Characterizing the Human Virome and Colorectal Cancer Link Using Machine Learning and Network Theory

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, 48108

[2]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, 48108

[3]Department of Family Medicine, University of Michigan, Ann Arbor, Michigan, 48108

[*]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

Text Length: 34,182 / ~39,000 Characters

* *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. Although the cause of colorectal cancer remains unclear, it has been strongly linked to gut bacterial communities. Viruses are another important component of the gut microbial community that have yet to be studied in colorectal cancer, despite their oncogenic potential. We evaluated the gut virome (virus community) role in colorectal cancer using a cohort of 90 human subjects with either healthy, precancerous, or cancerous colons. We utilized 16S rRNA gene, whole shotgun metagenomic, and purified virus metagenomic methods to compare the virome's role to that of the bacterial community. We found that alpha and beta diversity metrics were insufficient for detecting an association between the virome and colorectal cancer, but virome-based classification models were highly associated with colorectal cancer. Bacteriophages, not eukaryotic viruses, made up the majority of the CRC-associated virome, suggesting the community was indirectly linked to colorectal cancer, modulating bacterial community structure and functionality. Phages with broader host ranges had a more significant role in cancer development. These results provide foundational evidence that phage communities are associated with colorectal cancer, and that broadly tropic phages play more significant roles. Because of its importance, the virome should be considered in future colorectal cancer studies, as well as other cancer types.

Word Count: 229

Significance Statement

Colorectal cancer is a leading cause of cancer-related death in the United States and worldwide. It's progression and severity is linked to gut bacterial communities. Little is known about the cancer-associated gut virus communities and their influence on bacteria. We characterize this influence of gut virus community structure on colorectal cancer in humans. The link between the gut virome and colorectal cancer was established by using virome-based machine learning models to accurately classify patient cancer status. The cancer-virus link was driven primarily by bacteriophages (bacterial viruses). Wider phage host ranges were more strongly linked to colorectal cancer. The results suggest an indirect role for the virome impacting colorectal cancer by modulating their associated bacterial community.

Word Count: 119

Introduction

Cancer remains a devastating and persistent plague on humanity. Although cancer is still a primary cause of morbidity and mortality worldwide, we have made considerable therapeutic progress in recent decades. Perhaps one of our most impactful advances has not been in cancer treatment, but rather in detecting cancers at early stages, thereby improving treatment efficacy. This has been evident in a variety of prominent cancer types, with one of the most notable being colorectal cancer.

Colorectal cancer is the second leading cause of cancer-related deaths in the United States (1). The US National Cancer Institute estimates over 1.5 million Americans have been diagnosed with colorectal cancer in 2016, and over 500,000 Americans will have died from the disease (1). Although it remains a major health problem, the impact of colorectal cancer has been reduced by improved screening and prevention efforts (1, 2).

Development of colorectal cancer is a stepwise process that begins when healthy tissue develops into a precancerous polyp (i.e. adenoma) in the large intestine (3). 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 (2, 4). Survival for colorectal cancer patients may exceed 90% when the lesions are detected early and removed (4). Screening methods are effective, but their invasiveness has created a lack of compliance, creating a need for accurate, non-invasive screening methods. One such method is screening associated gut microbial communities.

Although the cause of colorectal cancer remains unclear, it has been strongly associated with gut bacterial communities (5–8). This association has allowed researchers to leverage bacterial community signatures as biomarkers to provide accurate, noninvasive colorectal cancer detection from stool (7, 9). 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 gut microbiome. Viruses are another important component of the gut microbial community that have yet to be studied in the context of colorectal cancer.

Due to their mutagenic abilities and their propensity for functional manipulation, human viruses are strongly associated with, and in many cases cause, cancer (10–13). Additionally, because bacteriophages are crucial for bacterial community stability and composition (14–16), and because bacteria have been implicated as oncogenic agents (7, 8, 17), bacteriophages are potentially indirectly linked to cancer. The gut virome (the virus community of the gut) has the potential to impact health and disease (e.g. cancer), and has been associated with diseases including periodontal disease (18), HIV (19), antibiotic exposure (20, 21), urinary tract infections (22), and inflammatory bowel disease (23). We aim to take this field of research further by beginning to assess the role of the virome in colorectal cancer.

Here we present a study of the colorectal cancer virome, highlighting the viruses most associated with the cancer state and the virome's utility for prognosis and diagnosis. We report that, like the association between the bacterial community and colorectal cancer was driven primarily by *Fusobacterium*, the association between the virome and colorectal cancer was driven by broadly infectious, bacteriophage hubs within the phage-bacteria ecosystem. By creating effective classification models using virus community signatures, we were able to accurately classify stool samples as cancerous, precancerous, or healthy. The implications of these findings 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. We expect this study will provide opportunities for continued research into the role of the virome in human cancer development.

Results

The Colorectal Cancer Virome Cohort

The study cohort consisted of 90 human subjects, 30 of which served as healthy controls, 30 of which had adenoma lesions consistent with a precancerous state, and 30 which had carcinoma lesions consistent with colorectal cancer **(Figure 1)**. Half of the stool was used to sequence the bacterial communities using both 16S rRNA gene and shotgun sequencing techniques. The 16S rRNA gene sequences were reported in a previous publication but re-analyzed here (7). The other half of the stool samples were purified for virus like particles (VLPs) and consequent genomic DNA extraction, followed by shotgun metagenomic sequencing. The VLP purification allowed us to observe the *active virome* because we only sequence those viruses that are encapsulated.

Virus DNA was purified by re-suspending the stool in buffer and removing contaminating cells (e.g. human, bacteria, etc) by filtration followed by contaminating cell lysis and degradation of the released genomic DNA **(Figure 1)**. The resulting genomic DNA was used to prepare a shotgun metagenomic sequencing library that was sequenced on the Illumina HiSeq4000 platform. Each run was performed with a blank control to detect any contaminants from reagents. Only one of the nine viral controls detected DNA, which was of a minimal concentration, providing initial evidence of successful sequencing of VLP genomic DNA over potential contaminants **(Figure S1 A)**. As was expected, these controls were sparsely sequenced and were mostly removed while sub-sampling to even depths **(Figure S1 B)**.

