Viral and Bacterial Communities of Colorectal Cancer

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 and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, 48109

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

[*]To whom correspondence may be addressed.

Corresponding Author Information

Patrick D Schloss, PhD
1150 W Medical Center Dr. 1526 MSRB I
Ann Arbor, Michigan 48109
Phone: (734) 647-5801
Email: pschloss@umich.edu

Journal: PNAS (*Preparation Details*)

Major Classification: Biological Sciences

Minor Classification: Microbiology

Keywords: Colorectal Cancer, Virome, Machine Learning

* *Figures included for internal editing purposes*

Abstract

Colorectal cancer is the second leading cause of cancer-related death in the United States and is a primary cause of morbidity and mortality throughout the world. Its progression has been linked to changes in colonic bacterial community composition. Viruses are another important component of the colonic microbial community, however they have yet to be studied in colorectal cancer despite their oncogenic potential. We evaluated the colorectal cancer virome (virus community) using a cohort of 90 human subjects with either healthy, adenomatous (precancerous), or cancerous colons. We utilized 16S rRNA gene, whole shotgun metagenomic, and purified virus metagenomic sequencing methods to compare the colorectal cancer virome to the bacterial community. We found that alpha and beta diversity metrics were insufficient for detecting virome changes in colorectal cancer, but more sophisticated random forest models identified striking changes in the virus community. The majority of the cancer-associated virome consisted of temperate bacteriophages, suggesting that the community was indirectly linked to colorectal cancer by modulating bacterial community structure and function. Our data suggested that the influential phages did not exclusively infect influential bacterial, but rather acted through the community as a whole. These results provide foundational evidence that bacteriophage communities are associated with colorectal cancer and likely impact cancer progression by altering the bacterial host communities.

Word Count: 213 / 250

Significance Statement

Colorectal cancer is a leading cause of cancer-related death in the United States and worldwide. Its progression and severity have been linked to colonic bacterial community composition. Although viruses have been linked to other cancers and diseases, little is known about colorectal cancer virus communities. We addressed this knowledge gap by identifying changes in colonic virus communities in colorectal cancer patients and how they compared to bacterial community changes. The results suggested an indirect role for the virome impacting colorectal cancer by modulating their associated bacterial community. These findings both support a biological role for viruses in colorectal cancer, as well as provide a new understanding of basic colorectal cancer etiology.

Word Count: 112 / 120

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 are crucial for bacterial community stability and composition (5–7), and because bacteria have been implicated as oncogenic agents (8–10), bacteriophages have the potential to indirectly impact cancer. The gut virome (the virus community of the gut) therefore has the potential to impact health and disease (e.g. cancer). Altered human virome composition and diversity have been identified in diseases including periodontal disease (11), HIV (12), antibiotic exposure (13, 14), urinary tract infections (15), and inflammatory bowel disease (16). The strong association of bacterial communities with colorectal cancer and the precedence for the virome in impacting other human diseases together 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 (17). The US National Cancer Institute estimates over 1.5 million Americans have been diagnosed with colorectal cancer in 2016, and over 500,000 Americans have died from the disease (17). One component of colorectal cancer etiology is variation in colorectal bacterial community composition (8, 10, 18, 19). This work 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 (18). This association has also allowed researchers to leverage bacterial community signatures as biomarkers to provide accurate, noninvasive colorectal cancer detection from stool (8, 20). While an understanding of colorectal cancer bacterial communities has proven fruitful both for disease classification and understanding underlying 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 is a stepwise process that begins when healthy tissue develops into a precancerous polyp (i.e. adenoma) in the large intestine (21). If left untreated, the adenoma will 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 (22, 23), and survival for colorectal cancer patients may exceed 90% when the lesions are detected early and removed (22). 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 changes might impact cancer progression and severity. We also aimed to evaluate their potential as a reservoir of diagnostic biomarkers. The implications of this study are threefold. *First*, this supports a biological role for the virome in colorectal cancer development and suggests that more than bacteria 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.

Results

Cohort Design, Sample Collection, and Processing

The study cohort consisted of 90 human subjects, 30 of which had healthy colons, 30 of which had adenomas, and 30 which had carcinomas **(Figure S1)**. Half of the stool was used to sequence the bacterial communities using both 16S rRNA gene and shotgun sequencing techniques. The other half of the stool samples were purified for virus like particles (VLPs) before genomic DNA extraction and shotgun metagenomic sequencing. The VLP purification allowed us to observe the *active virome* because we only sequenced those viruses that were encapsulated.

Virus DNA was purified prior to sequencing, allowing us to analyze DNA exclusively from within virus capsids **(Figure S1)**. Each extraction protocol was performed with a blank 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 initial evidence of successful sequencing of VLP genomic DNA over potential contaminants **(Figure S2 A)**. As was expected, these controls yielded few sequences and were mostly 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 1kb **(Figure S3)**. Because contigs only 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)**.

Unaltered Virome Diversity in Colorectal Cancer

Microbiome and disease associations are often described as being of an altered diversity (i.e. “dysbiotic”). We therefore initially evaluated the changes in virome OVU diversity in colorectal cancer. We evaluated differences in communities between disease states using the Bray-Curtis, Shannon entropy, and richness diversity metrics. We observed no significant alterations in either Shannon entropy or richness in the diseased state **(Figure S5 C-D)**. There was also no statistically significant clustering of the disease groups (ANOSIM p-value = 0.432, **Figure S5**). It is worth noting that there was a significant difference between the few blank controls that remained after rarefying the data, and the other study groups (ANOSIM p-value = 7.18 × 10-28, **Figure S6)**. This further supported the quality of our sample set. Similar to what was previously seen when analyzing the 16S rRNA gene sequences from the same samples (8), standard alpha and beta diversity metrics were insufficient for capturing virus community differences between disease states.

Altered Virome Composition in Colorectal Cancer

16S rRNA gene relative abundance profiles are effective feature sets for classifying stool samples as originating from individuals with healthy, adenomatous, or cancerous colons (8, 20). 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 16S rRNA gene model to a model built using virome relative abundance. Statistically, the viral model performed better than the bacterial model (corrected p-value = 0.369), although the results were practically similar, with the viral and bacterial models achieving mean AUC (area under the curve) 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 statistically significant but small performance improvement beyond the viral (corrected p-value = NA) and bacterial (corrected p-value = 0.00155) models, yielding an AUC of 0.816 **(Figure 1 A - B)**. This suggested that features from the virus and bacterial communities interact to improve the 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 relative abundance profiles from bacterial metagenomic shotgun sequencing data. This model performed more poorly than any of the other models (mean AUC = 0.505) **(Figure 1 A - B)**. Further 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% **(Figure S7 B)**. Although the features in the viral model were also 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.

