sequence information while minimizing numbers of lost samples.

Operational Genomic Unit Classification

Much like operational taxonomic units (OTUs) are used as an operational definition of similar 16S rRNA gene sequences, we defined closely related bacterial contig sequences as operational genomic units (OGUs) and virus contigs as operational viral units (OVUs) in the absence of taxonomic identity. OGUs and OVUs were defined with the CONCOCT algorithm (v0.4.0) which bins related contigs by similar tetra-mer and co-abundance profiles within samples using a variational Bayesian approach[50]. CONCOCT was used with a length threshold of 1000 bp for virus contigs and 2000 bp for bacteria.

Diversity

Alpha and beta diversity were calculated using the operational viral unit abundance profiles for each sample. Sequences were rarefied to 100,000 sequences. Samples with less than the cutoff were removed from the analysis. Alpha diversity was calculated using the Shannon diversity and richness metrics. Beta diversity was calculated using the Bray-Curtis metric (mean of 25 random sub-sampling iterations), and the statistical significance between the disease state clusters was assessed using an analysis of similarity (ANOSIM) with a post-hoc multivariate Tukey test. All diversity calculations were performed in R using the Vegan package[51].

Classification Modeling

Classification modeling was performed in R using the Caret package[52]. OTU, OVU, and OGU abundance data was preprocessed by removing features (OTUs, OVUs, and OGUs) that were present in less than thirty of the samples. This served both as an effective feature reduction technique and made the calculations computationally feasible. The binary random forest model was trained using the Area Under the receiver operating characteristic Curve (AUC) and the three-class random forest model was trained using the mean AUC. Both were validated using five-fold cross validation. Each training set was repeated five times, and the model was tuned for mtry values. For consistency and accurate comparison between feature groups (e.g., bacteria, viruses), the sample model parameters were used for each group. The maximum AUC during training was recorded across twenty iterations of each group model to test the significance of the differences between feature set performance. Statistical significance was evaluated using a Wilcoxon test between two categories, or a pairwise Wilcoxon test with Bonferroni corrected p-values when comparing more than two categories.

Taxonomic Identification of Operational Genomic Units

Operational viral units (OVUs) were taxonomically identified using a reference database consisting of all bacteriophage and eukaryotic virus genomes present in the European Nucleotide Archives. The longest contiguous sequence in each operational genomic unit was used as a representative sequence for classification, as described previously[53]. Each representative sequence was aligned to the reference genome database using the tblastx alignment algorithm (v2.2.27) and a strict similarity threshold (e-value $<$ 1e-25)[54]. Annotation was interpreted as phage, eukaryotic virus, or unknown.

Ecological Network Analysis & Correlations

The ecological network of the bacterial and phage operational genomic units was constructed and analyzed as previously described[34]. Briefly, a random forest model was used to predict interactions between bacterial and phage genomic units, and those interactions were recorded in a graph database using *neo4j* graph databasing software (v2.3.1). The degree of phage centrality was quantified using the alpha centrality metric in the igraph CRAN package. A Spearman correlation was performed between model importance and phage centrality scores.

Phage Replication Style Identification

Phage OVU replication mode was predicted using methods described previously[29,31,32]. Briefly, we identified temperate OVUs as representative contigs containing at least one of three genomic markers: 1) phage integrase genes, 2) prophage genes from the ACLAME database, or 3) genomic similarity to bacterial reference genomes. Integrase genes were identified in phage OVU representative contigs by aligning the contigs to a reference database of all known phage integrase genes from the Uniprot database (Uniprot search term: “organism:phage gene:int NOT putative”). Prophage genes were identified in the same way, using the ACLAME set of reference prophage genes. In both cases, the blastx algorithm was used with an e-value threshold of 10e-5. Representative contigs were also identified as potential lysogenic phages by having a high genomic similarity to bacterial genomes. To accomplish this, representative phage contigs were aligned to the European Nucleotide Archive bacterial genome reference set using the blastn algorithm (e-value $<$ 10e-25).