The high quality phage and bacterial sequences were assembled into highly covered contigs longer than 1kb **(Figure S2)**. Because contigs only represent genome fragments, we further clustered related contigs into operational genomic units (OGUs) with the majority containing hundreds of related contigs **(Figure S2 - S3)**. These operational units, which are conceptually similar to the operational taxonomic units (OTUs) used in 16S analysis, allow us to study bacterial and phage entities as highly related genomic entities.

Diversity is Insufficient for Virome-Based Cancer Classification

Microbiome and disease associations are often described as being of an altered diversity (i.e. "dysbiosis"). We therefore initially evaluated the diversity of virome OGUs and their association with colorectal cancer. We utilized the Bray-Curtis dissimilarity metric to evaluate the differences in communities between disease states. To control for uneven sequencing depths, we subsampled to a minimum depth that maintained most samples while excluding the sparsely sequenced negative controls.

There was no statistically significant clustering of the disease groups, as visualized by NMDS ordination (Anosim p-value = 0.432, **Figure S4 A**). An Anosim test with a post hoc multivariate Tukey test was used to calculate the statistical significance of the differences between the disease groups based on the variance around the cluster centroids **(Figure S4 B**). There were no significant differences between the disease groups, although there was a strongly significant difference between the blank controls (those few that remained after sub-sampling) and the rest of the study groups, further supporting the quality of our sample set (Anosim p-value = 7.18 × 10-28, **Figure S7)**.

In addition to beta diversity, we also calculated the differences in virus alpha diversity associated with colorectal cancer. Again we found no significant alterations in either Shannon entropy or richness of the virus communities **(Figure S4 C-D)**. Overall, standard diversity metrics were insufficient for capturing the differences in the virus communities between disease states. This suggested to us that a more sophisticated approach for understanding the microbial community may be required, such as machine learning classification.

Virome-Based Machine Learning Provides Accurate Cancer Classification

Previous work has shown that 16S rRNA community signatures are effective for classifying stool samples as originating from healthy, precancerous, or cancerous individuals (7, 9). This is valuable because it presents a potential alternative screening approach to an invasive and expensive colonoscopy. The exceptional performance of bacterial signatures in these predictive models also supports a role for bacteria in colorectal cancer. Here we built off of these findings by evaluating the ability of virus community signatures to classify stool samples and compared performance to models built using bacterial community signatures.

We built and tested random forest models to classify stool samples as belonging to either cancerous or healthy individuals. These models were based on virus metagenomic community signatures or bacterial 16S rRNA gene signatures. We also included a whole metagenomic sequence set to ensure viral metagenomic observations were not a trait of metagenomics in general. Each operational units' relative abundance was used in the feature set. The same model approach was used for all three datasets, and the only difference was the data used to learn the model. We confirmed that our model using bacterial 16S data replicated the performance of the original report which used logit models instead of random forest models **(Figure 2 A)** (7).

We compared the bacterial 16S rRNA gene model to a model built using the virome signatures. The viral model performed significantly slightly better than the bacterial model (corrected p-value = 0.00236) with the viral and bacterial models achieving mean AUC (area under the curve) values of 0.825 and 0.799, respectively, after fifteen random forest iterations **(Figure 2 A - B)**. To confirm that this observation was due to the virome signature itself, and was not a trait of metagenomic datasets in general, we built a model using whole metagenomic community signatures. This model performed poorly with a mean AUC of 0.49, lending support to our observation that classification performance is a trait unique to the virus metagenome **(Figure 2 A - B)**.

To evaluate the synergistic capabilities of the bacterial and viral signatures within the model, we built a combinatory model using both bacteria community and virome data. The combination model yielded modest but significantly improved performance beyond the virome (corrected p-value = 0.00999) and bacterial (corrected p-value = 2.32 × 10-6) models, yielding an AUC of 0.85 **(Figure 2 A - B)**. This suggests the virus and bacterial communities may have synergistic capabilities for classifying stool as belonging to cancerous individuals.

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 iteratively removed. *Fusobacterium* was the primary driver of the bacterial association with colorectal cancer, which is consistent with it's previously described oncogenic potential **(Figure 2 C)**(5). The virome signature was also driven primarily by a few operational genomic units, suggesting a role in cancer development **(Figure 2 D)**. The identified important 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 viral dark matter (uncharacterized virus genomic space) associated with the human virome. This is common in the virome, and 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, two bacterial and two viral organisms primarily drove the community association with cancer **(Figure 2 E)**. The most important microbes were two unidentified viruses, followed by *Bacteroides* and *Fusobacterium*.

The Gut Virome Classifies Cancerous, Pre-Cancerous, and Healthy States

After evaluating our ability to classify samples as cancerous or healthy, we incorporated the precancerous adenoma samples into the model and evaluated our ability to classify all three states in our total sample set. We used three-class random forest models for the bacterial 16S and viral sample sets. The bacterial signature model yielded an AUC of 0.779 and outperformed the viral community model which yielded an overall AUC of 0.698 (p-value = 1.08 × 10-5, **Figure 3 A-C**). Both models were best able to classify cancer samples from healthy or precancerous samples, but struggled to distinguish precancerous from healthy or cancerous **(Figure 3 A-B)**. The cancerous signal was the most discriminatory of the three sample types.

Many of the microbes important for the two-class (cancer vs healthy) bacteria and virus models were also important for the three-class model **(Figure 3 D-E)**. The most important bacterium was the same *Fusobacterium* between the two and three class models, supporting its significance to the association between cancer and the bacterial communities **(Figure 2 C, Figure 3 D)**. Unlike the two-class virome model, the viruses most important to the three-class model were identified bacteriophages **(Figure 2 D, Figure 3 E)**. A *Podoviridae* was most important, followed by two *Siphoviridae* phages.

The classification model determined cancer state by incorporating the relative abundance profiles of the microbes within each community. The signatures ranged from notably high abundance of some OGUs, low abundance of some OGUs, and an absence of other OGUs **(Figure 3 F)**. Not all important OGUs were of increased abundance. The viral classification model depended on the unique signatures of these different abundance profiles to accurately classify each sample.