The association between the two communities and colorectal cancer was driven by a few important microbes, measured using the mean decrease in model accuracy when each was removed. *Fusobacterium* was the primary driver of the bacterial association with colorectal cancer, which was consistent with its previously described oncogenic potential **(Figure 1 C)**(18). The virome signature was also driven by a few OVUs, suggesting a role for the viruses in cancer development **(Figure 1 D)**. The identified viruses were bacteriophages, belonging to *Siphoviridae*, *Myoviridae*, and orphan phage taxa without taxonomic identifiers (denoted “unclassified”). Many of the important viruses were unidentifiable (denoted “unknown”), suggesting they are members of the abundant unknown viral population associated with the human virome. This is common in the virome; studies can have as much as 95% of virus sequences belong to unknown genomic units (24, 25). When the bacterial and viral community signatures were combined, both bacterial and viral organisms drove the community association with cancer **(Figure 1 E)**.

Shifted Phage Influence Across Cancer Stages

Because our cohort included people with healthy, adenomatous, and cancerous colons we were able to evaluate shifts in influential phage identities during cancer progression. 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 16S rRNA gene 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 in the transitions from 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 model yielded a mean AUC of 0.771 and outperformed the viral community model which yielded a mean AUC of 0.699 (p-value = 1.08 × 10-5, **Figure S9 A-C**). The microbes important for the cancer vs healthy and healthy vs 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 also identified as bacteriophages **(Figure 1 D, Figure S9 E)**. Not all important OVUs were of increased abundance in the diseased state **(Figure S9 F)**.

Bacteriophage Dominance in Colorectal Cancer Virome

Changes 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 are 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.

We evaluated whether the phages in the community were primarily lytic (replicate by lysing their hosts) or temperate (lysogenic; able to integrate into their host's genome, as well as lyse the cell). 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, as previously described (25, 26). We found that the majority of the colon phages were temperate, and that the overall fraction of temperate phages remained consistent throughout the healthy, adenomatous, and cancerous stages **(Figure S10 E)**. Thus the majority of the OVUs were temperate bacteriophages and not eukaryotic viruses, indicating that the association between the virome and colorectal cancer was reliant on bacteriophage communities that can lie dormant in bacterial genomes. These findings were consistent with previous reports suggesting the gut virome is primarily composed of temperate phages (12, 16, 26, 27).

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 the bacterial and phage relative abundances **(Figure 2 A,C)**. There was an overall absence of correlation between the most influential phage 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. This hypothesis would be supported by showing that the influential bacteriophages were community hubs within the bacteria and phage interactive network. We investigated the infectious capabilities and potential host ranges of all phage OVUs using a random forest model to predict which phages infected which bacteria in the overall community (**Cite Network Preprint**). The predicted interactions were then used to identify community hub phages. This revealed a wide tropism range for the bacteriophages within the community **(Figure 3 A)**. We calculated the alpha centrality (measure of importance in the ecosystem network) of each phage OVU's connection to the rest of the network, and compared the centrality to the importance of each OVU in the colorectal cancer classification model. The phages with high centrality values were defined as community hubs. Phage OVU centrality was significantly and positively correlated with importance to the disease model (p-value = 0.0173, R = 0.14), suggesting that phages important in driving colorectal cancer were also more likely to be community hubs **(Figure 3 B)**. Together these findings supported our hypothesis that influential phages were hubs within their microbial communities.

Discussion

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, meaning bacteriophages may play an indirect role in promoting carcinogenesis by influencing bacterial communities (8–10). Despite their carcinogenic potential and the strong association between bacteria and colorectal cancer, the 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 colorectal cancer. Our findings support a working hypothesis for oncogenesis by phage-modulated bacterial community composition.

Our data allowed us to begin delineating the role the colonic virome played in colorectal cancer **(Figure 4 A)**. We found basic diversity metrics of alpha diversity (richness and Shannon entropy) and beta diversity (Bray-Curtis dissimilarity) were insufficient for identifying virome community changes 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 exclusive 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, 19). This provides a precedent for phage indirectly influencing colorectal cancer progression by altering the bacterial community composition. Overall, our data supports 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.

In addition to modeling the basic potential connections between virus communities, bacteria 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)** (18). We hypothesize that the broadly infectious phages in the colon began lysing, and thereby disrupting, the bacterial communities to open a niche in which opportunistic bacteria (such as *Fusobacterium nucleatum*) were able to colonize. Once the initial influential bacteria had established themselves in the epithelium, other opportunistic bacteria were able to adhere to the driver, colonize, and begin establishing a biofilm. Phages may have played a role in biofilm dispersal and growth by lysing bacteria within the biofilm, a process important for effective biofilm growth (28). The oncogenic bacteria were then 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 abundance was most influential. As the bacteria entered their oncogenic synergy with the epithelium, we hypothesize the phages could have continued mediating biofilm dispersal, as well as support the colonized oncogenic bacteria by lysing competing cells to maintain the niche and provide nutrients to the other bacteria. In addition to highlighting the most likely mechanisms by which the colorectal cancer virome is interacting with the bacterial communities, this outline will guide future research investigations of the role of the virome in colorectal cancer.

A notable observation from our analysis was the poor performance observed using bacterial metagenomic methods compared to the performance of models using viral metagenomes or 16S rRNA gene sequences. This observation highlights the importance of high sequencing coverage in bacterial metagenomic studies, and the advantage of 16S rRNA gene sequencing over whole metagenomic shotgun sequencing. We found that there were six bacterial OTUs that drove the performance of the 16S rRNA gene classification model, and these OTUs were all sparsely present and lowly abundant. Filtration of OTUs with a median relative abundance of zero resulted in the removal of the six important OTUs and reduced model performance to being nearly random like the bacterial metagenomic model. The bacterial metagenomic OGUs represented only the most abundant taxa, which was not informative for this application. There has been some success in using shotgun metagenomic approaches for stool colorectal cancer classification, but these approaches did not utilize OGU clustering like we did here, and the models only performed **as well** as the 16S rRNA gene model (29). Thus the targeted 16S rRNA gene sequencing approach, which yielded only a fraction of the bacterial metagenomic sequences, was more effective for detecting colorectal cancer in stool samples. Despite a 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.

In addition to the therapeutic ramifications for understanding the colorectal cancer microbiome, our findings provide a proof-of-principle that viruses, while under-appreciated and understudied in the human microbiome, are an important contributor to human disease that has the potential to provide an abundance of information that supplements that 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 thrive without each other (6). Not only is the human virome an important part of human health and disease, but it appears to have a particular significance in cancer research.

Methods

Analysis Source Code & Availability

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

https://github.com/SchlossLab/Hannigan_CRCVirome_PNAS_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 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) (30, 31). Briefly, the sequences were de-replicated, aligned to the SILVA database (32), screened for chimeras using UCHIME (33), 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 (20).