Supplemental Figures

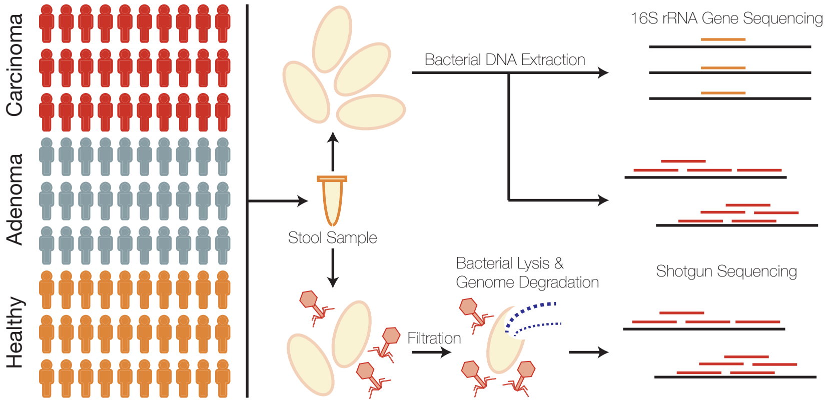


Figure S1: *Cohort and sample processing outline. Thirty subject stool samples were collected from healthy, adenoma (pre-cancer), and carcinoma (cancer) patients. Stool samples were split into two aliquots, the first of which was used for bacterial sequencing and the second which was used for virus sequencing. Bacterial sequencing was done using both 16S rRNA amplicon and whole metagenomic shotgun sequencing techniques. Virus samples were purified for viruses using filtration and a combination of chloroform (bacterial lysis) and DNase (exposed genomic DNA degradation). The resulting encapsulated virus DNA was sequenced using whole metagenomic shotgun sequencing.*

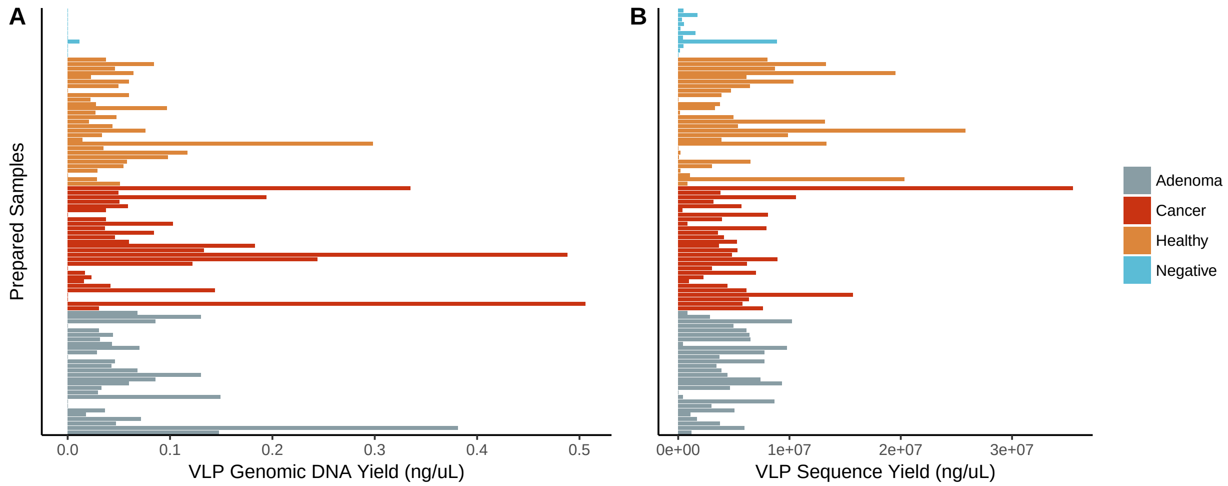


Figure S2: *Basic Quality Control Metrics. A) VLP genomic DNA yield from all sequenced samples. Each bar represents a sample which is grouped and colored by its associated disease group. B) Sequence yield following quality control including quality score filtering and human decontamination.*