Bacteriophages Drive Link Between Virome and Colorectal Cancer

The virome-based model was able to accurately classify stool samples as cancerous, precancerous, or healthy. Not only is this important for establishing an alternative diagnostic model, but it also suggests an underlying biological importance for viruses in colorectal cancer. Above we used our classification models to evaluate which virus OGUs were most highly linked to colorectal cancer. We were able to further characterize these important OGUs to better understand the underlying biological link between the virome and colorectal cancer.

The role of the virome in colorectal cancer 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 utilized the longest sequences from the OGUs as the representative sequences to be used for taxonomic classification. These sequences were aligned to a set of all reference virus genomes, including bacteriophages and eukaryotic viruses. A strict e-value threshold of 1e-25 was used to improve our confidence in the matches between the genome sequences. The most important viruses to the classification model were identified as bacteriophages (**Figure 3 E**). Overall we were able to identify 78.8% of the OGUs as known viruses, and 93.8% of those viral OGUs aligned to bacteriophage reference genomes. Thus the majority of the OGUs are bacteriophages and not eukaryotic viruses, indicating the association between the virome and colorectal cancer is reliant on bacteriophage communities. This is consistent with previous reports suggesting the gut virome is primarily phages (19, 23, 26).

Broadly Infectious Phages Play Greater Role in Colorectal Cancer

Reference-based analyses of the virome perform poorly due to sparse reference databases and genomic modularity. Much of the virome is described as genomic “dark matter” and taxonomic identities can be largely uninformative (e.g. *Siphoviridae* describes the phage morphology under an electron microscope). Sometimes these identities are also used to infer bacterial hosts. Instead of relying on genome alignments to infer the identities and host ranges of the phage OGUs, we employed a network-based technique to understand the roles of the OGUs in the greater microbiome context, as previously described (cite other network preprint here). We implemented a random forest model to predict which bacterial OGUs are infected by which phage OGUs, resulting in an ecological network of the bacteria and phages within the community **(Figure 4 A)**. We calculated the alpha centrality (measure of importance in the ecosystem network) of each phage OGU’s connection to the rest of the network, and compared the centrality to the importance of each OGU in the colorectal cancer classification model. We found that phage OGU centrality is significantly positively correlated with importance to the disease model (p-value = 0.0173, rho = 0.14), indicating that phage network hubs may play more important roles in colorectal cancer **(Figure 4 B)**.

Discussion

Because of their propensity for mutagenesis and capacity for modulating their host functionality, many viruses are oncogenic (10–13). Because some bacteria also have oncogenic properties, bacteriophages may play an indirect role in promoting carcinogenesis by altering bacterial communities (7, 8, 17). Despite their carcinogenic potential, the link between virus gut communities (the human gut virome) and cancer has yet to be evaluated. Here we show that, like gut bacterial communities, the gut virome is associated with colorectal cancer. Our findings support a model for oncogenesis by phage-modulated bacterial community composition.

Previous work has supported a causative role for bacterial community composition and colorectal cancer progression, as well as an indirect role for those antibiotic agents which alter that composition (6). Zackular *et al* showed that administering antibiotics (A) and thereby killing subsets of bacterial communities (B) was sufficient for altering colorectal cancer progression (C) in an inflammation-based murine model for colorectal cancer. These findings established a model of cancer dependence on altered bacterial community composition, which in turn relied on antibiotic administration.

$$A \rightarrow B \rightarrow C$$

In addition to bacterial communities, our findings highlight a strong link between gut virus communities and colorectal cancer. The signal is dominated by bacteriophages, suggesting the virome may indirectly promote cancer by modulating bacterial community composition within the gut. These results have led us to suggest an alternative model for microbiome-driven colorectal cancer that depends on bacteriophages (P), instead of antibiotics, as community modulators.

$$P \rightarrow B \rightarrow C$$

Our work supports a model in which bacteriophage communities play a role in colorectal cancer progression by altering bacterial community composition, which is already known to impact colorectal cancer. Alpha and beta diversity of the phage communities were not different between cancer states, but a more sophisticated random forest approach suggested that abundance relationships between phage taxa are highly linked to cancer status. Some select viral genomic units (phages) stood out as being highly important to the cancer-phage link, and this importance was correlated with centrality in the bacteria-phage ecological network. This suggests a model in which the phages most associated with colorectal cancer are also hubs within the bacteria-phage ecological network, thus allowing their abundance changes at the center of the interaction network to ripple throughout the rest of the community and promote the cancer-related ecosystem **(Figure 5)**.

When interpreting these results, it is worth noting that we are studying these communities as operational units, making us unable to draw conclusions with a high taxonomic resolution. This is especially true for bacteriophages, which are poorly annotated and have taxonomy of limited utility. Bacterial operational taxonomic units often represent relatively large categories of bacterial taxa, such as genera, that can represent a variety of species and strains. Likewise, metagenomic operational genomic units represent operationally informative classifications for understanding the community, but should not be interpreted as single phages or bacterial strains.

Further work will be required to support the potential causal relationships outlined in this study. Although we provide evidence for gut bacteriophage communities to alter bacterial communities toward a cancerous state, we must follow up with experimental validation of these observations. The clear pathway toward these answers will be the utilization of murine models as described by Zackular *et al* which showed the link between antibiotic administration and colorectal cancer development (6).

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 contributer 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 (15). 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 and Makefile are available for review at the following GitHub repository: https://github.com/SchlossLab/Hannigan-2016-ColonCancerVirome.

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 (7). 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 Data Acquisition & Processing

The 16S rRNA gene sequences associated with this study were previously reported (7). 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 analytical toolkit (v1.37.0) (27, 28). Briefly, the sequences were de-replicated, screened for chimeras using UCHIME (29) and the SILVA database (30), and binned into operational taxonomic units (OTUS) using a 97% similarity threshold. Abundance was normalized for uneven sequencing depth by randomly sub-sampling to 10,000 sequences, as previously reported (9).

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 (~0.1g) was resuspended in SM buffer and vortexed to facilitate resuspension. The resuspended stool was centrifuged to remove major particulate debris, followed by filtering through a 0.22µm 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 (31). 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 (32). Reads mapping to the human genome were removed using the DeconSeq algorithm (v0.4.3) and default parameters (33).