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. 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 (25). Briefly, an aliquot of stool (ca. 0.1g) was resuspended in standard SM buffer (Crystalgen) and vortexed to facilitate resuspension. The resuspended stool was centrifuged to remove major particulate debris, followed by filtering through a 0.22-μm filter to remove smaller contaminants. The filtered supernatant was treated with chloroform to lyse contaminating cells including bacteria, human, fungi, etc. The exposed genomic DNA from the lysed cells was degraded by treating the samples with DNase. The DNA was extracted from the purified VLPs using the Wizard PCR Purification Preparation Kit (Promega). Disease classes were staggered across purification runs to prevent run variation as a confounding factor. Purified DNA was used to prepare a shotgun sequencing library using the Illumina Nextera XT library 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 (25). 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 metagenomic sample sets were subjected to the same quality control procedures. The sequences were obtained as de-multiplexed fastq files from the HiSeq platform 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 (34). The FastX toolkit (v0.0.14) was used to quality trim the reads to a minimum length of 75bp and a minimum quality score of 30 (35). Reads mapping to the human genome were removed using the DeconSeq algorithm (v0.4.3) and default parameters (36).

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 (37). 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 (38). 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 removing samples with less total samples 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 in absence of taxonomic identification, we operationally 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 (39). CONCOCT was used with a length threshold of 1000 bp for virus contigs and 2000 bp for bacteria due to computational limitations.

Diversity

Alpha and beta diversity were calculated using the operational genomic 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 Entropy 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 (40).

Classification Modeling

Classification modeling was performed in R using the Caret package (41). OTU and OGU abundance data was preprocessed by removing features (OTUs and OGUs) that were present in less than half 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 ROC 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 across five iterations of mtry values. For consistency and accurate comparison between feature groups (e.g. bacteria, virus), the sample model parameters were used for each group. The maximum AUC during training was recorded across 10 iterations of each group model creation 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 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. 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) (42). Annotation was interpreted as phage, eukaryotic virus, or unknown.

Ecological Network Analysis & Correlations

The ecological network of the bacterial and phage operational genomic units were constructed and analyzed as previously described (cite network preprint here). 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

OVU replication sytle was identified using methods described previously (25, 26, 43). Briefly, we identified lysogenic OVUs as representative contigs containing at least one of three genomic markers: 1) phage integrase genes, 2) prophage genes from the ACLAME database, 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 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).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors thank the Great Lakes-New England Early Detection Research Network for providing the fecal samples that were used in this study. GD Hannigan was supported by the Michigan Molecular Mechanisms of Microbial Pathogenesis Fellowship. 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).

Figures

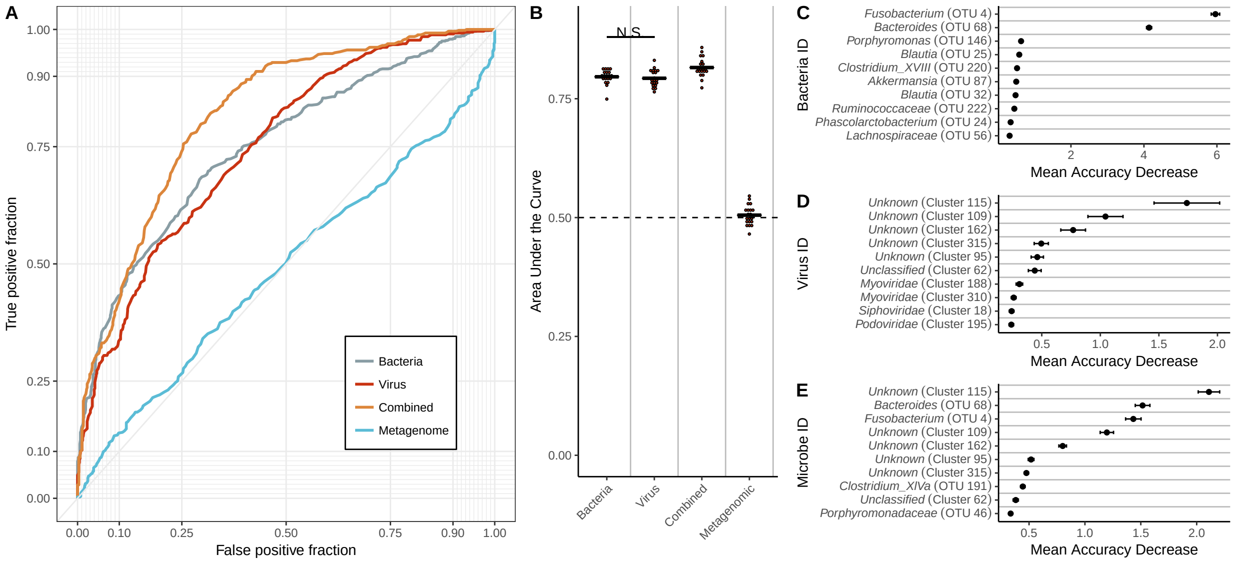


Figure 1: *Results from healthy vs cancer classification models built using virome signatures, bacterial 16S rRNA gene sequence signatures, whole metagenomic signatures, and a combination of virome and 16S rRNA gene sequence signatures. A) An example ROC curve for visualizing the performance of each 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. Results based on 15 iterations. Mean is represented by a point, and bars represent standard error. D) Mean decrease in accuracy of each operational genomic 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.*