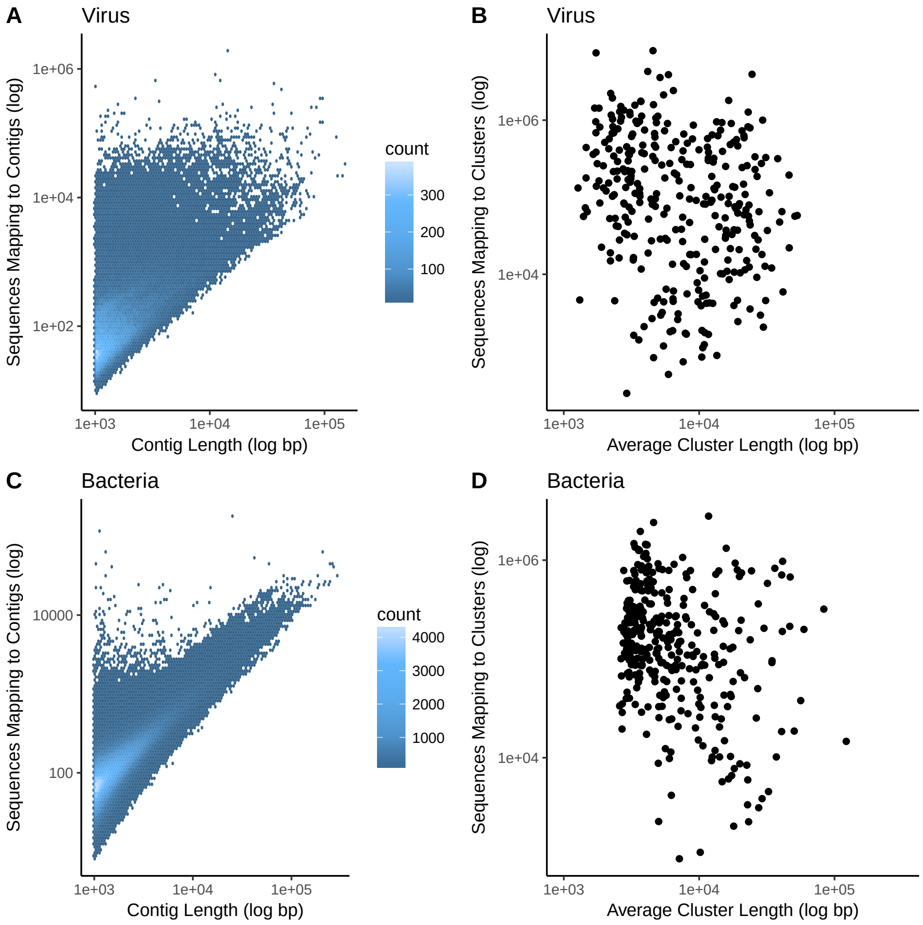


Figure S3: *Length and coverage statistics. A) Heated scatter plot demonstrating the distribution of contig coverage (number of sequences mapping to each contig) and contig length for the virus metagenomic sample set. B) Scatter plot illustrating the distribution of operational viral unit (OVU) length and sequence coverage for the virus metagenomic sample set. C) Heated scatter plot demonstrating the distribution of contig coverage and length for the whole metagenomic sample set. D) Scatter plot illustrating the distribution of operational genomic unit (OGU) length and sequence coverage for the whole metagenomic sample set.*

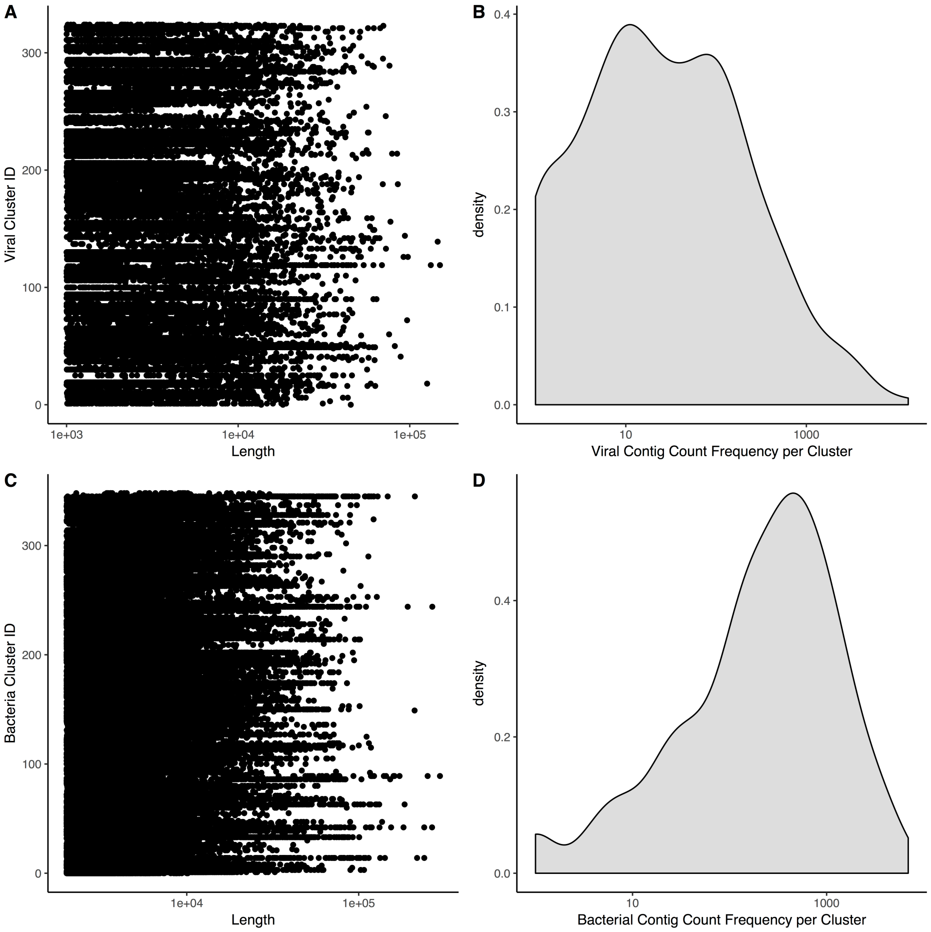


Figure S4: *Operational genomic unit composition stats. A) Strip chart demonstrating the length and frequency of contigs within each operational genomic unit of the virome sample set. The y-axis is the operational genomic unit identifier, and x-axis is the length of each contig, and each dot represents a contig found within the specified operational genomic unit. B) Density plot (analogous to histogram) of the number of virome operational genomic units containing the specific number of contigs, as indicated by the x-axis. C-D) Sample plots as panels C and D, but for the whole metagenomic sample set.*