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 (34). 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 (35). 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 1e6 and 5e5 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 (OGUs) are used as an operational definition of similar 16S rRNA gene sequences in absence of taxonomic identification, we operationally defined closely related contig sequences as operational genomic units (OGUs) in the absence of taxonomic identity. OGUs 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 (36). 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 sub-sampled down 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 [(37).

Classification Modeling

Classification modeling was performed in R using the Caret package (38). 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

Viral operational genomic units (OGUs) 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) (39). 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

Phage OGU replication sytle was identified using methods described previously (25, 26, 40). Briefly, we identified lysogenic phage OGUs 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 OGU 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

We thank the members of the Schloss lab for their underlying contributions. GD Hannigan was supported in part by the Michigan Molecular Mechanisms of Microbial Pathogenesis Fellowship.

396 # Figures

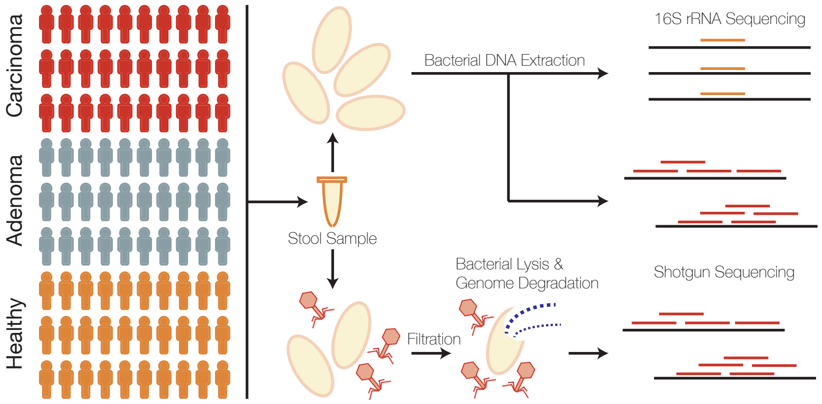

Figure 1: *Cohort and sample processing outline. Thirty subject stool samples were collected from healthy, adenoma (pre-cancer), and carcinoma (cancer) patients. Stool samples were split into two aliquots, the first of which was used for bacterial sequencing and the second which was used for virus sequencing. 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.*

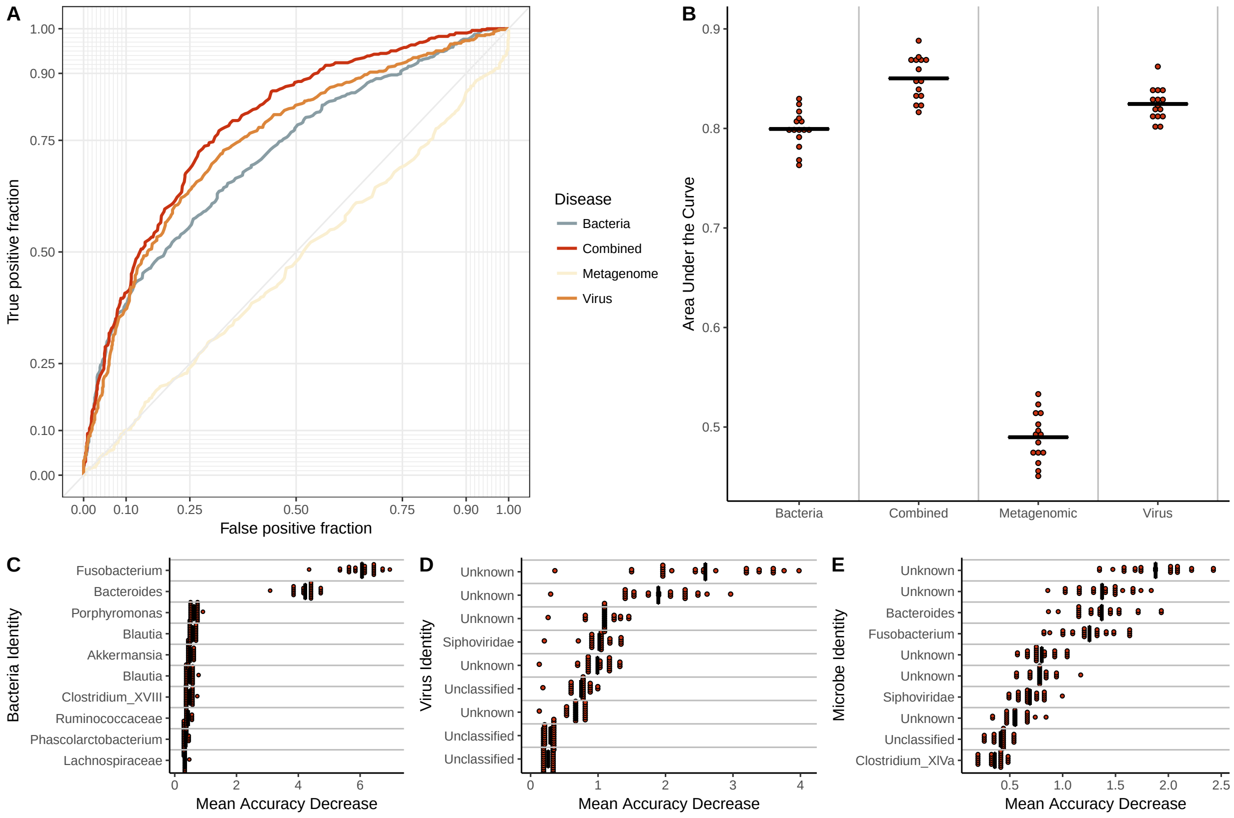

Figure 2: *Results from healthy vs cancer classification models built using virome signatures, bacterial 16S signatures, whole metagenomic signatures, and a combination of virome and 16S signatures. A) ROC curve for visualizing the performance of each of the models for classifying stool as coming from either a cancerous or healthy individual. B) Quantification of the AUC variation for each model, and how it compares to each of the other models based on 15 iterations. A pairwise Wilcoxon test with a Bonferroni 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 classification model when removed from the classification model. Results based on 25 iterations. OTU features are colored by taxonomic identity. 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 and virome features.*