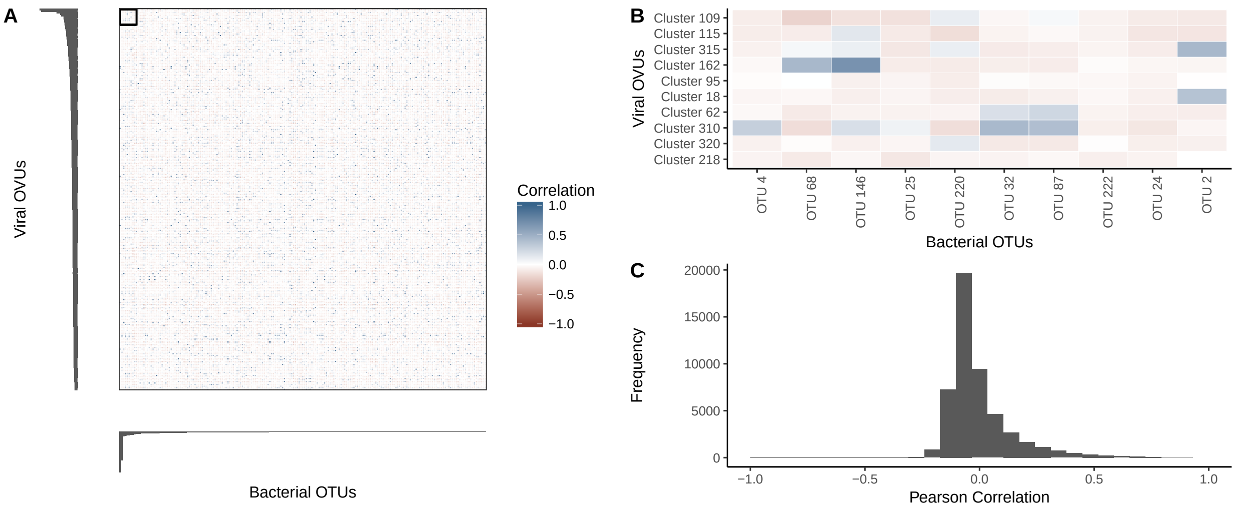


Figure 2: *Relative abundance correlations between bacterial OTUs and virome OVUs. A) Pearson correlation coefficient values between all bacterial OTUs (x-axis) and viral OVUs (y-axis) with blue being positively correlated and red being negatively correlated. Bar plots indicate the viral (left) and bacterial (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.*

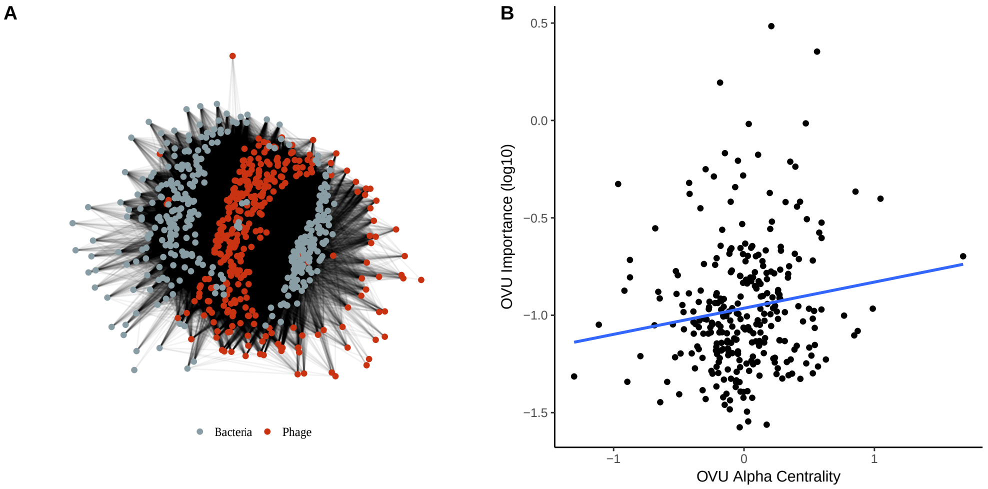
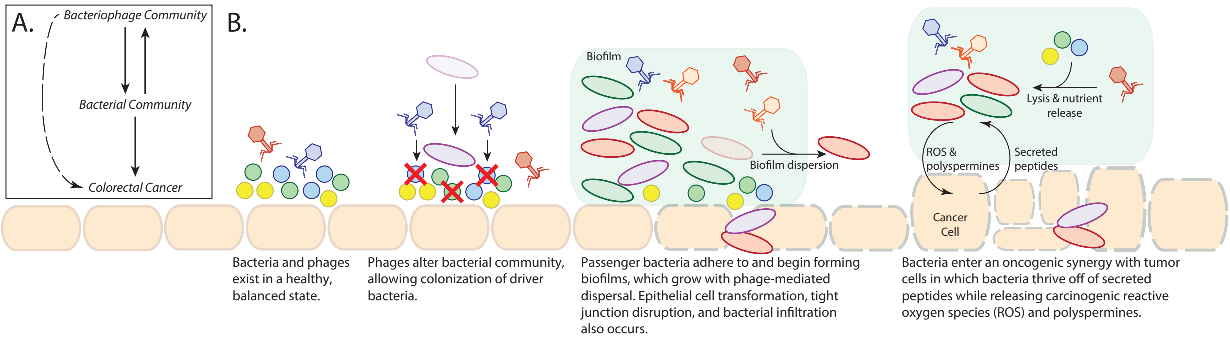


Figure 3: *Community network analysis utilizing predicted interactions between bacteria and phage operational genomic units. A) Visualization of the community network for our colorectal cancer cohort. B) Scatter plot illustrating the correlation between importance (mean decrease in accuracy) and the 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 Figures

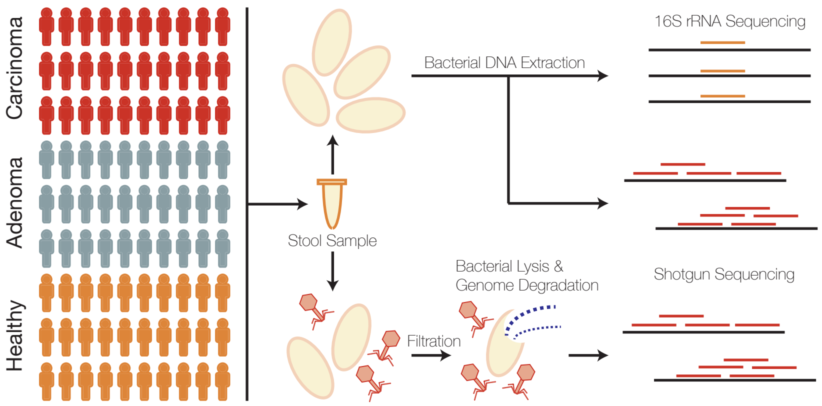


Figure S1: *Cohort and sample processing outline. Thirty subject stool samples were collected from healthy, adenoma (pre-cancer), and carcinoma (cancer) patients. Stool samples were split into two aliquots, the first of which was used for bacterial sequencing and the second which was used for virus sequencing. 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.*