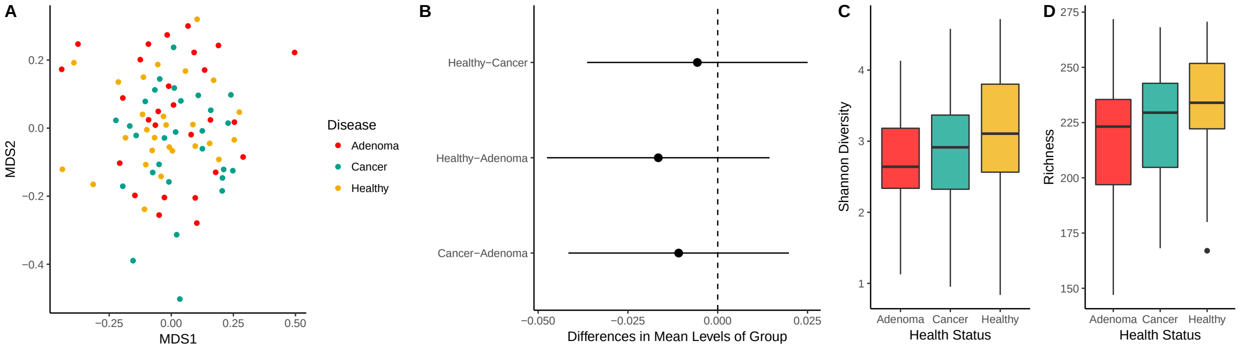


Figure S5: *Diversity calculations comparing cancer states of the colorectal virome, based on relative abundance of operational genomic units in each sample. A) NMDS ordination of community samples, colored for cancerous (green), pre-cancerous (red), and healthy (yellow). B) Differences in means between disease group centroids with 95% confidence intervals based on an ANOSIM test with a post hoc multivariate Tukey test. Comparisons (indicated on y-axis) in which the intervals cross the zero mean difference line (dashed line) were not significantly different. C) Shannon diversity and D) richness alpha diversity quantification comparing pre-cancerous (grey), cancerous (red), and healthy (tan) states.*