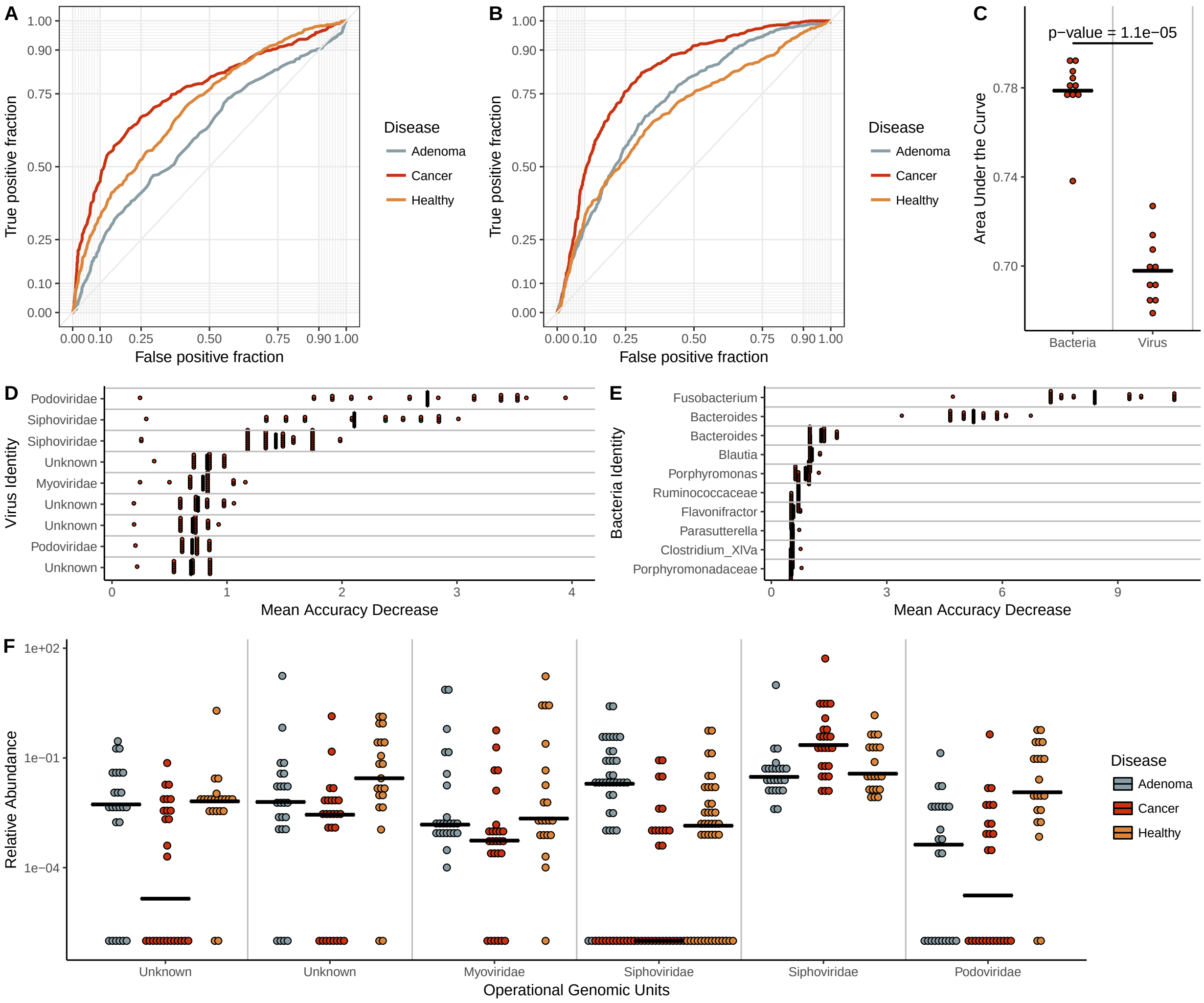

Figure 3: *ROC curves from A) virome and B) bacterial 16S three-class random forest models tuned on mean AUC. Each curve represents the ability of the specified class to be classified against the other two classes. C) Quantification of the mean AUC variation for each model based on 10 model iterations. 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 OGUs in the model, with the most important on the right. Line indicates abundance mean.*

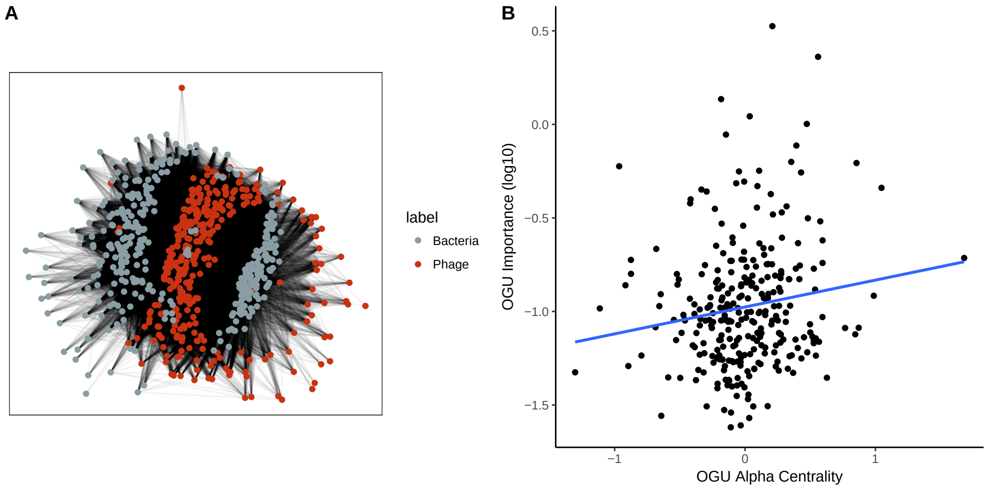

Figure 4: *Community network analysis utilizing predicted interactions between bacteria and phage operational genomic units. A) Visualization of the community network for our colorectal cancer cohort. B) Scatter plot illustrating the correlation between importance (mean decrease in accuracy) and the degree of centrality for each OGU. 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, rho = 0.14).*