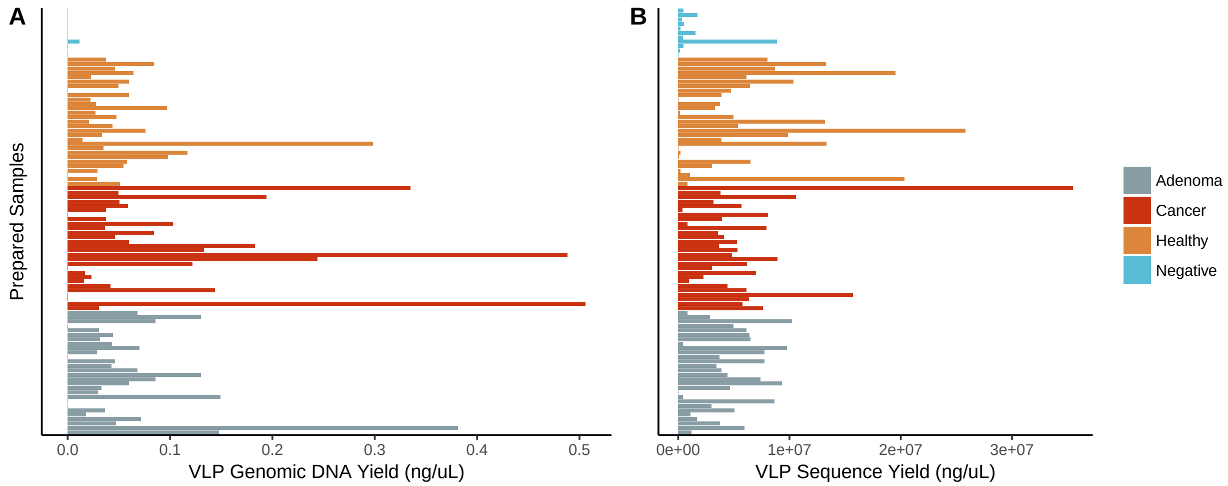


Figure S2: *Basic Quality Control Metrics. A) VLP genomic DNA yield from all sequenced samples. Each bar represents a sample which is grouped and colored by its associated disease group. B) Sequence yield following quality control including quality score filtering and human decontamination.*

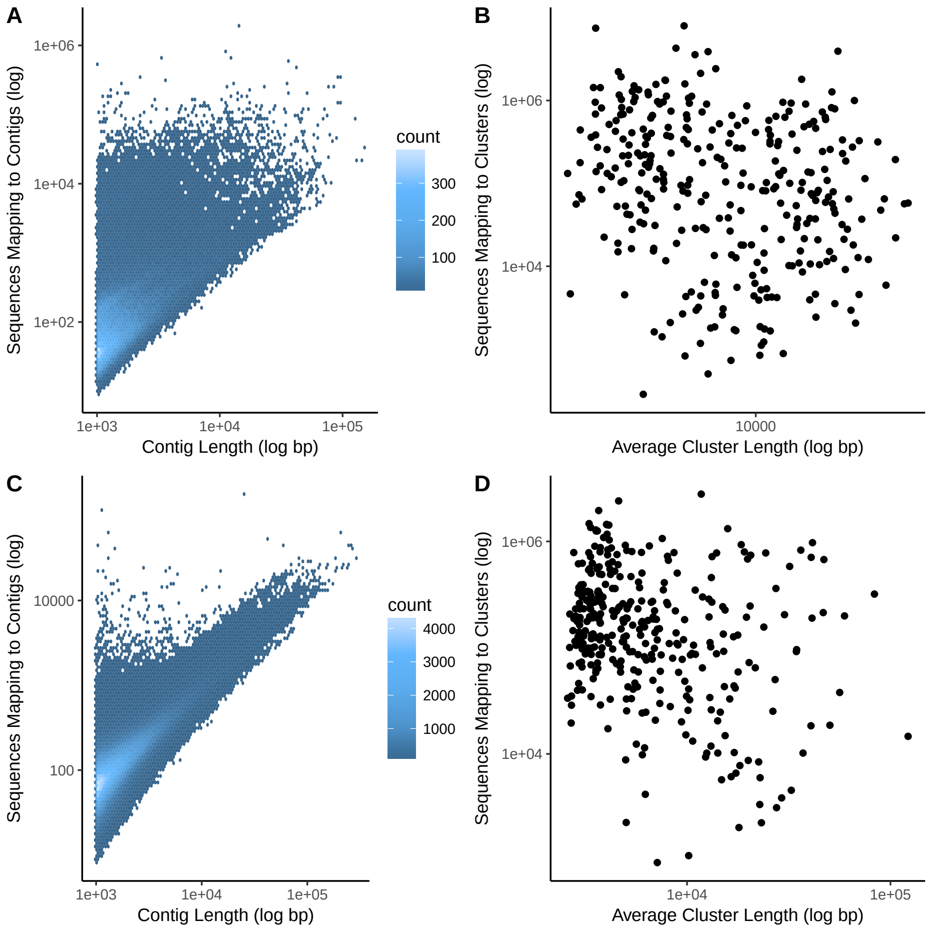


Figure S3: *Length and coverage statistics. A) Heated scatter plot demonstrating the distribution of contig coverage (number of sequences mapping to each contig) and contig length for the virus metagenomic sample set. B) Scatter plot illustrating the distribution of operational viral unit (OVU) length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

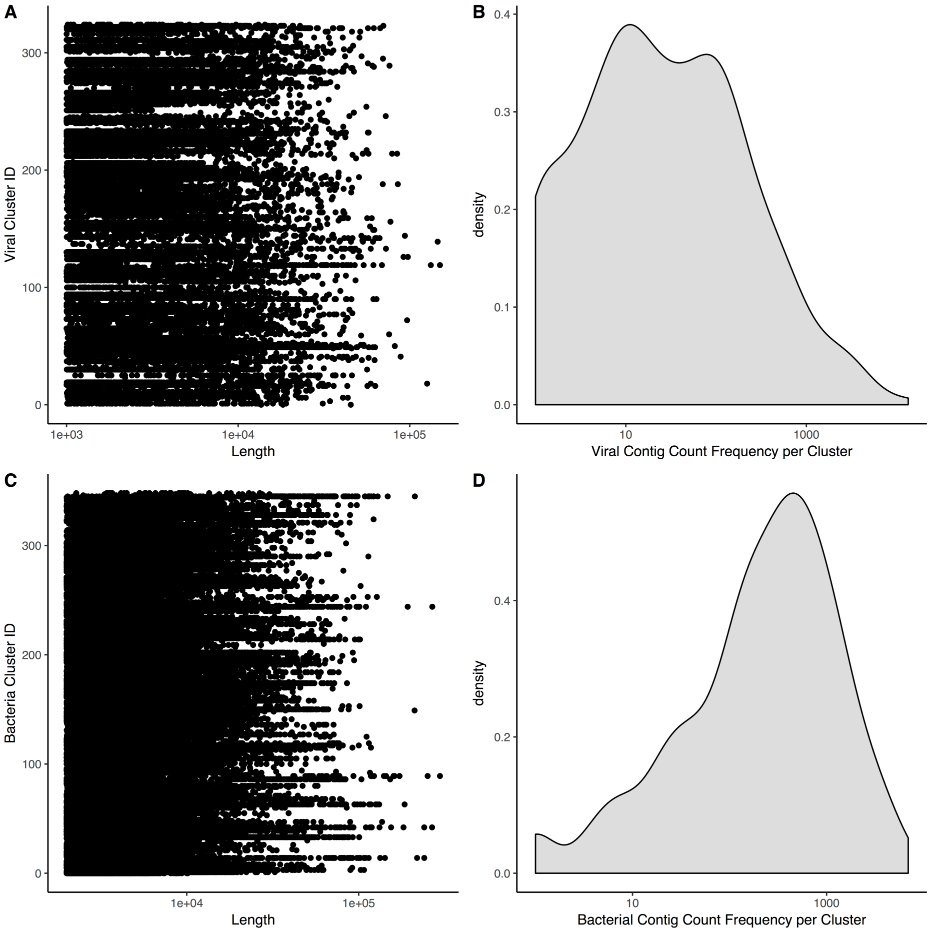


Figure S4: *Operational genomic unit composition stats. A) Strip chart demonstrating the length and frequency of contigs within each operational genomic unit of the virome sample set. The y-axis is the operational genomic unit identifier, and x-axis is the length of each contig, and each dot represents 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.*