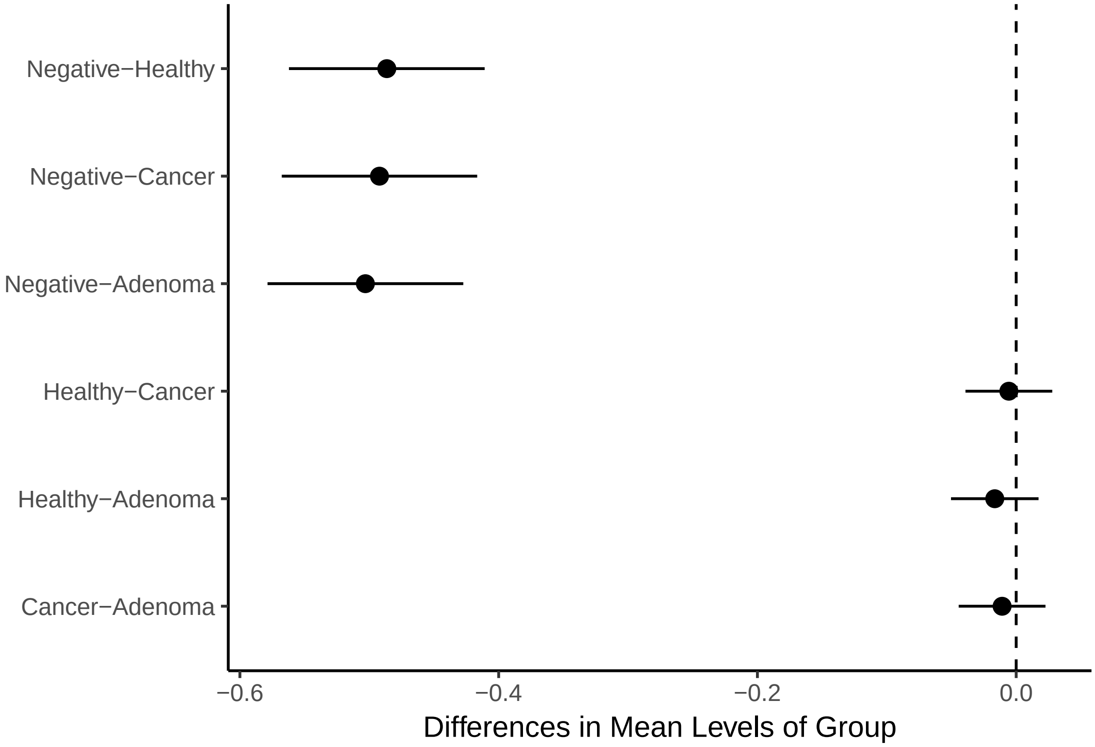


Figure S6: *Beta-diversity comparing disease states and the study negative controls. Differences in means between disease group centroids with 95% confidence intervals based on an ANOSIM test with a post hoc multivariate Tukey test. Comparisons in which the intervals cross the zero mean difference line (dashed line) were not significantly different.*

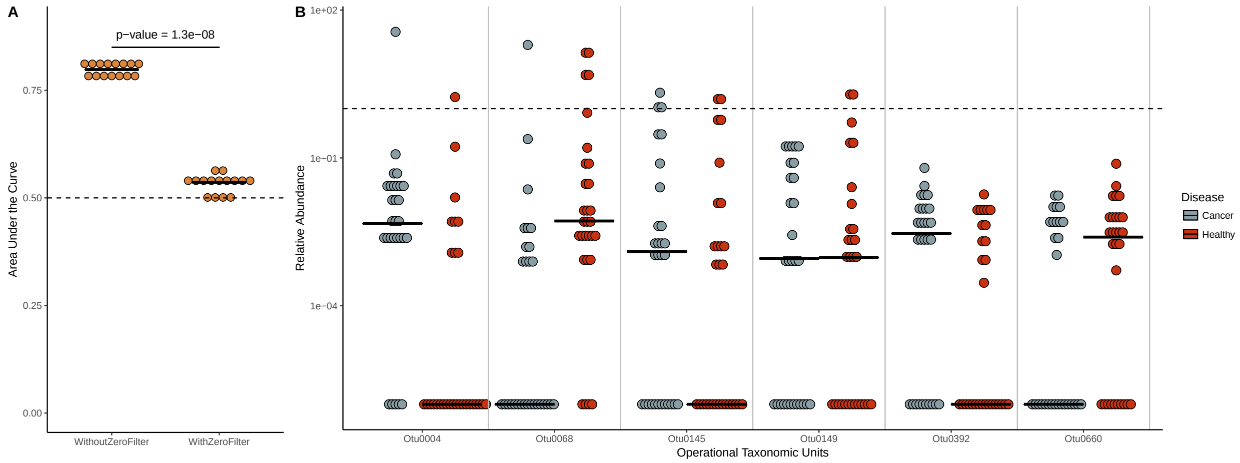


Figure S7: *Comparison of bacterial 16S rRNA classification models with and without OTUs whose median relative abundance are greater than zero. A) Classification model performance (measured as area under the curve) for bacteria models using 16S rRNA data both with and without filtering of samples whose median was zero. Significance was calculated using a Wilcoxon rank sum test, and the resulting p-value is shown. The random area under the curve (0.5) is marked with a dashed line. B) Relative abundance of the six bacterial OTUs removed when filtered for OTUs with median relative abundance of zero. OTU relative abundance is seperated by healthy (red) and cancerous (grey) samples. Relative abundance of 1% is marked by the dashed line.*