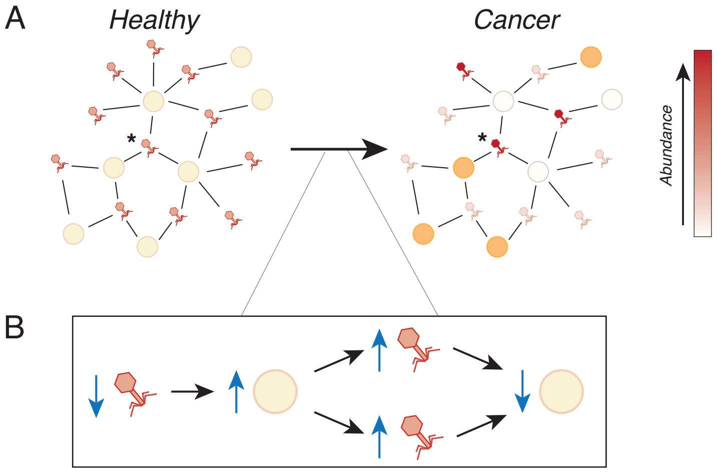

Figure 5: *Example visualization of the proposed model for bacteriophage modulation of colorectal cancer. A) Example network diagrams illustrating the impact of changes on keystone phages (* asterisk) within the microbiome. A hypothetical ecological network may being at a healthy equilibrium but an alteration to the abundance of a keystone phage can impact the remainder of the network. Highly connected phage host abundances may be altered by the original disruption, which would ripple out to the remainder of the model. B) Another visualization of the impact a change in keystone phage abundance could have on the remainder of the network.*

Supplemental Figures

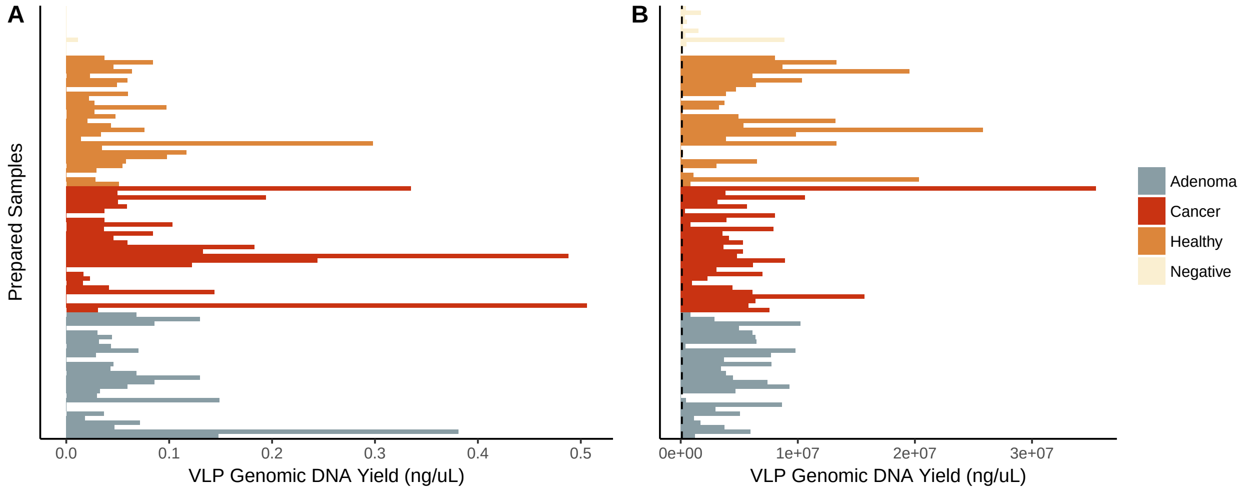

Figure S1: *Basic Quality Control Metrics. A) VLP genomic DNA yield from all sequenced samples. Each bar represents a sample which is grouped and colored by its associated disease group. B) Sequence yield following quality control including quality score filtering and human decontamination.*