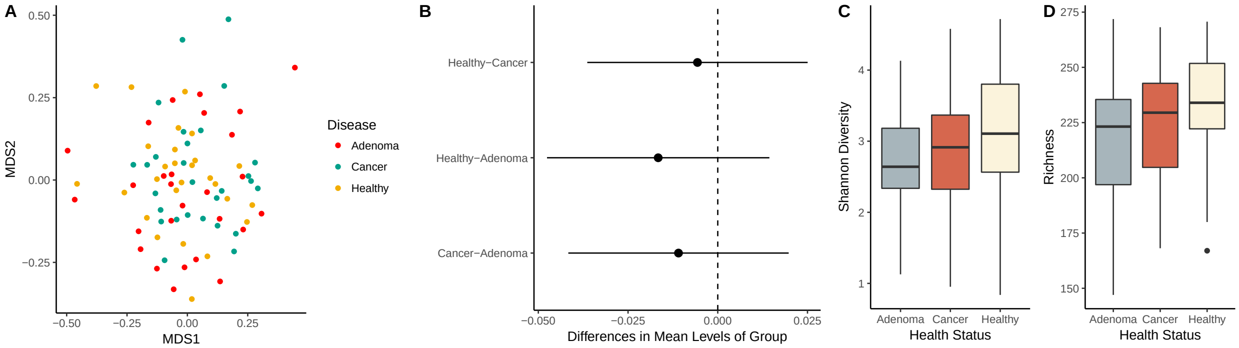


Figure S5: *Diversity calculations comparing cancer states of the colorectal virome, based on relative abundance of operational genomic units in each sample. A) NMDS ordination of community samples, colored for cancerous (green), pre-cancerous (red), and healthy (yellow). B) Differences in means between 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.*

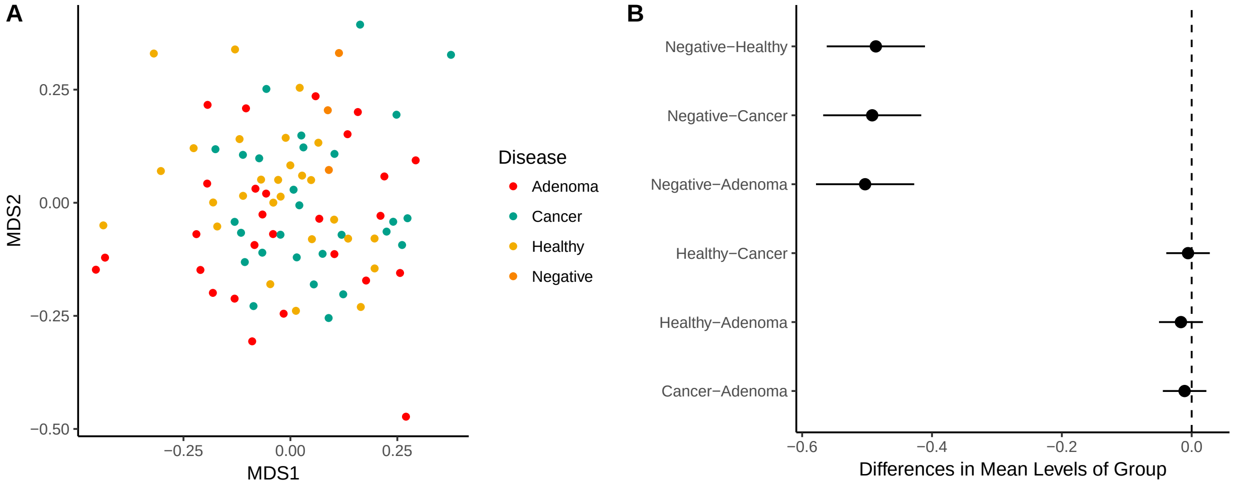


Figure S6: *Beta-diversity comparing disease states and the study negative controls. A) NMDS ordination of community samples, colored by disease state. B) Differences in means between disease group centroids with 95% confidence intervals based on an ANOSIM test with a post hoc multivariate Tukey test.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

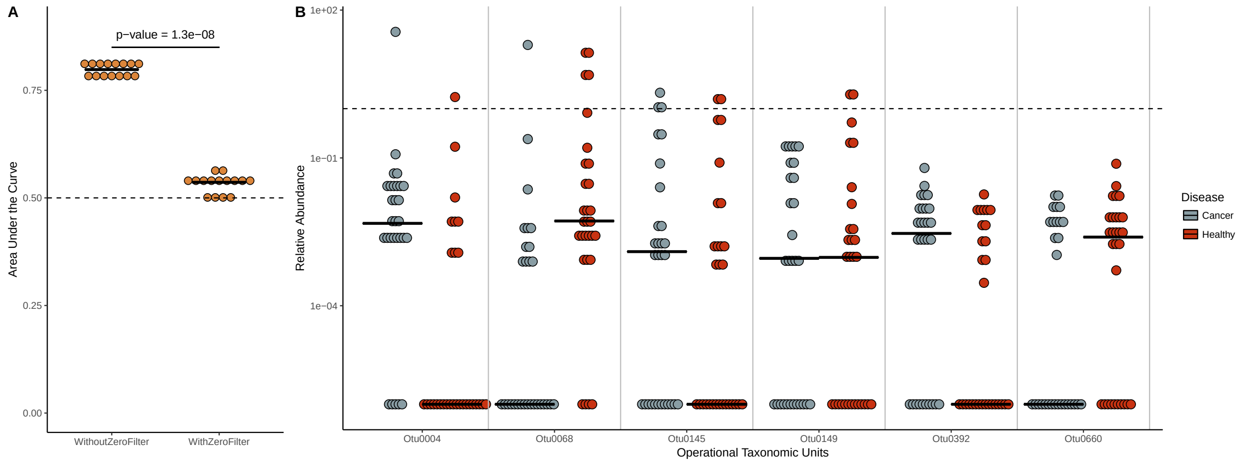


Figure S7: *Comparison of bacterial 16S rRNA classification models with and without OTUs whose median relative abundance are greater than zero. A) Classification model performance (measured as area under the curve) for bacteria models using 16S rRNA data both with and without filtering of samples whose 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.*