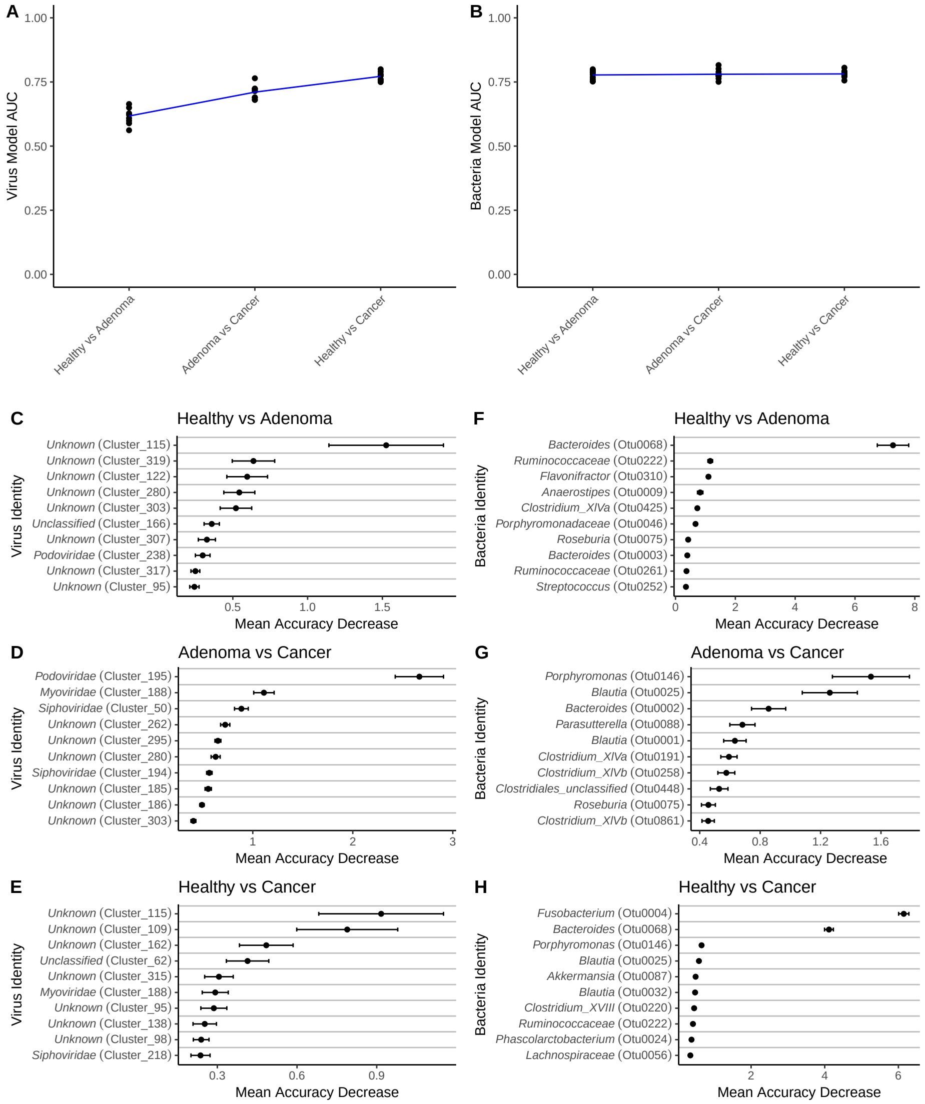


Figure S8: *Transition of colorectal cancer importance through disease progression. A) Virus and B) 16S rRNA gene model performance (AUC) when discriminating all binary combinations of disease types. Blue line represents mean performance from multiple random iterations. C-E) Top ten important phage OVUs when classifying each combination of disease state, as measured by the mean decrease in accuracy metric. Mean is represented by a point, and bars represent standard error. Disease comparison is specified in the top left corner of each panel. F-H) Top ten important bacterial 16S rRNA gene OTUs for classifying each disease state combination.*