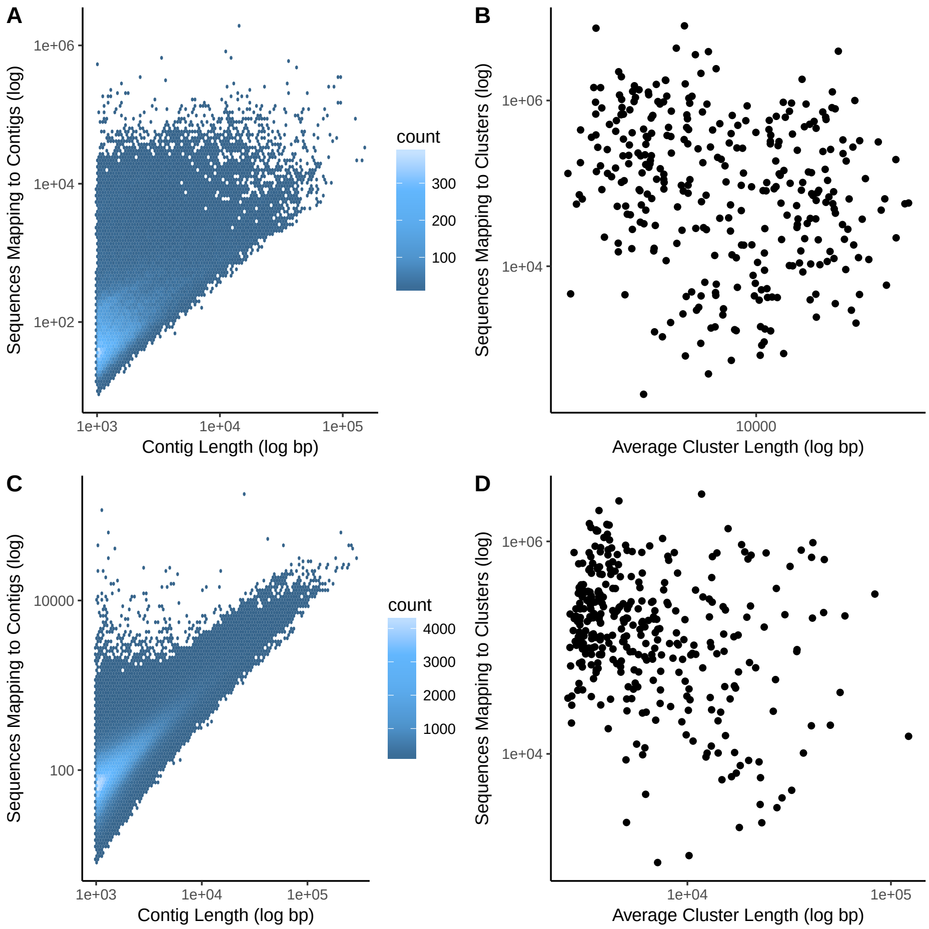

Figure S2: *Length and coverage statistics. A) Heated scatter plot demonstrating the distribution of contig coverage (number of sequences mapping to each contig) and contig length for the virus metagenomic sample set. B) Scatter plot illustrating the distribution of operational genomic unit (OGU) length 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.*

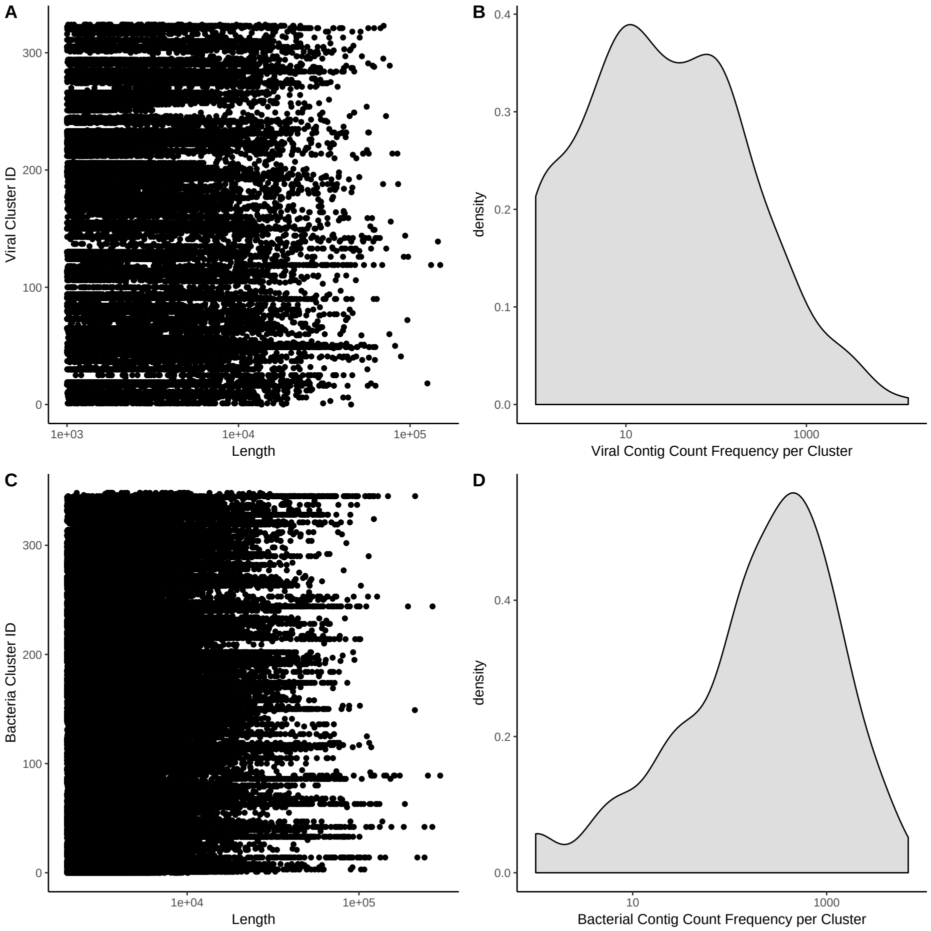

Figure S3: *Operational genomic unit composition stats. A) Strip chart demonstrating the length and frequency of contigs within each operational genomic unit of the virome sample set. The y-axis is the operational genomic unit identifier, and x-axis is the length of each contig, and each dot represents 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.*

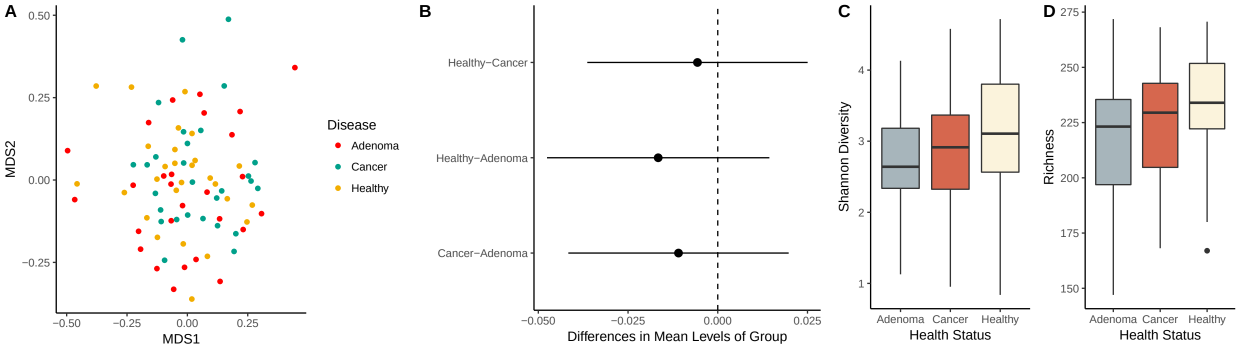

Figure S4: *Diversity calculations comparing cancer states of the colorectal virome, based on relative abundance of operational genomic units in each sample. A) NMDS ordination of community samples, colored for cancerous (green), pre-cancerous (red), and healthy (yellow). B) Differences in means between 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.*