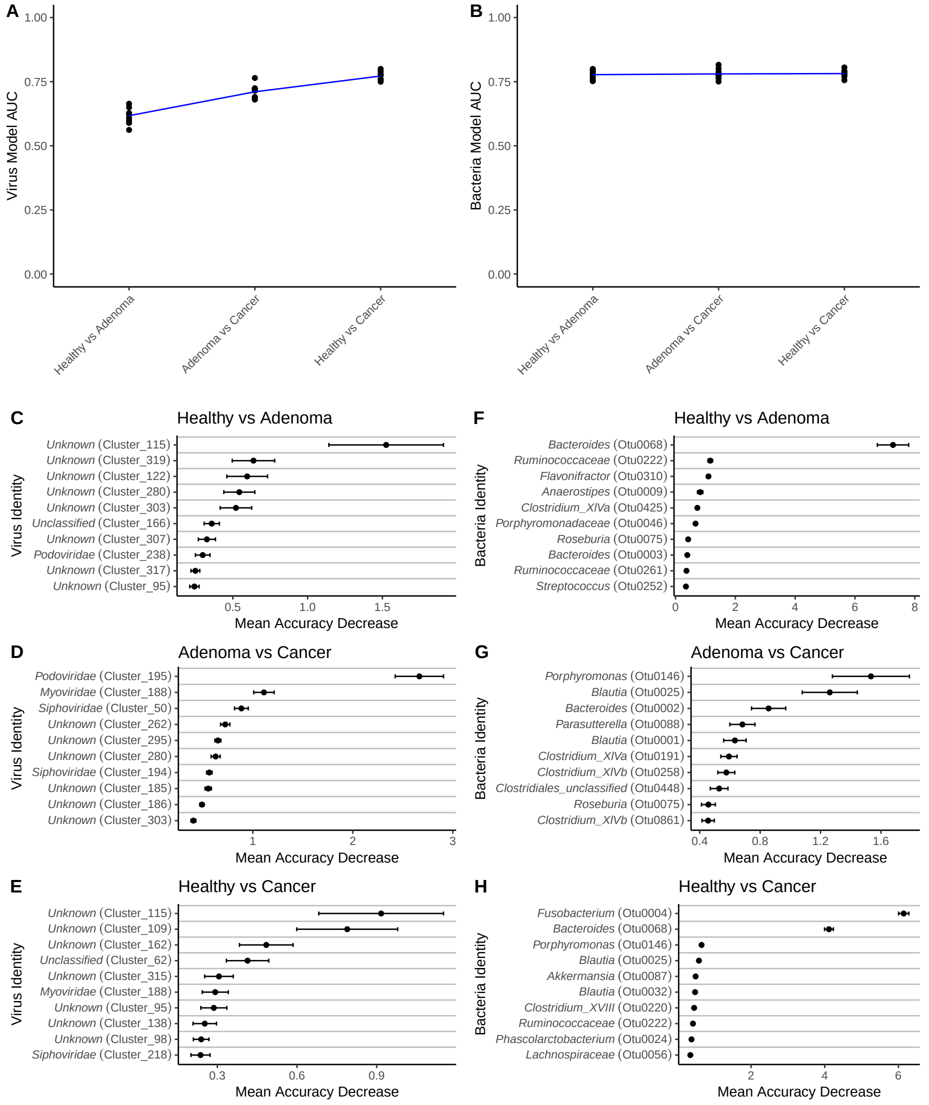


Figure S8: *Transition of colorectal cancer importance through disease progression. A) Virus and B) 16S rRNA gene model performance (AUC) when discriminating all binary combinations of disease types. Blue line represents mean performance from multiple random iterations. C-E) Top ten important phage OVUs 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.*

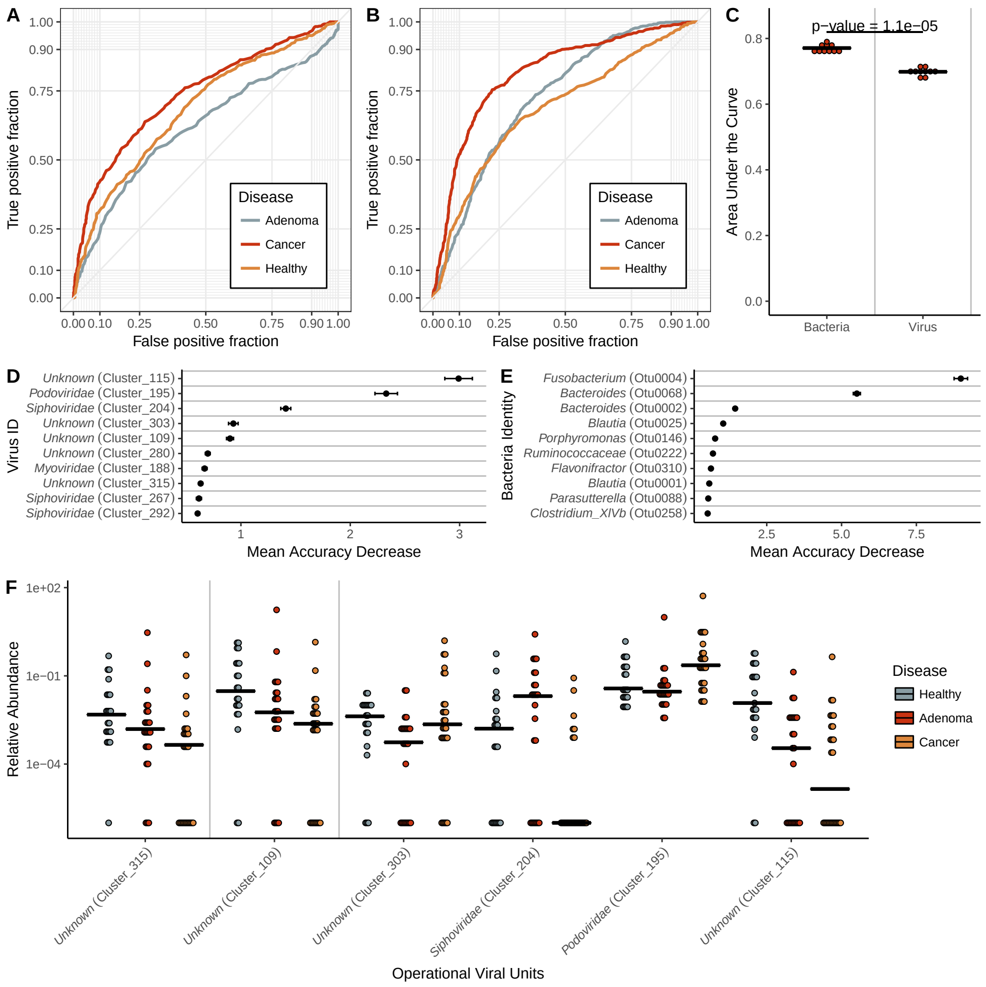


Figure S9: *ROC curves from A) virome and B) bacterial 16S three-class random forest models tuned on mean AUC. Each curve represents the ability of the specified class to be classified against the other two classes. C) Quantification of the mean AUC variation for each model based on 10 model iterations. 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.*