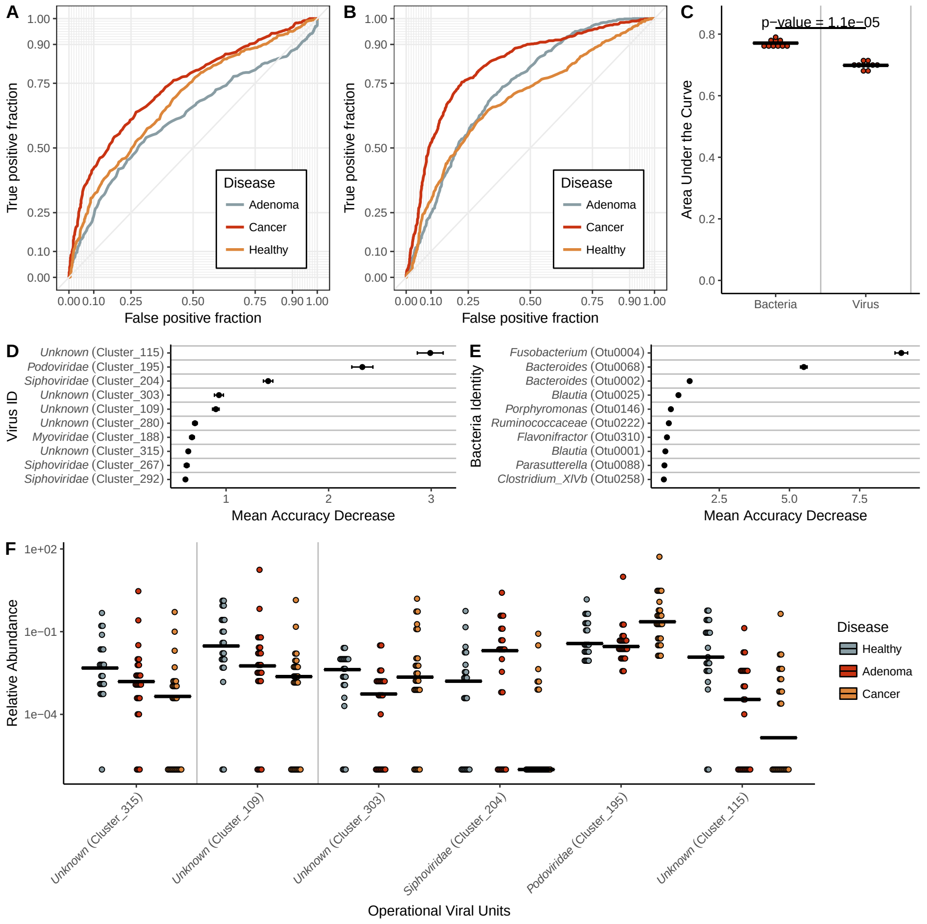


Figure S9: *ROC curves from A) virome and B) bacterial 16S three-class random forest models tuned on mean AUC. Each curve represents the ability of the specified class to be classified against the other two classes. C) Quantification of the mean AUC variation for each model based on 10 model iterations. A pairwise Wilcoxon test with a Bonferroni multiple hypothesis correction demonstrated that the models are significantly different (alpha = 0.01). D) Mean decrease in accuracy when virome operational genomic units and E) bacterial 16S OTUs are removed from the respective three-class classification models. Results based on 25 iterations. F) Relative abundance of the six most important virome OVUs in the model, with the most important on the right. Line indicates abundance mean.*

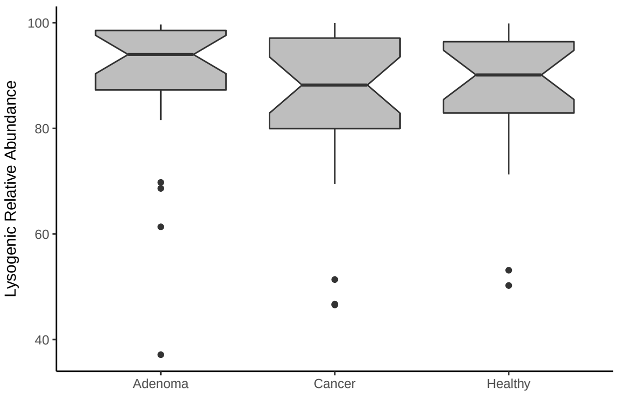


Figure S10: *Lysogenic phage relative abundance in disease states. Phage OVUs were predicted to be either lytic or lysogenic, and the relative abundance of lysogenic phages was quantified and represented as a boxplot. No disease groups were statistically significant.*

References

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

2. Shuda, M., Kwun, H. J., Feng, H., Chang, Y. & Moore, P. S. Human Merkel cell polyomavirus small T antigen is an oncoprotein targeting the 4E-BP1 translation regulator. *Journal of Clinical Investigation* **121,** 3623–3634 (2011).

3. Schiller, J. T., Castellsagué, X. & Garland, S. M. A review of clinical trials of human papillomavirus prophylactic vaccines. *Vaccine* **30 Suppl 5,** F123–38 (2012).

4. Chang, Y. *et al.* Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* **266,** 1865–1869 (1994).

5. Harcombe, W. R. & Bull, J. J. Impact of phages on two-species bacterial communities. *Applied and Environmental Microbiology* **71,** 5254–5259 (2005).

6. Rodriguez-Valera, F. *et al.* Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology* **7,** 828–836 (2009).

7. Cortez, M. H. & Weitz, J. S. Coevolution can reverse predator-prey cycles. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 7486–7491 (2014).

8. Zackular, J. P., Rogers, M. A. M., Ruffin, M. T. & Schloss, P. D. The human gut microbiome as a screening tool for colorectal cancer. *Cancer prevention research (Philadelphia, Pa.)* **7,** 1112–1121 (2014).

9. Garrett, W. S. Cancer and the microbiota. *Science* **348,** 80–86 (2015).

10. Baxter, N. T., Zackular, J. P., Chen, G. Y. & Schloss, P. D. Structure of the gut microbiome following colonization with human feces determines colonic tumor burden.

Microbiome **2,** 20 (2014).

11. Arthur, J. C. *et al.* Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338,** 120–123 (2012).

12. Ly, M. *et al.* Altered Oral Viral Ecology in Association with Periodontal Disease. *mBio* **5,** e01133–14–e01133–14 (2014).

13. Monaco, C. L. *et al.* Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome. *Cell Host and Microbe* **19,** 311–322 (2016).