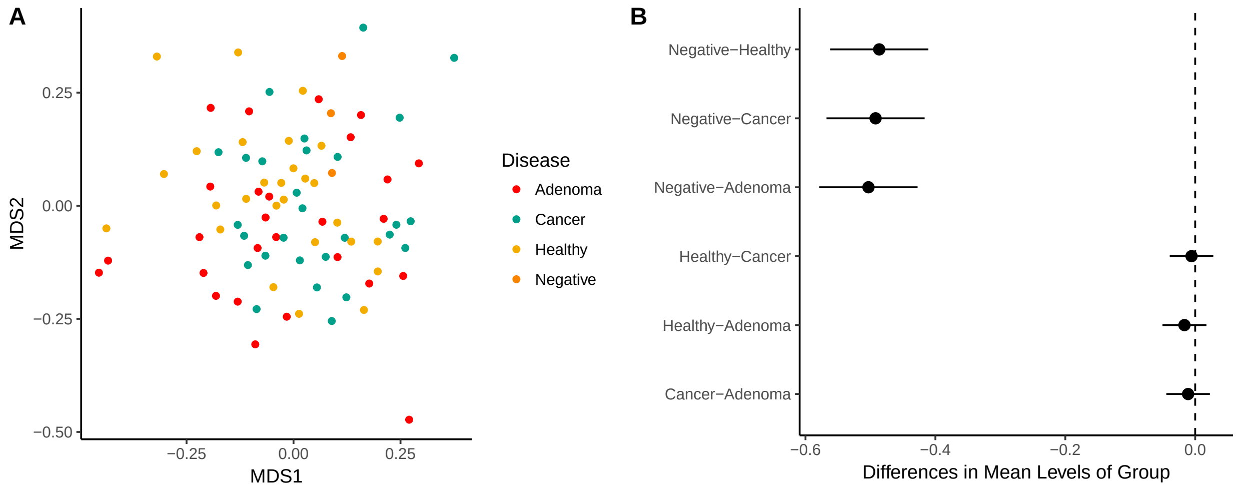

Figure S5: *Beta-diversity comparing disease states and the study negative controls. A) NMDS ordination of community samples, colored by disease state. B) Differences in means between disease group centroids with 95% confidence intervals based on an Anosim test with a post hoc multivariate Tukey test. Comparisons in which the intervals cross the zero mean difference line (dashed line) were not significantly different.*

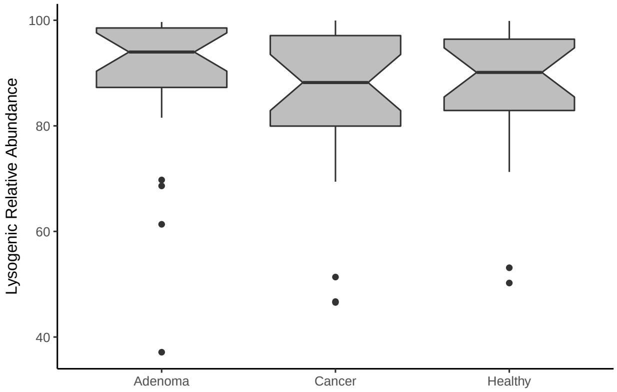

Figure S6: *Lysogenic phage relative abundance in disease states. Phage OGUs were predicted to be either lytic or lysogenic, and the relative abundance of lysogenic phages was quantified and represented as a boxplot. No disease groups were statistically significant.*

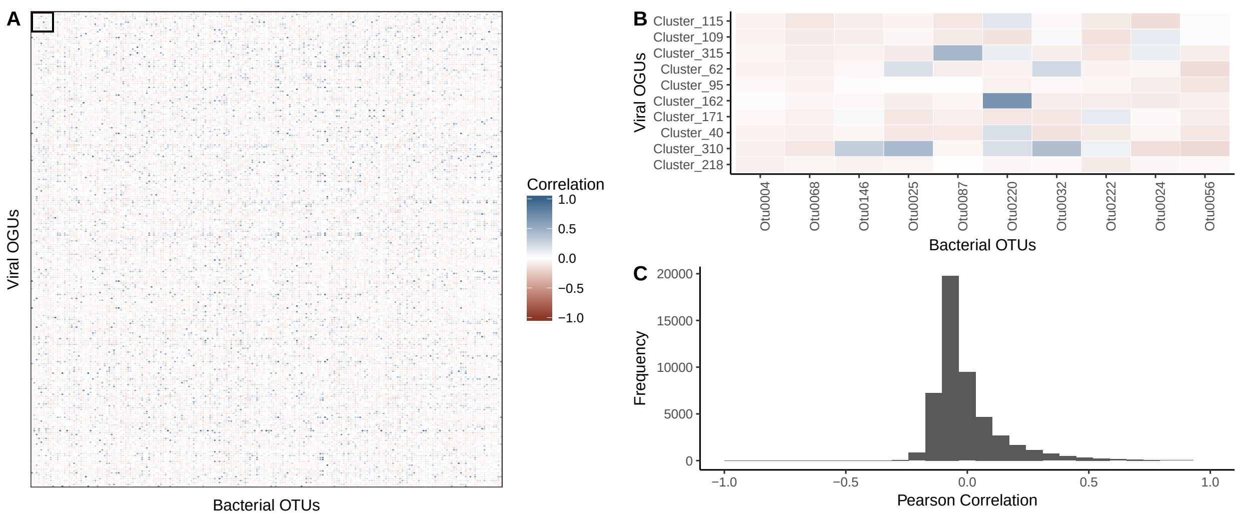

Figure S7: *Relative abundance correlations between bacterial OTUs and virome OGUs. A) Pearson correlation R values between all bacterial OTUs (x-axis) and viral OGUs (y-axis) with blue being positively correlated and red being negatively correlated. Operational units are organized by importance in their 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 OGUs. 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 OGUs.*

References

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

2. 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.

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

4. 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.

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

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

7. 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.

8. 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.

9. 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.

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

11. 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.

12. Schiller JT, Castellsagué X, Garland SM (2012) A review of clinical trials of human papillomavirus

prophylactic vaccines. *Vaccine* 30 Suppl 5:F123–38.

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

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

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

16. 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.

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

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

19. 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.

20. 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.

21. 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.

22. 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.

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

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. Sze MA, Schloss PD (2016) Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. *mBio* 7(4):e01018–16.

28. 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.

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

30. 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.

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

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

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

34. 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.

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

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

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

38. Kuhn M caret: Classification and Regression Training.

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

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