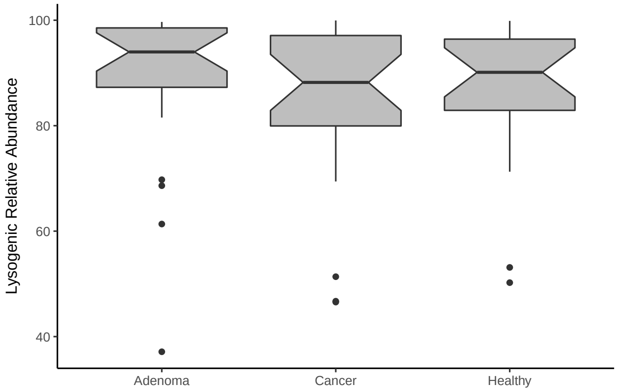


Figure S10: *Lysogenic phage relative abundance in disease states. Phage OVUs were predicted to be either lytic or lysogenic, and the relative abundance of lysogenic phages was quantified and represented as a boxplot. No disease groups were statistically significant.*

References

1. Feng H, Shuda M, Chang Y, Moore PS (2008) Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* 319(5866):1096–1100.

2. Shuda M, Kwun HJ, Feng H, Chang Y, Moore PS (2011) Human Merkel cell polyomavirus small T antigen is an oncoprotein targeting the 4E-BP1 translation regulator. *Journal of Clinical Investigation* 121(9):3623–3634.

3. Schiller JT, Castellsagué X, Garland SM (2012) A review of clinical trials of human papillomavirus prophylactic vaccines. *Vaccine* 30 Suppl 5:F123–38.

4. Chang Y, et al. (1994) Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* 266(5192):1865–1869.

5. Harcombe WR, Bull JJ (2005) Impact of phages on two-species bacterial communities. *Applied and Environmental Microbiology* 71(9):5254–5259.

6. Rodriguez-Valera F, et al. (2009) Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology* 7(11):828–836.

7. Cortez MH, Weitz JS (2014) Coevolution can reverse predator-prey cycles. *Proceedings of the National Academy of Sciences of the United States of America* 111(20):7486–7491.

8. Zackular JP, Rogers MAM, Ruffin MT, Schloss PD (2014) The human gut microbiome as a screening tool for colorectal cancer. *Cancer prevention research (Philadelphia, Pa)* 7(11):1112–1121.

9. Garrett WS (2015) Cancer and the microbiota. *Science* 348(6230):80–86.

10. Baxter NT, Zackular JP, Chen GY, Schloss PD (2014) Structure of the gut microbiome following colonization with human feces determines colonic tumor burden. *Microbiome* 2(1):20.

11. Ly M, et al. (2014) Altered Oral Viral Ecology in Association with Periodontal Disease. *mBio* 5(3):e01133–14–e01133–14.

12. Monaco CL, et al. (2016) Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome. *Cell Host and Microbe* 19(3):311–322.

13. Abeles SR, Ly M, Santiago-Rodriguez TM, Pride DT (2015) Effects of Long Term Antibiotic Therapy on Human

Oral and Fecal Viromes. *PLOS ONE* 10(8):e0134941.

14. Modi SR, Lee HH, Spina CS, Collins JJ (2013) Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* 499(7457):219–222.

15. Santiago-Rodriguez TM, Ly M, Bonilla N, Pride DT (2015) The human urine virome in association with urinary tract infections. *Frontiers in Microbiology* 6:14.

16. Norman JM, et al. (2015) Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160(3):447–460.

17. Siegel R, Desantis C, Jemal A (2014) Colorectal cancer statistics, 2014. *CA: a cancer journal for clinicians* 64(2):104–117.

18. Flynn KJ, Baxter NT, Schloss PD (2016) Metabolic and Community Synergy of Oral Bacteria in Colorectal Cancer. *mSphere* 1(3):e00102–16.

19. Zackular JP, Baxter NT, Chen GY, Schloss PD (2016) Manipulation of the Gut Microbiota Reveals Role in Colon Tumorigenesis. *mSphere* 1(1):e00001–15.

20. Baxter NT, Ruffin MT, Rogers MAM, Schloss PD (2016) Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome medicine* 8(1):37.

21. Fearon ER (2011) Molecular genetics of colorectal cancer. *Annual review of pathology* 6(1):479–507.

22. Levin B, et al. (2008) 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. *CA: A Cancer Journal for Clinicians* (The University of Texas MD Anderson Cancer Center, Houston, TX, USA. John Wiley & Sons, Ltd.), pp 130–160.

23. Zauber AG (2015) The impact of screening on colorectal cancer mortality and incidence: has it really made a difference? *Digestive diseases and sciences* 60(3):681–691.

24. Pedulla ML, et al. (2003) Origins of highly mosaic mycobacteriophage genomes. *Cell* 113(2):171–182.

25. Hannigan GD, et al. (2015) The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome. *mBio* 6(5):e01578–15.

26. Minot S, et al. (2011) The human gut virome: Inter-individual variation and dynamic response to diet. *Genome*

Research 21(10):1616–1625.

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

28. Rossmann FS, et al. (2015) Phage-mediated Dispersal of Biofilm and Distribution of Bacterial Virulence Genes Is Induced by Quorum Sensing. *PLoS Pathogens* 11(2):e1004653–17.

29. Zeller G, et al. (2014) Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology* 10(11):766–766.

30. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) 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(17):5112–5120.

31. Schloss PD, et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75(23):7537–7541.

32. Pruesse E, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* 35(21):7188–7196.

33. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27(16):2194–2200.

34. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17(1):10.

35. Hannon GJ FASTX-Toolkit. GNU Affero General Public License.

36. Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLOS ONE* 6(3):e17288.

37. Li D, et al. (2016) MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies

and community practices. *METHODS* 102:3–11.

38. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9(4):357–359.

39. Alneberg J, et al. (2014) Binning metagenomic contigs by coverage and composition. *Nature Methods*:1–7.

40. Oksanen J, et al. vegan: Community Ecology Package.

41. Kuhn M caret: Classification and Regression Training.

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

43. Hannigan GD, et al. (2017) Evolutionary and functional implications of hypervariable loci within the skin virome. *PeerJ* 5(4):e2959.