14. Willner, D. *et al.* Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLOS ONE* **4,** e7370 (2009).

15. Abeles, S. R., Ly, M., Santiago-Rodriguez, T. M. & Pride, D. T. Effects of Long Term Antibiotic Therapy on Human Oral and Fecal Viromes. *PLOS ONE* **10,** e0134941 (2015).

16. Modi, S. R., Lee, H. H., Spina, C. S. & Collins, J. J. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499,** 219–222 (2013).

17. Santiago-Rodriguez, T. M., Ly, M., Bonilla, N. & Pride, D. T. The human urine virome in association with urinary tract infections. *Frontiers in Microbiology* **6,** 14 (2015).

18. Norman, J. M. *et al.* Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160,** 447–460 (2015).

19. Siegel, R., Desantis, C. & Jemal, A. Colorectal cancer statistics, 2014. *CA: a cancer journal for clinicians* **64,** 104–117 (2014).

20. Zackular, J. P., Baxter, N. T., Chen, G. Y. & Schloss, P. D. Manipulation of the Gut Microbiota Reveals Role in Colon Tumorigenesis. *mSphere* **1,** e00001–15 (2016).

21. Dejea, C. M. *et al.* Microbiota organization is a distinct feature of proximal colorectal

cancers. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 18321–18326 (2014).

22. Flynn, K. J., Baxter, N. T. & Schloss, P. D. Metabolic and Community Synergy of Oral Bacteria in Colorectal Cancer. *mSphere* **1,** e00102–16 (2016).

23. Baxter, N. T., Ruffin, M. T., Rogers, M. A. M. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome medicine* **8,** 37 (2016).

24. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology* **10,** 766–766 (2014).

25. Fearon, E. R. Molecular genetics of colorectal cancer. *Annual review of pathology* **6,** 479–507 (2011).

26. Levin, B. *et al.* Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. in *CA: A cancer journal for clinicians* 130–160 (The University of Texas MD Anderson Cancer Center, Houston, TX, USA. John Wiley & Sons, Ltd., 2008).

27. Zauber, A. G. The impact of screening on colorectal cancer mortality and incidence: has it really made a difference? *Digestive diseases and sciences* **60,** 681–691 (2015).

28. Pedulla, M. L. *et al.* Origins of highly mosaic mycobacteriophage genomes. *Cell* **113,** 171–182 (2003).

29. Hannigan, G. D. *et al.* The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome. *mBio* **6,** e01578–15 (2015).

30. Brum, J. R. *et al.* Ocean plankton. Patterns and ecological drivers of ocean viral

communities. *Science* **348,** 1261498–1261498 (2015).

31. Minot, S. *et al.* The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Research* **21,** 1616–1625 (2011).

32. Hannigan, G. D. *et al.* Evolutionary and functional implications of hypervariable loci within the skin virome. *PeerJ* **5,** e2959 (2017).

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

34. Hannigan, G. D., Duhaime, M. B., Koutra, D. & Schloss, P. D. Biogeography & Environmental Conditions Shape Phage & Bacteria Interaction Networks Across The Human Microbiome. *bioRxiv* 1–40 (2017).

35. Lengeling, A., Mahajan, A. & Gally, D. L. Bacteriophages as Pathogens and Immune Modulators? *mBio* **4,** e00868–13–e00868–13 (2013).

36. G rski, A. *et al.* in *Bacteriophages, part b* 41–71 (Bacteriophage Laboratory, Ludwik Hirszfeld Institute of Immunology; Experimental Therapy, Polish Academy of Sciences, Wrocław, Poland. agorski@ikp.pl; Elsevier, 2012).

37. Rossmann, F. S. *et al.* Phage-mediated Dispersal of Biofilm and Distribution of Bacterial Virulence Genes Is Induced by Quorum Sensing. *PLoS Pathogens* **11,** e1004653–17 (2015).

38. Brockhurst, M. A. & Koskella, B. Experimental coevolution of species interactions. *Trends in ecology & evolution* **28,** 367–375 (2013).

39. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology* **79,** 5112–5120 (2013).

40. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent,

community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75,** 7537–7541 (2009).

41. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35,** 7188–7196 (2007).

42. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27,** 2194–2200 (2011).

43. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nature protocols* **4,** 470–483 (2009).

44. Kleiner, M., Hooper, L. V. & Duerkop, B. A. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* **16,** 7 (2015).

45. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).

46. Hannon, G. J. FASTX-Toolkit. GNU Affero General Public License

47. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLOS ONE* **6,** e17288 (2011).

48. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *METHODS* **102,** 3–11 (2016).

49. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).

50. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nature*

Methods 1–7 (2014).

51. Oksanen, J. *et al.* vegan: Community Ecology Package.

52. Kuhn, M. caret: Classification and Regression Training.

53. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532,** 465–470 (2016).

54. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 1 (2009).